The Institute For Research In Cognitive Science

Proceedings of the IRCS Workshop on Prosody in Natural Speech

University of Pennsylvania
3401 Walnut St., Suite 400C
Philadelphia, PA 19104-6228

P E N N

August 5--12, 1992

Site of the NSF Science and Technology Center for Research in Cognitive Science

IRCS REPORT NO.: 92--37

THE EPISODE IN NARRATION: THE INTERACTION OF PROSODY AND DISCOURSE MARKERS

Janet Bing

Old Domion University
Norfolk, VA 23529-0078
(JMB100F@ODUVM.BITNET)

ABSTRACT

The proposed algorithm segments a narrative into episodes (units larger than sentences) using both discourse markers and prosodic cues: fundamental frequency, pause, and declination. The episodes identified are not merely the result of physiological needs, but are thematically unified. In addition to identifying episodes, a combination of prosodic cues and discourse markers also identifies the major divisions of the narrative.

1. EPISODES AS PROSODIC UNITS

This paper presents a line-by-line algorithm which uses prosodic cues and discourse markers to identify the boundaries of prosodic units. Lehiste [1], Brown and Yule [2], Kumpf [3], Chafe [4], Coulthard and Brazil [5], and Schuetze-Coburn et al. [6] have all discussed how pause, high and low fundamental frequency, and declination serve as cues to the boundaries of prosodic units larger than sentences. Discourse analyses by Schiffrin [7], Polanyi and Martin [8], Hirshberg and Litman [9], and others have shown how discourse markers (hesitation forms, clue words, cue phrases, particles) signal discourse structure. The proposed algorithm assumes that no single cue is sufficient; declination, pause, phrase-final lowering and discourse markers all interact to organize a narrative into prosodic units.

I propose that these prosodic units are narrative units with thematic unity similar to van Dijk's [10] *episodes* and I have adopted this term. Using the proposed algorithm to identify episodes in a narrative collected by Mary O'Mally[1], I apply a procedure from Polanyi [11] to show that these units reflect the structure of that narrative and are not simply the result of physiological factors. The text of the narrative was originally transcribed and divided into lines occurring between pauses of .3 seconds or more. Using phonetic data like that in Figure 1, the text was annotated giving:

(1) a. the highest fundamental frequency (f_o) in brackets at the beginning of the line;

b. the terminal f_o at the end of the line in brackets;

c. the length of the pause (in seconds) in parentheses.

For example, the information in Figure 1 is represented as in (2).

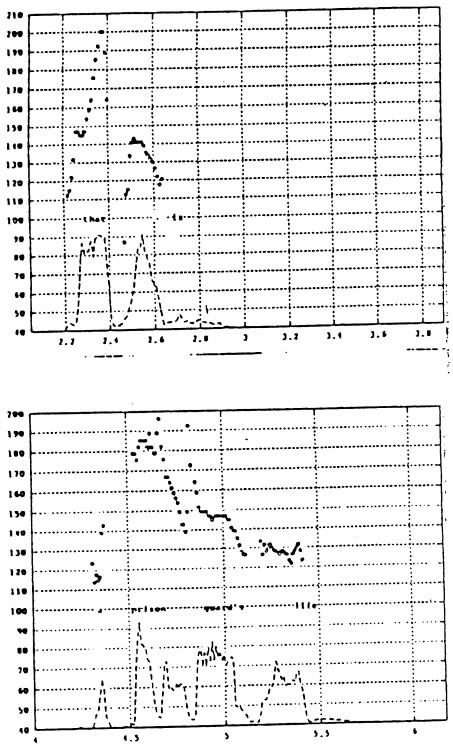


Figure 1

(2) [200] that is [120] (1.6)

[190] a prison guard's life [130] (1.)

Assuming that the first episode begins with the first line of text, each subsequent line is tested as a possible closing line. Excluding lines which contain only discourse markers, a line is judged to be the final line if it meets three or more of the following criteria:

(3) a. the following text line (n+1) contains a higher f_o than the line being tested (n);

b. the line (n) ends at 105 hz. or below;

c. the line (n) is followed by a pause of at least 1 second;[2]

d. the line (n) is followed by one or more discourse markers (n+1=d.m) (uh, and uh, well, so, but, but uh or a repetition) plus a pause of at least .2 seconds.

Excluding discourse markers, the line following the last line of one episode becomes the first line of the next episode.

The use of an algorithm which requires three out of four characteristics is not as ad hoc as it may seem, because prosody and discourse markers, including the prosody on the discourse markers themselves, can serve the same function. Pause and discourse markers followed by pause both give the narrator time to plan the next episode. An episode which does not fall to 105 hz. may be followed by a discourse marker which does. A resetting of register or pitch within an episode is usually the signal that a new episode has begun, but it may be the resetting after a parenthetical or direct quotation.

The proposed algorithm segments the text, "Bread," into episodes, as shown in (4). Each new episode identified by (3) is assigned a letter; a later revision will add other divisions; these are marked with letters and one or more primes. Lines which serve only as discourse markers are identified as **d-**. The symbol **?** indicates rising intonation. In lines where the highest f_o occurs on a word other than the first content word in the line, the word with the highest f_o is underlined. The **{?}** indicates where the interviewer has asked a question which cannot be heard on the tape; the following line is the narrator's response. The original text including the line divisions is not changed, but in a few cases phonetic data is missing.

(4)

A.1. [200] that is [120] (1.6)
2. [190] a prison guard's life [130] (1.)
3. [160] if I had the talent for writing [80] (1.0)
4. [150] you could [100] (2.2)
5. {?}
d- 6. [130] uh [100] (1.8)

B. 7. [170] not quite fifteen years [90] (2.2)

C. 8. [200] if I had the talents for writing [100](.8)
9. [160] I got more stories [105] (1.0)
10.[140] from [95] (.5)
11.[150] guys that I was good to [95] (1.1)
d-12.[120] and uh [100](.9)

D.13.[150] you know heard about [110] (2.)
14.[150] how they got in trouble and uh [100] (.4)
15.[150] how their lifes changed and so forth [100] (2.2)

E.16.[290] and then of course [180] (.7)
17.[240] you bump into people in [160] (.5)
18.[230] in prisons that uh [150] (2.5)
19.[222] just no matter how good you are to them [160] (2.5)
21.[170] they just don't understand it they would uh [120] (.4)
22.[160] they go out of the- out of their way to hurt you [110] (1.)
23.[140] like uh there was uh [100] (3.0)

F.24.[180] a Puerto Rican [90] (.4)
25.[130] we had [100] (1.0)
d-26.[130] and uh [100] (4.4)

G.27.[190] the bread that they had in jail was terrible [95] (2.2)
28. {?}
d-29.[110] bread yeah [80](1.0)

H.30.[120] I uh (3.0)
31.[115] I couldn't eat a whole piece of that bread it was so (1.0) [100]
32.[110] so awful (.8) [80]

I. 33.[120] uh [90] (.2) [225] I could get [130] (2.1)
34.[210]I could buy my lunch [120] (.9)
35.[205] for about a quarter at that time (2.) [90]
37.[160] save me the trouble of carrying it up there and everything [140] (.6)
d-.38.[150] but [180] (1.1)

I'. 39.[240] (1.) I like a couple of pieces of bread with my lunch (.4) [90]
40.[240] well I couldn't stand that bread [80] (1.4)
d-41.[155] so [112](1.6)

J. 42.[220] I used to uh carry my lunch(.8) [100]
43.[170] my wife would [115] (1.0)
44.[170] pack me a couple of sandwiches [150]_(1.)
d-45.[150] and uh [110] (2.2)

K.46.[270] the only time I enjoyed the jailhouse lunch [115](.5)
47.[127] was [115] (.2) uh (.9)
48.[150] when uh [115](.5)

49.[220] the cook up there [95] (.6)
50.[170] his name was black [100] (1.5)
51.[160] when he made irish stew [85] (.9)
52.[150] and uh he made [130] (.5)

53.[180] he made terrific stew [90] (1.2)
d-54 .[130] and uh [100] (.5)

L.55.[200] when he had that [130] (1.9)
56.[190] the inmates would [110] (.5)
57.[170] they'd find out about it before me [90] (.8)
d-58. [130] and uh [110] (.8)
59.[160] they'd come and tell me [110] (.5)
60.[160] that there was gonna be irish stew the next day [80] (.6)
d- 61.[130] well [180] (7.5)
d- 62.[130] oh [100] (.2)
d- 63.[160] there was gonna be irish stew that day [85]

M.64.(data missing) well if there was irish stew I'd give them my lunch.

N.65.[135] well I would have [90] (.9)
66.[295] bread from home see [100] (.6)
67.[160] and uh that's what they wanted uh [80] (2.4)
68.[150] the bread mostly [95] (.6)
d-69.[260] and then [140] (.7)

O.70.[245] it got so (1.0) that uh [100] (2.1)
71.[150] I uh [130] (1.5)
72.[240] I would take two three four pieces of bread extra and stick it in my lunch [100] (1.3)
73.[200] and they just [170] (.9)
74.[210] they'd make a sandwich in the kitchen [200](1.6)
75.[220] bring it out [155] (.5)
76.[210] and throw that bread away and use my bread [80] (1.1)
77.[170] because [130] (.6)
78.[150] they hated that bread in there too [105] (1.)
d-79.[175] so uh [130](1.0)

P.80.[232] the reason I'm telling this story [100](.9)
81.[160] about the bread is [120] (1.)
82.[240] this Puerto Rican kid (2.0) [130]

Q.83.[160] I uh
84.(data missing) I gave these other kids some white bread
85.[260] and I didn't have no more [95](.7)
86.[335] I didn't give him any well I didn't have no more I couldn't give him any (1.3) [90]
d-87.[200] and [150] (.7)

R.88.[330] I didn't know he was real mad [100-170] (1.5)
89.[245) he uh [180] (.5)
90.[245] he told these other kids [140] (.9)
91.[170] that uh [160] (.3)
92.[230] he's gonna really uh [130] (.9)
93.[232] get even with me [90] (1.)
d-94.[140] so uh [110](2.1)

S.95.[175] I didn't hear nothing about that [120] (1.0)
96.[220] well the next day [130] (1.5)
97.[170] I uh I was working four to twelve [110] (.6)
98.[220] when I come in [120] (1.)
99.[170] he said uh [110] (1.)
100.[130] uh [120] (1.4)

T.101.[230] the first thing I did when I come in on that shift [130] (.7)
102.[140] was to uh [100](.8)
103.[180] check the coal bin [90] (1.2)
104.[160] see we uh [100](1.8)
105.[170] I had coal I had uh boiler room duties [90] (1.2)
d-106.[140] and uh [105] (.90)

U.107.[180] I checked the coal bin to make sure there was enough coal in there to do for the shift (.6) [110]
108.[160] and uh they knew that was the first thing I did was check the coal bin (.4) [85]
d- 109.[120] well [105] (.3)

V.110.[170] the coal bin was uh (.6) [90]
111.[200] you entered a small door and looked in there [130](.6)
d-112.[120] and uh [100] (1.5)
113.[210] I'd be in there by myself all the time just checking to see how much was in there to see if we needed any [100] (1.0)

W. 116.[220] one of my [120] (1.0)
117.[200] friendly prisoners [130] (.5)
118.[168] run up to me [138] (.3)
119.[138] and said uh [105] (.9)
W' 120.[180] mister thompson? [140] (1.5)
121.[190] so-and-so is waiting for you with a knife in the coal bin [110] (1.3)
W" 122.[300+] I said what ? [300] (1.0)
W" '123.[200] he said don't look at me I'm not supposed to be telling ya [140] (.7)
W""124.[180] I thought how do you like this [80] (1.6)
d- 125.[130] so uh [110] (1.3)

X. 126.[200] well what I should have done [100]
127.(data omitted) was not go in the coal bin
128.[179] and go and get a couple more guards [105] (.7)
129.[181] and take them in with me [100] (.7)
d- 130.[120] and uh [100] (1.7)

Y. 131.[200] well he might have got killed then [100] (1.8)
132.[200] but uh I was just so mad [100] (1.0)
133.[180] when I found out who it was [110] (.5)
134.[220] I had never done anything to him [130] (.9)

135.[200] the only thing I did was run out of bread and didn't have any more [110] (1.6)
136.[220] but I was so mad I thought well I'll go in and see what he's gonna do with that knife [85] (1.2)
d-. 137.[130] well [120] (1.1)

Z. 138.[160] I opened the door and walked in [100] (1.2)
d- 139.[100] and uh [80] (1.4)

AA.140.[150] I didn't see nobody in there [100] (1.4)
d- 141.[110] but uh [100] (2.)

BB. 142.[190] as I strolled [140] (1.1)
143.[168] in further [120]
144.(data missing) fortunately there was a shovel laying there
145.[140] so I had enough sense to pick the shovel up [90] (1.4)
d- 146.[110] and uh [105] (.3)

CC.147.[180] soon after I picked the shovel up he stepped out [95] (3.7)

DD.148.(data missing) and he had this knife
149.[170] you know where they get the knives [100] (2.)
150.[150] ya know the bed spring [120] (2.3)
151.[180] and they take and they cut them off somehow or other or they break them off [90] (1.1)

EE. 152.[190] and they get somebody that works in the machine shop to sharpen it up [100] (1.9)
d- 153.[110] and uh [100] (1.)

FF. 154.[160] that's uh that's a jailhouse knife [100]
155.(data missing) it'll kill you in a minute
156.[170] it's good and sharp and everything they sharpen it up [100](.4)
d-. 157.[232] Well anyhow [115] (1.4)

GG.158.[200] he started cursing me out and telling me how he was gonna cut my throat (.2] [240] and [170]
(data missing) this that and the other thing.

HH.159.[155] I said you are ? [230] (.8)
160.[270] so I said well come on [148] (1.1)
161.[180] he was so mad he didn't notice the shovel in my hand [95]

II. 162.(data missing) so he runs over
163.[170] raises a knife [105] (1.0)
164.[170] and as he raised the knife he got the shovel on his head [85] (1.2)
d- 165.[110] and uh [110] (1.4)

JJ. 166.[180] I banged him a couple of times knocked him down [110] (.6)
167.[240] and I was about to push it into his face but I thought no [100] (.6)
168.[150] it's liable to kill him [90] (1.1)
d- 169.[140] so uh [105] (1.2)

KK.170.[185] he was unconscious then [100] (1.0)

An examination of the proposed episodes suggests that other factors in addition to prosody and discourse markers must be considered. Episodes W and HH both contain direct quotations by new speakers, and it is reasonable to expect that these will begin new episodes, comparable to new paragraphs in written discourse. Expressions such as "he said" and "I said" operate as discourse markers in these cases.

In episode I, the discourse marker but, with its markedly rising intonation, seems to signal a new episode should beginning on line #39, particularly because of the resetting of register on that line. Rising intonation in the upper voice range makes discourse markers more salient, as in lines #16 and then [160-290 hz.], #69 and then [140-260 hz.], and #157 well [120-230 hz.], all of which occur at episode boundaries. Because of the rising intonation, an episode boundary is stipulated at line #39.

The summary chart in (5) shows the characteristics of the episodes identified by the algorithm in (3) plus the proposed revisions. The chart identifies the first and last lines of each episode, the terminal pitch, the length of the pause after the last line, the resetting of pitch within the episode, and the following discourse marker plus fo and pause.

(5) Unit	Line	end fo	pause	reset?	next dm.
A	1-4	100	(2.2)		uh [100] (1.8)
B.	7	90	(2.2)		
C.	8-11	95	(1.1)	yes	and uh [100] (.9)
D.	13-15	100	(2.2)		
E.	16-23	100	(3.0)		
F.	24-25	100	(1.0)		and uh [100] (4.4)
G.	27	95	(2.2)		
H.	30-32	80	(.8)		uh [90] (.2)
I.	33-37	140	(1.1)		but [180] (1.1)
I'	39-40	80	(1.4)		so [112] (1.6)
J.	42-44	150	(1.0)		and uh [110] (2.2)
K.	46-53	90	(1.2)		
L.	55-60	80	(.6)		well/oh [100] (.2)
M.	64	data missing			
N.	65-68	95	(.6)	yes	and then [140] (.7)
O.	70-78	105	(1.0)	yes	so uh (2.1)
P.	80-82	130	(2.0)		(whole unit=d.m.)
Q.	83-86	90	(1.3)	yes	and [150] (.7)

Unit	Line	end fo	pause	reset?	next dm +pause
R.	88-93	90	(1.0)	yes	so uh [90] (2.1)
S.	95-99	110	(1.0)	yes	uh [120] (1.4)
T.	101-105	90	(1.2)	yes	and uh [100] (.9)
U.	107-108	85	(.4)		well [105] (.3)
V.	110-113	100	(1.0)	yes	
W.	116-119	105	(.9)		said uh [105] (.9)
W'	120-121	110	(1.3)		I said
W"	122	300	(1.0)		he said
W"'	123	140	(.7)		I thought
W""	124	80	(1.6)		so uh [110] (1.3)
X.	126-129	100	(.7)		and uh [100] (1.7)
Y.	131-136	85	(1.6)	yes	well [120] (1.1)
Z.	138	100	(1.2)		and uh [80] (1.4)
AA.	140	100	(1.4)		but uh [100] (2.0)
BB.	142-145	90	(1.4)		and uh [105] (.3)
CC.	147	95	data missing		
DD.	148-151	90	(1.1)	yes	
EE.	152	100	(1.9)		and uh [100] (1.0)
FF.	154-156	100	(0.4)		well anyhow (1.4)
GG.	158		data missing		I said
HH.	159	95	data missing		
II.	162-164	85	(1.2)		and uh [110] (1.4)
JJ.	166-168	90	(1.1)	yes	so uh [105] (1.2)
KK.	170	100	(1.0)		so uh [100]

In several cases, because of missing phonetic data the boundary between episodes is assigned based on the evidence available.

2.THE SEMANTIC UNITY OF EPISODES

Having proposed that a revised algorithm can correctly identify most episodes, it is still necessary to show that these units have semantic unity. Grimes [12] and others [13],[2],[14], have proposed a "discourse paragraph," "paratone" or "center of interest" and have assumed that these units larger than sentences have semantic unity. Van Dijk [10:177] defines *episodes* as "coherent sequences of sentences of a discourse, linguistically marked for beginning and/or end, and further defined in terms of some kind of 'thematic unity'--for instance, in terms of identical participants, time, location or global event or action."

However, as Schuetze-Coburn et al. [6:230-31] point out, it is possible that the declination units identified by (3) are merely the result of a speaker's physiological needs or diminishing breath supply. In order to show that the episodes in "Bread" are also discourse units, a procedure for identifying the structure of a narrative was adapted from chapter 2 of Polanyi [11]. First, the underlying complete and incomplete propositions were identified. Based on the identified descriptions and events, each episode was assigned one or more overt or implied topic sentences or topics which generalized the information in the original propositions. For example, the sentence, "N usually checked the coal bin alone," is the topic sentence for the four propositions in (6) which are found in Episode V.

(6) N (usually) enters small door in coal bin.
N looks in coal bin.
N is usually alone.
N usually checks amount of coal.

In addition to the episodes, there are larger prosodic units in "Bread" identified either by extended pauses (three seconds and longer), salient discourse markers (such as "well anyhow") or explicit reference to the structure of the discourse, such as the entire episodes, P and FF. The boundaries of these major divisions are marked with the symbol +++ in (7) and assigned roman numerals and names.

Following Polanyi [11], durative-descriptive propositions are labelled <DD> mainline story event propositions <e1, e2>, and negative events <-e>. Discourse markers and episodes functioning only as discourse operators are classified as operators <o>. Boundaries marked by salient discourse markers or extended pauses are marked by three pluses (+++). The resulting structure of the full narrative is (7):

(7) A. <o>A prison guard's life **I.Orientation**
B. <o> Fifteen years
C. <DD> N, the narrator, has friendly prisoner stories.
D. <DD> N has stories about their troubles and lives.
E. <DD> N also has stories about unfriendly prisoners.
++(3.0)

F. <o> a Puerto Rican **II.Protagonist**
++(4.4)

G. <DD> Prison bread was terrible.**III.Bread**
H. <DD> N couldn't eat the prison bread.
I.. <DD> Prison lunch was cheap and easy.
I' <DD> N liked bread, but not prison bread.
J. <DD> N carried his lunch from home with sandwiches.
K. <DD> The only good prison meal was Irish stew made by prison cook.
L. <DD> Inmates told N when Irish stew was planned.
++(7.5)

IV. N's Kindness
M. <DD> When there was Irish stew N gave prisoners his home lunch.
N. <DD> Inmates wanted N's bread from home.
O. <DD> N gave N's extra bread to inmates.
P. <o> The bread and the Puerto Rican (PR) are related.
++ (4.0)

V. Complication

Q. <e1> N couldn't give home bread to the PR prisoner.
R. <e3> The PR threatened N.
S. <-e> N didn't hear the threat.
<o> The next day N.worked from 4 to 12.
T. <DD> N checked the coal bin first.
U. <DD> Prisoners knew that N checked coal bin first.
V. <DD> N usually checked coal bin alone.
W. <e4-8> A friendly prisoner warned N.
X. <irrealis> N should not have gone alone.
Y. <irrealis> The PR could have been killed.
<DD> N hadn't done anything to hurt the PR.
AA.<e11> N entered the coal bin.
BB.<e16> N picked up a shovel in the coal bin.
CC.<e17> The PR approached N with a prison knife.
+++ (3.7)

VI. Suspension

DD.<DD> Jailhouse knives are broken off bedsprings.
EE.<DD> Jailhouse knives are made in the machine shop.
FF.<o> Jailhouse knives are dangerous.
+++

VII. Resolution

GG.<e18> The PR threatened N.
HH.<e20> The PR didn't see the shovel.
II.<e21> The PR and N fought.
JJ.<e22> N knocked the PR down with the shovel.
KK.<DD> The PR was unconscious.
LL. N called for help.
MM. The PR was punished.
+++

VIII. Coda

NN. The PR didn't return.
OO. Such things happen in prison.

Except for two of the episodes, S and Y, the episodes in "Bread" are consistent with van Dijk's [10] definition of episode. Although some details and events are omitted in (7), the listing of the topic sentences includes all of the essential information and events of the story.

Episode O is a potential counterexample since it appears to violate the unity of a single location. However, because the narrator is talking about a typical pattern of behavior rather than actual events, it is not a real counterexample. However, Episodes S and Y are genuine counterexamples, and episode boundaries are expected at lines #96 and #132. Some of the boundary cues are present. At the end of #95 there is a long pause; line #96 begins with a discourse marker and a shift in register (as well as in time). Similarly, line #131 ends with a long pause and low pitch. Line 132 begins with *but uh* and is preceded by a 1.3 second pause. In both cases, there is reason to suspect a boundary even though no boundary was identified by (3). With more data, it may be possible to refine the algorithm so that the boundaries which probably occur before lines #96 and #132 are identified.

The algorithm in (3) is a first approximation based on limited data. Even though this algorithm twice fails to correctly identify episodes, it is successful enough to suggest that the task of identifying the boundaries of episodes in narrative is possible if both discourse markers and prosodic information are used. The episodes identified by (3) come close to reflecting the structure of the discourse and are not merely arbitrary breath groups. In the narrative, "Bread," not only are there overt cues to episode boundaries, but also identifiable boundaries of larger units; an extended pause (three or more seconds) or a salient discourse marker, such as *well anyway*" almost always signals these boundaries between units that Labov and Waletzky [15] call parts of the "normal form" of a narrative.

3. IMPLICATIONS FOR FURTHER RESEARCH

This preliminary procedure of identifying episode boundaries in a single narrative suggests a number of hypotheses to be tested against more data.

(8) a. Discourse markers followed by pause signal boundaries different from those not followed by pause.

b. Short pauses (.2 seconds or more), long pauses (1.0 second or more) and extended pauses (3.0 seconds or more) have different communicative functions in narration.

c. In narration, the domain of declination is the episode and not the clause or sentence.

If further evidence is found to support the hypothesis in (8), these may or may not hold for other speakers and other types of discourse. The hypothesis in (8c) is particularly interesting, since it challenges Pierrehumbert [16], who claims that declination is a strictly local phenomenom. The hypothesis supports the models proposed by Thorsen [17] and Garding [18], who claim that declination is a global rather than local phenomenom. This issue, particularly, merits further investigation.

Notes:

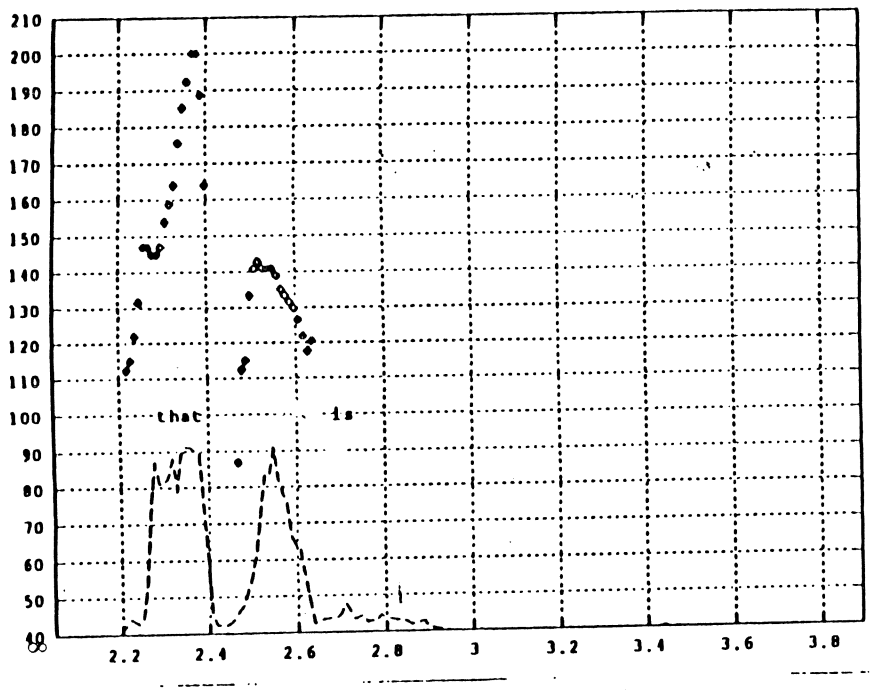
[1]I would like to thank Mary O'Mally and Bill Reynolds, who collected two versions of the narrative "Bread" for a class taught by William Labov, as well as my colleagues Charles Ruhl, John Broderick and Carol Hines for their assistance and suggestions. I am particularly grateful to Mark Liberman for sharing both the text and acoustic data on which this paper is based.

[2]One second as the minimal "long pause" was based on the fact that one second is the median and the mode for all types of pauses for this speaker and is also the definition of long pause used by Brown and Yule [2] and Chafe [14].

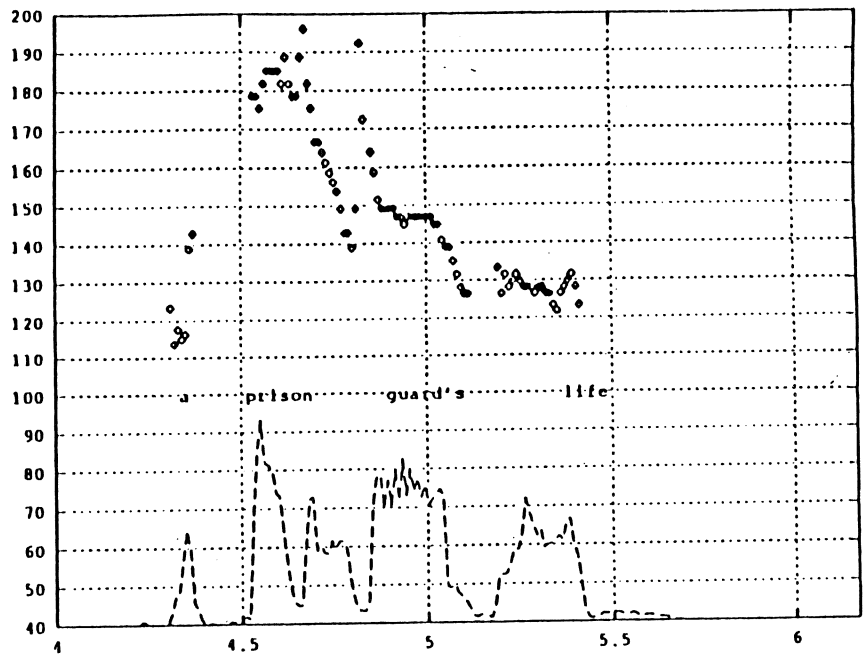
REFERENCES

1. Lehiste, Ilse. "Sentence Boundaries and Paragraph Boundaries--Perceptual Evidence," *The Elements: A Parasession on Linguistic Units and Levels.* Chicago Linguistic Society, Chicago, 1979.
2. Brown, Gillian and George Yule. *Discourse Analysis.* Cambridge University Press, Cambridge, 1983.
3. Kumpf, Lorraine E. "The Use of Pitch Phenomena in the Structuring of Stories," Russell Tomlin (ed.) *Coherence and Grounding in Discourse.* John Benjamins, Amsterdam, 1987.
4. Chafe, Wallace. "The Flow of Thought and the Flow of Language," Talmy Givòn (ed.) *Syntax and Semantics Volume 12: Discourse and Syntax.* Academic Press, New York, 1979.
5. Coulthard, Malcolm and David Brazil. "The Place of Intonation in the Description of Interaction," Deborah Tannen (ed.) *Analyzing Discourse: Text and Talk.* Washington, D.C.: Georgetown U. Press, 1982.
6. Schuetze-Coburn, Stephan, Marian Shapley and Elizabeth Weber. "Units of Intonation in Discourse: A Comparison of Acoustic and Auditory Analyses," *Language and Speech*, 34 (3), 207-234, 1991.
7. Schiffrin, Deborah. *Discourse Markers.* Cambridge: Cambridge University Press, Cambridge, 1987.
8. Polanyi, Livia and Laura Martin. "On the formal treatment of discourse particles: the case of Mocho la in narrative discourse." Unpublished paper, 1990.
9. Hirschberg and Litman. "Now let's talk about now: Identifying cue phrases intonationally," *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 1987.
10. van Dijk, Teun. "Episodes as Units of Discourse Analysis," Deborah Tannen (ed.) *Analyzing Discourse: Text and Talk.* Georgetown U. Press, Washington, D.C., 1981.
11. Polanyi, Livia. *Telling the American Story: A Structural and Cultural Analysis of Conversational Storytelling.* Ablex Publishers, Norwood, NJ, 1975.
12. Grimes, Joseph E. 1975. *The Thread of Discourse,* Mouton, The Hague, 1975.
13. Longacre, Robert E. *The Grammar of Discourse.* Plenum Press, New York, 1983.
14. Chafe, Wallace. "The Deployment of Consciousness in the Production of a Narrative," Wallace Chafe (ed.) *The Pear Stories.* Ablex, Norwood, NJ, 1980.
15. Labov, William and Joshua Waletzky. "Narrative Analysis: Oral Versions of Personal Experience," June Helm (ed.), *Essays on the Verbal and Visual Arts: Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society.* American Ethnological Society, Seattle, 1967.
16. Pierrehumbert, Janet. *The Phonology and Phonetics of English Intonation.* MIT Ph.D. dissertation, 1980.
17. Thorsen, Nina. "Two issues in the prosody of Standard Danish," A. Cutler and D.R. Ladd (eds.) *Prosody: Models and Measurements.* Springer-Verlag, Berlin, 1983.
18. Gårding, Eva. "A Generative Model of Intonation," A. Cutler and D.R. Ladd (eds.) *Prosody: Models and Measurements.* Springer-Verlag, Berlin,1983.

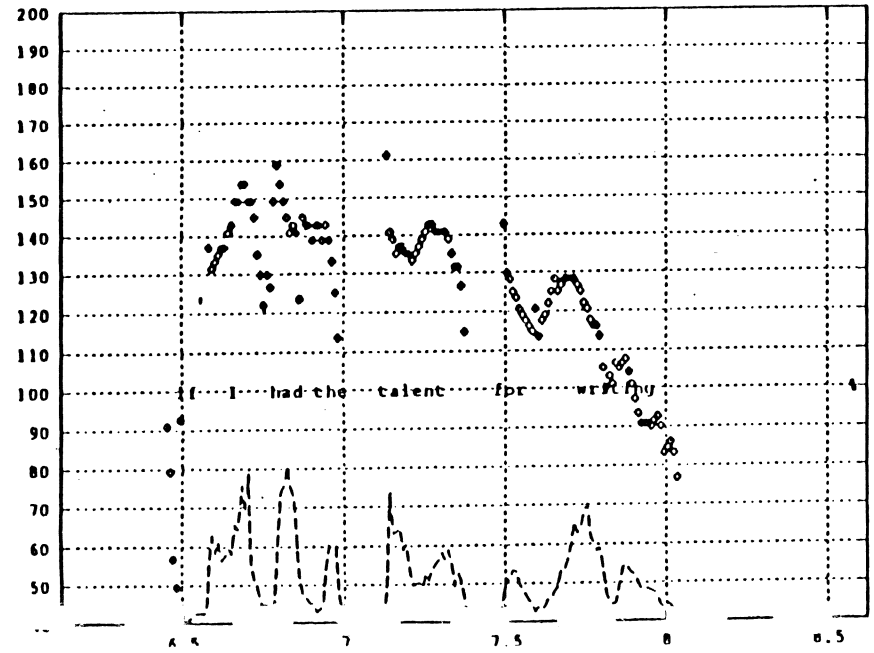
j11



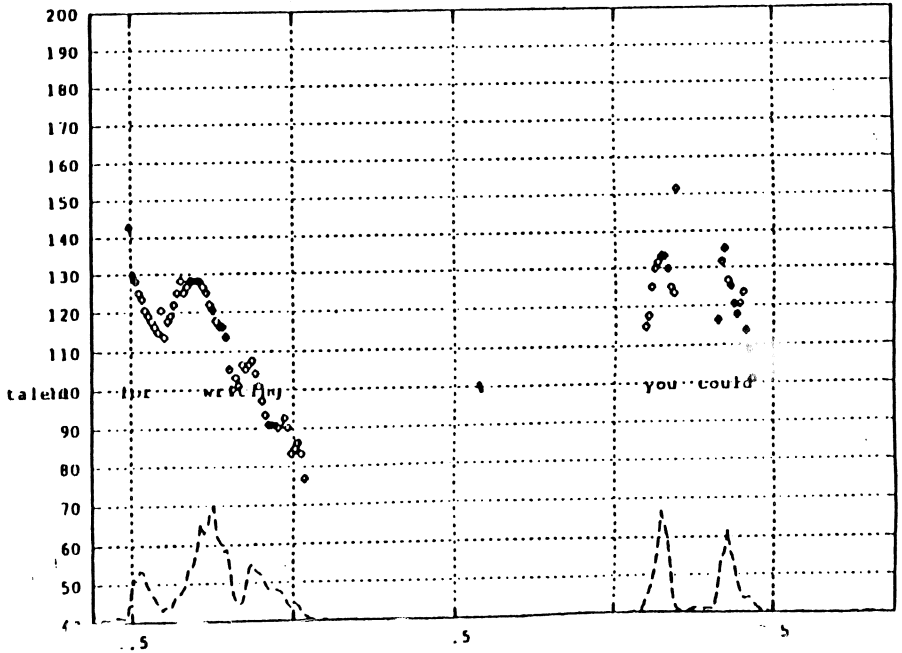
j12



j13 (1)



j13 (2)



Intonation as a Presentational Resource in Conversation

Susan E. Brennan

Psychology Department
State University of New York
Stony Brook, NY 11794
brennan@psych.stanford.edu

Workshop on Prosody in Natural Speech, U. Penn, August 5-12, 1992

KEYWORDS: *Intonation, prosody, mutual knowledge, communication, conversation, collaboration, grounding, backchannels*

1. Communication as collaboration

The collaborative view of language use holds that speakers and addressees are jointly responsible for contributions to conversations (Clark & Wilkes-Gibbs, 1986). They do not simply produce and comprehend utterances autonomously; instead, they achieve a state of mutual knowledge by exchanging evidence that they've understood one another (Brennan, 1990; 1992; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). This process of exchanging evidence is called *grounding*. Evidence of understanding can take many forms, including an appropriate second part in an adjacency pair, such as the answer to a question (Levinson, 1983), a relevant next turn in a conversation (Clark & Schaefer, 1987, 1989), backchannels (Yngve, 1970), or eye contact (Goodwin, 1981). The evidence used to ground utterance meaning can be provided verbally or visually (Brennan, 1990; 1992), and conversation is shaped in part by by the resources available for grounding in a particular communication medium (Clark & Brennan, 1991; Whittaker, Brennan, & Clark, 1991). And conversational partners appear to set higher or lower *grounding criteria* for the amount of evidence they seek and provide before concluding they understand one another well enough for the purpose at hand (Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs, 1986).

In the *contribution model* of Clark and Schaefer (1987, 1989), every contribution to a conversation has two phases: a presentation phase and an acceptance phase. Every utterance is itself a presentation; it does not become a contribution to the conversation until its acceptance phase is complete, that is, until the addressee provides evidence that he believes he understands what the speaker meant, and the original speaker ratifies that evidence. So even though a speaker may have an intention in mind when she presents an utterance, her utterance does not stand as a contribution to the conversation until she has evidence from her addressee that he has understood[1]. For instance, consider this example from the Lund corpus (Svartvik & Quirk, 1980), where A and B are two people engaged in conversation:

```
A: is term OK

B: what

A: is term all *right*

B: *yes* it seems all right so far
   . touch wood
```

Here, A presents an utterance that may have been intended to function as a question. But after A's first turn, B's utterance provides evidence that he does not yet understand. There are of course many possibilities, including these – perhaps he didn't hear her, perhaps he didn't understand what she meant by "OK," perhaps he didn't catch the word "term." A repairs her original presentation by repeating it with a word change - from "OK" to "all right." The end of A's utterance is overlapped by the beginning of B's (overlapping parts are indicated by the asterisks). Since B begins his utterance early, it could be that the problem was with the word "term." In any event, it is not until the fourth turn, when A hears B's relevant response, that she can surmise that her question has been properly understood by B. At this point A can go on with another relevant presentation. In doing so, she communicates to B that she is satisfied that the acceptance phase for the utterance she first presented is complete; after that, a contribution to the conversation has been made.[2]

[1] For clarity, I will use the convention that speakers are female and addressees are male.

[2] See Brennan & Cahn, 1992, for a discussion of the temporal asymmetry in A's and B's roles in producing a contribution to the conversation.

At each moment in a conversation, an individual can provide evidence to a partner, or seek evidence from that partner (Brennan, 1990; 1992). The exchange of evidence happens in a systematic way: A speaker presents constituents of an appropriate size, depending on the grounding criterion and on the communication medium, and an addressee typically provides appropriate evidence of understanding as soon as he concludes that he has understood a speaker's presentation well enough for current purposes. If an addressee believes that he does not understand, he will withhold the expected positive evidence, or provide explicit negative evidence (e.g. with a clarification question, a puzzled frown, etc.). If the expected evidence of understanding from an addressee is not forthcoming, a speaker will pursue it (Brennan, 1990; 1992; Pomerantz, 1984).

In this paper I present some data about the intonational resources that people can use for grounding the meanings of utterances. My claim is that intonation *not only* conveys information about syntactic constituents (Cruttendon, 1986) and the speaker's intention (Sag & Liberman, 1974; Liberman & Sag, 1974; Pierrehumbert & Hirshberg, 1990), *but also* can be used to manage the exchange of evidence between two people in conversation, en route to achieving mutual understanding. In particular, I examine phrase final rising intonation. It has been proposed by some that such intonation serves an interactional purpose (Brennan, 1990; McLemore, 1991), e.g. to elicit the attention of addressees, or to pursue a response. I will bring behavioral evidence to bear on the hypothesis that speakers use rising intonation to actively seek evidence of understanding from their addressees.

2. Intonation as a presentational resource

My corpus consists of stereo recordings of conversations about map locations. Pairs of people did a matching task using pictures of the same map displayed on two computers networked together. Since the task was to get both of their cursors located in the same target location on the map, the degree to which they understood one another was indexed by the distance between their cursors. Throughout their conversation, a log was kept of cursor position, and this log was later synchronized with the conversational transcript. This technique enabled a continuous online measure of understanding in conversation.

2.1. Method

Subjects were 24 Stanford graduate students between the ages of 21 and 32, all native speakers of American English. They participated as same-sex pairs who had never met one another before. There were eight women and 16 men, from 13 different academic departments. They were recruited through posted advertisements or electronic bulletin boards, and participated in exchange for a small honorarium.

Pairs of subjects in adjoining cubicles used computers networked together to do a matching task. Each partner was seated in front of an identical computer graphics display of a map. Each display had a small icon of a car. The task was for one person (the director, D) to convey a target location to the other (the matcher, M), and for the matcher to position a car icon in the target location by dragging it with his mouse. The task was done 80 times by each pair. They could talk to each other as much as they liked, but they could not see each other.

There were two experimental conditions, *visual evidence* and *verbal-only evidence*. That is, in half of the trials, the director could see the position of the matcher's icon on the screen, and so had visual evidence of exactly what the matcher understood; in the other half of the trials, there was no visual evidence (the director could see only her own icon). After the matcher "parked" his car in a location by clicking his mouse, the director initiated a new trial by clicking on her icon, which then moved by itself to a new preprogrammed location. Subjects' displays were always identical, except for their icon positionss. Maps of the Stanford University campus and of Cape Cod were used as graphic backgrounds for the trials. After every block of ten trials, the pair of subjects alternated evidence conditions, maps, or director/matcher roles.

2.2. Analysis

Speech transcripts. The conversations of six of the 12 pairs of subjects were chosen randomly for detailed transcription, yielding 480 descriptions of map locations. These descriptions were transcribed in segments that corresponded roughly to one phonemic clause per line (that is, a short sequence of words separated by a pause, and generally containing one primary pitch accent (Rosenfeld, 1987; see also Boomer, 1978; Dittmann & Llewellyn, 1967)). Each line was punctuated in order to categorize its clause-final prosody approximately: "." for final pitch lowering, "?" for final rising, "," for the end of a tone unit (if mid-clause) or else for list-like intonation (when at the end of a clause), "-" for a sudden self-cutoff on a level pitch, and no punctuation for clauses fitting none of these criteria. The clauses sometimes had extreme final lengthening, or drawled syllables, which were denoted by ":" following the letter that

most closely matched the sound being drawn out (ye:s for "yeeees," vs. yes: for "yesss"). Overlapping speech was transcribed using single or double asterisks to enclose the beginning and ending of both stretches of simultaneous talk. Unintelligible speech was enclosed in brackets. All transcripts were double checked for accuracy.

In order to conduct a detailed analysis of individual conversational interchanges, I took a random sample of 48 of the 480 transcribed interchanges. One item (that is, one location on the map) was chosen at random from each cell in the counterbalanced design of the experiment. In this smaller sample, each pair of subjects contributed interchanges concerning the same 8 map locations. I then coded whether or not the director in each interchange used the final rising intonation typical of question intonation in presenting a description, in the intial period before the matcher had made any verbal response.

Action transcripts. During each trial, the x and y coordinates of matcher's icon were recorded and time-stamped, to provide a record of the matcher's understanding of the location of the target. For the small sample of 48 trials, the distance between the matchers' icon and the target location (the director's icon) was plotted over time, to provide a visible display of on-line understanding (convergence between the two icons) in the conversational interchange.

Then the 48 action transcripts were synchronized with the speech transcripts. A naive coder, with copies of the language transcripts in front of her, listened to the tapes of the conversational interchanges using a videotape player that was equipped with a seconds counter. For each of the 48 trials in the sample, she zeroed the counter at the start of the trial (marked by a short beep) and recorded an integer at every 1.0 second interval by writing the integer over its corresponding word on the transcript. It took several passes over the tapes to record the seconds intervals and to check the synchronization of each trial. We estimate that this procedure was accurate well within a half-second. Synchronization of these action transcripts with the speech transcripts is shown using matching superscripts over the speech transcripts and the time-distance plots (see Figure 1).

Did the directors use rising intonation differently in their presentations to matchers when they could see what the matchers were doing vs. when they could not? The collaborative view predicts that they should; the exchange of evidence via intonation should be managed differently in a medium where visual evidence is available than in one where visual evidence is not (Clark & Brennan, 1991). I coded whether or not D used final rising intonation, often associated with questions in English, in any of the descriptions she uttered initially in an interchange, *before* the point where she got a verbal response from M. Then I coded whether or not M had moved his icon before responding verbally or before D's use of question intonation (whichever came first). The expectation was that D could use either M's verbal response, or M's icon movement (in the visual evidence condition) to conclude that M had understood her description of a map location.

2.3. Results and discussion

While the baseline frequency of using final rising intonation was the same across both evidence conditions – that is, directors were just as likely to use final rising intonation before the matcher's first verbal response in the verbal-only condition as they were in the visual condition (58.3 % to 41.7 %, n.s.), final rising intonation was distributed differently in the two evidence conditions. D's use of final rising intonation was related to whether M had moved his icon yet in the visual evidence condition, but not in the verbal-only evidence condition. With visual evidence, there was a correlation between D's use of final rising intonation and M's lack of icon movement was ($r_\phi = .48$, $p < .02$). That is, when the directors could monitor the matchers' icon visually and the matchers hadn't yet made any progress toward the target during the directors' initial descriptions, the directors were likely to use final rising intonation, possibly to pursue a response from the matchers. There was no such systematic relationship in the verbal evidence condition, where the directors weren't able to see whether the matchers had moved yet ($r_\phi = .17$, $p < .50$).

Let us take a closer look at the relationship between D's intonation and M's icon movement. Consider this example where D did *not* use any final rising intonation.

```
Example 1:

D: ok
   now we went
   south,
   we're
   about halfway down the screen
   in the electronics lab,
   we're in the southern most
   wing of the electronics lab
   good
   good

M: I think that's where my office is.
(Pair 3, Location 33, Visual evidence condition)
```

In this example, M began to move his icon immediately after D said "south" and made fairly steady monotonic progress toward the target (except for during the point when D was pronouncing "southern most" – see the time-distance plots for this example and others in the Appendix). M arrived at the target just before D 's second use of the word "lab," and D, since she had visual evidence that M's hypothesis about what she meant was correct, acknowledged this with "good good." A similar pattern was found in the next example (Example #2, Appendix).

Example 2:

D: uhh
Terman Engineering
just to the lower right of the five.

M: ok, Terman?
to the lower right?

D: yah
bingo.

M: right here?

D: right there.
(Pair 4, Location 25, Visual evidence condition)

In this example, M also made early progress toward the target, and D could monitor this and did not use final rising intonation. M sought further evidence about D's meaning by taking a verbal turn. In addition, this pair explicitly acknowledged having visual evidence by using a deictic strategy with: "bingo. - right here? - right there."

In the next example, M did not start moving his icon until after D said "Terman Engineering?"

Example 3:

D: um
now it's over:
near the right edge
near Terman Engineering?
right next to the number five?
below and to the right
of of the five?
right there.
(Pair 2, Location 25, Visual evidence condition)

As M started moving, D used final rising again. At the point where D said "number five?" M had gone just past the correct location. D provides a more detailed description: "below and to the right of of the five?" and about half a second after that, M's icon arrives in the target location.

Directors also used final-rising intonation when they could *not* see the matchers' icon. In the next example, from the verbal-only evidence condition, D happened to use final rising intonation before M moved. Here, D had started out with a very general description, "you're going down," and then explicitly pursued information from M before providing a more explicit description:

Example 4:

D: ok
you're going down to
uhh
do you know where
the electronic
labs are?
S E L?

M: yah I see it

D: ok....[continues]
(Pair 2, Location 33, Verbal evidence condition)

In this example, M responded both verbally and by starting to move his icon at roughly the same time, just after D's first final rising intonation, and as D finished saying "S E L?" Despite examples like this one, there was no significant correlation between D's intonation and M's icon movement in the verbal evidence condition, nor would we expect there to be, since D had no direct evidence of whether M had moved yet.

3. Conclusions

Conversations can include both verbal and visual evidence, as two people try to reach a point where they believe they understand one another well enough for their current purposes. The grounding process is both flexible and opportunistic; when visual evidence is available, it can "fill in" for verbal evidence – that is, it can substitute for a verbal turn in the conversation (Brennan, 1990; 1992). When a speaker gets an appropriate response from an addressee, she can conclude that the utterance she presented has been accepted; when she doesn't get the evidence she expects, she can actively seek such evidence.

In this study, speakers often used final rising intonation when their addressees had not yet provided any evidence

of understanding. It is possible that speakers did this in order to pursue a response from their addressees, since it was correlated with the addressee's lack of movement in the visual condition, and uncorrelated with the addressee's lack of movement in the verbal condition. During the initial presentation of map location descriptions, the baserates for final rising intonation were just as high in the verbal condition as in the visual one, but in the verbal condition it was distributed randomly, (or at least it was not predicted by the addressee's icon movement).

There is a difference between when an addressee has a hypothesis about what a speaker means, and when they can conclude that they understand one another (Schober & Clark, 1989). In this study, the time-distance plots of the convergence of the two icons show a distinct elbow when the matcher's icon arrives within close range of the director's icon. When the director can see the matcher's icon, the conversational interchange is brought to a rapid conclusion, often by the director using deictic means (e.g. "park right there!"). When the director cannot see the matcher's icon, there is a relatively long period where they must continue grounding, until they can conclude they've reached a state of mutual understanding (Brennan, 1990; 1992).

For the future, I plan to examine additional prosodic aspects of these dialogues having to do with turn placement. These include early or overlapping turns that may provide early evidence that M believes he has understood D's presentation, and also pauses that may mark a presentation as problematic and in need of repair (see Levinson, 1983; Jefferson, 1989).

4. References

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, 8, 148-158.

Boomer, D. S. (1978). The phonemic clause: Speech unit in human communication. In A. W. Siegman and S. Feldstein (Eds), *Nonverbal behavior and communication.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Brennan, S. E. (1990). Seeking and providing evidence for mutual understanding. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Brennan, S. E. (1992). On the time course of understanding in conversation. Manuscript in submission.

Brennan, S. E. and Cahn, J. (1992). An architecture for contributions and repair in a natural language interface. Manuscript in preparation.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In L.B. Resnick, J. Levine, and S.D. Teasley (Eds.), *Perspectives on Socially Shared Cognition.* Washington, DC:APA.

Clark, H. H. and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2, 19-41.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259-294.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Cruttenden, A. (1986). *Intonation.* Cambridge, UK: Cambridge University Press.

Dittman, A. T. and Llewellyn, L. G. (1967). The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology*, 6, 341-349.

Goodwin C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic Press.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger and P. Bull (Eds.), *Conversation.* Philadelphia: Multilingual Matters Ltd.

Levinson, S. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Liberman, M. and Sag, I. A. (1974). Prosodic form and discourse function. In *Papers from the Tenth Redional Meeting*, Chicago Linguistics Society, Chicago, IL.

Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15, 19-44.

McLemore, C. (1991). The pragmatic interpretation of English intonation: Sorority speech. Unpublished doctoral dissertation, University of Texas, Austin, TX.

Pomerantz, A. (1984). Pursuing a response. In J. M. Atkinson and J. Heritage (Eds.), *Structures of social action.* Cambridge: Cambridge University, University Press.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In: P. R. Cohen, J. Morgan, and M. Pollack (Eds.). *Intentions in Communication.* Cambridge, MA: MIT Press.

Rosenfeld, H. M. (1987). Conversational control functions of nonverbal behavior. In A. W. Siegman and

S. Feldstein (Eds.), *Nonverbal behavior and communication.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sag, I. A. and Liberman, M. (1975). The intonational disambiguation of indirect speech acts. In *Papers from the Eleventh Redional Meeting*, Chicago Linguistics Society, Chicago, IL.

Svartvik, J. and Quirk, R. (Eds.)(1980). *A corpus of English conversation.* Lund, Sweden: Gleerup.

Whittaker, S. J., Brennan, S. E., and Clark, H. H. (1991). Coordinating activity: An analysis of interaction in computer-supported cooperative work. Proceedings, CHI '91, Human Factors in Computing Systems, ACM Press, pp. 361-367.

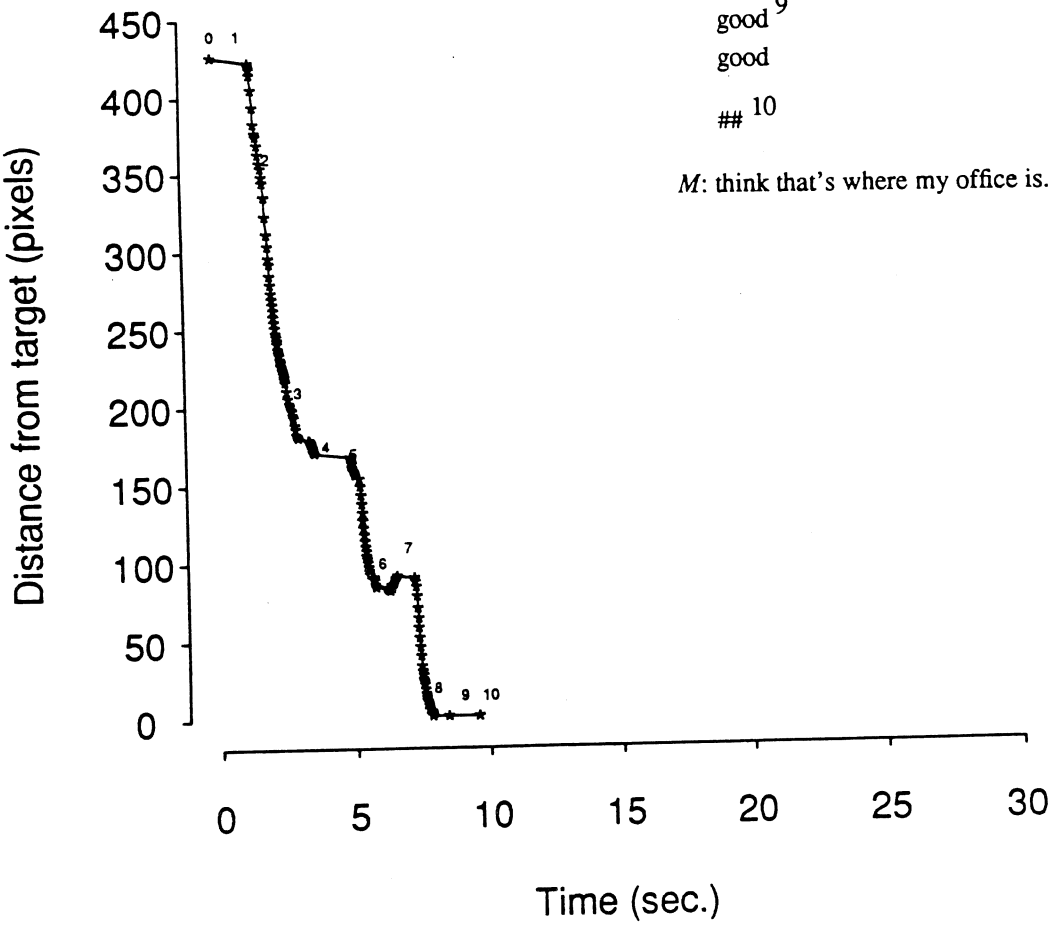
Wilkes-Gibbs, D. (1986). Collaborative processes of language use in conversation. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of Chicago Linguistic Society* (pp. 567-578). Chicago: Chicago Linguistic Institute.

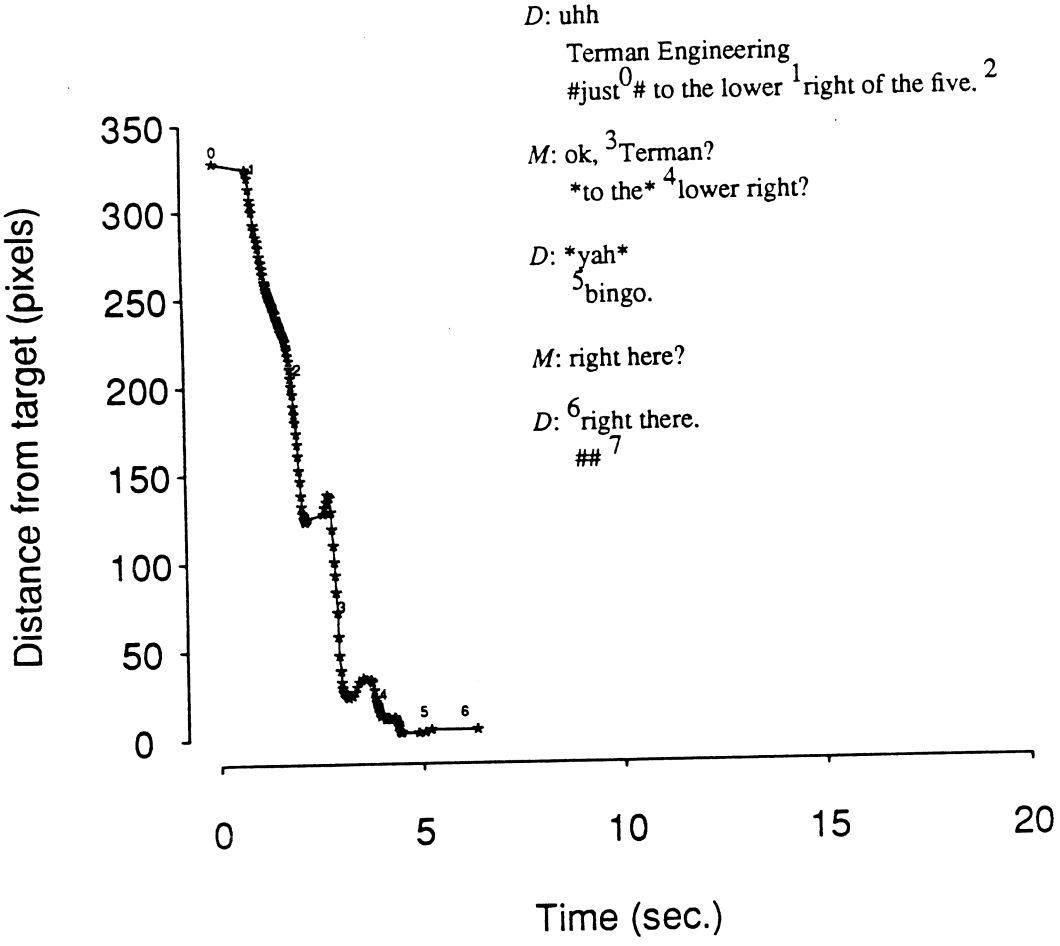
Appendix

D: ok
#[0]now we# went
[1]south,
[2]we're
about halfway down the [3]screen,
in the elec[4]tronics lab, [5]
we're in the [6]southern most
[7]wing of the electronics [8]lab,
good [9]
good
[10]

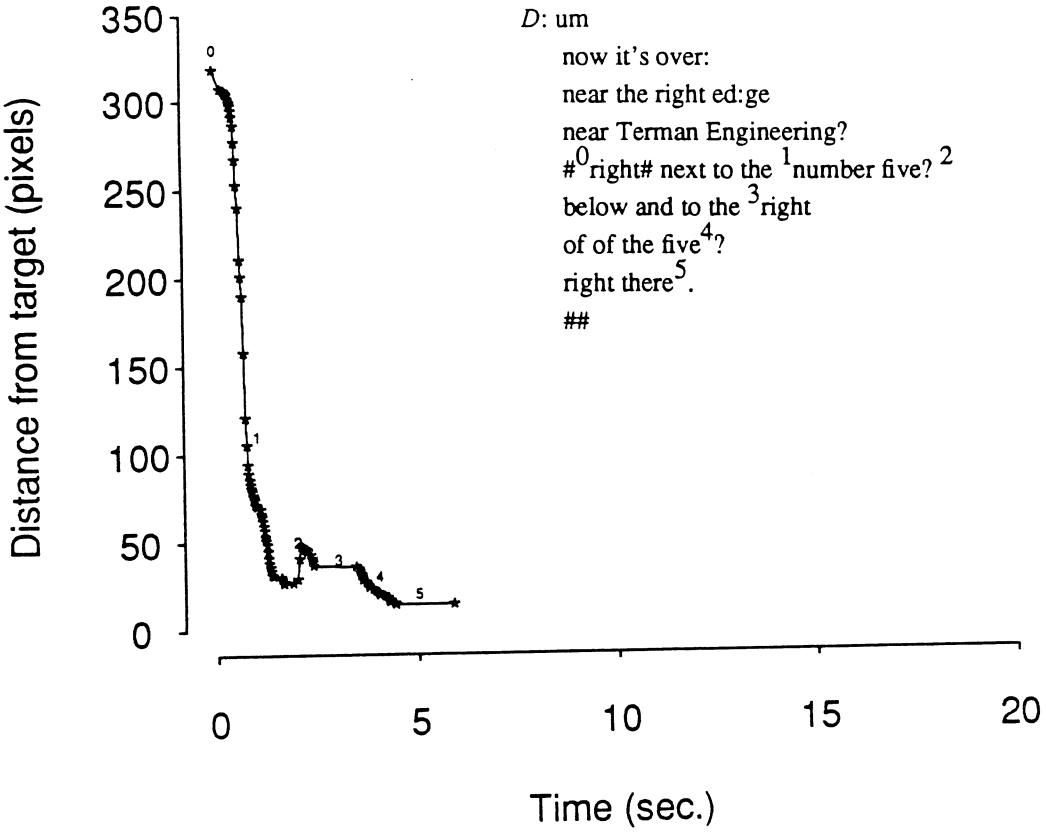
M: think that's where my office is.



Example 1 Visual evidence, Stanford map, Item 33, Pair 3



Example 2 Visual evidence, Stanford map, Item 25, Pair 4



D: um
now it's over:
near the right ed:ge
near Terman Engineering?
#[0]right# next to the [1]number five? [2]
below and to the [3]right
of of the five[4]?
right there[5].
##

Example 3 Visual evidence, Stanford map, Item 25, Pair 2

D: ok
you're going #[0]down to:#
[1]uhh
do you know wher:e
the [2]electronic
[3]labs are?
[4]S E L?*

M: *yah I see [5]it

D: ok,
[6]it's:
[7]right above
the:e
[8]TRON
[9]in
Elec[10]tronics,
and [11]it's
on that lower E[12]
part of the building. [13]

M: wait[17], it's ju- just above[15] thuuh

D: [16]Tron.

M: the [17]tronn?

D: yah, [18]on the building?

M: [19]what is the tron? [20]

D: oh-
in the
[21]elecTRONics labs?

M: [22]oh,
oh the TRON. [23]

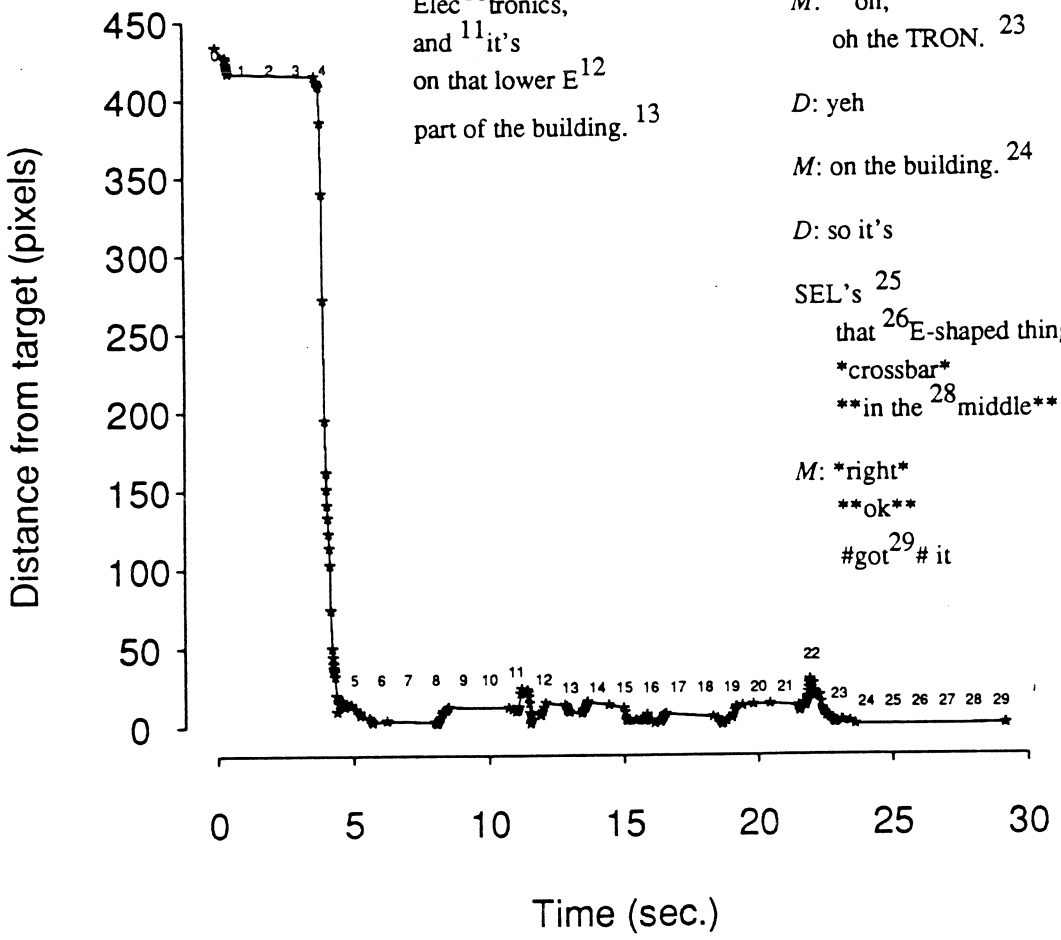
D: yeh

M: on the building. [24]

D: so it's

SEL's [25]
that [26]E-shaped thing's [27]lower
crossbar
in the [28]middle

M: *right*
ok
#got[29]# it



Example 4 Spoken evidence, Stanford map, Item 33, Pair 2

An investigation into the correlation of cue phrases, unfilled pauses and the structuring of spoken discourse

Janet Cahn

Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
cahn@media.mit.edu

ABSTRACT

Expectations about the correlation of cue phrases, the duration of unfilled pauses and the structuring of spoken discourse are framed in light of Grosz and Sidner's theory of discourse and are tested for a directions-giving dialogue. The results suggest that cue phrase and discourse structuring tasks may align, and show a correlation for pause length and some of the modifications that speakers can make to discourse structure.

1. Introduction

Because an utterance is best understood in the context in which it is delivered, its interpreters must be able to identify the relevant context and recognize when it is altered, supplanted or revived. The transient nature of speech makes this task difficult. However, the difficulty is alleviated by the abundance of lexical and prosodic cues available to a speaker for communicating the location and type of contextual change. The investigation of the interaction between these cues presupposes a theory of contextual change in discourse. The theory relating attention, intentions and discourse structure[3] is particularly useful because it provides a computational account of the current context and the mechanisms of contextual change. This account frames the questions I investigate about the correlation between between lexical and prosodic cues. In particular, the theory motivates the selection of the *cue phrase*[3] — a word or phrase whose relevance is to structural or rhetorical relations, rather than topic — and the *unfilled pause* (silent pause) as significant indicators of discourse structure.

2. The tripartite nature of discourse

To explain the organization of a discourse into topics and subtopics, Grosz and Sidner postulate three interrelated components of discourse — a linguistic structure, an intentional structure and an attentional state[3]. In the *linguistic structure*, the linear sequence of utterances becomes hierarchical — utterances aggregate into discourse segments, and the discourse segments are organized hierarchically according to the relations among the purposes or *discourse intentions*[1] that each satisfies.

[1] Discourse intentions are those goals or intentions intended to be recognized by each participant as the purpose to which the current segment of talk is devoted.

The relations among discourse intentions are captured in the *intentional structure*. It is this organization that is mirrored by the linguistic structure of utterances. However, while the linguistic structure organizes the verbatim content of discourse segments, the intentional structure contains only the intentions that underlie each segment. The supposition of an intentional structure explains how discourse coherence is preserved in the absence of a complete history of the discourse. Rather, discourse participants summarize the verbatim contents of a discourse segment by the discourse intention it satisfies. The contents of a discourse segment are collapsed into an intention, and intentions themselves may be collapsed into intentions of larger scope.

The discourse intention of greatest scope is the Discourse Purpose (DP), the reason for initiating a discourse. Within this, discourse segments are introduced to fulfill a particular Discourse Segment Purpose (DSP) and thereby contribute to the satisfaction of the overall DP. A segment terminates when its DSP is satisfied. Similarly, a discourse terminates when the DP that initiated it is satisfied.

The *attentional state* is the third component of the tripartite theory. It models the foci of attention that exist during the construction of intentional structures. The global focus of attention encompasses those entities relevant to the discourse segment currently under construction, while the local focus (also called the *center*[2]) is the currently most salient entity in the discourse segment. The local focus may change from utterance to utterance, while the global focus (i.e., current context) changes only from segment to segment.

The linguistic, intentional and attentional components are interrelated. In particular, the *attentional state* describes the processing of the *discourse segment* which has been introduced to satisfy the current *discourse intention*. The functional interrelation is expressed temporally in spoken discourse — the linguistic, intentional and attentional components devoted to one DSP co-occur. Therefore, a change in one component reflects or induces changes in the rest. For example, changes ascribed to the attentional state indicate changes in the intentional structure, and moreover, are recognized via qualitative changes in the linguistic structure. It is because of their interdependence and synchrony that I can postulate the hypothesis

that co-occurring linguistic and attentional phenomena in spoken discourse — cue phrases, pauses and discourse structure and processing — are linked.

The part of the theory most directly relevant to my investigation are those constructs that model the attentional state. These are the *focus space* and the *focus space stack*. The *focus space* is the computational representation of processing in the current context, that is, for the discourse segment currently under construction. Within a focus space dwell representations of the entities evoked during the construction of the segment — propositions, relations, objects in the world and the DSP of the current discourse segment.

A focus space lives on a pushdown stack called the *focus space stack*. The progression of focus in a discourse is modeled via the basic stack operations — pushes and pops — applied to the stack elements. For example, *closure* of a discourse segment is modeled by *popping* its associated focus space from the stack; *introduction* of a segment is modeled by *pushing* its associated focus space onto the stack; *retention* of the current discourse segment is modeled by leaving its focus space on the stack in order to add or modify its elements.

The contents of a focus space whose DSP is satisfied are accrued in the longer lasting intentional structure. Thus, at the end of a discourse the focus space stack is empty while the intentional structure is fully constructed.

The focus space model abstracts the processing that all participants must do in order to accurately track and affect the flow of discourse. Thus, it treats the emerging discourse structure and the changing attentional foci as publicly accessible properties of the discourse. However, although the participants themselves may act as if they are manipulating public structures, the informational and attentional properties of a discourse are, in fact, modeled only privately.

In explaining certain lexical and prosodic features of discourse, it is often useful to return to these private models. For a speaker's utterance is conditioned both by the state of the her own model and by her beliefs about those of her interlocutors. The time–dependent nature of speech emphasizes the importance of synchronizing private models. Lexical and prosodic focusing cues hasten synchronization. In particular, they guide the listeners in updating their models (among them, the focus space stack) to reflect the attentional changes already in effect for the speaker.

For my analysis, the most relevant private model belongs to the *current speaker*, whose discourse intentions guide, for the moment, the flow of topic and attention in a discourse and whose spoken contributions provide the richest evidence of attentional state. If cue phrases and unfilled pause durations can be shown to correlate with attentional state (and by definition, the intentional and linguistic structure), the attentional state they reveal belongs to the current speaker, and the attentional changes they denote are the ones the speaker makes in her own private model.

3. Main Hypotheses

The theory of the tripartite nature of discourse frames my hypotheses about the correlation of cue phrases, pause duration and discourse structure. The main hypotheses are these: that particular unfilled pause durations tend to correlate with particular cue phrases and that this correlation is occasioned by changes to the attentional state of the discourse participants, or, equivalently, by the emerging intentional structure of a discourse.

Cue phrases Changes to the attentional state occur at segment boundaries. Cue phrases by definition evince these changes — they are utterance– and segment-initial words or phrases and they inform on structural or rhetorical relations rather than on topic. Thus, for cue phrases, the question is not whether they correlate with attentional state, but how. To answer this question, we ask, for each cue phrase (e.g., *Now, To begin with, So*), whether it signals particular and distinct changes to the attentional state.

Pauses The correlation of unfilled pauses with attentional state is less certain because pauses appear at all levels of discourse structure. They are found within and between the smallest grammatical phrase, the sentence, the utterance, the speaking turn and the discourse segment. Their correlation is mainly with the cognitive difficulty of producing a phrase or utterance[1]. To link this correlation with the task of producing discourse structure, we must posit a variety of attentional operations with corresponding variability in cognitive difficulty. Specifically, we construct the chain of assumptions that:

- More than one attentional operation exists (e.g., initiation, retention, closure).
- The different attentional operations are distinguished by their effect on the attentional state and by the cognitive difficulty of their production.
- The amount of silence preceding an attentional operation is correlated with the greater or lesser demands it makes on mental processing.

To link unfilled pause duration to discourse structure we must first establish that operations on the attentional state can be distinguished sufficiently to explain the different demands that each operation makes on discourse processing and which, therefore, might be reflected in the duration of segment-initial unfilled pauses.

4. Auxiliary hypotheses

The linking of pause duration to the processing of discourse segments motivates some auxiliary hypotheses that refine notions about the kinds of mental operations sanctioned by the focus space model and about the internal structure of a discourse segment. These auxiliary hypotheses are developed in this section.

4.1. Attentional operations

In the theory of discourse structure, changes to the attentional state are modeled as operations on the focus space stack. These operations appear reducible to four distinct sequences of stack operations that correspond to four distinct effects on the attentional state, as follows:

- One push — *Initiate* a new focus space.
- No push, no pop — *Retain* the current focus space.
- One or more pops — *Return* to a previously initiated[2] focus space.
- One or more pops followed by a push — *Replace* a previous focus space(s) with a new one.

The arrangement is asymmetrical in that it is possible to pop more than one focus space per operation, but to push only one, as shown in Table 1.

Operation	Focus space stack before operation	Focus space stack after operation	Summary
Initiate	FS_1	FS_2 / FS_1	One push.
Retain	FS_1	FS_1	No push, no pop.
Return	FS_2 / FS_1	FS_1	One or more pops.
Replace	FS_1	FS_2	One or more pops, followed by a push.

Table 1: **The effect of the four focusing operation on the focus spaces (FS) in the pushdown focus space stack.**

The decomposition of focus space operations into stack operation primitives is not merely an attempt to impose a computational patina on descriptive terms. Rather, it suggests that operations that differ in kind and number place different requirements on mental processing for both speaker and therefore might be accompanied by lexical and acoustical phenomena that also differ.

4.2. Structure of a discourse segment

To further motivate the particular usefulness of cue phrases and unfilled pauses as locators of discourse segment boundaries and markers of attentional state, it is useful to distinguish among three phases in the life of a discourse segment (and its focus space counterpart) — its initiation, development and closure. We make the additional assumptions that each phase may be marked *explicitly* or *implicitly* and by *lexical* and *acoustical* phenomena.[3]

From inspection of dialogue, it appears that the development phase must be instantiated explicitly with lexical contributions, while the boundary phases need not be. However, while lexical marking of segment boundaries is optional, prosodic marking is not. Thus, at initiation of a discourse segment we find, for example, an expanded pitch range[12] and at its closure, phrase-final lowering[5] and syllable lengthening [6].

Sometimes, the same structural cue is implicit for one segment yet explicit for another. For example, in a *Replace* operation, explicitly marked closure of one segment implicitly permits the initiation of the next. Conversely, an explicitly marked *Initiation* of the current segment testifies implicitly to the closure of the previous one.

Boundary phenomena are of special relevance toward retrieving discourse structure from a multiplicity of lexical and acoustic clues. The distinction between explicit and implicit correlates for each phase of segment construction admits four classes of segment boundary phenomena — phenomena that are: explicit and segment-initial; implicit and segment-initial; explicit and segment-final; and implicit and segment-final. An investigation of how cue phrases and unfilled pauses reflect discourse structure and the state of its processing is thus an investigation of the *explicit* and *segment-initial* evidence of focus space initiation.

The selection of segment-initial phenomena in no way implies that segment-final phenomena are any less crucial to the communication and recognition of discourse segment boundaries. Nor does the selection of cue phrases and unfilled pauses minimize the contributions of other lexical and prosodic phenomena. Rather, these selections are motivated by features of the focus space model that both cue phrases and unfilled pauses might specially illuminate, and conversely, by features of the model that might specially illuminate the discourse function of cue phrases and unfilled pauses. These features are described in the following two sections.

[2] When at least one focus space remains on the stack, the discourse continues. When none remain, however, the discourse is ended.

[3] Gestural correlates of discourse structure and processing are outside the scope of this investigation.

5. Cue phrases, discourse markers and attentional state

Cue phrases are those words or phrases which introduce an utterance — e.g., *To begin with, First of all, Now, But* — and coordinate the flow of conversation and focus rather than contribute directly to the topic at hand. They provide broad, topic independent indications of how the speaker intends to relate the current utterance to those preceding it, thus locating the utterance in the discourse structure. The information they convey is attentional, intentional or both.

The study of cue phrases and their correlation with discourse structure and focus of attention is most extensive for the *discourse marker*[10] subcategory. Schiffrin's work in particular, is the basis for my predictions about the structural effects of cue phrases on the focus space model.

5.1. Discourse markers

Discourse markers are generally single word phrases, such as *Well, Now, Then* or *So*, whose pragmatic role in a discourse usually follows from their syntactic and semantic role in a grammatical phrase. That is, if a word in semantic guise relates *propositions in a grammatical phrase*, it marks in its pragmatic guise the same or similar relation between *utterances in a discourse*. For example[10]:

- **And**, as a discourse marker, indicates connectedness, conveying the speaker's view that the utterance it heads is connected to the prior discourse. The connection may be to the immediately previous utterance or to the speaker's prior [interrupted] turn.
- **But** also a marks connectedness, but connects utterances in a *contrast* relation. The contrast may be structural (resumption after a digression or interruption) or rhetorical. Like **well**, it introduces unexpected or undesired material, but in a less cooperative manner.
- **I mean** precedes a repair or modification of the speaker's own contribution or highlights something to which the speaker believes the hearer should attend.
- **So** may precede a presentation of a result, and indicates transitions to a higher level, in contrast to "**because**" which indicates progressive embedding.
- **Now** emphasizes what the speaker is about to do, and is often used to introduce evaluations.
- **Well** is often used in response, when the possibilities offered by the previous speaker are inadequate. It indicates an awareness of conversational expectations but also heralds a violation of the previous speaker's expectations.
- **You know** indicates an appeal to shared knowledge and mutual beliefs.

5.2. Discourse markers reinterpreted

Some of the observations about the conversational role of discourse markers invoke structural effects (embedding, return to a higher level) although without detailing the structure in question. A more unified and computationally driven account might be posed in terms of operations on the focus space stack, as follows:

- **And** (connectedness): *Retain, Return.*
- **But** (contrast): *Retain, Replace* or *Return.*
- **I mean** (modification or repair): *Initiate, Retain.*
- **So** (presentation of a result): *Return, Replace.*
- **Because** (progressive embedding): *Initiate.*
- **Now** (what the speaker is about to do): *Replace.*
- **Well** (inadequate options): *Replace.*
- **You know** (appeal to shared knowledge): *Retain*, or *Initiate* when it precedes an aside.

In addition, there are the cue phrases that highlight structural or propositional ordinality. The first use of such a phrase (e.g., *To begin with, In the first place,*) is likely to denote a focus space *Initiation* while subsequent uses (e.g., *Secondly, Finally,*) denote a focus space *Replacement.*

These formulations are not deterministic. They illustrate, however, the hypothesis that certain of the discourse markers are more likely to betoken certain focusing operations. Under what conditions might such correspondences exist? Clearly, features of the context in which a cue phrase is used might constrain its effect on focusing, and so explain how conversants are able to track focus from cues that, by themselves, are ambiguous.

Thus, to select the probable from the possible, corroboration from other quarters is required. Lexical corroboration may be semantic, from domain specific evidence of topic change or continuation. Or it may be syntactic, from those syntactic distributions that tend not to cross segment boundaries (tense, aspect and the scope of referring expressions[3])[4] Alternatively, prosodic features are likely to better identify the current use of a cue phrase from those that are possible.

6. Unfilled pauses and attentional state

The most useful prosodic correlates of discourse segmentation occur at segment boundaries and indicate either the opening of a new segment, closure of the old or both. For example, a phrase-final continuation rise forestalls segment closure while phrase-final lowering confirms it[9]. And expanded pitch range tends to mark the introduction of new

[4]For example, Walker and Whittaker observe that deictic pronominal reference may cross segment boundaries, while nondeictic pronominal reference does so only rarely[13].

topics, while reduced pitch range marks subtopics and parentheticals. Similarly, voice quality changes, e.g., from normal to creaky voice, may accompany attentional and intentional changes.

Filled pauses (e.g., *Um, uh*) and unfilled pauses appear at segment boundaries but are also found within a discourse segment and in the smaller groupings it contains. In contrast to the propositional and attentional accounts of intonational cues[9], accounts of pausing invoke the demands of cognition and pragmatics. For example, the duration of unfilled pauses has been observed to correlate with the cognitive difficulty involved in producing an utterance[1], while filled pauses may function as a floor holding device[7], or perhaps, correlate with the speaker's emotional response to topic[1].

As corroborators of attentional interpretations of cue phrases filled pauses are less useful than unfilled pauses because they overlap with cue phrases in both form (partially lexicalized) and function. A more independent measure is provided by unfilled pauses which are not lexicalized and therefore carry neither lexical nor intonational propositions. Rather, as correlates of the cognitive processing, they may also correlate with the specific differences among stack operations, which, after all, are cognitive operations, albeit idealized.

The selection of unfilled pause duration as a possible marker of attention and segmentation also has the practical advantage of being easy to locate instrumentally and easy to check perceptually. Moreover, its measurement is unambiguous instrumentally and requires less from perception, than, for example, intonational prosodic cues. For, while intonational features are categorical according to their type (combinations of the L, H and * tokens[8]) and the structure to which they apply (word, intermediate phrase, intonational phrase), pause duration is ordinal and is measured on the same continuous linear scale for all levels of linguistic and intonational structures.

7. Questions and predictions

My investigation is inspired by the theory relating attentions, intentions and discourse structure[3]. To the more specific observations linking cue phrases to attentional state[3, 10] and the duration of unfilled pauses to increased cognitive difficulty[1], I add the assumption of four fundamental focusing operations. Together, they motivate my hypotheses that:

(1) Specific cue phrases betoken specific focusing operations.

(2) Differences in the cognitive difficulty of the focusing operations are reflected in the duration of the pauses that precede them.

From these hypotheses come the specific questions that guide the research:

- Is there a correlation between the focusing operations and the duration of the pause that precedes it?
- Are cue phrases correlated with focusing operations — how often and under what circumstances?
- What is the relation of pausing and cue phrases — do they substitute for each other, compliment each other or play different roles such that one is required or allowed where the other is not?
- Is there a unique minimum cognitive cost for each stack primitive (Push, Pop) of which focusing operations are composed, and that would therefore explain differences in segment-initial pause duration?

In addition, the hypotheses raise questions not immediately answerable:

- If there are indeed patterns of usage, do they differ predictably for different discourse features, for example, by format (monologue or dialogue) or according to the planning effort (prepared or extemporaneous) required in formulating each utterance?
- If on the other hand, correlations are partial at best, can other lexical or prosodic features provide the missing correlates?

Research into these questions is not without its biases. Thus, I expected to find in my discourse samples the following correlations:

- *Unfilled pause duration and focusing operation are correlated.*
- *Cue phrases are correlated with focusing operations.* (The particular predictions are discussed previously in Section 5.2.)
- *Cue phrase type and unfilled pause duration are correlated as well.*

The hypothesized correlation of unfilled pause duration with focusing operations is based on assumptions about variations in complexity among the operations, such that longer pauses will accompany more complex operations. Complexity is conjectured to correlate with kind and number. That is, it varies according to whether the operation decomposes into pops, a push or both and it increase with the number of segments opened or closed in one operation.

This produces the particular predictions that:

- *Retentions will be preceded by pauses of the smallest duration* because they induce neither a push nor pop and therefore are the least costly of the focusing operations.

- *Pause duration is positively correlated with the number of segments affected in one focusing operation.* That is, the more segments opened or closed, the longer the preceding pause.
- *Pops are more costly than pushes.* This follows from an assumption that adding information (a push) builds on what is currently established and accessible, while removing information (one or more pops) makes the production of subsequent utterances more difficult.

8. Data

I analyzed two discourse samples — three minutes of a directions discourse and seven minutes of a manager–employee project meeting. The segmentation of the second proved difficult and is still in progress, so I report results only for the first.

In the directions discourse, Speaker B provides Speaker A with walking directions to a location on the M.I.T. campus. The discourse takes the form of an expert-client dialogue. Although Speaker A initiates the dialogue, most of the discourse segments and their intentions are introduced by Speaker B, the expert.[5]

9. Methods

The search for correlations among cue phrases, unfilled pauses and discourse structure generated three data collection tasks:

- Identification of cue phrases;
- Identification and measurement of unfilled pauses;
- Segmentation of the discourse via the identification of the focusing operations that effected the segmentation.

9.1. Cue phrase identification

The main challenge of cue phrase identification lay in distinguishing cue from non-cue uses of a phrase. Usually, cue uses are utterance- or segment-initial, while non-cue uses are not. However, this is not a reliable criterion for the connectives, *And* and *But*, which may head an utterance or phrase as either a cue phrase or a syntactic conjunctive. In cases where the usage was unclear, I decided against the pragmatic usage if the phrase in question provided syntactic coordination of two semantically related propositions. If even this judgment proved difficult, I applied the intonational criteria that distinguished cue and non-cue uses of *Now*[4]. Thus, if the cue phrase candidate was deaccented, or accented with L* tones or uttered as a complete intonational phrase, it was classified as a cue phrase.

[5]The conversation occurred in a face-to-face encounter and was recorded on a hand-held cassette recorder.

9.2. Pause location and measurement

Pauses were identified by ear and corroborated and measured using the waveform and the energy track displays of two signal processing programs.[6] The locations of all unfilled pauses were recorded, as were their durations, rounded to the nearest one tenth of a second.

In general, the procedure was straightforward. The only confusion was presented by the silence between the closure and release phase of plosives. This silence was not counted as a genuine unfilled pause.

9.3. Discourse segmentation

An accurate discourse segmentation falls out of an accurate classification of the focusing operations by which the segments have been constructed. The tasks are interrelated and both are difficult. Therefore, in this section I will discuss in detail the task, its difficulties and the classification criteria I developed to enhance the accuracy of my judgments.

The task The segmentation of a completed discourse is equivalently the task of recapturing the attentional state that accompanied each successive utterance. Attentional cues are especially important because topical relations do not always predict discourse structure. The points at which discourse structure diverges from the organization of information in the domain may be precisely the points at which attentional cues are most appropriate.

Segmentation of a completed discourse is most straightforward for expository text. In such discourse, domain and attentional hierarchies often coincide — the relations among segments and of each segment to the overall Discourse Purpose are clear. In spoken and impromptu discourse, however, the alignment of DSPs is not always so felicitous. Even in the task-oriented directions discourse, the relations among steps in the task did not conclusively determine the relations of the discourse segments in which these steps were described.

The particular segmentation difficulties presented by my sample(s) led to the development of explicit criteria for isolating the corroborating features of attentional operations and discourse structure. The criteria help clarify confusion from two sources — the distinction between attentional and domain hierarchies and the interpretation of underspecified lexical and prosodic attentional cues.

Separating the attentional from the topical. In prepared discourse (written or spoken) the intentional structure is tightly coupled to the Discourse Purpose. In contrast, impromptu discourse exhibits a looser coupling, owing to its

[6]*SPIRE*, written for the LISP machine by Victor Zue's group at M.I.T. and *dspB* (digital signal processing workBench) written for the DECstation by Dan Ellis at the M.I.T. Media Laboratory.

real-time and situated nature. In such discourse, the maintenance of coherence requires the real-time management of cognitive resources upon which competing demands may be made. As a consequence, influences outside the ostensible DP must be managed in support of continuing the conversation at all. Because DSPs that are ostensibly outside the current DP can become temporarily relevant, provision must be made for their principled incorporation into the attentional state and in the linguistic and intentional structures.

This is accomplished via attentional constructions that are more likely to occur in spoken discourse, for example, flashbacks, digressions and interruptions[3]. Their relation to the discourse in which they occur illustrates the difficulty of segmenting in hindsight a discourse whose DSPs may satisfy multiple DPs. This recommends against reliance on domain knowledge, since one discourse may invoke more than one domain.

Therefore, to locate segment boundaries, I use criteria that emphasize focusing operations independent of the ostensible DP. For example, although the succession of two topically unrelated segments might suggest a *Replace* operation, it is treated as an *Initiate* in the presence of explicit indicators of linkage or in the absence of explicit indicators of separation. Consequently, successive segments may be linked hierarchically in the attentional and linguistic structures despite their topical independence.

For example, in the following section of the directions discourse (1) is a topic introduction, (2) a digression and (3) an elaboration, i.e., a subtopic:

(1) To your left,
(2) if you have followed these directions faithfully,
(3) y'know you'll be facing a wall straight ahead of you,

Although (2) is a comment on discourse processing, it functions neither as a cue phrase nor a synchronization device. The digression it represents it not topically subordinate to (1), nor is (3) topically subordinate to (2). However, they are attentionally subordinate to the utterances they follow, as indicated by the continuation rises at the end of (1) and (2). While the semantic and topical differences between successive utterances argue for segment separation, the acoustical concomitants argue against. Therefore, the attentional moves that introduce (2) and (3) contain no pops. Instead, they are *Initiations*, producing the following segmentation:

Replace (1) To your left,
Initiate (2) if you have followed these directions faithfully,
Initiate (3) y'know you'll be facing a wall straight ahead of you.

Interpreting underspecified cues Even when successive utterances are aligned attentionally and topically, their cue phrase and prosodic markings may not conclusively reveal their exact place in discourse structures. The underspecified nature of cue phrase correspondences to focusing operations is discussed in Section 5.2. Prosodic marking is similarly underspecified, and on two counts. First, a particular intonational feature at the (e.g., phrase-final lowering, phrase-initial pitch range expansion) can felicitously indicate more than one focusing operation; second, the intonation at a phrase boundary often indicates stack primitives (push, pop, null) more reliably than the composite focusing operations from which discourse structure is deduced.

For example, in the directions discourse, the cue phrases *So, But* and *And* often indicated pops, as did the prosodic changes that accompanied them, e.g., expanded pitch range and a shift from L* to H* tones. However, these cues did not reveal exactly how many segments were popped nor whether a push followed the sequence of pops. Thus, it was not always easy to distinguish a *Return* (one or more pops) from a *Replace* (one or more pops, followed by a push).

Neither domain nor syntactic knowledge were conclusive in this regard. For example, domain and syntax dictated the following segmentation:

Return (4) And you need to turn left and then walk along Building Five.
Initiate (5) And you'll be walking through the architecture lofts.

but in contraindication to what was specified intonationally:

Return (4) And you need to turn left and then walk along Building Five.
Retain (5) And you'll be walking through the architecture lofts.

(The intonationally driven segmentation, in contradiction to the structure of knowledge in the domain, may account for the listener's subsequent confusion about the very point made in this section of the discourse.)

Classification criteria Because semantic clues to attentional state can be confusing and lexical and prosodic markings inconclusive, it became necessary to standardize the procedure and criteria for classifying the focusing operations. An accurate classification depends on the answers to two questions for the phrase undergoing classification: Has a new focus space been opened? Has an old focus space been closed? Most useful in this regard are the lexical and prosodic phenomena within and around the phenomena currently under evaluation for their attentional effect.

What constitutes current phenomena, and what might constitute its surrounds? I selected as *current* the speech fragment that begins with one of five fragment-initial tokens and whose

terminating boundary is marked by the occurrence of the next fragment-initial token. These tokens are:

- The unfilled pause;
- The filled pause;
- A cue phrase;
- An acknowledgment form: *Ok, Sure, Uh-huh,* etc.;
- Or the unmarked case: any other sentence-initial grammatical constituent, e.g., a noun phrase, auxiliary verb, complementizer or adverb.

My demarcation of the relevant surrounding phenomena was less bound to structure than to function. For both prior and subsequent phenomena, I selected the smallest speech fragment that could be distinguished by its discourse function, i.e., by its attentional, coordination or topical role. I assign five classifications:

- A cue phrase;
- An acknowledgment or prompt;
- A segment closure (e.g., *Good!*);
- A repair;
- Or the unmarked case — development of the topic.

The lexical and acoustical features of prior, current and subsequent speech fragments are treated as corrobating evidence for the attentional operation associated with the *current speech fragment*. Often this evidence indicated a stack primitive — push or pop — rather than a full-fledged focusing operation. This is illustrated in Table 2, which catalogues the lexical and prosodic features exhibited by prior, current and subsequent speech, and the stack and focusing operations for which each is considered evidence.

Coding the data The data relevant to every speech fragment was coded for later statistical analysis. This translated into two tasks — identifying the prior, current and subsequent speech fragment and for each current fragment, recording:

- The duration of the preceding unfilled pause;
- The type of fragment-initial constituent, either:
  - A cue phrase;
  - An explicit acknowledgment form (e.g., *Ok, Sure.*);
  - A filled pause;
  - Or any other sentence-initial syntactic form whose function is primarily topical, not pragmatic.
- The co-occurring focusing operation.
- The embedding of the current segment in the linguistic structure (number of levels).
- The number of segments opened or closed in the focusing operation.

GIVEN:		CONCLUDE:	
SPEECH EVIDENCE	FEATURE	STACK PRIMITIVE	FOCUSING OPERATION FOR CURRENT FRAGMENT
Prior speech	Falling phrase-final intonation, acknowledgment, lexical/semantic closure.	Pop of co-occurring segment(s).	Replace.
	Phrase-final continuation rise.	Null	Retain.
Current speech	Pronominalization, reduced pitch range, nonstandard phonation, many L* accents (parentheticals), relative clause, *Now, Y'know,* Ordinal cue phase.	Push or Null for co-occurring segment.	Initiate, Retain.
	Nonpronominalized repetition (e.g., segue), expanded pitch range, reintroduction of normal phonation, *So, But.*	Pop of previous segment(s).	Return, Replace.
	Falling phrase-final intonation, acknowledgment, prompt, lexical closure, phrase-final creaky voice.	Impending Pop of co-occurring segment(s).	Retain (but an impending Return or Replace).
Subsequent speech	Nonpronominalized repetition (e.g, a segue), expanded pitch range, normal phonation, *So, But, Now.*	Pop of previous segment(s).	Return, Replace.

Table 2: **The lexical and acoustical features that support classifications of stack primitives and focusing operation(s).** A co-occurring segment denotes the segment containing the speech (prior, current, subsequent) under examination. The focusing operations, however, describe the attentional interpretation that such speech indicates for the *current* speech fragment.

- The discourse function of the immediately prior speech (cue phrase, acknowledgment, closure, filled pause, repair, topical but none of the above).

- The discourse function of the immediately subsequent speech (same categories as for prior speech).
- Whether the speaker was initiating or continuing a speaking turn with the current fragment.

Using this metric, one hundred speech fragments were identified according and their features coded. The coded representation of the discourse was then analyzed for distributions and statistical correlations.[7] The results are reported in the next section.

10. Results

In this section I summarize the raw data, report the results of statistical tests and offer an explanation of the findings.

10.1. Data

The segmentation of the discourse was reconstructed according to the focusing operations indicated both lexically and acoustically. The segmentation described a discourse with two top level segments. Within the first, the overall task was defined; within the second, it was executed. The task definition segment was itself composed of two top level segments, while the execution segment is composed of nine.

The key elements of the coding scheme were, of course, the focusing operation, the fragment-initial token and the duration of the unfilled pause preceding the fragment. Distributions for these categories are catalogued in Table 3.

10.2. Statistical analyses

The predictions were analyzed via statistical tests on the coded representation of the discourse.

Pause duration and focusing operation A comparison of the mean pause duration for each focusing operation showed a significant difference among the operations ($F(3,96)=7.31$, $p<.001$). The data in Table 4 point to the *Replace* operation as most different from the other three operations in this regard.[8]

Pause duration and number of segments affected in a focusing operation Longer pauses were positively correlated with the number of segments opened or closed during one focusing operation ($r = .357$, $p<.001$). This finding might partially explain the long pauses that appear before a *Replace*, since a *Replace* is the focusing operation most likely to affect the most focus spaces, By definition, it requires [almost] everything to be popped from the focusing before the initiation (push) of a new focus space.

[7] The discourse function classifications and the within-/between-turn distinctions were recorded to track the features influencing the judgment of focusing operation, but were not included in any calculations.

[8] However, the importance of this observation is offset by the small sample size and large standard deviation.

Category	Feature	Number of Occurrences	
		Marked	*Unmarked*
Focusing operation	Initiate	13	10
	Retain	18	37
	Return	6	5
	Replace	7	4
	ALL	44	56
		Initial	*Internal*
Fragment-initial constituent	And	3	4
	But	2	1
	Now	2	–
	Oh	2	–
	So	3	2
	Well	2	–
	Y'know	2	–
	Ordinal cue phrase	1	–
	Acknowledgment	2	7
	Filled Pause	7	4
	Unmarked	19	37
	ALL	45	54
	Seconds	*Initial*	*Internal*
Unfilled pauses	0.0	5	15
	0.1	6	11
	0.2	3	15
	0.3	4	5
	0.4	11	5
	0.5	1	5
	0.6	3	4
	0.7	4	2
	0.8	1	–
	0.9	1	–
	1.7	1	–
	2.0	1	–
		41	62
	Average	0.422 seconds	.224 seconds

Table 3: **Distributions of fragment-initial constituents, focusing operations and pause durations.** Separate counts are taken for *segment-initial* and *segment-internal* phenomena and for *marked* and *unmarked*. A *marked* focusing operation begins with a cue phrase, an acknowledgment form or a filled pause, while an *Unmarked* operation does not.

Focusing Operation	Number of Tokens	Mean Pause Duration (seconds)	Standard Deviation
Initiate	23	0.3217	0.2173
Retain	55	0.2091	0.1818
Return	11	0.2545	0.2505
Replace	11	0.6500	0.6727

Table 4: **Mean pause durations for each focusing operation.**

INITIAL TOKEN	INITIATE	RETAIN	RETURN	REPLACE	ALL
And	0.43 3	0.25 2	0.25 2	–	0.33 7
But	–	0.70 1	0.00 1	0.10 1	0.27 3
Now	–	–	–	0.55 2	0.55 2
Oh	–	0.00 2	–	–	0.00 2
So	–	0.15 2	0.15 2	0.05 1	0.13 5
Well	–	–	–	0.20 2	0.20 2
Y'know	0.40 2	–	–	–	0.40 2
Ordinal	0.40 1	–	–	–	0.40 1
Acknow-ledgment	0.10 1	0.20 7	–	0.90 1	0.27 9
Filled Pause	0.23 6	0.05 4	0.00 1	–	0.14 11
Unmarked	0.35 10	0.23 37	0.40 5	1.15 4	0.33 56
ALL	0.32 23	0.21 55	0.26 11	0.65 11	0.29 100

Table 5: **The mean duration, in seconds, of the pause preceding fragment-initial tokens and focusing operations that co-occur.** The number of tokens in the calculation follows the mean value.

Pause duration and depth of embedding A correlation of pause duration and the depth of embedding in the linguistic structure (or equivalently, the number of focus spaces still on the stack) showed no significant effect on pause duration ($F(1,98) = 0.1861, p<.7$).

Pause duration, cue phrase and focusing operation The directions dialogue contained too few fragment-initial tokens to calculate meaningful statistics about their relation to focusing operations. Therefore, the best course was to select from the raw data (see Table 5) the patterns that were likely candidates for further testing. For example, *So* was never associated with an *Initiate* operation and also was preceded by the smallest mean pause durations (0.13 seconds). A filled pause, with a similar mean pause duration (0.14 seconds) was primarily associated with *Initiates* and *Retains* but never with *Replace*. And, while *And* shared the same focusing operations as a filled pause, its mean value for pause duration was more than twice as large (0.33 seconds).

Pause duration and marked/unmarked To compensate for the small sample sizes of the cue phrase data, all explicit lexical markers of structure (cue phrase, acknowledgment, filled pause) were collapsed into the category, *marked*. The data in this category were compared to the data for lexically *unmarked* fragments. Because the longest pauses preceded unmarked *Returns* and *Replacements*, I predicted that unmarked operations would in general be preceded by longer pauses than marked.

The results are in the direction predicted and are summarized in Table 6. The average duration for pauses preceding a marked focusing operation was 0.24 seconds (standard deviation = 0.24), while the average for pauses preceding unmarked

SPEECH FRAGMENT	INITIATE	RETAIN	RETURN	REPLACE	ALL
Marked	0.30 13	0.17 18	0.13 6	0.36 7	0.24 44
Unmarked	0.35 10	0.23 37	0.40 5	1.15 4	0.33 56
ALL	0.32 23	0.21 55	0.26 11	0.65 11	0.29 100

Table 6: **The mean duration, in seconds, of the pause preceding focusing operations and marked or unmarked speech fragments that co-occur.** The number of tokens in the calculation follows the mean value.

operations was 0.33 seconds (standard deviation = 0.36). Statistically this approaches significance ($T(96) = 1.58, p = .12$).

10.3. Discussion

Thus far, analysis of the data identifies significantly longer pauses for the *Replace* operation than for any other and shows that pause duration is positively correlated with the number of segments affected by one focusing operation. These findings begin to distinguish the focusing operations quantitatively, by number of focus spaces affected, and qualitatively, by whether they occur within an established context (*Initiate, Retain, Return*) or at its beginning (*Replace*).

Although, the raw data in Table 3 appears to show patterns for specific segment-initial tokens, the number of tokens is insufficient for establishing a correlation between cue phrase and focusing operations, let alone a three-way relationship among cue phrase, pause duration and focusing tasks.

The categorical classification present particular problems. For, uncertainties arose even with the application of a classification metric. Perhaps these uncertainties should have been incorporated into the coding scheme or perhaps the categorical classifications should have been abandoned[9] in favor of additional and quantifiable acoustical and lexical features.

10.4. Refining the original hypotheses

Only partial conclusions can be drawn from the data. However, the results are useful toward refining the original hypotheses and determining the content of future research. The distinction between the pause data for marked and unmarked fragments is a case in point. For each focusing operation, the difference between mean pause durations at best only approaches significance (see Table [24 6). However, because the values for all focusing operations are always greater for unmarked utterances, a hypothesis is suggested: that, given a speech fragment and the focusing operation it evinces, the preceding unfilled pause will be longer if the fragment is lexically unmarked.

[9] at least, in this stage of the investigation

If this hypotheses is correct, two accounts can be constructed that would jointly predict the appearance of cue phrases. One account emphasizes the processes involved in choosing and communicating the state of global focus. The other emphasizes the mutually recognized (by speaker and hearers) attentional and intentional state of the discourse. Together they identify the factors that would impel a speaker to precede an utterance with a cue phrase, an unfilled pause or both.

The influence of the speaker's internal processes and conversational goals If an unfilled pause preceding a lexically unmarked fragment is significantly longer, we might assume that a particular focusing operation is executed in a characteristic amount of time (given adequate consideration of other contextual features). Within this time, we might observe silence, a cue phrase or both. [10]

Because both pause and cue phrase can appear at the same location in a phrase, we ask if their functions are equivalent, or instead, complementary. My hypothesis selects the second option, that they are complementary in the cognitive processing each reflects and in the discourse functions each fulfills. For, if the duration of an unfilled pause is evidence of the difficulty of a cognitive task, a cue phrase is evidence of its partial resolution.

As a communicative device, cue phrases are more cooperative than silence. In silence, a listener can only guess at the current contents of the speaker's models. With the uttering of a cue phrase, the listener is at least notified that the speaker is constructing a response. The minimal cue in this regard is the filled pause. *Bone fide* cue phrases, however, herald not only an upcoming utterance, but a particular direction of focus and even a propositional relation between prior and upcoming speech.

Cue phrases serve not only the listener but also the speaker. Because they commit to topic structure, but not to specific referents and discourse entities, they buy additional time for the speaker in which to complete a focusing operation and formulate the remainder of the utterance.

The influence of the state of the discourse The account of the influence of the currently observable state of the discourse rests on two patterns in the data: (1) the difference in pause durations for marked and unmarked *Initiates* and *Retains* is minimal; and (2) the difference between marked and unmarked *Returns* and *Replaces* is greater. If these patterns can be shown to be significant, they suggest that remaining in the current context is less costly than returning to a former context, or establishing a new one. The corollary is the claim that an expected focusing operation need not be marked, while an unexpected operation is most felicitous when marked.

[10]The discussion will focus on cue phrases, even though the points are relevant to other lexical markers of discourse structure and processing.

In other words, remaining in the current context or entering a subordinate context is expected behavior, while exiting the current context is not. Exiting the current context (focus space) carries a greater risk of disrupting a mutual view of discourse structures. The extent of risk is assessed for the listener by the difficulty of tracking the change and for the speaker, by the difficulty of executing it. The risk originates in the nondeterministic definitions of *Return* and *Replace* operations — both contain in their structure one or more pops. In addition, these operations can be confused because both begin identically, with a series of pops.

Because closing a focus space is a marked behavior, the clues to changing focus are most cooperative if they guide the listener toward re-invoking a prior context (i.e., a *Return*) or establishing a new one (Replace). Thus, certain clues are more likely to mark a return to a former context (e.g., *So, Anyway, As I was saying*), while others (*Now*, the ordinal phrases) mark a *Replace*.

Future work The goal of future investigations is to establish the bases for predicting the appearance of particular acoustical and lexical features. The speculations presented in this section provide a theoretical framework. If borne out, they can be refashioned as characterizations of the circumstances in which cue phrases and unfilled pauses are most likely to be used.

11. Conclusion

The relationships among cue phrases, unfilled pauses and the structuring of discourse are investigated within the paradigm of the tripartite model of discourse. Within this model, the postulation of four focusing operations provides an operational framework to which can be tied the discourse functions of cue phrases and the cognitive activity associated with the production of an utterance. Especially, the difficulty of utterance production might be explained by the complexity of the co-occurring focusing operation. Such a correspondence is, in fact, suggested by the positive correlation of pause duration and the number of focus spaces opened or closed in one operation on the focus space stack.

However, because the classification of focusing operations is uncertain, more data and better tests are required to characterize the relationships among the lexical and acoustical correlates of topic and focus. In addition, the aptness of the tripartite model itself is not assured. The idealizations it contains may undergo modification in light of new results, or be augmented by other accounts of discourse processing. On the other hand, the analysis of more quantitative data may confirm the implications of the model, and its appropriateness as the foundation for investigating the lexical and prosodic features of discourse.

12. Acknowledgments

Many thanks to Susan Brennan who selected and ran the statistical tests on the data and to Stephen Lines for numerous helpful comments on this paper. Various stages of this work were supervised in turn by Chris Schmandt and Ken Haase, both of the M.I.T. Media Laboratory. Their support is gratefully acknowledged as well.

References

1. Goldman–Eisler, F., A comparative study of two hesitation phenomena. In *Language and Speech (4)*, 1961, pp. 18-26.
2. Grosz, B. J., Joshi, A. K. and Weinstein, S., Towards a computational theory of discourse interpretation. *Draft*, 1989.
3. Grosz, B. J. and Sidner, C. L., Attention, intentions, and the structure of discourse. In *Computational Linguistics (12:3)*, 1986, pp. 175-204.
4. Hirschberg, J. and Litman, D., Now let's talk about now: identifying cue phrases intonationally. In *Association for Computational Linguistics (25)*, July, 1987, pp. 163-171.
5. Hirschberg, J. and Pierrehumbert, J., The Intonational Structuring of Discourse. In *Proceedings of the Association for Computational Linguistics*, July, 1986, pp. 136-144.
6. Klatt, D. H., Vowel lengthening is syntactically determined in a connected discourse. In *Journal of Phonetics(3:129)*, 1975, pp. 129-140.
7. Maclay, H. and Osgood, C. E., Hesitation phenomena in spontaneous English speech. In *Word(15)*, 1959, pp. 19-44.
8. Pierrehumbert, J. B., The phonology and phonetics of English intonation. *Ph.D. Thesis*. Massachusetts Institute of Technology, Department of Linguistics, 1990.
9. Pierrehumbert, J. and Hirschberg, J., The meaning of intonation contours in the interpretation of discourse. In *Intentions in Communication*. Edited by Cohen, P. R., Morgan. J. and Pollack, M. E., 1990, pp. 271-311.
10. Schiffrin, D., *Discourse Markers*, Cambridge University Press, 1987.
11. Sidner, C. L., Focusing in the comprehension of definite anaphora. In *Readings in Natural Language Processing*. Ed. by Grosz, B. J., Sparck-Jones, K. and Webber, B. L, Morgan Kaufman Publishers, Inc., 1986, pp. 363-394.
12. Sorensen, J. M. and Cooper, W. E., Syntactic coding of fundamental frequency in speech production. In *Perception and Production of Fluent Speech*. Ed. by Cole, R. A., published by Lawrence Erlbaum, 1980, pp.399-440.
13. Walker, M. A. and Whittaker, S., Mixed initiative in dialogue: an investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990, pp.70-79.

FRENCH LIAISON IN NATURAL DISCOURSE

Troi C. Carleton

Department of Linguistics
University of Texas at Austin
Austin, Texas 78712

1. INTRODUCTION

French liaison is one of the more well discussed phenomena in linguistics. It is the peculiar distribution and behavior of liaison which have led to numerous attempts to characterize and account for it in discourse. Traditionally, liaison was accounted for category by category, and even word by word in some cases. Various factors such as historical influence, phonetics, pragmatics, and syntax played a role in predicting the distribution of liaison in this approach (see DeLattre 1947, 1955, 1956, 1966). This typological approach has been abandoned in the past few decades in exchange for proposals consistent with current thought on the prosodic hierarchy and a shift towards more generalized theories. These types of proposals account for liaison in a less holistic way than proposals along the lines of DeLattre, appealing primarily, though not exclusively, to syntax based arguments. Selkirk (1972, 1974) presents a comprehensive proposal of the distribution of liaison based on X-Bar theory, thus rejecting for the most part the notion of the categorically (with respect to grammar) and lexically governed distribution of liaison. However, while the syntactic based account of liaison presented by Selkirk is an attractive proposal in that it is seemingly simple, it fails to accurately account for the distribution of liaison in spoken colloquial French. The traditional approach has a lot more to offer with respect to accounting for the distribution of liaison, in that the traditional approach looks to more than syntax for an explanation, but it too falls short of accurately accounting for the phonological distribution in natural discourse. A major reason why both of these approaches fall short has to do with the very basic assumptions these proposals make with respect to the distribution of liaison.

Essentially, all accounts of liaison begin by making the following two assumptions. First, there are three distinct styles of speech in French — casual, careful, and poetic — and these three styles are marked by their liaison distribution.[1] Liaison distribution falls out from the second assumption, namely that there are three distinct liaison categories — obligatory, prohibited, and optional. Casual speech is marked by a restricted liaison inventory, while careful speech may optionally include more liaison environments, and finally, formal or poetic speech has virtually no restrictions on where a liaison *may* occur. Herein lies the problem.

Since casual and formal speech are marked by *optional* increase in liaison use, it is, in fact, conceivable that what appears to be a restricted casual use of liaison, may in fact, be a case of careful speech where option to liaise is not taken. If this is true, then we must necessarily question the validity of style defined by the distribution of liaison. In turn we must question the status of the various liaison categories, since they rely crucially on style. Furthermore, since these models do rely crucially on the notion of style, we are forced to question the adequacy of these models at the very base level. Without the notion of style playing a key role in characterizing liaison, the optional liaison category becomes more of an escape-hatch for instances of liaison or lack of liaison which the current theory or model is unable to explain. Furthermore, even if we did accept the theoretically weak assumptions that these proposals base themselves on, they still fail to make accurate predictions concerning the distribution of liaison in the corpus of data under investigation in this study.

Our task must be to find a way to flush out the category of *optional* liaison as best we can by finding alternative explanations for the seemingly unpredictable behavior of liaison. In doing so we must look beyond the grammatical component. The immediate goal in this study is to arrive at an observationally adequate account of liaison. This we achieve by an inductive analysis of the distribution of liaison in a corpus of natural discourse. Crucial to this study and others like this is the corpus of natural discourse. It is only by looking at natural discourse that we can begin to see where current proposals fall short and new approaches are in order. An inductive approach here will allow us to capture generalizations in the data that simpler more popular models may miss. It is only by inductively approaching a corpus of natural data that we may begin to understand what role such phonological phenomenon, such as French liaison, play in prosodic phrasing.

---

[1]Some approaches may assume as many as four distinct styles – casual, careful or elevated, formal, and poetic.

Phonologists are often interested in the segmental process that has something to do with prosody. In fact, often times certain phonological processes are viewed as autonomous structures of prosody. That is to say, it is often assumed that segmental processes, such as French liaison, follow from a general structure, rather than act alone as separate pieces to the whole. While it is true that different phonological features will have something to say about phrasing, our approach shows that phonological features are not tools which simply organize content. This study does not assume there to be an underlying shape of discourse which relies exclusively on the grammatical component. The fact that there is sometimes an iphomorphism between content and shape, does not lead us to the conclusion that content is the shape. My work indicates that there is sensitivity to linguistic structure, but that the linguistic structure and the phonological features are not in lockstep. Other factors which we will be discussing in this paper must be taken into account in order to explain the behavior of these features. By using natural discourse as a model and by approaching the data inductively, we begin to understand what devices are used in organizing discourse.

2. A BRIEF OVERVIEW OF COMPETING PROPOSALS: SELKIRK VS. DELATTRE

2.1 SELKIRK'S PROPOSAL FOR LIAISON [2]

Selkirk's account for the occurrence and distribution of liaison relies crucially on syntax. Her central claim is that a liaison will occur within phonological words and not between them. The domains of these phonological words are determined according to the basic principles of SPE (Chomsky&Halle, 1968). A word is defined by a boundary symbol # which is inserted by rule at the beginning and end of certain syntactic domains. In essence, main lexical categories get boundaries at both the beginning and end, and non-lexical categories only get boundary symbols at the beginning.

> "the boundary # is automatically inserted at the beginning and end of every string dominated by a major category, ie, by one of the lexical categories "noun", "verb", "adjective" or by a category such as "sentence" and "noun phrase", "verb phrase" which dominates a lexical category." (taken from SPE (p.366))

According to this account, liaison will occur between two words when there is only one boundary symbol separating them, and a liaison will not occur when there are two boundary symbols separating them. In the example below, liaison occurs between *des* and *enfants*, but not between *enfants* and *anglais*.

(1) ex. ##des#enfants##anglais##

In the case where liaison occurs, *des* is a determiner, and thus a non-lexical category. It only gets a boundary marker at the beginning of it. *Enfants*, on the other hand, is a lexical category, so it is flanked by boundary symbols. Therefore, there is one boundary symbol, namely that of *enfants*, which separates *des* and *enfants* . Hence, liaison is licensed between the two words. Between *enfants* and *anglais*, there are two boundary symbols because both words are lexical. Therefore, a liaison is blocked in this case. It is not to say, however, that a liaison between *enfants* and *anglais* can **never** occur. This is not the only boundary configuration that this string of words can assume. That is to say, while a liaison does not occur between *enfants* and *anglais* in style I (casual speech), it may occur in style II (formal speech) and style III (poetic speech) by applying a "readjustment rule".

The notion of readjustment rules allows Selkirk to account for liaisons which occur across two word boundaries. The purpose of a readjustment rule is to manipulate the word boundaries, reducing ## to # in certain contexts (contexts being defined as Style I, II or III) , thus licensing liaison. Readjustment rules, as shown in example 2, allow her to explain deviant liaison behavior by putting the weight of the explanation on style.

(2) Style I: ##des#enfants##anglais## --->des enfants || anglais
Style II: ##des#enfants##anglais## --> (readjustment rule applies)
##des#enfants#anglais##-->des enfants^anglais

Crucial to the notion of the readjustment rule in light of the behavior of liaison is the optional status liaison hold in the readjusted environments. That is to say, a readjustment rule does not *necessitate* a liaison in the readjusted

[2]For a more detailed account of Selkirk's proposal see Selkirk (1972, 1974), Morin & Kaye (1982), Carleton (1992b)

contexts, it *permits* a liaison. As we have pointed out, it is precisely this dependence on style and optionality for prosodic explanation that this study questions.

2.2 DeLattre's Typology

Delattre (1947) presented the general tendencies and classification of the french liaison. Like the other proposals, he assumes the distribution of liaison to be predictable based on speaker style. His account differs slightly from other accounts in that he proposes *four* distinct styles. 1) familiar conversation, 2) elevated conversation, 3) formal , and 4) verse or poetic. Based on these assumptions he makes predictions with respect to the distribution of liaison in the following sentence: *Des hommes illustres ont attendu* , 'the illustrious men have waited'. Conversational french would only have a liaison between the article and the noun [3] — Des^hommes || illustres || ont || attendu. Elevated conversation could additionally have a liaison between the auxiliary and the participle —Des^hommes || illustres || ont^attendu. Formal speech could add a liaison between the noun and the adjective—Des^hommes^illustres || ont^attendu. Finally, poetic speech could liaise at every possible chance — Des^hommes^illustres^ont^attendu.

Like Selkirk's proposal, all but the first liaison in the example is optional in all styles except conversational style. Unlike Selkirk's proposal, however, DeLattre takes many factors other than syntactic relationships into account in describing liaison tendencies. As he explains, "la frequence avec laquelle se fait telle liaison facultative semble dependre de nombre facteurs." (p.44, 1955). 'The frequency of the optional liaison seems to depend on a number of factors.' He states five factors which influence the occurrence of liaison.

The first factor is stylistic, which was described above. The second factor he considers is syntax. He considers syntactic sequences and rates them from 1 to 10 , where 10 is the most frequent and 1 is the least. These sequences include determiner^noun, adjective^noun, personal pronoun^verb, etc. This sequential approach is the basic approach that my study has taken in initially classifying liaison distribution. Third, he defines the prosodic factor. Under this, he includes the length of the elements to be liaised, the intonation of the phrase, and focus.[4] The fourth factor influencing the distribution and frequency of liaison is the phonetic factor. For example, he states that liaisons are most likely to occur after a vowel than after a consonant, and furthermore, liaison is easier to make after one consonant than after two. The fifth and final factor effecting the frequency of liaison is a historic factor. For example, an historic influence might be the aspirated 'h'. An 'h' which was aspirated historically will not liaise.

While DeLattre's analysis of the distribution of liaison is not entirely consistent with what the corpus under investigation reveals, the spirit behind his analysis is consistent with the spirit behind this one. That is to say, he clearly understood the complexity of the distribution of liaison, and clearly saw that several forces combined to account for the behavior of liaison. Morin and Kaye (1982), in response to Selkirk, call for a more traditional approach to accounting for liaison, and by that they meant something more along the lines of DeLattre's proposal.

3. The Study

3.1 Methodology

This study is based on a tape recorded conversation between three French women and one American woman who speaks fluent french (the American woman's data were disregarded in this study.). The data were collected in 1983 at the University of Minnesota by Betsy Barnes, and came to me through Knut Lambrecht at the University of Texas, Austin. I received the orthographic transcripts of the conversation along with the tapes. The conversation I analyzed is part of a bigger corpus collected over several casual meetings between the four women and Betsy Barnes in her living room while enjoying lunch. Time and space prohibit me from discussing the entire corpus under investigation here. What I present in this working paper, is a slice of the corpus I have already analyzed (see Carleton, 1992b for more in depth discussion of the corpus).

---

[3] " || " indicates that no liaison occurs between two constituents (ex. hommes || illustres), and
" ^ " indicates there is a liaison relationship between two given constituents (ex. des^hommes).

[4] It is not within the scope of this study to show exactly where DeLattre's account does not fit my corpus; however, it is the fact that what he proposes here is inconsistent with what my data show. This thread can be picked up and illustrated in the next step of this study. The purpose would be to show with evidence the shortcomings of his analysis. Like Selkirk, my aim is not to completely discredit his word, but rather to use it as a spring board for further research into the nature of liaison.

I recorded each and every occurrence and non-occurance of liaison in the corpus. A non-occurance is defined as every phonological environment in which liaison could have taken place and didn't. Initial categorization of these tokens concentrated on syntactic sequences rather than syntactic domains.[5] Any cases in which it was difficult to ascertain whether or not a liaison occurred was thrown out, regardless of how "predictable" liaison was in the particular context. Most of the tokens I threw out were discarded because either the speaker was laughing , eating, or talking at the same time as someone else, and I couldn't hear it. Liaison omission as a result of fast speech was not discarded.

All of the occurrences and non-occurrences of liaison are divided into three major categories— 1) Always, 2) Never, and 3) Contingent. The contingent category basically replaces the optional category put forth in previous accounts. The label *Contingent* was selected for mnemonic purposes reminding us that the cases that fall into this category can be explained *contingent* on factors beyond syntax. The factors under consideration include simple grammatical/lexical factors, phonetic factors, phonological factors, historical factors, prosodic factors, pragmatic factors, and individual speaker variation. The fact that liaison occurrences can be explained by looking at these additional factors allows us to divide the Contingen*t* category into three sub-categories ALWAYS, NEVER, UNCLASSIFIED. The primary basis for categorization into the three major categories is syntactic sequencing, while the basis of categorization for the contingent sub-categories may be any one of the additional factors mentioned above. Sub-categories ALWAYS and NEVER refer to the cases where a liaison will always or never occur contingent on some factor/s beyond syntactic sequencing. At this stage of the study, UNCLASSIFIED remains an escape-hatch of sorts for tokens I am still unable to explain, due mostly to lack of data. With respect to this study, nothing was considered a performance error.

In the next section of this paper, I will discuss in detail the distribution of liaison in the context of the impersonal *c'est*. Liaison between *c'est* and a following constituent is *contingent* in French. I discuss these contingent cases and show how these cases can be generalized and accounted for by considering the effects of factors beyond word boundaries and syntactic domains. Previous accounts have placed *c'est* in the optional category because it has not been possible to predict its distribution within the parameters of the models proposed. Because of the approach taken in this paper, "prediction" of liaison behavior is not a goal; rather, the concern in this paper focuses on "explanation". We are specifically interested in explaining this corpus of data. I will show that for the most part, what has been considered by many to be a problematic glich in liaison literature is not so difficult to explain, if we are willing to complicate our account a bit by considering factors beyond the grammatical component which are relevant and important in characterizing the behavior of this phonological phenomenon.

3.2 THE ANALYSIS OF IMPERSONAL C'EST

The impersonal *c'est* corresponds very roughly to "that's..." or "that is..." , where the impersonal *ce* contracts with the copula *etre* and is followed by prepositional phrases, adjective phrases, or noun phrases. On a strictly sequential level, *c'est* participates in the following configurations:

1) C'est + preposition
2) C'est + interrogative
3) C'est + det + X
4) C'est + noun
5) C'est + adjective
6) C'est + adverb
7) C'est + conjunction

In the corpus under investigation, there are 32 cases in which a liaison occurs between *c'est* and the following constituent, and there are 25 cases in which liaison does not occur. While Selkirk's proposal would licence liaison between *c'est* and a following constituent, her proposal falls short of being able to predict which of the 57 cases would liaise and which would not. She acknowledges its unpredictable behavior by delegating *c'est* to the optional category. In this way, she escapes having to offer an explanation for something her model clearly cannot explain.

---

[5]Initial organization based on syntactic sequencing was done for simplicity's sake. With the amount of data under investigation, it was simply easier to break it up into syntactic sequences. I make no claim at this point in the study regarding the specific role syntax plays in prosodic structuring. As the study matures, we will be better able to ascertain where syntax fits in, if at all.

These cases *can* be explained, however, by looking beyond the model. The table below categorizes the fifty-seven tokens of *c'est* in liaison environments. Consider the table below.

Table 1. NON-LIAISON AND LIAISON ENVIRONMENTS WITH IMPERSONAL *C'EST* : ORGANIZED SEQUENTIALLY CATEGORY BY CATEGORY

C'EST + PP	C'EST + PP
C'est ll a toi C'est ll en rafia C'est ll en fin de compte C'est ll apres justement C'est ll en Farenheit	
C'EST + INTERROGATIVE	C'EST + INTERROGATIVE
C'est ll ou?	
C'EST + NP	C'EST + NP
C'est ll un^apartement C'est ll un ^autre boulet C'est ll un, un tres beau, ..parterre C'est ll un TOAST	C'est^un beau cadeau "that's a beautiful present" C'est^un gaspillage C'est^un chose C'est^un fruit C'est^une piece C'est^un musicien C'est^un sorte de soupe
C'EST+ NOUN W/O ARTICLE	C'EST + NOUN W/O ARTICLE
C'est ll Uptown C'est ll "ee" C"est ll "e" C'est ll autre chose	C'est^elle "that's her"

C'EST + ADJECTIVE	C'EST + ADJECTIVE
C'est ll incroyable "That's unbelievable" C'est ll horrible C'est ll horrible C'est ll horrible C'est ll agreable C'est ll extrodinaire	C'est^affreaux C'est^enorme C'est^incroyable C'est^imbecil Il est^enorme Il est^evident C'est^interessant C'est^excellent C'est^international

C'EST + ADVERB	C'EST + ADVERB
C'est ll encore autre chose C'est ll enfin tu.. C'est ll assez cher "that's expensive enough" C'est ll un peu comme lutherens C'est ll un petite peu different C'est ll un peu comme Jean Marc "that's a little like Jean Marc"	C'est^effectivement bon a jeter C'est^effectivement un'autre "that's effectively another" C'est^assez connue "that's well known enough"
C'EST + CONJUNCTION	C'EST + CONJUNCTION
C'est ll aussi un peu Francais "that's also sort of French"	C'est^aussi la vie "that's also life"

What is most noticeable about the above set of data is that there are no cases in which a liaison *always* occurs in a given sequence. There are, however, certain sequences in which a liaison will *never* occur — namely, when *c'est* is conjoined with a preposition, or an interrogative. Therefore, we can predictably account for six of the fifty-seven tokens by referring to the syntactic sequencing at the grammatical category level.

Sequencing at this level will not, however, explain the behavior of liaison in the remaining fifty-one tokens. These are the cases where *c'est* is conjoined with determiners, nouns, adjectives, adverbs , and conjunctions. We begin our account by addressing *c'est* + determiner. Consider the table below.

TABLE 2. C'EST + DETERMINER + X

C'est ll un^apartement C'est ll un ^autre boulet C'est ll un, un tres beau, ..parterre C'est ll un TOAST	C'est^un beau cadeau C'est^un gaspillage C'est^un chose C'est^un fruit C'est^une piece C'est^un musicien C'est^un sorte de soupe

There are two things which determine whether or not liaison will surface in the above environment. The first is the initial segment of the noun, and the second is the status of the noun in the French language. Notice in Table 2 that liaison does not occur when the noun is vowel initial. All of the cases in which liaison occurs in this environment are cases in which the noun is consonant initial. One speculative explanation may be that sequences of liaisons are avoided in casual speech [6]. This is a simple explanation based on a pattern which could be generalized from looking at the data inductively. [7] This explanation is not a valid one in the simpler more elegant models where such patterns and generalizations are not recognized. In those models impersonal *c'est* must necessarily be placed in the optional category along with the other cases in which no "predictability" is possible. We can begin to see how such models do not stand a great of a chance of shedding much light on the real nature of these phonological phenomenon.

The other case in which liaison does not occur in the *c'est* + determiner environment is where the noun is a loan word. Loan words appear to have a different status in French. It is consistent with the rest of the data (ie., the larger corpus) that liaison behavior with loan words is distinct from liaison behavior of French words. In fact, there is no case in which a loan word liaises with anything. While it is the determiner which is in liaison relationship with the *c'est* constituent, it is possible, again speculative, that the loan word is indirectly effecting liaison relationship. We

[6]This is one of the things that we want to test with native speakers. How are multiple liaison sequences perceived? Are they natural or marked in casual conversation? If they are marked, what do they mark, and when would they be likely to occur?

[7] By "generalize", I mean to simply make a statement about the data under investigation. Due to the size of the corpus, qualitative generalizations are not specific goals of this study.

have only one token of this type, so more than a speculative account is irresponsible. What is important here is not that we arrive at the exact reason for the lack of liaison, but rather that we arrive at *a* reason. That is to say, we can clearly show that the distribution of liaison in the *c'est* + noun phrase context *can*, in fact, be explained contingent on factors beyond syntactic domains. In this case, the factors include phonological conditions influenced possibly by stylistic considerations and lexical status. Table 3 shows liaison behavior with *c'est* + noun without an article. Here we use different criteria in explaining the distribution.

TABLE 3. C'EST + NOUN (W/O ARTICLE)

C'est ll Uptown C'est ll "ee" (fem. past participle morpheme) C"est ll "e" (masc. past participle morpheme) C'est ll autre chose	C'est^elle

In case there is no article associated with the noun following *c'est* , liaison may or may not occur between *c'est* and the noun in question. As we discussed above, loan words don't liaise. So, in the case of the loan word "Uptown", no liaison takes place between *c'est* and Uptown. Secondly, a liaison will not occur between *c'est* and a so called open class of nouns which surface without an article. This is shown by the examples with "ee" and "e".[8] The generalization that can be made here is that if the lack of article is optional, as it is in the cases of "ee", "e" and "autre chose", then liaison does not occur[9]. It must be emphasized that the generalizations that are made here are in no way definitive. The opposite is true. With more data, it may very well be that these generalizations won't hold at all. Crucial here is that a generalization, however spurious, is observed.

The final generalization that can be made with respect to this set of data is that a liaison will occur in cases where the noun is a pronoun which cannot take an article. Now let us consider what happens when *c'est* comes together with a preposition. See table 4 below.

TABLE 4. C'EST + PREPOSITION

C'est ll a toi C'est ll en rafia C'est ll en fin de compte C'est ll apres justement C'est ll en Farenheit	

Like loan words, prepositions behave consistently throughout the corpus. That is to say, prepositions do not liaise with anything directly to their left, which in this case includes the impersonal *c'est* [10]. In fact, if we look back at table (1) we will find that the other examples of cases in which *c'est* never liaises with either a certain category or word are all cases which behave consistently throughout the corpus. That is to say, interrogatives never liaise, and *un peu* , funnily enough, never liaises. These factors appear to be more language specific rather than case specific. Hence, these factors which explain the distribution of liaison in the context of *c'est*+ preposition are more generalized than the factors involved with the other syntactic sequences in the context of *c'est*. This is primarily because the syntactic sequence itself explains the occurrence of liaison in cases like *c'est* + preposition, or interrogative, where as factors *beyond* the syntactic sequence explain the occurrence of liaison in the cases involving noun phrases, nouns, etc..

Now let us consider a new set of data, namely tokens of *c'est* + adjective. Study Table 5 below.

[8]"ee" and "e" are orthographic representations of the past participle morphemes. Both are pronounced /e/. The speaker is a French teacher who was describing the difficulty her students had with these morphemes. " That's "ee" and that's "e" ? "

[9]Actually, in the case of 'autre chose', I am not sure whether or not that is idiomatic. This needs to be checked out with a native speaker. If it is the case, then a different generalization needs to be made. Really, the crucial point is that we can make generalizations on some level, and thus account for in some crude fashion the distribution of these cases of liaison.

[10]The exception to this is 'a', which is discussed in the longer version of this paper (Carleton, 1992).

Table 5. C'EST + ADJECTIVE

C'est ll incroyable	C'est^affreaux
C'est ll horrible	C'est^enorme
C'est ll horrible	C'est^incroyable
C'est ll horrible	C'est^imbecil
C'est ll agreable	Il est^enorme
C'est ll extrodinaire	Il est^evident
	C'est^interessant
	C'est^excellent
	C'est^international

C'est + Adjective poses an interesting twist to our story so far. There is no clear cut distinction between those cases in which we get liaison and those cases in which we do not. One generalization that can be made is that no liaisons occur with "h" initial adjectives. They may be due to hypercorrection of the aspirated "h" liaison prohibition, an historical influence on French phonology. This, however, does not explain why the other three cases show up without a liaison — namely, *c'est* + e*xtrodinaire, incroyable* and *agreable.*

One generalization that can be made about the above cases is that all of the tokens in which no liaison occurs between the *c'est* and the adjective are spoken by the same speaker. It is also the case that all but one of the cases in which there *is* a liaison between *c'est* and the adjective are spoken by the other two speakers. In fact the generalization can be supported further by the fact that the exception to liaison occurrence in other contexts can consistently be traced to this same speaker. In general, her speech is very fast and not well articulated. She tends to reduce her words and mutter. It seems to be true that faster speech is more likely to drop a liaison than slower speech. In this case, we must attribute the lack of liaison to individual speaker style. Therefore, we will assume that the *unmarked* case liaises *c'est* with an adjective. Individual speaker style can then override the positive liaison status that this context projects.

Now let us turn to the context of *c'est* + adverbials. Consider table 6.

TABLE 6. C'EST + ADVERB

C'est ll encore autre chose	C'est^effectivement bon a jeter
C'est ll enfin tu..	C'est^effectivement un'autre
C'est ll assez cher	C'est^assez connue
C'est ll un peu comme lutherens	
C'est ll un peu comme Jean Marc	
C'est ll un petite peu different	

The first generalization that can be made with respect to liaison status between *c'est* and adverbs is that lexical adverbs always liaise in this context. With respect to non-lexical adverbs, or closed class adverbs, it is the case that with the exception of *assez* closed class adverbs never liaise with *c'est.* In fact, a further generalization can be made about closed class adverbs, namely that there is a set of closed class adverbs which never liaise in any context. Like prepositions, they are language specific, not context specific factors. These adverbs include all temporal adverbs, like *ici, aurjourd'hui, hier,* etc.. In addition, this set includes *encore,* and as mentioned previously, *un peu.*

Assez poses an interesting case because we must consider influencing factors which are much more deeply embedded. Consistent with liaison exception between *c'est* and adjectives, the same speaker who didn't liaise with adjectives, does not liaise here with *assez. Assez* has rather complicated behavior, and time did not permit a full investigation into its distribution. However, for the purpose of the analysis of the distribution of liaison in the *c'est* environment, we can speculatively attribute the liaised and non-liaised *assez* to both individual speaker variation, and pragmatics.

In the case where we have no liaison between *c'est* and *assez*, the speaker is reporting something about the price of apartments in Minneapolis. Everyone is comparing rents and she is contributing to the general conversation. She has the floor, but she does not appear really invested in the subject. The topic is impersonal. In the case where liaison does occur between *c'est* and *assez*, the speaker is drawing focus upon the fact that a cooking school that she went to was well known, and that, therefore, qualifies her as a good cooking teacher. In this case, she most definitely holds the floor, and plans to hold the floor for a lot longer, since the information about here cooking school is simply lead-in information to what she wants to talk about, which is her cooking.

The case of *assez* is a nice one because it really points out the complexity of the variables involved in whether or not a liaison occurs. The example I have in my data is not enough to draw any conclusions with respect to *assez*, however, it is enough to show that any theory which ignores pragmatic interpretation and speaker variation will fail to capture many generalizations or patterns in the distribution of these liaisons. These theories, will, in fact, have no alternative but to classify these cases as optional and unpredictable. As it is, there are so many variables involved that it will take careful investigation of a much larger corpus to tease out exactly which variables are the crucial ones, and where they stand in relation to one another with respect to this context.

Let us now finally turn to the last two sequences under investigation here, namely *c'est* + conjunction, and *c'est* + interrogative. Consider Table 7.

Table 7. C'EST + CONJUNCTION

C'est ll aussi un peu Francais	C'est^aussi la vie

This case is a bit similar to the *assez* case. Again, it is the same speaker who never liaises who doesn't liaise here. In addition, it is the same speaker who liaises most frequently who liaises in this case. It is clear from the rest of the corpus however, that it is the marked case to liaise in this context. It is hard to say in this case whether or not the lack of liaison here is due to speaker variation. There is just too little data. Speaker variation is, after all, the last resort taken for explaining liaison distribution. The best we can do in this case is choose one of the variables and categorize it.

Table 8. C'EST + INTERROGATIVE

C'est ll ou?	

Interrogatives behave the same way as prepositions. The only thing to note in this case is that the prohibition against liaising with an interrogative has always been considered from the right end only. That is to say, descriptions have traditionally noted that interrogative + X does not liaise, not that X + interrogative does not liaise. Given the data above, which gives us a case of X + interrogative, it is clear that liaison does not occur on either side of the interrogative.

This concludes our discussion of the impersonal *c'est*. What I hoped to show by discussing this was that the "optional " category is not "optional" at all, but is better described by the term "contingent". In being contingent, those contexts which fall into this category are influenced by factors beyond syntax and word boundaries. They are no less explainable than other liaison contexts; they simply depend upon more factors. As we saw, some of the contexts were easily explained based on categorical or lexical sequencing. Others were dependent on more context specific factors like phonological conditioning, as we saw with vowel initial and consonant initial nouns, or on lexical status of the word in French, as we saw with loan words. Still other factors which influence the distribution of liaison include individual speaker variation, and pragmatics. Certain factors like speaker variation, pragmatics, and phonetic effects like speech rate and reduction can override liaison status in the contingent category (ex. speaker variation overrides positive liaison status of *c'est* + adjective). These factors are much less generalized than factors like lexical status, phonological factors, and historical factors, which are in turn less generalized than syntactic sequencing. What is important here is 1) there seem to be distinct levels of prominence concerning factors which influence liaison activity, some being much easier to define than others, 2) the factors involved cover a wide range of linguistic considerations and, 3) it is the presence or absence of these factors which help us arrive at a plausible explanation with respect to the distribution of liaison.

5. CONCLUSION

Having looked at the impersonal *c'est* in detail , and having done the preliminary work of developing an accurate typology and description of the distribution of liaison in natural discourse, I think we can clearly see that models which attempt to account for the distribution of liaison in a simple and more elegant manner will most definitely miss crucial generalizations, and thus, be unable to account for what is really happening. I have shown that we can account for liaison much more adequately by taking an inductive approach. As this study continues, one very important step will be to look more systematically at how all of the various factors apply to each and every token. In addition, a larger corpus will help us to see a great deal more about the nature of liaison.

Natural discourse is messy. It is only by looking at natural discourse that we can clearly see where the flaws of the simpler theoretical models are. A clear and accurate description of the facts at hand gives us the real and complete picture of what we have to account for. This study simply shows where some of the flaws in previous studies surface when put to the test of having to account for natural discourse. In order to continue to make progress in this area, it is clear that we must challenge all current theories to stand up to the rigor of language in use. This study and studies like this are a challenge to both current prosodic theories, and formal linguistic theories in general.

6. REFERENCES

Carleton, T.C. (1992b). "An Inductive Approach to the French Liaison in Natural Discourse." unpublished manuscript, University of Texas at Austin.

Chomsky, N. and M. Halle (1968). *The Sound Pattern of English.* New York: Harper & Row.

Delattre, P. (1947). "La liaison en francais, tendances et classifcation." *French Review* 21. 148-157.

Delattre, P. (1955). "Les facters de la liaison facultative en francais." *French Rreview* 29.42-49.

Delattre, P. (1956). "La frequence des liaisons facultatives en francais." *French Review* 30. 48-54.

Delattre, P. (1966). *Studies in French and comparative phonetics.* The Hague: Mouton.

Morin, Y.-C., and J.D. Kaye. (1982). "The Syntactic Bases for French Liaison." *Journal of Linguistics* 18. 291-330.

Selkirk, E. (1981). *The Phrase Phonology of English and French.* Indiana University Linguistics Club; Bloomington.

Selkirk, E. (1974) "French Liaison and the X-bar Convention." *Linguisitc Inquiry* 5. 573-590.

INTONATION UNITS AND PROMINENCES IN ENGLISH NATURAL DISCOURSE

Wallace Chafe

Department of Linguistics
University of California
Santa Barbara, CA 93106

ABSTRACT

This paper emphasizes the need to understand prosody with relation to all of language functioning, and to take account of all observable properties of sound, including the basic dimensions of frequency, intensity, and duration, but also various derived properties such as tempo and voice quality, while at the same time attending to both the physical and perceptual manifestations of these properties. The discussion focuses (1) on the identification of intonation units, which are seen as reflecting a universal inability of consciousness to focus on more than a small amount of information at a time, and (2) on the complexly interrelated acoustic and perceptual manifestations of prominences within intonation units, as well as the functions of these prominences.

INTRODUCTION

One's understanding of prosody is inevitably colored by the direction from which one approaches it. A great deal of the recent literature seems to assume without question that prosody is a part or extension of a "phonological component" that is attached to a "syntactic component" within a generative model of language. For one who finds the generative approach to be fundamentally flawed it is difficult to relate to the literature that assumes it. In what follows I will be coming at prosody from a different direction, while continuing to hope for an eventual reconciliation of any and all approaches within a more complete picture of what prosody is and how it functions.

For many years I have listened to and attempted to transcribe and analyze tape-recorded conversations, stories, and rituals in several languages -- most extensively in English and two American Indian languages, Seneca and Caddo. This activity belongs to a long and honorable tradition in which prosodic concerns have always had an important place, even if they have not always been clearly articulated. I was taught from the beginning to pay close attention to accents, pauses, and pitch contours that seemed to contribute to the form and function of language. Most of these observations were made "by ear", although on various occasions I consulted spectrograms that helped to clarify both segmental and prosodic phenomena. Lately I have been more than a little grateful for the ability to observe fundamental frequency, intensity, and duration more easily than was possible in the days when spectrograms were all we had.

It was always clear, even with respect to segmental phenomena, that visual displays of acoustic data never provided all the answers; that there was no one-to-one relation between physical sound and the way we perceive it; that our auditory apparatus and our brains do not simply record sound, but interpret it. That is certainly as much the case with prosody as it is with vowels and consonants. Furthermore, it is plausible to suppose that what is ultimately functional in language is what we perceive, and not the physical sound. It follows that if transcriptions and analyses are intended to capture what is functionally relevant, they need to give precedence to perceptual observations. But there is an interesting problem here, stemming from the fact that it is not always a simple matter to become consciously aware of what it is that we perceive. That fact becomes obvious as soon as one spends a few minutes with students who have trouble distinguishing rising from falling pitches, even though they know there is some kind of difference. The ability to pinpoint the nature of perceived prosody appears to be a skill, as much as the ability to perform phonetic transcription of any kind, or the ability to transcribe music in a musical dictation class.

Visual displays of sound can be a great help in this situation, because they make it possible to "see" whether a pitch goes up or down, whether one syllable is louder or longer than another, the precise length of a pause, the location of a breath, the beginning and end of overlapping speech, and so on. If these representations do not tell us what we perceive, at least

they show what enters our ears. One of the major current needs in prosodic research is to establish the relation between acoustic prosody and perceptual prosody. The more we know about this, the better we will be able to make sense of acoustic data. Even now, however, acoustic displays provide an invaluable way to sharpen our understanding of what we hear.

One of my major interests for some time has been to relate the flow of language to the flow of conscious experience. It seems obvious that both our thoughts and the language that organizes and expresses them are in constant flux, and that changes in the one are inseparably linked to changes in the other. Within this picture there is a crucial place for the kind of linguistic unit my colleagues and I have been calling an *intonation unit.* Intonation units emerge with relatively satisfying clarity from natural speech. I say "relatively" because, while their boundaries are easy to determine in the majority of cases, there remain other cases where the evidence for them is less than overwhelming. One of my aims here is to review the kinds of evidence that can lead us to identify the boundaries of intonation units, even in the more difficult cases.

I believe that the importance of intonation units stems above all from the hypothesis that each such unit expresses the information on which a speaker is focusing his or her consciousness at the moment the intonation unit is being produced, information the speaker hopes to introduce into the consciousness of the listener. From a careful study of intonation units, furthermore, it appears that each one expresses no more than one new idea. There is no room here to explore the ramifications of this "one new idea hypothesis" in detail, but attempts to clarify the meanings of "one", "new", and "idea" have shed interesting light on several important aspects of conceptualization and language. Here I can only allude to the usefulness of intonation units as catalysts to a variety of related discoveries.[1][2]

It appears that intonation units are only one of several levels that are defined by prosodic aspects of natural speech. More inclusive units delimited by pauses, inhalations, declining levels of pitch and loudness, shifts to and from higher or lower baselines of pitch and loudness, various changes in voice quality, and probably other criteria may all have functional significance of some kind as well. Here we will notice some of these other units in passing, but the discussion will be centered on intonation units.

Within intonation units, some words or syllables are more prominent than others. Prominence may be a matter of significantly higher or lower pitch, lengthening, loudness, or some combination of these properties. Among other things, prominence signals "non-given" information -- information the speaker judges not to be in the focus of the listener's active consciousness at the moment. I will point to examples of this function, along with a few others.

Besides delimiting separate foci of consciousness and signaling properties such as non-givenness, prosody also conveys what might in a most general way be called *attitudes.* An obvious case is the distinction between the high rising terminal contour used in English for the expression of yes-no questions, in contrast to the falling contour used for assertions and question-word questions. But there are a variety of other contours that express a variety of other attitudes, most of which are more difficult to characterize than those just mentioned. This aspect of prosody seems particularly resistant to systematic understanding, and it is another area where future research can be especially rewarding.

INTONATION UNITS

Observations of natural speech confirm the ubiquitous presence of intonation units, as well as their association with minimal chunks of information. The general picture is one in which a speaker focuses on the idea of an event or state or isolated referent, or sometimes on nothing more than an attitude or a connection between successive ideas. The speaker may then express that idea, attitude, or connection in an intonation unit that may be clearly defined by a variety of prosodic criteria or, in cases where the idea is especially closely related to an adjacent idea, by only one or two criteria. But intonation units are always bounded prosodically in some respect. If they were not, we could not say that the flow of separate ideas through consciousness is consistently reflected in the flow of prosody.

The features that characterize intonation units may relate to duration (perceived in terms of tempo and

lengthening), fundamental frequency (perceived as pitch), intensity (perceived as loudness), voice quality, the alternation between silence and vocalization, and/or change of turn. Examples of these features will be given here in two formats: the transcriptions that are cited in the text, and the displays of acoustic data that are available in Figures 1-5 at the end. The transcription conventions are largely, but not entirely those set forth in [3]. The figures were produced by the Summer Institute of Linguistics' CECIL system. It would be especially useful to hear the examples as well, but within the wholly visual context of a written article the transcriptions and figures will have to suffice.

Figure 1 shows a well-defined intonation unit whose boundaries are confirmed in a variety of ways. I will transcribe it as follows, using conventions that will be explained as we proceed:

(1) a .. and so the **háll** is réal ló=ng%.
b ...(.36) [next intonation unit]

Preceding the vocalization in (1)a is a very brief pause of about .07sec. We have been transcribing pauses of less than .10sec simply with two dots. Following (1)a is a much longer pause of .36sec, transcribed with three dots followed in parentheses by a measurement of its length. (An accuracy to hundredths of a second seems appropriate for such measurements.) By convention, boundary pauses are shown at the beginning of each intonation unit. Among other things, then, (1)a is set off by pauses.

One of the major cues to intonation unit boundaries is tempo, captured to at least some degree in the notion of "anacrusis".[4] Intonation unit (1)a begins with a sequence of three accelerated syllables (*and so the*) occupying roughly .10sec each. I have transcribed accelerated syllables in smaller type. After that there are two words (*hall* and *real*, separated by a rapid *is*) whose duration lies in the range from about .20 to .30sec, a normal length for one-syllable words. The intonation unit ends with a word of extended length (*long*), occupying .43sec, as shown with the equals sign. This pattern of *acceleration-deceleration*, proceeding from reduced length syllables up to about .15sec, through normal length syllables from about .15sec to .35sec, to extended length syllables longer than .35sec, is characteristic of many intonation units, and may in some instances be the best evidence for an intonation unit boundary.

When it comes to pitch, it happens that (1)a coincides with a "declination unit".[5] There are three words with high pitch (*hall*, *real*, and *long*), with a decline in the pitch of each (maxima of 299Hz, 211Hz, and 192Hz respectively), a good illustration of what is often called downstep. (I will return below to the unusually high pitch on *hall*.) It is noteworthy that the mid pitches of the accelerated first three words of this intonation unit are equal to or even higher than the high pitch on the last word. In addition to the declining pitch throughout, the end of the intonation unit shows a falling terminal contour, transcribed with a period.

One of the prosodic features in the category of voice quality is creaky voice, laryngealization, or vocal fry. It is conspicuous here at the end of the word *long*, where it is transcribed with a percent sign. Intonation units often end and sometimes begin with creaky voice, which thus provides still another clue to their boundaries.

In short, the identification of 1(a) as a coherent intonation unit is supported by a convergence of (1) the pauses preceding and following it, (2) the tempo of acceleration-deceleration, (3) the decline in pitch level, (4) the falling pitch contour at the end, and (5) the creaky voice at the end.

Intonation units are not always this well defined, and the example that follows was chosen in part to illustrate some less clear cases. It consists of eleven intonation units distributed within three breath units (a-f, g-h, and j) that are bounded by inhalation pauses lasting between .40sec and .50sec. The sequence belongs almost entirely to Speaker A, who does all the talking except for two brief responses from Speaker B (in i and k). Speaker A is relating her experiences in helping a friend move from one apartment to another, and here she is listing some of the factors that contributed to the confusion surrounding the move:

(2) a(A) ...(h)(.42) **Plus** the twó **dò=gs-**
b(A) and the **cà=t-**
c(A) %and the **kíds=,**
d(A) .. and the **s=créaming-**
e(A) and **dróp**ping things,

f(A) <LO <P and só %it %was a real-
g(A) ...(h)(.40) %it was a %méss=.
h(A) %it %was [%a] real %méss=. P> LO>
i(B) [Huh.]
j(A) ...(h)(.49) So it tóok us a long time%=.
k(B) Yeah.

Figure 2 shows intonation units (2)a-c. After the inhalation pause, (2)a begins with a loud connective *plus* followed by a noun phrase (*the two dogs*) whose syllables become increasingly long up to the final word *dogs*, which lasts .50sec. That word is spoken with a low and level terminal pitch contour that seems to mean here something like 'this was still one more thing (that caused confusion)'. I have shown the level contour with the final hyphen (see below for the use of the grave accent mark on *dogs*). Intonation unit (2)b begins with two accelerated syllables (*and the*) followed by the lengthened word *cat*, which shows the same low and level contour as (2)a. Intonation unit (2)c has the same accelerated-decelerated tempo, with the accelerated words *and the* followed by a lengthened *kids*, where much of the length is in the final sibilant. The terminal contour of (2)c is different from that of (2)a and (2)b, consisting of a fall-rise (shown with the comma), a more standard list intonation that assigns the kids to a different category from the dogs and the cat. An additional boundary feature in (2)c is the creaky voice on its initial word *and*.

Figure 3 shows intonation units (2)d-f. After a brief pause, (2)d continues the acceleration-deceleration pattern, with extended length focused on the initial consonant and first syllable of the word *screaming*. However, the final syllable of *screaming* is long enough to accommodate the terminal pitch contour, which represents a return to the low level pattern of (2)a and (2)b. Without the duration and final pitch pattern of the final syllable of (2)d, (2)d and (2)e together would probably be perceived as a single intonation unit. Intonation unit (2)e follows the same general pattern, though the tempo differences are not as great. What seems to be significant in (2)e is the fact that the two final unaccented syllables (*ping things*) are as long as the accented syllable (*drop*). As a result, the final syllable has room to accommodate the same fall-rise terminal contour that we saw in (2)c. Intonation unit (2)f begins with five accelerated syllables (*and so it was a*), followed by a single normal-length syllable (*real*), which carries a level terminal contour. The beginning of (2)f is signaled further by a drop to a low pitch level (whose extend is shown by "LO" in angle brackets) and low volume (shown by "P" for *piano* in angle brackets), both of which continue until the end of (2)h. The reduced vigor of (2)f is also manifested in the creaky voice on the words *it was*. With this attempt at a final summing up, the speaker was approaching the end of a larger, more sentence-like unit. It is worth noting, however, that even with allowance for the truncated nature of (2)f, the syntax of (2)a-f (or even of (2)a-h) is not what would ordinarily be expected of a sentence.

Figure 4 shows intonation units (2)g-h, which begin with an inhalation pause. The low pitch and volume levels introduced in (2)f are continued. In both of these intonation units everything is accelerated except for the extended final word (*mess* in both). Both show a falling terminal contour. Both begin and end with creaky voice. Intonation unit (2)i is a backchannel *huh* by Speaker B, coinciding with Speaker A's indefinite article in (2)h.

Figure 5 shows intonation units (2)j-k, which also begin with an inhalation pause. After a normal length *so*, (2)j contains four accelerated syllables followed by a normal-length *long* and then an extended-length *time*, which is creaky at the end. It ends with the same falling pitch contour as (2)g and (2)h. It is immediately followed by (2)i, consisting of Speaker B's *yeah*, whose much lower pitch reflects the fact that Speaker B was a man.

In short, one of the most consistent features of these intonation units is the pattern of acceleration-deceleration. Each intonation unit also ends in an identifiable terminal pitch contour. Several end and/or begin with creaky voice. The insertion of an inhalation pause sets off the beginnings of a, g, and j, and the ends of f and h. A low level of pitch and loudness sets off f-h from e and j. Finally, the backchannel response in k also serves to terminate j. The entire sequence demonstrates well how the boundaries of intonation units are identifiable on multiple grounds, and further that while some boundaries are signaled in multiple ways, others are signaled by only one or two of the features discussed.

The division of (2) into intonation units illustrates well

the restriction of intonation units to no more than one new idea, particularly with the list of confusions in (2)a-e, each of which was a new idea and each of which apparently had to be expressed in a separate intonation unit for that reason. The longer intonation unit in (2)j shows how a separate verb (*took*) and object (*a long time*) may combine to express a single lexicalized idea (*took a long time*) without violating the one new idea constraint. Lexicalization is one of several important aspects of language and thought on which the one new idea hypothesis sheds useful light.

It is worth noting that the length of intonation units, as most easily measured by the number of words per intonation unit, is strongly limited by the one new idea constraint. In English, substantive intonation units (excluding backchannel responses and other units that serve only to regulate the flow of information) have a rather sharply defined modal length of four words. Other prosodically defined segments of discourse -- breath units, declination units, prosodic sentences defined by final falling pitch, etc. -- are more variable in length and content. They appear to be less dependent on any wired-in constraint, but rather to involve various kinds and degrees of coherence between the minimal foci of consciousness that are verbalized in intonation units. For example, of the three breath units in (2), the first is centered on listing the confusions:

(2) a(A) ...(h)(.42) **Plus** the twó **dò=gs-**
b(A) and the **cà=t-**
c(A) %and the **kíds=,**
d(A) .. and the **s=créaming-**
e(A) and **dróp**ping things,
f(A) <LO <P and **só** %it %was a real-

the second succeeds in articulating the evaluation that was too hastily attempted in (2)f:

(2) g(A) ...(h)(.40) %it was a %méss=.
h(A) %it %was [%a] real %méss=. P> LO>

and the third focuses briefly on another aspect of the total experience:

(2) j(A) ...(h)(.49) **So** it tóok us a long time%=.

It may be noted that the first and second of these breath units together constitute a single prosodic sentence -- a coherent depiction and summarization of the experience. Intonation units, then, reflect a strict and unavoidable constraint on the flow of discourse, and ultimately on the flow of consciousness. Larger prosodic units, defined in other ways, are more variably determined by the structure of what is being talked about.

PROMINENCES

Besides perceiving speech as segmented into intonation units, we perceive certain elements within an intonation unit as more prominent than others. The acoustic correlates of prominence are also complex and variable. There are degrees of prominence, and there are several ways in which prominence may be realized. In what follows I will use the term *accent* for prominences that are realized as pitch deviations from a mid or neutral baseline, usually a higher pitch but sometimes a lower one. In the transcriptions I have represented such pitch deviations with accent marks: acute for a significantly higher pitch and grave for a significantly lower one. When one of these accented elements is also either lengthened or loud (or both), I will say that it has a primary accent. A pitch deviation alone, without lengthening or loudness, will be called a secondary accent. Of course an element may be either lengthened or loud without showing a pitch deviation, and in such cases I will say simply that it is lengthened or loud, but not accented.

As a first illustration of prominences, we can look again at the intonation unit that was cited in (1)a, repeated here (see Figure 1).

(1) a .. and so the **háll** is réal ló=ng%.

Three of these words -- *hall*, *real*, and *long* -- are accented, all showing high pitch (with downstep). However, not only is the pitch of *hall* inordinately higher than that of the other two words, to an extent that is unexplainable on the basis of downstep alone, but *hall* is also significantly louder, as shown by the boldface type. Its greater prominence in both respects can be attributed to the fact that it expresses contrastive information. It is not new information -- the idea of this hall was introduced eight intonation units earlier -- but here the hall is being contrasted with the living room, the bedroom, and the bathroom, all of which had been introduced in the meantime. It

is not unusual for contrastive elements to show exaggerated loudness and/or pitch.

The new (previously unactivated) information in (1)a is expressed by the predicate *is real long*, in which the heaviest load is carried by the word *long*, which is both high pitched and lengthened. (It is not accidental that *long* also occurs at the end of this intonation unit.) The intensifier *real* is high pitched but neither loud nor lengthened. Thus, we find three different manifestations of prominence in this intonation unit: the high pitched and loud *hall* expressing contrastive information; the high pitched, lengthened, and final *long* expressing new information; and the high pitched *real* intensifying the meaning of *long*. We can say that *hall* and *long* have primary accents (high pitch along with loudness or lengthening), as is typical for both new and contrastive information, and that *real* has a secondary accent (high pitch alone), as is often the case with modifiers of various kinds.

Turning now to sample (2) and Figure 2, we find the complexity of prominence well illustrated in its first intonation unit:

(2) a(A) ...(h)(.42) **Plus** the twó **dò=gs-**

The words *plus*, *two*, and *dogs* all have some kind of prominence in terms of duration, pitch, and/or loudness: *plus* is loud, *two* is high-pitched, and *dogs* is loud, low-pitched, and lengthened. Functionally, it would appear that *plus* is loud (but nothing else) because it serves as a connective that introduces a new breath unit and a new subtopic in the conversation. The word *dogs* has a multiply marked primary accent because it expresses new information. The word *two* has a high pitched secondary accent because it modifies the idea of the dogs, much as *real* modifies the idea of being long in (1)a. The fact that *dogs* is low-pitched rather than high-pitched may be attributed to the terminal contour, not the accenting per se. In other words, accented elements are typically high pitched, but may be low pitched when a terminal contour demands it.

The same pattern of accents, minus the numeral, is observable in intonation units (2)b-d (Figures 2 and 3). Something like this pattern is also present in (2)e, where the primary accent is on the categorization of the event as an instance of *dropping*, followed by a normal length but unaccented generalized object *things*:

(2) b(A) and the **cà=t-**
c(A) %and the **kíds=,**
d(A) .. and the **s=créaming-**
e(A) and **dróp**ping things,

In both (2)f and (2)j (Figures 3 and 4) it is interesting that the connective *so* is loud but lacking in pitch prominence, thus mirroring the loudness of *plus* in (2)a:

(2) a(A) ...(h)(.42) **Plus** the twó **dò=gs-**

(2) f(A) <LO <P and **só** %it %was a real-

(2) j(A) ...(h)(.49) **So** it tóok us a long %tíme=.

Intonation unit (2)g (Figure 4) shows a primary accent on the word *mess*, expressing in this case *accessible* (or reactivated) information, since the characterization of the move as a mess had already been established earlier in the conversation. *Mess* shows both pitch and durational prominence, but is not loud, presumably because the entire intonation unit is spoken with reduced volume:

(2) g(A) ...(h)(.40) %it was a %méss=.

Intonation unit (2)h reinforces (2)g by repeating it with the addition of the word *real*, which was of course already foreshadowed in the truncated (2)f:

(2) h(A) %it %was [%a] real %méss=. P> LO>

Intonation unit (2)j (Figure 5) communicates a new idea that is captured in the lexicalized phrase *take (someone) a long time*. As is typical of such phrases, the primary accent falls on the final content word (*time*), but there is a high pitch (with no other sign of prominence) on the other content word (*took*), which is part of the accelerated buildup. Thus, *took* can be said to have a secondary accent:

(2) j(A) ...(h)(.49) **So** it tóok us a long tíme%=.

In short, all of the substantive intonation units in this sample show primary accents at or near their conclusions. Primary accents express non-given (new

or accessible) or contrastive information, and contrastive elements may show exaggerated pitch deviation as well as loudness. This pattern fits the British view that tone groups (or whatever they may be called) build up to and trail away from a nuclear accent. On the other hand, the presence of two clear, if differently motivated primary accents in (1)a shows that an intonation unit can contain more than one nucleus. Some intonation units exhibit (pitch only) secondary accents on subsidiary content words (*real*, *two*, *took*). Introductory connectives like *plus* and *so* may be loud but unaccented in terms of pitch.

SUMMARY

It is probably the case that all languages are produced in the format of a succession of intonation units, each expressing no more than one new idea. These intonation units are produced with a tempo of acceleration followed by deceleration, with a set of intonation-unit-final pitch contours whose specific manifestations vary with the language, and less consistently with pausing, changes in the level of pitch and/or loudness, changes in voice quality, and sometimes a change in speaker.

Prominence is more of a mixed bag. In English, high (or sometimes low) pitch combined with extended length and/or loudness is regularly associated with the expression of non-given information as well as contrastiveness, the latter sometimes showing exaggerated pitch height. English has various intensifiers, enumerators, and low content verbs (like *real*, *two*, and *took*) that typically show high pitch alone -- what I am calling a secondary accent. Introductory connectives like *plus* and *so* may be prominent with respect to loudness alone. High pitch, loudness, and lengthening are also associated with the expression of affect. Prominence, in short, has a variety of forms and functions.

REFERENCES

[1] W. Chafe: "Cognitive constraints on information flow", in R. Tomlin (ed.), Coherence and Grounding in Discourse, pp. 21-51. Amsterdam: John Benjamins (1987)

[2] W. Chafe: Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. (to be published)

[3] J. Du Bois, S. Cumming and S. Schuetze-Coburn: Discourse Transcription (to be published)

[4] A. Cruttenden: Intonation. Cambridge University Press (1986)

[5] S. Schuetze-Coburn, M. Shapley, and E.G. Weber: "Units of intonation in discourse, a comparison of acoustic and auditory analyses", Language and Speech 34, pp. 207-234 (1991)

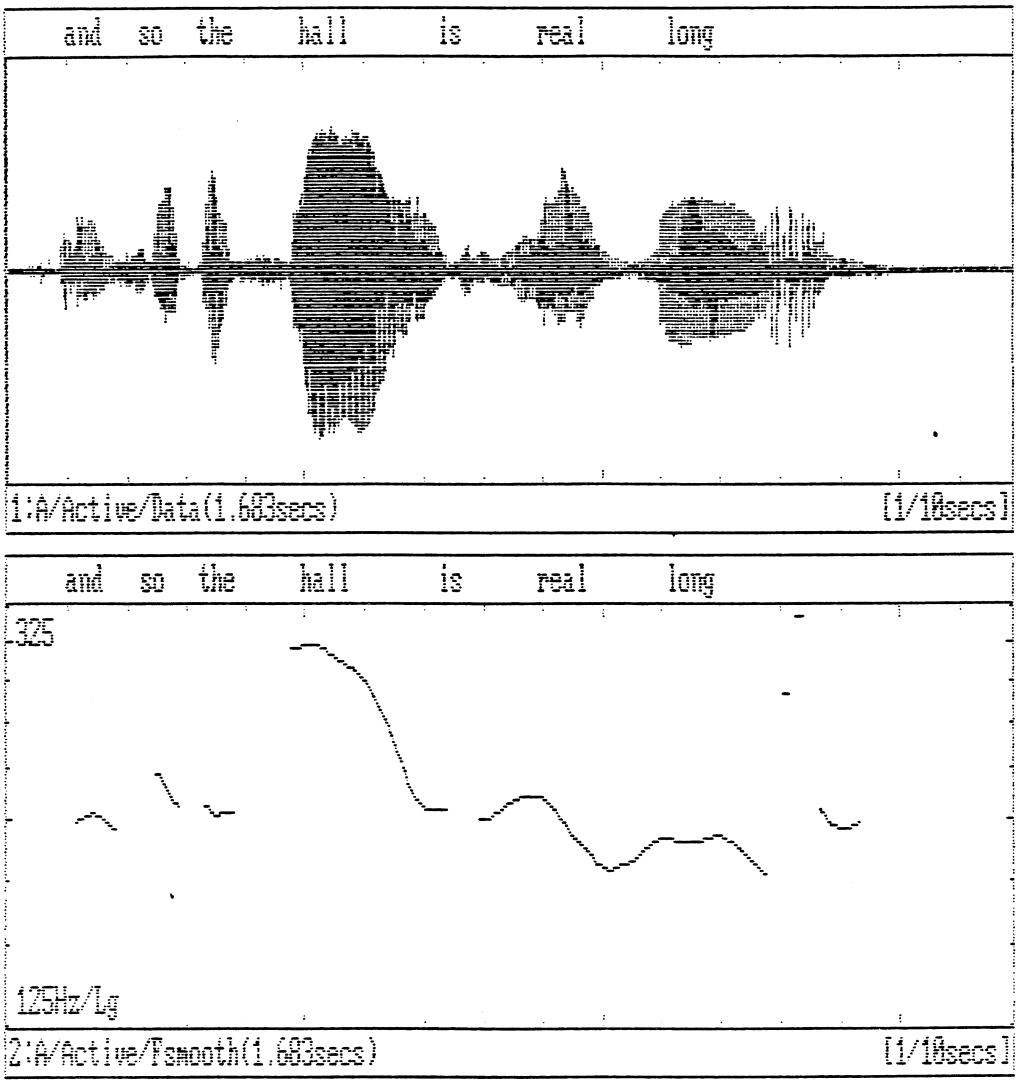


Figure 1

(1)a .. and so the **háll** is réal ló=ng%.

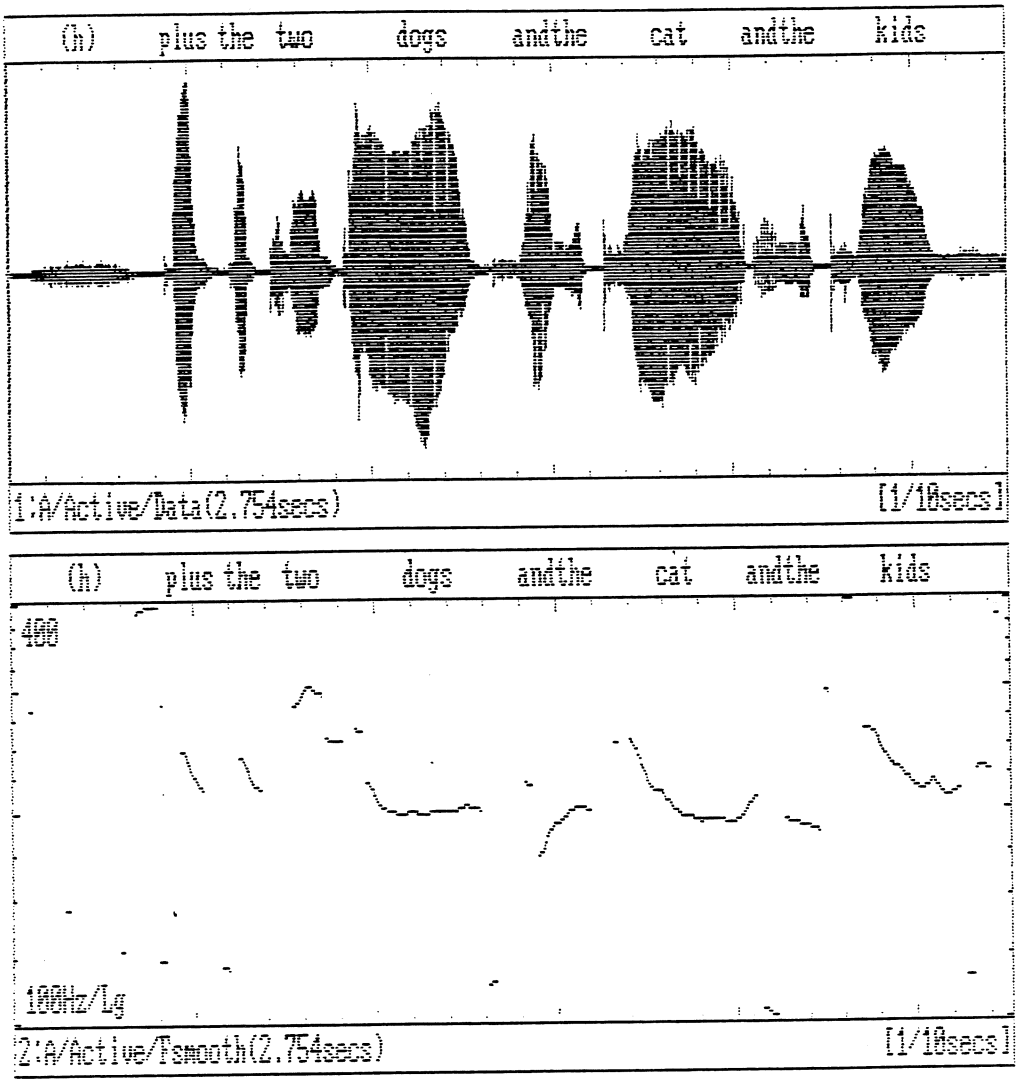


Figure 2

(2) a(A) ...(h)(.42) **Plus** the twó **dò=gs-**
b(A) and the **cà=t-**
c(A) %and the **kíds=,**

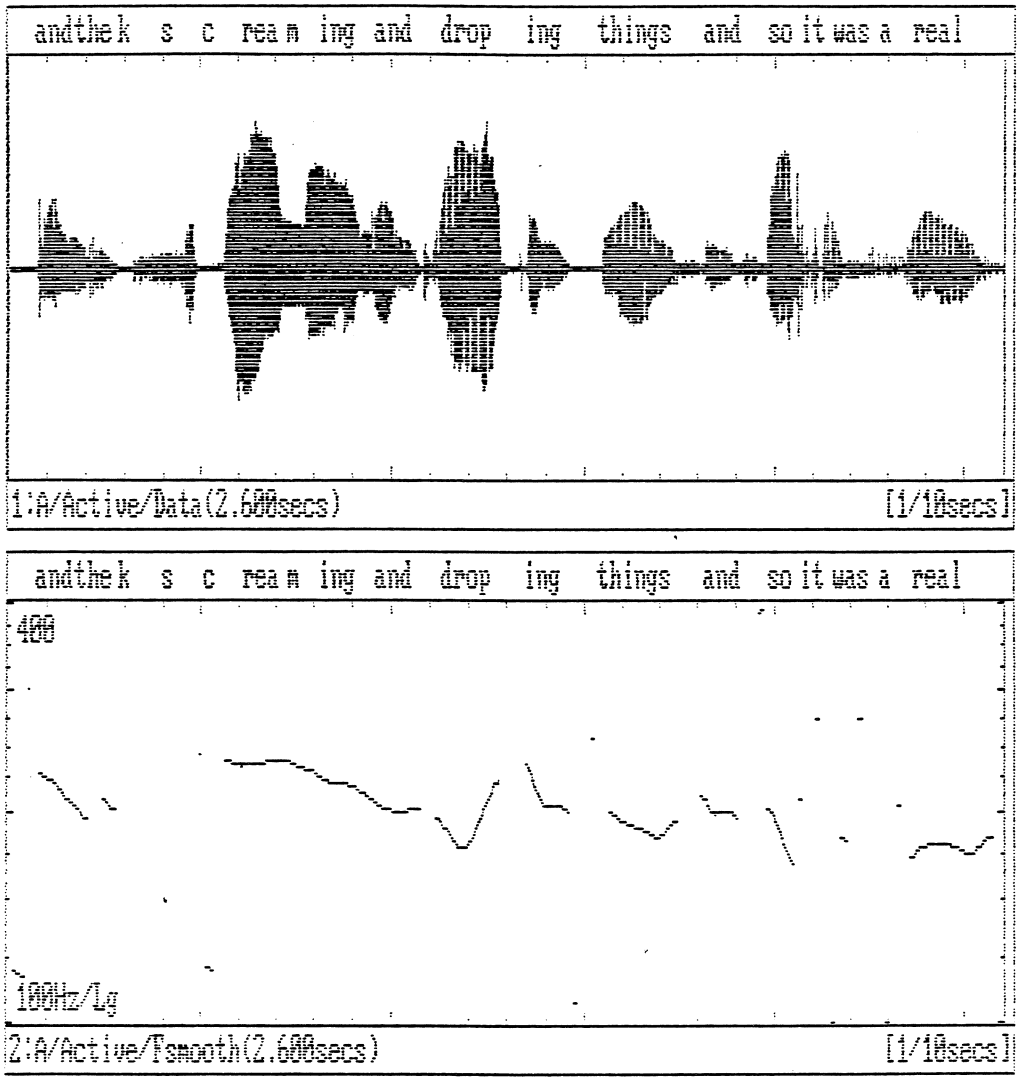


Figure 3

(2) d(A) .. and the **s=créaming-**
e(A) and **dróp**ping things,
f(A) <LO <P and só %it %was a real, —

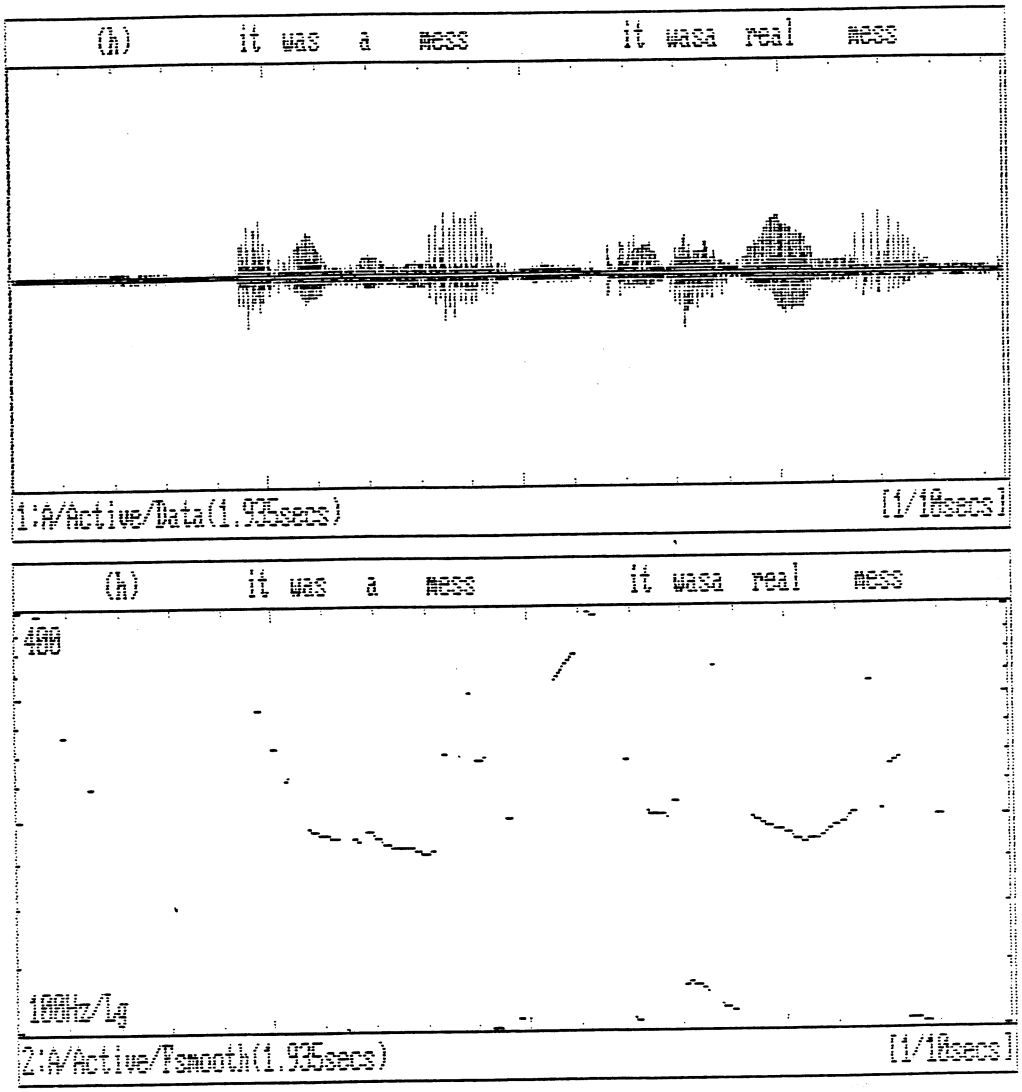


Figure 4

(2) g(A) ...(h)(.40) %it was a %méss=.
h(A) %it %was [%a] real %méss=. P> LO>
i(B) [Huh.]

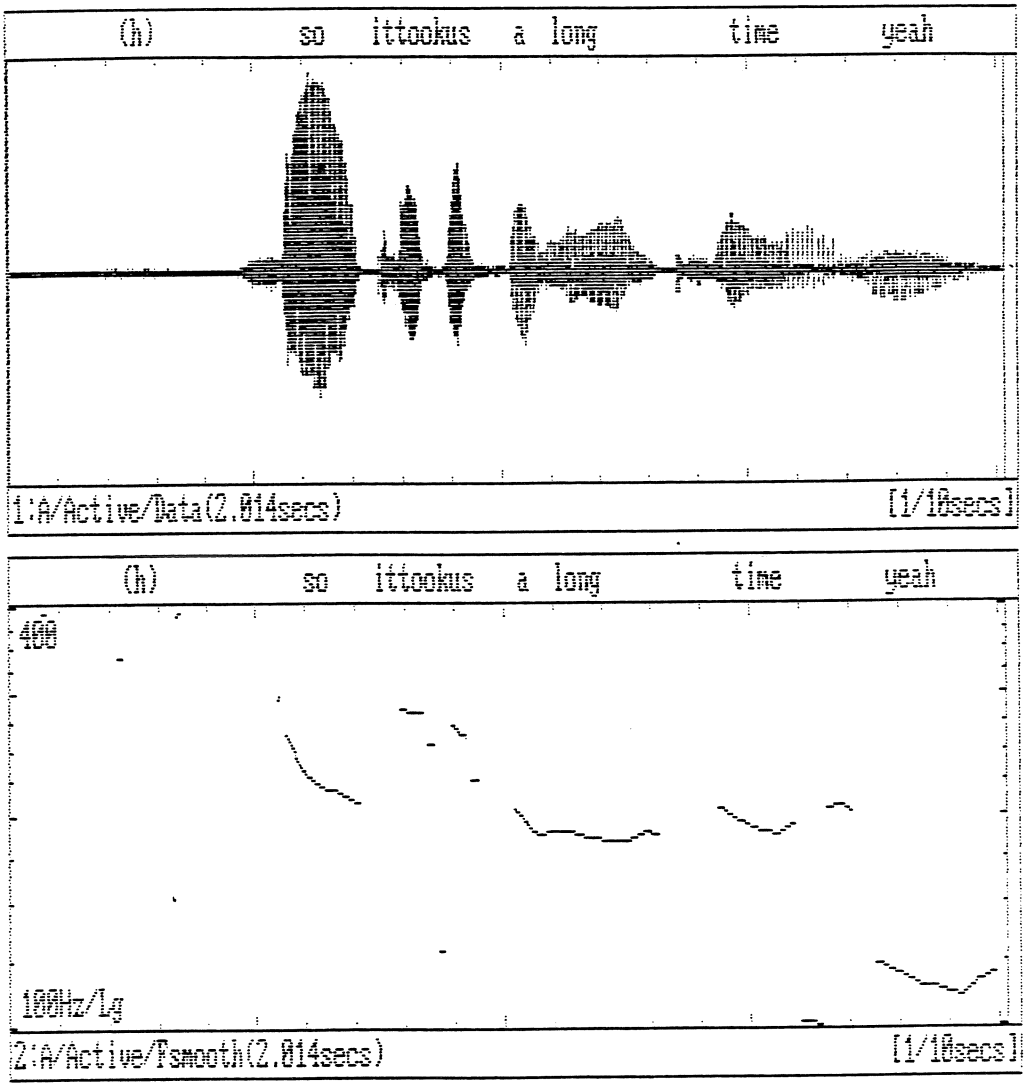


Figure 5

(2) j(A) ...(h)(.49) So it tóok us a long tíme%=.
k(B) Yeah.

PHONETIC INTERPRETATION OF TONE FEATURES IN PEÑOLES MIXTEC

John Daly

Summer Institute of Linguistics

ABSTRACT

Surface representations of tone in Peñoles Mixtec are derived from underlying representations which are determined largely by considerations of simplicity in the account of tone sandhi. There is tension between underlying tonal representations which simplify a description of tone sandhi and surface representations which lend themselves to straightforward phonetic interpretation. These demands are met within a theory of tone in which primary and register features are on separate planes and in which register has a cumulative effect.

1. Introduction

A coherent tone system of Peñoles Mixtec (PM)[1] emerges with the postulation of two primary features High (H) and Low (L) and two register features high (h) and low (l). These features make possible an insightful account of tone sandhi that is compatible with a relatively simple, although unexpected, account of the phonetic manifestation of tone.

The framework for this description of tone is Snider (1990), where the tone features are on two tiers (and planes) meeting at a Tone Node (TN). Every tone is specified for a primary feature (Snider's 'modal feature') and a register feature. Register features are cumulative in their effect, seen in tones in successive l registers being on successively lower levels.

In my adaptation of the model, tones in underlying structure are underspecified. Only the primary feature H and the register feature h appear, and subsequently the primary feature L is introduced by default. The introduction of default L is delayed until after all the rules have applied which do not require the presence of this feature, and which if it were present, would unnecessarily complicate the rules.

In underlying structure a register h feature is associated or unassociated. If it is unassociated, it is subject to association to a following vowel by rule. A register h feature is introduced in H tone assimilation following the last assimilated H tone, and this feature is also subject to association by rule. An associated register h feature, either underlying or derived, will have a primary L feature associated to its TN by default. The combination of the register h feature and the primary L feature, I call a h register L tone, symbolized as L^h.

The register feature l is not present in underlying structure but is introduced by default after all the rules accounting for tone sandhi have applied. Introducing this feature facilitates the specification of the phonetic value of H tone. H tones receive different values depending on whether they are in h or l register. Furthermore, l register accounts for the successive lowering of tones pertaining to a series of l registers by register features being defined as having a cumulative effect.

2. Tone Sandhi

The benefit of the tone features in tone sandhi is most clearly seen in tone alternation in the eight basic tone patterns on disyllabic morphemes. An informal representation of these patterns is: H H, H L, L H, L L, L^hH H, L^hH L, L^h H and L^h L. In one environment (following a subclass of L L morphemes) each of the first four tone patterns becomes one of the second four. In a different environment (following the pattern L^h L) each of the second four tone patterns becomes one of the first four. The same tone patterns are paired in whichever direction the change occurs. The two kinds of change are accounted for by the association or delinking of a register feature h.

The tone changes are summarized in (1). Following the subclass of L L morphemes which condition a change (morphemes with a floating register h feature), the

[1] Peñoles Mixtec is spoken in the village of Santa María Peñoles, located in the mountains to the west of Oaxaca City in the state of Oaxaca, Mexico. PM, along with a number of mutually intelligible varieties of Mixtec spoken in other villages of the same area, group together into what I have called Eastern Mixtec. Eastern Mixtec is sufficiently different from other dialects of Mixtec to be viewed as a distinct language. It is one of 30 or more Mixtec languages spoken in the states of Oaxaca, Puebla, and Guerrero, that are differentiated by mutual unintelligibility.

patterns on the left become the patterns on the right. Following the pattern L^h L, the patterns on the right become the patterns on the left.

(1)	H H	<-->	L^hH H
	H L	<-->	L^hH L
	L H	<-->	L^h H
	L L	<-->	L^h L

The phonetic evidence that the tones of the L^hH glide are the appropriate ones is that these tones on one vowel have the same phonetic value as when they occur on two vowels. L^hH H, where the first two tones are on a single vowel, have the same phonetic values as L^h H H, where the first two tones are on separate vowels. The essential difference in two tones on a single vowel is the glide from one level to the next.[2]

An example of the tones L^h H on a single vowel is in (2a) and on two vowels in (2b). H tone is represented by (´), L tone by (¯) and L^h tone by (ˇ). The combination of the tones L^h and H on a single vowel is represented by (ˇ).

(2) a. úú tătá
two father
'two fathers'

b. úú čìbá -dé
two goat -his
'two of his goats'

Additional evidence for positing the sequence L^hH on a single vowel is that following tones have the same values as when the tones L^h H occur on two vowels. In (2) the second H tone, for example, is one step up from the preceding H whether the preceding H is on the same vowel as the L^h or on a separate vowel.

3. Tones in h and l Registers

Given the above assignment of tone that facilitates a description of tone sandhi, it remains to be seen whether this analysis is compatible with a reasonable account of the phonetic manifestation of tone. A way must be found to relate the underlying representations to surface representations which receive a straightforward phonetic interpretation.

The phonetic data to be accounted for can be divided into tone sequences which are preceded by a L^h tone and those which are not. The tone sequences preceded by L^h in my analysis are of h register and other sequences are of l register. Consider first the data in (3).[3]

(3) a. čìū^{n4}
work

b. čìū^n -dé
work -his

c. k^{w}àžú
horse

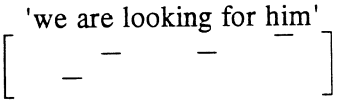
d. sàkú -dé
CON.laugh -he
'he is laughing'

[2] In Daly (1977) I postulated four tones: modified H, modified L, unmodifed H and unmodified L, differentiated by the features High and Modify, adapted to the phonetic requirements of PM (cf. Woo 1969). A modified H tone is equivalent to what I now analyze as a sequence of the two tones L^h and H, which more directly represent the phonetic facts; and a modified L tone is equivalent to the L^h tone.

[3] The data are based on my auditory impressions, supplemented by tracings of F0 in the CECIL system.

[4] Nasalization is represented by (^{n}) following the last vowel of a morpheme. It spreads left to the adjacent vowel. It also spreads to a second vowel to the left if the two vowels are adjacent to each other or are separated only by a glottal stop.

e. ndùkū -ndɨ́-dé
CON.look.for -we.ex -him
'we are looking for him'

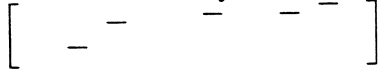


The following observations of the data in (3) must be taken into account: Immediately following a L^h tone, a L tone begins a step up and ends at extra low pitch before pause (3a). If a H tone follows the L tone, the L tone and the H tone are both level tones one step up from the L^h tone (3b). Immediately following a L^h tone, a H tone is one step up (3c). If there is a second H tone, it is one additional step up (3d). Two H tones are also a step apart, if a L tone intervenes between the L^h and the two H tones (3e).

Comparing (3b) and (3c) shows that one H tone following L^h has the same value whether there is an intervening L tone or not, and comparing (3d) and (3e) shows that two H tones have the same values whether there is an intervening L tone or not.

These sequences of tones can be expanded by adding any number of H tones (within the limits of a phonological phrase). H tones intermediate between the first and last H tone may be on the same level as the last H tone, but more often are on successively higher levels between the level of the first H tone and the last H tone. An approximation of the typical manifestation of the intermediate H tones is given in the graph in (4) where these tones are shown to be at the same level, halfway between the level of the first H tone and the last H tone.

(4) k^wìní -dé dí?ú^n
CON.want -he money
'he wants money'



A sequence of any number of L tones following a L^h tone begins and ends at the same phonetic levels as a single L tone: If it is phrase final, the sequence begins a step up from the L^h tone and ends at extra low pitch (5a). If it is followed by one or more H tones, the L tones are level tones a step up from the L^h tone (5b).

(5) a. ndùkū -šī ɨ̄ɨ̄^{n} kōlō
CON.look.for -she one turkey'
'she is looking for a turkey'

b. sà?nī -šī ɨ̄ɨ̄^{n} čótó
CON.kill -she one rat
'she is killing a rat'

These data demonstrate the possible combination of tones and the typical manifestation of tones in h register. Zero or more L tones followed by zero or more H tones pertain to the same h register and have the phonetic values described above.

All tones which do not pertain to h register pertain to l register. Examples of tones pertaining to l register are given in (6).

(6) a. ɨ̄ɨ̄^{n} njūšī
one chicken

b. úú dítá
two tortilla

c. kā?nī -šī úní čótó
POT.kill -she three rat
'she will kill three rats'

l register consists entirely of L tones (6a), entirely of H tones (6b), or of L tones followed by H tones (6c). L tones followed by H tones are level tones.[5]

The L^h tone, which contributes a register h feature to following tones, is itself of l register. It may be the

[5] The phonemic distinction between L and H is retained because of the potential contrast between them in most contexts–the contrast arising from the variants of L not being the same as the variants of H.

only L tone at l register (7a), or it may be the last of two or more L tones at l register (7b).

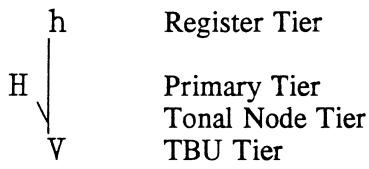
(7) a. čìbá
goat

b. ɨɨⁿ kʷàžú
one horse

4. Tone Representation in Surface Structure

In surface structure, every Tone Node (TN) has associated to it a primary feature and a register feature. A TN is associated to a vowel, the Tone Bearing Unit (TBU). The geometry of tone is illustrated in the diagram in (8) where a H tone is at h register.

(8)



The same primary features in the same combinations pertain to a single register, whether h or l. The tones of a single register conform to the template in (9) of zero or more L tones followed by zero or more H tones. Thus, a register may consist of L tones followed by H tones, of only L tones or of only H tones.

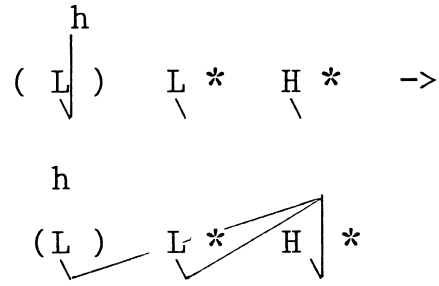
(9) L* H*

In order to make this generalization, I create a node for each sequence of tones which is to become h or l register. A register h feature is disassociated from a L^h tone and is placed at a following register node.[6]

Register nodes which do not become h register become l register by default.

The rule to create a register node is:

(10) Register Node Formation



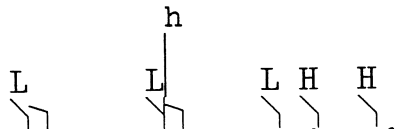
The rule applies twice. The first time, it applies to the more restricted environment--to each sequence of zero or more L tones plus zero or more H tones which is preceded by a L^h tone. The rule creates a register node which is associated to the L and H tones, and delinks the register h feature. The second time the rule applies, it applies to each remaining sequence of zero or more L tones and zero or more H tones, creating a node which is associated to each sequence of L and H tones.

An example of two applications of the rule of Register Node Formation is given in (11). In (11a) no rules of tone sandhi have applied, so the underlying representation is the same as the intermediate representation to which further rules apply. (The diacritics on the vowels in each of the following diagrams represent the tones of the intermediate representation.)

On the first application of the rule to (11a), two L tones followed by two H tones are associated to a register node, and a register h feature is delinked, shown in (11b). On the second application of the rule, the L tone which had a register h node before it was delinked and the preceding L tones are associated to a register node, shown in (11c).

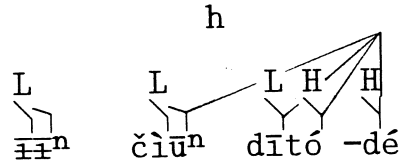
[6] The association of the register h feature with a primary L feature (L^h) and the subsequent delinking of the register h is an artifact of the underspecification of tone. The register h could be floating and not require delinking if instead of being associated to a primary L , a register l were associated to the L, the surface representation of this tone. The floating register h (on the same tier as the l) would follow the associated l, thereby assuring the proper ordering of the floating h with respect to following L and H tones to which the h comes to be associated. This possible simplification, however, is more than offset by the advantages of underspecification.

(11) a.

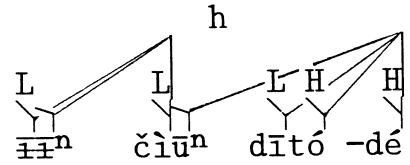


one work uncle -his
'his uncle's one job'

b.

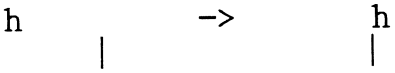


c.



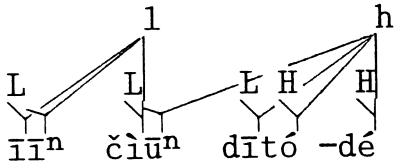
The rule which places the now floating register h feature at a following node is:

(12) Register h Labeling



This rule applies to the representation in (11c), and default l register is introduced to give (13). The second register node is labeled with the register h feature, and the first register node is labeled with the default l.

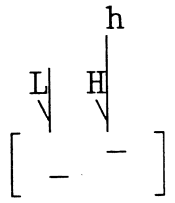
(13)



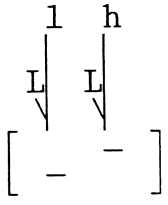
5. Phonetic Value of Features

The combination of a primary tone feature and a register feature determines the height of a tone. A register feature shifts a tone to a higher or lower phonetic level by the same degree that two primary tone features differ in height. A register h feature specifies the level of a tone as one step higher than the same tone in a preceding register, and a l register feature specifies the phonetic level of a tone as one step lower than the same tone in a preceding register. The relationship in the theory between primary tone features and register features is illustrated in (14). In (14a) a L tone and a H tone, which are in the same h register, are shown to be on two phonetic levels; the H tone is one step up from the L tone. In (14b) there are two L tones at two different pitch heights, the second on a separate h register. The second L tone is one step higher than the preceding L tone to the same degree that the L and H tones in (14a) are a step apart.

(14) a.



b.



The phonetic value of tone features illustrated in (14) conform to Inkelas (1987) and Snider (1988, 1990). In PM an innovation in the model is needed for a H tone to be either at the highest level of its register or at its lowest level. The first H tone following a L tone does not have the value shown in (14a)--only subsequent H tones do. The first H tone in h register is lowered to the lower end of the register and is at the same level as the preceding L tone, although it may vary to a somewhat higher pitch level.

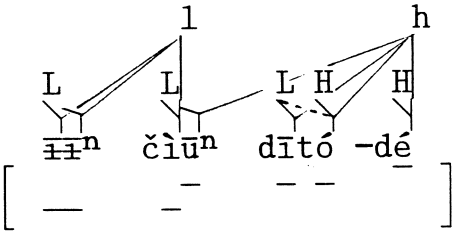
The lowered H tone, as will be seen, cannot simply be H or L, but must be both H and L. To obtain this combination of features, a primary L feature is spread to the tone node of a following primary H feature to form a merged L:H. By spreading L, (15) is derived from (13) above.

In (15) the primary L feature of the first vowel of dītó spreads to the tone node of the primary H feature of the second vowel of this morpheme to lower the H tone to the same level as the preceding L tone. The H tone of -dé is not lowered; it is at the highest level of h register.

As the diagram indicates, the phonetic difference between the two H tones in h register is as great as the phonetic difference between the two L tones of čìūn, which are of two registers. The difference between the two H tones is also as great as the difference between the first L tone of čìūn and the H tone of dītó. Put another way, the allophonic distinction between the

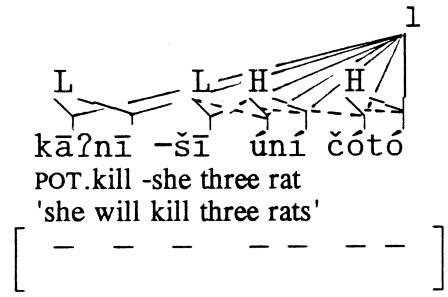
H tone of dītó and the H tone of -dé is as great as the phonemic distinction between the first L tone of čìū^{n} and the H tone of dītó.

(15)



It is only the first of two H tones that is lowered to low pitch level in h register. Contrast this with the level of H tones in l register. Not just the first, but every contiguous H tone following a L tone is at a lower level, so a primary L feature is spread to the tone node of each following H tone. In (16) the H tones of úní and of čótó are lowered and are each specified by the features L:H.[7]

(16)

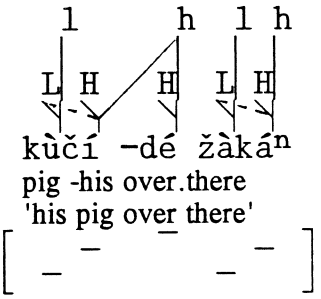


An alternative to the merged L:H might seem to be to spread a L tone and to delink a H tone, with the delinked H accounting for the derived L not gliding lower. This alternative is not feasible because a L tone maintains its identity into the phonetic component even when followed by an anchored H tone of the same register. While in one context the L tone is always a level tone, indistinguishable in pitch from a following H tone, in other contexts it fluctuates between being a level tone and being a drifting tone. In some contexts it is more likely to be a level tone, and in other contexts it is more likely to be a drifting tone. This kind of fluctuation is more appropriately handled in the phonetic component rather than by a change from one contrastive tone to another.[8]

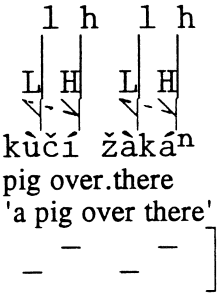
A second alternative to the merged L:H is to place the two H tones which are phonetically a step apart on separate h registers. However, in PM this is not an option. If two H tones are of two h registers, following tones should be shifted to a higher level from what they are when there is only one H tone because of the cumulative effect of register, but tones following two H tones and tones following a single H tone are at the same level. The two H tones must therefore pertain to one h register, as does a single H tone.

Examples showing following tones being unaffected by the number of preceding H tones at h register are in (17). In (17a) there are two H tones in the initial h register, and in (17b) there is one H tone in the initial h register. This difference does not affect the level of the following L^{h} H pattern.

(17) a.



b.

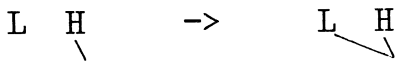


[7] Spreading L to more than one following H violates the no-crossing constraint. A possible remedy is to place L on a separate plane from H in the same way that register features are on a separate plane (cf. Coleman and Local 1991).

[8] A single L tone may drift lower, and a series of L tones may drift to successively lower pitch levels. Relevant to the likelihood of L tone drift is the presence or absence of a following H tone of the same register and the presence or absence of a preceding H tone. A following H tone of the same register diminishes the likelihood of L tone drift, and a preceding H tone increases the likelihood of L tone drift.

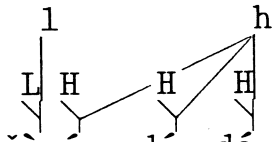
The association of a L feature to the first following H tone in h register, or to every following H tone in l register, is done by the same rule. In the case of h register the rule applies once, and in the case of l register it applies iteratively. The rule is:

(18) Primary L Spread



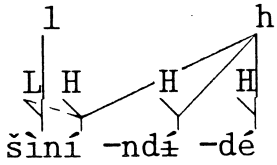
The Primary L Spread rule is applied to the first H tone of h register in (19). The primary L feature of šìní is spread to the primary H features of this same morpheme.

(19)



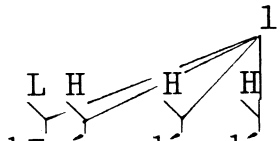
->

CON.know/see -we.EX -him
'we know/see him'



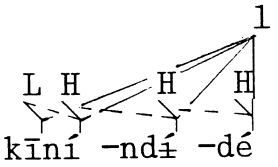
The rule is applied iteratively to each H tone of l register in (20). The primary L feature of kīní is spread to the primary H feature of this same morpheme and to the primary H features of -ndɨ́ and -dé.

(20)



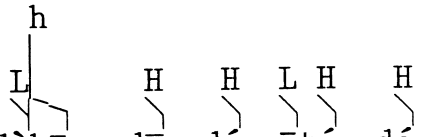
->

POT.see -we.EX -him
'we will see him'



The application of the Primary L Spread rule, as well as the Register Node Formation and Register h Labeling rules previously introduced, is illustrated in the following derivation. Beginning with an intermediate representation (21a), Register Node Formation applies once to delink a register h and to create a register node associated with following tones (21b), and applies a second time to create two additional register nodes (21c). Register h Labeling applies next (21d), and then register l is introduced by default (21e). Primary L Spread applies in h register (21f) and then in l register (21g).

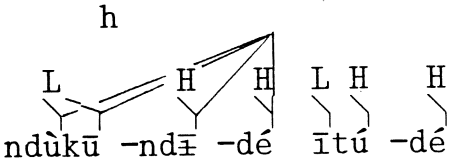
(21) a.



CON.look.for -we.EX -him corn.field -his
'we are looking for him in his corn field'

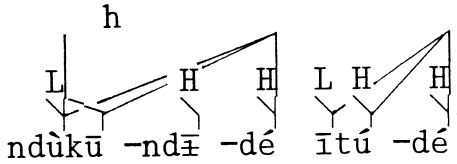
intermediate representation

b.



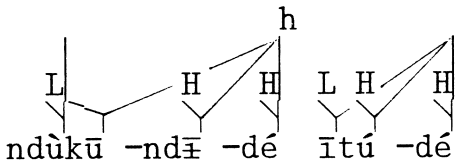
Register Node Formation
(first application)

c.



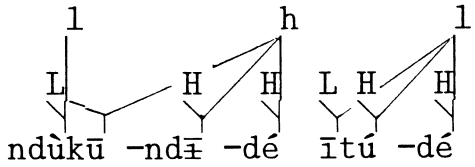
Register Node Formation
(second application)

d.



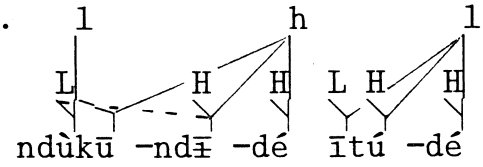
Register h Labeling

e.



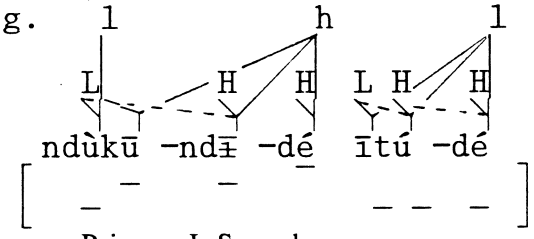
Default l Register

f.



Primary L Spread
(h register)

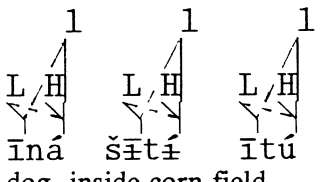
g.



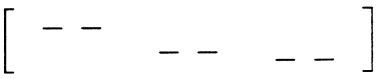
Primary L Spread
(l register)

A series of l registers occurs when L and H tones alternate. Each sequence of one or more L tones plus one or more H tones pertains to a single l register that places them lower than the tones of the immediately preceding l register, as exemplified in (22).

(22)

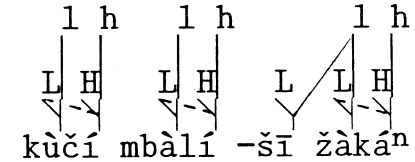


dog inside corn.field
'dog in a corn field'



There is no case of h register being cumulative, although this possibility is allowed for in the definition of register. Two h registers are always separated by a l register. Thus, the effect of h register on following tones is always offset by l register. In (23) three L^h H sequences are shown to be identical phonetically (apart from optional L tone drift).

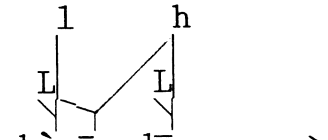
(23)



pig co-godparent -her over.there
'pig of her co-godparent
over there'

There is one important conditioned variant that has not been accounted for. A L tone always glides to extra low at the end of a phonological phrase. The model of tone followed here makes it possible to give a distinctive representation to a final L tone by adding a second L tone to the final vowel and assigning to this L tone a register l feature all its own. Thus, a final L tone ends at a register a step down from its beginning point. For example, the representation of a morpheme such as -dɨ̄ 'animal' in (24), which has a single L tone at h register, will come to have two L tones in surface structure, one at h register and the other at l register.

(24) a.

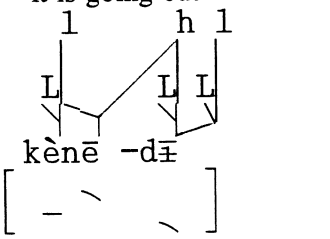


->

CON.go.out -AN

'it is going out'

b.



6. Conclusion

The choice of tone features and their assignment to strings of tones in PM are governed by considerations of simplicity and coherence in the description of tone sandhi and in the phonetic interpretation of tone. Simplification in the description of tone sandhi leads to a specification of the phonetic value of tone that is relatively straightforward.

The features which meet these requirements are the primary features High (H) and Low (L) and the register features high (h) and low (l). The primary features specify tones which are on two phonetic levels a step apart. The register features shift the primary features a step up or a step down. Tones at h register are a step higher than the tones of a preceding register, and tones at l register are a step lower than the tones of a preceding register.

The cumulative nature of tone register accounts for the lowering of a series of alternating L and H tones by a single register feature l being associated to each sequence of L plus H tones. The cumulative nature of register is not seen in h register although at first sight it may appear to be. Instead, every series of tones at h register is preceded by one or more tones at l register.

Underlying forms are underspecified for tone. They have only the primary feature H and register feature h, associated or floating, the latter associated by rule in the course of a derivation. The primary feature L and register feature l are filled in by default. The register feature h is associated only to the first vowel of underlying disyllabic forms, and comes to be associated with the feature L by default.

The benefit of these tone features is seen in tone sandhi where a floating register h associates to a following vowel and where an associated h is delinked. A third major process of sandhi is the assimilation of L to H and the introduction of a floating register h following the last H tone.

After rules of tone sandhi have applied, the tone representations are prepared for phonetic interpretation. First, zero or more L tones followed by zero or more H tones are identified as pertaining to the same tone register. Second, a register h, disassociated from its primary L feature is associated to the following tones pertaining to the same register. Third, all remaining tone registers are specified as l by default.

Finally, a primary L feature is spread to every following H tone of l register and to the first H tone only of h register. The need for this additional rule is to differentiate two levels of H tones which are both of the same h register, the first lowered to the lower end of h register and the last at the upper end of h register. It is seen that the two levels of H cannot be accounted for by the first H tone being at one h register and the last H tone at a second h register because following either one or two H tones, tones of a following l register maintain their same phonetic pitch level and are not shifted upward as they would be following two h registers. Tones intermediate between the first and last H tone are typically on successively higher pitch levels, ending below the level of the last H tone.

One or more L tones of the same register may be level or may drift lower. One or more merged L:H tones may be at the same phonetic level as a preceding L tone of the same register or may vary to a higher pitch level. If a L does not drift lower or a L:H does not vary to a higher level, the contrast between L and H tones may be lost. Nevertheless, the distinction between L and H tones is maintained until they are interpreted phonetically because in most contexts the L tones may be manifested as either level or drifting tones. Furthermore, the manifestation of L tones is on a continuum between their being clearly level or clearly drifting that makes it impossible to determine in every case whether the L tones are the same or different phonetically than H tones.

REFERENCES

1. Archangeli, D. and Pulleyblank, D., *Grounded Phonology*, MIT Press, Cambridge, to appear.
2. Clements, G. N., "The Hierarchical Representation of Tone Features," pp. 145-76 in I. R. Dihoof, ed., *Current Approaches to African Linguistics*, v. 1, Foris Publications, Dordrecht, 1983.
3. ---, "The Geometry of Phonological Features," pp. 225-52, Phonology Yearbook 2, 1985.
4. Coleman, J. and Local J., "The 'No Crossing Constraint' in Autosegmental Phonology," pp. 295-338, Linguistics and Philosophy 14, 1991.
5. Daly, J. P., "A Problem in Tone Analysis," pp. 3-20 in W. R. Merrifield, ed., *Studies in Otomanguean Phonology*, Summer Institute of Linguistics and University of Texas at Arlington, Dallas, 1977.
6. ---, "The Role of Tone Sandhi in Tone Analysis," Notes on Linguistics, to appear.
7. Goldsmith, J. A., *Autosegmental Phonology*, MIT PhD Dissertation, 1976.
8. ---, *Autosegmental and Metrical Phonology*, Blackwell, Cambridge, 1990.
9. Hyman, L., "The Representation of Multiple Tone Heights," pp. 109-52 in K. Bogers, H. van der Hulst and M. Mous, eds., *The Phonological Representation of Suprasegmentals*, Foris Publications, Dordrecht, 1986.
10. ---, "Register Tones and Tonal Geometry," 1989, to appear in K. Snider and H. van der Hulst, eds., *The Representation of Tonal Register*.
11. Inkelas, S., "Tone Feature Geometry," pp. 223-37, Proceedings of NELS 18, University of Massachusetts, Amherst, 1987.
12. Leben, W., *Suprasegmental Phonology*, MIT PhD Dissertation, 1973.
13. Pike, K. L., *Tone Languages*, University of Michigan Press, Ann Arbor, 1948.
14. Pulleyblank, D., *Tone in Lexical Phonology*, Reidel Press, Dordrecht, 1986.
15. Snider, K., "Towards the Representation of Tone: A Three-dimensional Approach," pp. 237-69 in H. van der Hulst and N. Smith, eds., *Features, Segmental Structure and Harmony Processes*, v.1, Foris Publications, Dordrecht, 1988.
16. ---, "Tonal Upstep in Krachi: Evidence for a Register Tier," pp. 453-74, Language 66, 1990.
17. Woo, N., *Prosody and Phonology*, MIT PhD Dissertation, 1969.
18. Yip, M., *The Tonal Phonology of Chinese*, MIT PhD Dissertation, 1980.
19. ---, "Contour Tones," pp. 149-74, Phonology 6, 1989.

PROSODIC TOPIC- AND TURN-FINALITY CUES

Ronald Geluykens and Marc Swerts *

Institute for Perception Research (IPO)
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

This paper describes an acoustic analysis of prosody in a variety of experimental types of dialogue. Subjects cooperatively had to perform a speaking task (i.e. describe simple rows of differently colored figures and signal their structure) and a listening task (i.e. respond to discourse boundaries in the speech produced by the interlocutor and take over as soon as the other had finished). It was found that the demarcation of discourse units by means of various intonation contours and accent shifts is largely dependent on the kind of discourse setting, in that speakers clearly take into account whether a conversational partner is likely to interrupt or not. Moreover, subjects appear not just to exploit local cues to signal the boundaries of larger-scale units. Our study reveals that they also have at their disposal: (1) a specific type of intonation contour (a level tone), occurring well before the actual end, that pre-signals that a unit will soon be rounded off; (2) topline-declination over the course of a topical unit that is different in final position than in non-final position; (3) a gradual shift in prominence in a NP from the adjective to the noun position over the course of a discourse unit.

1. INTRODUCTION

1.1. The problems

Conversation can be taken to be a kind of talk in which two or more participants alternate in speaking about particular topics (after Levinson 1983). Such a definition implies that a dialogue has both informational and interactive aspects: not only do interlocutors exchange ideas (topic dimension), they must also do so in an organized manner by regulating that no two speakers talk simultaneously (turn-taking dimension). Various conversation and discourse analysts have tried to find out what the linguistic devices are for signalling topical coherence in verbal com-munication on the one hand, and for governing the turn-taking mechanism on the other hand,

According to many researchers (e.g. Brown, Currie & Kenworthy 1980, Johns-Lewis 1986), prosody has a predominant role in structuring the flow of interactive spoken discourse. In particular, it is generally taken for granted that such suprasegmental features as speech melody, tempo, pause, etc. are used to demarcate both topical units and turns. Research has mainly concentrated on how the end of such units is signalled, as marking of finality is considered most relevant.

This paper also deals with the prosodic demarcation of discourse units in a conversation context and primarily focusses on finality cues. It addresses some problems largely neglected in the literature. Firstly, the notions 'turn' and 'topic' clearly are non-overlapping. For instance, it is quite conceivable that a speaker changes a topic within his turn or that a topic is continued over the turns of the participants. Therefore, we feel that it needs to be investigated how the prosodic structuring of topic flow interferes with that of the turn-taking, and vice versa (see section 2). Secondly, from experimental studies we know that a discourse unit may be rounded off on a global level too (e.g. Leroy 1984; Swerts, Bouwhuis & Collier (in prep.)). Hence, we will also try to find out whether there are some more indications that finality may be signalled prosodically well before the end of a unit (see section 3). The study reported upon here is experimental in nature, concentrates on Dutch and is mainly production-oriented. It includes both instrumental and auditory analyses.

1.2. Notes on methodology

Any experimental investigation on the role of prosody in governing both information flow and interaction in natural speech data is likely to run into methodological problems. On the one hand, in daily conversations, there are many potentially relevant variables involved that may all contribute to prosodic structure and that are difficult to vary independently. On the other hand, an experimental setting might be that far removed from ordinary dialogue that it gives us little insight in how the latter really works. In another paper (Swerts & Geluykens 1992), we show how one could get rid of some of these limitations for a monologue situation; in this paper, we develop a methodology which tries to find a compromise between, on the one hand, data which are spontaneous but too 'open-

ended' and uncontrollable, and, on the other hand, an experimental situation in which spontaneous interaction is still to some extent possible. In our dialogue set-up, we will attempt to vary the two levels mentioned above — topic flow and turn taking— independently.

1.3. The data

From a total of ten test subjects, two participants at a time were seated in front of each other in a sound-treated studio. In between them, a screen was placed, so that they could only hear, but not see each other. Together they had to perform five slightly different experiments that consisted of a speaking and a listening instruction: while one of the two was speaking, the other had to execute a listening task. Also, in all but the first experiment, there was a constant switching of roles, in that, during a test, speakers at particular times became hearers, and vice versa. (A schematic representation of the different discourse settings is given in Table 1; see appendix for all tables).

	instruction to speaker	instruction to listener
exp 1 (monologue)	- signal breaks between series	- indicate perceived breaks on sheet
exp 2 (dialogue)	- signal end of row	- take over at end of turn
exp 3 (dialogue)	- signal breaks between series - signal end of row (end of row always equals end of series)	- indicate perceived breaks on sheet - take over at end of row
exp 4 (dialogue)	- signal breaks between series - signal end of row (end of row never equals end of series)	- indicate perceived breaks on sheet - take over at end of row
exp 5 (dialogue)	- signal breaks between series - signal end of row (end of row sometimes does, sometimes does not equal end of series)	- indicate perceived breaks on sheet - take over at end of row - indicate whether end of row is equal to end of series or not

Table 1: Schematic representation of the different discourse settings used (further explanations in the text)

The general speaking task in the different experiments was to convey relatively simple chunks of information, viz. describe from left to right rows of differently colored geometrical figures (see also Swerts & Collier, in press; Levelt 1989). Care was taken that no figure or color occurred in two successive positions, to avoid effects of given-new information on production. Within each row, some successive figures were visually presented as belonging together by drawing connecting lines in between them; subsequent series (which we will label 'topics') were being presented as being unconnected there being no visual link (see figures 1A-B for examples). In this way, series of two to seven figures were created randomly, and, except in experiment 2, each row consisted of at least two of such series (subjects were not told how many series to expect). When describing the rows with series of figures, a speaker was not allowed to use lexical or syntactic cues to clarify the structure of the rows, and therefore could only exploit prosody to, for instance, signal the breaks between series or indicate the end of a row.

In experiment 1, a monologue setting designed to test information flow structuring independently from turn-taking considerations, the speaker was instructed to describe rows with series of geometrical figures to his partner in such a way that the major breaks between successive series became apparent. His partner, the hearer, had to try and detect these breaks, indicating this on an aswer sheet in rows with numbers (see figure 1C).

In experiment 2 a kind of enforced turn-taking was introduced. Both of the participants were given an instruction sheet, that consisted of five rows with geometrical figures and five rows with numbers, and each row of figures alternated with a row of numbers. Rows with figures were used for the description task, rows with numbers for the listening task. Subject A received a sheet with a first row of figures, subject B had a sheet with a first row of numbers. There was no subdivision in series present in the rows of figures, so speakers only had to signal when the other had to take over. The hearer just had to count the figures described, indicate this in his row with numbers and start describing his row with figures as soon as he thought that the other had stopped describing his.

In experiment 3, the instruction sheets given to the subjects were identical to those of experiment 2, except that the rows of figures were subdived into smaller series of geometrical figures. The task assigned to the speaker was now twofold: he had to make clear to the hearer when his row was completed, so that the other could start describing his row with figures; he also had to indicate where in his row the breaks occurred between successive series. The task to the hearer was also twofold: he had to take over, as soon as he thought that the other had rounded off a row, and he had to indicate on his answer sheet where he heard the major breaks. Note that in this experiment, the end of a row always coincided with the end of a series.

Experiment 4 was the same as experiment 3, except that a row of figures ended in an incomplete series (indicated visually as in figure 1B), so that the interlocutor had to, as it were, finish this series after taking over the floor. In this experiment , all the rows ended with such an incomplete series.

Experiment 5, finally, was a combination of experiment 3 and 4. The speaking and listening tasks again were the same as in experiment 3, only this time rows either could end in an incomplete or a complete series.

Speakers were asked to make this difference clear to the hearer, who had to try and indicate this on an answer sheet, by using either '>' for continuation or 'll' for finality (figures 1 C-D). As this task was more complex, subjects were asked to do this experiment twice (data from both sessions were analyzed).

Data from experiments 1 to 5 were auditorily analyzed to investigate the interference of topic and turn demarcation (see section 2). Measurements of global indicators of finality (see section 3), are based on an auditory analysis of experiments 3 to 5, and on an instrumental investigation of the speech materials from experiment 3.

2. INTERFERENCE OF PROSODIC DEMARCATION OF TOPICS AND TURNS

In this section, it is discussed how the prosodic demarcation of topical units may interfere with cues that signal turn-taking, and vice versa. In 2.1., we treat the distribution of various intonation contours as a function of discourse position. In 2.2., we look at accent structure in relation to the topic- and turn-dimensions.

2.1. Final contours

Since the distinction between falling versus rising contours is often claimed to be a powerful marker of finality versus continuation in discourse, we have auditorily determined the shape of the contour at several crucial locations in the patterns produced by our speakers. Contour shapes are depicted in figure 2, which has to be consulted together with Table 2 to get the full picture. For our purposes, it appears to be sufficient if we classify contours according to two parameters: type of final movement, which determines the contour label, and end-point of that movement in the pitch range. The 'normal' pitch range is divided into a low, mid and high part, to which two marked values are added, viz. very high and very low. Table 2 only depicts the major trends for each speaker.

Three rising contours ending in mid-position can be distinguished which are used topic-internally in all settings. As there appears to be no systematic difference in the distribution of these three rises, and since they do not differ as to direction of movement or end-point, all are referred to as rise-to-mid (RM) in Table 1. (Note that some series-internal contours are systematically different, however; these will be discussed in section 3.1., and are left out of the picture for now.) For some turn-internal finality markings, a rise-to-high (RH) contour is employed, especially in the more complex settings. For most outspoken finality, speakers mostly use a fall-to-low (FL).

When only one kind of finality has to be marked, as is the case in experiments 1 and 2, this is rather consistently done through a FL. Note, however, that some speakers try to signal the additional hierarchical organization in experiment 1 by using a RH for 'minor' topic finality; this indicates that, even in such a monologue setting, there is no simple correspondence between falling tones on the one hand, and finality on the other hand. In experiment 3, FLs are reserved for the [+topic][+turn]-final positions, whereas turn-internal topics are marked by a RH (except for one speaker). To conclude from this that falling intonation is primarily reserved for marking turn finality, however, would be a mistake, as can be deduced from the results of experiment 4, in which both RHs and FLs are used to signal within-turn topic finality. Even in those cases where RHs are employed, however, speakers very rarely use a fall to signal turn-finality, presumably because in this setting, it is in conflict with topic-continuation.

couple:	1		2		3		4		5	
speaker:	1	2	3	4	5	6	7	8	9	10
exp 1										
[-topic]	RM	RM	RM	RM	RM	RM	RM	RM	RM	RM
[+topic]	RH	RH	FL	FL	FL	RH	FL	FL	RH	RH
[+topic series]	FL	FL	FL	FL	FL	FL	FL	FL	FL	FL
exp 2										
[-turn]	RM	RM	RM	RM	RM	RM	RM	RM	RM	RM
[+turn	RH	RH	FL	FL	FL	FL	FL	FL	FR	FL
exp 3										
[-turn] [-topic]	RM	RM	RM	RM	RM	RM	RM	RM	RM	RM
[-turn] [+topic]	RH	RH	RH	RH	RH	RH	FL	RH	RH	RH
[+turn] [+topic]	FL	FL	FL	FL	FL	FL	FL	FL	FL	FL
exp 4										
[-turn] [-topic]	RM	RM	RM	RM	RM	RM	RM	RM	RM	RM
[-turn] [+topic]	RH	RH	RH	RH	FL	RH	FL	FL	RH	RH
[+turn] [-topic]	L	FL	RF	RF	RH	RH	RH	RH	L	L
exp 5										
[-turn] [-topic]	RM	RM	RM	RM	RM	RM	RM	RM	RM	RM
[-turn] [+topic]	RH	RH	RH	RH	RH	RH	FL	FL	RH	RH
[+turn] [-topic]	L	RV	RF	RF	L	L	RH	RV	RV	RH
[+turn] [+topic]	FL	FL	FL	FL	FV	FL	FV	FL	FL	FL

Table 2: pitch movements in topic- and turn-final positions (see also figure 2)

Speakers use various alternatives in the most complex setting 5. On the whole, there seem to be two major strategies: (i) speakers either create another rise or fall level by going beyond the commonly employed pitch range, resulting in both falls-to-very-low (FV) and rises-to-very-high (RV); (ii) or they create another major tone by varying pitch movement, mostly resulting in a level tone (L), with a pitch which stays at mid-level towards the end, or a rise-fall (RF), which can be defined as a falling movement preceded by a rise in pitch. In these more complicated tasks, speakers appear to have no problems signalling these different categories (they also score well perceptually, cf. section 5).

Generalizing, one can say that speakers reserve a FL for the cases where the two types of finality occur together. In cases where there is only topic-finality ([+top][-turn]), a RH is most often used (as in exp 3); in cases where there is only turn-finality ([-top][+turn]), a variety of patterns occurs, but most speakers are very well able to keep these three levels distinct.

It can be concluded, then, that prosodic demarcation by means of various intonation contours is largely dependent on the type of discourse setting: speakers clearly take into account whether a conversational partner is likely to interrupt him or not, which manifests itself in the intonational characteristics of his utterances. Summarizing, final falls are regularly used to signal both topic- and turn-finality when they are not in conflict; otherwise, low falls are reserved for the 'deepest' finality level, whereas high rises and/or other tones serve to signal other finality dimensions, both informationally and interactionally. It would thus be a mistake simply to equate 'falling' prosody with 'finality' without being more specific.

2.2. Accent structure

Having established that speakers use different types of pitch contours to structure their speech and signal various kinds of finality, we will now investigate a second prosodic dimension which appears to be relevant for the demarcation of discourse units, viz. accent placement. We have pointed out in the data description (see section 1.3.) that, in order to avoid interference from the given-new structure of the discourse, each string of figures was constructed in such a way that in each figure, both adjective and noun could be considered non-recoverable (Geluykens 1988, 1991, 1992), or 'new' information, in the sense of not being predictable from the preceding context. In such a situation, given the rules for 'neutral' accenting, one would expect the noun to carry the strongest accent, with perhaps a secondary accent on the adjective. An auditory analysis of the data, however, shows a different picture. For each of the experiments, the description of each figure was auditorily evaluated (independently by both authors, with a third session in cases of doubt), and put into one of two categories: adjective-noun compounds with strongest accent on the adjective (A), and compounds with the strongest accent on the noun (N). The percentage of 'untypical' A-accents was then calculated for each figure, relative to its position in a topical string. Results can be found in Table 3.

Table 3 shows several things. First of all, it was clear from listening to the data that not all speakers treated accentuation the same way: although the majority of them appeared to vary accent placement, and put the main accent sometimes on the adjective rather than on the noun, there were three speakers ('N-N') who rather consistently accented the noun (as we expected all of them to do). We thus distinguished two categories of speakers. Secondly, the 7 'variant' speakers ('A-N') all exhibited a clear pattern, in that there was a clear shift from a high percentage of accents on the adjective in initial position in a series, through a slightly lower percentage in mid-positions, to a very low percentage of accents on the adjective in final position. In Table 3, the averages for initial positions, final positions, and all mid-positions (i.e. positions 2-3-4 for a string of five figures, etc.) are depicted.

	initial position	mid-positions	final position
exp 1			
7 A-N speakers	32.4 %	28.6 %	2.7 %
3 N-N speakers	6.5 %	11.7 %	3.2 %
exp 2			
7 A-N speakers	45.1 %	37.2 %	0.0 %
3 N-N speakers	6.7 %	10.6 %	0.0 %
exp 3			
7 A-N speakers	70.3 %	58.1 %	7.2 %
3 N-N speakers	10.0 %	28.0 %	6.0 %
exp 4			
7 A-N speakers	48.6 %	34.8 %	2.9 %
3 N-N speakers	33.3 %	39.5 %	6.7 %
exp 5			
7 A-N speakers	66.9 %	58.2 %	3.3 %
3 N-N speakers	16.3 %	11.2 %	2.0 %
exp 3 + 4 + 5 (mean)			
7 A-N speakers	61.9 %	50.4 %	4.5 %
3 N-N speakers	19.9 %	26.2 %	4.9 %

Table 3: percentage of A-accents in three major positions (auditory analysis)

The table shows clearly that the behavior of the two main groups of speakers is strikingly different (1st vs 2nd rows). This is in itself a most interesting finding, as it indicates clearly that there are no clear-cut 'rules' for signalling discourse structure through accentuation. From the data of the 7 A-N speakers, we learn, however, that they can make subtle use of accent placement to provide an extra cue to the hearer as to discourse structure. Since this shift in prominence cannot really be due to interference from the given-new structure of these strings (as in each string both color and figure were different from, and unpredictable from, the preceding context), it must be concluded that this prosodic dimension is used as a device for bringing out topical structure: speakers use it, as it were, to highlight the extreme ends of a discourse unit.

Note also that neither topic structure (exp 1) nor turn structure (exp 2) in itself appears to be sufficient to cause a significant shift to A-accents (Table 3). When the two factors are combined, A-N speakers do show a striking increase in A-accents, which seems to imply that this is a matter of both information flow and interaction (the lower figures for experiment 4 are somewhat puzzling in this respect): the greater the complexity of the task involved,

the greater a need speakers appear to feel to exploit all prosodic variables to the full.

3. GLOBAL CUES TO FINALITY

A second aspect we wanted to address in this paper is the globality of finality-cues. The underlying question is whether finality of discourse units is signalled well before the actual end, so that hearers to some extent are enabled to predict when a speaker will round off a unit. In 3.1., we discuss a specific kind of intonation contour (a level tone), which appears to function as a non-local finality cue. In the subsequent sections, we present some results of acoustic measurements on data of experiment 3. In 3.2., the phenomenon of topline declination is treated. In 3.3., we embark on the gradual shift in prominence strength from the adjective to the noun over the course of a topical unit.

3.1. Analysis of non-final contours

Turning now to non-local signalling of discourse finality, there is one intonational cue which appears to be very prominent, and which we think may well be perceptually relevant. We have indicated in Table 1 that non-final figures are consistently marked with a rise-to-mid contour (RM). However, in experiments 3 through 5, each row of figures consisted of more than one series. In the final one of those strings, i.e. the string just before turn-taking occurs, internal figures tend to be marked differently, not with a low rise but with a kind of level tone. This pattern can be clearly distinguished from the RM: in the level tone, there is absence of outspoken pitch movement on the second accent, whereas in the 'real RMs' one observes a clear accent-lending fall or rise there. Note, though, that this level (L) tone has the same end-point as both RMs depicted in figure 2; in other words, we observe some prosodic similarity between all internal tones. Informal listening to these contours yield the strong impression that these contours pre-signal that the series is the last one in the turn, independentely from its final pitch contour. We have planned to investigate to what extent this factor is perceptually relevant.

3.2. Topline declination

Another prosodic dimension which we looked at, is the relative height of the Fo peaks in each string (topline declination). We limited the acoustic measurements to the speech materials from experiment 3. To calculate the values in table 4, we have selected the highest Fo on every A-N compound, irrespective of whether this occurred on the adjective or the noun, and compared this to all the other figures in the same string. Since strings consisted of two up to six figures (seven-figure strings were not used, as they are too rare), results were calculated for series of various lengths (see Table 4). The way this was done is the following: a mean peak height score in Hz for each speaker was calculated, and this was subtracted from actual peak heights in each position, to allow inter-speaker comparison. In Table 4, a positive figure thus indicates a peak height above mean peak height, a negative figure indicates peak height below mean peak height. Table 4 also distinguishes between turn-final (F) and non-turn-final (NF) strings, in order to assess the potential relevance of peak height variation in pre-signalling turn-finality.

position:	1st	2nd	3rd	4th	5th	6th
series of 2						
NF	+19.1	+5.9	—	—	—	—
F	+14.8	-37.4	—	—	—	—
series of 3						
NF	+16.4	+2.9	-1.5	—	—	—
F	+6.6	-9.1	-15.8	—	—	—
series of 4						
NF	+20.7	-0.5	-4.6	-2.3	—	—
F	+9.6	+2.5	-11.0	-23.2	—	—
series of 5						
NF	+19.5	-2.1	-5.4	-13.2	-5.6	—
F	+11.4	-15.7	-17.8	-6.7	-33.2	—
series of 6						
NF	+15.1	-7.5	-6.7	-11.6	-15.5	-6.1
F	+21.8	+11.8	+9.6	+9.2	-7.8	-23.8

Table 4: relationship between Fo peaks in final vs non-final series (see explanation in text)

Table 4 shows, first of all, that generally speaking there is indeed top-line declination present in each topical string; the first element in each string receives the highest Fo peak, and peak height then gradually declines. This appears to be independent of the actual length of the series. Moreover, the degree of declination seems to differ between turn-final and non-turn-final strings: whereas for turn-final strings there is indeed gradual declination up to and including the last item, in non-turn-final strings it is often the case that it is the before-last item which has the lowest peak; even in instances where this is not true in absolute terms, final peaks are still significantly higher than they are for turn-final strings. Note also that initial peak heights in non-final series are, generally speaking, higher than those in final series (apart from series of 6). Though results are not very conclusive, relative peak height of Fo peaks does appear to be important in two ways. Firstly,

peak height declination signals to some extent the topic structure of each turn, highest peaks occurring on the first item. Secondly, peak height at the end signals to some extent turn-finality (though peak height comparisons are by no means easy turn-finally, as we are dealing with a different intonation contour, viz. a FL). Once again, we are thus dealing with potential global prosodic cues for signalling discourse finality, as final series differ from non-final series with respect to some properties of the topline declination.

3.3. Relative differences in height of maxima in pitch accents

In our auditory analysis of accent positions (see 2.2.), it struck us that not all adjectival and nominal accents appeared equally strong. To give some acoustical support to our impression, we calculated, for each A-N compound, the difference in semitones in pitch height between the peak on the noun (if present) and the peak on the adjective (if present), assuming that this measure somewhat reflects the relative strength of the accents in each string. A high average peak on the adjective thus gives a negative value (N<A), a high peak on the noun a positive one (N>A). Measurements are only performed on the speech data of experiment 3 of 7 'A-N' speakers (see 2.2.). Results are represented in Table 5.

position:	1st	2nd	3rd	4th	5th	6th
series of 2						
F	-1.53	+0.07	—	—	—	—
NF	-3.07	+2.82	—	—	—	—
series of 3						
F	-1.35	-1.25	+1.04	—	—	—
NF	-1.92	-0.69	+4.00	—	—	—
series of 4						
F	-1.46	-0.04	-0.62	+1.34	—	—
NF	-1.31	+0.99	-0.76	+3.97	—	—
series of 5						
F	-2.35	-0.73	-0.52	-0.75	+0.95	—
NF	-2.42	-0.44	-0.36	+0.36	+2.44	—
series of 6						
F	-1.60	-1.57	+0.67	-1.57	+0.87	+1.70
NF	-1.17	-0.01	-0.13	-0.94	-0.01	+3.22

Table 5: relationship between A- and N-accent (in semitones) (7 speakers) (see text)

The results of this instrumental analysis are once again striking. Table 5 confirms the impression that, even in those cases where speakers, say, consistently place the accent on the adjective except for the last figure (resulting in e.g. an A-A-A-A-N series), it is the initial position which receives the strongest accent (highest Fo peak in relation to the noun), while subsequent accents appear to be less outspoken, and decrease gradually. In some cases, peaks on adjective and noun are about equally strong; although we forced our data into either an A- or an N-category for the auditory analysis, a subtler transcription method seems to be in order here. In other words, we observe a gradient shift in prominence strength from the accent to the noun over the course of a topic; we thus have found another global characteristic of a discourse unit. Moreover, there seems to be a difference between final and non-final series, in that the last NPs of final series have a less prominent noun-accent than the last NPs of non-final series (though, of course, still more prominent than the adjectival accent).

To supplement the data in Table 5, we have once again calculated the averages for all initial positions in a series, all final positions, and all intermediary positions, irrespective of the length of the series. This gives the situation presented in Table 6.

		initial position	mid-positions	final position
7 A-N speakers	F	-1.62	-0.55	+1.02
	NF	-1.98	-0.20	+3.29
3 N-N speakers	[F+NF]	+1.84	+1.65	+1.13

Table 6: relationship A- and N-accent in three major positions (in semitones)

Table 6 confirms, first of all, the findings of the auditory analysis for experiment 3 depicted in Table 3 above: the 7 A-N speakers show negative values in both initial and mid-position, while final position is highly positive, as a result of the much higher peak on the noun. Moreover, the average peak on the adjective is much higher in first position than it is in mid-position, confirming the finding that initial A-accents seem much more pronounced than intermediate A-accents. Those tendencies are more outspoken in NF- than in F-series, especially for the last positions in the rows. Table 6 thus brings out the results of Table 5 even more clearly. The 3 N-N speakers, which we have also included here for comparative reasons (we have added up F- and NF-series here), have positive values in all three locations, reflecting a higher average peak on the noun in all positions.

4. CONCLUSIONS

The preceding sections have provided both auditory and instrumental evidence for the claim that prosody is indeed used for structuring spoken discourse, both on the level of information flow (signalling topical units) and the level of interaction (signalling turn-taking). First of all, it has been shown that, from the speaker's point of view, prosodic

demarcation is largely dependent on the type of discourse setting: they clearly take into account whether his conversational partner is likely to interrupt him or not. Secondly, we found that both local (final intonation contours) and global (pitch range, accentuation) cues appear to be employed to structure spoken discourse. By combining information and interaction in a relatively simple interactive experimental setting, the pure contribution of prosody to the structuring of the discourse could be studied easily, although the precise impact of the different prosodic features requires further investigation.

This study can be extended in a number of ways. First of all, further perceptual research is needed to determine the relevance of these prosodic cues to the hearer. A kind of informal 'on-line' perceptual analysis was, of course, provided by the second participants in all five experimental settings, since, on top taking over turns, they also had to mark the discourse structure of the speech produced by their interlocutor. This gives us the chance to evaluate to what extent information flow was deduced successfully. Despite the difficulty of some of the tasks, success rate was quite high. Even for the most complex task in experiment 5, viz. deciding whether the turn-final series was 'complete' or not, subjects scored significantly above chance level (about 80 % correct). The perceptual efficiency of the turn-taking cues was also tested on-line, of course, by virtue of the fact that interlocutors had to react immediately by taking over the floor. Here, too, very few problems occurred. One could argue, however, that pause duration, or other factors, still functioned as important cues here apart from intonation, although the fluency as regards turn-taking is, on the whole, striking. Further perceptual experimentation is clearly needed here to determine the relative values of these different cues.

Secondly, it needs to be emphasized that both information flow and interaction have been kept relatively simple here; this research needs to be extended to more naturally occurring data. This will pose methodological problems, given the inherent contradiction between, on the one hand, collecting spontaneous data and, on the other hand, collecting data over which some variable control is possible. In another paper (Swerts & Geluykens 1992), we have shown how a compromise might be struck for a monologue setting. Similar methods wll have to be developed for the study of prosody in interactional settings.

(*) Both authors are also affiliated with the Belgian National Science Foundation (NFWO) and with the University of Antwerp (UFSIA and UIA, respectively).

REFERENCES

Brown, G., K. Currie & J. Kenworthy (1980). *Questions of intonation.* London: Croom Helm.

Chafe, W.L. (1987). Cognitive constraints on information flow. In: R.S. Tomlin, ed., *Coherence and grounding in discourse. Amsterdam: Benjamins, 21-55.*

Geluykens, R. (1988). Five types of clefting in English discourse. *Linguistics* 26: 823-841.

Geluykens, R. (1991). Topic management in conversational discourse: The collaborative dimension. *CLS* 27 (I).

Geluykens, R. (1992). *From discourse process to grammatical construction: On left-dislocation in English.* Amsterdam: Benjamins.

Johns-Lewis, C. (1986). *Intonation in discourse.* London: Croom Helm.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Leroy, L. (1984). The psychological reality of fundamental frequency declination. *Antwerp Papers in Linguistics* 40, University of Antwerp, Belgium.

Levinson, S.C. (1983). *Pragmatics.* Cambridge [...]: C.U.P.

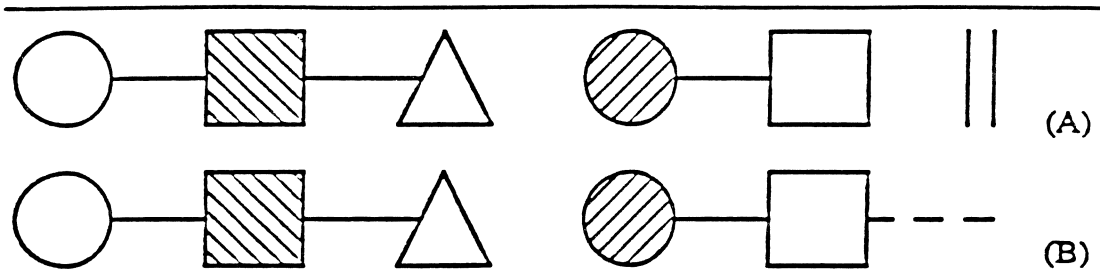
Sacks, H., E.A. Schegloff and G. Jefferson (1984). A simplest systematics for the organization of turn taking in conversation. *Language* 50: 696-735.

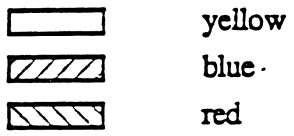
Swerts, M., D. Bouwhuis & R. Collier (in prep.). End(s) of intonation: A perceptual study of melodic cues to finality. Ms.

Swerts, M. & R. Collier (in press). On the controlled elicitation of spontaneous speech. To appear in *Speech Communication.*

Swerts, M. & R. Geluykens (1992). The prosodic structuring of information flow in spontaneous sppech. Paper presented at the Workshop on Prosody in Natural Speech Data, University of Pennsylvania, August 1992.

Swerts, M., R. Geluykens & J. Terken (in press). Prosodic correlates of discourse units in spoken monologue. To appear in *Proceedings of the 1992 ICSLP,* Banff, Canada.





1 2 3 I 4 5 II 6 7 8 9 10 ... (C)

1 2 3 I 4 5 > 6 7 8 9 10 ... (D)

Figure 1: example of production strings and perception tasks employed in exp 1-5

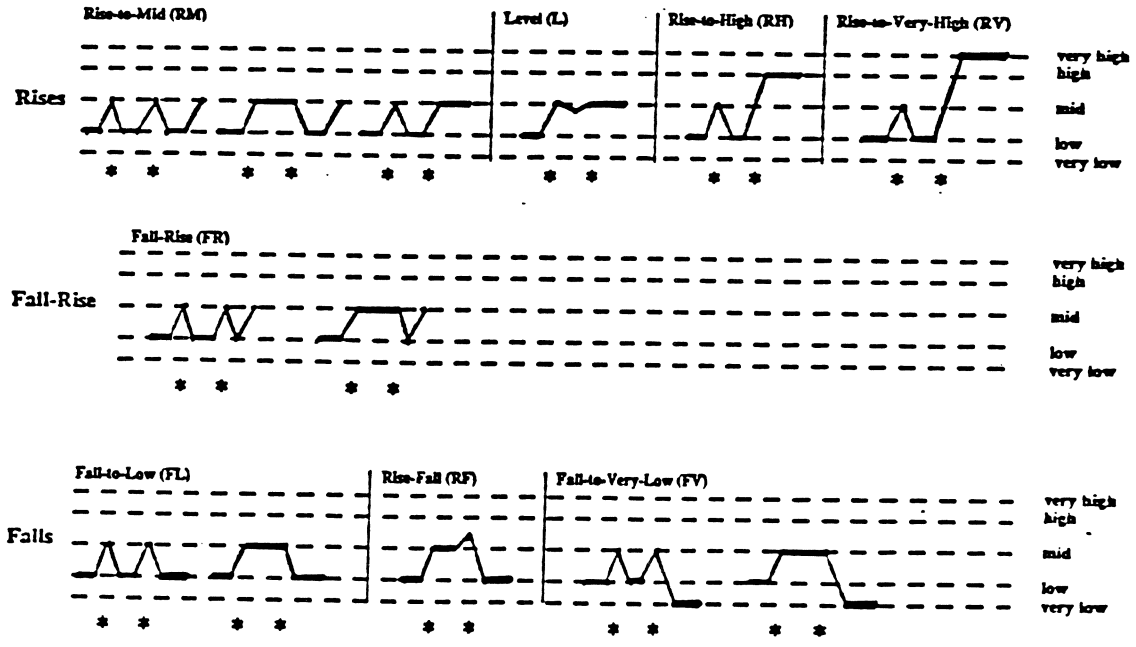


Figure 2: Detailed description of intonation contours. * represent the accents in the contours

Prosody and the interpretation of cue phrases

Beth Ann Hockey

Department of Linguistics
University of Pennsylvania
Philadelphia, PA

ABSTRACT

Cue phrases such as *okay* and *uh-huh* are often multiply ambiguous. Native speakers' intuitions are that the various interpretations of these items are distinguished prosodically. Studies by Hirschberg and Litman [1, 2] confirm these intuitions for cue and non-cue uses of several items. This study shows that various cue uses of an item can also be distinguished prosodically. Based on data from task oriented dialogs, three recurring pitch contours were found to correlate with the presence or absence of two features of the discourse: pronominal anaphora and turn taking.

1. Introduction

Certain linguistic expressions, termed 'cue phrases' [3], or 'discourse markers' [4], convey information about the structure of a discourse rather than contributing to the semantic content of a sentence. Since cue phrases overtly mark discourse information they have great potential as a diagnostic for discourse structure. However a property of cue phrases, noted by [3] and [1], is that they are generally ambiguous at least between discourse (cue) and sentential (non-cue) uses, and often among multiple cue uses as well. Hirschberg and Litman[1] and Litman and Hirschberg[2] report that cue and non-cue uses of many items can be distinguished by a combination of intonational phrasing and type of pitch accent.

Native speakers have strong intuitions that cue phrases such as "okay" can have many interpretations and that the various interpretations can be distinguished prosodically [5]. Given these intuitions, it seems likely that prosody can contribute to distinguishing multiple cue uses from each other as well as distinguishing them from non-cue use. If speakers' intuitions on the disambiguating effect of prosody with relation to cue phrases are accurate, one expects to find at worst each prosodic category correlated to a relatively small number of interpretive categories, so that the prosodic information at least narrows the available choices of interpretation.

Cue use interpretations can range over at least semantic, pragmatic, discourse and interactional factors [4] [3] [6] [7] [8] [9] [10] [11] [12]. Rather than hypothesize at the outset about interpretations, I will focus on identifiable features of the context that correlate with the differential distribution of various pitch contours. One of the goals in developing this type of classification technique is to produce theory-independent and relatively objective diagnostics of discourse structure. The descriptive results of such techniques can then be used to investigate the adequacy of a variety of discourse models in relation to actual discourses.

2. Methods

Data for the study is from taped dialogs generated by a task requiring two participants separated by a barrier to cooperatively reconstruct a paperclip design. These conversations are each about twenty minutes long and provide a fairly large number of cue phrases. This paper examines three of these paperclip task conversations with a total of four speakers. Speakers are identified by first initial and number. The conversations are represented in the tables in the next section by a sequence of two speaker identifications. The speaker who started with the completed design is listed before the speaker who was trying to reconstruct the design.

The relation between prosody and interpretation of cue words is investigated by forming natural groupings of F0 contours and by coding for certain properties of contexts, and identifying correlations between the F0 groupings and the context properties.

Grouping of F0 contours was done using characteristics such as relative F0 height of the first and second syllables and general shapes of the two syllables (e.g. rise, fall, level, degree of rise or fall). For each lexical item it was relatively easy to divide tokens into natural intonational classes by sorting pitch contours visually and auditorily, without relying on any previously-assumed system of description. This classificatory independance is an advantage since the fit between existing descriptive systems and natural data is often dubious [13]. Only tokens that were the sole items in some level of intonational phrase were used. This includes all tokens that constitute an entire utterance by themselves and tokens with sufficient phrasal separation so as not to be part of

a larger pitch contour.

Analysis of contexts was done with as few assumptions about discourse structure as possible and without reference to a specific theory of discourse. Two factors were considered in the classification of contexts:

1. distribution of pronominal anaphora and
2. turn taking behavior

As was done in Walker and Whittaker[14], I take the distribution of anaphora to be an indirect indicator of discourse structure. The arguement for using anaphora distribution as an indicator of discourse structure is based on the widespread observation (e.g. [3] [15]) that pronominal anaphora to an antecedent outside the discourse segment(s) containing the pronoun is generally not possible. Therefore, in relation to cue phrases, one expects that if a particular instance of a cue phrase is associated with a discourse segment boundary, a pronoun following that cue phrase should not be able to have an antecedent which preceeds the cue phrase. Each conversation used as data was coded from the transcript for the locations of anaphoric elements and their antecedents. Both pronouns and definite noun phrases were included in the anaphora database. Using the anaphora database, cue phrases could then be examined for whether either pronominal or definite NP anaphora took place across them. Notice that this analysis of anaphora does not take into account any details of discourse structure but only whether any pronoun following a cue phrase has an antecedent anywhere in the discourse preceding the cue phrase.

The turn taking behavior was coded for changes of speaker vs. continuation of a turn. The binary distinction is based on whether the same speaker talked immediately following the item or whether the other speaker talked.

Analysis was first done on G1B1 for the lexical item "okay". The results of that analysis were then tested on B2C1 and C1B2 for "okay" and on all three data sets for the item "uh-huh".

3. Results

3.1. G1B1 "okay"

A striking result from the intonational anaylsis of this first data set is that three especially clear contours emerge from the visual and auditory classification procedure.

One F0 contour type (ct1) is flat. The two syllables have very close F0 values and each syllable remains at its value for most of the syllable duration.

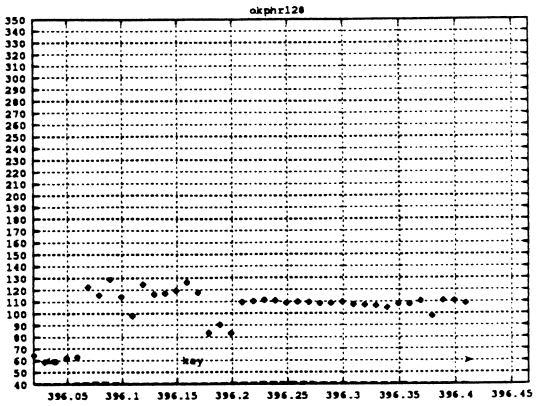


Figure 1: ct1

The second contour type (ct2) has a first syllable higher than the second with an abrupt transition. Both syllables have constant F0 value so are basically flat.

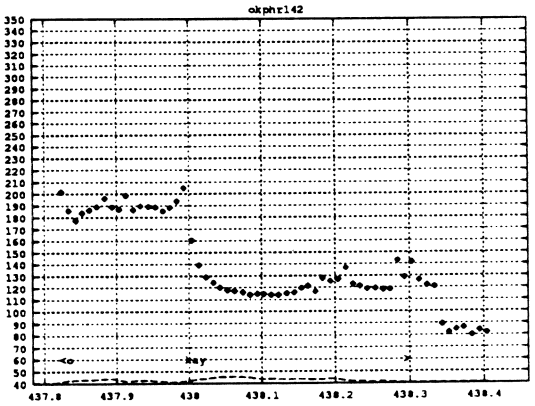


Figure 2: ct2

In the third contour type (ct3), the first syllable is flat or slightly falling. The second syllable is rising. The second syllable begins higher than the end of the first and ends considerably higher than any point in the first syllable.

The results on the first data set show that each of the three most prevalent F0 contours correlates with a distinct context that can be identified by a combination of pronominal anaphora phenomena and and turn taking behavior. One F0 contour type(ct1) is flat. The two syllables have very close F0 values and each syllable remains at its value for most of the syllable duration.

Pronominal anaphora occurs across none of the 8 tokens of ct2, while it does occur in 5 of 13 tokens of ct1 and ct3. This supports the claim in [16] that ct2 was associated

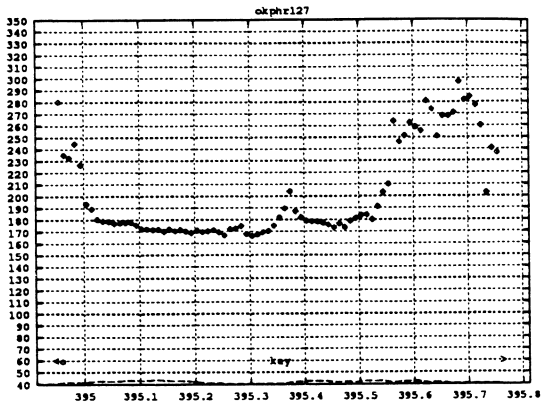


Figure 3: ct3

with the end of a discourse segment. Ct1 is always a turn continuation while ct3 is always a turn change, so ct1 and ct3 can be distinguished on that basis (see table 1 below).

G1B1	pronominal anaphora	turn change
ct1	25%	0%
ct2	0%	0%
ct3	60%	100%

Table 1: Occurance of pronominal anaphora and turn change across contour types in dialog G1B1

3.2. B2C1 and C1B2 – "okay"

The table below shows the distribution of contour types across instances "okay" for all three data sets.

		ct1	ct2	ct3
	G1B1	8	8	5
okay	B2C1	3	10	13
	C1B2	7	4	8

Table 2: Occurance of contour types with *okay* across dialogs

The two tables below show the occurance of pronominal anaphora and turn change for the second and third data sets. The categorical association of ct3 with a turn change observed in G1B1 is also present in B2C1 and C1B2.

B2C1	pronominal anaphora	turn change
ct1	1 (34%)	2 (67%)
ct2	1 (10%)	6 (67%)
ct3	4 (31%)	13 (100%)

Table 3: Occurance of pronominal anaphora and turn change across contour types in dialog B2C1

C1B2	pronominal anaphora	turn change
ct1	1 (34%)	2 (67%)
ct2	2 (50%)	3 (75%)
ct3	1 (13%)	8 (100%)

Table 4: Occurance of pronominal anaphora and turn change across contour types in dialog C1B2

3.3. Uh-huh

Two of the three recurrent contours found with *okay* also occur with *uh-huh*. The minor differences between a contour type occuring on *okay* and the same contour type occuring on *uh-huh* can be attributed to segmental effects. As can be seen in table 5 below ct2 does not occur with *uh-huh*. The number of occurances of ct1 and ct3 varies considerably across speakers.

		ct1	ct2	ct3
	G1B1	12	0	40
uh-huh	B2C1	2	0	12
	C1B2	1	0	10

Table 5: Occurance of contour types with *uh-huh* across dialogs

Only data for turn changes is shown in the table 6 below since the presence or absence of pronominal anaphora only distinguishes ct1 and ct3 from ct2 but not from each other. Pronominal anaphora did occur across both ct1 and ct3 with *uh-huh* as was the case for *okay*.

	turn change
ct1	13 (87%)
ct3	61 (98%)

Table 6: Occurance of turn change across contour types for *uh-huh*

4. Discussion

4.1. Ct3 and turn change

Across four speakers and two different lexical items ct3 categorically marks a turn change. The one instance of ct3 on "uh-huh" which is not listed in the table as turn change is actually a coding dilemma since in this case immediately following the ct3 marked "uh-huh" , both speakers talk simultaneously. Rather than being a counter example, this instance gives direct evidence that the other speaker actually did interpret the "uh-huh" with ct3 as signaling a turn change. The transcript of this instance is shown below, brackets indicate simultaneous talk.

(1)

222 B1: uh-huh. [fat?]

223 G1: [and the] fat end will be facing away from you

4.2. Ct2 and the absence of pronominal anaphora

This data clearly shows that the presence of ct2 on an instance of "okay" proscribes pronominal anaphora across that instance. There are no instances of "uh-huh" with ct2. This is consistent with an analysis of ct2 as marking a discourse segment boundary. Intuitively "uh-huh" cannot be used to mark discourse segment boundaries. We would therefore predict that "uh-huh" would be incompatible with ct2, and this is in fact the case. What particular properties of "uh-huh" are responsible for this difference in distribution between it and "okay" is a topic of ongoing research.

Three instances of ct2 in B2C1 and C1B2 seem to contradict this analysis of ct2 as a boundary marker. However, in examining these 3 instances, it is clear that they are only apparent counter examples to the generalization that ct2 marks the end of a segment.

(2)

25 B2: black white red yellow

26 C1: black white red yellow should these be linked?

27 B2: and these are all linked right

28 C1: okay

29 C1: black white red yellow, okay

30 B2: okay and I would say these put them from the the distance from the top to the bottom should be maybe ten inches

(3)

99 C1: uh and the black one is going to come up just about below just a hair below where the other black one is in the middle

100 B2: okay

101 B2: so it's about on the same level as the blue one in the in the far line

(4)

216 B2: [over an inch to the right] [(())]

217 B2: okay

218 C1: you got that

219 B2: yeah

All three examples are explained if a speaker can only end his or her own discourse segment. So in (2) the speaker ends the segment consisting of his utterance in line 29. Since line 29 is embedded in B2's larger segment, which includes lines 25, 26, 27 and 30, the closure of 29 does not affect the ability of "these" in 30 to have "these" in 27 as an anaphoric antecedent. Similarly, the "okay" on line 217 in (3) has only itself to end, so does not efect the subsequent anaphora to C1's utterance. In (4) the

"okay" can end at most itself and line 216. Notice that in this case C1 seeks additional feedback after B2's "okay" before continuing, since he does not have the clear turn change marking that would have been provided had B2 used ct3 instead.

4.3. Ct1

The clear difference in turn taking behavior between ct1 and ct3 in G1B1 was not observed in the other two data sets. So for the data as a whole ct1 and ct3 cannot be distinguished by the discourse features discussed in this paper in cases where there is a turn change. Ct1 simply conveys nothing to a hearer about whether not a turn change will follow. Obviously there are many discourse features that have not been discussed in this paper that might differentiate ct1 and ct3 in all environments. Research is underway to investigate such additional discourse features. Another question which needs to be addressed in relation to ct1 is why such a striking difference between G1B1 and the other two data sets. The answer to this question may involve issues of individual and sociolinguiguistic variation. Another possibility is that there is a difference related to percieved need for clarity by the speakers. If speakers have to make extensive corrections of prior discourse perhaps the frequency of the ambiguous ct1 would be reduced in the corrective dialog. These issues need to be addressed by future research on additional discourse features and on additional data.

4.4. Prosody and interpretation

The data shows how two of the contours discussed in this paper reduce the range of interpretation a hearer can have for the cue phrases with which the contours are associated. In working toward results that can explain how a hearer can use prosody, it is valuble to start from a data driven analysis of the prosody. The prosody is part of the speech signal in a way that abstract discourse categories are not. We would like to have an account which explains how the prosody provides information about interpreting the discourse and not the other way around. It is unlikely that a top-down approach to prosody would have led to investigating the particular contours discussed in this paper. It seems even less likely that top-down approach based on discourse categories would.

In addition to showing how prosody can reduce the range of interpretation of a cue phrase, these results suggest that in some uses of cue phrases the prosody matters more than the particular lexical item. For example, "Okay" and "uh-huh" function in the same way when uttered with ct3. Although the number of analyzable tokens of other cue phrases in this data is too small to arrive at any conclusions, analysis of the usable tokens suggests that a number of other items, such as "right", "yeah", "alright" and "so", can serve the same function as "okay" and "uh-huh" when uttered with ct3. It also appears that like "okay", "right" and "alright" with ct2 can function as discourse segment boundaries. Cue phrases are not entirely vacuous semantically and the semantics of the item can interact to generate implicatures when the semantic content of the item is not being used directly. For example, if "so" is used with ct3, the function is the same as with other items such as "okay" and "uh-huh" namely to prompt the other speaker to talk, to pass up a turn. But with "so" the semantics of "so" which is to conjoin a fact, action or event with its result [4] comes through as an implicature that what should follow should be a result of the prior turn. The person who utters "so" with ct3 when they could have used "okay" or "uh-huh" seems to be conveying that the other person should go on AND get to the point(=result). Ct3 can be thought of as marking the item to which it attaches for an interactional interpretation.

References

1. Julia Hirschberg and Diane Litman. Now let's talk about *now*: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 1987.
2. Diane Litman and Julia Hirshberg. Disanbiguating cue phrases in text and speech. In *Proceedings of Coling90*, 1990.
3. Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the stucture of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
4. Deborah Schiffrin. *Discourse markers.* Cambridge University Press, 1987.
5. Beth A. Hockey. An experimental approach to invesitgating the role of prosody in the interpretation of cue phrases. Draft, 1992.
6. Lawrence C Schourup. *Common discourse particles in English conversation.* PhD thesis, Ohio State University, 1982.
7. E. A. Schegloff. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In *Analyzing Discourse: Text and Talk.* Georgetown University Press, 1982.
8. Robin Cohen. A computational theory of the function of clue words in argument understanding. In *Proceedings of Coling84*, 1984.
9. Robin Cohen. Analyzing the stucture of argumentative discourse. *Computational Linguistics*, 13:11–25, 1987.
10. Steve Whittaker and Phil Stenton. Cues and controls in expert-client dialogues. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, 1988.
11. Rachel Reichman. *Getting computers to talk like you and me: discourse context, focus, and semantics.* MIT Press, 1985.

12. M. Merritt. On the use of 'okay' in service encounters. In J. Baugh and J. Sherzer, editors, *Language in use*, pages 139–47. Prentice-Hall, Englewood Cliffs, NJ, 1984.

13. Cynthia Ann McLemore. *The Pragmatic Interpretation of English Intonation: Sorority Speech.* PhD thesis, University of Texas at Austin, 1991.

14. Marilyn A. Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proc. 28th Annual Meeting of the ACL*, 1990.

15. Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Towards a computational theory of discourse interpretation. Draft, 1986.

16. Beth A. Hockey. Presented at the AAI Fall Symposium, Asilomar CA, November 1991.

okay	talked over	part of larger phrase	**entire utterance**	**own phrase**	total	**percent used**
G1B1	15	44	22	33	114	48%
B2C1	41	42	42	46	170	49%
C1B2	14	10	30	11	65	63%

Table 7: Frequency and phrasing of *okay*

uh-huh	talked over	part of larger phrase	**entire utterance**	**own phrase**	total	**percent used**
G1B1	13	2	60	8	83	82%
B2C1	8	0	25	1	34	76%
C1B2	3	0	13	0	16	81%

Table 8: Frequency and phrasing of *uh-huh*

COMPARING INTONATIONAL FORM WITH DISCOURSE FUNCTION: A STUDY OF SINGLE WORD UTTERANCES*

Jacqueline C. Kowtko†

Human Communication Research Centre
University of Edinburgh, 2 Buccleuch Place
Edinburgh EH8 9LW U.K.

ABSTRACT

Recent attempts to analyze the function of intonation in discourse (both monologue and dialogue) classify the data according to type of intonational tune [4, 7] and make a more or less general characterization of the discourse function associated with utterances containing the particular tunes [8, 5]. The literature shows convincingly that intonation signals boundaries in discourse structure, but lacks a clear specification of discourse function. A suitable discourse taxonomy is needed to fine-tune the relationship between intonation and discourse function. A recent analysis of dialogue [6] provides a framework of conversational games which allows more fine-grained examination of prosodic function. The current paper introduces an intonational analysis of single word utterances based upon such a framework and compares results in progress with previous work on intonation.

1. INTRODUCTION

Recent approaches to the analysis of intonational function within dialogue include an examination of the tunes carried by single-word *cue phrases* (e.g. *now* [4], *okay* [5], and others [7]) across different discourse situations. The literature also includes a more sweeping approach toward classifying phrase-final tunes which presents broadly generalized discourse functions for each of three types of intonational tune: phrase-final *rise*, *level*, and *fall* [8]. Since there is currently no commonly accepted *grammar* of discourse, these studies devise their own relevant discourse categories. Hockey [5, p. 1] reflects upon the problem, with reference to cue phrases. She states that cue phrases

> ...convey information about the structure of a discourse rather than contributing to the semantic content of a sentence. ... Context and prosody are major factors contributing to differences in interpretation among various instances of a cue phrase. In order to investigate the connection between prosodic features and uses of a cue phrase, uses must be identified.

The above is partly a response to Hirschberg and Litman [4, 7] who limit their description to a binary discourse/sentential distinction. Litman and Hirschberg [7] leave the analysis of cue phrase function to the interpretation of various specific discourse approaches and instead focus on validating their prosodic model of cue phrase use [4] with additional data from monologue. The model specifies that a cue phrase in discourse use will occur either alone in a phrase (with unspecified tune) or initially in a larger phrase (deaccented or with a low tone). Thus, Litman and Hirschberg leave open the question of how their prosodic model could further specify discourse function.

McLemore [8] approaches discourse as structured by topics and interruptions. Her data includes announcements given at Texas sorority meetings and conversation between members. She finds that phrase-final tunes indicate certain general functions: *rising* tune *connects*, *level* tune *continues*, and *falling* tune *segments*. Context determines how each of these tunes operates. For instance, phrase-final rise, indicating non-finality or connection, can manifest itself as turn-holding, phrase subordination, or intersentential cohesion. Likewise, the other tunes perform slight variations on the function of *continue* and *segment* according to context.

Hockey [5] admits to settling upon an arbitrary system of discourse classification after attempting to adopt a previous analysis based upon a somewhat similar set of speech data (trying to map discourse categories from conversation at a library reference desk to talk arising from a paperclip task). She focuses on task oriented dialogue and attempts to specify discourse function of the cue phrase *okay*. She presents her results in terms of intonational contours and their corresponding discourse categories, finding that they correlate with McLemore's [8] results: 89% of *rising* contours occur where the speaker was *passing* up a turn and letting the other person continue; 86% of *level* contours serve to *continue* an instruction; 88% of *falling* contours mark the *end* of a subtask. But her

*An earlier version of this paper appears in the Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, 1992, 282-284.

†A UK Overseas Research Student Award provides partial support. I would like to thank my advisors Stephen Isard and D. Robert Ladd for advice on this work.

categorization of discourse is still weak in that it is not replicable.

Admittedly, there are a limited number of intonational tunes (low rise, high rise, level, fall, etc.). But limitation in intonational tune should not force a limitation in discourse category. Detailed understanding of intonational function is necessarily linked to a more robust view of discourse structure. These previous studies provide good intonational analysis but within vague discourse structures.

2. CONVERSATIONAL GAMES IN DIALOGUE

The analysis offered by Kowtko, Isard, and Doherty-Sneddon [6] provides an independently defined taxonomy of discourse structure which allows a closer examination of how intonation signals speaker intention within task oriented dialogue. In the analysis, linguistic exchanges termed *conversational games* (from a tradition of literature originating in [9]) embody the *initiation–response–feedback* patterns which relate to underlying non-linguistic goals. It is through the framework of games and their components, *conversational moves*, that the intonation of single word utterances can be compared with their discourse function, as intended by the speaker.

A conversational game is defined as consisting of the turns necessary to accomplish a conversational goal or sub-goal. The initiating utterance determines which game is being played and is similar to the *core speech act* in Traum and Allen [10]. In the terms of Clark and Schaefer [3], the initiating utterance serves as *presentation* phase and the ensuing *response* and *feedback* moves function primarily as *acceptance* phases. Implicit, mutually agreed rules dictate the shape of a game and what constitutes an acceptable move within a game. These rules embody procedural knowledge which speakers employ in everyday conversation.

The repertoire of games and moves in Kowtko *et al.* [6] is based upon a map task (see [1], for a detailed description): One person is given a map with a path marked on it and has to tell another person how to draw the path onto a similar map. Neither participant can see the other's map.

The nature of the map task is such that from the conversations the speaker's intentions remain fairly obvious. Kowtko *et al.* [6] report that one expert and three naïve judges agree on an average of 83% of the moves classified in two map task dialogues. Six games appear in the dialogues: Instructing, Checking, Querying-YN, Querying-W, Explaining, and Aligning. They are initiated by the following moves:

INSTRUCT	Provides instruction
CHECK	Elicits confirmation of known information
QUERY-YN	Asks yes-no question for unknown information
QUERY-W	Asks content, *wh-*, question for unknown information
EXPLAIN	Gives unelicited description
ALIGN	Checks alignment of position in task

Six other moves provide response and additional feedback.

CLARIFY	Clarifies or rephrases given information
REPLY-Y	Responds affirmatively
REPLY-N	Responds negatively
REPLY-W	Responds with requested information
ACKNOWLEDGE	Acknowledges and requests continuation
READY	Indicates intention to begin a new game

Since the map task involves one player instructing another on how to draw a path, the conversation naturally consists of many Instructing games. The structure of games allows for looping of response and feedback moves within a game and nesting of games.[1]

The prototypical game consists of two or three moves: initiation, response, and optionally feedback. The large majority of games (84% from a sample of 3 dialogues, $n = 65$) match the simple prototype. Games that do not match the prototype are still well-formed, having extra response–feedback loops, nested games, or extra moves. Very few games (less than 2%) break down as a result of a misunderstanding or other problem.

Here is an example of a prototypical Instructing game. The vertical bar indicates the boundary of a move:

A: Right,|| just draw round it.
READY || INSTRUCT
B: Okay.
ACKNOWLEDGE

Conversational game structure offers a taxonomy which specifies both the function and context of an utterance, as move x within game y. This facilitates the study of the function of intonational tune, since the tune reflects

[1] As a comparison with Clark and Schaefer [3] embedded games often coincide with instances of embedded contributions in the acceptance phase.

an utterance's conversational role.

3. INTONATION IN GAMES

Using data from map task dialogues [1], I have been analyzing single words which compose moves within themselves: *right, okay, aye,*[2] *yes, no, mmhmm,* and *uh-huh.* They typically surface as 5 of the 12 moves in the games analysis [6]: READY, ACKNOWLEDGE, ALIGN, REPLY-Y, and REPLY-N. The current data set consists of 56 out of 80 single word moves spoken by 3 of the 4 conversants in 2 dialogues. For purposes of this study, I am excluding words which form partial utterances (24 of the 80), thus avoiding any interference with accents in the speakers' larger intonational phrases. I have intonationally transcribed each word as high level (H), low level (L), rise (LH), fall (HL), rise-fall (LHL), and fall-rise (HLH).

In order to compare my results with those of McLemore [8] and Hockey [5], I have tried to interpret each utterance as belonging to one of the three general, functional categories. Certain trends become visible: ACKNOWLEDGE moves after EXPLAIN or INSTRUCT, which interrupt the speaker without taking control, typically *connect*; READY and ACKNOWLEDGE moves which precede other moves by the same speaker *continue*; REPLY-Y, REPLY-N, and ACKNOWLEDGE after EXPLAIN or a response move (specifically elicited moves) *segment.*

The data yield the results shown in Table 1.[3]

Table 1: Intonational Tune vs. Dialogue Function

	Connects	*Continues*	*Segments*	
Rising	**1**	0	5	17%
Level	12	**3**	20	9%
Falling	1	0	**14**	93%
	7%	100%	36%	

From the table, we see that 17% of *rises* appear as *connecting* moves, 9% of *levels* as *continuing* moves, and 93% of *falls* as *segmenting* moves. Only the last category matches other published results. Similarly, analyzing the data according to general discourse function (looking down the columns) reveals that only one of the three categories appears to match previous results: *continuing* moves have a *level* intonational tune. It is possible that my classification of utterances would not be corroborated and cause some of the disagreement. Also, it is possible that dialectal variation would account for some of the difference, but I believe that these factors do not account for the difference in results.

These results reflect an intonation-based approach. Information may be lost in the process of collapsing various discourse contexts into three intonational categories (as in [8]) and then limiting discourse categories to match those three existing intonational categories (as in [5]). Using independently motivated discourse categories, in a discourse-based approach, should allow one to see clearer, more detailed results.

When categorized according to *move* (specific function) and position in *game* (discourse context), trends begin to emerge from the data. Granted, the numbers for each category are currently small and not statistically reliable, but some trends are striking and suggest that more data will prove to yield interesting results. Of the 56 data points considered here, three moves are represented: REPLY-Y, REPLY-N, and ACKNOWLEDGE. We find that when one of the utterances appears as a REPLY-Y move in an *Aligning* game, the tune will be *level* if the game is embedded, otherwise *falling.* REPLY-Y and REPLY-N moves within *Querying-YN* games vary according to the previous speaker's last accent. The tune is *low level* when the previous speaker ends low and *falling* when the previous speaker ends high. A single word appearing as an ACKNOWLEDGE in an *Explaining* game generally carries a *low level* tune. When in an *Instructing* game, it carries a *falling or rising* tune after an ALIGN move or continued INSTRUCT move, and otherwise a *level* tune. Within a *Querying-YN* game, there are not yet any clear patterns for the ACKNOWLEDGE move, as half of the tunes are *level* and half *falling.* Within a *Querying-W* game, the tune is *rising.* These results are summarized in Table 2.

I have considered two theories as to why the previous speaker's last accent influences the tune of a conversational move. Firstly, there is the possibility that both speakers cooperatively maintain an overall tune (through *pitch concord* which matches the final key of one move and the initial key of the next move [2]). However, if this were the case, we would expect to see more influence from the previous speaker's accents in other categories of conversational move. More likely is the second possibility that the difference in last accent represents a different nuance of meaning, to which the hearer then responds appropriately. The question of what influences the previous speaker's last accent in a move remains unknown.

[2] Participants in the map task were taken from the population of undergraduates at Glasgow University, and the dialogues consequently contain Scottish English.

[3] The score of 93% is significant ($p < .01$). The 7% is also significant ($p < .01$) and the 9% borderline ($p < .05$), although opposite to predicted results. All other results are statistically non-significant ($p > .05$), according to the Kolmogorov-Smirnov One-sample Test.

Table 2: Intonation Associated with Move X in Game Y

Move	*Game*	*Tune*	*Additional Factors*	*Data*
REPLY-Y	Aligning	H/L	game is embedded	5 of 5
		HL	otherwise	1 of 1
REPLY-Y/N	Querying-YN	L	prev. speaker ends low	5 of 5
		HL	prev. speaker ends high	5 of 6
ACKNOWLEDGE	Explaining	L		3 of 4
ACKNOWLEDGE	Instructing	LH/HL	after ALIGN or continued INSTRUCT (i.e. elicited)	8 of 10
		L/H	otherwise	15 of 18
ACKNOWLEDGE	Querying-YN	L/H	no clear pattern	3 of 3
		HL	no clear pattern	3 of 3
ACKNOWLEDGE	Querying-W	LH		1 of 1

Work is progressing on other dialogues, amassing enough pitch trace data to allow clear patterns to emerge for each type of move in each game context. The goal is, within a discourse context, to be able to predict an utterance's function or *move*, given the intonation, and, conversely, predict intonation, given the type of move.

References

1. Anderson, Anne H., Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert, "The HCRC Map Task Corpus," *Language and Speech*, Vol. 34, 1991, No. 4, pp. 351-366.
2. Brazil, David, Malcolm Coulthard and Catherine Johns, *Discourse Intonation and Language Teaching*, Longman, London, 1980.
3. Clark, Herbert H. and Edward F. Schaefer, "Collaborating on contributions to conversations," *Language and Cognitive Processes*, Vol. 2, 1987, No. 1, pp. 19-41.
4. Hirschberg, Julia and Diane Litman, "Now let's talk about *now*: Identifying cue phrases intonationally," *Proceedings of the 25th annual Meeting of the Association for Computational Linguistics*, 1987, pp. 163-171.
5. Hockey, Beth Ann, "Prosody and the interpretation of 'okay'," Presented at the *AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, November, 1991.
6. Kowtko, Jacqueline, Stephen Isard and Gwyneth Doherty-Sneddon, "Conversational games within dialogue," Research Paper HCRC/RP-31, Human Communication Research Centre, University of Edinburgh, 1992.
7. Litman, Diane and Julia Hirschberg, "Disambiguating cue phrases in text and speech," COLING-90 *Proceedings*, 1990, pp. 251-256.
8. McLemore, Cynthia A., *The Pragmatic Interpretation of English Intonation: Sorority Speech*, Ph.D. dissertation, University of Texas at Austin, 1991.
9. Power, Richard, *A Computer Model of Conversation*, Ph.D. dissertation, University of Edinburgh, 1974.
10. Traum, David R. and James F. Allen, "Conversation Actions," *Proceedings of the AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, November, 1991, pp. 114-119.

THE PHONETIC INTERPRETATION OF TONE IN IGBO

Mark Liberman, J. Michael Schultz, Soonhyun Hong, Vincent Okeke

Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104-6305

ABSTRACT

Igbo, a language of the Kwa branch of the Niger-Congo family, is spoken by about 15 million people in southeastern Nigeria. Its phonology, morphology and syntax have been widely studied (e.g. [1, 2]), especially with reference to the intricate patterning of lexical tone. This paper is a preliminary study of the phonetic interpretation of Igbo tone. We use an experimental method first applied to English ([4, 5]), in which a speaker varies pitch range orthogonally with variation in tonal material, and we compare the success of different models in characterizing the interaction of tone identity, phrasal position, tone sequence, and pitch range in determining patterns of measured F0 values. From the statistical structure of these data, we draw several conclusions about Igbo tone and its phonetic interpretation.

(A shorter version of this paper was published as [3])

1. Lexical Tone in Igbo

In this section, we will introduce the system of lexical tone in Igbo, especially as it relates to the design and interpretation of the experiments we have done. We pass over in silence most of Igbo's intricate and fascinating tonal phonology, since our immediate concern is with the phonetic interpretation of the surface tonal categories. However, the basic nature of these surface categories will emerge as a crucial question.

An Igbo syllable can have one of three tonal categories, known as "high" (H), "low" (L), "mid" (M). The H and L tones occur freely, but the M tone can only occur following an H tone or another M tone. Thus there are five possible tone patterns for two-syllable words: (1) HH, (2) HL, (3) LH, (4) LL, (5) HM. Monosyllabic words are rare, and all of those that can occur in isolation have high tone.

Among the Igbo as elsewhere, speakers are free to change their pitch range, and do so for many reasons. We may ask whether tonal distinctions are nevertheless maintained—for instance, is HH in a low pitch range distinguished from LL in a higher one? Are HM and HL kept distinct? If so, how? Our experiments provide a tentative answer: contextual effects in the system of tonal interpretation largely maintain the distinctions among bitonal patterns as pitch range varies.

There are several general effects that modify the realization of Igbo tones in phrasal context. One of the most important of these is known as "downdrift," which progressively lowers H and L tones when they occur in sequence. Thus on one typical pronunciation of the phrase Àbànóbì ọ́ma, "good Abanobi," the successive minimum and maximum F0 values were as follows:

TEXT:	A	ba	no	bi	ọma
TONES:	**H**	**L**	**H**	**L**	**H**
F0:	237	178	210	158	181

A model for the phonetic interpretation of Igbo tone must obviously take account of downdrift. The term "downstep" is used to refer to a different circumstance. Although some phonologists have treated Igbo "mid" tones as a third tonal category, distinct from high and low (e.g. [6], [7]), others (e.g. [1]) have interpreted the restricted distribution of the "mid" tone to mean that it is actually just a "high" tone that happens to be "downstepped", i.e. realized at a lower pitch. On this view, HM is actually H ! H, where the raised exclamation point marks the downstep location. This situation (or its counterpart in other tone languages) has been given various phonological interpretations:

1. the downstep marker may be reified as a separate phonological entity, as suggested by the exclamation-point diacritic;
2. downstep may be viewed as the consequence of a "floating" low tone between the two high tones, which thus triggers lowering by the more general process of downdrift, but is not otherwise realized ([8], [9]);
3. the downstepped sequence HM may be viewed as the expression of two distinct high tones, whereas an HH sequence is viewed as a single high tone spread over two syllables ([1]).

One obvious question is whether downdrift and downstep are phonetically the same, as we might expect from the

second hypothesis given above. One of our experiments suggests that Igbo downdrift and downstep are phonetically different, in a way that may be interpreted to support a variant of Clark's theory in [1].

2. Design of the experiments

We will mark Igbo tone using the system of [22, 11], which expresses the concept of downstep made explicit in Clark's theory: "high" is written with an acute accent; "low" is written with a grave accent; an unmarked syllable continues the previous tone; and a repeated "high" tone mark is interpreted as "mid." In this system, as in all the similar tonal orthographies known to us, downdrift is not marked, because it is assumed to be automatic (or if variable, only expressively so).

In this notation, the materials for our first experiment consisted of the 13 words:

HH:	ísi *head*	óke *male*	íre *tongue*
HL:	ísì *odor*	ókè *boundary*	írè *to be effective*
LH:	ìsîì *six*	òké *rat*	
LL:	ìsi *blindness*	òke *share*	ìre *effectiveness*
HM:	ísí *to cook*		íré *to sell*

On each trial, the subject was asked to read a word or phrase in one of three modes: addressed quietly to someone seated nearby; addressed to someone seated on the other side of a broad table, a little more than a meter away; or addressed to someone at the other end of a room, about 10 meters away. 195 utterances (five repetitions of each phrase in each mode) were elicited in random order. The subject was a man aged 45, from the village of Awo-Omamma in Imo State, about halfway between Owerri and Onitsha.

This procedure produces a good deal of pitch range variation, of the kind involved in thus "raising" or "lowering" the voice. Comparable measurement points in different tokens of the same tonal type have F0 values up to about an octave apart, which is several times larger than the difference between lexically-distinct tone categories in a given pitch range. As a result, we often find (for instance) that an initial L tone in a wide-pitch-range utterance is actually higher than an initial H tone in a narrow-pitch-range utterance.

Loudness and duration also vary in an experiment of this type, giving an independent indication of the speaker's level of vocal effort. We would like to point out that there are many other functional dimensions that are often associated with F0 effects that could be described in terms of "pitch range" variation. Examples include the speaker's overall level of arousal, the topic structure of the discourse, and the relative prominence of particular words and phrases. It should not be assumed that all such F0 effects will show the same patterns as the effects we have studied here, which are linked with the distance of an interlocutor.

We will not model the duration and amplitude effects of the range variation that we have induced, nor will we try to model all aspects of the F0 contour. We will represent the pitch of each syllable by a single value, taken automatically from the F0 time-function. We have tried various definitions for this representative value, including the value at the syllable mid-point and the average or median value; we find that the general structure of the data is the same, but the cleanest patterns result when we pick the maximum value for H tones, and the minimum value for L tones. We will discuss this point somewhat further when we consider the question of whether high tones are raised before low tones.

Like many (but apparently not all) tone languages, Igbo shows a pattern in which tones occuring later in a sequence are sometimes lowered relative to the values for the tones occuring earlier. In order to model these downtrends, we need to look at F0 patterns in longer sequences of tones. Thus our second experiment used the 13 phrases shown below, each produced six times in each of the three pitch-range modes. These phrases are personal names, or concatenations of personal names with "nà" *and*, a construction chosen because it is semantically flat, and also fails to induce the complex tone changes that occur with many syntactic combinations.

1	Díké nà Áma	HM L HH
2	Íke nà Áma	HH L HH
3	Íbè na Áma	HL L HH
4	Ńne Ụ́ba nà Íbè	HH MH L HL
5	Ụ́gwụ nà Íbè	HH L HL
6	Áma nà Íbè	HH L HL
7	Áma nà Íke	HH L HH
8	Íke nà Áma nà Ába	HH L HH L HH
9	Ọ́nụ́ ọma	HM HH
10	Ọ́lụ̀ká ọma	HLH HH
11	Ọ́nwụ̀ká ọma	HHM HH
12	Ábànóbì ọ́ma	HLHL HH
13	Ábàríkwú ọma	HLHM HH

The term *downdrift* applies to the regular lowering of H and L in alternating sequence, while *downstep* is used for the so-called "mid" tone, which is thus considered to be a lowered version of H. The materials in Experiment 2 are designed to let us characterize and compare these phenomena.

3. Results: Experiment 1

Figure 1 shows the some data from Experiment 1 as a a scatter plot in which the X axis shows the F0 value of the first syllable, and the Y axis shows the F0 value of the second syllable. The $y = x$ line is plotted as well.

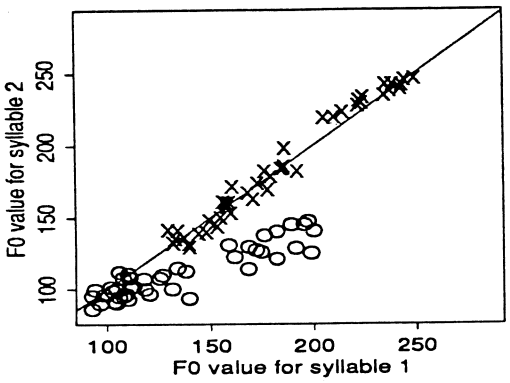


Figure 1: Igbo HH (X) LL (O) Disyllables.

We can see several things in Figure 1. First, although the points span a broad range, their relationship is quite a tight one: the correlation of the HH measurements is .988, while the correlation of the LL measurements is .905.

Second, the two H tones are generally about equal in value, but the second of the two L tones is usually lower than the first. Functionally, this helps distinguish an HH sequence in a low pitch range from an LL sequence in a high pitch range. We can express the same difference between the HH and LL data by saying that the LL trend is best fit by a line with a slope less than .5 and a non-zero intercept, while the HH data is not statistically distinguishable from $y = x$. Regression on the HH data gives a slope of 1.0 (standard error .025), and an intercept of -6.8 (s.e. 4.7). Regression on the LL data gives a slope of .44 (s.e. .032), intercept 51.9 (s.e. 4.5).

Figure 2 shows that the H/L proportion is systematically greater in the HL order than in the LH order, consistent with the fact that L is lower in the second position of LL sequences. Also, the HL data is clearly requires a non-zero intercept, while the LH data does not. Thus regression on the HL data gives a slope of .25 (s.e. .03), with an intercept of 60 (s.e. 6.1), while the LH data gives a slope of 1.3 (s.e. .08), with an intercept of -5.7 (s.e. 12). The HM data gives a slope of .92 (s.e. .03), and an intercept of -5.6 (s.e. 5.4). The fact that the HM slope is so much greater than the LL slope calls somewhat into question the interesting suggestion of Stewart ([12]) that the lowering of final low syllables is to be interpreted as "another manifestation of key lowering," a cover term in which he includes downstep.

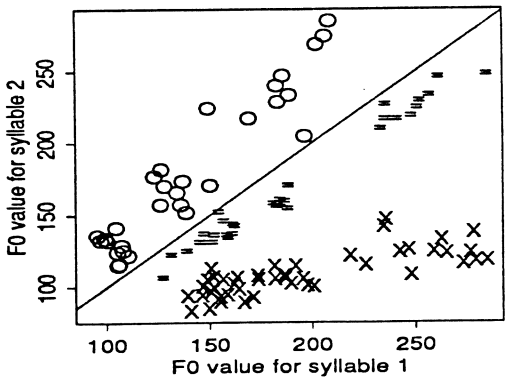


Figure 2: Igbo HL (X) LH (O) HM (=) Disyllables.

The patterns in the data of Experiment 1 suggest a model in which the F0 is predicted to be $T + T\,R\,F\,D$. Here T is a factor that has one value for H and another for L; F is a factor that is < 1 for L in second position and 1 otherwise; D is a factor that is < 1 for a "downstepped high" (here mid) tone; and R is a "latent variable" representing the pitch range of a given utterance. This model yields the following expressions for predicting the syllable 2 pitch (y) from the syllable 1 pitch (x) in a given utterance:

$$\begin{array}{ll} \text{HH} & y = x \\ \text{HL} & y = (FL/H)x + (1-F)L \\ \text{LH} & y = (H/L)x \\ \text{LL} & y = Fx + (1-F)L \\ \text{HM} & y = Dx + (1-D)H \end{array}$$

The model suggests values for the L and H parameters which are quite reasonable, given the speaker's observed F0 in low pitch-range utterances. The model also predicts significant intercept terms in the LL and HL cases, where we did find them, and no intercepts in the HH and LH cases, where we didn't. Unfortunately, it predicts an intercept in the HM case as well, where we didn't find one; the predicted intercept is small, since $1 - D$ appears to be small, but this may point to a problem in the model. $(T+TRF)D$ would predict no intercept for the HM case; space does not permit further exploration.

We should point out that this model embodies some probably wrong assumptions about the nature and environment of the lowering conditioned by D, namely that (in disyllables) it applies only to the HM sequence, and not (for instance) to the L in the HL sequence.

Having decided on the structure of our model (right or wrong), we can optimize its parameters with respect to the whole data set.[1] Doing this on the data of Experiment

[1] Treating the R parameters as latent variables, we use the downhill simplex method ([13]) to perform the optimization.

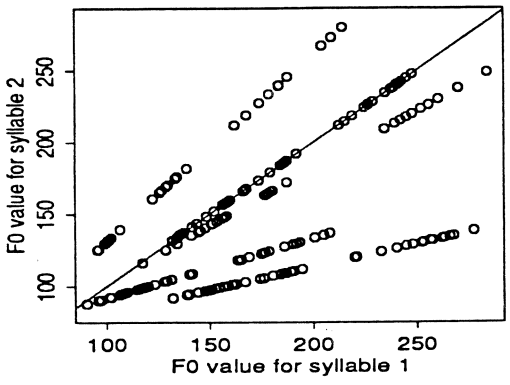


Figure 3: Model Values for Igbo Disyllables

1, assuming the $T+TRFD$ model with D applying only to the HM case, we get parameter estimates H=112, L=86, F=0.42, D=0.80, and an RMS error across all the data of 6.2 Hz. Figure 3 plots the result of replacing each data point with its model counterpart. In this plot, the basic features of the data set are well captured. Given space limitations, we leave it to the reader to convince herself that certain other simple functional forms for the model, such as $TRFD$, $T + R + F + D$, and so on, produce qualitatively wrong predictions. Their overall prediction error is significantly higher, but just as important, some of the key qualitative aspects of the data are misrepresented. We find it especially interesting that the all-multiplicative model makes such obviously wrong predictions, since this failure calls into question the popular practice of using a semitone scale for F0 contours.

4. Results: Experiment 2

The second experiment permits us to explore some other interconnected questions about the phonetics of Igbo tone. For instance, we can ask whether downdrift affects both H and L tones in the same way; whether downstep and downdrift are the same; and whether H is raised in front of L.

4.1. Raising of H before L

We take up the last point first, since it affects the measurements used in evaluating the other issues. In two independent studies of Yoruba, Connell and Ladd ([14]) and Laniran ([15]), suggest that H tones are raised before L. They offer various arguments, notably the fact that the final H in sequences such as HHL, HHHL, HHHHL, etc. is markedly higher than the non-final occurrences, and that this effect is not seen in sequences such as HHM, HHHM, HHHHM, etc.

There are no exactly comparable cases in Igbo, since Yoruba has a true, freely-distributed M tone, whereas Igbo's M tone is a downstepped H. Still, we ought to be able to see the effect in Igbo, *mutatis mutandis*.

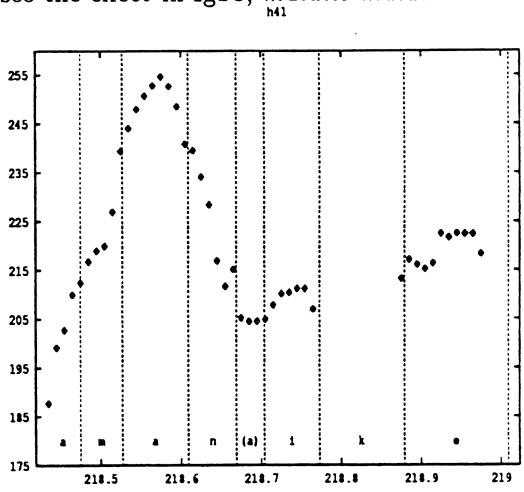


Figure 4: **Áma nà Íke** (far interlocutor)

Some examples seem to suggest that raising of H before L also holds in Igbo, as figure 4 clearly exemplifies. This figure shows the phrase "Áma nà Íke" (tone pattern HH L HH) spoken in the version with a distant interlocutor (and thus a high and broad pitch range). We can clearly see that the second syllable of "Áma" is much higher than the first.

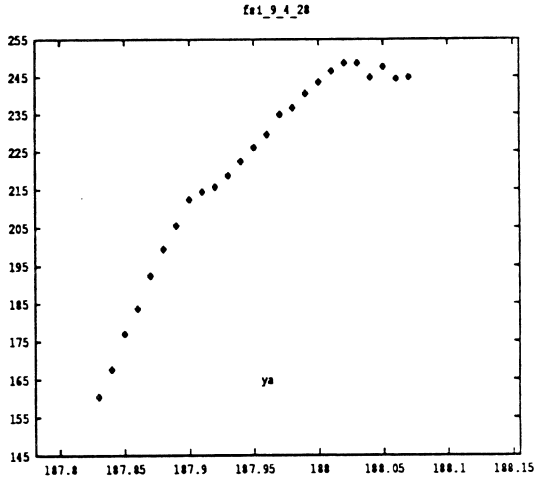


Figure 5: **yá** (FSI 9.4)

However, we are not convinced that this fact should be interpreted in terms of a rule of high-before-low raising. Essentially the same pattern for "Áma" would be possible in the absence of a following L tone. Indeed, the presence of more than one H syllable is by no means required. For instance, the F0 track (shown in figure 5) for the isolated monosyllable "yá" (*he*) shows a rising contour that is rather similar to the rising pattern seen on "Áma"

in figure 4, and quite similar rising patterns could be seen in utterances that begin with three H syllables.

In general, the starting F0 for utterance-initial H-tone stretches in Igbo is variable, but almost always lower than the peak F0, which is usually not reached until near the end of the sequence of H-tone syllables. It is not clear what governs the variability in this matter, but stretches of different numbers of H syllables show qualitatively similar patterns, whether in isolation, before L, or before M.

Furthermore, this behavior seems to be a particular case of a much more general non-equivalence of F0 values realizing adjacent equal tones. Similar phenomena in Yoruba are extensively discussed by Laniran. In our much more limited examination of the Igbo case, these phenomena seem consistent with a system in which a block of contigous "same-toned" syllables (or other tone-bearing units) have just the same number and type of F0 targets that a single syllable in the same context would have. This view is a sort of phonetic version of the phonological principle known as the Obligatory Contour Principle, or OCP. The phonological OCP principle has the consequence that a stretch of segments with the same value of a phonological feature must represent a single feature spread over the whole sequence. Our phonetic version of this principle suggests that a tone spread over a stretch of tone-bearing units will only be interpreted phonetically one time, regardless of the nature and quantity of the units that it is associated with (abstracting away from coarticulatory effects that may arise when targets are crowded too close together, or physical interactions with other on-going articulations). This hypothesis contrasts with systems in which each tone-bearing unit in such a block is subject independently to F0 interpretation.

Laniran's treatment of Yoruba is intermediate between such a "phonetic OCP" system and a system in which each tone-bearing unit is given independent F0 interpretation. The raising of the last H in a sequence when an L tone follow is one key case in which additional targets seem to be required. This is not the place to examine this matter in any detail, but it seems to us that the Yoruba examples might be amenable to a treatment in which the H-tone targets in the sequences HL, HHL, HHHL, etc. have the same number and the same timing principles regardless of the count of H tones.

Note that we accept Laniran's argument that the H tone target is raised in Yoruba before L vs. M. However, our Igbo data provide quite a strong argument against the view that a similar process operates there. This argument arises from comparing the relationship of the two H tone targets in HLHL sequences and in HLHM sequences. If H were raised before L to anything like the same extent that Laniran suggests for the case of Yoruba, the relationships should be significantly different in the two cases.

Figure 6 presents a scatter plot of this relationship in our data. As this figure suggests, the two cases are not statistically distinguishable. This makes it rather unlikely that Igbo has a general rule of H raising before L, and suggests that the apparent existence of such an effect in examples like figure 4 must be explained in terms of other F0-interpretation principles.

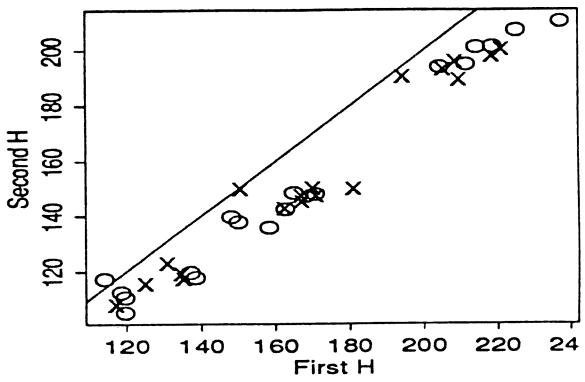


Figure 6: H/H Relationship in HLHL (O) vs. HLHM (X)

Applying this assumption, we will treat patterns such as HLH, HHLHH, and HLLHH in the same way for purposes of evaluating the scaling of downdrift, choosing the F0 minima and maxima (wherever they occur) as the points of reference.

4.2. Effect of downdrift on H and L

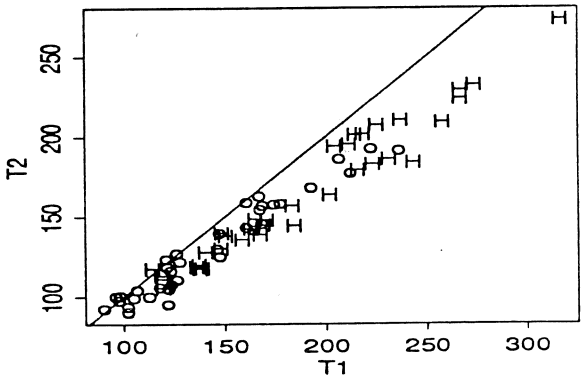


Figure 7: Downdrift on L (o) and H (H)

Does downdrift affect H and L tones in the same way? The answer appears to be "yes" in Igbo. Figure 7 shows successive L tones in an HLHL sequence are clearly lowered. Furthermore, the relationship between pairs of downdrifting H tones in an HLHL sequence (plotted with

character "H"), and the relationship between pairs of downdrifting L tones further along in the same sequence (plotted with character "o"), appear to be the same. Naturally H tones are higher than L tones, other things equal, and so the L tone pairs tend to be drawn from a lower region of the distribution—but the distribution seems to be the pretty much the same for both kinds of tone pairs.

4.3. Downstep vs. downdrift

Is downstep (the lowering of M tones after H) the same as downdrift (the lowering of the second H in an HLH sequence)? Our data suggest that it is not. We take downdrift data from sentences 2, 3, 5, 6, 7, 8, 10, 12 and 13 of experiment 2, using only the first downdrift in sentence 8. We take downstep data from the HM words of experiment 1, and from sentences 1, 9, and 11 of experiment 2. As before, the F0 relations are remarkably homogeneous and well controlled: $r = .978$ for the downdrifts, and .980 for the downsteps. The slopes and intercepts derived from regression, and the standard errors of the estimates, are given below:

	Intercept	std. err.	Slope	std. err.	N
HLH:	7.65	2.74	0.824	0.014	162
HM:	-2.30	4.14	0.924	0.021	85

It is pretty clear that the slopes are different in the two cases, and indeed it seems that downdrift imposes roughly twice as much lowering as downstep does. Thus we can provisionally reject the hypothesis that (from a phonetic point of view) downstep is "the same thing" as downdrift.

This is a surprise, since there have been taken to be good typological, historical and phonological reasons to equate the two processes. What to do? We see three paths: to accept that the processes are distinct; to salvage their phonological equivalence by excusing their phonetic distinctness on some independent grounds; or to take the view that downdrift is two units of downstep. This last move strikes us as the most interesting one. To make its content clearer, we return to our simple-minded model $T + TRFD$. If the D parameter is to be used to model either downstep or downdrift, it will have to take on a sequence of successively lower values as we accumulate lowerings in a phrase like HLHLHLH or HMMMM. The obvious way to do this is to write the formula as $T+TRFD^N$, where N starts at 0 and increments by 1 for each unit of lowering. Then we might increment N in several different ways, among them:

Tones:	H	L	H	L	H	L	H	H	M
N.1:	0	1	1	2	2	3	3	0	1
N.2:	0	0	1	1	2	2	3	0	1
N.3:	0	1	2	3	4	5	6	0	1

The first two are traditional "downstep equals downdrift" theories, differing in whether the lowering occurs on the L tone or on the H tone. Phonologists who have explicitly considered the issue have generally opted for N.1 (e.g. [1, 7]). The last idea (labelled N.3) is an example of a way of counting that makes a downdrift worth two downsteps—it says, basically, that N counts the distinct (in the OCP sense) tones in the string. As far as we know, it has not previously been suggested.

Note that all three of these ways of counting N predict that H and L tones "see" downdrift in the same way. They differ in the predictions they make about the relationship between downdrift and downstep, and the specific relationship between H and L tones (although the last point can only be explored given independent constraints on the functional form of the model, and on the other model parameters). Production scaling experiments, like those discussed in this paper, offer a good opportunity to compare such hypotheses in a quantitative way. Space does not permit a consideration of this comparison here, but it represents an interesting example of how phonetic evidence can be brought to bear on what has been taken to be a phonological issue.

As we noted earlier, there are several different ideas in the literature about the nature of downstep (i.e. "mid" tone) in Igbo. Some authors treat downstep as a case of downdrift in which the L tone between two H tones is "floating" (i.e. unassociated with any segmental material), and as a result is not realized phonetically except by virtue of causing the H tone that follows to be lowered. Others argue against this treatment, and suggest that the "mid" tone is either a third phonological category, or else simply an independent H tone, which is interpreted phonetically at a lower pitch value than the H tone that precedes. This point of view (suggested notably in Clark ([1]) is represented graphically in figure 8:

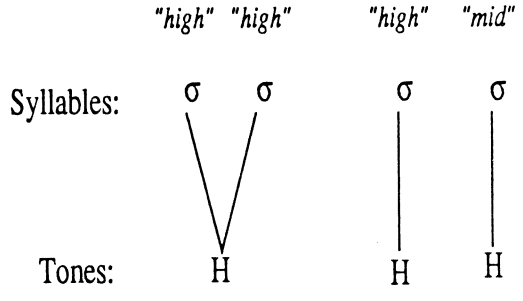


Figure 8: Clark's theory of the Igbo mid tone

Clark's arguments against the "floating low tone" theory of Igbo downstep are entirely phonological in character. She points to the regular appearance of a downstep at the boundary between cyclic domains when we would expect the tones on each side to be simply H; and also the regular

disappearance of a downstep in environments that can be simply characterized in terms of a rule "delete the middle of three adjacent H tones." She argues that both these classes of phenomena require unmotivated complexities to state if downstep in Igbo is treated in terms of the presence of a floating L tone.

Clark is still able to maintain the identity of downstep and downdrift:[2] she expresses them as a rule of REGISTER LOWERING, which says:

> Lower the high (and low) pitch registers at the juncture between a high tone and any following tone within the same phrase.

Our evidence with repect to this question is purely phonetic, and somewhat indirect. First, we observe that downstep and downdrift do not lower an H tone by the same amount; thus the floating L tone theory, which treats downstep as downdrift of H tones across an invisible L, must provide some excuse for this difference in values. Second, we observe that L tones are downdrifted in exactly the same manner, quantitatively, as H tones. Finally, we observe that downstep lowers successive H tones by about half as much as downdrift does. All of this comes together nicely if each successive new tone, whether H or L, causes the whole tonal system to "deflate" by a fixed amount. Then both downstep and downdrift are just the tick of passing tones.

Our version of Clark's REGISTER LOWERING rule says something like

> Increment the exponent of the D parameter whenever a new tone occurs.

It is equivalent, but more palatable psychologically, to maintain the current D value by successive multiplications, and reset it (partly or completely) at phrasal boundaries. Note that this formulation of the "deflation" principle differs from Clark's, in that hers still maintains the phonetic equivalence of downstep and downdrift. However, our formulation is consistent with Clark's phonological analysis, and inconsistent with the floating-L-tone analysis. In that sense, our findings support her position.

There is a small remaining problem: is H lowered after a phrase-initial L? Emenanjo ([16]) says that it is not, and Hyman and Schuh ([17]) claim that a dialect difference exists on this point. Since the amount of lowering involved would not be very great, and pitch values are also affected by overall pitch range, vowel type, etc., it is not clear how to tell if lowering occurs. Note that according to our model, the predicted effect (the change in F0 value of the first H tone in a phrase due to the presence or absence of an preceding L tone) is about 20% of the first H tone's difference from the base of the H register. In conversational speech for a typical male speaker, this difference might about 20 or 30 Hz., so that the predicted effect will be in the range of 4 to 6 Hz, which could easily be missed.

We do not believe that our data provides any way to check this question. If it turns out that H is not "deflated" at all following initial L, then our rule would have to be modified to exclude this case.

A more fundamental question is: why are there no downstepped low tones? Perhaps the lowered final L tones should be analyzed as a kind of downstep, but even if this is true, our treatment of L and H is not at all parallel. The floating-low-tone account of downstep offers a reason for this asymmetry, but at too great a phonological and phonetic cost. In effect, our account (in common with Clark's) requires that high tones are exempt from OCP restrictions in certain cases, while low tones never are.

There are many possible treatments for this state of affairs, among them a metrical account such as that offered in [2], [18]: H tones might be exempt from OCP violations just in case they are the heads of separate tonal "feet." Our experimental evidence does not bear directly on such explanations, but just helps to pose more clearly the problem that they aim to solve.

5. Summary of Experimental Conclusions

From the intricate and well-controlled patterns that emerge when pitch range is varied against tonal sequence in Igbo, we derive a number of tentative conclusions.

First, the scaling of Igbo tones requires a model that is neither multiplicative nor additive; we can describe it by saying that increasing pitch range adds to a basic H or L tone value, in units that are proportional to the basic value of the tone type in question. This produces relationships among tone values in sequence that are qualitatively different from those that would arise in purely multiplicative or additive models. Second, Igbo L tones appear to be lowered in final position. Third, Igbo H tone does not appear to be raised before L (as opposed to before M). Fourth, Igbo downdrifted H and L tones seem to behave identically. Fifth, downstep and downdrift are quantitatively distinct: downdrift imposes a significantly greater degree of lowering. This difference is consistent

[2] The main reason to maintain this equivalence seems to be a typological one. Downstep and downdrift, both very common but not ubiquitous, apparently have an implicational relationship—two-tone languages with downstep always have downdrift, but not vice versa.

with treating downstep and downdrift as two symptoms of a process that deflates both high and low pitch registers by a constant amount every time a phonologically-distinct tone is encountered.

These conclusions must be considered tentative for a number of reasons. We have looked at data from only one speaker; the task represents only one (artificial) style of speech; there are other ways to describe and explain each bit of evidence we have presented, specifically alternative statements of the environments and alternative functional forms for modeling. Still, we feel that our approach reveals things about the tonal system that would not come out without pitch range variation. We try to maximize (rather than minimize) such variation, and then use the rich statistical structure of the resulting data to distinguish among alternative hypotheses about the nature of the underlying system.

6. Tone and Intonation

The phonetic realization of lexical tone, in Igbo and in general, deserves careful consideration on its own merits. Aside from the intrinsic interest of the phenomena, we may cite both the theoretical importance of the phonological questions that arise, and also the methodological benefits of treating the phonetic interpretation of F0, which is relatively easy to measure, and behaves in a delightfully lawful way in controlled experiments. In considering the difficult problems of intonational analysis in natural speech, studies of lexical tone take on an additional significance.

One of the central questions of intonational research is whether intonation has a phonology, and if so, how we can decide what its categories and relations should be. The phenomenon of *word constancy* means that lexical tones divide into fairly clear surface phonological categories, much like other distinctions in lexical phonology. Despite their many other disagreements, the general run of humanity (linguists among them) generally agree on what tokens count as instances of the same lexical type. In plain language, any dolt can recognize the word "dog" when he hears it. The categories of lexical phonology, whether tonal or not, inherit a considerable amount of this comforting constancy. We may not agree on the feature content of an English /g/ or an Igbo mid tone, but we usually know one when we hear it. By contrast, there is relatively little agreement (among ordinary folk and linguists alike) about what utterance tokens count as instances of the same intonational type. As a result, we are likely to disagree both on how a "high rise" should be analyzed, and also on when we have heard one. This may be a fact about language (namely that intonation is not based on the same sort of categories as lexical phonology), or a fact about consciousness (namely that intonational categories are not accessible to reflection, whereas words are—for some reason perhaps connected to the phenomenon of reference).

Despite these difficulties, many linguists have analyzed intonation in terms of categories just like those used for the analysis of lexical tone, although several extensive traditions exist that are largely or completely non-phonological (e.g. [19]). In developing and evaluating tonal theories of intonation, it is natural to turn to instrumental analysis to bolster our categorically-weak perceptions, and to hope that patterns of objective measurements will help validate the postulated distinctions. Lexical categories of tone provide a phonological anchor point for studies of the interplay of tone, structure, and rhetorical or stylistic modulation. Looking at such patterns in a language with lexical tone, we see what an intonational language might be like if its pitch contours were properly analyzed as the interpretation of similar phonological structures.

This approach is rendered more difficult by the fact that lexical tone languages are typologically diverse and individually complex. There are many open questions (and perhaps more yet to be asked) about their phonological structure. The phonetic interpretation of lexical tone in phrasal context is still mostly *terra incognita*, especially when the many sources of rhetorical and stylistic variation are considered.

Some of the liveliest theoretical problems in intonational studies today concern the nature of phrasal downtrends in F0, and the status of what seem to be "mid" F0 values. These are questions that many lexical tone languages also bring to the fore, as we have seen—although it must be kept in mind that the facts and their explanations are apparently somewhat varied across languages (cf. [20]).

Let us continue to use the term "downdrift" to describe successive lowering in sequences of H and L tones, and the term "downstep" describe the middle case in a situation in which there are three distinctive tone levels after H, but only two after L. Let us add the term "F0 decay" to describe a gradual decay in F0 in sequences of like tones, whether H or L, such as has been reported for Luo in [24]. Then from a simple descriptive point of view, tone languages with two basic tone levels may have downdrift, downstep and F0 decay (e.g. Luo); downdrift and downstep but not F0 decay (e.g. Igbo, Efik); downdrift only (e.g. Hausa); or none of these (e.g. Vai). Languages with more than two basic tone levels may have downdrift (e.g. Yoruba) or lack it (e.g. Nupe). Those languages with downdrift may suspend it in utterances with a certain pragmatic force, typically described as "yes/no question" (e.g. Hausa),[3] or they may not (e.g. Igbo). Downdrift-

[3] Apparently such suspension in languages with both downdrift

suspension languages may also have the addition of a final high tone (Hausa), a final low tone (e.g. Kolokuma Ijọ), or no extra final tone (e.g. Efik) to phrases in which downdrift is suspended. Vai, which has no downdrift, also adds a final high tone in yes/no questions.

The treatment of such phenomena in lexical tone languages has been mirrored in various treatments of certain phenomena in English. We might mention two cases:

- the treatment of "stepping" contours in which a succession of accented syllables are realized on successively lower pitch values, with or without intervening dips;
- the treatment of the final mid-level F0 target in the contours variously called "vocative," "warning/calling", "stylized," etc.

The stepping contours have often been treated as cases of downdrift; their presence or absence might arise by using the "switch" that lets certain tone languages suspend downdrift to express (some pragmatic signal described as) a yes/no question. Here the common existence of downdrift in two-tone languages, and its suspendability in some of them, give us license to treat somewhat similar intonational phenomena in an analogous way. However, English then seems to require a rather fine granularity of downdrift suspension—in some proposals, every pitch accent is marked for presence or absence of a catathesis (i.e. downdrift) feature. This sort of freedom has never been reported for any lexical tone language—downdrift (if present) is automatic in sequences of H and L tones, and can apparently be suspended only on a phrase-by-phrase basis, whereas downstep applies only to H tones following other H tones. This does not prove the proposed descriptions of English are wrong, but it does remove the "blessing" of similarity to tone language phenomena.

What about the final mid-level F0 target in English "vocative" contours? The fact that it sounds something like an Igbo "mid" tone helped to license by analogy its treatment in [21] as a high tone downstepped by a floating low tone (not otherwise realized) associated with the previous pitch accent. If it were true that downstep in all terrace-tone languages is always an expression of floating low tones, then such an analysis for English would gain plausibility. If data in other terrace-tone languages turn out, like Igbo, to disconfirm the floating-low-tone theory of downstep, this treatment loses plausibility. The final vocative target in English might still be an Igbo-like downstepped high, but to maintain the analogy, any H phrase accent should be downstepped after any pitch accent ending in H. The only way to avoid this would be to analyze the donwstepped H phrasal tones as independent, while non-downstepped H phrasal tones (e.g. those in rising or high level contours) are just spread from the tone in the fore-going pitch accent. While not to be dismissed out of hand, such an analysis seems odd. Basically, English HM contours (such as the vocative) are just not distributionally similar to the similar-sounding sequences in Igbo.

If (as discussed under the heading of downdrift) we add a catathesis feature to every pitch accent (or even every tone) in English, we make it easy to treat final target in the vocative contour as a downstepped H—but the price is a high one. This move is informationally equivalent to doubling the language's tonal inventory, and will often be redundant with independently-needed marking of emphasis and pitch-range changes.

As an alternative, we might observe that English vocative contours also sound somewhat like H M sequences in languages like Yoruba, where the mid tone is a genuine independent category. Thus one might attempt an analysis of the vocative contour using an independent mid tone category. The fact that tone languages can easily have three or four paradigmatically distinct categories of tonal targets helps license this line of investigation. Of course, introducing a new tonal distinction raises the question of its distribution—if it occurs freely, then the multiplicity of resulting distinctions must be motivated. If its distribution is restricted to those cases where we feel we need it, then we need to explain this.

Of course, it is quite possible that the English vocative contour is neither like HM in Igbo nor like HM in Yoruba, since there are plenty of other models available among lexical tone languages. In analyzing English, we must deal one way or another with the fact (noted in [25]) that final falling contours seem to fall into at least three classes:

- ordinary terminal falls, where the endpoint is at the bottom of the speaker's pitch range;
- non-terminal falls, where the endpoint is somewhat above the bottom of the speaker's pitch range;
- vocative contours, where the endpoint is lower than the peak, but seems higher than either of the other two cases.

We could deal with these observations in many ways, using one, two or three phonological categories, with a wide choice of intensional and extensional definitions in each

and downstep affects only downdrift, and leaves downstep in place (e.g. in Efik). If this is true, then it is a bit of a problem for theories under which downstep and downdrift are the same thing. However, no such case has been modeled quantitatively as far as we are aware, and it is often reported in such cases that the pitch range as a whole is raised and narrowed, so it seems wise to reserve judgment.

case. We could treat all three cases as gradient scaling of a single phonological category; we could call the vocative target an Igbo-like downstepped H (or a Yoruba-like M), and distinguish the other two cases as gradient scaling of a single phonological category; we could call the vocative and non-terminal targets M, and distinguish them gradiently (as in [26]); we could treat all three cases as different phonological categories of fall (as in [25]); we could treat the non-terminal fall as a phonetic variant of fall-rise; and so on. All of these treatments (and more) might be licensed by appeal to the facts of some lexical tone languages.

No doubt tone languages will continue to be a source of inspiration to students of intonation. The careful comparative study of languages with lexical tone should give us a sense of what phonological and phonetic resources are available for the analysis of intonational phenomena, whether intonation contours turn out to be homologies or merely analogies of lexical tone contours. However, at the present stage of development of our knowledge, we should be careful not to accept superficial analogies too easily, whether from tone to intonation or from one tone language to another. Interest in universal principles needs to be tempered by respect for the facts of each language, and by willingness to recognize that the number of carefully-modeled tone systems is small compared to the apparent diversity of human potential in this area.

References

1. Clark, M. M., *The Tonal System of Igbo*, Foris Publications, Dordrecht, 1990.
2. Manfredi, V., *Agbo and Ehugbo: Igbo Linguistic Consciousness, its Origins and Limits*, Harvard University PhD Thesis, 1991.
3. Liberman, M. Y., Schultz, J. M., Hong, S., and Okeke, V., "The Phonetics of Igbo Tone," pp. 743-746, ICSLP 92, Alberta, 1992.
4. Liberman, M. Y., and Pierrehumbert, J. B., "A Metric for the Height of Certain Pitch Peaks in English," JASA 66 (S1), 1979, S130.
5. Liberman, M. Y., and Pierrehumbert, J. B., "Intonational Invariance under Changes in Pitch Range and Length," pp. 157-234 in M. Aronoff and R. Oehrle, eds., *Language Sound Structure*, MIT Press, Cambridge, 1984.
6. Carrell, P. L., *A Transformational Grammar of Igbo*, West African Language Monographs, no. 8, The University Press, Cambridge, 1970.
7. Goldsmith, J., *Autosegmental Phonology*, MIT PhD Thesis, 1976.
8. Clements, G. N., and Ford, K. C., "Kukuyu Tone Shift and its Synchronic Consequences," *Linguistic Inquiry* 10:179-210, 1979.
9. Pulleyblank, D., *Tone in Lexical Phonology*, Reidel, Dordrecht, 1986.
10. Welmers, W.E., and Welmers, B. F., *Igbo: a Learner's Dictionary*. UCLA, Los Angeles, 1968.
11. Nwachukwu, P. A., *Towards an Igbo Literary Standard*, Kegan Paul International, London, 1983.
12. Stewart, J. M., "Niger-Congo, Kwa," pp. 179-212 in Thomas Sebeok, ed. *Current Trends in Linguistics*, v. 7, Mouton, The Hague, 1971.
13. Press, W. H., Flannery, B.P., Teukolsky, S. A., Vetterling, W. T., *Numerical Recipes*, Cambridge University Press, 1988.
14. Connell, B., Ladd, D. R., "Aspects of Pitch Realization in Yoruba," pp. 1-30, Phonology 7, Cambridge University Press, 1990.
15. Laniran, Y., *Intonation in Tone Languages: The Phonetic Implementation of Tones in Yorùbá*, Cornell University PhD Thesis, 1992.
16. Emenanjo, E. N., *Elements of Modern Igbo Grammar*, Oxford University Press, Ibadan, 1978.
17. Hyman, L. M., and Schuh, R. G., "Universals of Tone Rules: Evidence from West Africa." *Linguistic Inquiry* 5:81-115, 1974.
18. Manfredi, V., "The Limits of Downstep in Agbo Sentence-Prosody," IRCS Workshop on Prosody in Natural Speech, August 1992.
19. 't Hart, J., Collier, R., and Cohen, A., *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*, Cambridge University Press, 1990.
20. Dunstan, E., ed., *Twelve Nigerian Languages*, Africana Publishing Corporation, New York, 1969.
21. Pierrehumbert, J. B., *The Phonology and Phonetics of English Intonation*, M.I.T. PhD Thesis, 1980.
22. Welmers, W. E., *A Grammar of Vai*, University of California Publications v. 84, University of California Press, 1976.
23. Elimelech, B., *A Tonal Grammar of Etsakǫ*, University of California Publications v.87, University of California Press, 1978.
24. Tucker, A. N., and Creider, C. A., "Downdrift and Downstep in Luo," pp. 125-134 in R. K. Herbert, Ed., *Proceedings of the Sixth Conference on African Linguistics*, OSU Working Papers in Linguistics no. 20, 1975.
25. Pike, K., *The Intonation of American English*, University of Michigan Press, 1945.
26. McLemore, C. A., "Prosodic Variation across Discourse Types", *Proceedings of 1992 IRCS Prosody Workshop*, IRCS Technical Report, University of Pennsylvania, 1992.

A PROSODIC ANALYSIS OF TWO EARTHQUAKE NARRATIVES

Margaret Luebs

University of Michigan
Ann Arbor, MI 48109

ABSTRACT

This paper analyzes the prosody of two narratives of the 1989 San Francisco earthquake, in order to show that a consideration of prosody can be an important part of narrative analysis. The focus is on two aspects of the narratives: their structure, and the humor used in them. It is shown that prosody plays an important role in delineating the structure of a narrative, and perhaps should be used as a criterion when choosing a theory of narrative structure. It is also shown that prosody has an equally important but less easily described role in signalling attempts at humor.

1. INTRODUCTION

This paper is a continuation of work I began in a paper titled "Earthquake Narratives" (Luebs 1992). In that paper, I proposed that the stories people tell about their earthquake experiences form either an identifiable sub-genre of the well-known discourse genre "narratives of personal experience" (Labov 1972) or part of a network of types of such narratives. In my analysis of earthquake narratives I discussed narrative structure, but did not address issues of intonation or other prosodic features. In the current paper I attempt to show how a consideration of prosody affects and interacts with a more traditional non-prosodic narrative analysis.

There has been a great deal of work done on spoken narrative, but much of it omits discussion of prosody. An important exception to this is the work of several anthropological linguists (e.g. Tedlock 1983, Sherzer 1990, Woodbury 1987), but their work differs from mine in that it usually deals with formal tellings (or rather retellings) of traditional stories, whose prosodic patternings are (relatively) fixed and easy to recognize. However, the prosody of casual narrative deserves study as well, since discourse analysts who study casual conversation have shown that prosody plays an important role there (e.g. Gumperz 1982, Sacks, Schegloff & Jefferson 1974, etc.).

The paper is divided into two main sections: narrative structure and humor. In the first section I begin with a brief discussion of theories of narrative structure and then examine my data for evidence of prosodic correlations with the theories. In the second section I discuss the use of prosody in expressions of humor and other dramatic devices present in the data.

2. DATA

The current paper analyzes sections of two earthquake narratives (see transcripts in Appendix A), told by a 29 year old woman (Sarah) and a 28 year old man (Dan). (This is a subset of the original data, the narratives of 14 people who experienced the 1989 earthquake in northern California.) The speakers are both white, both native speakers of American English, and both veterans of previous quakes (although the man is not a native of California). I chose to focus on these two mainly because the recordings are of better quality than some of the others, and also because it seemed interesting to examine tapes of both a man and a woman, of similar age. I have pitchtracks of both narratives.*

I taped the narratives about two months after the quake, which meant that the experience was still fresh in people's minds but the stories had been told many times already and have a "practiced" quality about them.

The taping was done in formal interview style, with the participants seated together at a table, and the subject given the cue "Tell me about your earthquake experience." However, other factors help to make the narratives more "natural". For example, since all subjects were either family members or close friends of mine, there was a good rapport between us. Also, the subjects knew that because I had not experienced the earthquake, I was truly interested in hearing their stories, not just in collecting them for a project. In fact, because of my great interest I often stepped out of the role of passive listener to ask questions or make comments. These interruptions may seem like annoying breaks in the narrative flow, but would be quite common in any truly "natural" narrative-in-conversation.

A note on notation: I have not tried to represent prosody in the transcripts in Appendix A, except for impressionistic details such as italics for extra stress. Within the body of the paper I occasionally represent intonation using boldface for a fall in pitch and all caps for a rise.

* I am grateful to Cynthia McLemore for pitchtracking the narratives, and for giving me advice and encouragement.

3. PROSODY AND NARRATIVE STRUCTURE

3.1 Theories of Narrative Structure

In my earlier paper I discussed Labov's (1972) classic 6-part structure of narrative (abstract, orientation, complicating action, evaluation, result/resolution, coda -- all of which are optional except complicating action), rejecting it in favor of Johnstone's revision (1990). Johnstone shows that a story consists of a mix of orientation clauses and narrative clauses, that result/resolution is simply a part of the narrative core, and that the abstract and coda function as the entry to and the exit from the story. Both Labov and Johnstone make it clear that evaluation should not be conceived of as a formal section of a narrative, but rather a functional strategy which may take many forms throughout a narrative.

I then argued that because Labov's and Johnstone's theories of narrative structure are so general, it is hard to use them to study particular types of narrative. Labov's theory in particular seemed suspect to me, because he based it mainly on one type of narrative, what he calls the "Danger of Death" narrative (subjects were asked "Were you ever in a situation where you were in danger of being killed?"). If it could be shown that different types of narratives have their own different types of structures, this would be evidence for the need for a new theory of narrative structure. Accordingly, I went on to propose my own very specific structure of an earthquake narrative, as follows: (1) orientation; (2) quake begins; (3) speaker responds to the quake, as do objects and other people; (4) quake ends; (5) speaker sees-realizes-finds out about-responds to things.

This structure works very well -- all the quake stories I collected from adult speakers follow it almost exactly. But what is particularly interesting about it is that it is so dependent on the actual physical occurrence being talked about (i.e. the earthquake). The earthquake really seems to drive the story, determine how it will be told and in what order. Intuitively this does not seem odd, but it is in conflict with the vast amount of linguistic literature which claims that the content of an utterance is totally irrelevant in a discussion of its structure.

3.2 Prosodic Correlations

When the data is examined for prosodic correlations with narrative structure, there are some interesting results. Perhaps the most surprising is the correlation with the narrative structure of earthquake stories. The five part structure is of course not five equal parts, but rather three sections divided by two boundaries: (2) quake begins, and (4) quake ends. The boundaries themselves are not always well marked, even lexically: Sarah does say "and all of a sudden we felt this shaking" (line I.10) but Dan does not specifically introduce the quake; Dan does say "right after it.. after it.. stopped.." (line II.72) but Sarah never specifically refers to the ending. However, there is a difference in the prosody of part (3) compared to parts (1) and (5) which seems to be related to the topic being discussed.

In part (3) the subjects are describing the quake and their immediate response to it, and in this section they sound decidedly more animated than they do in sections (1) and (5). This makes sense intuitively: the events described are exciting, what with all the unusual movement going on, and so the speakers use prosody to try to convey this. One of the things they do prosodically is to extend their range, bouncing from their normal low pitch to a higher than usual high pitch and then back down again, in many of the intonation units in this section. For instance, Sarah's average high pitch (i.e. highest pitch in each intonation unit divided by number of intonation units) is 280hz in part (1) and 260hz in part (5), but in part (3) it is 335hz. Likewise, Dan's average high pitch is 138hz in part (1), and 139hz in part (5), while in part (3) it is 165hz. For both speakers, the average low pitch stays about the same in each section. These numbers are shown in the chart in Table 1 below. Average difference refers to the difference between the highest and lowest points in each intonation unit divided by number of intonation units.

I. Sarah			
	avg high	avg low	avg difference
prequake	280	155	125
midquake	335	156	179
postquake	260	163	97
II. Dan			
	avg high	avg low	avg difference
prequake	138	96.5	41.5
midquake	165	92.5	72.5
postquake	139	91.3	47.5

Table 1: Differences in pitch range within intonation units in different parts of the narratives.

This result is only a very rough suggestion of iconicity between prosody and content or structure, but it is certainly interesting. There may be other things going on as well: for instance, I believe loudness and speed also increase in part (3) but I have not measured this. One real problem with my calculations is the question of what an intonation

unit is: for this study I used very loose criteria for deciding what one is, and it is possible that the results would change if the units were divided differently.

There are no such obvious correlations between Labov's narrative structure and the prosody in these narratives. In the past I have tried to use Labov's structure to analyze narratives and have had trouble deciding whether a line was complicating action or orientation, for instance. Although I had hoped that prosodic factors could make this easier, they do not seem to. Although one might expect some prosodic distinction between orientation and complicating action, since the latter is moving the action forward but the former is not, I have not yet discovered any such distinction.

The other clearly present prosodic correlation with narrative structure is that of prosodic paragraphs or paratones (Brown 1977, p 86), discussed in greater detail by Janet Bing (this volume). I have not yet delineated these precisely in the earthquake narratives; however, it seems clear to me that the text is divided into a series of these little episodes, which are marked prosodically as well as lexically. The beginning of each episode is typically preceded by a long pause, and then has a dramatic rise and fall (although the line may not be very exciting in content), plus a discourse marker such as "so" or "and then". The introductory lines of these speech paragraphs are particularly obvious in Dan's narrative because he keeps restarting -- he seems to begin his story, but then backs away and gives some more background information, which tends to be lower-pitched overall, with less-dramatic falls and rises. An example is given in Figure 1 (line II.35).

4. PROSODY AND HUMOR

In this section I discuss some of the speakers' attempts at humor in their narratives, and how this involves prosody. Although I had originally hoped to be able to be able to identify the prosodic cues most typical of humor, this turned out not to be possible. Several different prosodic cues are associated with the examples of humor in these narratives; it will take more research to determine exactly how the system works.

Humor is a logical area in which to study prosodic cues, since it is an accepted bit of folk linguistics that prosody is a part of what makes funny things funny. I have found that often people will not think a transcript of a narrative is funny at all, but when they hear the tape they laugh out loud. I have been particularly interested in the humor in these narratives, as one might not expect humor in stories about something that killed 65 people and caused $7 billion of property damage. However, humor is in fact a common response to disaster, at least in our culture (see e.g. Oring 1987, Wolfenstein 1957, for discussion). The humor in earthquake stories seems to be both a way of coping with this disaster (by belittling it) and a way of defining how the community will approach the on-going threat of earthquakes.

In another earlier paper (Luebs 1991) I discussed the different examples of humor that are present in the earthquake data. By humor I do not mean specifically jokes, but rather more subtle attempts at humor (something like what Long & Graesser 1988 call "wit"). Typical quake humor seems to deal mainly with (1) the "stupid" things people do and think during quakes, (2) the absurd things or situations caused by quakes, and (3) ridiculously unsafe (considering that this is earthquake country) buildings or other structures, and people's lack of preparedness for quakes (despite the ever-present threat).

One thing I have found difficult is how to decide what in the narrative is actually intended to be humorous. In this paper I will simply assume that something is humorous when I or the subject laugh at it. I am sure this criterion leaves out some attempts at humor and includes some things not meant to be humorous; however, I think it is adequate for a preliminary look at humor and prosody. In this section I will first describe the prosody of utterances which provoke laughter, and then discuss the similarities and differences.

4.1. Sarah's narrative

In Sarah's narrative, laughter occurs at lines I.7-8, I.22-28, and I.48-51. The first of these is in the pre-quake section, while the other two are mid-quake. The pre-quake humor, line I.7: "in one of those *old army buildings* (laughs) made of cement" (see Figure 2) is an ironic comment about the unsafeness of the building she and her co-workers use as an office (humor type 3). Line I.8 "you don't know how well they're reinforced" explains the joke more clearly. Sarah pauses briefly before line I.7, puts extra emphasis on the words "old army buildings" but pitches them fairly low for her (180-200hz) in a kind of monotone, and lengthens the word "old" so that it is about as long as the following two-syllable words, causing it to stand out. She also laughs, probably the safest way of telling someone that something is supposed to be funny.

In lines I.22-28 she is describing the foolish action of a friend and her and others' response to it (humor type 1). In line I.25 she tells what the friend did ("*stuck his head* out of the WINdow"), in line I.26 she explains why this was dumb (the windows "slam down"), and in lines I.27-28 she describes their reaction to this. She slows down a little and stresses "stuck his head," with a HL on "head" and on "window"; she also slows down and stresses both words in "*slam down.*" Her voice is pitched very high through this whole section, mostly in the high 200's, 300's and above, and she is also laughing. When she imitates their screaming she goes even higher, perhaps into falsetto, laughing hard.

In lines I.48-51 she is also "reporting" speech, first the voice of a teacher who did not understand why her students were running outside, and then the (collective) voice of a group of young students defending themselves (humor type 2). For the teacher's voice Sarah goes up high and stays there, in almost a monotone, and stretches out the words a little so that sounds such as the vowel in "what" are much clearer than they would be in normal speech (see Figure 3). When she imitates the students she lowers her voice a little (not much) and puts in more normal intonation, but it still sounds different from normal speech (or normal screaming). She also laughs during her imitation of the students.

What these three examples have in common is some distortion of the "normal" speech patterns, but the nature of the distortion varies. Extra stress and volume are common, as is a reduction in speed. Sarah's cues aren't subtle; she has road signs all over saying "this is funny -- be sure to laugh at this." However, she uses these prosodic devices in other parts of her narrative where she is not trying to be funny (i.e. lines I.32-33, line I.44). She seems to use unusual intonation as often for excitement or emphasis as for humor.

Sarah avoids humor in the post-quake part of her narrative. In lines I.67-73 I think I was expecting her to be leading into humor, making fun of herself and the other teachers, but instead she ends line I.72 with a rise which makes her sound serious, hinting at the disaster she was soon to discover. Later, in lines I.85-89 she continues this tone, sounding almost ominous as she describes the "black puff of smoke" that rose "up from over the hill". She slows down and speaks softly but intensely. In addition, the words "black puff of smoke" are spoken with slight pauses between them. (see Figure 4).

4.2. Dan's narrative

In Dan's narrative, laughter occurs around lines II.15-20, lines II.45-48, line II.57, lines II.66-71, and lines II.82-83. He has pre-quake, mid-quake, and post-quake examples of humor. The first is his first attempt at beginning the story. He starts out in a somewhat dramatic voice "OcTO**ber** **seventeenth**, NINEteen eighty-**nine**" but then, perhaps because he thinks he sounds pompous, begins to get silly -- he says "I.. Dan SULliVAN" with exaggerated intonation, does a dramatic swooping HL on "FOR**mer** housemate" and then lapses into teasing with "erstwhile lover." This doesn't fit into my 3 types of humor -- it is more the nervous joking of someone embarrassed about being taped -- but it may also be related to the struggle my subjects experienced in talking about a frightening experience while trying to maintain their control over it.

Line II.45 caused me to laugh during the taping, but I am not sure Dan meant it to, because after he says it he appears to try to defend himself against my laughter. However, this "defense" is also amusing. The intonation in line II.45 is notable because of its *lack* of drama: the line is funny because it is incongruous for Dan to have been making such an exciting announcement so casually. He returns to signalling humor with exaggerated intonation in the following lines, using extreme high-lows on "that was really GO**od**", "because they didn't KNO-**ow**", and "that it was an EARTH**quake**" (see Figure 5).

Line II.57, "uh.. I said "I think we should get under the TA**ble** now" (see Figure 6) is funny because as a polite suggestion it is an understatement of what someone might be likely to say in this situation (such as "ohmigod get the hell under the table!). Dan signals that it is funny by using unusual intonation for a suggestion: staying fairly low (90-110hz) and unstressed until the bounce up to 240hz and then back down to 100hz for "table".

Lines II.66-71 are more subtle humor. The most marked line, II.66 "HE **was** kind of SMI**ling** at first", is again marked as funny by intonation -- the steep rise on "he", and the rise and fall on "smiling" -- I think "he" is also lengthened a little. Dan also laughs a little at line II.69 "and the smile went away".

In lines II.82-83 he is making fun of himself (humor type 1). In line II.82, "uh and THEN I **realized** it PRO**bably** wasn't so SMART **to go out** on the BAL**cony**," he stresses "probably" "smart" and "balcony" and he laughs a little, as do I.

So Dan relies mainly on intonation, and also somewhat on stress. Although I have not measured it exactly, I believe he does not make much use of changes in speed or loudness. He is more subtle than Sarah, at least in this brief section, and does not use as many cues.

4.3 Comparison

Although it is evident that Sarah and Dan use somewhat different strategies for humor, there are also similarities. The humorous remarks are always marked prosodically in some way -- extra stress, slowing down and/or lengthening words, unusual or exaggerated intonation. Dan seems to prefer marked intonation, while Sarah, who uses dramatic intonation quite often when she is not being funny, seems to prefer extra stress and length.

It is also apparent that the strategies used for humor are not so different from the strategies used for showing excitement or emphasis. This helps explain why people sometimes "miss" a joke -- the cues are so numerous and varied, and so similar to cues for other things. On the other hand, none of the examples of humor in these narratives is hard for me to spot (but then again, these are my friends). Both speakers made ample use of cues like laughter (and smiles,

as I recall), which may be needed to make it perfectly clear what is intended to be funny.

Sarah's post-quake seriousness has interesting prosody as well, particularly in the way it contrasts with the prosodic strategies she uses for humor. It seemed to me, in my earlier study of the earthquake narratives, that many of the speakers switched between humor and seriousness in different parts of their narratives, depending, perhaps, on whether they were focussing on the humorous aspects of the earthquake or on its disastrous effects. It would be interesting to go back to those other narratives and see what prosodic devices are used by other speakers to convey seriousness or tragedy.

5. CONCLUSION

Although this exploration of prosody in narrative is only preliminary, it does raise some provocative points. First, besides adding to evidence for the existence of paratones, it suggests that prosody and narrative structure are even more closely connected than has been previously thought. It strengthens my argument for a specific structure for earthquake narratives and leads to the question of what other structures may be reflected in prosody. It also suggests that researchers interested in narrative and other linguistic structures larger than the clause would benefit from including prosody in their analyses.

The section on humor is more tentative than that on narrative structure, but I believe this is a reflection of the complexities of the subject. Although my data do not identify exactly which prosodic cues signal humor, they do show that humor can be signalled in a number of ways and that prosody is always involved. This does not mean, however, that any and every type of prosody can signal humor. One interesting phenomenon in the data is the evidence that different speakers have favorite prosodic strategies for humor, which perhaps are more or less successful. A possible future research project in this area might study professional comedians, and/or ordinary people rated as funny or not funny, to see how they use prosody differently.

REFERENCES

1. Brown, Gillian. 1977. *Listening to Spoken English.* London: Longman, as discussed in Brown, Gillian and George Yule. 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.

2. Gumperz, John J. 1982. *Discourse Strategies.* Cambridge: Cambridge University Press.

3. Johnstone, Barbara. 1990. *Stories, Community and Place: Narratives from Middle America.* Bloomingston: Indiana University Press.

4. Labov, William. 1972. "The transformation of experience in narrative syntax" in *Language in the Inner City.* Philadelphia: University of Pennsylvania Press.

5. Long, Debra L. and Arthur C. Graesser. 1988. "Wit and humor in discourse processing." *Discourse Processes* 11:35-60.

6. Luebs, Margaret. 1992. "Earthquake narratives." Paper presented at the 18th annual meeting of the Berkeley Linguistic Society.

7. Luebs, Margaret. 1991. "Earthquake stories." Unpublished paper.

8. Oring, Elliott. 1987. "Jokes and the discourse on disaster." *Journal of American Folklore* 100:276-286.

9. Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. "A simplest systematics for the organization of turntaking for conversation." *Language* 50: 696-735.

10. Sherzer, Joel. 1990. *Verbal Art in San Blas: Kuna Culture through its Discourse.* (Cambridge studies in oral and literate culture: 21) Cambridge: Cambridge University Press.

11. Tedlock, Dennis. 1983. *The Spoken Word and the Work of Interpretation.* Philadelphia: University of Pennsylvania Press.

12. Wolfenstein, Martha. 1957. *Disaster: a Psychological Essay.* Glencoe, Ill: The Free Press.

13. Woodbury, Anthrony C. 1987. "Rhetorical structure in a central Alaskan Yupik Eskimo traditional narrative" in *Native American Discourse: Poetics and Rhetoric,* ed. Joel Sherzer and Anthony C. Woodbury, pp 176-239. Cambridge: Cambridge University Press.

Appendix A: Transcripts

I. Sarah

S OK.
we were all having a meeting,
it was in the afternoon after we had let the kids off
and they were all playing.. in.. whatever place they were playing,
and we were all sitting inside in the.. small conference room,
(1.5) in one of those *old army buildings* (laughs) made of cement,
you don't know how well they're reinforced?
we were sitting in there in our meeting,
and all of a sudden we felt this shaking,
and all of us.. turned to each other and we said: "an earthquake"
and then.. it kind of dawned on us,
wow,
this is really.. *bigger* than we've ever felt *before*,
and so we kind of went "an *earth*quake",
you know.. with.. more emphasis,
and we all started running,
towards the door and outside..
because if we're out.. far enough there's kind of a.. grassy field?
and we could just kind of escape to that grassy field.
so we started running outside,
and one of my friends,
who's really a smart guy,
but he was really interested in seeing the earth move,
and he *stuck his head* out of the window,
we have these windows that kind of (laughs) you know *slam down* (laughs)
and we were all screaming "Steve!
Get your head out of the window it's gonna" (I: oh no) "fall down on you" (laughs)
So out we ran,
onto the lawn,
and I guess there was just a few seconds,
between that first initial *jolt* (I: uh huh)
and then when it started really shaking,
and we felt that shaking
and it was *really*.. quite significant you know (I: uh huh)
we could really feel it.. rolling,
nothing broke--
nothing went out--
all the lights stayed on--
no problem,
but the *kids*,
they were in the dorm--
that *were* in the dormitories--
went rrrrunning outside to the parking lots--
and this teacher started screaming at 'em,
cause I don't know what the teacher had been doing--
but didn't.. she didn't *realize* that there had been an earthquake,
and.. so she was like.. "what are you doing running out of the dorms,
you're supposed to be in there at quiet time".
and they're all like "no it's an earthquake (laughs)
we're doing what we're supposed to".
so that was kind of.. you know.. interesting.
and then.. they were from
I Could you see the ground moving?
people have told me about seeing that.. but I'm --
S I have sort of a vague memory w- sitting in the room--
seeing it move a little bit.
when we were outside I know I definitely didn't,
I just felt a very definite shake. (I: uh huh)
but.. inside I sort of have a vague?
it seemed like things were moving,
but it wasn't really clear.. if they were or not?
I Um hm.. well you were pretty far..
S Yeah.. we were..
yeah..
so we all.. came back inside to finish our meeting up?
and we.. were kind of like joking about well,
where do you think the epicenter was,
we were all guessing--
and how much do you think it was,
and we were all guessing--
and nobody was even close? you know,
as to how significant it was?
and then,
I Did you still have power?
S And we had power and everything.
and.. a few minutes later it got to be dinnertime,
and we were.. we were sitting around,
and we decided.. you know.. we maybe we should monitor,
cause these kids live up in Piedmont? (I: uh huh)
and we wanted to make sure that *their* homes were OK, (I: yeah)
cause.. um we weren't really sure.. so,
we.. we were starting to listen to the radio,
and just before we started to listen to the radio..
up from over the hill,
where you look over towards the m-- you know.. San Francisco Marina area--
there was this black puff of smoke, (I: ohhh)
and it started rising--
and it went across the sky,
and that's when we really got frightened,
oh my gosh.. you know,
something significant has happened..
and that's when we turned on the radio,
and.. started monitoring and listening and stuff?

II. Dan

D [Reading] Earthquake tape two?
how many stories do you have--
I Um - t- - eleven.
You'll be twelve.
D Eleven.
I'll be twelve.
OK.
The twelfth person.
I've been meaning to write this down.
my friend Greg Myer wrote a letter to his parents,
you know, kind of detailing it-- (I: uh huh)
and I I should have done that while it was still fresh,
but I want my parents want to know about it too so, (I: yeah)
write it all down.
October 17th, 1989,
uh I..Dan Sullivan,
Margaret's former (I: laughs) housemate,
and uh.. erstwhile lover,
no no scratch that,
that's off the record, (I: laughs)
D (laughs) and uh (laughs)
I It's also a lie (laughs)
D (laughs) Dogs, go away.
and the dogs are harrassing us--
uh I was in class at the time,
um.. I have a.. a Tuesday-Thursday afternoon class.
and uh.. it was one of my seminar classes,
with the chairman of the department.
and there were just six of us,
in the first.. of the first year students. (I: uh huh)
We all sit around this big table,
and uh.. class goes from.. is it 3:15 to 5:30? or something like that?
I oh my God
D and uh.. yeah,
so it was toward the end,
and uh.. uh there're a couple.. there are two.. o- other Californians.. in the class,
and then everybody else is from oth-
there's a Korean woman, a Singaporean, and a a guy from from.. Milwaukee.
um tsk.. but it happened,
I I keep trying to go back and figure out exactly what happened at the time, um
you know it's it's like you *don't* want to relive it?
but you want to re-experience it again? (I: uh huh)
you want to know what happened? (I: uh huh)
but.. I think I stood up and I said-- um
"This is an earthquake" (I: laughs)
and kind of identify it for everybody. (I: uh huh, laughs)
and the people in the class afterwards said that that was really good,
because they didn't *know* .. that it was an earthquake, (I: laughs)
they were wondering what the hell was going on.
but you could really hear it,
um there was a re- there was a rumble,
there's that sound. (I: uh huh)
uh.. the room I was in had nothing.. free standing,
it was all fixed.. you know,
fixed lights and everything--
and so about probably three or five seconds into the earthquake,
uh.. I said "I think we should get under the table now."
because it was.. becoming uh apparent that it was a very strong one.
big thick slab table something like this one,
really heavy,
so we all.. got under the table,
very quickly.
and I remember looking at my professor,
who's this gen.. he's the genius in the department,
he's in you know a real luminary in the field.
he was kind of smiling at first.
and then it kept building,
and getting stronger and stronger--
and the smile went away,
but I'll never forget *looking* at *him* under the table,
along with the rest of us.
And uh.. right after it.. after it.. stopped,
I I ran out onto the balcony,
to look.. and see what else was going on,
I could see the palm trees in the Quad,
sha- uh sw- swaying back and forth and stuff,
and there were people outside--
and.. kind of panic-stricken looks on their faces,
and some people were.. smiling--
and.. some people were kind of.. whooping,
and.. you know there's that nervous reaction. (I: uh huh)
uh and then I realized it *probably* wasn't so *smart* to go out on the balcony..
so I went back inside, (I: laughs)
(laughs) and the alarms went off.. immediately.. in the building,
um.. the communication department was rebuilt a couple of years,
it's in the Quad,
and they uh gutted the building,
and rebuilt it from the inside out,
so it's very secure.
um.. there was no real *apparent* damage immediately,
uh.. but the University did suffer quite a bit though,

Appendix B: Sample Pitchtracks

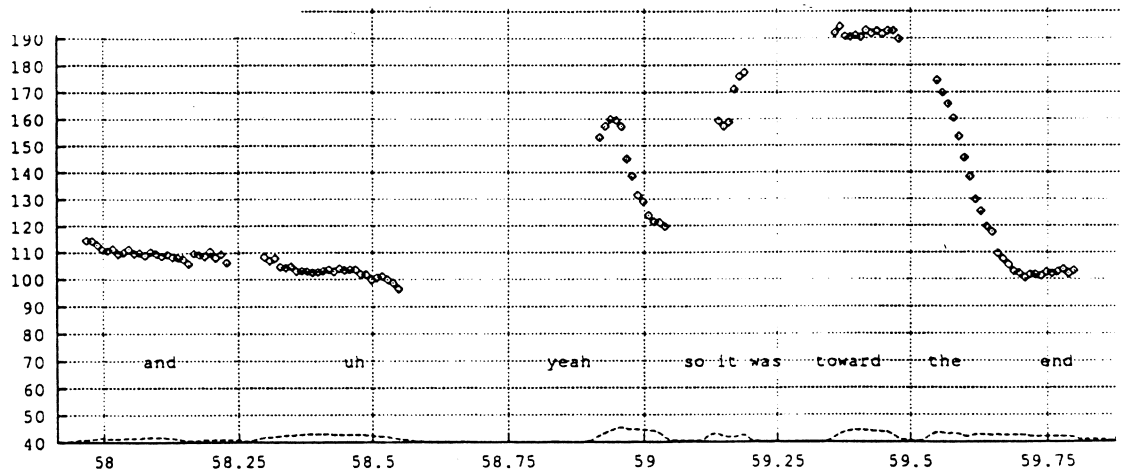


Figure 1 (line II.35)

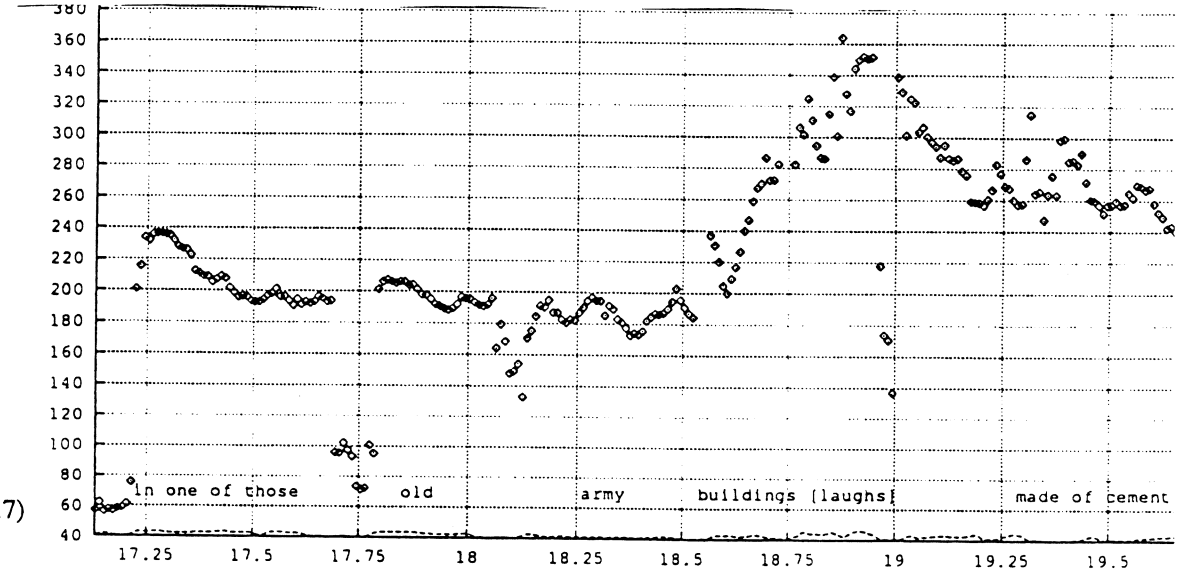


Figure 2 (line I.7)

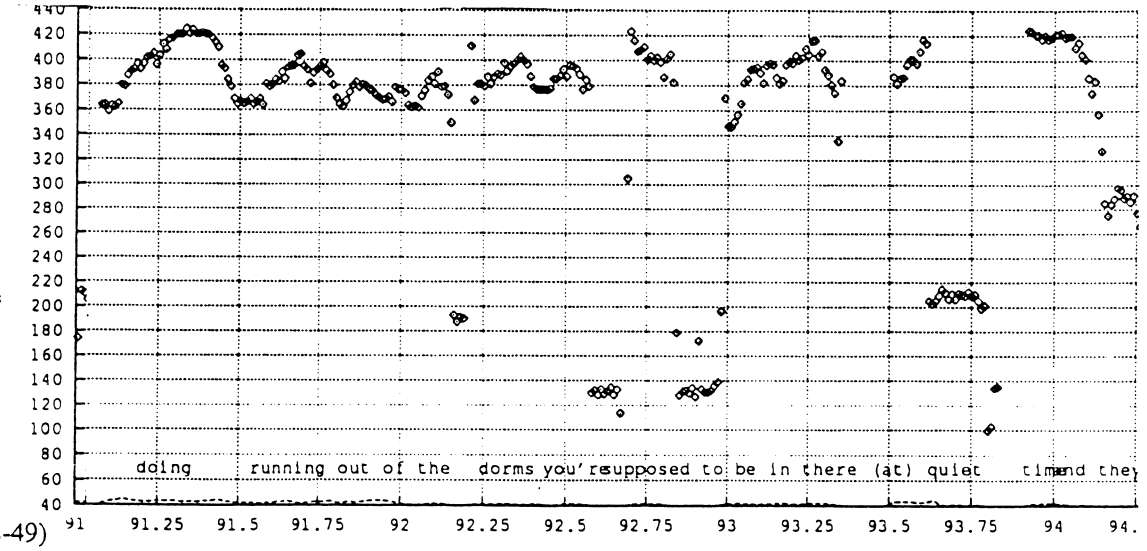


Figure 3 (lines I.48-49)

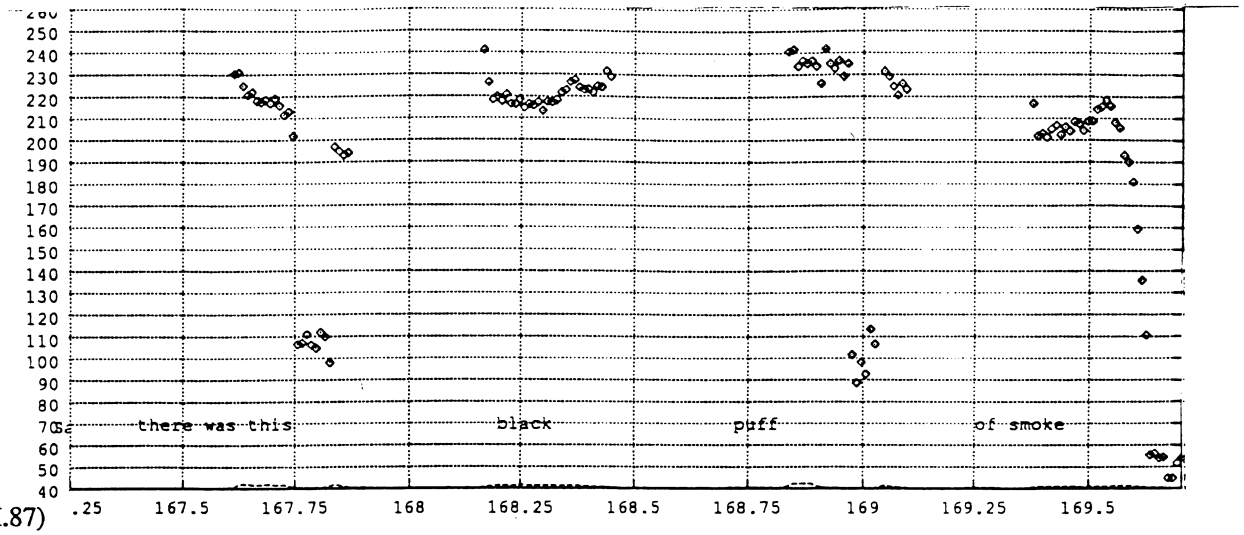


Figure 4 (line I.87)

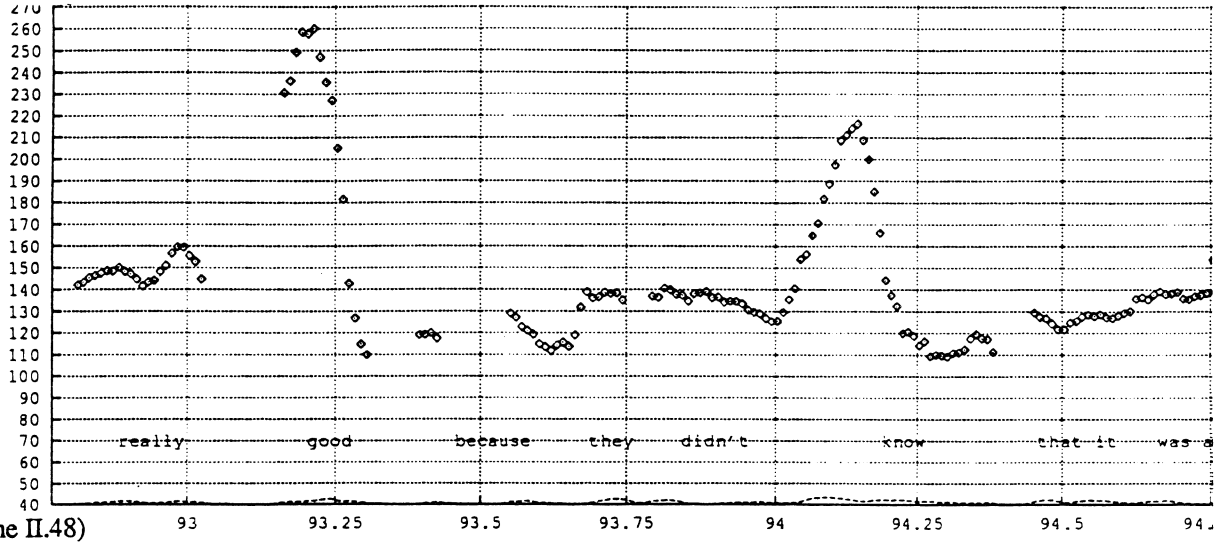


Figure 5 (line II.48)

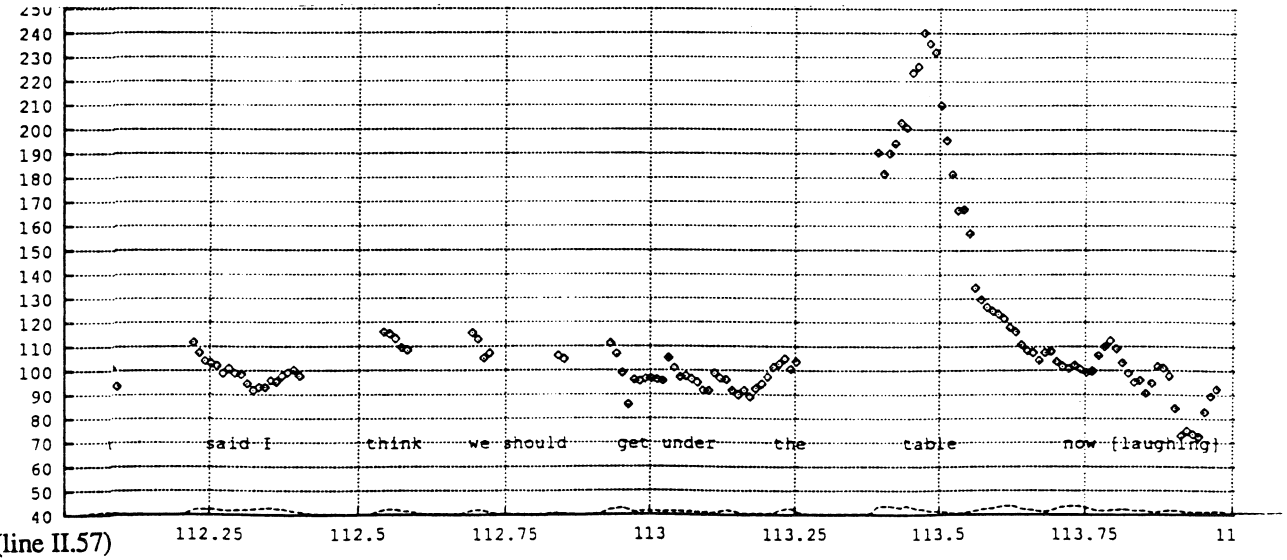


Figure 6 (line II.57)

THE LIMITS OF DOWNSTEP IN ÁGBỌ̀ SENTENCE-PROSODY*

Victor Manfredi
Boston University African Studies Center
270 Bay State RD
Boston MA 02215

ABSTRACT

A recorded corpus[1] of some 80 nonspontaneous Ágbọ̀ [2]examples shows systematic resetting of downstepped pitch within the minimal sentence. As this phenomenon is not independent of a preceding downstep, and can never cumulate upward, it is precisely not 'upstep' (pace Meir *et al.* 1975; Snider 1990) but rather *antidownstep* or *downstep-reset*. Contra expectations of the reigning phonological model of downstep (e.g. Clements 1981), *downstep-reset* is limited neither to clausal boundaries (where trivially it does occur) nor to performance contexts of maintaining adequate pitch range. A first, impressionistic pass over the Ágbọ̀ corpus readily identifies two linguistic contexts for *downstep-reset*:

- After word final downstep before phrase boundary (tracks 2, 3, 13, 26, 28, 31, 33, 41, 48, 50, 52, 63, 70-72, 74, 79, 80). Most examples of this *edge effect* involve a PP or serial VP — neither type containing a pause.
- After a verb in which lexical H and L are neutralized (tracks 21, 22, 28, 32-35, 37, 39, 41, 43, 45-47, 68-70, 72, 76, 77). This *architone effect* regularly occurs, inter alia, before the negative/relative suffix *-ní*.

In a framework of tone-metrical licensing (Bamba 1992, Manfredi 1992), the two *downstep-reset* contexts share one property: a H tone in a weak position. The configurations which predict weak H are found in surface syntax. Weak H also accounts for *downstep-reset* in the Àbánkeléle dialect—previously claimed to have a so-called 'upstep' juncture—and in standard Ìgbo.

1. GARDEN-PATH TONEMARKING

The problem addressed in this paper was noticed nearly 40 years ago. Transcribing some sentences of ShiTswa (a Benue-Congo language of Mozambique) in 1953, Welmers noticed a failure of deterministic tonemarking. Having convincingly assigned ShiTswa to the 'terraced-level' type later codified by Stewart (1965), based on the cumulative pitch lowering which occurs automatically between successive H-tone domains, he was surprised to observe

> a clear contrast... after low, between a nonlow at the same level as the preceeding nonlow and a nonlow at a slightly lower level. (1973: 87)

Such a contrast creates a garden path for the application of a standard tone orthography comprising three rules:

- H- and L-bearing syllables are individually marked [´] and [`] respectively.
- Downdrift (Stewart's "automatic downstep") occurs between H-bearing syllables across L-bearing syllables.
- ("Nonautomatic") downstep between two adjacent H-bearing syllables is marked [!].

To demonstrate the breakdown of tonemarking, Welmers (1973: 91*f.*) cites the following paradigm:

1a.	Vámùwóná mùfánà. 3pl.see... child	'They see [the] child'
b.	Vámùwóná mùfánà wa mùbìkì. 3pl.see... child of chief	'They see [the] cook's child'
c.	Vámùwóná mùfánà wa !hósí. 3pl.see... child of chief	'They see [the] chief's child'

The imparsable syllable is *wa* 'of' in (1b) and (1c): no available tone diacritic fits that word's pitch. Consider the possibilities. *Wa* can't be marked L: it is pronounced higher than the flanking L-bearing syllables in (1b), and higher than the downstepped H in (1c). Neither can *wa* be marked H: it is pronounced on the same pitch as the middle syllable of *mùfánà*—rather than on a lower pitch which it would be expected to have as the bearer of a well-behaved H tone. Thus,

*Thanks to A. Akinlabí, M. Bamba, Ụ. Íhìọ́nụ́, Y. Lánìran, M. Liberman, A. Nwáchukwu, J. Ògbú, H. Tada.

[1]Text given in full below, with four pitch tracks. The examples—elicited to test tone classes of monosyllabic verb roots—are either gnomic, quasi-proverbial sentences with no marked focus; or mini-discourses with controlled focus structure. A hifi recording of the corpus, spoken by one person (not in real time) on one occasion, has been deposited in the phonetics lab, Williams Hall, University of Pennsylvania. Track numbers refer to the file labelled "/home/myl/db/agbo".

[2]Ágbọ̀ is the westernmost form of Ìgbo in the historical sense. Colonial/federal governments and their missionary/academic allies carved the periphery of the Ìgbo-speaking area into ethnic districts (e.g. "Ìká", "Ìzíì", "Ìkwẹ̀rẹ") on ideological grounds (kinship, kingship, confession, lexicostatistics). In reality, many of the claimed unique peripheral characteristics are actually found throughout the area; many others are just borrowings from non-Ìgbo-speaking neighbors; thus, neither sort of evidence proves anything about Ìgbo-internal relationships (cf. Ọ̀nwụejíọgwụ 1975).

Welmers is constrained to leave *wa* without a tonemark, stipulating that this absence means 'same pitch level as nearest previous H'. The unmarked *wa* is not toneless; it implicitly bears its underlying H as expected, but is preceded by a special juncture which negates the downdrift (automatic downstep) which would ordinarily occur at that point.

As the anomalous, antidownstep juncture occurs only in possessive phrases, all of which are formed with the "associative" morpheme *wa*, Welmers (1973) conceives a morphological solution: a "phonemic upstep" is assigned to *wa* itself, as a kind of prosodic prefix whose bizarre nature is excused by its unique distribution. Though the mechanics of his 1973 proposal are certainly *ad hoc*, the intuition that the antidownstep juncture is construction-specific is consistent with a prosodic government approach—offering at least the prospect of an explanation based on principles of tone-syntax interaction. To explore this possibility, it is first necessary to review some of the elementary relationships of phonological government which pervade the languages of this great, transcontinental family.

2. TONAL PROSODY AS GOVERNMENT

Bamba (1989, 1992) shows that OCP-based, nonlocal pitch effects like downstep, as well as local pitch effects like raising and spreading, reflect the constituency of metrical domains. Bamba's framework is *prosodic* because the domains in question interact with surface syntax in predictable ways. The basis of this interaction is the core licensing principle which, by hypothesis, is shared by phonology and syntax: the government relation.[3] The overall goal of this section is to show that *downstep-reset* is an example of prosodic licensing in this sense. The first step in the demonstration is to survey some simple cases in the relevant languages.

2.1 Tone and locality

As extended to Benue-Kwa[4] languages by Manfredi (1988/1992), prosodic licensing in Bamba's sense is implied by cross-linguistic, and language-internal, distributions of (local) spreading and raising with respect to downstep.

	local				nonlocal	
	spreading		raising		H ! H interval	
	H /⊥	L /_H	H /⊥	L /_H	partial	total[5]
Standard Yorùbá[6]	+	+	+			+
Ágbò	+				+	
Ọ̀nicha					+	
Ḿbàisén					+	(*Auslaut*)
Àbánkeléke[7]			+			+
Ɣɔmalá-Yamba[8]		+		+	+	+
Ɣekoyó[9]		+				+

Table 1. Distribution across Benue-Kwa of some local and nonlocal tone effects

The table shows *inter alia* that L-spreading and L-raising—both being local L tone effects—are in complementary distribution with partial downstep—which is a nonlocal effect, since it cumulates over the entire sentence. It is important to realize that this implication holds robustly even in Ɣɔmalá-Yamba, where only strong L tones spread or raise, and only weak L tones qualify as partial downstep triggers.

---

[3]If, on the other hand, "phonology is different" (Bromburger and Halle 1989), the licensing principles of metrical domains have nothing in common with those of phrasal syntax. As their pessimistic premise rules out prosodic results in advance, one should reject it provisionally and seek generalizations until they appear or until one tires of the search.

[4]Benue-Kwa, the largest branch of Niger-Congo, extends from central Côte d'Ivoire (or perhaps from eastern Liberia) to eastern and southern Africa. To date, no phonological (as opposed to lexical) evidence for an internal subgrouping of Benue-Kwa has been offered. A potential candidate for a syntactic isogloss is the movement of a main verb to the position of inflection ("V-to-I movement" cf. Emonds 1978); this occurs in Ìgbo and eastwards, and in Ànyí (or perhaps Akan) and westwards, but not in a central zone extending from Gbè to Yorùbá and Ẹ̀dó (cf. Déchaine 1992).

[5]Total downstep lowers an H-tone to the pitch level of a non-H-tone in the same context; partial downstep doesn't.

[6]In Yorùbá, (nonautomatic) downstep occurs only after an elided L tone; it is a total downstep as defined in the preceding footnote, since a downstepped H is lowered at least to the level of M. According to Láníran (1992: 250), Yorùbá M is not downstepped, but the preceding H is raised; Yala-Ikom's 'downstepped M' (Armstrong 1975) may be similar.

[7]A.k.a. "Izi" or "Ìzíì", an ethnic label promoted in literacy materials, starting shortly before the Nigerian Civil War, by the Énugwú branch office of the Summer Institute of Linguistics (cf. Meir et al. 1975).

[8]A.k.a. "Dschang Bamileke"—studied (and, if I am not mistaken, spoken) by Tadadjeu (1974).

[9]A.k.a. "Kikuyu"—studied by Clements and Ford (1978).

The other complementarity in the table is between total and partial downstep. For nonfinal contexts, one can predict the occurrence of total downstep from L-spreading. In absolute final position (*Auslaut*), however, total downstep also occurs in Ḿbàisén (among several other southern dialects) which lacks L-spread. The multiple sources of total downstep suggest that it is a default which obtains wherever H tone is governed.

The distribution in Table 1 can be studied in terms of tone-metrical interaction. Consider the principles in (2).

2. *principles*[10] A metrical governor is stronger than its governee (H>L>M).[11]
[s] immediately dominates a metrical governor.
[w] is strictly adjacent to a metrical governor.
Tonal government iff [s].

The idea in (2), adopted from Bamba (1989/1992), is that two different kinds of licensing relation—respectively tonal government and metrical government—are separately responsible for the local and nonlocal phenomena referred to in (2). The generalization of complementarity follows from the fourth assumption, namely that tonal government (e.g. spreading, raising) is possible only if the tonal governor occupies in a strong metrical position. Since H is the metrical governor in the partial downstep relation, partial downstep excludes L from a strong position, hence L cannot be a tonal governor.

To accommodate the variation observed in Table 1, this framework must be supplemented by the parameters in (3).

3. *parameters* (i) The set of tonal governors is {H}, {L}, {H, L}.
(ii) Tonal government is expressed by {spread} {raise} {both} {neither}

The resort to parameters is, in general, problematic, unless (as suggested by Borer 1984, Fukui 1986) they can be reduced to learnable inventories of closed-class (i.e. 'functional') items. Minimally, one would hope that only tonal government needs to be parametrized, at least for the closely languages in question. The required parameter settings are listed in (4).

4. *settings*

	(i)	(ii)
Yorùbá	H, L	*both*[12]
Ágbò	H	*spread*
Ọ̀nịcha/Ḿbàisén	H	*neither*
Àbánkeléke	H	*raise*
Ɣɔmalá-Yamba	*some* L	*both*
Ɣekoyó	L	*spread*

For the present, I will set aside issues of parametric learnability or arbitrariness, and proceed to examine cases where syntactic government seems to affect the tonal and metrical relationships just outlined.

2.2 Prosodic government

The smallest assumption sufficient to explain downstep-reset is the failure of a licensing condition for downstep. Bamba defines downstep as a nonlocal government relation between tones, mediated by metrical constituency. If tonal government requires syntactic government, then downstep can't follow a tone which is not in a governing position.

5. *licensing* Locally, an element is ungoverned iff governing.
Unlicensed elements incorporate under the local licensed node, e.g.:
(a) Domain-initial L incorporates under following [s].
(b) Domain-final H incorporates under preceding [w].

(5a) accounts for initial L-raising (also in Ẹ̀dó, cf. Elugbe 1977). (5b) follows from the definitions in (2), and directly advances the goal of this paper to account for the possibility and distribution of weak H tones.

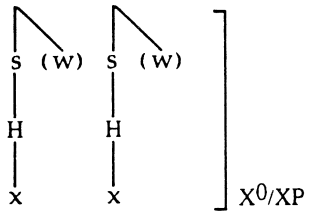
The consequence of (5b) is illustrated in (6a). The filled weak node is unlicensed: it doesn't govern anything because it is final, and it isn't governed since it is not weaker than the preceding strong node. Incorporation of stray H yields (6b).

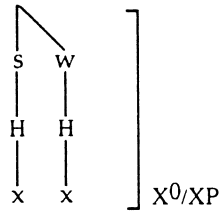
---

[10]Most of these principles simply recap the definitions of Liberman and Prince (1977).

[11]This hierarchy couldn't be valid in a true 'upstep' language, if any exists. No such language has yet been documented.

[12]Láníran finds L-raising only concomitant with H-raising; her algorithm (1992: 237*f.*) involves a relation called "upstep", which actually applies right-to-left (n.b. backwards in time) across tonal feet. That this is indeed an example of raising is shown by her observation that the first H's extra height factor does not affect the level of an initial L.

6a. 

b. 

Prosodic licensing has numerous empirical consequences in Ìgbo. For example, consider the well-known restriction of lexical downstep to the final syllable, cf. the Ọ̀nịcha forms in (7):[13]

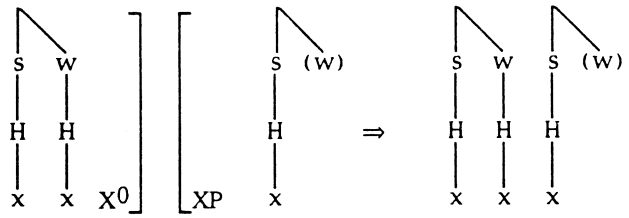
7. átụlụ́ 'sheep' ífẹlẹ́ 'shame' *v́cv́cv
ńkịtá 'dog' ọ́bẹlẹ́ 'small creature'

If these forms are composed of three H-bearing morphemes, the third and final morpheme is evidently weak, hence its H tone is exempt from the OCP. As is well known and ill understood, however, the final downstep of nouns drops phrase-internally:[14]

8. | | | | | | |
|---|---|---|---|---|---|
| ọ́nụ́ | 'mouth' | ụ́zọ̀ | 'path' | ọ́nụ ụzọ̀ | 'door(way)' | *ọ́nụ́ ụzọ̀ |
| ágụ | 'leopard' | áta | 'grassland' | ágụ atá | 'savanna leopard' | *ágụ atá |
| ọ́bẹlẹ́ | 'small creature' | nwá | 'child' | ọ́bẹlẹ nwá | 'dear little child' | *ọ́bẹlẹ́ nwá |
| ńkịtá | 'dog' | ụ́nụ̀ | '2pl' | ńkịta ụnụ̀ | 'your dog' | *ńkịtá ụnụ̀ |

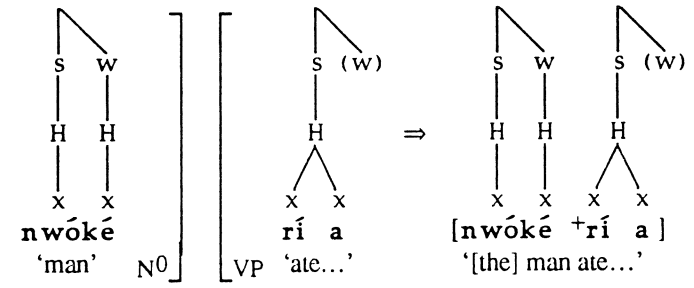
Whatever special licensing permits a word-final H to be weak in citation forms such as those in (7), (8) shows that this licensing is not available phrase-internally.

The Ágbọ̀ corpus, however, shows that a weak H is conserved in certain other contexts, which I have labeled architones. If (6b) is a negative verb plus its pronominal prefix, the corpus shows that in a larger verb phrase, the word-final weak H is equivalent to a weak L (the total downstep effect), and the initial H of the following word has higher pitch (the *downstep reset* effect).

9. 

What needs explaining in this framework, therefore, is the contextual difference between *downstep reset* in Ágbọ̀ and its absence (with corresponding loss of the word-internal downstep) in Ọ̀nịcha.

Some Àbánkeléke examples of (9) are given in (10) and (11).

10. 

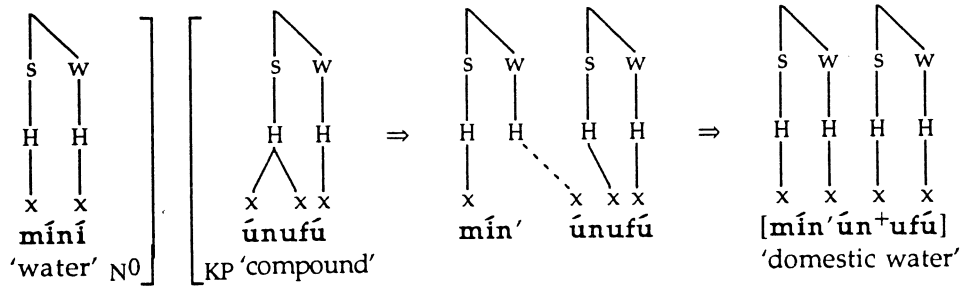
(The tone cliticization in (11) is driven by the elision of the last timing unit of *míní*.)

---

[13]The few exceptions in (i) are most likely exempted by internal structure.
(i) *ọ́ghẹ́(lẹ)* 'opening' (Ọ̀nịcha) *ókóro* 'young man' (Àbánkeléke), cf. *óke* 'male'

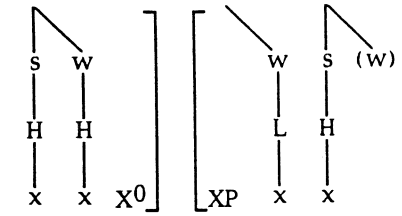
[14]One exception may be exempted by internal structure, cf. *dí* 'master':
(i) *ágadí* 'elderliness' (Ọ̀nịcha) *ágadí nwaànyị* 'old woman' **ágadi nwaànyị*

11.



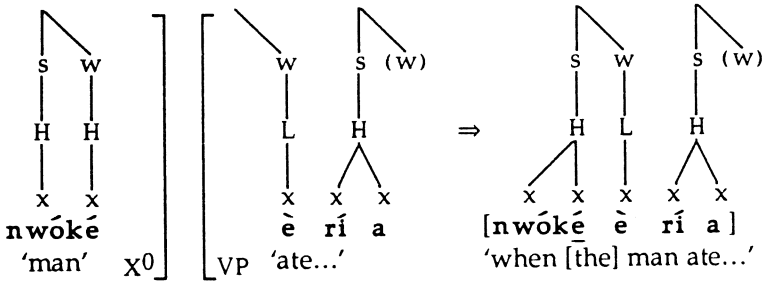
If the following phrase begins with L, another difference emerges, cf. (12).

12.



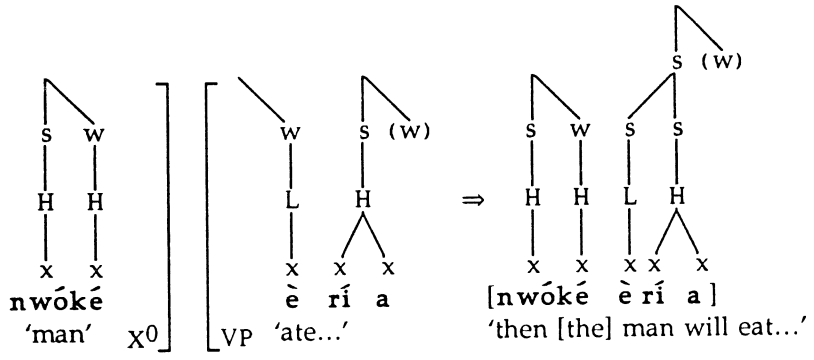
By stipulation in (4), tonal government takes the form of H spreading onto following L in Ágbọ̀, and H raising before L in Àbánkeléke. But by definition in (2), tonal government entails a strong position, so we might not expect a tonal government effect in either dialect. H spread doesn't occur in relevant Ágbọ̀ contexts, e.g. (24b), but H raising (notated by underlining) is reported by Meir *et al.* in corresponding Àbánkeléke examples, forcing a derivation like (13) which violates structure preservation.

13.



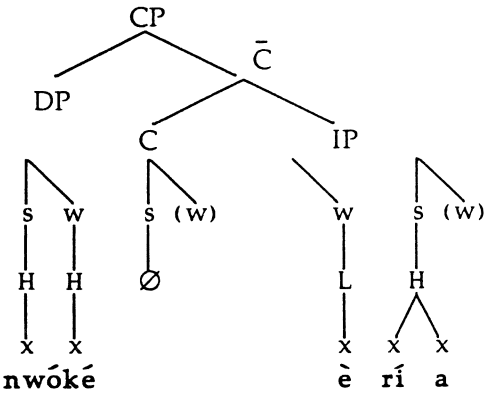
Fortunately, an alternative analysis is available; indeed it is required by the grammar. Meir *et al.* report an example minimally contrasting with (13):

14.



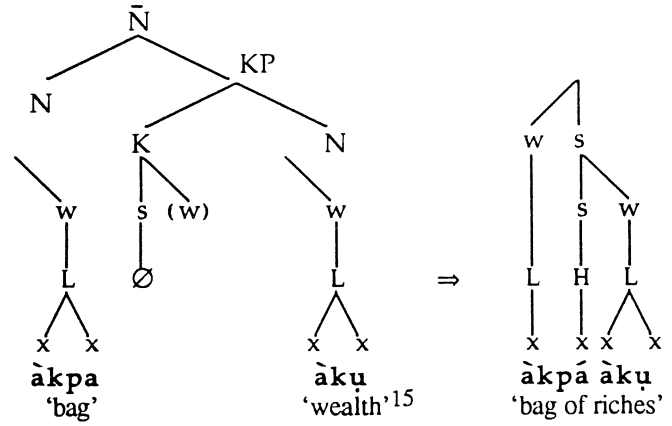
Within a principle-based framework, (13) and (14) cannot have the same syntax. Minimally, the conditional clause in (13) must include an additional head, plausibly a determiner, for compositional semantics. Independently, from the so-called associative construction, it is clear that the null Comp in Ìgbo relative clauses is spelled out on the surface with a H tone (see Excursus). It is unnecesssary to stipulate this, so long as the null Comp is metrically strong. This gives the conditional the s-structure in (15):

15.



How does (15) satisfy prosodic well-formedness? Examples of the genitive construction like (16) been argued to exemplify the principle in (17), cf. Manfredi (1992: 159).

16.



17. *prosodic cliticization* — An unassociated element acquires as its association domain the adjacent timing unit of its governing category.

In (15), cliticization of the null Comp creates the context for the observed raising. If this goes through, then tonal government in Àbánkeléke is structure-preserving.

A final question is why downstep reset occurs in Ágbọ̀ before the negative morpheme *ní*, which bears H tone, but not for example before the toneless *-ghị* of Standard Ìgbo (to which it is cognate). *Ní* is either a suffix or a left-branching phrasal head. We might suppose that *ní* as a phrasal head with inherent H is metrically strong. Then after a downstepped verb it will have the exactly the downstep reset configuration in (9). A related effect is seen in the Excursus, where a lexically unmotivated H tone appears in Ìgbo relatives as the content of null, strong Comp and Kase nodes.

3. SUMMARY

The above, preliminary analysis of prosodic licensing in Benue-Kwa languages takes off from the concrete and learnable disjunction between local and nonlocal tone effects, to posit quasi-syntactic relationships of constituency and government among tone elements, in the tradition pioneered by Bamba for Mandekan languages. Because government also forms an indispensable part of syntactic licensing, such an analysis offers the hope of explaining a wide range of phenomena which have heretofore inspired only bizarre diacritics of 'upstep' juncture. Equally importantly, it brings a rich array of phonological evidence, especially small parametric differences among closely-related languages, to bear on issues of syntactic representation.

In view of these results, Welmers' tonemarking puzzle (with which the paper began) counts as a monument to the keen linguistic intuition of that eccentric missionary, but also to the complacency of Africanist phonologists and syntacticians who have managed to preserve their respective specializations in pristine, obtuse segregation for too many decades.

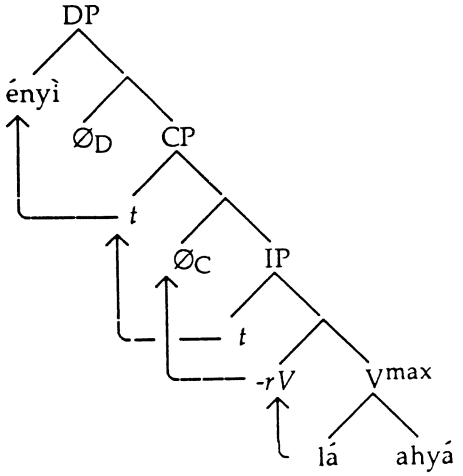
---

[15] *Àkụ* is, specifically, inert or non-reproducing wealth, as opposed to *ụ̀bá* which includes seed stocks and livestock.

EXCURSUS: PROSODIC MINIMALITY IN ÌGBO

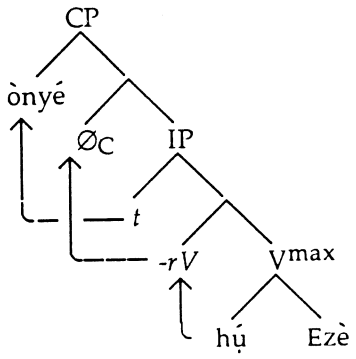
In Standard Ìgbo, an otherwise empty functional head is nevertheless strong in order to govern the head of an embedded constituent.[16]

18a.



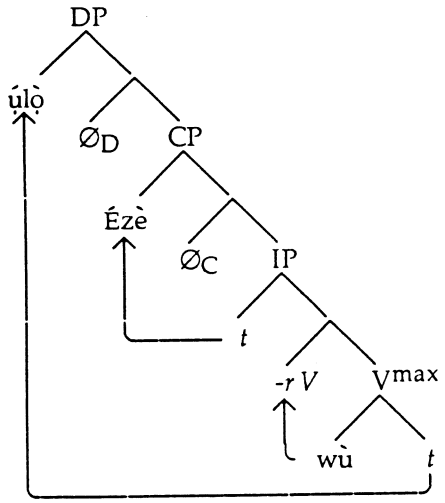
[ényí lára ahyá]
'the friend that left the market'

b.



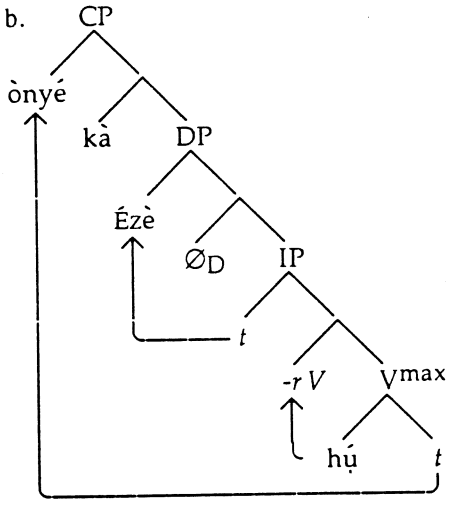
[ònyé hụ́rụ Ezè]
'Who saw Ézè?'

19a.



[úlọ́ Ezé wùru] 'the house Ézè built'

b.



[ònyé kà Ézé hùrụ] 'Who did Ézè see?'

EXCURSUS II: YORÙBÁ

Both L and H are necessarily strong in a surface three-tone system. That H also raises before L (Lániran 1992: 240), sentence-initial L does not downstep the following H (1992: 219), and spreading cannot cross M (1992: 199*fn.*), all follow from the presence of LH feet (1992: 251). Lániran (1992: 270) refutes Pierrehumbert and Beckman's (1988) claim—repeated e.g. by McCarthy (1988)—that declination is not computed over phonological tones.

EXCURSUS III: AGAINST REGISTER TONES

The register tone framework (Snider 1990) has no account for prosodic domains. Contour tones are overgenerated, unless markedness between 'modal' and 'register' tones is invoked to exclude possible but unattested contours. A "left-to-right implementation rule" (like Schachter and Fromkin's numerical algorithm) is also needed. The (non-arboreal) register formalism does not represent cumulation explicitly. The lack of symmetry between upstep and downstep is accidental.

---

[16]In Ágbọ̀, the empty head of a relative clause is spelled out with the copula *hụ́n*.

CORPUS

Speaker
Julius Ògbú
Ìdumu Úku, Ágbọ̀
June, 1977

Tone orthography
[´, `] = surface tones;
no mark = same as preceding tone;
[´] after [´] = downstep;
[+] = antidownstep

Track no.

1. Ǹ jné afya. Ó wí m ogné kịrị.
'I went to market; it took me a brief time'
2. Ń jné áfyá + ọ́nọbé tanì. [a copy of pitch track 2 follows below]
'I'll go to market after a little while today'
3. Nį́ ḿ jne afyá + éki ílẹ. [a copy of pitch track 3 follows below]
'Let me go to market tomorrow'
4. Ányụ àtú nkọ́, ì kebe gí ẹ́ be nknú.
'An axe is usually sharp before you use it to cut wood'
5. Àŋánį́ ọ nọ̀? Ọ̀ tụ́ nkọ.
'How is it?' 'It's sharp'
6. Ọ̀pya atú átụ, ì kebe gí é betúfú ùknuésù.
'A machete is usually sharp before you use it to cut open [a bundle of] yam pegs'
7. Àŋánį́ ọ nọ̀? Ọ̀ tụ atú.
'How is it?' 'It's sharp'
8. Ǹgbadna enwóke ákọ, ò kebe náhi ohúkpagha.
'An antelope is usually very clever, before it can escape a hunter'
9. Àŋánį́ ọ dnọ̀ náhi? Ò nwo akọ́.
'How did it manage to escape?' 'It's clever'

[transcription/translation of tracks 10-12 is missing]

13. Ékú ugbó wẹ gí eŋeré + kwá àkọ́ úkọ.
'A farm coat sewn with hide itches'
14. Àŋánį́ ọ mé i? Á á kọ m úkọ.
'How does it affect you?' 'It doesn't itch me'
15. Kí ọ mé ẹ́? Ọ̀ kọ́ á ụkọ.
'What does it do to him?' 'It itches him'
16. Kí i wẹtnafúnị a? Ọ̀ kọ akọ́.
'Why did you take it off?' 'It itches'
17. Ẹ́gẹdí aàja ánụ àja nį́ ọ márnị ọ̀sụ́ọ ọ̀bẹlẹzẹ́ẹ̀.
'An elder dices up meat so that s/he can know the sweet taste of "ọ̀bẹlẹzẹ́ẹ̀"'
18. Àŋánį́ ọ dnọ̀ kwádeme ẹ́? Ọ́ já anú; ò méyị ofigmò.
'How did s/he manage to prepare it? S/he diced meat; s/he added palm oil'
19. Àŋánį́ ọ kwadèmé ẹ? Ọ̀ já anụ àja.
'How does s/he prepare it?' 'S/he dices up meat'
20. Ńmúndù abú ẹbù ógné ilẹ́ ifnọ́ gì etí.
'Small children sing whenever the moon shines'
21. Kí wẹ́ me ẹ̀ wẹ́ gìlẹ + nį́ rahni? [a copy of pitch track 21 follows below]
'What did they do that they did not sleep?'
22. Ábú wẹ̀ ẹ́bù, étnè + nį́ wẹ̀ égú. [a copy of pitch track 22 follows below]
'They sang, they didn't dance'
23. Ógù ọ́mụmụ nwa ènyí nà éré.
'The birth medicine we received was effective'
24. Àŋánį́ ọ rnụnį́ i? Òre ere.
'How then did it work for you?' 'It was effective'
25. [incomplete transcription] Òré ère.
[...] 'It will be effective'
26. Òríri Nni Ugbó + ápú ọ̀-hú-mma.
'The Feast of Farm Food turned out well'
27. Ó pú kẹ̀ wẹ́ dnọ̀ kúu? Ọ̀ pụ apụ.
'Did it turn out as they said?' 'It turned out [well]'

28. Ǹké í + hnú lála + nîì? Ọ́ pú àpụ.
'Yours which is coming up? 'It will turn out [well]'

29. Mírni ezúe òsuó òhú. (possibly: Mírní + ézúe...)
'Rain fell [in] one area'

30. Nị́ mirní èzúe ugbó ò rúe m̀gbé ẹ̀nyasị.
'Rain must fall on the farm by evening'

31. Mírní + nị́ o zué ebe ndị ọhụ̀ụ̀.
'Rain will fall someplace'

32. Ányụ atnú + ní nkọ́.
'The axe isn't sharp'

33. Éyị̀lẹ́ m+ ányụ tnú lẹ + ní nkọ́!
'Don't give me an axe that's not sharp!'

34. Ọ̀pya átnú + ní àtnúụ̀.
'The machete isn't sharp'

35. Ánị̀lẹ m gí ọ̀pya átnú lẹ + ní àtnúụ̀!
'Don't have me use a machete that's not sharp!'

36. M̀gbadna áánwo ákọ.
'Antelopes aren't clever'

37. Ń sèka hú ùté m̀gbadná nwò lẹ + ní ákọ.
'I can see the track of an antelope that's not clever'

38. Èbulúku aákọ akọ́.
'[The ritual coat of an Ólokún priest] doesn't itch'

39. Nị́ ẹ́ yịmẹ ẹkwà kọ lẹ + ní ụkọ́. (speaker hesitates)
'Let him put on a cloth that doesn't itch'

40. Ńdị kikenì áája anú ǹkẹ́ ọ̀belezéè.
'People nowadays don't dice meat for "ọ̀bẹlẹzẹ́ẹ̀"'

41. Ánị̀lẹ onye ghàlẹ+ ní àja ánú + lẹ́ ḿ! (strong effect)
'Don't let someone who omits dicing meat host me!'

42. Ńmú ndù̀ áábụ ẹbù̀ íme isi àbalị.
'Children don't sing [on] moonless nights'

43. Ńdị ghàlẹ+ ní àbú ẹbù̀ ásekà tné egú.
'Those who omit singing cannot dance'

44. Ógù̀ áàre ere.
'[The] medicine is totally ineffective'

45. Á nì̀ lẹ m + gí ọgù̀ ẹlẹ+ ní ère ere.
'I won't use medicine that is totally ineffective'

46. Òríri apú + ní àpúụ̀.
'The feast flopped dismally' [did not turn out at all]

47. Hnú pù̀ lẹ + ní àpúụ̀ jọkọ anwozị́.
'What flops is going to have another [chance]'

48. Élẹ́ + ógné wẹ gì̀ gú gí + hnú aka ahnù̀ kẹ̀ wẹ́ gì̀ gú + ahyuá nì̀.
'It is not when they dug yams last year that they're digging yams this year'

49. Ógné wẹ gì̀ gú gí + wnú ogné mirní gì̀ lúa gu.
'The time they harvest yam is the time when rain has finished tapering off'

50. Ógné wẹ égi + gú gí wnù̀ ógné ọ-wnù̀-lẹ gha ekí + jnẹ́mẹ.
'The time they will harvest yam is any time after tomorrow and thereafter'

51. Ébe o wu ụzọ̀ chọ́ ewù wnụ epeté èpete.
'Where he stood seeking shade is muddy'

52. Ù̀bé o wu ụzọ̀ rú elú + ákpági.
'The ladder he stood upright broke'

53. Ébe o wu ụzọ̀ ché nmú a wnù̀ ahamáhà ụwáyà.
'Where he stood waiting for his children is in the middle of the road'

54. Ébe i ewu ụzọ̀ chéri wẹ́ wnù̀ ébe ụzọ̀ nọ́hìmé.
'Where you will stand waiting for them is where the path makes a bend'

55. Émụ̀ aknú ìhian aknụ́.
'Sickness troubles people'

56. Ọ̀banije esú ìhian esú.
'Sweat affects people greatly'

57. Ọ̀banije éèsú ìhian esúù.
'Sweat doesn't affect people at all'

58. Èzizá nkụ kà alị azáà. (why not: Èzizá nkú...)
'A broom of mature palm [branches] is best for sweeping the ground'

59. Èzizá ọ̀kịtị áàka alị azáà kárị èzizá nkú.
'A broom of baby palm [branches] doesn't sweep better than one of mature palm'

60. Wẹ́ amarị nwa èmé nwá.
'They know [how] the child will make itself'

61. Wẹ́ ámarị nwa èmé nwá.
'They don't know [how] the child will make itself'

62. Nwátá márị ihie èmé nwá.
'A child that knows something will mature'

63. Nwátá àmá ihié + á èmé nwá.
'A child that doesn't know something won't mature'

64. Ónye ehyù ẹkwá ọ̀hụhụ amárị ọnụ a.
'Someone who shops for hen's eggs knows their price'

65. Ónye eéhyù ẹkwá ọ̀hụhụ ámárị ọnụ a.
'Someone who doesn't..., doesn't...'

66. Éru eèpú ugbo wnú ekurù.
'The mushroom that appears on the farm is "ékurù"'

67. Éru aàfọdú nkú wnụ ekurù.
'The mushroom that grows on palm trees is "ékurù"'

68. Éru aàfọdú ọfya, ónobè nị́ ènyị́ húe + nị́ ẹ, ò réhi.
'The mushroom that grows in the woods, soon after we don't pick it'

69. Ékurù áàfọdú nkú onobé, ómeni ènyị́ húe + nị́ ẹ, ò réhi mgbé ènyasị.
'The "ékurù" that will grow on palm trees soon, if we don't pick it, it rots by evening'

70. Éru eépu ọfyá + ónobè; nị́ ènyị́ húe + nị́ ẹ, ò réhi.
'A [type of] mushroom will come out in the woods in a little while, if we don't pick it, it rots by evening'

71. Éru eèfie ẹnyí ugbó + wnú ekurù.
'The mushroom that eludes us in the farm is "ékurù"'

72. Éru eéfie ẹ́nyi +ọfya ekí +wnú ugu éni. Nẹ́dị ẹ́nyi aghọ́sị +nị́ ẹ́nyi kẹ wẹ́ àchọ́ á.
'The mushroom that will elude us in tomorrow's woods "ugu éni"'
'Our father didn't show us how to look for it'

73. Mánya aàsúọ ìkpohó wnù ọgọrọ.
'The wine that women like is "ọgọrọ"'

74. Mírní ezúe + íme àbalị. Mánya aàsúọ + tanì wnú nkụ elú.
'Rain fell during the night. The wine that will be sweet today is "ńkụ elú"'

75. Ánụ mẹ éke ẹsị rọ.
'The meat I [usually] share out is horse'

76. Ánụ mẹ éke wnù ẹsị ma ọ wnụ éfni, ẹ́lẹ́ + hnú ká ntị̀.
'The meat I [usually] share is horse or cow, it is not that which is smaller'

77. Ánụ mẹ éke + rị́ ndù kíkẹnì.
'The meat I will share out is alive now'

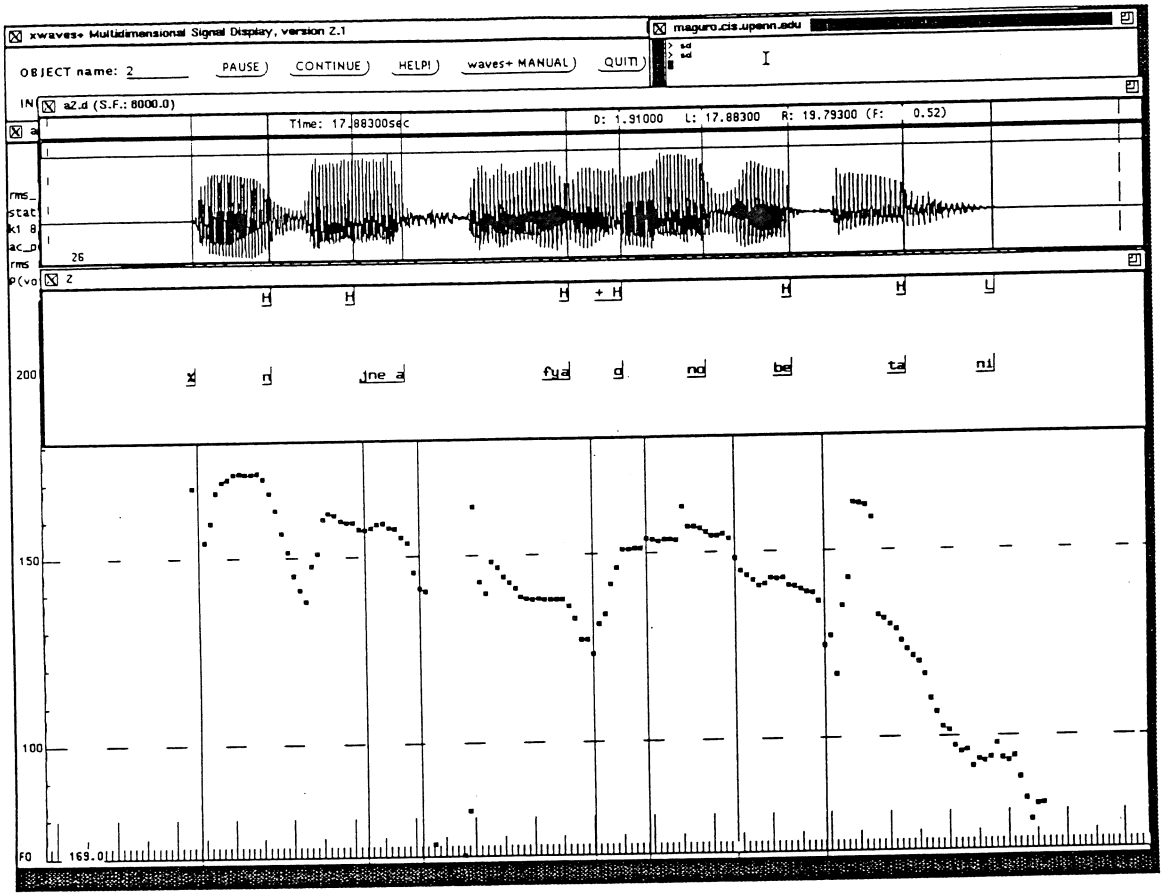
78. Mánya mẹ ára wnụ òzu nị́ nkụ elú, ẹ́lẹ́ ọ̀gọrọ.
'The wine I usually drink is "òzu" and "nkụ elụ", it is not "ọ̀gọrọ"'

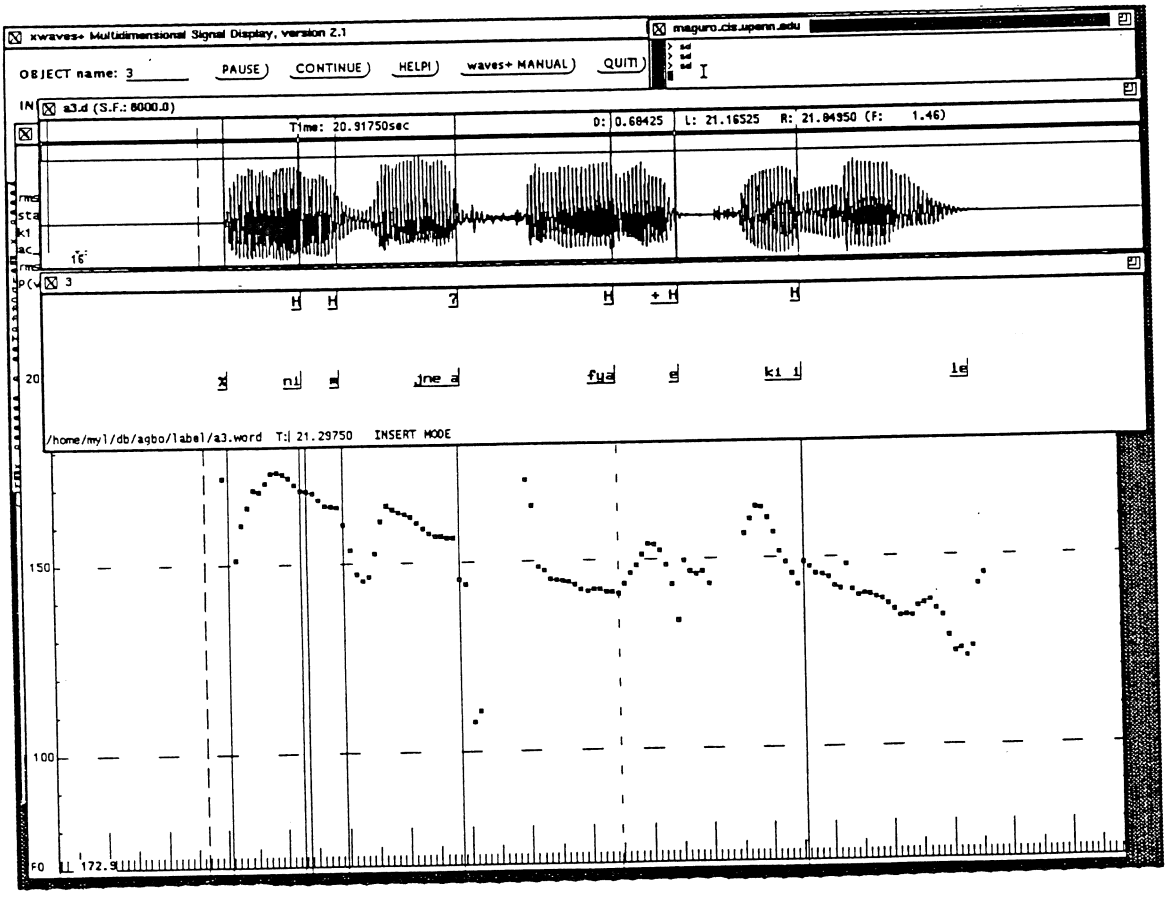
79. Mánya mẹ ára, è gí ḿ + dọ́nọ ọ súọ. (syntax unclear)
'The wine I will drink is claimed to be going to be sweet'

80. Ọ̀gwá o zùzú ẹnyi + nị́ ìyá, èrú ụkà a rị́ká.
'The meeting that includes us and her/him, it usually comes to a big argument'

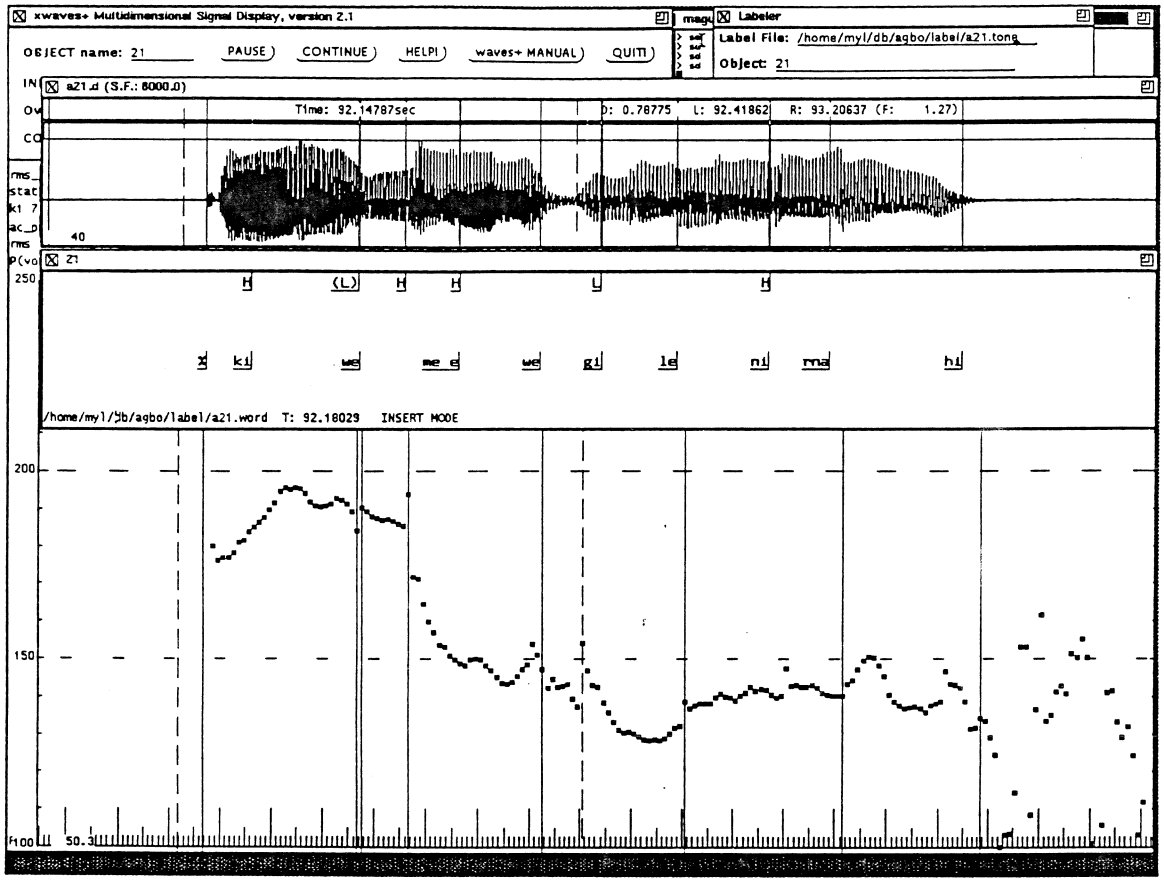
81. Ọ̀gwá oó zuzu ẹnyi, yá ebufùlẹ á.
'The meeting that will include us, let her/him not cancel it'

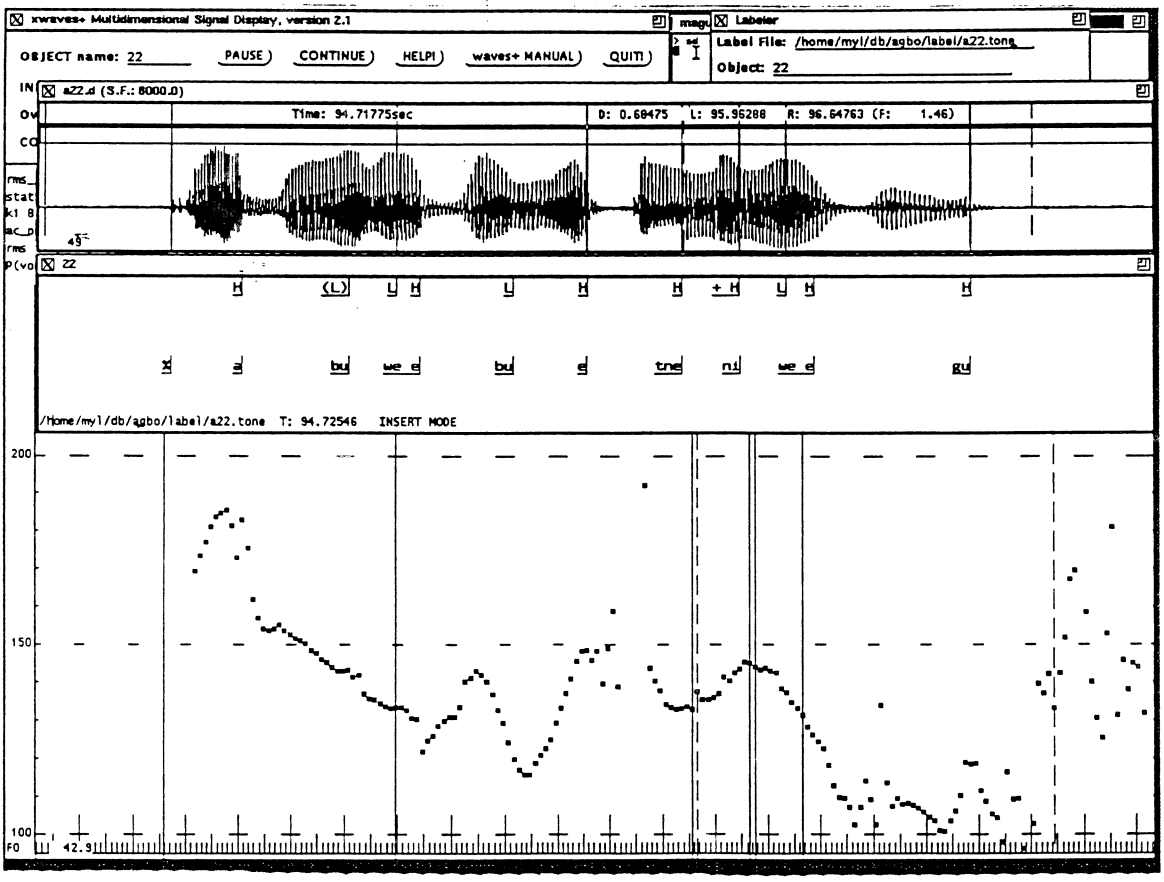
PITCH TRACKS 2-3





PITCH TRACKS 21-22





REFERENCES

Akinlabí, A. M. 1985 *Tonal underspecification and Yorùbá tone*. University of Ìbàdàn dissertation.

Armstrong, R. G. 1968 Yala (Ikom): a 'terraced level' language with three tones, *Journal of West African Languages* 5: 49-58.

Bamba, M. 1990 On downstep in the tonal system of Ojene Jula, in Hutchison and Manfredi (eds.) 1990: 1-14.

_____ 1992 Sur la rélation entre ton et accent, Université du Québec à Montréal dissertation.

Bendor-Samuel, J. T. (ed.) 1989 *The Niger-Congo Languages*. Lanham, Maryland: American Universities Press for Wycliffe Bible Translators, Inc.

Bromburger, S. and M. Halle 1989 Why phonology is different, *Linguistic Inquiry* 20: 51-70.

Clements, G. N. 1981/83 The hierarchical representation of tone features, *Harvard Studies in Phonology* 2: 50-108, also in Dihoff (ed.) 1983: 145-76.

Clements, G. N. and K. C. Ford 1978 On the phonological status of downstep in Kikuyu, in Goyvaerts (ed.) 1981: 309-57.

Collins, C. and V. Manfredi (eds.) 1992 *Proceedings of the Kwa Comparative Syntax Workshop. M. I. T. Working Papers in Linguistics* 17.

Déchaine, R.-M. 1992 Inflection in Ìgbo and Yorùbá, in C. Collins and V. Manfredi (eds.) 1992: 95-119.

Dihoff, I. (ed.) 1983 *Current Approaches to African Linguistics* 1. Dordrecht: Foris.

Elugbe, B. O. 1977 Some implications of low tone raising in southwestern Ẹ̀dó, *Studies in African Linguistics* Supplement 7: 53-62.

Emonds, J. E. 1978 The verbal complex V'-V in French, *Linguistic Inquiry* 9: 151-75.

Goyvaerts, D. L. (ed.) 1981 *Phonology in the 1980's*. Ghent: Story-Scientia.

Huang, C. T. J. 1980 The metrical structure of terraced-level tones, *NELS* 10: 257-70.

Hulst, H. van der and K. Snider (eds.) 1992 *The Representation of Tonal Register*. Berlin: Mouton de Gruyter.

Hutchison, J. and V. Manfredi (eds.) 1990 *Current Approaches to African Linguistics* 7. Dordrecht: Foris.

Kaye, J. D., J. Lowenstamm and J.-R. Vergnaud 1988 Constituent structure and government in phonology, *Linguistische Berichte*.

Láníran, Y. Ọ. 1988 Modeling intonation in a tone language: the Yorùbá example, paper presented at the workshop "Tone, Accent and Locality in Niger-Congo", University of Massachusetts, Amherst, 14 November.

_____ 1992 Intonation in tone languages: the phonetic implementation of tones in Yorùbá, Cornell University dissertation.

Liberman, M. and A. Prince 1977 On stress and linguistic rhythm, *Linguistic Inquiry* 8: 249-336.

Liberman, M. and V. Manfredi *in preparation* A formal and instrumental study of 'missing' downsteps in standard Ìgbo.

Mádùká, O. N. 1984 Ìgbo ideophones and the lexicon, *Journal of the Linguistic Association of Nigeria* 2: 23-29.

Manfredi, V. 1988/1992 Spreading and downstep: prosodic government in tone languages, paper presented at the workshop "Tone, accent & locality: towards a typology of Niger-Congo languages", Dept. of Linguistics, University of Massachusetts, Amherst, 14 November. Revised version in van der Hulst and Snider (eds.) 1992: 139-89.

McCarthy, J. J. 1988 discussant's commentary on Láníran 1988.

Meier, P., I. Meier and J. Bendor-Samuel 1975 *A Grammar of Ìzíì*. Norman, Okla.: Wycliffe Bible Translators, Inc.

Nwáchukwu, P. A. 1987 Questions, relative clauses and related phenomena in Ìgbo, *ms, Lexicon Project*, M. I. T. Center for Cognitive Science.

Ọ́nwụejíọgwụ̀ [Onwuejeogwu], M. A. 1975/1977 Some fundamental problems in the application of lexicostatistics in the study of African languages, *Paideuma* 21: 6-17 and *Ọ̀dụmá* 3.2: 29-36.

Pierrehumbert, J. B. and M. E. Beckman 1988 *Japanese Tone Structure*. Cambridge, Mass.: M. I. T. Press.

Snider, K. 1990 Tonal upstep in Krachi: evidence for a register tier, *Lg.* 66: 453-74.

Stewart, J. M. 1965 The typology of the Twi tone system, preprint from the *Bulletin of the Institute of African Studies* (Legon), 1: 1-27.

Tadadjeu, M. 1974 Floating tones, shifting rules and downstep in Dschang Bamiléké, *Studies in African Linguistics* Suppl. 5: 283-90.

Welmers, Wm. E. 1959 Tonemics, morphotonemics and tonal morphemes, *General Linguistics* 4: 1-9.

_____ 1973 *African Language Structures*. Berkeley: University of California Press.

Williamson, K. R. M. 1989 Niger-Congo overview, in Bendor-Samuel 1989: 3-45.

PROSODIC VARIATION ACROSS DISCOURSE TYPES

Cynthia A. McLemore

Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104
cam@arch.ling.upenn.edu

1. INTRODUCTION*

In this study, I compare the frequency and distribution of a small set of prosodic features in two different types of discourses, or speech activities. The goals of this investigation are to refine methodologies for transcribing and characterizing intonational regularities in natural speech, and to uncover the ways in which intonational forms are used for particular, situated ends.[1]

The data considered here are "natural" in the sense that they weren't elicited for the purpose of study; however, each dataset consists of speech performed in fairly constrained situations, rather than arising spontaneously in the course of conversation. Constrained data of this nature are useful in that they simplify the problem to some extent; the constraints themselves yield clues to prosodic patterning. Nevertheless, hypotheses about intonational function that are formulated on datasets of this kind are necessarily preliminary, if conversation is taken as the fundamental paradigm of language use (Fillmore 1981), which it undoubtedly should be. The data examined in this study include:

- A portion of a second-grade mathematics lesson ("Lesson")
- One exchange from a call-in radio talk show interview with a politician ("Interview")

Both speech activities are recurrent ones for the primary speakers, who occupy social roles with which these activities are associated: the teacher conducts daily lessons in the classroom, the politician gives interviews (more broadly: answers questions and espouses policies) on a regular basis.

Each dataset may be considered a token of a discourse type in a very broad sense: i.e. members of the English-speaking community, including the participants, have names for these speech activities, indicating that they are recognized and evaluated as distinct (see Silverstein 1979, Swales 1990:58). That is, a "lesson" is not an "interview", and neither discourse is a television commercial (Gumperz 1982:102-105), a sports commentary (Ferguson 1983), a meeting announcement (McLemore 1991a), a traditional narrative (Woodbury 1987), or the opening of a telephone exchange (Liberman & McLemore 1992).

While the two discourses examined here clearly belong to distinct genres, the distribution of prosodic features in them reflect more general characteristics that cross-cut genres. Biber (1988:170) distinguishes genre from text type: the latter represents groupings of texts based on their linguistic form, regardless of their genre. In order to arrive at such general principles, the discourse-internal correspondences between intonational form and function will be examined and related to previous findings, and the intonational features of the two discourses will be compared. The assumption motiving this approach is that, while the prosodic structure of a discourse may arise from the rational, more or less conscious, intentions of speakers, the meanings created by intonational choice ultimately can be understood only by an ordering of the facts of use.

2. INTONATIONAL DESCRIPTION

2.1 Segmentation

The segmentation of natural speech into discrete intonational phrases is far from straightforward (Du Bois et. al. 1991:100-114; McLemore 1991a:28-44).[2] In this study, segmentation has been based on sound structure without regard to syntactic, semantic, or pragmatic constituency (as much as possible); i.e. criteria for intonational juncture include pauses, pitch excursions and

* This research was supported by NSF STC grant number DIR89-20230, and OERI grant number R117610003-92.

[1] Two more long-term goals of this undertaking are to characterize sources of intonational variation in a principled way, and to shed light on the intonational phonology of English by examining recurrent patterns in phonetic data that correspond to apparent functions.

[2] Indeed, Liberman, McLemore & Woodbury (1991) argued that evidence for independent hierarchic prosodic units, such as the intonational phrase, generally is lacking; rather, such 'units' are motivated by local, gradient phonetic cues or phonological processes and other constituent structures (pragmatic, syntactic, semantic). See also Woodbury (1992).

salient changes in scaling values. This is an extremely narrow view of segmentation; however, without a more fully articulated account of syntactic, semantic, and pragmatic constituency in discourse, it seems wise to avoid attributing their effects to prosodic structure.[3]

All self-interruptions (preceding repairs, restarts, and so-called hesitation phenomena) were coded as junctural markers, since they disrupt the speech stream; ruling this type of juncture out *a priori* could only be motivated by considerations of function, which properly follow a formal characterization.

2.2 Tune

The tonal description used in this, and previous, research has its roots in early generative treatments (see Liberman & Pierrehumbert 1984), in that discrete target tones are used to describe rises (LH) and falls (HL, HM). However, the notational conventions used here are largely intended as a pre-theoretical discovery procedure, or null hypothesis, as in phonology generally, where systematic description of phonetic data is a prerequisite for theorizing. The tonal transcription differs from the revision of Pierrehumbert (1980) in Silverman et. al. (1992) primarily in that a minimal phonology is assumed here, consisting of one tone type with three categorical values and variable text-tune alignment (T, with the values H, L, or M, aligned with stress, T*, or not, T); i.e., no independent categories of phrase accents and boundary tones are postulated. For example, the different phonetic forms that would be described in those systems as H* L* L L% would be described here more specifically according to actual phonetic form: H* L*, with no implicit L L%; or H* L*– (where '–' indicates a simple temporal function corresponding to a sustained final L*). In addition, no tone is designated as the "nuclear accent," although for the most part the stress immediately preceding the juncture is the only one considered for the purposes of this study (the complete transcripts are fully notated for stress).

Sustained tones are notated with a following dash, T– (H–, L–, M–), indicating that the current pitch value is held relatively constant until the next (notated) tone, pause, or turn change. Tonal interpolation is otherwise a relatively direct path from one tone to the next (e.g., H L indicates a straight downward movement; L H indicates a straight rise from L to H).

In addition, a tone with unspecified value, T, has been used to notate contours set apart from surrounding speech by pauses or shifts in scaling, but which show no internal pitch change. Most such utterances are particles or 'discourse markers', e.g., *um*, *uh*, *and*, *well*, *so* (cf. Hockey 1991, 1992). While it is possible to specify a local tonal value based on the last value of the preceding phrase, this descriptive criteria often results in a counterintuitive specification. For example, if the preceding junctural tone is scaled very high, and the constant value on a following *and* is lower, it could be described as L, but might nevertheless sound quite high. Analysis in terms of scaling values rather than tonal category would be more useful at this stage of theorizing (cf. Shriberg 1992).

2.2.1 A Note on Mid

Mid is used as a descriptive category for reference to the set of values at endpoints of falls that sound non-Low. Phonetically, the criterion used for identifying junctural tones as Mid is primarily that the end value is scaled higher than a preceding L (usually in the same phrase, although in some cases the lower L was in the immediately preceding phrase). When a preceding L tone wasn't available (e.g., in turn-initial utterances), a tone was coded as Mid if it sounded Mid.[4] In the Lesson data, most of the analysis was performed auditorily rather than instrumentally, since many of the very final values for junctural tones were impossible to recover instrumentally.

It may be the case that the phonetic form of gradiently scaled Low tones conflate with that of target Mid tones. For the most part, the criterion used for identifying Mid excludes sequences (more than two) of non-low Low tones that are progressively, gradiently scaled (i.e. in a declining pitch range), since each Low in such a sequence would generally be *lower* than a preceding one, rather than higher. On the other hand, sequences of progressively declining Mid junctural tones in the tonal environment of LHM, where L is lower than M, would be identified as Mid.

In the Interview data, a comparison of L, M and H junctural tone values for the two individual speakers, "Caller" and "Mayor," shows that M values are consistently distinct from L and H junctural tone values when the immediately preceding H peak value is considered for each case (this H value was not used to code M):

[3] In segmenting phrases based on vowel length alone, Wightman et. al. 1991 categorize gradient junctural strength into five levels of phrasing. Although this description has since been incorporated into the intonational transcription in e.g. Silverman et. al. 1992, it has not been used here, since the segmentation criteria include vowel length and textual relations.

[4] In addition to Mid tones identified in terms of relative scaling values, forms that *sound* Mid, in fast speech at least, include falls followed by a slight rise whose value is less than a preceding H*, and falls to low in which the L is sustained (see Liberman & McLemore 1992 for examples). Several clear cases of the slight rises were notated as Mid. No cases of sustained L were notated as Mid, in order to allow investigation into the functional patterning of these different phonetic forms.

FIGURE 1:

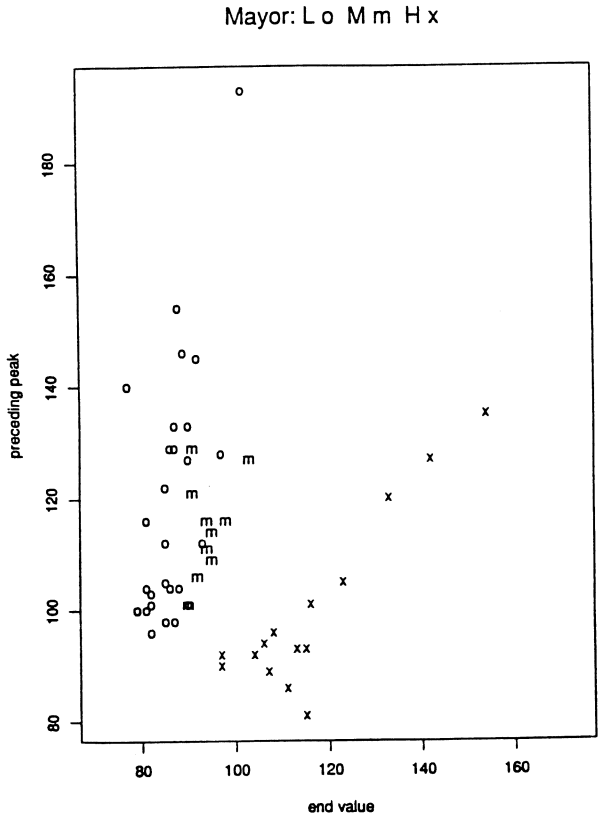
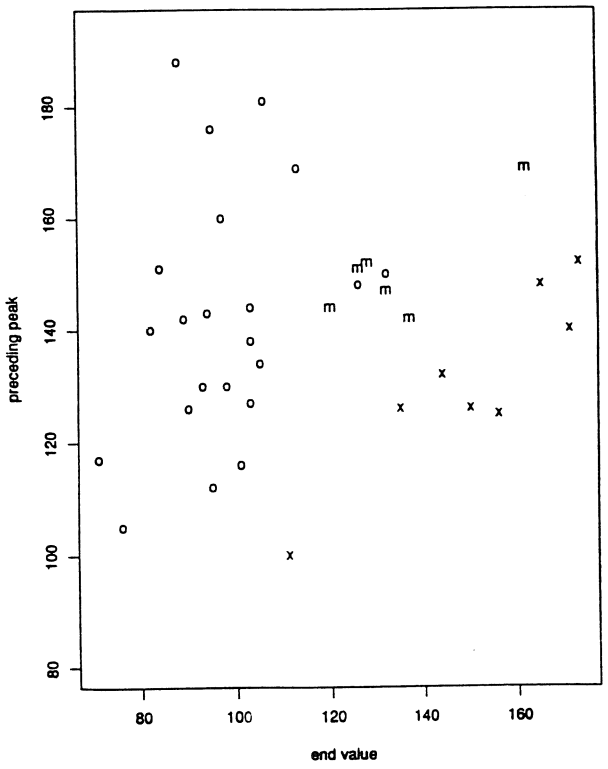


FIGURE 2:



2.3 Text-Tune Alignment

Tones that align with a stressed syllable are marked with an asterisk, T*, (H*, L*, M*), and referred to as *accents*. One consequence of the close phonetic description applied here is that junctures can occur immediately following accents. In fact, there are a number of phrases in the Lesson data that are overall rising or overall falling contours, with clear accents at each end point. In the data analysis that follows, I have made a distinction between junctural tones (rises, falls, and levels; i.e. pitch movements that help create a juncture) and what will be called "final accents," stress-aligned tones that precede a juncture created by a pause or e.g., scaling shift.

Thus, the contour referred to as the "vocative chant," which Liberman (1975) characterized as (L) H* M, has the following possible variants (ignoring the optional L):

H*M H*M– HM* HM*–

The analysis of intonational features in this study is limited to the two-tone sequence preceding a juncture, including also:

H*L H*L– HL* HL*–

L*H L*H– LH* LH*–

Sample pitchtracks, referred to in the analysis below, are shown in the Appendix.

3. INTONATIONAL FUNCTIONS

3.1 Rises and Falls

In McLemore (1991a, b) I argued that rising, falling, and level junctural tones have the fundamentally iconic[5] general functions of connecting, segmenting, and continuing, respectively. That is, as pitch excursions, both rises and falls segment the speech stream; but rises carry additional information: as the first part of an incomplete pitch peak that implicates a second part, they are used and interpreted as connecting to the second half of a dyad.

These abstract functions are essentially relational; the things related may fall primarily or simultaneously into the three general domains of interaction (e.g. turn-taking), text (textual relations or discourse structure), and information

[5]More specifically, junctural tones are diagrammatic icons arising from acoustic phonetic form, much like a map is a diagrammatic icon for relations between places. They aren't entirely "natural" or universal (cf. Bolinger 1980), but rather depend on the culture-specific evaluation of the signalling roles of intonational primitives in the system and cultural assumptions about which domains are relevant to interpretation.

structure (given/new, background/foreground). Rises, for example, function abstractly as connectives, and convey a broad range of more specific meanings depending on their textual, interactional, and discourse structural environments. They connect turns when the speaking floor is at issue, and textual units when the relation between such units is at issue; they connect participants, when address, participation, and attention are salient themes in context; and among at least some groups of speakers (e.g., the Texas sorority that I studied), they function much like text-aligned H accents to foreground new or exceptional information at the phrasal level when relative ranking of utterances in terms of shared knowledge is at issue (McLemore 1992c). (In the latter case, the relation is not so local and linear as the others, but rather paradigmatic, i.e. in the choice of contrastive tonal value). To summarize, junctural H tones:

- (segment and) connect
- implicate a second part, in associated text or interaction
- foreground associated information or action

In contrast, falls to low convey the least amount of relational information; in the sorority data, this form was found to generally segment textual phrases and turns, and under certain circumstances to co-occur with old, expected, or otherwise unexceptional discourse contributions. Unlike rises, falls alone do not elicit response (i.e. without additional conventions, textual information, or other cues coming into play). To summarize, junctural L tones:

- segment
- don't provide any information about what follows, in associated text or interaction
- background associated information or action

More specific interpretations of junctural tones, such as uncertainty, hesitation, conclusiveness, etc., arise from (more or less conventionalized) co-occurrence with text and aspects of context.

3.2 Levels

When junctural tones (H, L) are sustained (H–, L–), they become transparently iconic signs for continuation (as the tone is sustained, so is some aspect of the speech activity underway).

Tonal perseveration for H (H–) makes its general function continuative rather than connecting (i.e. current values for the speech activity underway are maintained, rather than changed as with H), and has the effect of changing its interactional value. That is, H– is more relevant to the interpretation of textual relations than to participant relations; unlike H, the intonation itself doesn't elicit response (i.e. but can if it co-occurs with text or other cues that do).

Perseveration of L (L–) also changes its general function from segmenting to continuing, (again, current values are maintained rather than changed, as with L). Like L, L– doesn't overtly cue interactional behaviors, although it can co-occur with them; like H–, L– cues a local continuative relation between textual units.

In the sorority corpus, both H– and L– junctural tones were found to co-occur primarily with old or expected information (consistent with Ladd's 1978 observations, as well as the data analyzed in Walker 1992), although within that functional space, H– still appears to mark information as foregrounded.

3.3 Falls to Mid

If falls to low segment the speech stream, and thereby text and interactional units, what do falls to mid do? As with an intonational rise, the form itself is incomplete when compared to a whole pitch peak; since function follows form closely in intonation, it isn't surprising that falls to mid seem to mark incompletion. The theory outlined above would predict that since falls to mid are falls that don't completely segment, they should share some characteristics with both rises and falls (see also Liberman 1975). The actual functional correlates of HM junctural tones and levels will be examined in the following sections.

4. COMPARISON OF DATASETS

The classroom data ("Lesson") consists of an excerpt of approximately 10 minutes of speech from a second grade mathematics lesson conducted in an inner-city Parish school in Pittsburgh, Pennsylvania. The portion of the lesson examined is the initial portion, called "pre-team" by participants, in which the teacher sets up the problem to be solved and works out preliminary solutions with contributions from the students. (Twenty-one of 225 phrases in the transcript are students' utterances, or joint teacher-student utterances; they have not been included in the intonational analysis because they are largely inaudible. An additional 2 phrases spoken by the teacher were not included because they were inaudible, resulting in a total of 202 phrases.)

The radio talk show ("Interview) dataset consists of an excerpt of approximately 4 minutes of speech from a call-in interview talk show aired on the public radio station in Austin, Texas. This particular exchange is between a caller (C), and the Austin mayor (M); the host's introduction and closing have been removed (for a total of 103 phrases). It is one of several exchanges between the interviewee and callers during the one-hour program. The intonational characterization below is of the entire exchange, including the speech of both the caller and the mayor, which display striking similarities.

In both the radio talk show interview and classroom contexts, participants bring knowledge about appropriate interactional behaviors to the verbal exchange. Call-in radio talk shows have a recurrent basic structure: the host introduces callers who direct questions or comments to the guest (and if there are no callers, the host plays this role); the guest responds; the host has the option of limiting the duration of either speaker's turn (by e.g., introducing another caller or requesting clarification). The nature of the communicative medium also imposes constraints on interaction: participants know they have a limited amount of time for the exchange, and that silence is to be avoided (see Coles 1991, Goffman 1981:197-330). This means that if one has the speaking floor, there is an especially urgent obligation to keep it filled with sound; and on the other hand, participants must respond promptly when a response appears to be called for.

A second grade classroom also has a recurrent, basic interactional structure (see Resnick et. al. 1991): the teacher has primary obligation to maintain the speaking floor, and allocates it either by calling on individual students (verbally or gesturally) or by indicating that a contribution from the class is required. Indeed, part of the lesson taught in the classroom is appropriate interactional behaviors generally. Furthermore, when verbal or nonverbal contributions, or attention more generally, are elicited from students, they are obligated to respond, or otherwise be (implicitly or explicitly) reprimanded.

4.1 Intonational Frequency Differences

The most significant (and obvious) difference between the two datasets is in the number of accents per intonational phrase. The pedagogical discourse shows fewer accents per phrase (1.5) than the interview discourse (2) – i.e. phrases are shorter and less complex intonationally. This difference is even more striking than the numbers suggest, since about half of the one-accent phrases in the Interview consist solely of discourse markers or so-called pause fillers, while only one third of the one-accent phrases in Lesson do.

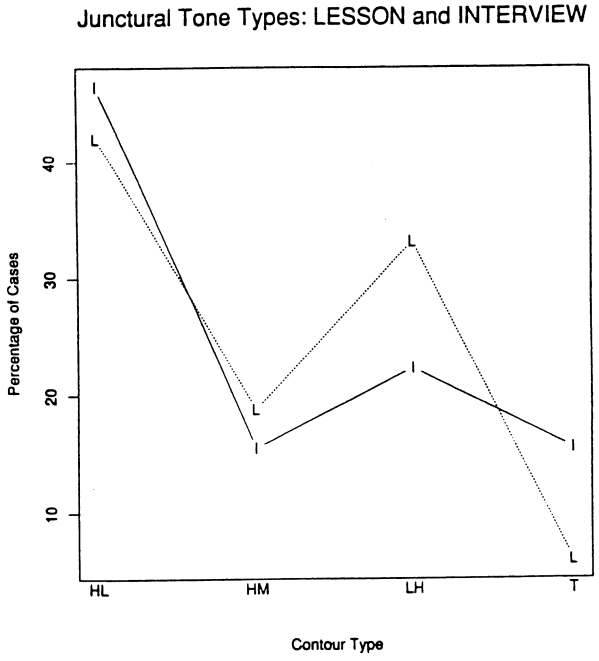
A comparison of overall contour types (tonal sequence corresponding to the whole 'phrase') in the two discourses indicated very little difference in the frequency of phrase-internal tonal elements. However, junctural forms are different at the level of better than p=.999. The greater variability in junctural forms than in phrase-internal composition is undoubtedly due in part to the fact that a majority of the intonational phrases segmented in Lesson contain *only* the junctural tone sequence.

The most striking and significant difference in junctural tone frequency across the two datasets is that forms of LH (rises and high final accents) occur more in Lesson than in Interview:

Proportions of Junctural Form Types for each dataset:

	HL	HM	LH	T
Lesson	.42	.19	**.33**	.06
Interview	.47	.15	**.22**	.15

FIGURE 3:



A higher proportion of phrases in Interview consist solely of level pitch, with no tonal value assigned (T above), than in Lesson. Since the T category itself is heterogeneous in pitch values relative to surrounding material (although constant in lack of pitch movement), it will be compared to level or sustained tones more generally (i.e. the class of H-, L-, M-, T).

The table below shows the proportion of junctural forms in each dataset according to stress alignment and tonal perseveration.

Proportions of Stress Alignment and Tonal Perseveration Variations:

	T*T	TT*	T*T–	TT*–	T
Lesson	.58	.20	.10	.06	.06
Interview	.58	.07	.05	.15	.15

A greater proportion of levels occur in Interview than in Lesson (i.e. the class of final level tones and phrases without tonal movement, T*T–, TT*–, and T below), and, as noted, there is a higher proportion of final accents in Lesson than in Interview.

These junctural form types are further distinguished by tonal value in the tables that follow; proportions shown are relative to the total of all three tables below (i.e. of 103 junctures in Interview, 202 junctures in Lesson).

Falls to mid occur more frequently in Interview than in Lesson, while rises occur more frequently in Lesson:

	FALLS		RISES
	H*L	**H*M**	**L*H**
Lesson	.30	**.06**	**.21**
Interview	.36	**.15**	**.07**

Among level junctural tones, sustained Mids occur more frequently in Lesson (especially H*M–), while sustained Highs occur more frequently in Interview:

	LEVELS					
	H*L–	HL*–	**H*M–**	HM*–	L*H–	**LH*–**
Lesson	.02	.03	**.07**	.02	.01	**.01**
Interview	.04	.02	--	.01	.01	**.12**

Finally, the two datasets also show a difference in the frequency of final accents, which are more common in Lesson overall: High final accents occur more frequently in Lesson than in Interview:

	FINAL ACCENTS		
	HL*	HM*	**LH***
Lesson	.07	.04	**.10**
Interview	.05	--	**.02**

To summarize, with respect to the general categories shown in Figure 3 above: In Lesson, most occurrences of LH are rises, L*H (about two thirds), while in Interview, most are high levels, LH*– (over half). The two datasets are similar in the proportion of falls to low; however, among variants of HM, falls to mid (H*M) occur more frequently in Interview, while a greater proportion of mid levels (HM–) are evident in Lesson. Among the greater proportion of final accents in Lesson, High final accents are far more frequent in Lesson than in Interview.

4.2 Intonational Distributions

Rises and high levels, and falls to mid and mid levels, are examined more closely below for their distribution within and between the two datasets, with respect to interactional behaviors (e.g. turn-taking), shared knowledge (including e.g., repetitions), and aspects of discourse structure (e.g., local textual relations).[6]

4.2.1 Rises (L*H)

In the Lesson dataset, rises pattern as cues to interactional behaviors in general, but underdifferentiate verbal interaction from nonverbal action or attention (i.e. connecting participants more generally rather than simply speaking turns). The majority of junctural tones preceding turn change are rises:

LESSON:

	H*L	L*H
Turn change	3	13
No turn change	57	30

There are 18 instances of student or whole-class response to the teacher: about two-thirds of the teacher's phrases (13 of 18) that elicit responses are spoken with L*H; 3 are spoken with H*L; one with LH* and one with T- (neither of which is shown in the table above)[7].

Of the rises preceding turn change, 4 occur on vocatives (*Nicole*?), 5 begin counting sequences which the students join in, and 5 occur on imperatives, questions, or variations on those forms, such as truncated *be* statements (e.g., *three groups of four is*?). In three of the latter 5 cases, the teacher names a student prior to the utterance that finishes her turn.

The 3 falls that precede turn change also occur on vocatives, but interestingly, differ from vocatives carrying rises in that they occur on names of boys rather than names

[6]Intonation is represented in the text as follows:
Falls to low are marked with a period ("."), falls to mid are marked with an addition sign ("+"), rises are marked with a superscript question mark ("?"), and sustained tones are indicated with a dash ("–"). Pauses are shown in angled brackets ("< >"), measured in seconds or noted as "<.>". CAPS indicates a High pitch excursion, **bold** indicates a Low (not necessarily marked).

[7]In Lesson, the single instance of T- that is followed by response occurs on a truncated statement, which the student, named in an earlier utterance, finishes (after "//"): *seVEN rows– / sound like more than– // three rows.*

of girls (e.g., *Martin.*).[8] The numbers are quite small and merely suggestive of gender socialization (cf. the finding that American women use intonational rises more frequently when talking with other women (Edelsky 1978), or report that they do (McLemore 1991a, b)). Some vocatives addressed to boys take rises when attention or other non-verbal action appears to be elicited (as in e.g., line 123 shown below).

In addition to cueing verbal interactional behavior, L*H is used on utterances that elicit specified behaviors, or attention more generally (e.g. with students' names; see also Figure 4 in the Appendix):

123 AntoNIO?
L* H

124 there's another TEAM?
L* H

125 team **four**?
L*H

It is important to note what is co-occurring in the classroom, as well as what follows use of the form: on videotape it is apparent that the teacher sometimes uses rising intonation on a current utterance, combined with gaze, to elicit attention from students who may have become distracted (cf. Keenan, Schieffelin & Platt, 1978).

Direction of attention or (verbal or nonverbal) action appears to be implicated in every occurrence of L*H in Lesson. This suggests that the children must attend to other cues (e.g., gesture, text) or conventionalized routines to determine the more particular significance of a rise for interactional behavior.

In the Interview dataset, the majority of junctural tones preceding turn change are falls to low, H*L, although 2 of the 5 rises that occur do precede turn change:

INTERVIEW:

	H*L	L*H
Turn change	8	2
No turn change	31	5

[8]In one of the cases mentioned above, L*H occurs on a boy's name, but is exceptional in being a vocative tag to an overall rising yes-no question rather than occupying its own phrase. In one of the instances of H*L, the same boy's name carries a fall when it occurs as a vocative tag to an overall rising WH question.

There are 12 changes of speaker: 8 follow intonational falls, and 2 follow intonational rises; one follows HL* and one HL*- (not shown in the table above). The 8 falls that precede turn change are not significantly different in form from the 31 falls that don't; they are scaled to quite low values, but similarly low values occur within single-speaker turns.

Both of the rises (L*H) are used by the Mayor, one in his initial single-phrase turn (*hello Rick*?), and the other in his second single-phrase turn, a yes-no question (*uh Rick is that area posted for no parking*?). In both cases, the rises redundantly cue the subsequent response.

No uniform pattern is evident in the additional 5 rises used in Interview within same-speaker turns, although all occur between clauses: in 2 cases L*H appears to create additional cohesion between clauses (*if* C_1 *then* C_2; C_1 *because* C_2), and in 3 cases L*H has the effect of foregrounding a new referent (see example below).

Why don't these rise cue a response? Because the particular domain to which an intonational form is interpreted as relevant, text or interaction, depends in part on the pragmatic and semantic import of the text with which the form is associated, as well as the interactional conventions at work in a given exchange. In short, when the 'connection' function is dropped into a textual location where turn-change is not plausible (either because there is no coherent meaning to respond to or because it's clear from textual structure that the speaker is not finished), that function is applied to textual units rather than turns or interactants.

4.2.2 High Levels (LH–)

High levels of the form LH*– constitute 12% of the total junctural forms in Interview, which makes them by far the most frequent form of level junctural tone in that dataset. In comparison, High levels constitute only 1% of junctural tones in Lesson. The following example from Interview is especially illustrative of the different uses of rises and high levels:

53 AUstin is **a great climate**?
H* L– L* H

54 It's **a good** CLIMATE FOR BICYCLING–
L H*–

55 uh THAT community **is growing**?
H* L L* H

56 OUGHT TO BE GROWING–
T*–

57 and it's **a** proDUCTIVE THING **to** DO 'N–
L H*– L H*–

58 it's less intrusive to our **en**VIRONMENT–
L H*–

59 uh more Energy efficient as **they** SAY–
H* L H*–

(See Figure 5 in the Appendix.)

The rises in 53 and 55 occur on a new referent or predicate in the discourse, while the High levels occur on discourse old or inferrable information (see Prince 1991 for given/new distinctions). The use of High levels on text framed as a listing sequence (i.e. in which the theme is continuous across junctures that correspond to a parallel textual frame) is similar to the use of phrase-final levels in the sorority planning meeting discourse reported in McLemore (1991a, c), in which the discourse is structured by a written program under discussion; there, too, L*H and LH– had distinct distributions, in which L*H occurred on exceptional or unordered items. In this case, the aspect of text relatively foregrounded by L*H happens to be newness (the material from which the subsequent list is formed), and the text relatively backgrounded by LH*– is a continuation of it.

Most of the High levels in Interview are used in this way – i.e., in a series of similar utterances containing old or inferrable information. (Two of the 5 repetitions in Interview take High levels, and a third is categorized as T, but sounds quite high). Two exceptions occur early in the exchange; they occur on new referents, and have a foregrounding effect similar to L*H (*I've noticed when I'm bicycling especially on SHOAL CREEK– / that the BIKE LANE–*).

High levels aren't generally used in Lesson. L*H is used on the parallel counting sequences, which are usually begun by the teacher and joined by the students (so elicitation of partipation is appropriate). There are 45 repetitions in Lesson; none take High levels. Except for the counting sequences, which take L*H, the majority of repetitions occur with falls (to mid or low), or (mid or low) levels.

L*H– doesn't precede a turn change in either Lesson or Interview.

4.2.3 Falls to Mid (H*M)

Like High levels, falls to Mid never precede a turn change in either dataset[9]; they also have a tendency to co-occur with old information. Occurrences of H*M constitute 15% of the total junctural forms in Interview compared to 6% of Lesson; in addition, their usage patterns are slightly different in the two datasets.

Uses involving self-interruptions and apparent disfluencies of various kinds (e.g. when the following utterance begins with or consists of *uh* or *um*) account for about half of the total number of H*M in Interview, compared to only about 1/6 of the occurrences of H*M in Lesson.

Consider the following general pattern of H*M use. The phonetic form results from physical contingencies when a speaker stops talking abruptly after a High peak (e.g., from Lesson: *SOMEone's+ / SOMEone's TALking while you're TALking*.) This would seem to instantiate neither a target Mid tone nor a higher-scaled Low tone; indeed, the Mid value could hardly result from tonal *target* per se at all. However, this kind of occurrence is in principle difficult to distinguish from the use of H*M in less mechanical cases of self-interruption (repair, disfluency, etc.) – which may also be used for deliberate communicative purposes, such as floor-holding. For example, consider lines 23 and 24 from Interview (which follow the Caller's report of a problem):

23 is there ANy**thing** you can do as MA**yor** to HELP us+
H* L H* L H* M

24 HELP us+ <.2>
H* M

25 RIders who **are** TRYing to **get** OUT there **and**. < 1.0>
LH* L H* L H* L

This usage is suggestive of an oblique interactional function of Mid (e.g., floor-holding); such a function is all the more oblique because of its affinity to relatively unplanned 'accidents' of speech. (Overt interactional behaviors such as speaker change aren't the extent of interaction in speech communication; all speech forms have interactional consequences, including not changing speaker turns. Intonation is an important resource in avoiding turn changes at points where there might otherwise be opportunities for them. See Sacks et. al. 1974.)

Another general pattern of H*M also has an instantiation that appears to result from physical contingencies; that is, when speech rate is accelerated through a juncture so that attaining a very low value for Low would be physically difficult (i.e. undershoot). This is especially apparent when the following phrase is shifted upward in pitch range. An

[9]Of course, given the appropriate textual material or local conventions, any junctural form can precede a turn change. In some dialects of American English, as well as Scottish and British English, H*M appears to be conventionally associated with certain types of interrogatives. (See e.g., Brown et. al. on Scottish English.)

example from Interview is shown in line 89 (see also Figure 6 in the Appendix):

88 AND **uh–**

89 CERtainly we'll **en**COUrage the poLICE to do everything **we** CAN+

90 we CAN'T **reLY upon** the poLICE **to** MAKE **people** **oBEY the** LAW **in** ALL CAses.

Again, however, the effect of this usage of H*M on both local textual relations and turn-taking (or lack of it) is indistinguishable from slower, seemingly more deliberate uses of H*M.

In both types of uses, whether deliberate or 'disfluent', H*M functions somewhat like a fall and somewhat like a rise, indicating a cohesive relation between two utterances (and their associated texts and acts) not by overtly marking connection or continuation, but by not quite segmenting the two utterances. In both uses, the material on which H*M occurs is in some sense treated by the speaker as less important than the material that follows it. There is a tendency for H*M to background, rather than foreground, when the relative rank of information is relevant to the communicative event. Consider the following example from Lesson:

60 but SEven+ <.>
H* M

61 **se**VEN **rows sound like** MORE.
L* H L– H* L

Here, the teacher responds to a students' answer by re-introducing the term *seven* into the discourse (*seven* ended a counting sequence several utterances prior, and other terms referring to the items counted have been used in the meantime, e.g., *number of cupcakes on this tray*). In line 60, *seven* is marked as salient by the High accent, but backgrounded by the M junctural tone.

In this usage, H*M is similar to the so-called 'backgrounding' or B contour examined in Liberman & Pierrehumbert (1984) and Steedman (1991) (who describes it as L+H* L H%). Based on the data analyzed here, as well as that in Liberman & McLemore (1992), it appears that in fact (...)H*LH and H*M are variations of the same form. (See McLemore 1992 for a more indepth discussion of formal and functional patterns of falls to mid and their variations.)

Finally, another general pattern of H*M use is closely related to the backgrounding function, but differs from it in that rather than occuring on old information or a theme or topic about which more follows, it occurs on utterances out of the blue that set up an expectation for more speech or action, e.g. in 2 instances from Lesson on *you know what*. Since this use of H*M, in combination with such textual phrases, has the effect of creating suspense, I will refer to it here as the *anticipatory* function of H*M. It exploits the locally cohesive function of H*M, as well as the backgrounding usage. Most occurrences of H*M in Lesson are backgrounding or anticipatory.

The anticipatory function of H*M is even more apparent when the final M is sustained as a level, i.e. H*M–, which is a frequent form of HM in Lesson, as discussed below.

4.2.4 Mid Levels (H*M–)

Mid levels of the form H*M– occur at 7% of junctures in Lesson; they are the most common form of level junctural tones in the Lesson dataset. None occur in Interview.

Nearly one third of H*M– occurrences are within a sentence, as in the following example (see Figure 7 in the Appendix):

33 LAUREN said+–
H*– M–

34 **o**KAY?
L H*

35 LAUREN said+–
H*– M–

36 "**there are** THREE **rows.** "
L– H* L

This excerpt is from part of the fictional narrative the teacher uses to set up the problem; line 36 contains the first mention of a number that will be crucial to the solution. The use of H*M– in the quotative frame of lines 33 and 35 is both backgrounding and anticipatory with respect to what follows (line 36). The sustained final Mid tone enhances the anticipatory function by simple temporal duration. (Note the relation between the intermediate scaling of Mid and its tonal lengthening; each aspect suggests continuation.)

As the *okay* with LH* in line 34 indicates, the teacher is concerned with holding the students' attention; indeed, in other occurrences of H*M–, the teacher is simultaneously performing gestures, such as holding up a certain number of fingers, to which the students must attend.

Recall that H*M and H*M– are never used to elicit response; they are also never used to directly elicit students' attention. Rather, the teacher uses L*H or LH* to overtly elicit attention or response, and forms of Mid to actively *hold* students' attention. That is, the use of Mid in some (but not all) cases indicates that the speaking floor may be in question and the current speaker is claiming it. That is

also the interactional function of level junctural tones, as noted previously, which makes sustained M– all the more effective in this regard.

4.3 Summary

In Lesson, phrases are shorter and simpler intonationally: many phrases contain one continuous pitch movement, overall falling or overall rising, some with emphasis at either or both ends (TT*). Although the teacher is the primary speaker, participation of the students is considered necessary to accomplishing the pedagogical purpose of the discourse, hence the frequent elicitation of students' attention, action or verbal response (L*H). At the same time, the teacher must maintain interactional order so that students follow the main points of the lesson (H*M, H*M–). The discourse is pre-planned for the most part; i.e. the teacher has outlined the lesson, if not the particulars of her speech, and the discourse incorporates speech routines that are used in other datasets from this classroom. When intonational junctures occur within otherwise cohesive textual units, or on repetitions, they are more often accompanied by enhancing gestures and actions than by indications of disfluency, suggesting a careful, emphatic speech style (H*M, H*M–); indeed, there are numerous repetitions and rephrasings (see Resnick et. al. 1991 on the forms and pedagogical functions of *revoicing* in this classroom). Consistent with this, there are no occurrences of *uh* or *um* in the dataset, but rather discourse markers that, together with the whole range of intonational forms, appear to play a role in eliciting attention and structuring the discourse and the activity overall.

In Interview, phrases tend to be long and intonationally complex. With some frequency, junctural forms break up otherwise cohesive textual units (H*M, T), often accompanied by self-interruptions and pause fillers. In part, this undoubtedly reflects the unplanned nature of the discourse, although these patterns also have useful interactional implications (i.e. preventing speaker change). Turn change is not cued explicitly with intonation, but apparently by other (semantic and pragmatic) sources of interactional information. Emergent parallelistic structure is evident in some same-speaker turns (LH*–), and given/new relations are marked intonationally in some of these sequences (L*H, LH*–).

5. CONCLUSION

The different junctural forms clearly have a high degree of overlap in aspects of their functions; e.g., H*M and LH*– are similar in (generally) not cueing interactional behavior, creating local textual cohesion, and co-occuring with non-new information. The distinction between them is subtle, and has as much to do with the associations they accrue from both general and local patterns of use (i.e. their more or less conventional indexical value) as with the fundamental differences in their (iconic) form.

Nonetheless, for individual junctural forms there are clear tendencies in usage patterns with respect to interaction, information, and textual relations. For the community of speakers from which the present data were drawn, it would be surprising to find uses (conventionalized or not) that aren't generally consistent with the patterns observed here; rather, local discourse variation is evident in the selection of specific forms, and the aspect of their function that is made salient by the contingencies of the communicative event itself (including, e.g., the domain a given use is primarily taken to comment upon, text or interaction).

Consider the case of H*M. In both Interview and Lesson, the patterning of H*M suggests that it is used, like levels, to continue across a juncture. However, like L*H and H*L, the phonetic form contains a pitch excursion which effectively segments the pitch stream; it differs from H*L in not segmenting all the way to Low, and unlike LH*–, it segments in part. Like L*H, it is used to mark more specific relations between the contours it partially segments, as well as their associated text and acts. In Interview, this cohesive function is most salient; i.e. it occurs primarily in self-interruption, mid-clausally, and on particles like *um*. Its use has interactive implications (i.e. in terms of floor holding), which are less direct than e.g., the use of L*H; it differs crucially from L*H in not explicitly cueing interactional behaviors. In Lesson, the backgrounding and anticipatory functions of the final mid value are most salient, and the interactional consequences (of floor- or attention-holding) are more clearly evident, since all forms of attention and interaction are visible. The most frequently occurring subtype of HM is a mid level, in which the anticipatory function is enhanced, with respect to both text and interaction.

REFERENCES

Brown, G., K. Currie, & J. Kenworthy. 1980. *Questions of Intonation*. Baltimore: University Park Press.

Bolinger, D. 1980. Intonation and "Nature." In M. Foster & S. Brandes, eds., *Symbol as Sense: New Approaches to the Analysis of Meaning*. New York: Academic Press.

Coles, F. 1991. Pre-Sequences in a Radio Talk Show. In *Texas Linguistics Forum 32: Discourse*, C. McLemore ed. Austin: University of Texas Department of Linguistics and the Center for Cognitive Science. 21-42.

Du Bois, J., S. Schuetze-Coburn, D. Paolino & S. Cumming. 1991. Discourse Transcription. University of California at Santa Barbara ms.

Ferguson, C. 1983. Sports Announcer Talk: Syntactic Aspects of Register Variation. *Language in Society* 12: 153-172.

Fillmore, C. 1981. Pragmatics and the Description of Discourse. *Radical Pragmatics*. Cole, P. (ed.). 143-166.

Goffman, E. 1981. Radio Talk: A study of the ways of our errors. In *Forms of Talk*, pp. 197-330. Philadelphia: University of Pennsylvania Press.

Gumperz, J.J. 1982. *Discourse Strategies*. London: Cambridge University Press.

Hockey, B. A. 1991. Prosody and the Interpretation of *okay*. Paper presented at the AAAI Fall Symposium, November 1991.

Hockey, B. A. 1992. Prosody and the Interpretation of Cue Phrases. Paper presented at the IRCS Workshop on Prosody in Natural Speech, Philadelphia, PA, August 5-12.

Keenan, E. Ochs, B. Schieffelin & M. Platt, 1978. Questions of Immediate Concern. In E. Goody (ed.), *Questions and Politeness*. Cambridge: Cambridge University Press. 44-55.

Ladd, D.R. 1978. Stylized Intonation. *Language* 54:3, 517-541.

Liberman, M. 1975. *The Intonational System of English*. MIT Ph.D. Dissertation. (Published by Garland, New York, 1979.)

Liberman, M. and C. McLemore. 1992. The Structure and Intonation of Business Telephone Openings. In *The Penn Review of Linguistics* 16: 68-83.

Liberman, M., C. McLemore & A. Woodbury. 1991. On the Nature of Prosodic Phrasing. Paper presented at the LSA Summer Institute Workshop on Grammatical Foundations of Prosody, July 5 1991, University of California at Santa Cruz, Santa Cruz, CA.

Liberman, M. and J. Pierrehumbert. 1984. Intonational Invariance Under Changes in Pitch Range and Length. In *Language Sound Structure*, M. Aronoff and R.T. Oehrle, eds. Cambridge, Mass: MIT Press.

McLemore, C. 1992. Discourse Context and Prosodic Function. Paper presented at NWAVE XXI, Ann Arbor, Michigan, October 1992.

McLemore, C. 1991a. *The Pragmatic Interpretation of English Intonation: Sorority Speech*. University of Texas Ph.D. dissertation.

McLemore, C. 1991b. The Interpretation of L*H in English. In *Texas Linguistics Forum 32: Discourse*, C. McLemore, ed. Austin: University of Texas Dept. of Linguistics and the Center for Cognitive Science. 175-196.

McLemore, C. 1991c. Boundary Tone Values and Shared Knowledge. Paper presented at Grammatical Foundations of Discourse and Prosody, LSA Summer Institute workshop, Santa Cruz, California, July 1991.

Pierrehumbert, J. 1980. *The Phonetics and Phonology of English Intonation*. MIT Ph.D. dissertation.

Prince, E. 1991. The ZPG Letter: Subjects, Definiteness, and Information-Status. In S.A. Thompson and W. Mann eds., *Discourse Description: Diverse Analyses of a Fund-Raising Text*. Amsterdam/ Philadelphia: John Benjamins.

Resnick, L., M. Leer, V. Bill, and L. Reams. 1991. From Cupcakes to Equations: The Structure of Discourse in a Primary Mathematics Classroom. University of Pittsburgh Learning Research and Development Center ms.

Sacks, H., E. Schegloff, and G. Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50: 696-735.

Shriberg, E. 1992. Intonation of Clause-Internal Filled Pauses. Paper presented at the IRCS Workshop on Prosody in Natural Speech, Philadelphia, PA, August 5-12.

Silverman, K, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg. 1992. TOBI: A Standard for Labeling English Prosody. *ICSLP Proceedings*.

Silverstein, M. 1979. Language Structure and Linguistic Ideology. In P. Clyne, et. al. (eds), *The Elements: A Parasession on Linguistic Units and Levels*. Chicago: Chicago Linguistic Society. 193-247.

Steedman, M. 1991. Structure and Intonation. *Language* 67.2: 260-296.

Swales, J. 1990. *Genre Analysis*. Cambridge, U.K.: Cambridge University Press.

Walker, M. 1992. When Given Information is Accented: Repetition, Paraphrase and Inference in Dialogue. Paper presented at the IRCS Workshop on Prosody in Natural Speech, Philadelphia, PA, August 5-12.

Wightman, C., S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. 1991. Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. Unpublished ms.

Woodbury, A. C. 1992. Prosodic Elements and Prosodic Structures in Natural Discourse. Paper presented at the IRCS Workshop on Prosody in Natural Speech, Philadelphia, PA, August 5-12.

Woodbury, A. C. 1987a. Rhetorical Structure in a Central Alaskan Yupik Eskimo Traditional Narrative. In J. Sherzer & A. C. Woodbury, eds., *Native American Discourse: Poetics and Rhetoric*. Cambridge: Cambridge University Press.

APPENDIX

Figure 4

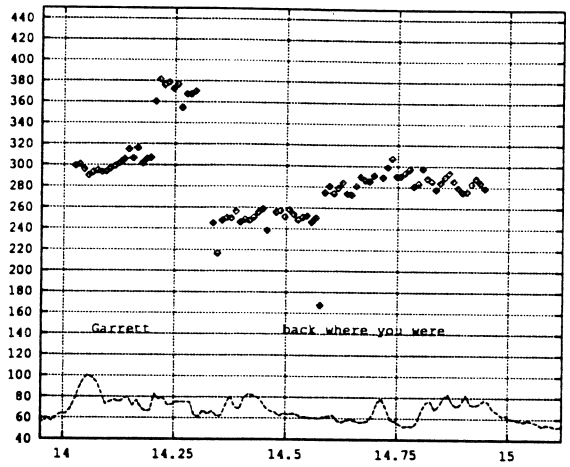


Figure 5

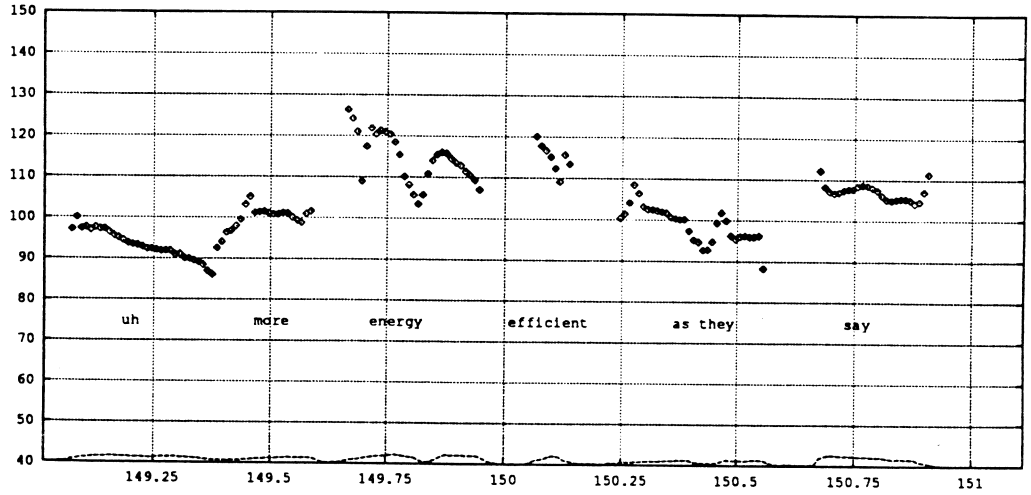


Figure 6

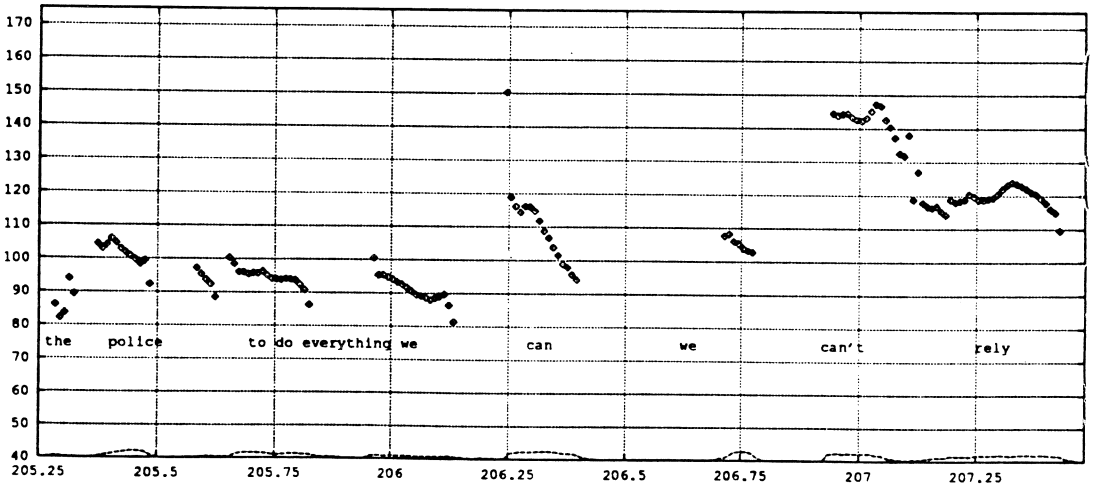
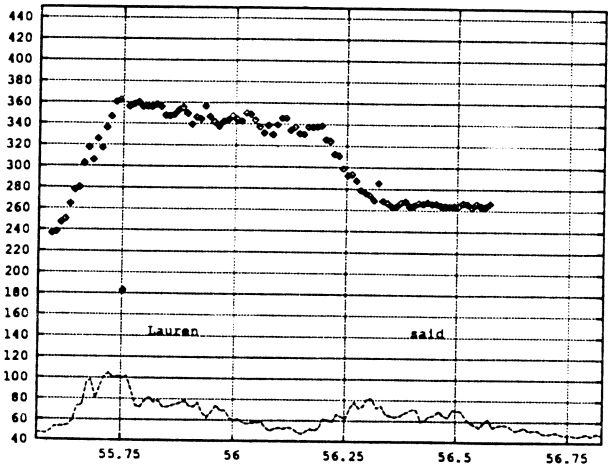


Figure 7



PROSODIC ASPECTS OF M.L. KING'S "I HAVE A DREAM SPEECH"

Corey A. Miller

Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104

ABSTRACT

This research examines the prosodic characteristics of Martin Luther King's "I have a dream today" speech in an effort to better understand both the prosody of oratory and the prosodic qualities of King's speech that move people. The peroration of the speech was digitized and analyzed using the Waves program on a Sun SparcStation. Among the salient findings were King's sustained high pitch, several recurrent pitch patterns and various special effects. Many of these features are exemplified with reference to pitchtracks. Some discussion of the characterization of oratory with respect to speech and nonspeech modes of perception ensues.

1. INTRODUCTION

Martin Luther King is probably the first person that comes to many Americans' minds when they think of a great orator. In this paper, I wish to explore some of the prosodic manifestations of King's oratory. Ultimately, such an enterprise might contribute to an understanding of what it is acoustically in such oratory that moves people, abstracting away from the content of the words. More narrowly, I hope this research contributes to what we know about English intonation in general, since we are all likely to employ some of the features discussed below, albeit in a modified way, for some communicative effect in our own speech. Viewed in this way, the register of the public oration might be looked at simply as an exaggeration of some of the prosodic tools that we already possess and use in everyday verbal interaction. More strongly, however, we may interpret the special prosodic features of the oration investigated here as cues to listeners that cause the acoustic signal to be perceived by a poetic, rather than a strictly speech, mode of perception.

King's oratory falls within the tradition of American black preachers. Many traits of his delivery can be seen in the sermons and political speeches of contemporary black clergymen and politicians such as Jesse Jackson and William Gray. The literature on black preaching and oratorical style only hints at the kinds of issues that are profitably investigated with the aid of acoustic analysis. For example, Boulware offers the following critique:

> King made great use of his nasal resonators, which enriched his vocal tones. These tones were slightly flat, because of his failure to make more oval the openings of his vocal outlets.[1]

The text chosen for analysis here is the peroration of King's "I Have a Dream Today" address to several hundred thousand people at the March on Washington on August 28, 1963.[2] This is probably his best known speech, and many of its passages are intimately familiar to many Americans. I have isolated several prosodic highlights of the speech which I will describe below, making use of pitch contours, amplitude contours and spectrograms produced with the Waves speech analysis software package on a Sun SparcStation.

2. HIGH PITCH

One of the first observations I made in the course of intonational analysis was the extremely high pitch at which "I Have a Dream Today" was delivered. King's pitch peaks generally average between 280-300 Hz; well above the average pitch for an adult male. The high pitch is most likely due to the high amplitude at which King delivered the speech. According to Cruttenden, "...producing syllables with extra loudness produces extra airflow through the vocal cords and pitch goes up accordingly."[3] It seems likely that the emotional content of the speech and the effect King wished to produce were responsible for the loudness, since his microphone would have obviated the need to shout. Nevertheless, King certainly would have had a motivation to talk over the muting effect of several hundred thousand people talking, coughing, etc.

---

[1] Boulware (1969), p. 250.

[2] I would like to thank Raymond Trent, of the Biddle Law Library for providing me with an excellent recording. I digitized the recording at 8000 Hz on a Sun SparcStation using the Waves speech analysis package.

[3] Cruttenden (1986), p. 50.

In order to see what King's pitch level would be in conversational circumstances, a brief analysis was made of his "Letter from Birmingham Jail," of April 16, 1963. While the letter is not exactly conversation, it has the advantage of dating from the same year as "I Have a Dream Today" and being delivered in the understandably sober tones of one who is in jail. In the "Letter," King's pitch ranges mostly between 80 and 120 Hz.

3. RECURRENT PITCH PATTERNS

I studied pitch contours from approximately the last five and one-half minutes of the "Dream" in order to note recurring fundamental frequency patterns. In the broadest sense, there are steadily decreasing and steadily increasing patterns. The steadily decreasing, or downstep, patterns are often characterized by a final pronounced fall. The steadily increasing patterns often have a final post-tonic fall, or a downstepping final pattern.

3.1. Steadily Decreasing and Steadily Increasing Pitch Patterns: Downstep and Upstep

Figure mlk5 provides an example of the steadily decreasing pattern on a fairly long phrase. In this case there is a brief rise up to the pitch accent, after which the steady decrease in fundamental frequency begins. In all examples of this kind, there is a pronounced fall on the last word, creating a break with the steadier decreases leading up to it. In figure mlk5, there is a rise up to the first *dream*, and then steady downstep to the final *dream*. A variant of the steadily decreasing pattern appears on some of the very short phrases that punctuate the "Dream," as exemplified by figure mlk16. If it is to be imbued with any consistent meaning in the "Dream," the downstepping patterns described here could often be said to conclude the passages in which they occur. This is frequently not the case, as downstepping patterns may be followed by upstepping patterns on the same theme. Nevertheless, series of upsteps and downsteps generally reach their final conclusion on a downstep.

Figure mlk8 is an example of upstep without a final fall. It is followed by a 1.5 second pause until the beginning of the next phrase. The upstepping pattern combined with a long pause is likely to be a suspense-creating device. King also achieves this effect by elongating final words in upstepping passages. Such long final words often exhibit a steady decrease in pitch.

3.2. Upstep followed by Downstep

When King's phrases are taken two at a time, a recurrent pattern emerges whereby an upstepping pattern is followed by a downstepping pattern, creating a fairly complete unit of thought. The upstepping phrase is often accompanied by final syllable lengthening and a pause of over 1 second, thereby creating suspense for the concluding downstepping phrase. For example, figure mlk8's upstep pattern is followed by a downstepping phrase with a pronounced final fall, figure mlk9. A large complex of up-down patterns, resulting in a frenzied, breathless passage is represented by Figures mlk39-mlk42. Figure mlk39 is an upstep without a final fall, followed by mlk40 which is a downstep that also does not dip particularly low, preparing the way for mlk41, which contains a brief upstep to the first syllable of *crooked*, which has one of the highest fundamental frequencies reached in the speech: 411 Hz. The latter half of figure mlk41 is a downstep without a pronounced final fall, leading to the the upstep in Figure mlk42, which *is* concluded by a low-dipping final fall on the last syllables of *together*, thus ending this powerful prosodic and thought unit.

3.3. Lists

Efforts to create parallels are present on both the prosodic and textual levels of the "Dream." One manifestation of textual parallelism might be termed the "list." The following passage contains a list of states and their characteristics which happen to break down roughly into pairs that can be analyzed according to the upstep-downstep pattern discussed in the previous section. In this way, the textual parallelism of high places in various states is overlaid with a prosodic parallelism of upstep-downstep patterns.

> UPSTEP: ...from the mighty mountains of New York
>
> DOWNSTEP: let freedom ring from the heightening Alleghenies of Pennsylvania
>
> UPSTEP: let freedom ring from the snow-capped Rockies of Colorado
>
> DOWNSTEP: let freedom ring from the curvaceous slopes of California

Another memorable list that King evokes is that of contrasting kinds of people whom King wishes to see brought together. Consider the following passage containing three lists of groups of people (italicized), "...we will be able to speed up that day when all of God's children, *black men and white men, Jews and Gentiles, Protestants and Catholics* will all be able to join hands..." King applies parallel downstepping pitch patterns and levels on the elements of the first two lists. For the last list, *Protestants and Catholics*, King employs a slightly different strategy. The unstressed, normally unpronounced orthographic *o* of *Catholics* is pronounced as a schwa,

giving *Protestants* and *Catholics* three syllables each. The insertion of an extra syllable into *Catholics* contributes to the staccato effect of *Protestants and Catholics*, picking up on the staccato effect of *black men and white men.*

4. SPECIAL EFFECTS

In addition to his manipulation of up- and downstepping pitch contours, King employs various prosodic special effects to help catapult his speech into the memorable oratorical register for which he is famous. In contrast to pitch contours, many of these effects are realized on single segments or words, including vibrato and the extreme lengthening of certain vowels and consonants. Other effects, such as breathiness, breathlessness, staccato, and the non-reduction of reduced vowels to achieve even timing, are associated with whole phrases.

4.1. Vibrato

Traditionally vibrato is divided into two categories: pitch vibrato and amplitude vibrato. King employs both, in addition to a vibrato that appears to manifest itself in a tradeoff of formant values in the course of a segment. *Made*, in figure mlk41, has a pitch vibrato with a frequency of approximately 10 Hz. Spectrographic analysis of the segment appears to show pulsating, varying energy in the formants, perhaps aiding the vibrato effect. *My*, in Figure mlk24, appears to feature a combination of pitch and amplitude vibrato, as can be seen in the pitch contour and the intensive cyclic pattern present in the amplitude contour.

4.2. Segment Lengthening

To effect vibrato on a segment requires that the segment be of longer than normal conversational length. In fact, all of the vibratoed words discussed above are realized as particularly long segments. *Down*, in Figure mlk30, is an example of extreme vowel lengthening: the word is 1.34 seconds long. Such a long word also provides a stage on which a steadily decreasing contour can be realized. Long vocalic segments often feature rising patterns as well, as evidenced by *my* in Figure mlk24 (.67 seconds) and *rise* (.52 seconds) in Figure mlk8.

Extremely long [s] segments at both the beginning and ends of words can be found in the "Dream". The phrase "sweltering with the heat of injustice" has an initial consonant that is .18 seconds long. The use of this device on this word has onomatopoeic effect, as it might conjure up the hiss of a furnace or a desert snake. Here, prosodic special effects are grouped near one another, as the final [s] of the above phrase lasts .34 seconds.

4.3. Breathiness, Breathlessness and Frenzy

Breathiness is an intermittent feature of King's speech in the "Dream." It is exemplified in the passage, "...from every state and every city," particularly on the word *city*. The combination of breathiness and low pitch of the final syllable (120-160 Hz) results in a particularly grave and grandiose effect. Contrastively, a lack of breath characterizes certain passages which I call frenzied. These are long, quickly spoken passages in which King hardly pauses between phrases. Such a passage is illustrated by figures mlk39 through mlk42, lasting 12.3 seconds, which have already been discussed regarding their repeating upstep-downstep pattern. Another such passage also lasting about 12 seconds consists of the words:

UPSTEP: ...we will be able to work together

DOWNSTEP: to pray together

DOWNSTEP: to struggle together

DOWNSTEP: to go to jail together

DOWNSTEP: to stand up for freedom together

DOWNSTEP: knowing that we will be free one day.

The first phrase has a rising pitch pattern, while all of the subsequent phrases exhibit falling pitch patterns. Instead of creating biphrasal pitch parallels as in mlk39-mlk42, King has simply created uniphrasal parallels in the latter passage.

5. CONCLUSION

Martin Luther King employs suprasegmental effects to a high degree of artistry in the "Dream." Through them he is able to affect his audience in ways that penetrate deeper than the semantic value of the words would normally imply. Tsur has introduced the concept of a poetic mode of speech perception[4] which, while embedded within the speech mode (as elaborated by Liberman and others[5]), is activated by particular acoustic events, thus allowing for the affective or right-hemispheric qualities of the speech signal to be more fully realized than they would be in normal speech. Viewed in this light, the "Dream's" lasting impact is owed to some extent to the ability it has to cause listeners to process the acoustic signal via this poetic mode, thus awakening some of the emotive power normally restricted to the nonspeech mode of perception.

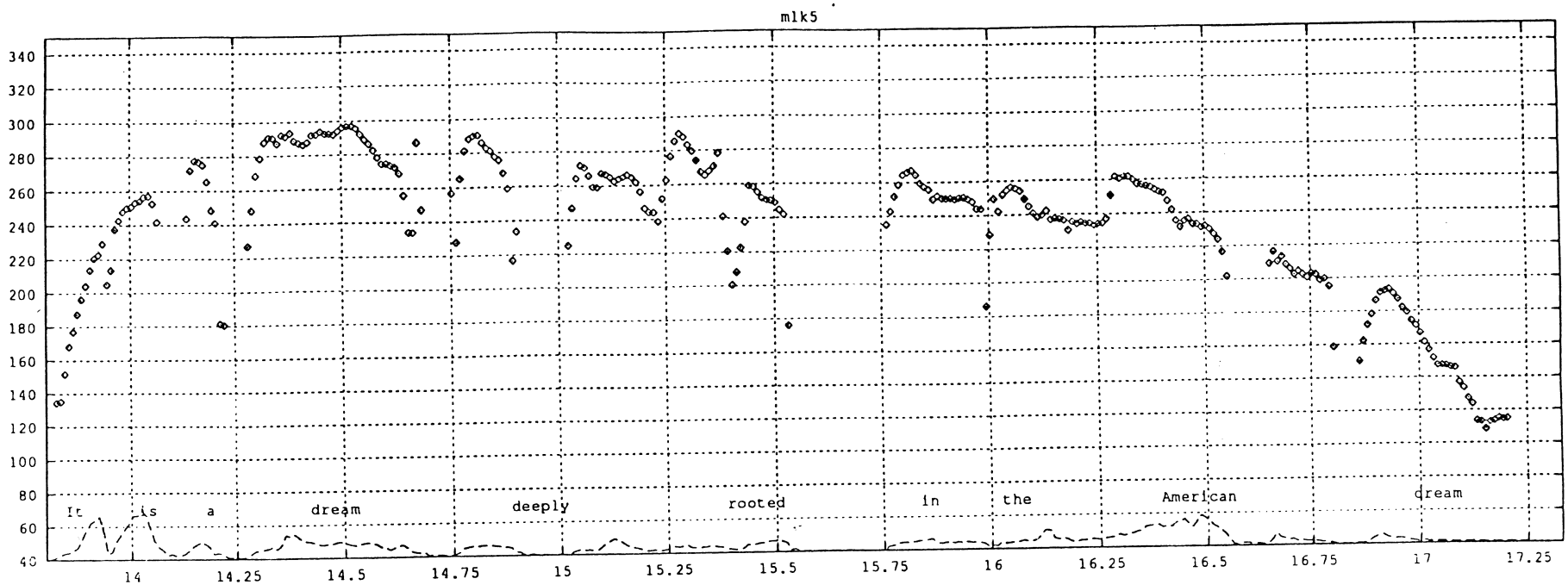
---

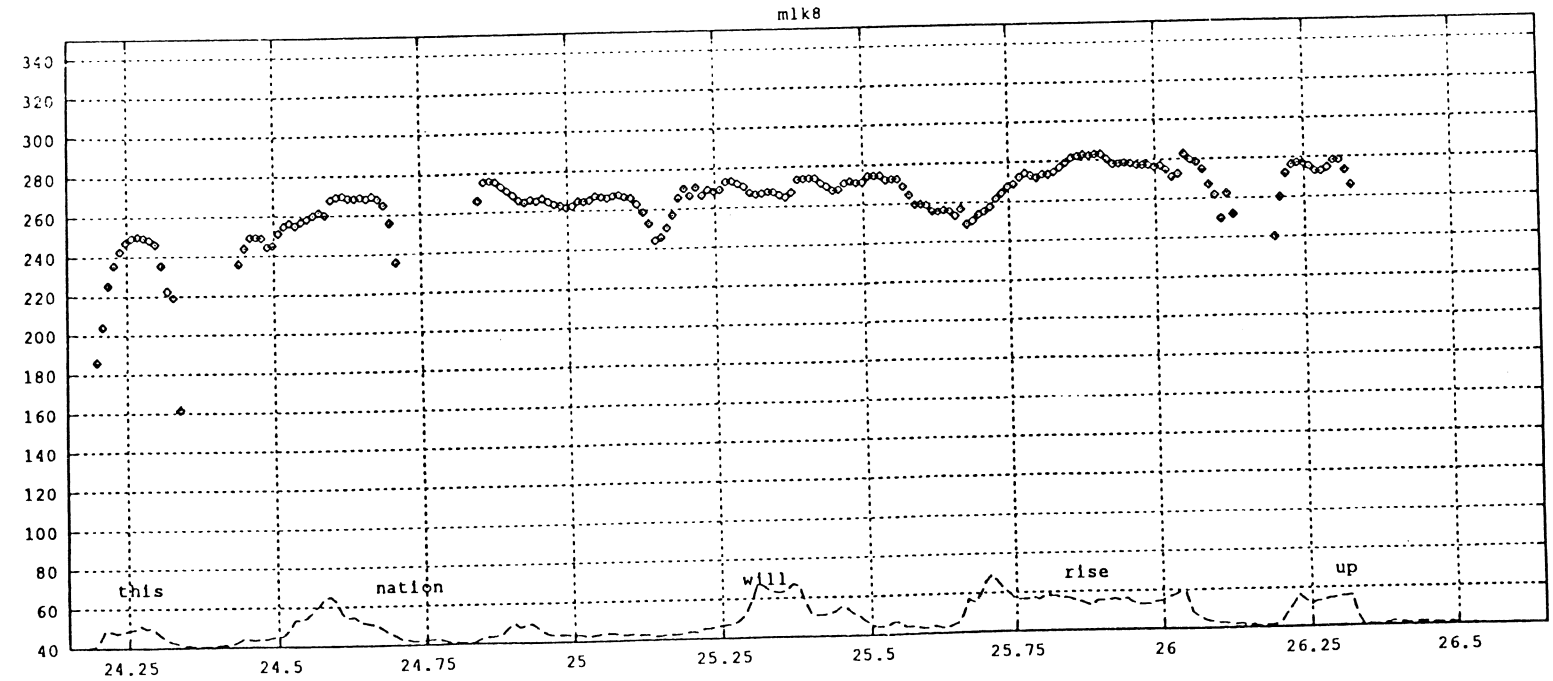
[4]Tsur (1992).

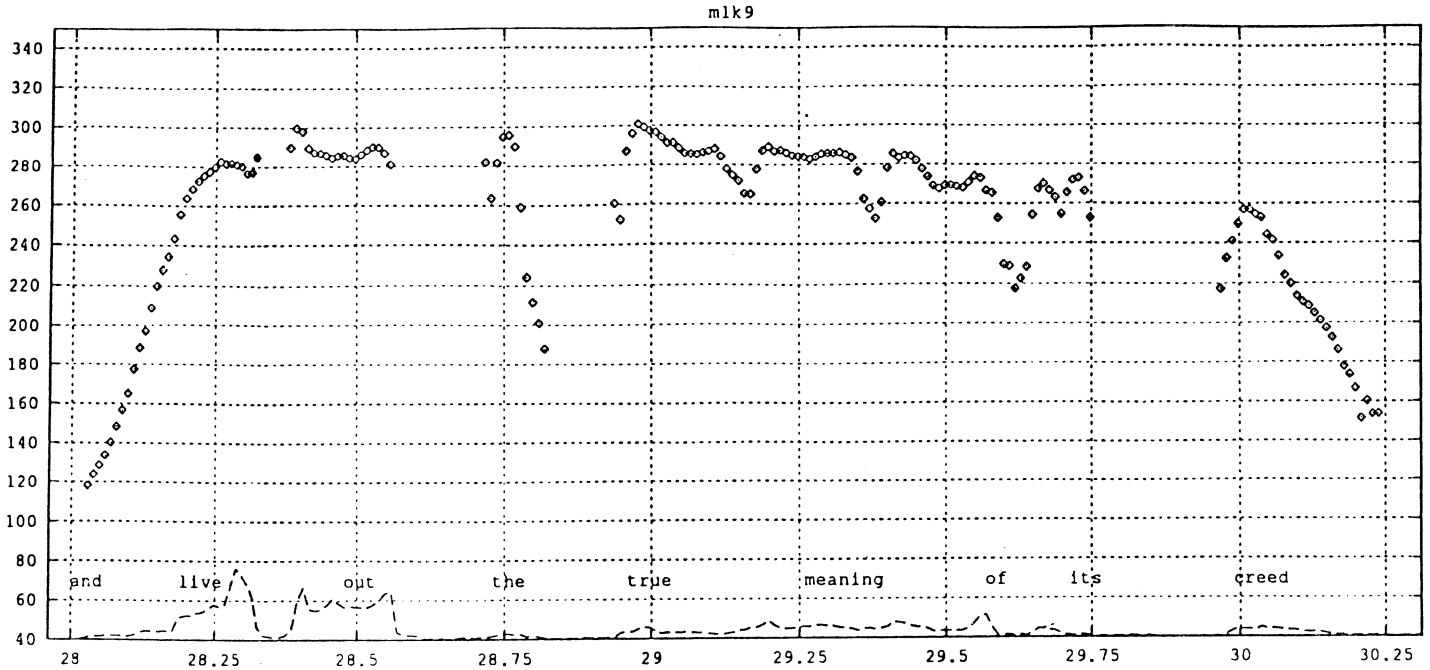
[5]See for example, Liberman *et al.* (1967).

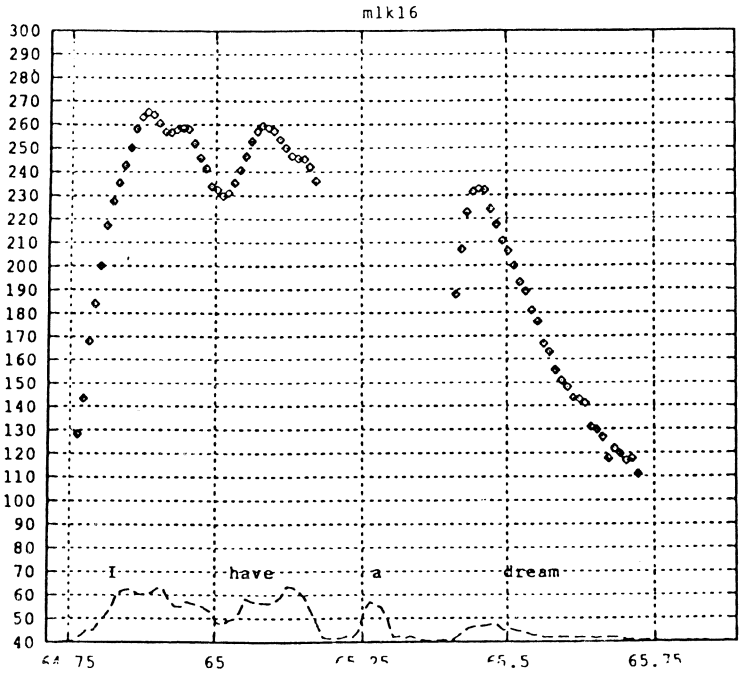
REFERENCES

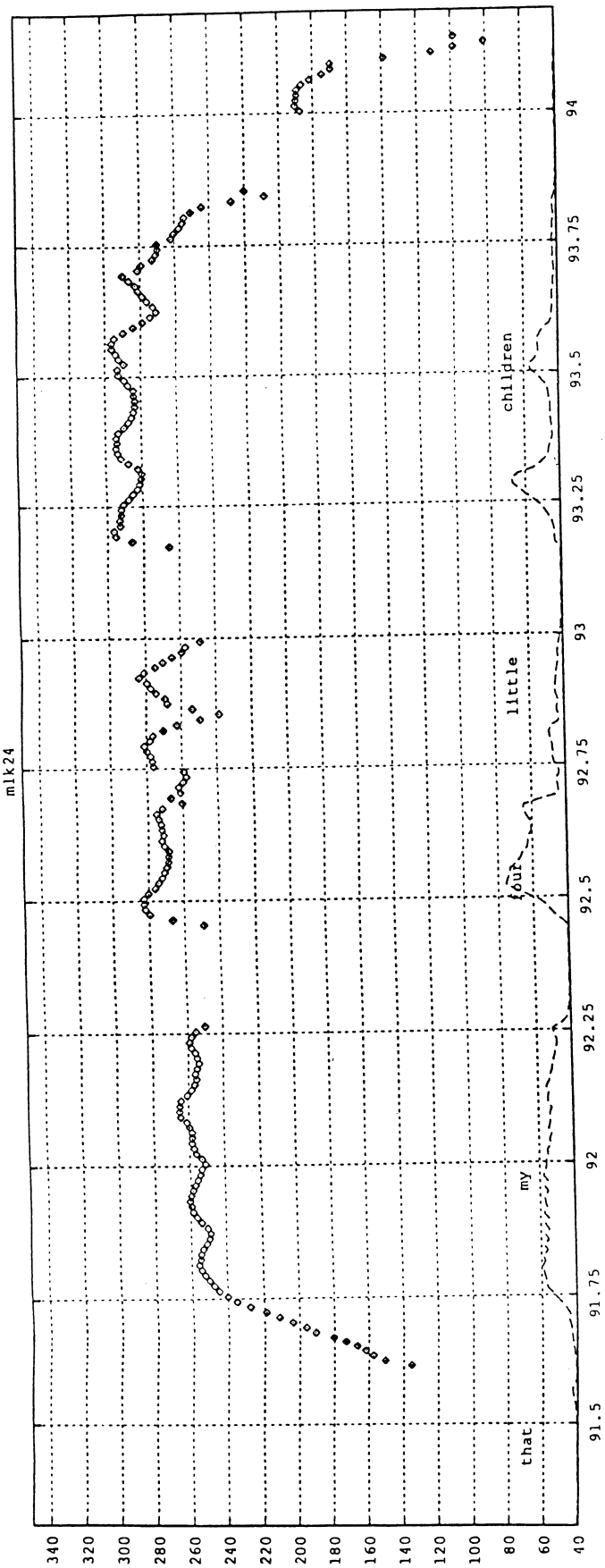
1. Boulware, Marcus, *The Oratory of Negro Leaders, 1900-1968,* Negro Universities Press, Westport, Conn., 1969.
2. Cruttenden, Alan, *Intonation,* Cambridge University Press, Cambridge, 1986.
3. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. "Perception of the Speech Code," *Psychological Review* 74: 431-461, 1967.
4. Tsur, Reuven, *What Makes Sound Patterns Expressive,* Duke University Press, Durham, N.C., 1992.

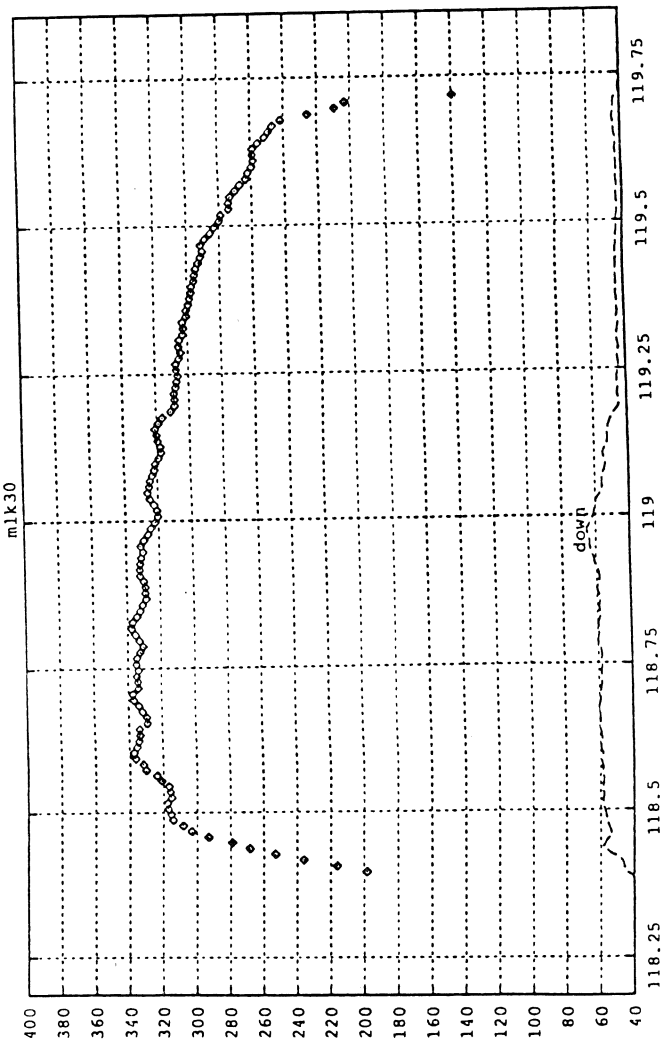

mlk5
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
It
is
a
dream
deeply
rooted
in
the
American
dream
14
14.25
14.5
14.75
15
15.25
15.5
15.75
16
16.25
16.5
16.75
17
17.25

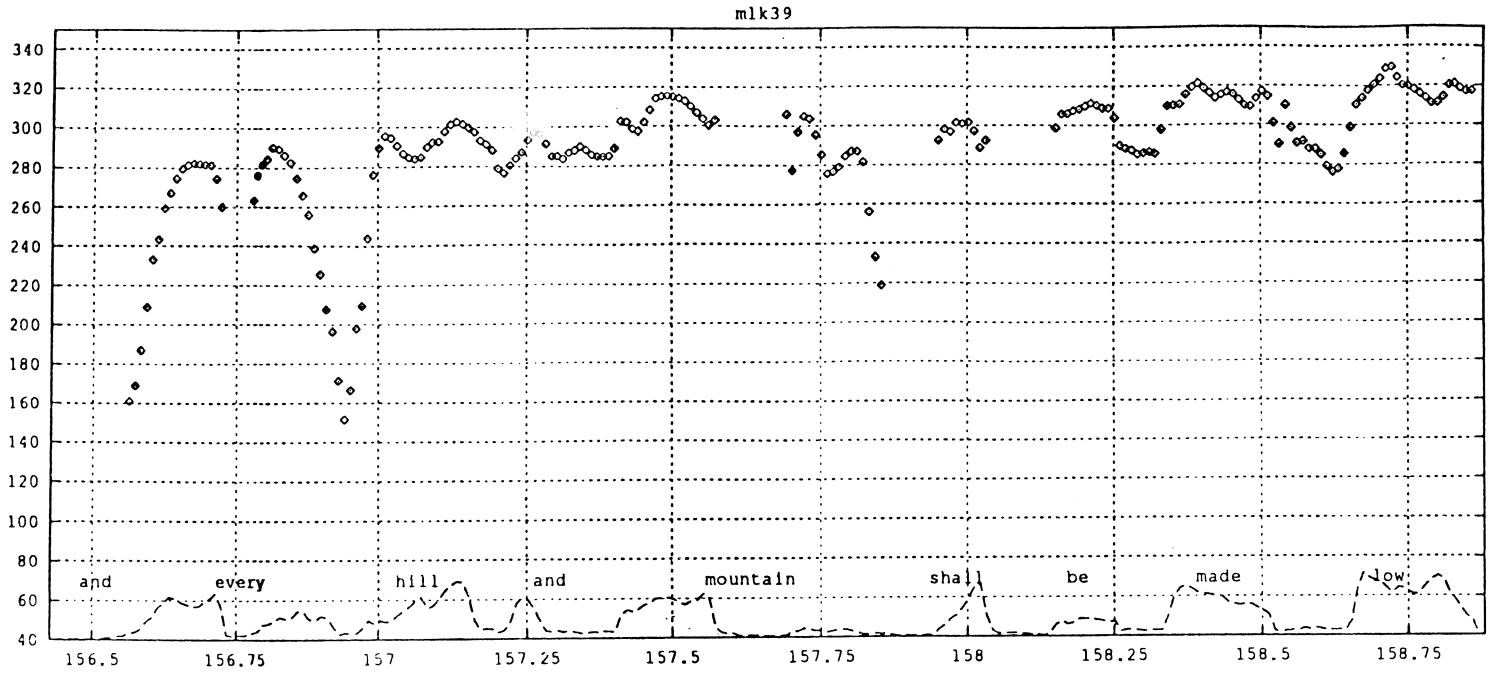

mlk8
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
this
nation
will
rise
up
24.25
24.5
24.75
25
25.25
25.5
25.75
26
26.25
26.5

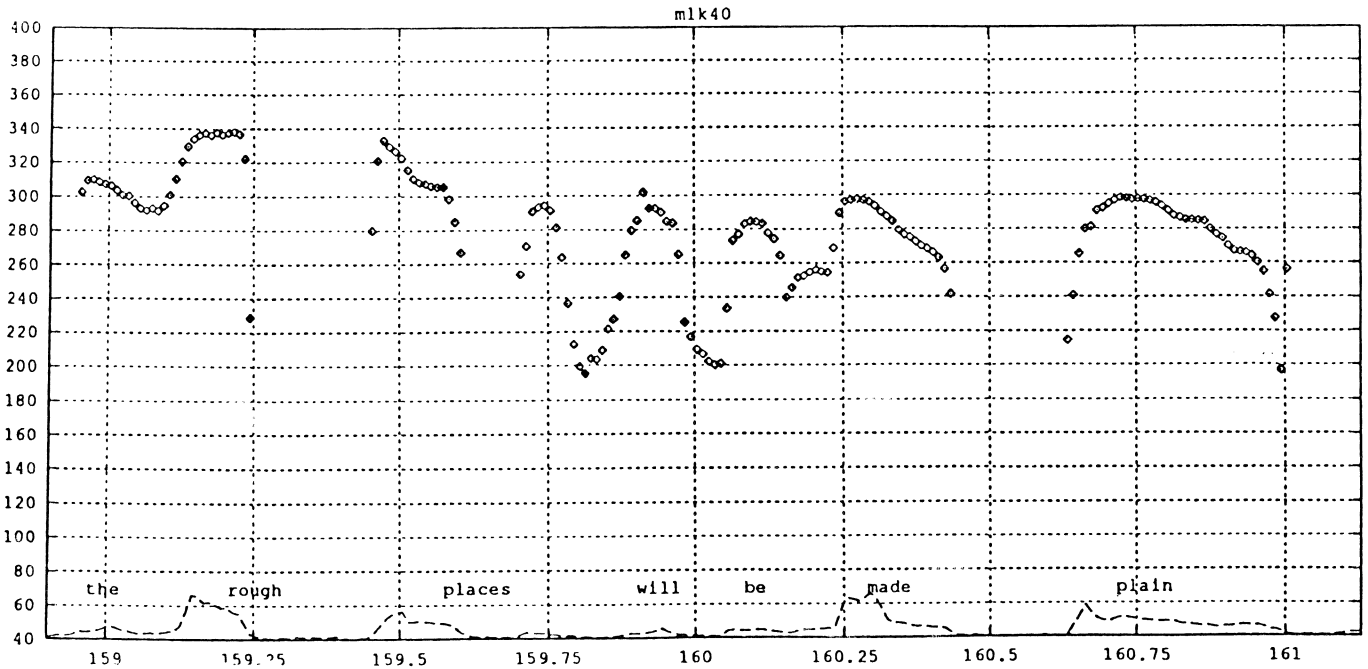

mlk9
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
and
live
out
the
true
meaning
of
its
creed
28
28.25
28.5
28.75
29
29.25
29.5
29.75
30
30.25

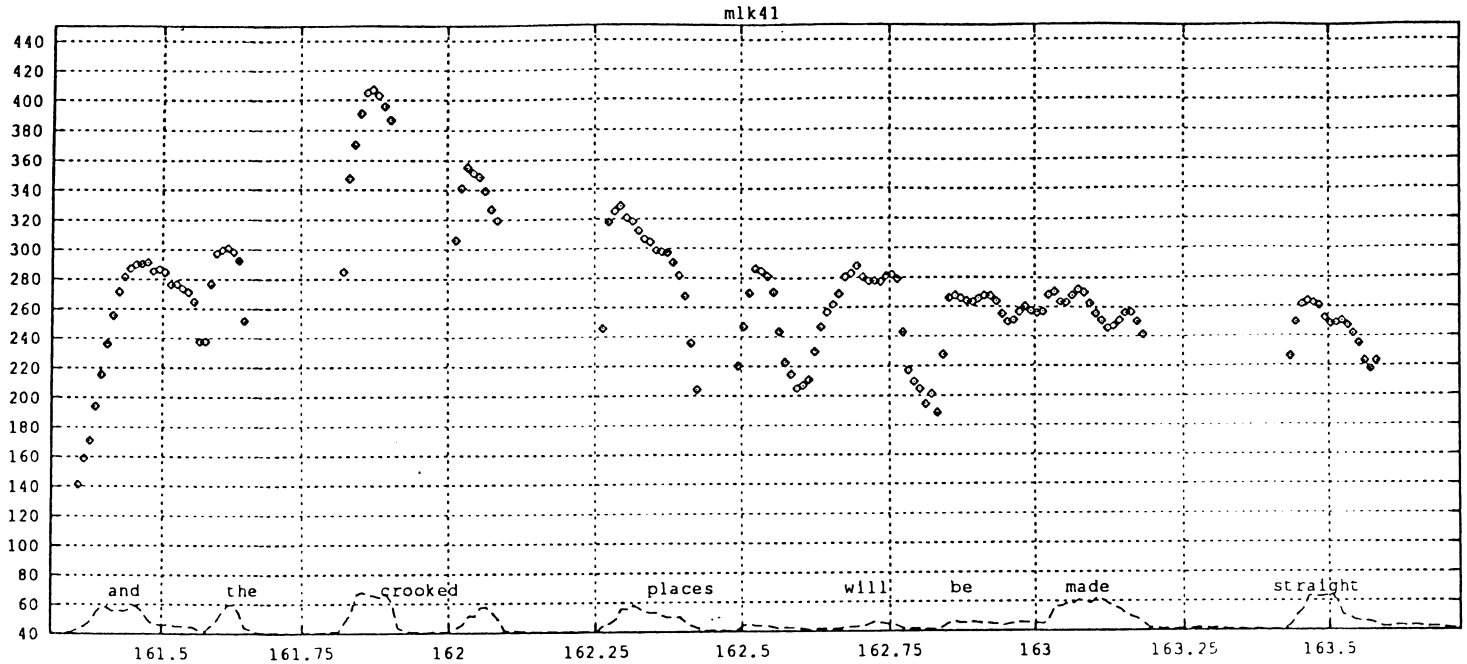

mlk16
300
290
280
270
260
250
240
230
220
210
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
I
have
a
dream
65
65.5
65.75

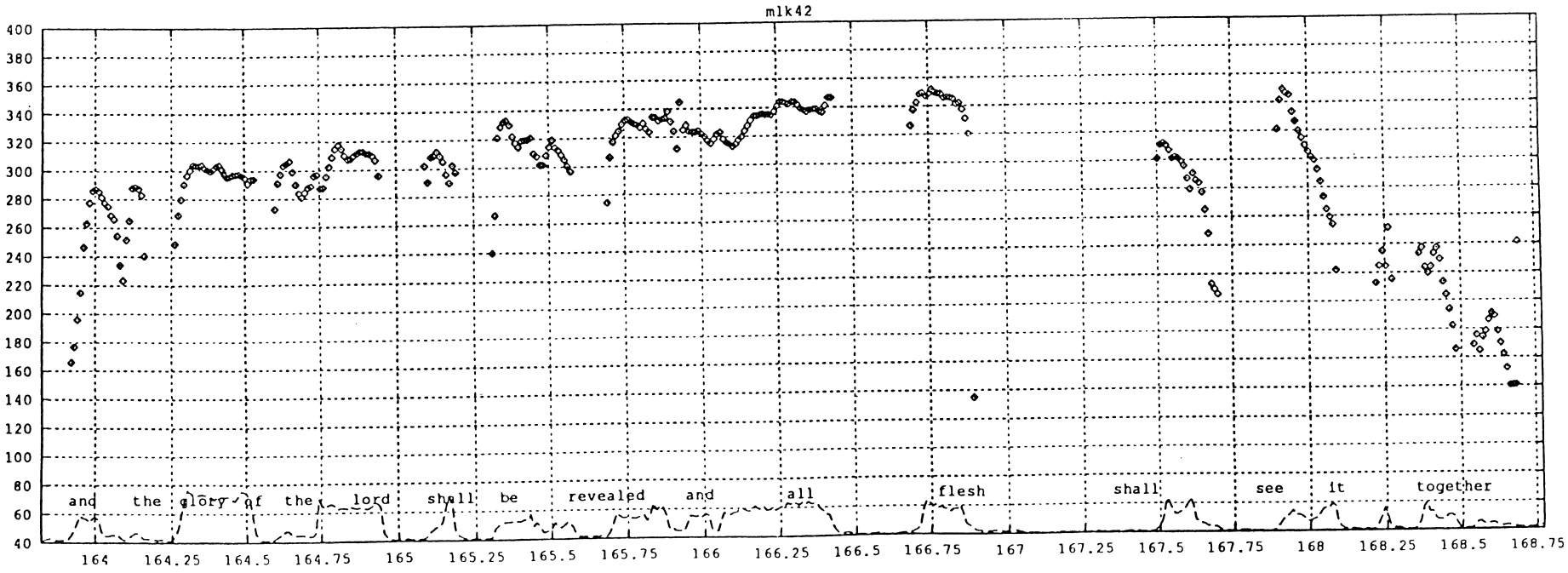

mlk24
that
my
four
little
children
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
91.5
91.75
92
92.25
92.5
92.75
93
93.25
93.5
93.75
94


mlk30
down
400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
118.25
118.5
118.75
119
119.25
119.5
119.75


mlk39
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
and
every
hill
and
mountain
shall
be
made
low
156.5
156.75
157
157.25
157.5
157.75
158
158.25
158.5
158.75


mlk40
400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
the
rough
places
will
be
made
plain
159
159.25
159.5
159.75
160
160.25
160.5
160.75
161


mlk41
440
420
400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
and
the
crooked
places
will
be
made
straight
161.5
161.75
162
162.25
162.5
162.75
163
163.25
163.5


mlk42
400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
and
the
glory
of
the
lord
shall
be
revealed
and
all
flesh
shall
see
it
together
164
164.25
164.5
164.75
165
165.25
165.5
165.75
166
166.25
166.5
166.75
167
167.25
167.5
167.75
168
168.25
168.5
168.75

Prosody and Information in Naturally-Occurring Discourse

R.T. Oehrle & M. Yaeger-Dror
Department of Linguistics, University of Arizona, Tucson, AZ 85721

We wish to construct an account of prosody which is both coherent with principles of grammatical analysis and responsible to naturally-occurring, contextually-situated speech. This setting provides a domain in which it is possible to test, refine, and extend theoretical hypotheses in the light of empirical data.

1. Definitions and preliminary observations

We assume (with Liberman [22]) that for any utterance u, it makes sense to regard its phonetic and phonological properties—jointly, *phon(u)*—as a combination of a lexical support *lex(u)* and a prosodic structure *pros(u)*. Motivation for this division of linguistic labor can be found in the distinct felt equivalences corresponding to the rows and columns below (using common orthographic conventions):

ED.	vs. ED?	vs. ED?!
ED spoke.	vs. ED? spoke	vs. ED?! spoke
Ed SPOKE.	vs. Ed SPOKE?	vs. Ed SPOKE?!

This point of view, while not the only possible one, raises two basic questions:

- how can we characterize the phonological/phonetic relation among the terms *lex(u)*, *pros(u)*, and *phon(u)*?
- how is the distinction between *lex(u)* and *pros(u)* related to other grammatical dimensions?

Both questions have attracted considerable attention. Since we do not characterize prosody in terms of particular phonetic or phonological properties (but rather in terms of the division of labor between prosodic and non-prosodic properties), the importance of the first question arises from the fact that the properties of *lex(u)* and the properties of *pros(u)* can (and do) affect the same phonetic parameters (frequency, intensity, duration), an entanglement between *phonetic realization* and *lexical/prosodic contribution* not peculiar to the point of view advocated here. With regard to the second question, it is a straightforward matter to see how *lex(u)* relates to other grammatical dimensions: any partition of an utterance into a sequence of morphosyntactic parts (such as words and affixes) must be consistent (in a way determined by the logic of phonological and phonetic realization) with the lexical support of the utterance. The contribution of these morphosyntactic parts to syntactic composition, semantic interpretation, and pragmatic force is a standard and central problem of linguistic analysis. It is equally straightforward to produce evidence that there are interesting relations between *pros(u)* and the properties of u in a number of other grammatical dimensions—evidence which supports the view that while *pros(u)* is not determined by the properties of any single non-prosodic dimension, the particular prosodic properties associated with an utterance impose constraints on its properties in other dimensions.

2. Theoretical desiderata

A theoretical framework for prosodic analysis will need certain properties.

2.1. Parallel architecture

To characterize relations between the phonological and phonetic properties of prosodic structures and their correlative properties in non-prosodic dimensions, we assume that the grammatical composition (or analysis) of complex expressions assigned properties in all dimensions simultaneously.

2.2. Structural flexibility

To accommodate the view that intonational phrases constitute units of prosodic analysis but need not correspond to standard syntactic constituents, a theoretical framework for prosodic analysis must make possible multiple analyses (in some sense) of a single utterance-type.

2.3. Dynamic discourse interpretation

To accommodate the fact that there is an interaction between the accentual properties of a constituent and the discourse information it conveys, it is reasonable to assume some form of dynamic interpretation, based on the insight that the interpretation of an utterance both depends on the context in which it is uttered and affects the very context as it is uttered.

2.4. Interactional dynamics

Spoken language provides interesting evidence of a highly structured relation between speaker (more generally, discourse participants), spoken code, social context, and interlocutors. This relation is reflected in the continuum of **style** [18, 19, 20, 34] ranging from self-consciously read materials to casual, unmonitored conversation—a scale characterizable as **speech-oriented** at one end and **task-oriented** at the other—and in the companion notion **register** [1, 2, 3, 5], which distinguishes register (from speech style) as the conventionally accepted way to speak in a specific situation (as due, for instance, to expectations within the culture or to specific audience design factors). It is also reflected in the relations between and purposes of the discourse participants, involving issues of power and solidarity [8], and distinctions such as **informational / social** interaction [35, 4] and **supportive / neutral / face-threatening** conversational acts [7]. Much of the interactional work done is conveyed through prosodic information. Labov and Fanshel [21] suggested that it may be the deniability of prosodic 'input' which predisposes speakers to use prosodic information for such socially sensitive tasks. To accommodate these distinctions, we will employ a terminology based on the proposals of [12, 13, 15, 7, 30].

2.5. An Integrated Framework

A natural formal setting which is consistent with these desiderata is the family of **multi-dimensional categorial grammars** [25, 23, 32], which incorporate **parallel architecture** in an essential way and permit different degrees of **structural flexibility** (depending on the details of the system in question). In this framework, each expression is identified with a 'vector' which characterizes the information associated with it in each relevant 'dimension'. This point of view allows the phenomena of **dynamic interpretation** and **interactional dynamics** to be integrated with other properties of linguistic composition.

3. Interpreting pitch-accent placement

The key concept of dynamic discourse interpretation relevant here is that the interpretation of an expression A in a context ϕ may both depend on ϕ and affect ϕ, which we represent as follows, using ϕ' to represent the resulting context:

$$\phi[\![A]\!]\phi'$$

This representation is heuristically useful: it allows a number of different existing theories to be formulated in a common frame and it suggests new lines of theoretical development. Two kinds of hypotheses will be considered here: first, the classification of expressions according to their interaction with contextual parameters; second, the classification of the discourse structures represented above by the variables ϕ and ϕ'.

A hypothesis which has guided a great deal of valuable work in the Generative Tradition is that pitch-accent placement can be characterized in a way that is indifferent to context. We believe that an adequate account of pitch-accent placement on this basis is unattainable.

3.1. Pitch-accent and information

A widely-held alternative view is that pitch-accents occur on 'new information' and do not occur on 'old information'. Formulating this theory in the simple framework sketched here requires a decision concerning the representation of 'information' and the 'old/new' contrast. A simple way to do this is to suppose that the information contained in a discourse $\mathcal{D}$ at a point x consists in a representation of the content of the discourse portion preceding x, together with a collection of 'discourse referents'. If A is an expression with non-dynamic interpretation a, we may represent the context-change potential of new information as $\phi[\![A]\!]\phi \sqcup a$ and the context-change potential of old information as $\phi \sqcup a[\![A]\!]\phi \sqcup a$ (The symbol $\sqcup$ stands for disjoint union of pieces of information.) To connect this account with **H*L** pitch-accents (say), we identify accented occurrences $\acute{A}$ of A with the structure $\phi[\![\acute{A}]\!]\phi \sqcup a$ and unaccented occurrences with the structure $\phi \sqcup a[\![A]\!]\phi \sqcup a$.

But this account cannot be correct. On the one hand, consider a narrative which begins: *once upon a time, there were two bears—'dum and 'dee.* 'DUM *was* OLD*er than 'dee* .... Although the second occurrence of *'dum* represents 'old information', it must be accented. On the other hand, consider a discourse in which one person rushes in to announce: GUESS WHAT! *my* BI*cycle's missing!* The bicycle's absence is new information in this context, but the expression associated with the introduction of this information into the discourse—namely, *missing*—need not be accented.

Intuitive judgments of this kind suggest that the relation between information structure and pitch-accent placement is more subtle than a simple dichotomy between new information and old information allows. If we assume that pitch-accent placement is interpreted relative to some informational domain, then there are two aspects to the problem: the first involves constraints on what components of an utterance may be taken to be prominent, relative to a fixed analysis, a given accentuation pattern, and a particular discourse context; the second involves how information is assumed to be structured in discourse. These two aspects of the problem suggest more plausible alternatives to the overly simple correlation of pitch-accents and information discussed above.

3.2. Pitch-accents and focus

The fact that pitch-accents are localized to syllables and not directly to some informational domain makes it necessary to characterize a relation between the syllables of any utterance and the linguistic structures whose content is relevant to that domain. In particular, for any expression e in a fixed context, a reasonable goal is to associate with each subset σ of the syllables of e a *focus set* of component parts of e which are the possible foci of e when every syllable in σ is accented. When the content of e itself is focused, we say that an utterance of e has a *wide-focus* interpretation. From this perspective, it makes sense to enrich the possible modes of interaction between expressions and context: expressions corresponding to simple types (such as proper names) may be treated exactly along the lines sketched above; but functor categories may be classified according to how they interact with the context-change potential of their arguments. For example, some one-place predicates may behave in a way that correlates the new/old contrast with the presense/absence of accent, but for other one-place predicates, it is possible that if the argument of a member of this class counts as new in a given context, the syllable within the argument which is (when accented) compatible with a wide-focus interpretation of the argument may also be (when accented) compatible with a wide-focus interpretation of the predicate-argument combination (even though the predicate itself is unaccented but new); but when the argument does not count as new, the preferred syllable for indicating a wide-focus interpretation may shift to the most prominent syllable within the predicate itself. Such an account ([26], for example) makes possible a more sophisticated account of functors like *be missing* than the simple correlation between accent and information discussed earlier. Moreover, the fact that certain functors need not be accented when their co-domain supports a wide-focus interpretation means that in many contexts, the choice of accenting them or not is accessible to pragmatic influences. (We consider one such case—the case of negation—in detail below.)

A complete account of the interpretation of pitch-accent placement depends not only on the relation between pitch-accent placement and the information-structure of particular utterances, but also on how that structured information connects with context. Another direction of research we hope to pursue further in the light of empirical investigation of naturally-occurring discourse is the possibility of endowing the theoretical representation of discourse context with richer structure. For example, an account along the lines of the centering model of Grosz, Joshi, and Weinstein [16] makes it possible to treat only topics as contextually de-accented. This richer articulation of discourse context (which has other advantages, as well) makes it possible to deal with cases like the *two-bears-'dum-and-'dee* example above, where an expression representing old information is obligatorily stressed when a topic shift is involved.

4. Quantifiable parameters

4.1. Pitch-accent 'prominence' measure

The hypotheses to be formulated are based on an assumed connection between subjective impressions of pitch-accent placement and an informational domain. Many linguists who have quantitatively analyzed linguistic data have found a correlation between 'focal' or 'new' information in a discourse and physical parameters such as duration [9, 10, 11] or pitch prominence [24]. Integration of these two perspectives—one intuitive and abstract, the other quantitative and concrete—requires the establishment of a correspondence between physical parameters and subjective impressions of prominence, on the one hand, and a common view of the informational domain. Although accent is acoustically produced with both pitch prominence and increased vowel duration and peripherality, the defining criterion for accent in our study is pitch prominence only. The primary motivation for this decision is that since our data are to be compared with the results found in earlier acoustic studies of negation (and to be compared with the algorithm proposed by Hirschberg [17] for synthesis) where the only criterion for prominence was determined by the pitch, pitch prominence is the sole criterion to be used here. Because the earlier studies did not clearly define their criterion for the determination of 'prominence', we use a very broad rule, to permit even a limited 'focal prominence' to be included: a token can be considered pitch prominent if the fundamental frequency on the vowel is raised relative to the fundamental frequency of immediately adjacent words.

4.2. Negation and disagreement

This study will concern itself primarily with the analysis of negatives in a discourse, and how they are realized intonationally. The study of the prosodic aspects of negation in discourse has given rise to two traditions with conflicting claims. On the one hand, some researchers who have analyzed negatives have found that negatives are realized with pitch prominence, and have attributed this finding to a correlation between negatives and 'new' or 'focal' information. We refer to this correlation as the linguistic **Focal Prominence Rule**. For example, O'Shaughnessy and Allen [29], in a study of read sentences, found that pitch prominence occurred on negatives, even when they were contracted. Hirschberg [17], who analyzed the speech of NPR announcers to determine a reasonable algorithm for synthesis, initially assumed that 'closed class words' should be unstressed, but concluded that negative bearing elements, even though closed class, should bear pitch prominence.

On the other hand, Schegloff, Jefferson and Sacks [31] presented evidence that in conversational speech, there is a 'preference for agreement', to which speakers adapt their speech.

We will refer to this as the **Agreement Rule**. The Agreement Rule would predispose speakers to use pitch prominence and a durational increment on negatives used for agreement, or used in a neutral-informational setting, but would predispose speakers to use a neutral pitch, and durational reduction (including contraction) on a disagreement or in the course of performing a face-threatening act in the context of a face-enhancing interaction. Both Yaeger-Dror [35] and Tottie [33] have found that pitch prominence is relatively rare on negative elements in actual discourse.

A particular goal of the present study is to determine the interaction of the Focal Prominence Rule (or other accounts of the relation between prominence and information) and the Agreement Rule. Of course, the simplest pattern occurs when the Agreement Rule can be neutralized. In interactionally neutral settings, where the negative is used informationally, the Agreement Rule is most likely to be neutralized, and the Focal Prominence Rule is dominant. Table 1 provides some examples of other possibilities as well:

Table 1. Interaction of interactional intent and pitch prominence, their relationship to the Focal Prominence Rule (FPR) and Agreement Rule (AR).

	+Prominent	**-Prominent**
Neutral	FPR dominates AR neutral	FPR contradicted AR neutral
Face threat (FTA)	FPR dominates AR contradicted	FPR contradicted AR dominates
Supportive exchange	FPR dominates AR dominates	FPR contradicted AR contradicted

One potential reason for variation in pitch prominence is related to the interactive intent, as shown on Table 1. On the one hand, Yaeger-Dror [35] and Tottie [33] have both shown that if the negative is used to agree rather than disagree, this is what has been called a supportive interchange, and prominence is most likely to occur because both rules favor pitch prominence in this case. In contrast, Yaeger-Dror & Nunamaker [36] showed that even in read dialogue, pitch prominence is least likely to occur when a statement is theoretically face threatening. Table 1 shows that in this case, the Agreement Rule is seen to 'overrule' the FPR.

4.3. Specific hypotheses to be tested

The hypotheses we wish to test, then, are the following:

Focal Prominence Rule:

- Pitch prominence is to be expected on a negative which supplies new information.

Agreement Rule:

- Prominence is to be expected on a negative which supplies an agreement with an earlier speaker.
- Pitch prominence is to be avoided when a possible interpretation could be a disagreement with a previous speaker within an interaction—except in specifically 'licenced' situations, where face threats are to be expected (e.g., debates, arguments between intimates).

We may test these hypotheses against the properties of naturally-occurring data in two ways. First, when the prosodic structures occur in the data chosen, do the data conform to the properties of these hypotheses? Second, do the hypotheses give a broad enough account of the prosodic structures that occur in the data. At the same time, it is necessary to consider what sort of data forms the most appropriate testing ground for hypotheses of the kind considered above.

4.4. Negation and focus structure

One other possible reason for a lack of pitch prominence is related to the specific syntactic focus intended, as shown on Table 2. In line with an understanding that closer attention to syntactic focus might differentiate between possible strengths of contradiction between the two rules, Table 2 shows that in fact, if there were narrow focus on some other word in the sentence, the FPR would be neutralized. The least likely locus for pitch prominence on negatives would thus occur in statements where there is narrow focus on another word; the most likely, would be in sentences with narrow or contrastive focus on the negative itself. Sentences with wide-focus interpretations might fall in between.

Table 2. Interaction of wide vs. narrow focus and pitch prominence, and their relationship to the Focal Prominence Rule (FPR) and Agreement Rule (AR), in the case of a face threatening act.

	+Prominent	**-Prominent**
Face threat		
Narrow focus on negative	FPR dominates AR contradicted	FPR contradicted AR dominates
Narrow focus on other word	FPR neutral AR neutral	FPR neutral AR neutral
Wide focus	FPR dominates AR neutralized (partly)	FPR contradicted AR dominates

The present study will attempt to determine the degree to which the overtly face threatening material will be pitch prominent, and the degree to which that prominence can be neutralized either by scope considerations or by interactional rules.

5. Choice of corpus

Data were collected from the following registers:

- informational data (from 'news' and from a tutorial)
- interactional data (from the political debate, most of which is face-threatening)
- Neutral & Supportive data will be cited using data from earlier studies.

Since self-conscious style does not always influence speech in a clearly defined way [34, 20], we considered it important to choose a corpus from speech which would be less self-conscious than read sentences and would be 'task-oriented' rather than 'speech-oriented'. To maintain consistency of register with many other recent analyses of naturally-occurring speech [17], we have chosen corpora broadcast over PBS, and will refer to this as *NPR-speak*. The primary corpus under discussion is a tape of a political discussion originally broadcast over the MacNeil-Lehrer report, generously provided by Karen Adams of ASU. It is reasonable to classify this material as *careful* in style, *NPR-speak* in register, and *confrontational* in interactive intent.

One of the rationales for using an NPR-speak corpus is to neutralize the vectors of power and solidarity, by assuring that the difference in power between speakers is minimal, and that the speakers are not too intimate ('solidary').

The interactional intent of most examples in such a corpus can be fairly easily distinguished. For example, there are very few 'neutral' factual uses of the negative in the segment transcribed here. Although there are few supportive uses of the negative in a debate, one occurs on line 24 (of the appended transcript). Most of the other negatives here (which occur in lines of slanted type in the transcript) are clearly face threatening to the other politician in the interaction; as noted above, face-threatening acts appear to be 'licensed' in this type of social situation. This we conclude from the data themselves.

6. Linguistic variables

The primary linguistic focus will be the prosodic and linguistic realization of negatives found in discourse.

6.1. Pitch variation

The primary goal of this paper is to determine the relative frequency of focal prominence occurring on a negative; in the process, it will be possible to compare the importance of the Focal Prominence Rule and the Agreement Rule. In cases where the Focal Prominence Rule is not dominant (i.e., where there is not pitch prominence on the negative) we will then determine, for a restricted segment of the corpus, whether scope considerations offer an account of the apparent anomaly.

As stated earlier, a token can be considered pitch prominent if the fundamental frequency on the vowel is raised relative to the fundamental frequency of immediately adjacent words. Often a pitch prominent token will also be produced with a pitch contour; for present purposes, if the pitch is raised on the vowel of *not*— as on lines 32, 57— or, in a contracted case, on the vowel of the auxiliary onto which the negative is contracted—as on lines 5, 11, 12, 19, 21, 23, 24, 54, and 55 (twice)—this will be considered evidence of pitch prominence, even if there is no contour. If there is a contour in the vowel, even if pitch is not higher than on the immediately preceding vowel, this is also categorized as 'pitch prominent' (line 21). If there is no pitch prominence—as in line 9 and the first example in line 12—this will be referred to as a 'neutral' pitch. It is also theoretically possible that cases would occur in which there would be negative prominence (that is, pitch lowering) on a token, as proposed by Bolinger [6]; however, we did not find any cases of this type of prominence. Specific techniques for analysis of pitch prominence will be explained in greater detail below.

If the pitch is nonprominent, and the negative is contracted, this corresponds to the assumptions of the Agreement Rule (for polite interactions), and minimizes any 'face threat' which might result.

6.2. Technique for the acoustic analysis

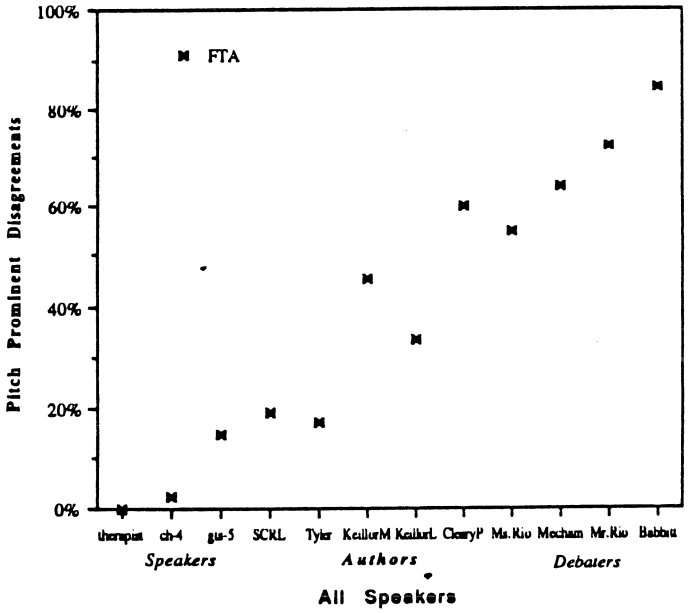
Segments from each of the tapes were run through either MacSpeechLabII[1] or Signalyze 1.1, using a Macintosh IIci with a MacRecorder interface. Note that this 8 bit A-D interface for the analysis of fundamental frequency was sufficiently clear for the high quality recordings. If the trackable pitch is no higher than the immediately preceding pitch,[2] and does not have any contour (as on line 9), then it is not categorized as pitch prominent. As described above, if a local pitch prominence occurs (even if that prominence is not the pitch peak for the entire sentence) the negative is classified as prominent. While this decision criterion clearly ignores other forms of prominence (amplitude and duration), it accurately compares the data with the template which would be provided

[1]Kerry Green and Tom Bourgeois of the University of Arizona Speech Research Lab were kind enough to give us access to their facilities.

[2]Note, not just 'not higher than the nearest local maximum', but 'not higher than the immediately preceding vowel'.

by the synthesis algorithm.

Figure 1.
Pitch Prominence in FTA-Disagreements



6.3. Quantifying the prominence × FTA results

Figure 1 presents the results of such an analysis, comparing casual conversational speech (reported in [35]), with readings of books and Keillor monologues (reported in [36]), and with the pitch tracks for the present corpus of political debates. If these results are representative, it is clear that therapeutic style and casual conversational style require the lowest percentage of pitch prominence in face threatening disagreements (between 0% for a therapist, and 18% for teenaged therapy patients with casual conversations in between). Data gathered from a DARPA tutorial shows that even there, face threatening disagreements are pitch prominent only 20% of the time. In read (adult) dialogue, the percentage is only somewhat higher, although in read dialogue-of-children, up to 60% can be pitch prominent. This is consistent with our understanding that in conversation, or even simulated conversation, the AR is stronger than the FPR for socialized adults. However, even the kiddie-FTA's have lower percentages than the political debates used for the present corpus, which range from 60-85% pitch prominent. We draw the conclusion that pitch prominence is licenced in this debate register, but that even in this register, the FPR does not account for 100% of the data: while the political debates have a much higher percentage of pitch prominent negations than the other corpora, 20-40% of the negations are still unaccountably non- prominent. In the next section, we will consider whether incorporating the scope into the analysis as a parameter provides an explanation for the 20-40 % gap between the FPR 'target' and the debate's negative realizations.

6.4. Quantifying the prominence × scope results for a subcorpus

In what follows, we examine each occurrence of negation in the segment of the transcript found in Appendix I and score it for various parameters: pitch [prominent or non-prominent], negation [old or new], focus [wide or narrow], accent [optional or obligatory], interaction [neutral, speaker-enhancing, addressee-enhancing, addressee-threatening]. In each case, we assess whether this profile is consistent or inconsistent with the Focus Prominence Rule and the Agreement Rule of §4.3. The lines discussed may be found in the appended transcription. Pitch-tracks of these lines may be found in Appendix II.

line 5: *No, it isn't a matter of whether I have regrets ...*

pitch	prominent
negation old/new	new
focus	narrow
accent	obligatory
interaction	self-enhancing

This is consistent with both the FPR and the AR.

line 9: *the governor does not have the power to create a state holiday.*

pitch	nonprominent
negation old/new	new
focus	wide
accent	optional
interaction	self-enhancing

This is inconsistent with the FPR (since negation is new but not accented), but consistent with the AR.

line 11: *You can't say there's any regrets.*

pitch	prominent
negation old/new	new
focus	narrow
accent	obligatory
interaction	self-enhancing

Consistent with both the FPR and AR.

line 12: *It isn't anything that—*

pitch	non-prominent
negation old/new	old?
focus	wide?
accent	optional
interaction	self-enhancing

Consistent with both FPR and AR. Note, however, that

the non-prominence of negation here is immediately self-corrected: see the next example.

line 12f.: *it isn't anything of my doing*

pitch	prominent
negation old/new	old
focus	wide
accent	optional
interaction	self-enhancing

Inconsistent with FPR (since negation is old but nevertheless prominent), but consistent with the AR. [Note that this is a self-correction of the immediately preceding fragment.]

line 19: *roughly 25% don't want a state holiday*

pitch	prominent
negation old/new	new
focus	narrow
accent	obligatory
interaction	pseudo-neutral

Consistent with both FPR and AR.

line 21: *some don't care*

pitch	prominent
negation old/new	new
focus	wide
accent	optional
interaction	enhances spkr's positive face

The optionality of prominence here is inconsistent with FPR and supports AR.

line 23: *isn't quite correct*

pitch	prominent
negation old/new	new
focus	wide
accent	optional
interaction	enhances spkr's positive face

The optionality of prominence here is inconsistent with FPR and supports AR.

line 24: *I don't think we meant to suggest that*

pitch	prominent
negation old/new	new
focus	narrow
accent	obligatory
interaction	face-enhancing to addressee i.e., supportive interchange

Consistent with both FPR and AR.

line 32: *I've not got into being concerned about that.*

pitch	prominent
negation old/new	new
focus	[complex]
accent	obligatory
interaction	on-record fta

Consistent with both FPR and AR.

line 54: *But he can't run away from the issue.*

pitch	prominent
negation old/new	new?
focus	narrow
accent	obligatory
interaction	on-record fta

Whether this is consistent with the FPR depends on whether we count negation as new (consistent) or old (inconsistent); consistent with AR. The rhetorical effect here goes beyond the distinctions that our parameters make.

line 55a: *he doesn't support*

pitch	prominent
negation old/new	new
focus	wide
accent	optional
interaction	on-record fta

The optionality of prominence here is inconsistent with FPR and supports AR. The rhetorical effect here goes beyond the distinctions that our parameters make.

line 55b: *and doesn't want*

pitch	prominent
negation old/new	old?
focus	wide
accent	optional
interaction	on-record fta

Inconsistent with FPR (since negation is old and pitch is prominent); consistent with AR. The rhetorical effect here goes beyond the distinctions that our parameters make.

line 57: *It's not political.*

pitch	prominent
negation old/new	new
focus	narrow
accent	obligatory
interaction	on-record fta

Consistent with FPR and AR.

Table 3. Focus and prominence.

	+Prominent		-Prominent	
	New	Old	New	Old
Narrow focus	6	0	0	0
Wide focus	4	1	1	1

7. Conclusion

The subcorpus was chosen, first, to neutralize the AR, second, to minimize cases of nonprominent pitch, and third, to determine whether a more complex account of the scope of focus can help to explain the cases where nonprominent pitch still occurs. The evidence supports the conclusion that the sentences which do permit non-prominent negations in the debate were those with the wide focus. We look forward to the opportunity of testing this conclusion on a larger corpus.

Appendix I

MacNeil/Lehrer interview with Evan Mecham and Bruce Babbit

Robin MacNeil:

In view of the fact that some Arizonans, at least, are unhappy about this decision are you—do you have any regrets about having rescinded the holiday, and are you reconsidering.

Evan Mecham:

No, it isn't a matter of whether I have regrets. It's a matter that as I see my responsibility it's to be respective of the law. Bruce and I, perhaps, have a difference of opinion here, but my attorney general tells me that this—*the governor does not have the power to creat a state* holiday. And consequently, I acted in the only rational *and responsible way that I can do so. You can't say that there's any regrets. It isn't anything that—it isn't anything of my doing. I just came into a situation that it* was my responsibility to correct. I noted, of course, that in the presentation as I watched here earlier says that all of these people feel differently about that, and yet the results of a poll here that was just published over the weekend points out that 25% of the people, or roughly a *quarter want a state holiday, and roughly 25% don't want a* state holiday. And then the others in the middle, some *don't care, and some says well it'd be nice to have some*thing. So I think the representation that everybody is *opposed to what I've done isn't quite correct.*

Robin MacNeil:

I don't think we meant to suggest that. We just meant that some are opposed to what you've done. Is your—Was your objection strictly a legal one, governor, though. I mean there have been a wide variety of things you've been quoted as saying. For example, that Martin Luther King was not of the stature of Washington or Lincoln and thereforde didn't deserve a birthday holiday like them. Did you say that, and do you believe that?

Evan Mecham:

I've not got into being concerned about that. I—This issue really has, you know, been blown up by others. My primary concern has been the fact that as for my responsibility it is to correct a thing that could be a sticky issue. We've got a—Some people feel that we have a holiday, or we did have in this state. And it's been my responsibility to correct that. Some have said, well, why don't you let it go to the court. Well, that's not a responsible action, either. I feel that Bruce—I might say my friend Bruce, because we're friends, although we're political—we have political differences. I think he acted in a totally political manner of which to do this. And I think that he as an attorney, I think he as an former attorney general, he knows the law. He knows how to read it. I think he could read the same thing into it. I'm sorry that he's started this great controversy. But it was up to me to take the only responsible action that was left to me to do.

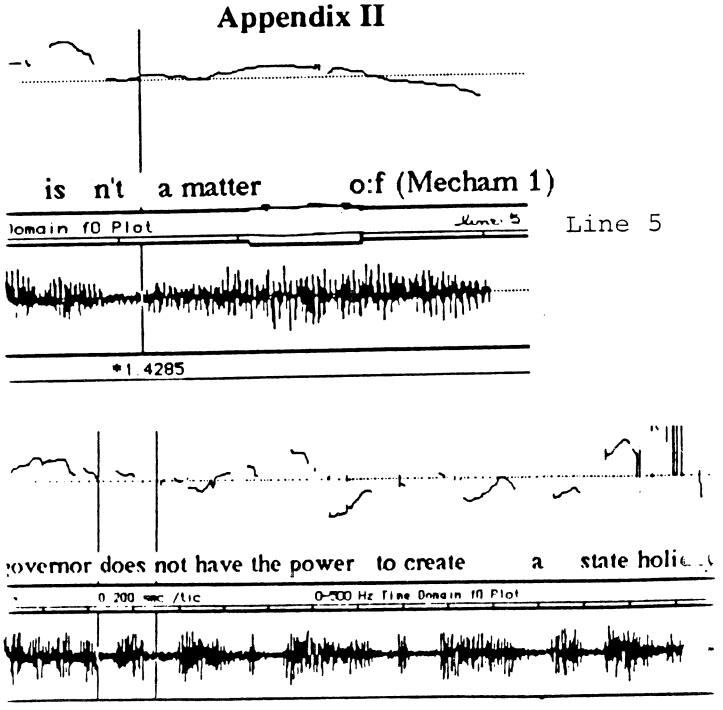
Robin MacNeil:

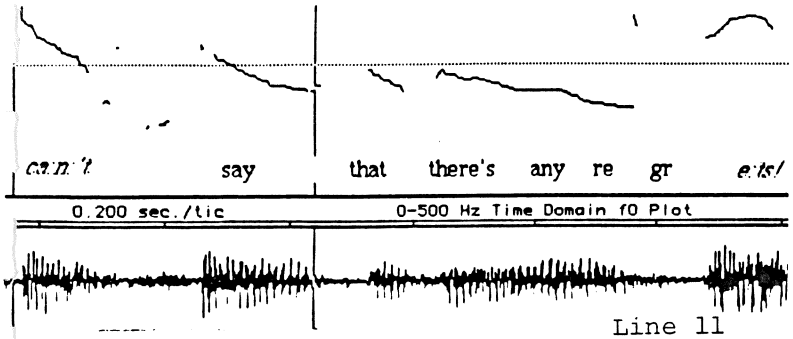
Governor Babbitt, or former Governor Babbitt, the present governor says what you did was just illegal.

Bruce Babbitt:

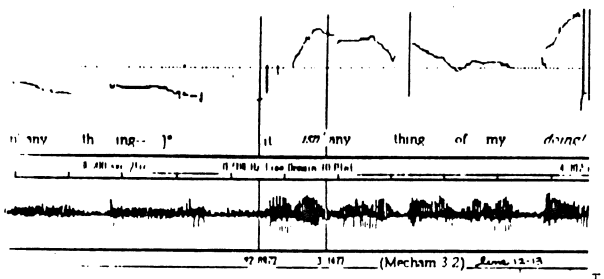
Well he can hide behind the lawyers. And there are a lot of lawyers on both sides of this issue. He can hide behind *them. But he can't run away from the issue. And that is, he doesn't support and doesn't want and is using his power* to thwart and oppose a Martin Luther King holiday. I sup*port it. It's not political. My involvement in this issue* began in Selma, Alabama, in 1965. It's continued to this day. Martin Luther King is a symbol of what America is all about. Of the ability to triumph over discrimination, over deprivation. It's the American story. It's a great symbol. I believe it ought to be a holiday. And the plain fact is that he's using his office to prevent it.

Appendix II

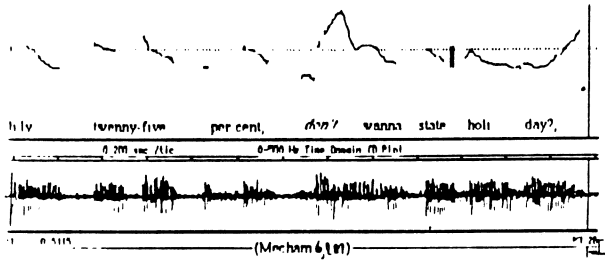




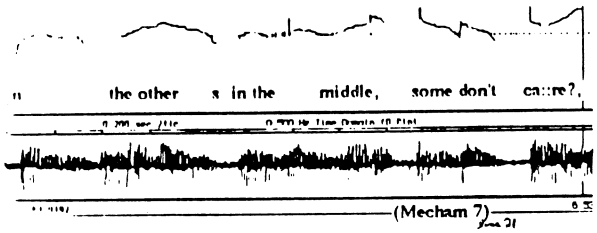
Line 11



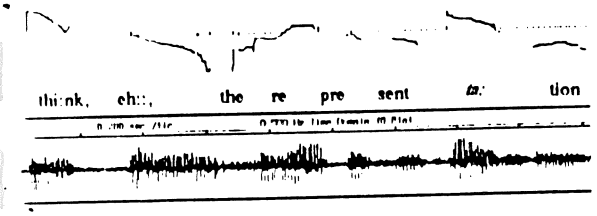
Line 12



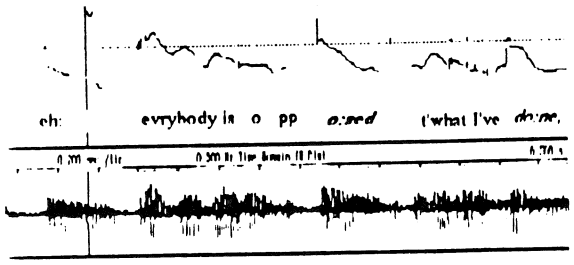
Line 19

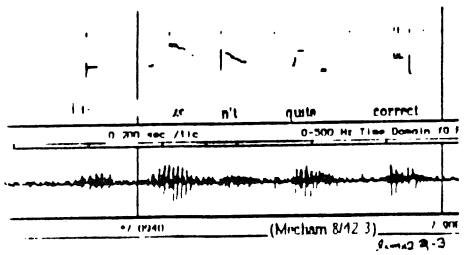


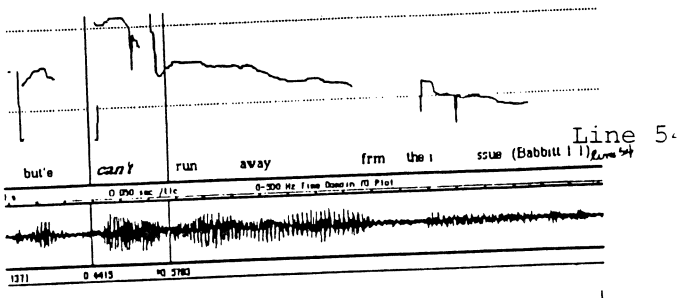
Line 21



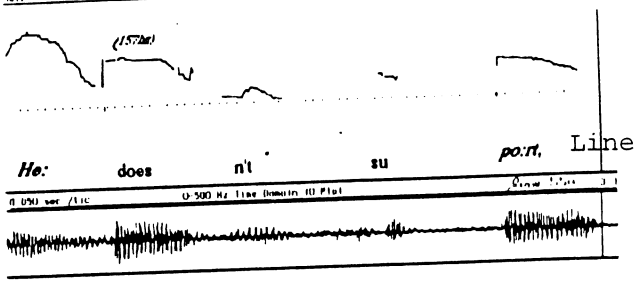
Line 22-



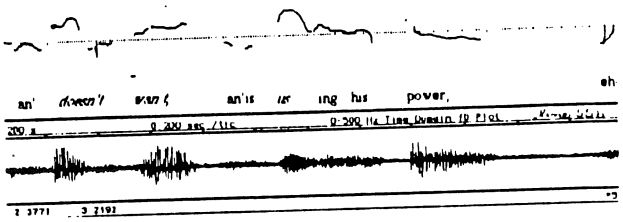


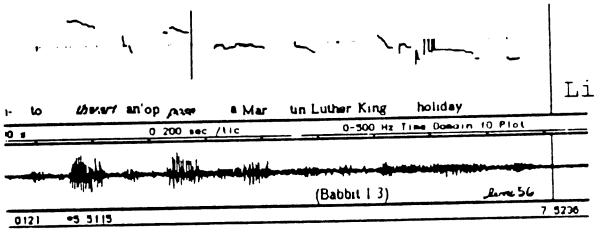


Line 5



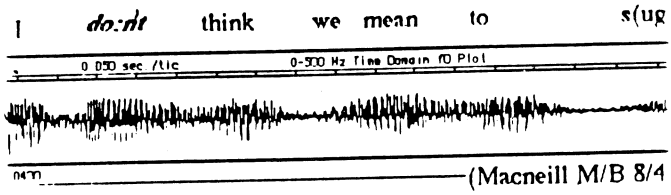
Line 5

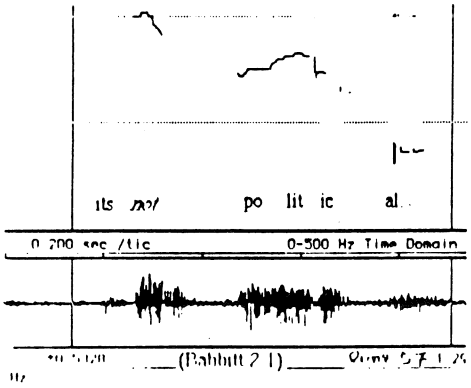




Line 5

Line 2





Line 5

References

[1] Bell, A. 1984. Language style as audience design. *Language in Society*. 13, 145-204.

[2] Bell, A. 1991a. Audience of accommodation in the mass media. In Giles, H., N. Coupland and J.Coupland (eds) *Contexts of Accommodation*, pp. 69-102. Cambridge: CUP.

[3] Bell, A. 1991b. *Language in the News Media*. Oxford: Blackwell.

[4] Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: CUP.

[5] Biber, D. & E. Finegan 1992. Parallel patterns in social dialect & register variation: Towards an integrated theory. In Biber, D. & Finegan, E. (eds), *Perspectives on Register: Situating Register Variation within Sociolinguistics*. Oxford: Oxford University Press.

[6] Bolinger, D. 1978. Intonation across languages. In Greenberg, J. (ed), *Universals of Human Language* vol. II, pp. 471-524. Stanford: Stanford University Press.

[7] Brown, P. & S. Levinson 1978. Universals of language usage: Politeness phenomena. In Goody, E. (ed.), *Questions & Politeness: Strategies in social interaction*, pp. 56-289. Cambridge: CUP.

[8] Brown, R. & A. Gilman. 1960. The pronouns of power and slidarity. In T.A. Sebeok, ed., *Style in Language*, Cambridge, Mass.: MIT Press, 253-276.

[9] Coker, C. & N.Umeda 1971. Toward a theory of stress & prosody in American English. *Proceedings of the 7th Intl. Congress of Acoustics*, pp. 137-140. ...Budapest.

[10] Fowler, C.A. 1988. Differential shortening of repeated content words produced in various communicative contexts. *Language & Speech* 31, 307-19.

[11] Fowler, C.A. & J. Housum 1987. Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *J. Mem.Lg.* 26, 489-504.

[12] Goffman, E. 1967. *The Presentation of Self in Everyday Life*. Harmondsworth: Penguin.

[13] Goffman, E. 1971. *Relations in Public*. NY: Harper & Row.

[14] Goffman, E. 1981. *Forms of Talk*. Philadelphia: UPenn Press.

[15] Grice, H.P. 1990. *Studies in the Ways of Words*. Cambridge, Massachusetts: Harvard University Press.

[16] Grosz, B., A. Joshi & S. Weinstein 1986, ms. Towards a computational theory of discourse information. CIS, Harvard University & University of Pennsylvania.

[17] Hirschberg, J. 1990. Accent & discourse context: Assigning pitch accent in synthetic speech. *Proceedings of the Eighth National Conference on Artificial Intelligence, II*, pp. 952-57. Cambridge: MIT Press.

[18] Labov, W. 1966 [1983]. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

[19] Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

[20] Labov, W. 1986. Sources of inherent variability in the speech process. In J.S. Perkell & D.H. Klatt, eds., *Invariance and variability in speech processes*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 402-423.

[21] Labov, W. & D. Fanshel. 1977. *Therapeutic Discourse*. New York: Academic Press.

[22] Liberman, M.Y. 1975. *The intonational system of English*. MIT dissertation; reproduced [1978] by IULC.

[23] Moortgat, M. 1989. *Categorial investigations: logical and linguistic aspects of the Lambek calculus*. Dordrecht: Foris.

[24] Nooteboom, S.G & J.G. Kruyt 1987. Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America* 82, 1512-1524.

[25] Oehrle, R.T. 1988a. Multi-dimensional compositional functions as a basis for grammatical analysis. In Oehrle, R.T.et al. (eds.), *Categorial Grammars & Natural Language Structures*, pp. 349-389, Dordrecht: D. Reidel.

[26] Oehrle, R.T. 1988b. Sources & structures of linguistic prominence in English.In Schiffer, S. & Steele, S. (eds.), *Cognition & Representation*, pp. 209-241. Boulder: Westview Press.

[27] Oehrle, R.T. 1991a. Prosodic constraints on dynamic grammatical analysis. In S. Bird, ed., *Declarative Perspectives on Phonology*. Edinburgh Working Papers in Cognitive Science. Volume 7. pp. 167-195. Centre for Cognitive Science. University of Edinburgh.

[28] Oehrle, R.T. 1991b. Grammatical structure and intonational phrasing: a logical perspective. In the working papers prepared for the 1991 AAAI Fall Symposium on Discourse Structure in Natural Language Understanding, Asilomar, November, 1991.

[29] O'Shaughnessy, D. & J. Allen 1983. Linguistic modality effects on fundamental frequency. *JASA* 74, 1155-1171.

[30] Sacks, H. 1992 (in press). *Harvey Sacks' Lectures on Conversation*, 2 volumes, edited by Gail Jefferson. Oxford: Blackwell.

[31] Schegloff, E. G. Jefferson & H. Sacks 1977. Preference for self-correction in the organization of repair in conversation. *Lg*. 53, 361-82.

[32] Steedman, Mark. 1991. Stucture & intonation. *Lg*. 67, 260-96.

[33] Tottie, Gunnel. 1991. *Negation in English Speech and Writing*. Pantopia: Academic Press.

[34] Yaeger, M. 1974. Speaking style: some etic realizations and their significance. *Pennsylvania Working Papers on Linguistic Change and Variation* I, 1.

[35] Yaeger-Dror, M. 1985. Intonational prominence on negatives in English. *Lg & Speech* 28, 197-230.

[36] Yaeger-Dror, M. & J. Nunamaker. 1992. Negatives and Register. *Journal of the Acoustical Society of America*, 91, S3288(A).

PROSODIC ORGANIZATION IN THE SPEECHES OF MARTIN LUTHER KING

Robin M. Queen
Linguistics Dept.
University of Texas at Austin

0. INTRODUCTION

The public speeches of Martin Luther King Jr. present an interesting juncture for the study of language as it pertains to culture (and vice versa) because of King's unique place within the cultural history of the United States and within the African-American community. There have been many studies of King's rhetorical style as well as the political and social implications of the content of his speeches; however, there has been very little work done in which the actual linguistic devices which he uses have been clearly identified and described with respect to both distribution and interpretation. This paper offers a first and preliminary account of certain aspects of King's language, with particular emphasis on his use of prosodic tools as a method of discourse organization and cultural reference.[1]

1. METHODOLOGY AND THEORETICAL FRAMEWORK

I have chosen an inductive approach to analyzing the speeches of Martin Luther King Jr. By this method, the patterns of his speech become apparent independently. Furthermore, it allows for more general interpretations of form which may occur within individual texts but also across the corpus as a whole.

The actual corpus of data was obtained from a sample of King's speeches. An audio recording was made from the original videotape *The Speeches of Martin Luther King Jr.*. (distributed by MPI home videos). The entire corpus was then transcribed, with each line corresponding to a pause boundary. A representative sample set of speeches was then chosen for specific analysis. Speeches were chosen based on a loose set of criteria which included length of the speech and setting in which the speech appeared. Once the sample set was completed, each speech was digitized using the equipment in the Phonetics Laboratory at the University of Texas at Austin, and subsequently, the sample set was pitch tracked using the linear predictive correlation pitch tracker of the Klattlamm program developed by Denis Klatt. In representing specific intonational contours, I will be drawing on the following set of symbols:

? indicates a phrase-final rise

– indicates a phrase-final level

# indicates a phrase-final fall

indicates a elongated fall

∞ indicates a phrase-initial rise and

an underlined string indicates a prominent stress.

1.1 Theories of prosody

Several theories have been employed for the purpose of the analyses of the speech of

[1]This study is largely the result of a course taught by Anthony Woodbury at the Univeristy of Texas at Austin in the Spring of 1992. I would like to thank Dr. Woodbury as well as Keith Walters, Troi Carleton, Mark Hewitt, John Kaufmann and Mark Louden for helpful comments during the course of this study. However, any errors in fact or interpretation are my own.

MLK, with the central analytic work being done through theories of phonetic implementation and the morphology of intonation. The theory of the phonetic implementation of intonation operates within the general framework of generative linguistics in which the elements of a finite grammar can be combined in such a way as to generate an infinite set of possible well-formed intonational tunes. The development of a finite grammar of intonation is largely due to early work by Mark Liberman (1975) and to work by Janet Pierrehumbert (1980) in which intonational contours are a result of the interpolation between a very small set of possible tones, basically H(igh) and L(ow).[2] These two tones align to phrasal units in one of three ways-as boundary tones (T%), phrase accents (T), (tones which fall between a boundary tone and a tonal accent), and pitch accents (T*) (tones aligned to stressed syllables). The rules of interpolation, as specified on a language specific basis, account for the movements of pitch between tonal accents, and are often specific according to the pragmatics or discourse structure.[3] .

Additionally, the theory of the representation and interpretation of intonation proposed in the current work of Cynthia McLemore, Mark Liberman and Anthony Woodbury provides the backbone for this analysis. On this theory, the interpretation and use of given linguistic forms (most notably intonation and other prosodic phenomenon) is partially specified by the particulars of culture and/or context. McLemore (1991) showed that the specific boundary tones used among sorority members (a socially homogeneous group) obtained their "meaning" partly through generalized interpretations and partly through their interaction with the specifics of context and cultural fine-tuning. Through this interpretation, prosody may be considered iconic, in that the use of a given prosodic feature can refer directly to cultural conventions. In short, the theory demonstrates the need for cultural and contextual knowledge as well as the need for knowledge of the formal linguistic structures in order to truly begin to be able to account for specific linguistic forms.

2. THE AFRICAN-AMERICAN CHURCH - AN OVERVIEW

As pointed out above, the role of culture and context appear quite important to the interpretation of certain linguistic phenomena such as prosody; therefore, a basic understanding of the context of the African-American church and the tradition of preaching in the African-American church proves crucial to understanding the linguistic form of King's speech. Geneva Smitherman writes "the traditional black church is the oldest and perhaps still the most powerful and influential black institution" (1977:90). Historically, the church has been the cornerstone of African-American social and political life, and it's roots as one of the avenues of change within the community go back to the days of slavery.

Within the social hierarchy of the church, the minister has a privileged position in that he ranks directly below God. Due to a calling from God, the minister carries the primary leadership position within the church and de facto, also one of the primary leadership positions within the community in general.[4] A minister in the African American church is responsible for mediating between the sacred and secular concerns of his congregation, thereby

[2]In Liberman's work, the tone inventory also included a mid-tone; however, since the work of Pierrehumbert a two tone system has become widely accepted and has been adopted here as well.

[3] See Pierrehumbert, 1980 for a consise summary of the shapes which interpolation rules cause

[4]As recent events in Los Angeles have shown, the church and the ministers within the church have been central to dealing with reconstruction efforts as well as coordinating

providing a connection between religious (and historical) teachings and the current experiences of the congregation.

2.1 The Performed sermon

Gerald Davis writes "The recognition of the organizing principles which support the sermon in performance is the key to the fruitful investigation of narrative creativity among African-Americans..." (1985:47). One of the principle organizing points of the sermon is the creation of tension between dual forces, most often between the sacred and the secular. Davis maintains that there are rhythmic differences between the two forces, with the sacred being irrhythmic while the secular is rhythmic (1985:60).[5] Important for this discussion is the fact that the tensions is created and maintained primarily through the use of specific prosodic tools, including elongation, enjambment, intonational contours, dramatic pausing and emphatic repetition. (1985:78)

The sermon itself is made up of structural units which have a distinct organization. The sermon begins with the identification of the theme through Biblical reference and proceeds with both broad and narrow interpretations of that theme. The sermon, in being formulaic and adhering to specific, culturally expected and defined modes of organization in terms of content, can be summarized as being comprised of several distinct narrative units. Davis says that the structure of the sermon serves as a mnemonic for the preacher, and that the preacher has great freedom for layering his own personal style over the basic structure.[6]

In addition to the basic infrastructure of the sermon, the African-American preacher generally follows several other rules of performance. For instance, the preacher is expected to use the vernacular. Furthermore, the preacher generally uses argumentation aimed at appealing to messages with which the congregation is already familiar and at appealing to his/her own believability or ethos.

3. Martin Luther King

Martin Luther King Jr.'s speech adheres to the basic form of an African-American sermon; thus the specific devices which he uses are not unique. King's uniqueness comes from his ability to mold the expectations placed upon him by his profession to the needs of an audience which goes beyond that of the congregation of an African-American church. King spoke largely in a political arena to an audience of mixed races and religious affiliations. He is less stringent about using the language of the African-American community, and instead uses something very close to a standard Southern English. His religious references are broad and easily identifiable and his use of actual Biblical quotation sparse. The tensions he creates are often less sacred/secular and more something like "us"/"them", or between the general and the specific. Thus, King is brilliant because he speaks to a broad audience whose expectations of him as a political speaker are quite different from the expectations of him as a minister, and still, he speaks to that audience by using the oratorical tools of the African-American church.

---

media coverage and arranging for dialogues with various gang leaders.

[5]Unfortunately, Davis is not explicit as to the actual criteria for being either rythmic or irrythemic, and it is my interpretation that he is using primarily impressionitic determinations for classification.

[6]For a more detailed and comprehensive account of the tradition of preaching in the African-American church as it pertains to prosodic tools, see Queen (1992).

4. RHETORICAL STRUCTURE IN THE SPEECHES OF MLK

The model of rhetorical structure as deveolped by Woodbury (1985, 1987) relies on independent components which organize a text both autonomously and in interaction with one another. The model in it's minimal form includes a prosodic, syntactic and thematic (or content) component, as well as other genre-specific components. For the analysis of Martin Luther King, the heaviest organizational burden is placed on certain prosodic features. It is important to note that, although the genre of speech has remained constant for the texts under discussion, each text is organized according to unique interactions among the various features (as well as other components such as the syntax). Therefore, while the tools themselves remain constant, their implementation is variable in many ways. This is an important factor for recognizing the relevance of whatever generalizations I have made.

5. PROSODIC ORGANIZATION

Martin Luther King's use of prosody makes its primary organizational contribution at the level of the line and achieves this state primarily through the use of structural pauses of varying length. In addition to pause phrasing, King also uses individual pitch contours which can either reinforce the phrasing created by the pauses, divide long pause phrases into units which adhere more to syntactic boundaries (usually S'), or act independently of phrasing. Additionally, King uses both bitonal and monotonal pitch accents as a way of denoting stress.

King's use of phrase-internal accents conforms for the most part to expectations about stress in English natural discourse.[7] In other words, an analysis of the meter does not reveal any unique poetic properties such as we expect from poetry proper. Therefore, I will not be discussing the metrical properties of his speech in any detail, other than to say that King uses stress (which not only includes pitch contours, but also duration and loudness) as a matter of emphasis, and thus, his patterns are highly variant across and within individual texts.

The basic prosodic devices which King has at his disposal include structured pauses, phrase-final pitch contours and phrase initial rises. In describing these devices, I will be taking the viewpoint that they are defined and interpreted by their form as well as by the cultural conventions and traditions surrounding their use. For instance, while a pause creates a moment of suspension of speech purely by definition (i.e. it's form), much of the determination of possible junctures for pausing as well as the interpretation of the suspension of speech are largely the responsibility of the context.

5.1 Stylistic differences

As with any speaker, one expects variation along some continuum of styles from Martin Luther King Jr. Since the range of his variation is highly constrained by the fact that much of what is recorded of his speech is fairly formal, it is pointless to try and draw specific stylistic profiles. Nonetheless, it does appear to be relevant and important to note the gradience in terms of his speech being more or less preacher like. In fact, the interpretation of which linguistic forms act as specific tropes or icons depends on their isolation as units

[7]It should be noted that any striking use of stress may in fact be attributable to King's native English which differs from my own. See Queen (1992) for a more in-depth discussion of the phrasal accents.

relevant to a specific genre, thus that genre must be identified as clearly as possible.

I will be discussing the features of his preaching throughout the remainder of this essay; therefore, I will now briefly outline the characteristics of his non-preaching. The most salient difference between the two genres is the lack of pitch excursions in the non-preaching style and the significantly lower pitch baseline. The characteristic phrase-initial rises and phrase-final falls are completely absent as are the elongated syllables. Pausing adheres strictly to larger syntactic boundaries (basically S'), and the seemingly deliberate and slow manner of speech does not occur. Furthermore, there is neither Biblical reference nor the creation of a polarity at the level of content. In general, there is nothing about his speech in the non-preaching style which exposes his expertise at oral performance.

5.2 Pause phrasing

Pauses are the primary indication of the line and act both independently of and in conjunction with pitch contours. In general, longer pauses tend to mark off units which are larger than a line, although this is by no means a steadfast rule. Furthermore, a pause may (but must not) indicate pragmatic salience. Often, a longer pause is implemented due to interaction from the audience. A note about audience interaction, however, is that, for the most part, King is in control of how long the audience responds to a given phrase, and given the expectations for call and response within the African-American preaching tradition, length of pause may in fact be "planned" in the sense that King structures in a "long" or a "short" pause. Audience response tends to correlate heavily with thematic considerations such that the introduction of a given trope may spark audience response. I find no specific way of determining which factor influences the other and am left to conclude that the two are simply correlated. The following example shows how pausing works within a given text.

1.
He was murdered by the irresponsibility (1.45)
of every politician from governors on down(1.61)
who have fed his constituents the stale bread of hatred and the spoiled meat of racism. (2.68)
He was murdered by the timidity (1.59)
of a Federal Government (1.65)
that can spend millions of dollars a day to keep troops (.59)
in South Vietnam (.47)
and can not protect the lives of its own citizens seeking the right to vote.

This particular text is interesting in several ways, most obvious of which is the tight parallelism which holds formally as well as thematically. In terms of the pause structure, it should be noted that the longest pause falls between the two largest units, whereas shorter pauses divide individual lines. The final clause receives the shortest structural pauses, a fact which corresponds with the salience of that clause as the "punchline".

5.3 Phrase-final contours

I will now turn to a specific description of phrase-final intonational contours. Martin Luther King uses three types of phrase-final contours-rising, level and falling. The falling contours are further categorized in terms of elongated falls and short falls. By far the most prevalent phrase-final contour is the non-elongated fall. This should not be surprising given the fact that phrase-final lowering is an extremely robust occurrence cross-linguistically (cf. Bolinger 1986 and Pierrehumbert and Hirschberg 1990). Of more interest to the present study are the

phrase-final levels and rises and the elongated falls.

In general, phrase-final rises occur very rarely, and I have thus combined them with the levels in terms of both description and distribution (but not necessarily interpretation). Phrase-final levels follow the form presented in Pierrehumbert (1980).[8] Within a text, however, there does not seem to be a clear-cut method for determining when a level boundary will occur. Furthermore, its distribution (as with the distribution of all of the intonational features) appears local in that given texts and given units within texts may vary with respect to the implementation of a level as opposed to a fall. Very often, however, levels are used between structurally parallel lines as seen in the following example.[9]

2.
And to be sure that (.7) --
all of the bags were checked (2.19) #
and to be sure that (1.09) --
nothing would be wrong in the plane (.67) ##

Here, the parallel lines are marked by a level boundary. Furthermore, note the longer pause which helps to demarcate the larger units of parallelism. Phrase-final levels may also occur when pausing breaks syntactic clauses, as is demonstrated in the following example.

3.

I speak out against this war because (.38) --
I am disappointed with America (.72) ##
There can be no great disappointment (.55) --
where there is no great love ##

Additionally, speeches by King which are not in the preacher style use the phrase-final level as the default boundary tone, rather than the phrase final fall.

A fall which lasts over .4 of a second from the final accent to the termination of speech is considered to be an elongated fall. The determination of .4 of a second was based on taking the average length of all phrase final falls and calculating any fall which was longer than that average as being elongated. Elongated falls generally occur on the final syllable of the phrase, but may also encompass the final foot or occasionally several syllables. This methodology may be a bit questionable, however, I was unable to determine some independent criteria for considering a fall to be elongated, although elongated falls are quite salient perceptually. Furthermore, making the distinction between the two is important in terms of the organization of individual texts. For instance, the following example shows the way in which elongated falls mark the orientation and coda of a short speech.

4.

And I oppose the war in Vietnam (.68) ##
because I love America (1.48) ##
I speak out against it not in anger but with anxiety (.54) #
and sorrow in my heart (.59) --
and above all with a passionate desire (.93) #
to see our beloved country stand (.38) --
as the moral example of the world (.98) ##

I speak out against this war because (.38) --
I am disappointed with America (.72) ##
There can be no great disappointment (.55) --
where there is no great love ##

---

[8] King uses H*HL% and HL* H L% almost exclusively. Levels occur most often at unit-internal points and must always occur text-internally.

[9] The reader should note the following transcription conventions: –indicates a level boundary; ? indicates a rise, # indicates a non-elongated fall and ## indicateds an elongated fall.

In this speech, elongated falls occur at the ends of the first two lines (the orientation to the speech), and again at the end of the first thematic unit. In the second thematic unit, elongated falls mark the semantically salient points. The next example demonstrates the way that elongated falls may work even more directly.

5.
And it seems that I can hear the God of History saying (.43) #
That was not enough (1.8) #
But I was hungry (1.32) #
and ye fed me not. (1.02) --
I was naked and ye clothed me not (1.75) --
I was devoid of a decent sanitary house to live in ##
and ye provided no shelter for me ##
and consequently you can not enter the kingdom of greatness (.90) ##
If ye do it unto the least of these, my bretheren, (1.83) #
ye do it (1.05) #
unto me #

Here, King uses elongated falls to mark the syntactic boundaries within a larger pause phrase. Furthermore, the use of the elongated fall corresponds with the final lines of a parallel set (note also the the first lines of the set are set off by level tones) as well as an enjambed pause phrase. The phrase itself is salient thematically because within it lies one of the primary messages of the speech.

5.4 Phrase-initial rises

The final prosodic trope which I will be discussing is the phrase-initial rise. This particular tool occurs comprehensively across the data and it's absence rather than it presence is what appears to be marked. The phrase-initial rise is characterized by it's occurrence following a pause break. The rise begins low, most often below 200 Hz and rises between 50-200 Hz, with the average rise being approximately 120Hz. The most perceptually salient rises are those which occur in the range of 100-200 Hz. Figures 1-2 in the appendix show pitchtracks of phrase-initial rises.

Rises occur on either the first syllable or the first foot of the pause-phrase, but may not occur if the first foot carries a prominent stress as the following examples show (the rise is marked with this [∞] symbol and the prominent stress is underlined).

6.
∞And we've had the plane (.84)
protected and guarded all night

7.
∞And I've seen (.62)
the promised land

8.
∞But it wasn't a victory for colored folk

9.
This is why I've said

Unfortunately, the case is not as simple as stress since there are also pause phrases where the rise is absent even though the stress does not fall on the first foot. Here, again, the absence of the rise on phrases where the rise is sanctioned seems to act in a way which marks salience. The general tendency in terms of the distribution of the phrase-initial rise appears to be related to information structure.[10] Furthermore, the use and absence of the rise is local in terms of organization. The following give such organizational examples:

[10]For instance, King's speeches never begin on a phrase-initial rise. See Queen (1992) for a more in-depth discussion of the relationship of phrase-inital rises to infromation structure.

10.
The pilot (.97)
said (.37)
over the (1.18)
public address system (.28)
We're sorry for the delay (1.6)
∞but we have Dr. Martin Luther King on the plane (1.83)

11.
∞that in order to get this bill through, (.28)
∞we've got to rouse the conscience of the nation ∞and we oughtta march to Washington, more than one-hundred thousand, in order to say (.8)
∞in order to say that we are determined (.61)
∞and in order to engage in a non-violent protest (.46)
to keep this issue before the conscience of the nation

The first three lines in 10 are constrained by the fact that there is phrasal stress on the first foot, however, the subsequent lines do not have the same constraint. In this example (and it is one of the few examples of this kind), it is the presence of the initial rise which is marked. Example 11 shows the more prevalent pattern of using the absence of the rise in a salient manner, in this case to mark the end of the content unit as well as the end of the speech as a whole. This second example also shows how the phrase-initial rise is not absolutely committed to occuring at the beginning of a pause phrase, but may also occur at the beginning of a syntactic clause and may in fact mark salient syntactic boundaries which occur within pause phrases.

6. PROSODIC INTERPRETATION AND CONCLUSION

Given that I have isolated a set of prosodic tropes or devices which Martin Luther King has at his disposal, where do these data lead? If we accept the culture and tradition of which King is a part as integral to determining what tools are acceptable for him to use, then we must also accept the fact that in some way those tools are indexical to the culture which defines them. The prosodic tropes which King uses have independent functions which are to a large degree defined by their very form (i.e. a rise by its nature signals movement from an endpoint while a fall by its nature signals movement to an endpoint), but they also have an indexical function. McLemore (1991) writes "the indexical function is basically one of evoked extralinguistic associations, in which a language feature cues one to look for the expected co-occurring aspect of context..." and Woodbury (1992) says of Stray footing in Nunivak,

> If we think of SF as a stylized slowing or rallentando, then it has naturalness as a marker of ending. However, it is purely a matter of convention that this device has become the pervasive, normal and expected marker of ending in the speech of Nunivakers. Moreover, the speaker (or child or linguist) must know what others will accept as 'whole and complete communicative acts'. And knowing that is a question of culture.
>
> (1992:8)

The prosodic devices used by King are acceptable and expected because culture and tradition define them as such. The question as to their inherent "meaning" in terms of the connection between a given sound or set of sounds and some definable concept becomes is in many ways less interesting.

I think this viewpoint becomes especially clear when we consider the phrase-initial

rise. This is an intonational contour which is unique in many ways. It has not been used in tests of computer-generated speech nor has it's interpretation as something emotive or something based on the internal state of the speaker been widely studied or discussed. The question is, then, what does it mean when Martin Luther King starts a phrase with an initial rise?[11] I contend that it doesn't necessarily "mean" anything. It is indexical or iconic and derives it's meaning from the fact that conventions for its use and interpretation have been established by the culture which uses it.

REFERENCES

1. Bolinger, D. 1986. *Intonation and it's parts: the melody of speech.* Stanford:Stanford University Press.

2. Davis, G. 1985. *I got the word in me and I can sing it, you know.* Philadelphia: University of Pennsylvania Press.

3. Hirschberg, J and J. Pierrehumbert. 1986. The intonational structuring of discourse. AT &T Bell Labs Technical Memorandum 11225-870325-08.

4. Hymes, D. 1968. The ethnography of speaking. In: *Readings in the sociology of language.* J. Fishman (ed.). The Hague: Mouton.

5. Liberman, M. 1978. *The intonational system of English.* Bloomington: Indiana University Linguistics Club.

6. MPI Home Videos. 1989. *The speeches of Martin Luther King Jr.* The great speeches series. MPI Home Videos:United States.

7. McLemore, C. 1991. The pragmatic interpretation of intonation: sorority speech. Doctoral Dissertation. University of Texas at Austin.

8. Mitchell, H. 1970. *Black Preaching.* Philadelphia:Lippencott.

9. Pierrehumbert, J. 1980. *The phonology and phonetics of English intonaton.* Doctoral Dissertation. MIT.

10. Pierrehumbert, J. and J. Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In: *Intentions in Communication.* P. Cohen, J.Morgan, and M.Pollock (eds.). Cambridge, MA: MIT Press. 271-311.

11. Queen, R. 1992. Rhetorical structure and Prosodic Organization in the Speeches of Martin Luther King Jr. unpublished ms. The University of Texas at Austin.

13. Woodbury, A. 1985. The functionsof rhetorical structure: A stucy of Central AlaskanYupik Eskimo discourse. *Language in Society* 14:153-190.

16. --1992. Utternace-final phonology and the Prosodic Hierarchy. Handout. Meeting of the Linguistic Society of America, 1992. Philadelphia.

---

[11]The use of this specific contour is not specific to Martin Luther King. Recently on *Saturday Night Live*, Jesse Jackson performed a parody of the speech of an African-American preacher in which one of the primary jokes was his exagerated use of prosody (including the phrase-inital rise).

APPENDIX

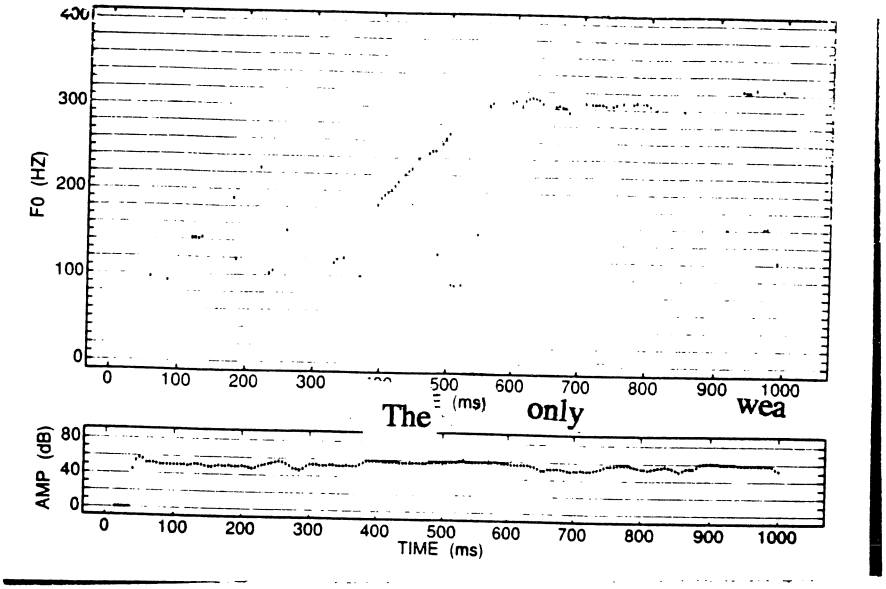


Figure 1

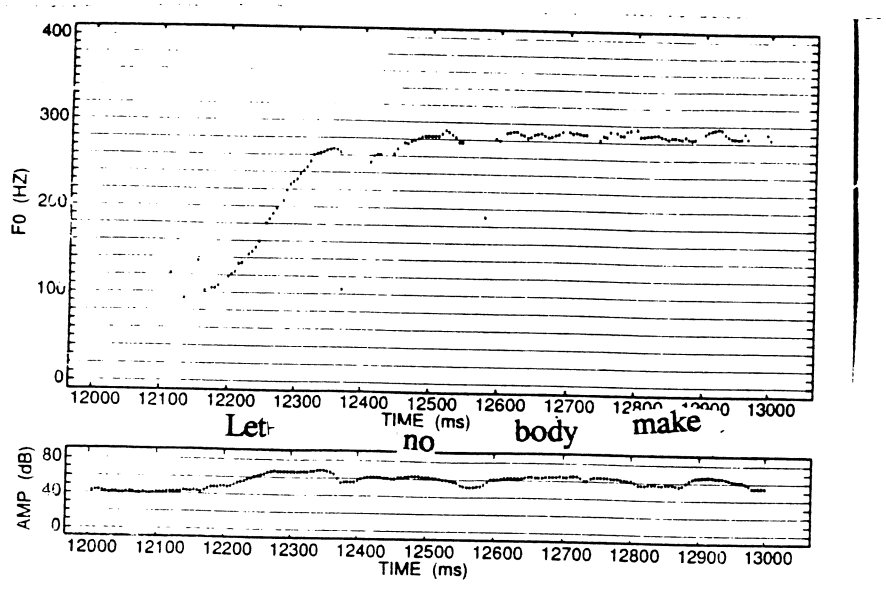


Figure 2

MARSEC: DESIGN OF A MACHINE-READABLE SPOKEN ENGLISH CORPUS OF BRITISH ENGLISH

Peter Roach, Nawal Ghali and Simon Arnfield

Speech Laboratory, Department of Psychology, University of Leeds, U.K.

1. INTRODUCTION

In our attempts to make generalisations about intonation we depend heavily on the validity of the method used to represent prosodic information. In order to evaluate the relationship between intonation transcription and the physical properties in the speech signal we need a large sample of transcribed recordings; this paper describes work on such a corpus which also provides a considerable amount of grammatical information.

Many methods have been developed for recording pitch movements as heard by the trained analyst. Some attempt to record pitch movements with maximal phonetic accuracy (and hence with some redundancy) while others rely on a prior phonological analysis to make possible a more economical coding with minimal redundancy. Ideally, any good intonation transcription should make it possible to generate an acceptable fundamental frequency contour that closely resembles that of the original speech, though explicit statements of this as a goal are comparatively recent. However, we do not know in quantitative terms how successful we can expect this process to be. Experimental evidence on the fallibility of human judgements about prosody tend to be based on rather small samples. What is needed is a large body of human-transcribed speech in computer-readable form that will enable us to explore in statistical terms the relationship between the trained expert's auditory transcription and the acoustic analysis of the same data carried out by computer.

We are working on a project (funded by ESRC and shared with Lancaster University) to convert the Spoken English Corpus (Knowles et al, forthcoming[1]) into a machine-readable database stored on CD-ROM. This corpus, the original work on which was funded by IBM UK, comprises around 6 hours of radio broadcasts and other talks, and has already been prosodically transcribed in its entirety by two experts; the text is in machine-readable form with numerical codes for tone-marks. The prosodic transcription that was chosen for the analysis is a variety of the type of transcription commonly referred to as "Standard British", but it differs in some significant respects from the most widely adopted versions such as O'Connor and Arnold [2], particularly in that it does not segment the intonation-unit into pre-head, head, nucleus and tail. Each pitch-accent is therefore marked with one of the available tones, with the convention that the last pitch-accent in the intonation unit is deemed to be the nucleus.

The corpus has been grammatically tagged word by word, and an automatically generated parse applied. The textual form of the corpus is therefore very richly annotated. Our project is currently working on the digitisation of the corpus and the acoustic analysis of it, and this will be

followed later this year by work on automatically aligning the text, syllable by syllable, with the acoustic signal, so that the recording corresponding to any portion of the corpus can be retrieved very easily from the disk.

The research work described in this paper is intended to produce three principal deliverables:

(i) A machine-readable version, stored on CD-ROM, of the six-hour corpus of digitised recordings on which the Spoken English Corpus is based. This machine-readable corpus is called MARSEC (MAchine-Readable Spoken English Corpus).

(ii) A set of linked files providing textual, grammatical, acoustic and prosodic annotations of the recordings, all with a common time reference.

(iii) A statistical methodology for examining the relationship between the expert auditory transcription of the existing SEC and the acoustic parameters that can be extracted from the recorded signal.

In addition, the project is evaluating alternative transcription systems for work of this sort. The existing SEC was transcribed using an approach that would not necessarily be regarded as ideal in the context of present-day prosodic research. This part of the research is primarily the field being worked on by our partners in the University of Lancaster. The following outline is based on the three headings given above, and follows them in the order given.

2. DELIVERABLE (i): *Machine-Readable Speech Data*

The corpus (i.e. the full set of recording) lasts for around 6 hours. We decided that it should be converted into digital form in pieces lasting no longer than 1 minute (to enable microcomputer-based speech workstations to handle chunks of the corpus without the need for further editing). One minute of speech takes up a little less than 2 mbyte.

2.1. Filenames and subdirectory structure:

There are 11 subdirectories, corresponding to the 11 categories of the corpus (A through M, excluding I and L). The files have been named according to the section number, the 1-minute chunks and, at the end, we have added either b (if the prosodic transcriber was BJW) or g (if the transcription was done by GOK). The extension .sig represents signal files. For example, the file A0101b.sig is part A, section 1, first minute, transcribed by BJW.

2.2 Digitising the recordings:

We used a PC configured as a SAM workstation, since we intend to use the SAM conventions and protocols as far as possible: these have become the de facto standard for most collaborative European speech research. The files were created with AU21DSK, a program produced by the

manufacturers of the OROS AU21 board that is used in the SAM workstation. We used a sampling rate of 16 KHz. The OROS board automatically applies appropriate anti-aliasing filtering.

2.3 Editing and storing:

We have used the PTS signal editing package to edit these files. Each file has been limited to within one minute, breaking at a pause (a clear silence). The recording includes also the first word following the pause, and the following file starts from the word before the pause; in this way we preserve the pause itself. The breaks have been marked on a master copy of the text.
The recording were initially archived on Exabyte magnetic tape, and was then transferred on to a single CD-ROM disk, which will be available through the ESRC Data Archive[1].

3. DELIVERABLE (ii): *Cross-referencing mechanism for the corpus.*

The MARSEC corpus will consist of several versions of the data as follows:

* acoustic waveform
* fundamental frequency waveform
* intensity waveform
* phonetic transcription
* syllabic division transcription
* prosodically annotated transcription
* punctuated transcription
* word tag transcription
* parse treebank

This section describes a mechanism for cross-referencing from any file to the equivalent position in any other file. To be able to do this several complicated indexes need to be created. The phonetic transcription is aligned automatically with the acoustic waveform by HMM. This is the key step in allowing cross-referencing between the acoustic data and the textual data. The next step is to match the phonetic transcription with the syllabic transcription. This should be fairly straightforward since the phonetic transcription will have been generated from the prosodic text (as will the syllabic transcription) although they might contain some minor differences. The syllabic transcription will have been generated from the prosodic transcription and as such will be easy to match backwards, especially if some re-alignment data (such as word boundaries) is included in the syllabic transcription (maybe only temporarily). The task of aligning the prosodic transcription with the word-tag

[1] ESRC Data Archive, University of Essex, Colchester CO4 3SQ.

transcription has mostly already been achieved. Matching the treebank and punctuated versions with the word-tag is relatively straightforward, and will be easier when the errors that have been introduced by post-editing of the punctuated version and the word-tag version have been resolved.

3.1 CROSS-REFERENCING

There are thirteen basic types of cross-reference possible, shown diagrammatically in Fig.1:

3.1.1.1 acoustic -> phonetic transcription
3.1.1.2 acoustic -> fundamental frequency / intensity.
3.1.1.3 acoustic -> syllabic, prosodic, word-tag, punctuated, treebank

3.1.2.1 F0/intensity -> acoustic
3.1.2.2 F0/intensity -> phonetic
3.1.2.3 F0/intensity -> syllabic, prosodic, word-tag, punctuated, treebank

3.1.3.1 phonetic -> acoustic
3.1.3.2 phonetic -> fundamental frequency / intensity
3.1.3.3 phonetic -> syllabic, prosodic, word-tag, punctuated, treebank

3.1.4.1 text -> acoustic
3.1.4.2 text -> fundamental frequency / intensity
3.1.4.3 text -> phonetic
3.1.4.4 text -> syllabic, prosodic, word-tag, punctuated, treebank

Full details of the interlinking of these levels are given in MARSEC documentation which will be distributed to users of the corpus.

3.2 FILE FORMATS AND NAMING CONVENTIONS IN THE SEC

The SEC consists of various versions of the data. This section explains what files exist and their file formats. The original version of the corpus was produced from tape recordings of radio broadcasts and some other talks. From these the unpunctuated transcription was produced which was punctuated by volunteers and prosodically transcribed by Gerry Knowles and Briony Williams. The punctuated version was then used to produce the word tag version, and later the parse treebank (not described in this paper). The new version of the corpus will also include five more versions: acoustic waveform, fundamental frequency waveform, intensity waveform, syllabic division transcription, phonetic transcription.

3.2.1 UNPUNCTUATED VERSION

Not distributed. This version is exactly the same as the punctuated version (below) except that no punctuation symbols are included. There are 53 files in 11 categories:

A01 A02 A03 A04 A05 A06 A07 A08 A09 A10 A11 A12
B01 B02 B03 B04
C01
D01 D02 D03
E01 E02
F01 F02 F03 F04
G01 G02 G03 G04 G05
H01 H02 H03 H04 H05
J01 J02 J03 J04 J05 J06
K01 K02
M01 M02 M03 M04 M05 M06 M07 M08 M09

Each category deals with a different type of speech style. See 'A Manual of Information to Accompany the SEC Corpus' for more information.

3.2.2 PUNCTUATED VERSION

In many instances these files were produced by volunteers punctuating the given unpunctuated text and represent the original "script" that the reader is supposed to have used in producing the recording. However, in some cases the original scripts were in fact available and have been used. This has allowed some variability to creep into the corpus where the original script was not followed exactly by the speaker. Due to the nature of the task of punctuating a text derived from a recording and not being able to hear the recording there will inevitably be some spurious punctuation. File name conventions are as for the unpunctuated version and the files exist in the 'pun' subdirectory. Some editing has been done and is noted by comments such as [change of speaker], [live commentary omitted], or [interview omitted]. Each file contains some header information contained within square brackets, stating the text number, title, speaker(s) and broadcast notes. For example:

[001 SPOKEN ENGLISH CORPUS TEXT A01]
[In Perspective]
[Rosemary Hartill]
[Broadcast notes: Radio 4, 07.45 a.m., 24th November, 1984]

Good morning. More news about the Reverend Sun Myung Moon,
...

3.2.3 PROSODIC VERSION

The prosodic version was split into three directories: bjw, gok and dup. bjw and gok contain the files prosodically transcribed by Briony Williams and Gerry Knowles respectively whereas dup contains sections from the corpus that have been transcribed by both bjw and gok. The same file name conventions have been used but with some slight changes. In some cases one of either bjw or gok will have transcribed the whole section and the other will have done a small section for comparison purposes. In these cases the whole transcribed section will occur in one of bjw or gok and the repeated section will occur in dup. In other cases each will have done half a section with a small overlap. In these cases the same file name will occur in each subdirectory bjw, gok with the overlap occurring in dup. It is therefore true to say that some of the information in dup is entirely repeated (as in the last case above) whereas some information (the first case above) only exists here. Header information is included as for punctuated files but with the addition of a line to indicate the transcriber. Unfortunately no indication is given as to where (in this recording) this file comes from. So, for example, in section C where the only recording is C01 which was split between the transcribers it is impossible to discover whether bjw/C01 precedes gok/C01 or vice-versa without examining the text for clues. The main differences in file format from the punctuated version is the omission of punctuation (except apostrophes in words such as "don't" and hyphens) and the inclusion of prosodic information.

Prosodic information is marked with a set of codes. There are three marks used for tone unit boundaries: | || and #240; the latter is also used to mean low rise-fall, but fortunately this only occurs once in the corpus on line 148 of bjw/F04: "#240following". Other prosodic symbols are _ (meaning low-level) and the remainder which comprise a # followed by a 3 digit number. The full list of prosodic symbols, as they appear in the corpus, is: | || _ #161 (high fall-rise), #162 (high rise-fall), #246, #247 (low fall-rise), #171 (low rise), #172 (high rise), #173 (low fall), #174 (high fall), #163 (high level), #248 (stressed unaccented syllable), #165 (pitch raising), #166 (pitch lowering), #240 (low rise-fall); #249 was originally used as synonymous with #248. The code #240 is used as a "hesitation tone-unit boundary" and was only transcribed by GOK. In addition to this (* bracketed words erased *) and (* unfinished tone group *) also occur.

Example of prosodic transcription file:
[001 SPOKEN ENGLISH CORPUS TEXT A01]
[In Perspective]
[Rosemary Hartill]
[Broadcast notes: Radio 4, 07.45a.m., 24th November, 1984]
[Transcriber: BJW]
#166Good #174morning || #165#174more #249news about the #163Reverend _Sun
#248Myung #174Moon | _founder of the Unifi#174cation #248Church | who's
#161currently in #248jail | for #174tax evasion || ...

3.2.4 WORD-TAG VERSION

This actually exists in two formats, "vertical" and "horizontal". In practice the horizontal version is only of use because it is easier to read and this will not be described here. The vertical tag format is that produced by the CLAWS (Garside et al [3]) tagging suite developed for the LOB corpus and exists in the vtag subdirectory. The horizontal tag files exist in the htag subdirectory. The same file name conventions used by the punctuated version are used. The format of the vertical tag file contains six columns, each line having a single entry such as a single word or punctuation symbol. Column 1 is the file name; column 2 is the line number in the punctuated version from which this entry came; column 3 is a three digit number, the first two of which indicate the word number on the line and the third digit is used to number each punctuation symbol following the word; column 4 contains the word-tag; column 5 is the word/punctuation entry; column 6 contains residual information such as marking manual editing (@), some enclitics (< and >), some compound/non-standard words (*), ditto forms etc. End of sentences are marked by inclusion of end of sentence markers. These are tagged as 5 hyphens and the word is 43 hyphens. A number of differences (from the punctuated version) have been introduced into the vertical tag format due to editing -- these largely take the form of inserted/deleted/modified punctuation. One notable difference is that A01--A06 have had square brackets changed to parentheses, whereas elsewhere they remain square. Header information is removed except for the title and author, and lines such as [change of speaker] and [speech extract omitted]. This information is tagged just as other information. Here is an example:

```
A01   2 001 (     (                                   @
A01   2 010 IN    In
A01   2 020 NP    Perspective
A01   2 021 )     )                                   @
A01   3 001 (     (                                   @
A01   3 010 NP    Rosemary
A01   3 020 NP    Hartill
A01   3 021 )     )                                   @
A01   5 001 ----- -------------------------------------------
A01   5 010 JJ    good
A01   5 020 NN    morning
A01   5 021 .     .
A01   5 022 ----- -------------------------------------------
A01   5 030 AP    more
A01   5 040 NN    news
A01   5 050 IN    about
A01   5 060 ATI   the
A01   5 070 NPT   Reverend
```

4. DELIVERABLE (iii): *Statistical Treatment*

This is the area in which least progress has been made so far, since it is intended that it will form the bulk of the work in the second year of the project. The problems to be addressed are familiar enough to anyone with experience of working with acoustic records of prosodic phenomena, and for the present we shall simply summarise the main problems as we see them. Each F0 and intensity file generated by acoustic analysis of the data files is in the form of a continuous vector corresponding roughly to pitch and loudness. The relationship between the two acoustic parameters on the one hand and intonation on the other is not fully understood, and will be a major focus of interest in the research. We feel that acoustic studies of intonation have concentrated too exclusively on fundamental frequency as the correlate of the auditory percept, and we hope to establish that some weighted function of fundamental frequency and intensity will produce a better match.

In carrying out auditory transcription, the auditory system appears to perceive pitch as continuously varying. In the SEC, major changes in the pitch are marked on the text with one of a fixed number of tone marks (e.g. / for rising movement, \ for falling). In the simplest kind of comparison, we could calculate *for a given speaker* what acoustic characteristics correspond to the tone marks \ and / on single syllables. We would then be able to predict what tone mark a human analyst would use to represent a given set of F0 values for such items. A number of factors cause complications:

1. Different speakers have different pitch ranges. In studying more than one speaker, therefore, we need to look not at *absolute* F0 values but at *relative* ones (relative to the speaker's normal pitch range.

2. Although on single syllables the relationship between (auditory) pitch and (acoustic) F0 is easy to see, the tone mark actually predicts pitch behaviour over all the syllables which follow the tone-marked syllable up to the next tone mark **or** the tone-unit boundary which follows. In terms of the auditory analysis of intonation used for the SEC, therefore, there is no categorical difference between the following:

/ no / nobody / nobody went there / nobody went there without a ticket

- though the F0 record will look very different. The same tone mark therefore represents a considerable variety of F0 contours.

3. The tone marks seem to imply a continuous pitch movement, but the F0 track stops and reverts to baseline when voicing ceases; this happens at silent pauses, of course, but also at voiceless consonants. So a sentence like 'Eat sweet potatoes' will have several large gaps in the F0 data that will cause problems for pitch-reading algorithms, though the human auditory system is able to "connect up" between the voiced parts.

4. It will not be sufficient to be able to state context-free rules to convert between acoustic and auditory values. The phonological categories used in intonation transcription are realised in very different ways in different contexts, particularly as regards position in the tone-unit. Consequently an analysis based on isolated fragments of speech could only be regarded as a tentative preliminary approach. One of the most valuable aspects of the material we are working with is the wealth of examples of relatively unrestricted connected speech. We hope that this will permit us to make more progress on this research area than has been possible in earlier work.

5. REFERENCES

[1] G.O.KNOWLES, L.TAYLOR AND B.WILLIAMS *The Lancaster/IBM Spoken English Corpus*, Longman (forthcoming).
[2] J.D.O'CONNOR AND G.F.ARNOLD *The Intonation of Colloquial English,* (2nd.Ed), Longman (1973).
[3] R.GARSIDE, G.LEECH AND G.SAMPSON *The Computational Analysis of English*, Longman (1987).

PROSODIC PHRASE AS A PROTOTYPE

Stephan Schuetze-Coburn

Department of Linguistics
University of California, Los Angeles
Los Angeles, CA 90024-1543

ABSTRACT

The linguistic unit 'prosodic phrase' has an underlying if not overt syntactic basis in many phonological and descriptive accounts of prosodic structure. On the other hand, phonetically oriented definitions are usually too limited or vague, so that they fail in the analysis of natural, connected speech. The basis for avoiding phonetic substance or for not providing adequate phonetic detail is the apparent lack of a clear set of invariant phonetic cues with which the category 'prosodic phrase' may be defined. It is suggested that while this may indeed be the case, there are alternatives to searching for criterial attributes. Viewing the category 'prosodic phrase' as a prototype is one way of shifting the perspective away from the expectation of necessary and sufficient conditions and towards a characterization of 'prosodic phrase' which more accurately reflects even the variation found in spontaneous speech. Properties of prototypes in linguistic theory are examined, and the implications of considering a prosodic phrase category as a prototype are explored in the context of a German conversational narrative which has been analyzed auditorily into 'intonation units'.

1. INTRODUCTION*

In this paper, I will be continuing the discussion of the prosody of 'intonation units' that opened the workshop with the presentations of Du Bois and Chafe. Initially, though, I will be framing the points that I have to make in general terms, rather than in terms which are relevant only to intonation units, so I prefer to call the units that I will be talking about in the first two sections of my paper 'prosodic phrases'. I'd like to think of 'prosodic phrase' as a cover term for the various phrase-length prosodic units that are commonly found both in discourse and phonologically oriented studies. Examples include the 'tone group' of Halliday (1963, 1967a, 1985), the 'tone unit' of Crystal & Quirk (1964) and Crystal (1969, 1975), and Brazil (1975, 1978, 1985), the 'intonation group' of Cruttenden (1970, 1986) and Fox (1984), the 'intonation phrase' of Pierrehumbert (1980), the 'intonational phrase' of Selkirk (1981) and Nespor & Vogel (1983), the 'γ-frame' of Gibbon (1984), the 'major phrase' of Ladd (1986), and many other loosely related expressions (for example, the units of Pike 1945, Trager & Smith 1951, and of other earlier works such as Palmer 1924). By grouping these units together, I don't want to imply that any pair of units is actually equivalent.[1] But there are obvious similarities, and I do think that they all are trying to get at a certain kind of prosodic organization—what can be called the basic phrasing of utterances.

My focusing exclusively on this structural level shouldn't be taken to mean that I believe other possible levels of prosodic organization aren't worth investigating—either in a phonological hierarchy, where especially smaller units are of interest (e.g. word, clitic group), or in a prosodic account of discourse structure (where the larger units, e.g. major paratone, pitch sequence, and so forth, find their place). But I do believe that prosodic phrases form a particularly interesting level of organization for a variety of approaches to the structure of speech, especially that which is natural, connected, and spontaneous. Units at this level have been claimed to function as a domain, for instance, for the information structure of discourse (e.g. Halliday 1967b; Kreckel 1981; Chafe 1987), and for speech production (e.g. Laver 1970; Svartvik 1982), as well as for various intonational features (e.g. declination, Pierrehumbert 1980) and phonological rules (e.g /t/-flapping in English, Nespor & Vogel 1982). They have also been shown to have interactional significance by contributing to turn organization (Orestrom 1983; Ford & Thompson 1992).

1.1. Prosodic phrase and data type

My point of departure for examining the nature of the category 'prosodic phrase' is a methodological one. How do linguists study intonation, and how does this bear upon the definition of prosodic units? While it is seldom acknowledged, when we construct phonological models or perform

* I wish to thank Jack Du Bois for helpful suggestions and the members of the workshop for their comments. Special thanks to Mark Liberman and Cindie McLemore for the use of the Linguistics Dept. Phonetics Lab at the University of Pennsylvania. I am especially grateful to Felicia Hurewitz for digitizing and pitch tracking the conversational excerpt discussed in this paper. The conversation was recorded in 1988 in Bielefeld, Germany, during a stay funded by the *Deutscher Akademischer Austauschdienst*. I also thank Marian Flaherty, Hartmut Kreft, Marlene Marlow, and Silvia Rode for their advice and help with the transcription; all responsibility for the 'final' stage is of course mine.

[1] Many researchers have assumed (or asserted) equivalence. As each of the above-mentioned units is defined differently (and serves various purposes), I prefer to be more cautious on this point (cf. Couper-Kuhlen 1986: 76).

perceptual experiments, the extent of a domain such as 'prosodic phrase' can be easily manipulated ('extent' as measured in terms of some linguistic construct). Example data that are to be accounted for in a model and stimulus materials in an experiment are not independently provided, but are of course selected specifically for their suitability in realizing the stated goals. But because the linguistic structures that these kinds of analyses are based on are fundamentally syntactic in origin, not prosodic, the syntactic structure ends up influencing—or even determining—the prosodic domain as well.[2] Typically, what one finds is a single grammatical phrase, simple sentence, or sentence pair with a 'normal' (or special) prosody projected onto it, rather than a natural prosodic structure, with the syntax only secondary or incidental. Another way to think about this state of affairs is to consider the number of intonational accounts available in which the syntactic structure is assumed and a prosody is subsequently assigned, versus those where the prosodic structure is given and a syntactic form is then selected (or alternately, accompanies the prosody). There is good reason to view such data with mistrust.[3]

Momentary reflection on data type is important because in working with extensive amounts of continuous, spontaneous speech, it becomes apparent that the extent of the domain 'prosodic phrase' does not correlate very well with how it is usually illustrated, but rather it exhibits much greater (unpredictable?) variation. Broadly speaking, each instantiation of a prosodic phrase is the product of a set of interlocutors—a speaker and the hearer(s)—in a specific context, so its scope emerges from the interplay of an array of factors. In other words, the extent of a prosodic domain is not given in advance in terms of only syntactic (or other such) constraints imposed by the researcher's uniform idea of what the domain should be; instead, it varies as required for the relevant communicative purposes.[4] The point here is that the effect of nonlinguistic parameters on the shape of prosodic phrases has been—it can be said—subverted entirely, and the influence of nonsyntactic ones has been largely ignored in linguistic description.

Crucially, what is evident in the first place is that the challenge of segmenting speech into prosodic phrases is frequently not met, but is neatly finessed by placing certain nonprosodic restrictions on the data. The homogeneous, sentence-oriented language that is often encountered in theoretical and descriptive accounts of intonation does not reflect the variable conditions and pressures that exist in natural language use: speakers and hearers regularly create phrasing that transcends the usual paradigms. Yet, as many of us will agree, if any concept of prosodic phrase is to be truly viable as a linguistic category, it must be possible to specify how all phrases are to be reliably identified in a prosodic analysis of natural, in addition to idealized, speech.

1.2. Defining prosodic phrase

A second orientation point is to consider previous attempts at defining the extent of prosodic phrases. The myriad units encountered in the literature provide one indication that there is little agreement on how prosodic phrasing is to be achieved, either theoretically or in the context of the analysis of particular language data. Many characterizations of prosodic phrase—whether phonologically or descriptively oriented—involve the identification of some unifying pitch sequence, at least indirectly. Two examples: Pierrehumbert (1980) considers the primary prosodic configuration of her phrasal unit to be the 'tune', i.e. a series of pitch accents followed by a phrase accent and a boundary tone. For Altenberg (1987), the defining configuration is 'a coherent intonation contour optionally bounded by a pause and containing (among other things) a salient pitch movement (the nucleus), normally at the end of the unit' (p. 47). While the conceptual notions contained in these characterizations might be clear enough, delimiting all actual instances of intonational coherence in an extended stretch of connected speech—and not just picking out the clear cases, or constructing their equivalent—proves to be a somewhat elusive task. It would seem that by specifying prosodic phrases largely in phrase-internal terms, one is able to cover only a subset of a given text. Yet I take it that a goal of an adequate prosodic description is to account for all utterances. Without a doubt, many—for some texts perhaps even a majority—of contours can be identified relatively unproblematically, either on a perceptual basis or with the help of instrumentation. But some phrases invariably defy identification using the definitive criteria that a focus on the notion 'intonation contour' demands.[5] Unfortunately,

---

[2] In some cases, the assumption is that syntactic structure is prior, since it serves as the input to a phonological component.

[3] Cf. Gibbon's (1988) discussion of data types and his skepticism of accounts based on isolated, invented items for ad hoc illustration (which he classifies as 'anecdotal').

[4] To be fair, some phonologists have recognized the inherent variability in basic phrasing (in the sense that they have consider it 'free'), e.g. Selkirk (1981: 130). Still, prosodic boundaries are held to align with syntactic boundaries.

[5] Most phrase-sized units do seem to have an identifiable coherent contour, but not all do. (Of course, whether the presence of a 'coherent contour' alone, i.e. without any concomitant features of the type mentioned below, is enough for a prosodic phrase to be perceived is an open question.) Three main classes of phrases regularly lack a contour in the usual sense:

(a) Uncompleted phrases. These are considered separate units whether a contour is present or not; i.e., they are not treated as 'residue' which can be incorporated into some other phrase, or ignored. They may exhibit some relative structure despite their lack of a contour. This is most clearly illustrated in sequences of uncompleted phrases.

attending only to highly prominent aspects of the prosody—i.e. the pitch accents—does not give adequate indication of where the boundary between phrases is to be drawn.

Faced with the difficulty of unambiguously distinguishing the component parts of 'coherent intonation contours' (e.g. boundary tones or nuclear accents), it is tempting to abandon the standard line of defining prosodic phrase altogether and to adopt instead a more reliable and 'objective' measure. Brown, Currie & Kenworthy (1980) took such an approach when they rejected a well-known phrasal unit based on nuclear tone in favor of a pause-based unit. But the simplicity of this kind of 'definition' of prosodic phrase is plainly spurious. Although pauses in both spontaneous and read speech are readily measurable (in relative terms), they do not all reflect a common origin (cf. Chafe 1980; Deese 1980; Goodwin 1981), so that strictly pause-based units are not necessarily meaningful when it comes to characterizing prosodic structure as a whole. More importantly, we should stop to consider whether there is any a priori reason to believe whether a single parameter, be it pause or pitch accent or whatever, might unambiguously identify coherent phrasing, either in principle or in practice. Given the complexity of the phenomenon, surely the answer must be no. While getting a firm grip on prosodic phenomena is notoriously difficult, concerns about how to make a category operational should not force us to abandon theoretical (and empirical) substance.

We might ask then: Why not make use of a wide range of prosodic features in defining prosodic phrases? Now, researchers have long noted that factors in addition to specific pitch patterns correlate with phrase boundaries to varying degrees. Prosodic features like silent and 'filled' pauses and other such vocalizations, 'anacrusis' and 'final' lengthening (and other features tied to the 'timing' of the speech), and certain voice quality features are often used to facilitate phrase boundary identification, even if their exact status vis-à-vis these boundaries remains unclear. (Other features such as local variations in pitch width and intensity, as well as largely segmental features like aspiration, also suffer from similar limitations.) Crystal (1969: 205), for instance, states: 'In fact, any process of intonation analysis will take simultaneous account of both boundary cues and internal structures ... and any comprehensive definition of the tone-unit must also have recourse to a complementarity of cues'. He then gives pitch reset and pause as the two primary criteria he uses. Yet, while these and other 'boundary cues' undoubtedly exist to some degree, they have been viewed by many as relatively marginal.

There are several reasons for this lack of interest. One difficulty in making serious use of such correlates is the 'optionality' of all of the above-mentioned phonetic features. In classical definitions, optional features cannot be defining. Thus, it would seem that the main problem to overcome in determining the phrasing of connected speech—which is the focal point here and one of the topics of the workshop—is the lack of a single, ever-present identifiable cue (or invariant set of cues) in the acoustic signal (or the perception thereof), either at phrasal boundaries or in conjunction with intonation contours. Given that received approaches have not been genuinely successful, it may prove fruitful to look at the problem of delimiting prosodic phrases from a fresh vantage point. Instead of searching for the 'correct' invariant components which could be forged into an viable definition, it is perhaps worth considering whether the customary way of defining linguistic units is suitable in this instance. Specifically, while necessary and sufficient criteria may seem adequate or appropriate for defining linguistic categories when some forms of language are examined (e.g. decontextualized language or spoken language which originates from written form, as in reading aloud or in text-to-speech systems), it may, however, also be the case that adherence to a rigid definition of prosodic phrase will never capture the variability which is an integral part of spontaneous discourse.

2. PROTOTYPES AND PROSODIC PHRASE

In recent years, the notions of 'fuzzy' categories and 'prototypes' have been exploited to account for an impressive range of linguistic data (Rosch 1978; Lakoff 1987; Rudzka-Ostyn 1988; Tsohatzidis 1990), including phonological categories (Jaeger 1986; Nathan 1986; Taylor 1989). I would like to suggest that the best way to treat the category 'prosodic phrase' is as a prototype along the lines of these previous studies. In doing so, I believe we come closer to balancing our desire to formulate explicit models of prosody with the practical concerns that arise in dealing with natural speech.[6] The general thrust of this proposal is, of course, not entirely new. Precursors to this idea include Chafe (1987), who describes intonation units using a schematic 'general format'. In subsequent work, Chafe

(b) Nonlinguistic vocalizations. A variety of audible vocal sounds are treated as separate phrases when they are not perceived as part of a larger phrase gestalt, including laughter, inhalation, and coughing.

(c) Short responses or backchannel utterances. Especially when low pitched, these frequently have no clearly identifiable primary prominence or contour, but are nevertheless perceived as phrases. These prosodic phrases need not contain any prominent ('accented') syllables at all.

The extent to which these 'deviant' phrases can be ignored depends on the assumptions and goals of one's approach.

[6] While it is also desirable that the psychological validity of this type of model be demonstrated for this category, such a claim must await later experimental confirmation; here, the precise cognitive representation is not an issue. I wish only to examine the plausibility of this type of model with regard to the characteristics of prosodic phrase.

(1992) calls this format 'the structure of a prototypical intonation unit'. Schuetze-Coburn, Shapley & Weber (1991) also discuss intonation units in terms of deviation from an (abstract) prototype, but there has been to date no detailed treatment of prosodic phrase in terms of a full range of characteristics attributed to prototype models.

2.1. Characteristics of prototype models

In order to evaluate how well a prototype model might define 'prosodic phrase', it is instructive to look at some primary characteristics of prototype categorization and consider their general relevance to the category. Geeraerts (1989) compiles a set of four properties which he says are typical of prototypicality. These include the notions of 'criterial attributes', 'family resemblance', 'centrality', and 'gradience'. Lakoff (1987) covers additional properties, including 'embodiment', and 'basic-level categorization', which I examine below as well.

The notion 'criterial attributes' has to do with the requirement of a 'checklist' of features, each of which must obtain for a definition to apply (cf. Fillmore 1975). The lack of such attributes defining a category is a salient feature of prototypicality. This notion is, of course, the catalyst to the present discussion, and so plays an obvious role here. The optional status of EVERY phonetic feature in a characterization of prosodic phrase undermines a normal definition in terms of necessary and sufficient conditions. If no feature is criterial, then no obvious distinction between essential and incidental features can be established. But in a definition constructed around a prototype, this property would pose no problem; each feature may serve as an attribute in phrase boundary production and perception. However, this is not to say that all features carry equal weight; some may seem more important than others, which the model captures through the property 'centrality' (see below).

'Family resemblance' refers to the idea that exemplars of a category may share individual category features with just a subset of members, yet the subsets of a category overlap so that there is enough similarity among members for each to be included in the category (cf. Wittgenstein 1953). Given the lack of criterial attributes mentioned above, this property is directly relevant to 'prosodic phrase' in that there is no feature which all instances of the category share. When a range of prosodic (and not just strictly intonational) features is taken into consideration, it is apparent that individual prosodic phrases cannot be structurally identical, but instead must resemble one another to greater or lesser degrees, depending on the features realized in any given phrase.

'Centrality' can be summarized in a maxim 'All phrases are not created equal'. That is, there are core and peripheral members of a category, with core members being more 'salient'. Alternately, centrality can be measured in terms of the frequency which member characteristics occur, with predominating features being salient. Centrality is thus a gradient notion concerning the relative degree of membership in a category or the importance of individual features in characterizing membership.[7] As applied to prosodic phrase, if it can be shown that phrases which are clearly instances of the category make better exemplars than others, or if the balance of features differs from phrase to phrase, then the prosodic phrase category would exhibit this prototype property.

'Gradience' refers to the idea that category boundaries are indeterminate, or 'blurred at the edges' (Geeraerts 1989: 593).[8] If we could point to peripheral cases where it could not be decided whether a given token was an instance of a (specified) category or not, then we could say that the category has no inherent clear-cut boundaries. This is one property which seems to be inapplicable to a prosodic phrase category. In order to speak of a category which contains actual prosodic phrase tokens as members, it will be necessary to draw boundaries to delineate one phrase from another. The physical (and psychological) integrity of prosodic phrases must be established in this way, unlike that of concrete objects, like 'cup', which I presume are more obviously independent entities. Note that the lack of (membership) gradience does not preclude gradience in terms of centrality, but it must be possible to decide (in principle at least) whether or not a phrase has been produced.[9]

Additional characteristics of prototype models have been summarized by Lakoff (1987). Two apparently relevant properties which provide necessary epistemological links in a cognitive model are the notions 'embodiment' and 'basic-level categorization. As these notions have less to do with the structure of a category, remarks here are only meant to be suggestive. Lakoff distinguishes two types of embodiment. The first, 'category (or conceptual) embodiment', refers to the biological and experiential grounding of categories, i.e. to their fundamentally nonautonomous nature. Regarding prosodic phrase, this grounding would mean that such a category could not be an independent linguistic (phonological) construct, but would have clear cognitive and social bases. I view the establishment of a prosodic phrase category founded in phonetic substance as a first step in the confirmation of this kind of grounding. In this regard it is interesting to note that some phonetic features used to identify prosodic phrases have been claimed to be 'language independent': they have been found to occur in a range of languages, so that there is reason to suspect that

---

[7] Lakoff (1987) differentiates two aspects of centrality: 'centrality' proper, and 'centrality gradience'.

[8] For Lakoff (1987), this property is 'membership gradience', to be distinguished from 'centrality gradience'.

[9] The question of 'intonational sandhi', in which two (abstract) prosodic phrases are intonationally merged, is an interesting one. On the surface, though, there is but one phrase.

more than structural factors are involved. Pause, F_0 declination, F_0 reset, diminution of F_0 range, prosodic lengthening, and intensity decrease are all such features (Vaissière 1983). Yet, while the inventory of phonetic features may be similar, their relative importance probably differs from language to language, so that a certain arbitrary (i.e. linguistic) component remains.

The second type of embodiment, 'functional embodiment', refers to the psychological status of the category. If a category can be shown to be employed by language users automatically, without conscious effort, then the category exhibits this prototype property. The observation that prosodic phrases are highly relevant to interactional behavior, without speakers and hearers making direct reference to them, points in this direction.

'Basic-level categorization' has to do with the hierarchical organization of categories; as the name implies, basic-level categories are claimed to be cognitively basic. That phrase-level units do not constitute simply one level in a prosodic hierarchy is evidenced indirectly by their falling near the middle of the hierarchy, rather than being the top or bottom level. On the other hand, very suggestive evidence can be found in the area of language acquisition. From my own preliminary observations, it is apparent that in the course of development from a one to two-word stage, prosodic properties of individual words—which constitute complete phrases (cf. Menn 1983)—are concatenated along with the lexical material. That is, the integrity of the prosodic phrase is at first respected, reflecting its fundamental nature. Only later, as language use becomes more sophisticated, are the prosodies melded into a single 'coherent contour', with standard accompanying phonetic cues.

(1)
1	C:	...(1.0) </ˈdədɪs/> ((looking))
2		... </ˈdʒuˈdʒu/> tòp
3		... òn
4		...(.7) ((taking top off)) tòp </əbə/>
5		.. òff
6		... top .. òff
7	P:	... top's òff
8		.. mhm̀

An instance of this process is given in example (1).[10] Each line of (1) constitutes a prosodic phrase, as indicated by the intonation contour gestalt, the distribution of pitch prominences, and pausing. The phrases of interest here are 4, 5, and 6, which contain two instances of the same 'topic + predication' structure *top off* by speaker C (age 25 months). Phrases 4 and 5 contrast with phrase 6. In the first instance, the topic + predication is distributed over two prosodic phrases, i.e. each part has its own intonation contour. This reflects the earlier one-word stage of speech production (cf. the similar pattern in phrases 2–3), even though the syntactic and semantic structure is now more complex. In the second instance, the topic + predication is contained in one prosodic phrase. Here, the prosodic structures of the two words have been integrated, mirroring the syntax and semantics of the construction.

In sum, on a preliminary assessment, the category 'prosodic phrase' appears compatible with a prototype model. Three of the four properties of prototypes discussed by Geeraerts (1989) seem applicable; two other nonstructural characteristics outlined by Lakoff (1987) seem to apply as well. These considerations, I believe, are promising in the development of a model of prosodic phrase that is anchored in phonetic substance.

2.2. Notions of prototype

At this point it will help clarify matters to mention that there are two general orientations when describing a prototype, as Cruse (1990) points out. Under one perspective, the relations between the members of a given category are in view: the focus is on the 'prototypical exemplar' (either as an idealization or as the 'most representative' member), against which the other (actual and potential) members of the category may be evaluated. In other words, the prototype serves as a cognitive reference point for the categorization of nonprototypical tokens. This is the notion of prototype advanced in Schuetze-Coburn et al. (1991) and Chafe (1992). Under the other perspective, the category as a whole is in view: the focus is instead on the prototypical characteristics of the category, and the properties which serve to define it in an intensional sense.[11] It is this latter perspective that is of primary interest here, as the ultimate goal is to advance a phonetically based characterization of the category prosodic phrase in terms of prototypical features. Regarding the applicable prototype properties discussed above, the phonetic features that will satisfy this goal will (a) be realized unpredictably (lack of criterial attributes), (b) cluster into reoccurring subsets (family resemblance), and (c) be associated with varying degrees of importance (centrality).

Numerous phonetic attributes occur with sufficient regularity, especially at the edges of phrases, so that they can be considered prototypical features of the prosodic phrase. These features include silent pauses (the absence of vocalization or 'offtime'), which occur BETWEEN phrases; pitch reset and accelerated speech at the BEGINNING of a phrase; lengthened speech and laryngealization or other low pitch phenomena at the END of a phrase; and various vocaliza-

[10] Transcription conventions: three dots indicate a silent pause (.3–.6 second); two dots, a shorter pause; durations for longer pauses are given in parentheses ( ). Utterances with no obvious lexical correspondence are enclosed in </ />. Prominence is indicated with a grave accent. Transcriber comments are given in double parentheses (( )).

[11] Cf. Geeraerts' (1989) feature analysis of prototype characteristics.

tions which generally indicate some sort of hesitation ('filled pauses'), such as *uhm*, usually occurring at the beginning of a phrase, but occasionally constituting a separate, delimiting phrase. Other features, such as overall intensity peak timing, intensity diminution, and voice quality modulation may have a scope which extends over the course of a phrase.[12] These phonetic features all play a role in addition to the phrase-internal feature 'coherent intonation contour', as manifested by a particular configuration of accents. All these cues have been discussed individually in some detail at one time or another by various authors. How they each make a contribution in the production and perception of prosodic phrases in connected speech is the idea worth exploring further.

3. INTONATION UNITS AND THE PROTOTYPE MODEL

I wish now to give a brief indication of the descriptive potential of a prototype-based model of prosodic phrase by examining its flexibility in a concrete application. A short (4′ 20″) exchange by two speakers of colloquial Standard German was selected from a longer recording of a spontaneous conversation. The text was transcribed and the utterances were divided into speaker turns. Subsequently, turn units were segmented into the prosodic phrases called 'intonation units' using the system of auditory analysis presented and outlined in Du Bois, Schuetze-Coburn, Cumming, and Paolino (1991; 1992); cf. also Chafe (this volume). The phonetic basis of intonation units makes this unit an appropriate selection as a representative of prosodic phrases as a whole.

3.1. Prosodic cues to intonation units

During the segmentation process, various phonetic cues are taken into consideration, primarily those mentioned above. In the transcription, systematic attention is devoted to four prosodic cues—silent pause, accelerated speech, lengthened speech, and laryngealization—which constitute the most important phonetic features for phrasing that we feel can be auditorily noted with adequate reliability.[13] In addition, some use is made of pitch reset (here, a marked shift in pitch, generally on a nonprominent syllable, at the beginning of an intonation unit). Where marked, it is derived from an inspection of the pitch tracks of the excerpt, together with a comparison of the pitch periods before and after the unit boundary by measuring the frequency of the glottal pulse from the digitized waveform. The gestalt perception of a 'coherent intonation contour' itself is difficult to quantify directly, but empirical evidence suggests that this is not necessary. In other words, while not all of the prosodic features that may be present in the signal are individually attended to in the transcription system, the segmentation process IS sensitive to factors which contribute to the perception of a phrasal-level contour gestalt, and cues that are important for the segmentation of each phrase are noted.[14]

3.2. Prosodic cue patterns

The transcribed excerpt contains 269 intonation units. Of these, 113 were not included in the tabulations presented below: 30 constitute solely 'nonlinguistic' vocalizations (primarily laughter); 83 include backchannel utterances, overlapping turns, and turn-initial intonation units—in short, all units that were already delimited by turn boundaries. The remaining 156 turn-internal intonation units were examined for the presence of the four main prosodic cues listed above; the results are given in Table 1. From the table, it it clear that the frequency of occurrence for the tabulated cues varies greatly in the selected excerpt. While about two-thirds of the intonation units are preceded by silent pauses, only one in ten is bounded by laryngealized speech.

Silent Pause	Acceler'd speech	Length'd speech	Laryng'd speech
105 (68%)	74 (47%)	34 (22%)	15 (10%)

Table 1. Occurrence of four prosodic cues in a conversational excerpt.

With regard to their place in a prosodic phrase model, it is noted that each feature exhibits the expected prototypical properties. While silent pauses of various lengths occur with some frequency, not every pair of intonation units is separated by a pause. Furthermore, phrase-internal pauses, though less common, also occur. Concerning accelerated and lengthened speech, while there is a tendency for the rate of speech to decelerate through the course of an intonation unit, tempo does not always vary in this way, and patterns are complex. That is, although stretches of accelerated speech often occur at the beginning of an intonation unit, rather than elsewhere, acceleration does not only correspond to unit beginnings: some units are perceived as consisting entirely of accelerated speech. Similarly, while segment and syllable lengthening does occur at the end of phrases, this cue is not limited to this position, and marked lengthening

[12] This list is not exhaustive, of course; one could point to additional regularly occurring phonetic cues as well.

[13] Pause durations, however, were checked instrumentally in conjunction with another study (Schuetze-Coburn, in progress); Estimated pauses proved to be valid; 96% of the transcribed pauses varied no more than ±.1 second from their acoustically measured counterparts (the claimed accuracy of the auditory judgements).

[14] Of course, other aspects of prosodic notation are not included here either; detailed representation of the pitch accents, for example, is left for other systems.

is much less common overall. Finally, while vocalization sometimes becomes laryngealized at the end of a phrase, or glottal constrictions occur phrase initially (or both the end of one phrase and the beginning of another are so marked), this feature also occurs within phrases, and its occurrence is fairly limited.

For the purposes of evaluating a prototype model, an aspect of feature occurrence more interesting than the frequency of individual features is the way which features pattern with each other. Combinations of prosodic cues for the tabulated intonation units are given in Table 2. In the table, feature combinations are read horizontally, with a minus sign indicating the absence of a feature, and a plus sign, its presence. Thus, the top row gives the number of cases where none of the four prosodic cues is present, i.e. [–PAUSE, –ACCELERATION, –LENGTHENING, –LARYNGEALIZATION], of which there are 12. That is, (reading across the table) out of the 51 intonation units which lack a preceding silent pause, 22 lack in addition initial accelerated speech; out of these 22 cases, 13 lack final lengthening; and out of these 13 cases, 12 have no laryngealization. In the second row, the number of cases where laryngealization is the only cue are given, i.e. [–PAUSE, –ACCELERATION, –LENGTHENING, +LARYNGEALIZATION], of which there is 1. And so on until the last row, which gives the number of cases where all four cues are present (i.e. 1 case).

Silent Pause	Acceler'd speech	Length'd speech	Laryng'd speech	
51 –	22 –	13 –	12 –	⇐ No cues
			1 +	
		9 +	8 –	
			1 +	
	29 +	21 –	20 –	(13%)
			1 +	
		8 +	8 –	
			0 +	
105 +	60 –	49 –	43 –	(28%)
			6 +	
		11 +	10 –	
			1 +	
	45 +	39 –	36 –	(23%)
			4 +	
		6 +	5 –	
			1 +	⇐ All cues

Table 2. Cooccurrence patterns for four prosodic cues in a conversational excerpt.
Feature combinations are read off horizontally; the top row is all cues absent; the bottom, all present. Percentages for combinations occurring in over 10% of tabulated intonation units are given to the right.

Certain configurations of cues are clearly much more common than others. Pause alone turns out to be the most frequent pattern, found in 28% of the tabulated intonation units. Pause plus accelerated speech is found in an additional 23%, and accelerated speech alone marks 13%. Together, these three feature configurations cover almost two-thirds of the tabulated units. Perhaps somewhat surprisingly, given the past emphasis on THE prototype for all prosodic phrases, it is also evident that the presence of all four features in a unit is a rare event (1%), just as it is relatively uncommon for none of these features to be present (8%). Instead, fully 85% of the tabulated units exhibit 1 or 2 prosodic cues. Thus, a TYPICAL intonation unit has at least one prosodic feature, but its boundaries are not marked to an extreme degree.

The best way to illustrate the variation in feature combinations which cue intonation units is to present an excerpt from the transcription and discuss the combination present in each instance. Example (2) is a stretch of nine intonation units by speaker A, who is talking about a friend planning to write funeral marches for rich Americans. (As before, each line in the transcription represents a separate intonation unit.)[15]

(2) 151 A: ~%also er würde er so~ /+
so he would he like

152 (H) «.3» ^ vòrher so 'n hàlbes Jàh=r /+ ◊
before like a half year

153 auf so (%) ‹% uh %› /--
on like uh

154 ^ bei dem lèb'm müss'n /+
by him live must

155 .. (H) «.4» ^ ~d'mit er sie richtig~ kènn'nlernt /+
so.that he them correctly gets.to.know

156 ^ bevòr die Leute stèrb'n /+
before the people die

157 ^ das wéiß [ja man] ja mèi=st'ns /+
that knows yes one yes mostly

158 B: [(CLEARS THROAT)]

15 Additional transcription conventions: accelerated speech is bracketed by tildes, lengthened segments are followed by an equal sign; laryngealization is indicated with percent signs. Pitch reset is indicated by a raised caret; 'focus' accent, by an acute accent mark. Inhalation is represented by (H), with its duration following in double angled brackets. Intonation units are marked for transitional continuity class as '+' (continuing) or '•' (final). Perceived terminal pitch direction is indicated with '\' (fall) or '/' (nonfall). Uncompleted units end in double hyphen '--'. Questionable unit boundaries are indicated with '◊'. Overlapping speech is enclosed in square brackets. Elision of segments is indicated by a "'. See Du Bois et al. (1991; 1992) for more details.

159 A: .. bevòr man so stìrbt /+
before one like dies

160 .. ~jedenfalls~ bèi% ◊ .. Lèuten die so länger kränk [sind \ nè] /+
in.any.case by people that like longer ill are ok

161 B: [jà] \+
yes

(A: 'So he would like / half a year before / on like uh– / have to live at his [house] / so that he really gets to know them / before people die / you usually know'/ B: (CLEARS THROAT)/ A: 'before one dies / at least with people who have been sick a long time' / B: 'yeah' /)

In this excerpt, there are eight intonation unit transition points that need to be discussed with respect to the prosodic cues signaling the phrase boundaries. Point 1 is the boundary between units 151 and 152. Here there is a short (.3 second) break in vocalization in conjunction with the inbreath between the units. Such a break is comparable in its timing to a silent pause; however, after an inhalation, there is a very strong tendency to reset one's pitch, and pitch reset is an important feature for cuing a new intonation unit boundary.[16] In this case, there is an obvious shift in pitch on *vorher* (of about 60 Hz, see Figure 1 @ time 149.25).[17] Point 2 is the boundary between units 152 and 153. One feature occurs here, lengthening on the final syllable of unit 152. While usually a fairly robust cue, the boundary in this instance is perceived to be rather weak, which is indicated by the diamond at the end of the line.

Point 3 is the boundary between units 153 and 154. Glottal constriction and creaky voice through *uh* is found at the end of unit 153 (which is perceived as uncompleted), followed by pitch reset on *bei* (not visible in Figure 1 @ time 150.75, as no pitch values registered for the preceding low pitched segments). Point 4 is the boundary between the next pair of units, 154 and 155. Here there is slight pause and then again an inbreath with a following pitch reset (of about 150 Hz, in this case a shift down, see Figure 1 @ time 152). In addition, there is an initial stretch of accelerated speech in unit 155.

Point 5 is the boundary between units 155 and 156. Of the prosodic features considered here, only pitch reset (of about 60 Hz, see Figure 1 @ time 153.5) is evident. Point 6 lies between units 156 and 157. Once again, the only cue present is pitch reset (of about 80 Hz, see Figure 1 @ time 154.5).

Point 7 is the boundary between units 157 and 159. Here there is both lengthening at the end of unit 157 and a slight pause between units. Point 8 is the boundary between units 159 and 160. There is a slight pause between units and accelerated speech at the beginning of unit 160 (one of the very common feature combinations). Note that in unit 160, glottal constriction occurs after *bei*, and a slight pause is perceived. An intonation unit boundary was considered at this point (indicated by the diamond in the line), but was ultimately rejected. (This situation contrasts with boundary point 2 above, where a boundary was indeed marked.)

4. SUMMARY

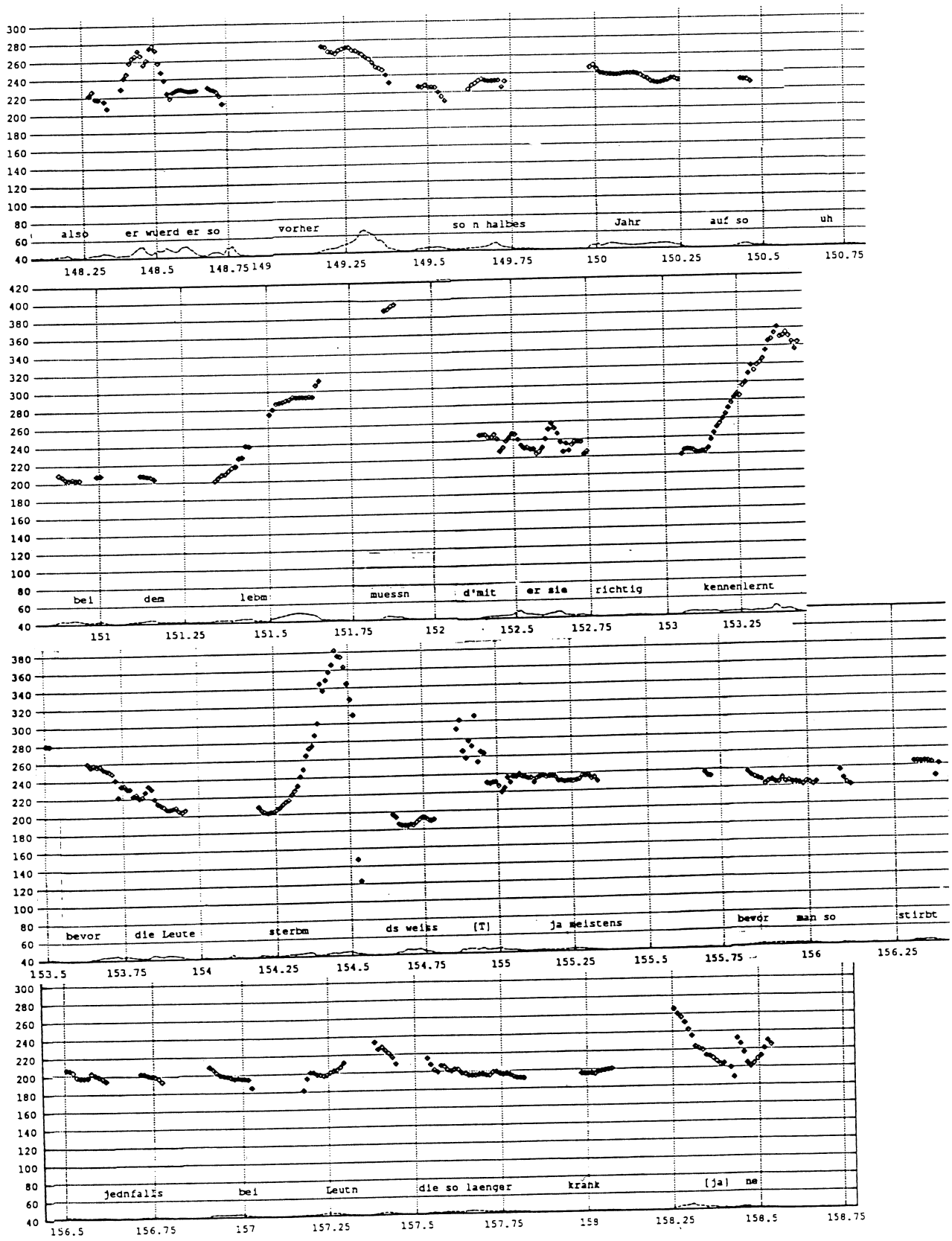
In the first section of my paper, I discussed some of the shortcomings of the usual approaches to a (general) linguistic unit 'prosodic phrase'. In the second section, I suggested that this category be treated as a prototype category, and I outlined typical characteristics that such a model might exhibit. In the third section, I examined the distribution of four prosodic features in an excerpt from a spontaneous German conversation which had been segmented into 'intonation units' in order to illustrate that these features meet expectations placed on prototypical characterizations of the category. It was found that an intonation unit will exhibit a variety of cues, yet rarely, if ever, do all phonetic features actually occur in any given instance. Nevertheless, it is the case that a conjuncture of cues is usually identifiable before a prosodic phrase is perceived. While a set of features may cooccur at particular points in a utterance—clearly identifying a prosodic boundary—it is also the case that individual features are found elsewhere, as illustrated above. That is, prosodic boundaries are manifested more or less strongly, depending in part on how many features are present, but the presence of a feature apparently does not guarantee a boundary. One problem, then, is that the feature threshold used to determine whether a boundary is identified—or is not selected—could (with our current understanding) be arbitrarily set by the researcher. The larger open question with phonetic cues is thus not their inventory, but their relative weight (e.g. pitch reset is arguably more central than laryngealization), and the interpretation of their interaction in influencing the perception of prosodic phrase boundaries. These are empirical questions which can be answered, but it will require a close analysis of a large amount of connected, preferably spontaneous, natural speech.

---

16 Schuetze-Coburn, Shapley & Weber (1991) found in a corpus of American English conversation that a new intonation unit was perceived every time the speaker's pitch was fully reset (but that this reset occurred on average only once every other intonation unit).

17 Octave errors in the pitch tracking have been adjusted manually.

Figure 1. F_0 tracings corresponding to example (2). Scale is Hz as a function of time in seconds.



REFERENCES

Altenberg, B. 1987. *Prosodic Patterns in Spoken English.* Lund: Lund University Press.

Brazil, D. 1975. *Discourse Intonation.* Birmingham Univ.

—— 1978. *Discourse Intonation II.* Birmingham Univ.

—— 1985. *The Communicative Value of Intonation in English.* Birmingham Univ.

Brown, G.; K. L. Currie & J. Kenworthy. 1980. *Questions of Intonation.* London: Croom Helm.

Butterworth, B. 1983 (ed.) *Language Production II.* London: Academic Press.

Chafe, W. L. 1980. Some reasons for hesitating. In Dechert & Raupach (eds.), 169-80.

—— 1987. Cognitive constraints on information flow. In Tomlin (ed.), 21-51.

—— 1992. Discourse, Consciousness, and Time. Ms., UCSB.

—— This volume. Prosody in English and Seneca natural discourse.

Couper-Kuhlen, E. 1986. An Introduction to English Prosody. Tübingen: Max Niemeyer.

Cruse, D. A. 1990. Prototype theory and lexical semantics. In Tsohatzidis (ed.), 382-402.

Cruttenden, A. 1970. On the so-called grammatical function of intonation. *Phonetica* 21.182-92.

—— 1986. *Intonation.* Cambridge: Cambridge Univ. Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English.* Cambridge: Cambridge University Press.

—— 1975. *The English Tone of Voice: Essays in intonation, prosody and paralanguage.* London: Edward Arnold.

Crystal, D. & R. Quirk. 1964. *Systems of Prosodic and Paralinguistic Features in English.* The Hague: Mouton.

Cutler, A. & D. R. Ladd. 1983 (eds.) *Prosody: Models and Measurements.* Berlin: Springer-Verlag.

Dechert, H. W. & M. Raupach. 1980 (eds.) *Temporal variables in speech.* The Hague: Mouton.

Deese, J. 1980. Pauses, prosody, and the demands of production in language. In Dechert & Raupach (eds.), 69-84.

Du Bois, J. W.; S. Schuetze-Coburn; D. Paolino & S. Cumming. 1991. *Discourse Transcription.* Ms., UCSB.

—— 1992. Outline of discourse transcription. In Edwards & Lampert (eds.).

Edwards, J. A. & M. D. Lampert. 1992 (eds.) *Talking Data.* Hillsdale, NJ: Lawrence Erlbaum.

Enkvist, N. E. 1982 (ed.) *Impromptu Speech: A symposium.* Åbo: Åbo Akademi.

Fillmore, C. J. 1975. An alternative to checklist theories of meaning. BLS 1, 123-31.

Ford, C. E. & S. A. Thompson. 1992. Projectability in conversation. Ms., UCSB.

Fox, A. 1984. *German Intonation.* Oxford: Clarendon Press.

Fretheim, T. 1981 (ed.) *Nordic Prosody II.* Trondheim: Tapir.

Geeraerts, D. 1989. Introduction: Prospects and problems of prototype theory. *Linguistics* 27.587-612.

Gibbon, D. 1984. Intonation as an adaptive process. In Gibbon & Richter (eds.), 165-192.

—— 1988. Intonation and discourse. In Petöfi (ed.), 3-25.

Gibbon, D. & H. Richter. 1984 (eds.) *Intonation, Accent and Rhythm.* Berlin: de Gruyter.

Goodwin, C. 1981. *Conversational Organization.* New York: Academic Press.

Halliday, M. A. K. 1963. The tones of English. *Archivum Linguisticum* 15.1-28.

—— 1967a. *Intonation and Grammar in British English.* The Hague: Mouton.

—— 1967b. Notes on transitivity and theme in English, Part 2. *Journal of Linguistics* 3.199-244.

—— 1985. *An Introduction to Functional Grammar.* London: Edward Arnold.

van der Hulst, H. & N. Smith. 1982 (eds.) *The Structure of Phonological Representations I.* Dordrecht: Foris.

Jaeger, J. J. 1986. Concept formation as a tool for linguistic research. In Ohala & Jaeger (eds.), 211-37.

Kreckel, M. 1981. Tone units as message blocks in natural discourse. *Journal of Pragmatics* 5.459-76.

Ladd, D. R. 1986. Intonational phrasing. *Phonology Yearbook* 3.311-40.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things.* Chicago: Chicago University Press.

Laver, J. 1970. The production of speech. In Lyons (ed.), 53-75.

Lyons, J. 1970 (ed.) *New Horizons in Linguistics.* Harmondsworth: Penguin.

Menn, L. 1983. Development of articulatory, phonetic and phonological capabilities. In Butterworth (ed.), 3-50.

Nathan, G. S. 1986. Phonemes as mental categories. BLS 12, 212-23.

Nespor, M. & I. Vogel. 1982. Prosodic domains and external sandhi rules. In van der Hulst & Smith (eds.), 225-55.

—— 1983. Prosodic structures above the word. In Cutler & Ladd (eds.), 123-40.

Ohala, J. J. & J. J. Jaeger. 1986 (eds.) *Experimental Phonology.* Orlando: Academic Press.

Oreström, B. 1983. *Turn-taking in English Conversation.* Lund: CWK Gleerup.

Palmer, H. E. 1924. *English Intonation with Systematic Exercises (2nd edition).* Cambridge: W. Heffer.

Petöfi, J. 1988 (ed.) *Text and Discourse Constitution.* Berlin: de Gruyter.

Pierrehumbert, J. B. 1980. The Phonology and Phonetics of English Intonation. Ph.D. diss., MIT.

Pike, K. L. 1945. *Intonation of American English.* Ann Arbor: University of Michigan Press.

Rosch, E. 1978. Principles of Categorization. In Rosch & Lloyd (eds.), 27-48.

Rosch, E. & B. Lloyd. 1978 (eds.) *Cognition and Categorization.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rudzka-Ostyn, B. 1988 (ed.) *Topics in Cognitive Linguistics.* Amsterdam: John Benjamins.

Schuetze-Coburn, S. In progress. The measurement and representation of pauses in natural discourse.

Schuetze-Coburn, S.; M. Shapley & E. G. Weber. 1991. Units of intonation in discourse. *Language and Speech* 34.207-234.

Selkirk, E. O. 1981. On prosodic structure and its relation to syntactic structure. In Fretheim (ed.), 111-40.

Svartvik, J. 1982. The segmentation of impromptu speech. In Enkvist (ed.), 131-145.

Taylor, J. R. 1989. *Linguistic Categorization.* Oxford: Clarendon Press.

Tomlin, R. S. 1987 (ed.) *Coherence and Grounding in Discourse.* Amsterdam/Philadelphia: Benjamins.

Trager, G. L. & H. L. Smith. 1951. *An Outline of English Structure.* Norman, OK: Battenburg Press.

Tsohatzidis, S. L. 1990 (ed.) Meanings and Prototypes. London: Routledge.

Vaissière, J. 1983. Language-independent prosodic features. In Cutler & Ladd (eds.), 53-66.

Wittgenstein, L. 1953. *Philosophical Investigations.* Oxford: Blackwell.

Pitch Accent Placement within Words

S. Shattuck-Hufnagel[1] M. Ostendorf[2] K. Ross[2]

October 29, 1992

[1]Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[2]Boston University, Boston, MA 02215 USA

Abstract

Two aspects of prosodic structure have been suggested as factors contributing to the early placement of prominence, sometimes called 'stress shift', in late-main-stress words: these two factors are *rhythmic regularity* and *onset location of the pitch accent* in its prosodic phrase. This paper reports data from a corpus of FM radio news stories showing support for both factors. When listeners label phrase-level prominences for each syllable in an utterance, they tend to report that the speaker has placed a prominence early in a late-main-stress target word when either (a) the first syllable of the following word also bears a prominence (rhythmic clash) or (b) the target word carries the first prominence of the prosodic phrase (onset marking). We also examined some of the acoustic correlates of early prominence labeling in a subset of the same target words. When a syllable early in the word (i.e. before the main-stress syllable) is labeled with a phrase-level prominence, it shows a substantial F0 movement compared with the same syllable in non-early-accent examples. These findings support the hypothesis that apparent stress shift is, at least in part, a matter of early pitch accent placement within the word.

1 Introduction

For many years observers have noted that some words of American English can be produced with their major prominence on a syllable to the left of their main-stress syllable, under certain conditions. The set of words that can undergo this apparent 'stress shift' include those with a) main stress located late in the word, and b) a full-vowel syllable earlier in the word, which can carry the prominence; examples often cited in the literature include *Massachusetts* (as in "Massachusetts miracle"), *thirteen* (as in "thirteen men"), and *Japanese* (as in "Japanese food").[1] Dictionaries

[1]Gussenhoven suggests that not all pre-main-stress full-vowel syllables are candidates for stress shift [1].

sometimes capture this phenomenon by marking both unreduced vowels as main-stress vowels, or by putting a main-stress marker in parenthesis on the earlier vowel. Phoneticians occasionally described this phenomenon [2], but the first systematic treatment in the context of a general theory of prosody was by Vanderslice and Ladefoged in 1972 [3]. They suggested that in words like *telegraphic*, both of the full-vowel syllables should be marked as accentable, "leaving the determination of which accents are ultimately to be realized phonetically to the care of a late rhythm rule..." (p. 829). In subsequent work, the two major accounts of apparent shift have been based on rhythm and on intonation. We will briefly summarize these two approaches before presenting the results of our perceptual and acoustic analyses.

2 Phonological Accounts of Apparent Stress Shift

2.1 Rhythm-based theories

Liberman (1975) [4] gave the phenomenon of stress shift an important role as evidence for his theory of metrical phonology. Since then, a number of theoretical proposals have suggested that stress shift provides a persuasive argument in support of the metrical grid, a representational device which indicates the degree of stress on a syllable by the number of marked cells in the vertical column above that syllable in a nucleus-based matrix [5, 6, 7, 8, 9]. In these models, lexical stress corresponds to the marking of a certain number of cells in the column; when words are concatenated into a phrase and phrase-level stress is assigned, the resulting pattern of columns of cell markings in the grid may create a *stress clash*. That is, two adjacent or nearly-adjacent syllables may be rhythmically strong, so that a tendency toward the placement of heavy stresses at more equal intervals requires that one of them be moved. Observation has shown that the left-hand member of a pair of clashing stresses, which (according to the Nuclear Stress Rule of English) is normally the weaker one, appears to shift further left, away from the clashing (and stronger) stress on its right. An example that is often cited to illustrate the claims of this approach compares *the Mississippi legislator* with *the Mississippi legislation*. The former phrase embodies a stress clash, and so might be expected to undergo stress shift; the latter would not.

In these stress-based theories, rhythmic stress is viewed as a unitary phenomenon, whether it occurs at the lexical or the phrasal level; all x-marks in the cells of the grid are of the same variety, and differences in degree of stress are reflected simply in the number of such marks in the vertical columns. Phrase-level stress occurs on the main-stress syllables of words, except when this results in rhythmic irregularity, in which case it can move to an earlier syllable in its word. The acoustic correlates of the moved prominence are not specified.

2.2 Pitch-accent-based theories

Concurrent with these developments in phonological theory, relevant developments were also occurring in a different domain: models of intonation. In 1965, Bolinger [10] suggested that there were two main kinds of markers for phrasal prominence: one pitch marker or accent early in the phrase, and another late in the phrase. He noted that the tendency to put the first pitch accent as early as possible in the phrase might lead to its placement on a syllable to the left of the main-stress syllable, for late-stress words. In 1975, 't Hart and Collier [11] described the IPO model of intonation for Dutch, based on perceptual equivalence of simplified F0 contours, and noted that the common 'hat pattern' for declarative sentences included an 'onset rise'. Although they did not explicitly discuss the relevance of the onset rise to apparent stress shift, their model for Dutch reflects the same observation that Bolinger reports for English: the common occurrence of a pitch marker early in the phrase.[2]

Shattuck-Hufnagel (1988) [15], building on this observation, proposed an account of apparent stress shift in speech production planning. In this model, speakers have the option of placing pre-nuclear pitch accents on full-vowel syllables to the left of the main-stress syllable of a late-main-stress word, and they exercise this option with particular frequency when the target word carries the first pitch accent of its phrase. Subsequent expansions of this proposal [16, 17] have suggested that this tendency, combined with the location of phrase-final pitch accents on the main-stress syllable, might help the listener to identify the first and last pitch accents of a phrase.

In 1980, Pierrehumbert [18] proposed a grammar of intonation for American English that combined an account of pitch accent prominence with an account of the tonal markers of prosodic boundaries. In 1986 Beckman and Pierrehumbert [19] elaborated this model to include two kinds of prosodic phrases: the **intermediate phrase**, marked by at least one pitch accent and by the presence of a phrase accent that determines the intonation contour between the final pitch accent and the right boundary, and the **intonational phrase**, consisting of one or more intermediate phrases and marked by an additional boundary tone at its right boundary, realized on the final syllable of the phrase. Beckman and Edwards (forthcoming) [20], working in this framework, have suggested that apparent shift may occur because speakers tend to accent the first accentable syllable in a new intermediate phrase. Gussenhoven (1991) [1] has proposed a phonological model which reaches a similar result via a different mechanism. In his model, all accentable syllables receive an accent and the accents on alternate syllables are removed starting from the right-most accent. Monaghan (1990) [21] has also proposed an algorithm of alternate-accent deletion for prosody synthesis.

In all of these approaches, whether from the point of view of cognitive processing, phonological theory or synthesis algorithms, apparent stress shift is said to result

[2]The IPO model was subsequently extended to British [12, 13] and to American [14] English.

from a) the placement of a pre-nuclear pitch accent on the early full-vowel of a late-main-stress target word, plus b) the disappearance of the nuclear (or final) pitch accent from its main-stress syllable. For example, in the single-word phrase

```
Massachusetts
*     **
```

the pre-nuclear accent might occur on **Ma-** and the nuclear accent on **-chu-**. But in the longer phrase

```
The Massachusetts miracle
*              **
```

the nuclear pitch accent no longer occurs on **-chu-**; by the Nuclear Stress Rule of English, it occurs on **mir-** in the following word. Since the pre-nuclear accents in both the single-word phrase and the longer phrase are optional, predictions about acoustic comparisons between the two cases are not straightforward; before testable predictions can be made, we must to determine which options the speaker selected for pitch accent location in the particular pair of utterances being compared.

In these pitch-accent-based theories, prominence is not viewed as a unitary phenomenon. Instead, the type of prominence (and its acoustic correlates) varies systematically with the level of constituent in the prosodic hierarchy; phrase-level prominence is a matter of pitch accents, whether those pitch accents occur on the lexically main-stressed syllable or on other full-vowel syllables in the word. Such theories predict that apparent stress shift will be associated with the kinds of F0 markers that normally signal a pitch accent.

Empirical studies relevant to the stress-based and pitch-accent-based approaches to apparent stress shift rely on two kinds of observations: judgments of prominence placement [22], and measurements of acoustic parameters believed to be correlates of perceived prominence, e.g. duration and F0 [23, 24]. Few studies have combined these two approaches [25]. As a result, if the theories do not predict with complete accuracy where shift will occur, the acoustic measurements that have been reported may not always correspond to actual cases of perceived shift. As part of a larger study of prosody, we have begun to address this problem in an extended analysis of a corpus of FM radio news stories. We have begun with perceptual labeling of perceived prominence, major boundary tones and the prosodic boundaries between each pair of adjacent words, as described in Price *et al.* (1992) [26]. Eventually we plan to have the corpus labeled for pitch accent type, boundary tones and break-index values, as described in [27], and to have acoustic analyses of segment and syllable duration and F0 patterns. Here we report results from the initial prominence and break index labeling effort for a portion of the corpus, and from acoustic analyses of a smaller subset of stress-shift candidate words, that are relevant to the issue of where pitch accents appear within words. The questions we asked were 1) Does onset position in the prosodic phrase, as well as rhythmic clash, influence the early

placement of pitch accents in the word? 2) Does the final pitch accent of the phrase occur on the main-stress syllable of its word? and 3) When a syllable is judged to have early accent placement, does it show a substantial F0 marker, compared to non-early-accent utterances of the same word?

3 Analysis of Radio News Speech

3.1 Methods

Both studies reported here are based on a corpus of recorded FM radio news broadcasts spoken by two female radio announcers. A total of twenty-five radio news stories (7880 words) were available for this study. The database is described in more detail in Ross *et al.*, (1992) [28] where some of these results are also presented. For the acoustic measurements, we focus on a small subset of this corpus: 44 instances of the early-accent candidate word *Massachusetts* that occurred in these stories. This data set is useful because it offers the control of a single word, and yet the word appears in naturally occurring paragraphs of speech, as opposed to isolated laboratory sentences. The stories were initially hand-labeled for phrasal prominences (syllables are labeled either prominent or unmarked) and seven levels (0-6) of prosodic boundaries, based on listening only. For this study, only the intermediate phrase boundaries (3 or higher) are relevant. The words in these stories that were candidates for early accent placement (i.e. late-main-stress words with an earlier unreduced vowel[3]) were then verified by listeners with access to visual displays of F0 contours, and in some cases additional pitch accents were labeled.

3.2 Study 1: Phrasal prominence location within words

We classified the words that were candidates for early accent placement according to their accent labeling. Table 1 shows the categories in columns 1, 2 and 3: accent on the early syllable only, on both the early and main-stress syllables and on the main-stress syllable only. A fourth category, for words that did not receive a phrasal prominence label, is omitted here. In addition, we classified the target words according to the serial position of their accent in the intermediate phrase: first accent (but not last), middle accent (neither first nor last), last accent (but not first), with a separate category for target words that contain all the accents in the phrase. Finally, we distinguished among different contexts: clash context (prominence on the initial syllable of the following word in the phrase), possible clash (prominence later in the following word) and no clash (no prominence on the following word). These distinctions allow us to ask the following questions.

1) Does early accent placement occur with greater frequency under conditions of rhythmic clash?

[3]See Shattuck-Hufnagel (1992) [16] for more detailed discussion of the criteria for words included as early accent candidates.

Accent Location in Phrase	Accent Location in Next Word		Early Accent	Two Accents	Main Stress	Total
			1	2	3	
First	First Syll.	A	20	3	6	29
	Non-first Syll.	B	3	4	8	15
	No Accent	C	12	7	13	32
Medial	First Syll.	D	12	1	2	15
	Non-first Syll.	E	5	2	7	14
	No Accent	F	6	0	10	16
Final	N/A	G	4	7	62	73
All	N/A	H	5	12	33	50

Table 1: Pitch accent placement in early-placement candidate words in a corpus of FM radio speech. Horizontal divisions show the serial position of target-word pitch accents in phrases (first (but not last), middle (neither last nor first), last (but not first), all). Rows within horizontal divisions show position of the pitch accent in the following word, reflecting clash context. Columns 1, 2 and 3 indicate the position of the pitch accent within the target word (early syllable only, main-stress syllable only, double accent). Deaccented words are not included.

In our terms, rhythmic clash occurs when the word following the late-main-stress target word has a phrasal prominence on its initial syllable. The entries in rows A and D, column 1, of Table 1 show that when the following word has a prominence on its initial syllable, early placement is likely to occur (32 of 44 words, or 73%). When the following word does not have initial prominence (rows B, C, E and F), early placement (column 1) is less likely to occur (26 of 77 words, or 34%). Clearly, early placement is more likely to occur under conditions of accent clash in this corpus.

2) Does early accent placement occur, even without a clash, when the accent is the initial accent in an intermediate phrase? n Comparing the entries in row C, columns 1 and 2, with row F, columns 1 and 2, we see phrase-initial accents occur early in their words in 59% of cases (19 of 32 words), while middle accents occur early in only 38% of cases (6 of 16 words.) For this comparison, we include the double-accented words, because the early-pitch-accent model gives the speaker the option of placing multiple accents on a single word. The stress-based account has more difficulty accounting for these instances, since the prominence on the early syllable is assumed to have been moved there from its former location on the main-stress syllable.

As others have pointed out [3], phonological theory suggests that phrase-final or nuclear accents will occur on the main-stress syllable of their words. The entries in rows G and H of Table 1 show that this is the case: 93% (114 of 123) of the

Data set	Early Accent	Two Accents	Main Stress	No Accents
Massachusetts	15	3	17	9
25 Stories	67	36	141	65

Table 2: Number of occurrences of different classes of accenting for early accent candidate words in two data sets.

phrase-final accents in this corpus appear on a main-stress syllable. We include the double-accented words here, because their second accent always appears on the main-stress syllable. Finally, when the target word contains all of the accents in its intermediate phrase (row H of Table 1), it often has two accents (24%, or 12 of 50, compared with 12% for the rest of the examples.) This finding supports the claim that speakers tend to place an onset accent on an early full-vowel syllable, and a final accent on a main-stress syllable of the words in an intermediate phrase.

3.3 Study 2: Acoustic correlates of early-accent syllables

If perceived stress shift is at least partly a matter of early pitch accent placement within the word, then syllables labeled with early prominence should show the acoustic correlates of a pitch accent. To test this prediction, we carried out acoustic analyses of 44 instances of the word *Massachusetts* that occurred in the labeled corpus. The phrases containing *Massachusetts* were analyzed interactively using the Klattools developed by Dennis Klatt for use on the Vax. Time offsets were obtained for each segment in the target word, and F0 contours were estimated using Klatt's method, which compares the spacing between individual harmonics in adjacent pitch periods. From these analyses we obtained 1) the size and direction of the largest continuously-measurable F0 change in each syllable, and 2) the duration of each syllable. Here we will compare these measures for the word-initial syllable *Ma-* in two types of utterances: those labeled with a prominence only on the main-stress syllable *-chu-* (17 examples) and those marked with a prominence on the early syllable *Ma-* (15 examples on *Ma-* only, 3 examples with double accent.) The distributions of accents within words for the larger set of early-accent candidates and for the subset of occurrences of *Massachusetts* are shown in Table 2.

F0 comparisons. One of the 18 examples of early accent labeling was omitted from this analysis because diplophonia made it difficult to estimate the F0 contour on the initial syllable of the target word. The remaining 17 showed a substantial F0 rise in the first syllable of the word. The extent of this rise ranged from 7 Hz to 86 Hz, with a mean of 44 Hz. A typical example is shown in Figure 1. In contrast, F0 in the initial syllable of non-early-accent examples did not show a consistent pattern. Here again, one utterance had to be omitted because of diplophonia. Of the remaining 16 examples, 6 had a rising F0, 7 showed a fall, one showed a fall-rise

and two showed little or no change. The absolute value of the largest continuous F0 change in these syllables, for those that had one, ranged from 6 to 32 Hz, with a mean of 16 Hz. A typical example is shown in Figure 2. For these two FM radio speakers, there appears to be no question that listeners' judgments of early prominence location corresponds to a substantial F0 marker of the kind that might be expected for a pitch accent. Corresponding syllables in utterances without an early accent do not have this F0 marker.

For all of the 17 examples with early pitch accent placement, including the three with double accent placement, the F0 pattern for the main-stress syllable *-chu-* showed a substantial fall (see Figure 1). The size of the fall ranged from 11 Hz to 46 Hz, with a mean of 23 Hz. Since the break index following these target words was 2 or less, it is unlikely that this fall in F0 was associated with a low boundary tone or phrase accent. Moreover, the F0 values for the unaccented main-stress syllable generally overlapped with those in the accented first syllable. In contrast, for the 16 examples with main-stress-only accent placement, the F0 of the accented main-stress syllable was substantially higher than in the first syllable; in most cases, the lowest measurable F0 in the accented syllable 3 was 20 or more Hz higher than the highest measurable F0 in the unaccented syllable 1 (see Figure 2). This contrast suggests that, if voicing had not been interrupted by the *-ch-* at the syllable onset, we might have seen a substantial pitch marker (i.e. a rise in F0) in the accented third syllable. The F0 of an accented syllable is not, of course always higher than for an unaccented syllable. But in this case it does suggest that the contrast between early-prominence-only and main-stress-prominence-only words lies in the placement of pitch accents.

Duration comparisons. Preliminary comparison of syllable durations for *Ma-* in *Massachusetts* showed no noticeable difference between the early-accent and non-early-accent cases. Since other analyses we have carried out both in this corpus and other corpora suggest that there may be a difference in duration between accented and unaccented syllables, the question requires further study.

4 Discussion

Results of our analysis of perceived prominence location and of syllable duration and F0 contours for 44 utterances of the word **Massachusetts**, as well as for the larger corpus of FM radio speech, show that early-prominence location in early-accentable words is influenced by two tendencies exhibited by speakers: to avoid the rhythmic clash that would arise if two pitch accents occurred on nearly-adjacent syllables in adjacent words, and to place the first accent of an intermediate phrase on the earliest possible syllable of its word. In addition, results show that this early prominence corresponds to a substantial rise in F0 on the prominent syllable, suggesting that apparent stress shift is, in many cases, the result of patterns of pitch accent placement.

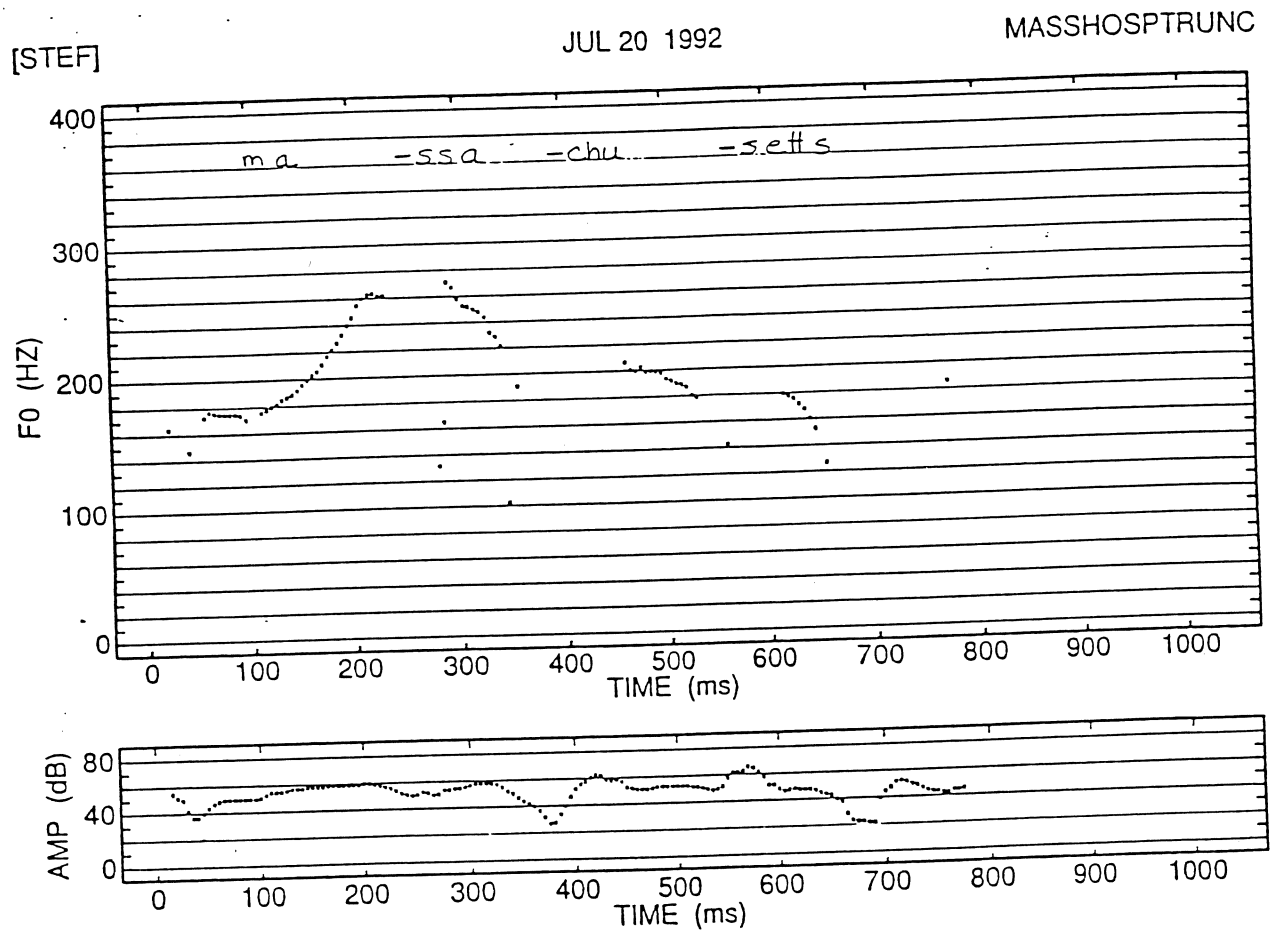


Figure 1: F0 contour for an example of the word *Massachusetts* labeled with an accent on the first syllable. Note the large F0 rise on the accented syllable *Ma-*. The small fall on the non-accented vowel *-u-* lies within the F0 values defined by the first-syllable rise.

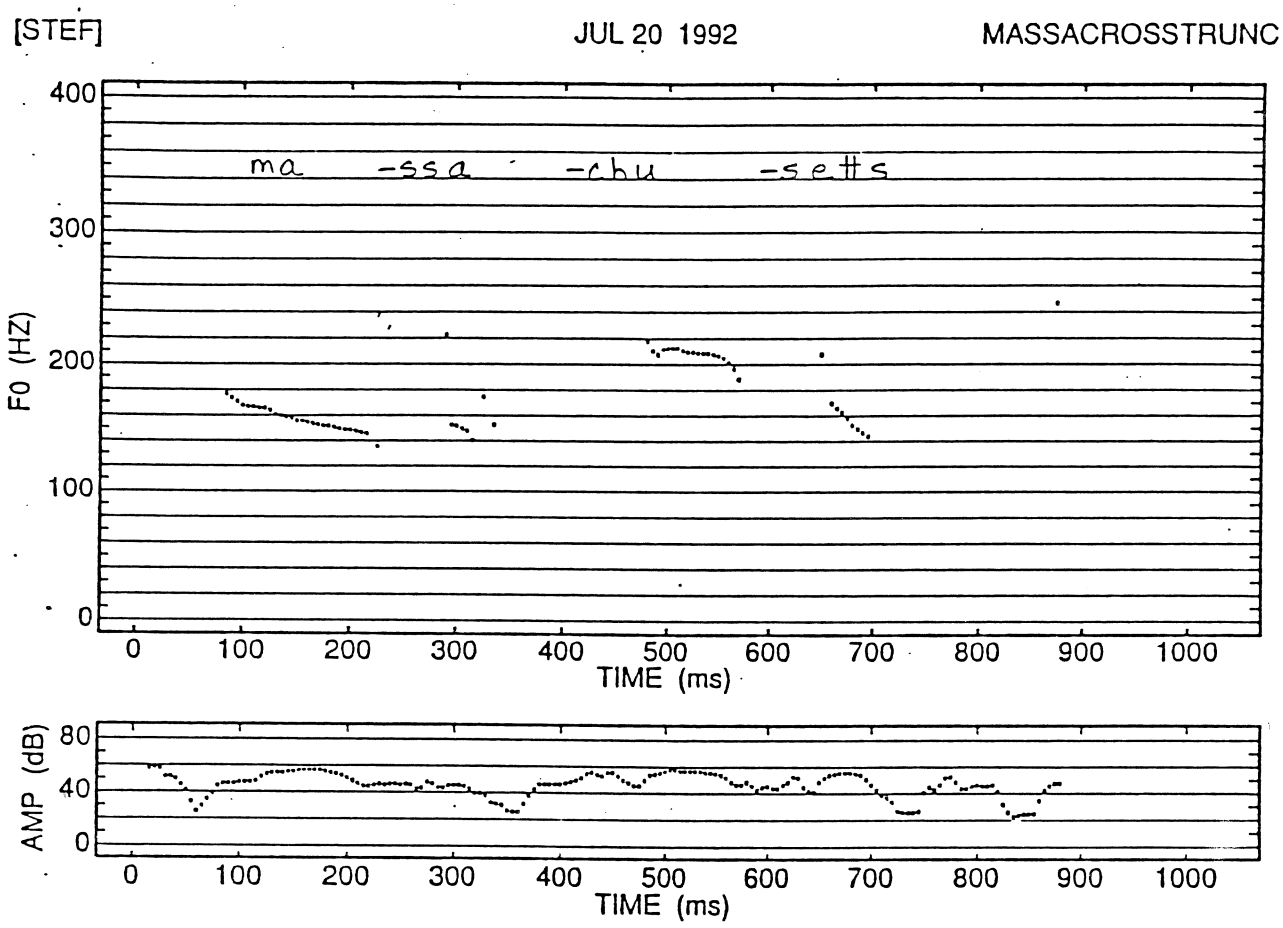


Figure 2: F0 contour for an example of the word *Massachusetts* labeled with an accent on the third (main-stress) syllable. The F0 fall on the non-accented *Ma-* is moderate in size. The F0 fall on the accented vowel *-u-* is also moderate but lies well above the F0 values for *Ma-*.

Acknowledgments

This research was funded by NSF under grant number IRI-8805680 with additional support coming from NIH, award number NIH 8-RO1-DCO-0075, and the Department of Education under the Graduate Assistance Applied to National Needs Program, award number P200A90080.

References

[1] C.Gussenhoven, "The English Rhythm Rule as an Accent Deletion Rule," *Phonology* 8, 1-35, 1991

[2] D. Jones, *An Outline of English Phonetics,* Cambridge: Heffer, 1917, 9th edition 1964

[3] R. Vanderslice and P. Ladefoged, "Binary Suprasegmental Features and Transformational Word-Accentuation Rules," *Language,* 48, 819-838, 1972

[4] M. Liberman, *The Intonational System of English,* Massachusetts Institute of Technology Ph.D. Dissertation, 1975

[5] M. Liberman and A. Prince, "On Stress and Linguistic Rhythm," *Linguistic Inquiry,* 8, 249–336, 1977.

[6] A. Prince, "Relating to the Grid," *Linguistic Inquiry,* 14, 19-100, 1983

[7] E. Selkirk, *Phonology and Syntax: The relation between sound and structure,* Cambridge, Mass.: MIT Press, 1984

[8] B. Hayes, "The Phonology of Rhythm in English," *Linguistic Inquiry,* 15, 33-74, 1984

[9] M. Nespor and I. Vogel, *Prosodic Phonology,* Dordrecht: Foris, 1986

[10] D. Bolinger, "Pitch Accent and Sentence Rhythm," *Forms of English: Accent, Morpheme, Order,* ed. I. Abe and T. Kanekiyo, Tokyo: Hokuou, 1965

[11] J. 't Hart and R. Collier, "Integrating different levels of intonation analysis," *Journal of Phonetics,* 3, 235-255, 1975

[12] J.R. de Pijper, *Modeling British Intonation,* Dordrecht: Foris, 1983

[13] N. Willems, *English Intonation from a Dutch Point of View,* Dordrecht: Foris, 1982

[14] S. Maeda, "A Characterization of Fundamental Frequency Contours of Speech," *MIT RLE Quarterly Progress Report 114,* 193-211, 1974

[15] S. Shattuck-Hufnagel, "Acoustic-Phonetic Correlates of Stress Shift," *Journal of the Acoustical Society of America,* 84, S98, 1988

[16] S. Shattuck-Hufnagel, "Stress Shift as Pitch Accent Placement," *Proceedings of the 1992 International Conference on Spoken Language Processing,* Banff, 747-750, 1992

[17] Shattuck-Hufnagel, S. "Stress shift as early pitch accent placement: a comment on Beckman and Edwards," *Proceedings of Labphon III,* ed. P. Keating *et al.,* forthcoming

[18] J. Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Massachusetts Institute of Technology Ph.D. Dissertation, 1980

[19] M. Beckman & J. Pierrehumbert, "Intonational Structure in Japanese and English," *Phonology Yearbook 3,* ed. J. Ohala, 255-309, 1986

[20] M. Beckman and J. Edwards, "Articulatory Evidence for Differentiating Stress Categories," *Proceedings of Labphon III,*, ed. P. Keating *et al.,* forthcoming

[21] A. Monaghan, "Rhythm and stress-shift in speech synthesis," *Computer Speech and Language,* 4, 71-78, 1990

[22] M. Beckman, M.G. Swora, J. Rauschenberg and K. De Jong, "Stress Shift, Stress Clash and Polysyllabic Shortening in a Prosodically Annotated Discourse," *Proceedings of the 1990 International Conference on Spoken Language Processing,*, Kobe, Japan, 1990

[23] W. Cooper and S.E. Eady, "Metrical Phonology in Speech Production," *Journal of Memory and Language,* 25, 369-384, 1986

[24] M. Horne, "Empirical Evidence for a Deletion Formulation of the Rhythm Rule in English," *Linguistics* 28, 959-981, 1990

[25] S. Shattuck-Hufnagel, "Acoustic Correlates of Stress Shift," *Proceedings of the XII International Congress of Phonetic Sciences,* Aix-en-Provence, 1991

[26] P. Price, M. Ostendorf, S. Shattuck-Hufnagel and C. Fong, "The Use of Prosody in Syntactic Disambiguation," *Journal of the Acoustical Society of America,* 90, 2956-2970, 1991

[27] K. Silverman *et al.,* "A Standard Scheme for Labeling Prosody," *Proceedings of the International Conference on Spoken Language Processing,* Banff, 867-870, 1992.

[28] K. Ross, M. Ostendorf, S. Shattuck-Hufnagel, "Factors affecting pitch accent placement," *Proceedings of the International Conference on Spoken Language Processing,* Banff, 365-368 1992

VARIATIONS OF THE MANDARIN RISING TONE*

Chilin Shih
Department of East Asian Languages and Cultures
Rutgers University
College Avenue Campus
New Brunswick, NJ 08903

Richard Sproat
Linguistics Research Department
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

ABSTRACT

This paper uses the Mandarin rising tone (tone 2) as an example to illustrate the range of tonal variation in speech, from words read in isolation, words read in sentence frames, to words produced in conversation. It is shown that the 2nd Tone Sandhi Rule (2TS) described by Chao [1] is the result of a phonetic implementation rule, which applies to the low target of a rising tone in high tone context when the rising tone in question is in prosodically weak positions. The amount of pitch drop to the low target of a rising tone varies with the prosodic strength of the syllable. As a result, the pitch contour of an extremely weak rising tone in high tone context approaches the shape of a high level tone.

1. Background

Mandarin has four lexical tones: High level, Rising, Low, and Falling. The following table lists the tones in the traditional nomenclature, pitch contour, and tonal targets in H (high) and L (low).

Name	Pitch Contour	Targets
Tone 1	High level	H
Tone 2	Rising	LH
Tone 3	Low	L
Tone 4	Falling	HL

Table 1: Mandarin Tones

Tone 3 has other variants. Most notably, a falling-rising shape may surface in the sentence final position. Therefore the tone is also known as a dipping tone. In addition to the four tonal categories, some suffixes in Mandarin do not have an inherent tone. A syllable without tone is weak; its duration is typically only half of a corresponding tone-carrying syllable. The pitch value of a toneless syllable is determined primarily by the preceding tone. The lack of tone on a syllable is traditionally referred to as neutral tone, or tone 0.

We take Y.R. Chao's ([1]) description of the 2nd Tone Sandhi (2TS) as a starting point:

> "A tone sandhi of minor importance has to do with the change of the 2nd to a 1st Tone in three-syllable groups. If in a three-syllable word or phrase ABC, A is in the 1st or 2nd Tone, B in the 2nd Tone, and C in any except the neutral tone ... then B changes into the 1st Tone for speech at conversational speed, but does not change at a more deliberate speed." (p. 27-28).

The examples Chao gave include the following. (We use the Pinyin romanization system through out this paper. Angle brackets highlight a *changed* tone: [1] indicates a phonetic tone 1 with an underlying tone 2 source).

1. xi1 yang2 shen1 → xi1 yang[1] shen1
   west-ocean-ginseng
   "western ginseng"
2. cong1 you2 bing3 → cong1 you[1] bing3
   green-onion oil pancake
   "green onion pancake"
3. long2 fu2 si4 → long2 fu[1] si4
   Long Fu Temple
4. Shei2 neng2 fei1? → Shei2 neng[1] fei1
   who can fly
   "Who can fly?"

* This research is supported by National Science Foundation Grant BNS-9021274 and Rutgers University Research Council Grant 2-02364 to the first author. Many thanks to John Kingston and all the participants of the Prosody Workshop for valuable discussions.

5. hao3 ji3 zhong3 → hao[2] ji[2] zhong3
→ hao[2] ji[1] zhong3

quite several kind
"quite a few kinds"

Examples (1) and (2) show the rule applying after a tone 1 syllable, while (3-5) show the rule applying after tone 2. (4) and (5) show that the rule is not restricted to lexical items. It applies to sentences, questions as well as statements. Finally, (5) shows that the rule can affect a derived tone 2. The intermediate step in (5) shows the effect of another tone sandhi rule, Mandarin 3rd Tone sandhi (3TS), which changes the first of two tone 3 syllables to tone 2. In (5), the 3TS feeds the 2TS, allowing further change of the middle syllable to tone 1.

Furthermore, Cheng [2] shows that the 2TS may apply to a sequence of tone 2's derived via the 3TS, creating a stretch of tone [1]'s from underlying tone 3's. Cheng also notes that such a chain effect is only found in short phrases, which are not likely to extend beyond 6 or 7 syllables.

6. Lao3 Li3 mai3 hao3 jiu3
Old Li buy good wine
"Old Li buys good wine."

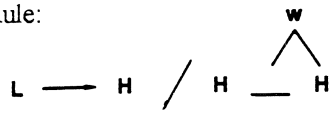
[2] [2] [2] [2] 3 --After 3TS
[2] [1] [1] [1] 3 --After 2TS

Our experimental results will be shown to support Chao's description in full, that the 2TS is frequently found in conversational speech, but much rarer in deliberate speech. The focus of this paper is the reasons behind the peculiar rule description. There are two issues involved: First, what is the motivation behind the rule description? Second, is the 2TS a phonological rule? Alternatively, could it be the result of phonetic implementation?

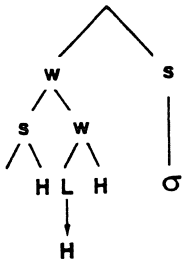
If the 2TS is a phonological assimilation rule, in which the L target of a rising tone assimilates to surrounding H targets, then two syllables in the rule description should be enough. Why should the presence of a third syllable be crucial, and yet the value of its tone have no effect? If the tone on the third syllable has no function, why can't it be a neutral tone?

Yip [3] and Zhang [4] both provide a stress account to explain the requirement of the third syllable. They both recognize that the real conditioning factor of the 2TS is stress, instead of syllable count. Their rules below specify that the L target of a Tone 2 will change to H only if the syllable is dominated by a weak (w) node. In other words, the 2TS only applies to prosodically weak syllables. Yip's rule does not explicitly refer to a third syllable as the conditioning factor; the rule suggests that the stress condition is all that is needed. Zhang's stress tree depicts the typical stress relation of a three syllable word in Mandarin: the middle syllable is weak, therefore allowing the 2TS to apply.

7. Yip's 2TS Rule:



8. Zhang's 2TS Rule:



Even though distinction of stress levels in a tone language such as Mandarin is subtle, it is generally agreed that a Mandarin word typically has final stress, and a three syllable structure has the stress pattern *secondary, tertiary, primary* [5, 6]. Hoa [5] has demonstrated the derivation of such stress relation in a left-branching word following the metrical structure and the rhythm rule of Liberman and Prince [7]. In a right-branching structure, similar stress relationship can be obtain by restructuring (Zhang [4], Hoa[5]), or by treating the three syllable structure in question as a single domain, ignoring internal branching (Shih [8]).

Interpreting stress as the conditioning factor of the 2TS explains successfully why the third syllable cannot have a neutral tone. Since neutral tone syllable are weak and cannot carry stress, its presence in the final position forces the middle syllable (being the last stress/tone carrying unit) to receive primary stress, therefore destroying the 2TS environment defined in Yip and Zhang.

A question left unanswered by Chao, Yip, and Zhang concerns the reason why a L target will change to H in a prosodically weak position. We believe that the answer lies in the way tonal targets are implemented in general. Turn to the second issue concerning Chao's description, whether the 2TS is a phonological rule, we will confirm with experimental data that 2TS is the consequence of phonetic implementation. We will show that the resulting tone 1 of the 2TS is a coincidence, so to speak, expected of an extremely weak L target in a high tone environment, where the strength of a L target is reflected in its distance away from the surrounding H tone environment. In H tone context, a weak L tone is implemented with higher

F0 value while a strong one has lower F0 value, as proposed in Pierrehumbert and Beckman [9] for Japanese.

Furthermore, we consider the 2TS to be part of a more general picture: a weak tone has less strength to break away from the interpolation line of surrounding strong targets of all types. The phenomenon is not limited to H L H sequences. Figure 1 shows the pitch track of the utterance *fan3-ying4 su4-du4*, "reaction speed", from our conversation data.

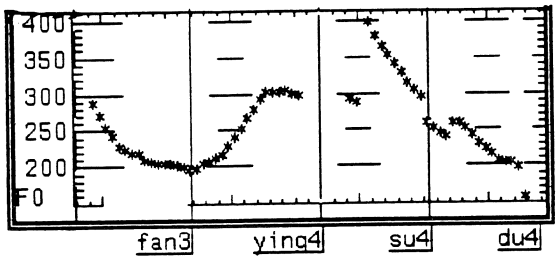


Figure 1: Pitch Track of *Fan3-Ying4 Su4-Du4*

The second syllable is in a weak position, and the falling tone (tone 4) surfaces with a rising shape. The surprising outcome is made possible when the syllable in question is so weak as to lose its own targets. The resulting pitch is just the transition from the previous syllable to the following syllable. In the pitch track shown, the preceding tone is low and the following tone is high, accounting for the transitional rising shape, which turns out to be the opposite of what is expected of the lexical tone. Since Mandarin speakers are typically sensitive to *wrong* tones, we play this speech segment to eight native speakers, asking them to identify the tone sequence, and ask them whether there is any thing wrong. All eight speakers identify the tone sequence as 3 4 4 4, the correct underlying tones. They didn't notice any aberrant behavior of the second syllable when the whole word was played to them, and were surprised later to hear a rising tone when the second syllable alone was played.

The following experiment is designed to investigate the two issues on the 2TS discussed above. First, we ask if the 2TS should be considered as a phonetic implementation rule rather than a phonological rule. Having answered this question in the affirmative, we will show that the phonetic implementation rule in question is sensitive to stress. We follow a straightforward working hypothesis of Liberman and Pierrehumbert [10] about the nature of phonological rule versus phonetic implementation: a phonological rule typically creates distinct classes on the surface, while phonetic implementation creates a continuum.

2. Experiment Design

We designed two sets of stimuli. The first, the control group, consisted of three and four syllable words with lexical high tones (tone 1) on every syllable. The second, the test group, were three or four syllable words with lexical high tones surrounding a single lexical rising tone (tone 2). These stimuli are described below:

Stimuli

I. Control group:

- 14 left branching three-syllable words with 1 1 1 tone pattern
- 9 right branching three-syllable words with 1 1 1 tone pattern
- 5 four-syllable words with 1 1 1 1 tone pattern

II. Test group:

- 22 left branching three-syllable words with 1 2 1 tone pattern
- 9 right branching three-syllable words with 1 2 1 tone pattern
- 9 four-syllable words with 1 2 1 1 tone pattern
- 6 four-syllable words with 1 1 2 1 tone pattern

In practice we found no effect of branching in the results for the three syllable words, so we will henceforth collapse the two conditions.

A female speaker of Beijing Mandarin was asked to produce utterances containing these words in the following contexts. Words with three and four high tones (1 1 1 and 1 1 1 1) are not used in reading contexts (IIb) and (IIc). Test words and sentences in each context are presented on cards to the speaker in random order. Contexts (IIb) and (IIc) are mixed. Each word/sentence is read once.

Contexts

I. Isolation

II. Read sentences with three different frames:

a. Ta1 shuo1 ____ chu1 mao2 bing4 le0.
"He says ___ went wrong."

b. Ta1 ke1 ABC de0 A zi4.
"He carves the character A of ABC"

c. Ta1 ke1 ABC de0 B zi4.
"He carves the character B of ABC"

III. Conversation

The isolated syllables *A* and *B* in contexts (IIb-c) are referred to as the *frame words*. The frame words are prosodically strong, carrying both the word stress and the sentential focus. In the case of 4 character words, the frame word B is the tone 2 syllable; it could be the second or third syllable of the word. The speaker pronounced the syllable *kel* "carves" in (IIb-c) with a falling tone, which has noticeable lowering effect on the test words in (IIb-c) contexts, in comparison to the test words in the (IIa) context.

The conversation data are collected by having the speaker discuss the experiment with the experimenter after she had finished with the isolation and read sentence conditions for all words. Some of the typical questions addressed to the speaker include: "Do you like ___?" and "How often do you read ___?" All occurrences of words with the appropriate tonal combination produced by the speaker in the conversation are collected for analysis. Naturally, we were not able to elicit utterances for every stimulus word under these conditions. We have also collected a few words that are not used in the isolation and read sentence conditions.

One F0/time pair is measured for each syllable of the word for analysis. The pitch contours of most tone 1's and tone 2's in our data have an overall concave shape. Typically, we measure the F0 and time value of the F0 minimum for both types of tones. The lowest point of a tone 2 lies close to the center of the vowel region, and is considered to represent the L target of the rising tone. Occasionally, a tone 1 will have a convex shape; under those conditions, we measure the time and F0 value for the F0 maximum.

3. Results

Figure (2) shows three occurrences of the same word *zangl mao2-yil*, "dirty sweater", all collected from the conversation session. These pitch tracks show quite a variation in the implementation of the L target of the rising tone. Figure (2a) shows a pitch drop of around 50 Hz from the H target of *zangl* to the L target of *mao2*. The pitch drop in (2b) is considerably less, at around 25 Hz. Finally, there is hardly any pitch drop in (2c). The L target of a rising tone is realized at around the same pitch height of the surrounding H targets.

If the 2TS were a phonological rule, we would expect most of the samples to be similar to either Figure (2a), where the rule has not applied (note that the rule is optional), or to Figure (2c), where the rule has applied. Cases like Figure (2b) are not expected, so the frequency should be low, if they occur at all.

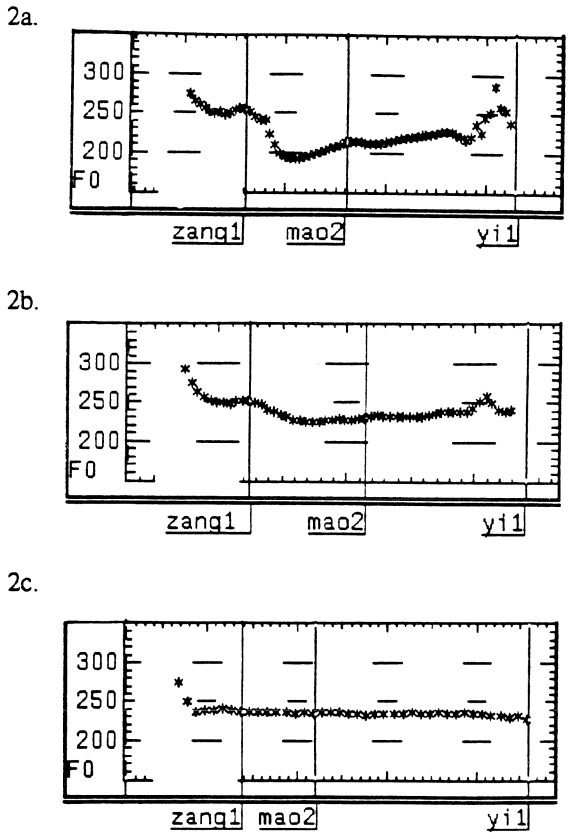


Figure 2: Variations in Pitch Drop

In the following, we will show that the amount of pitch drop is indeed gradient, and cases like (2b) are the norm, not the exception, supporting the view that the 2TS is the result of phonetic implementation. It is reasonable to ask whether the variations in pitch drop can be derived from duration: if the duration is short, there may not be sufficient time to reach down to a L target. We cannot find clear evidence supporting this hypothesis. Finally, we test whether stress is a factor, and find strong support in the sense that the amount of pitch drop correlates well with the phonologically defined stress levels.

3.1 Gradient Implementation

Figure 3 shows averaged tonal trajectories for all 1 1 1 and 1 2 1 words divided by condition. The 1 1 1 cases are represented by black plotting symbols, while the 1 2 1 cases are represented by open symbols; for example, the black triangle trajectory represents the average of all 1 1 1 words in the "Read Sentence" condition, and shows that the first syllable's high tone has a mean pitch value of 304Hz, the second syllable's high tone has a mean pitch value of 288Hz and the third syllable's high tone has a mean pitch value of 280Hz. The 1 1 1 words, with a string of H tones, show very similar behavior under all

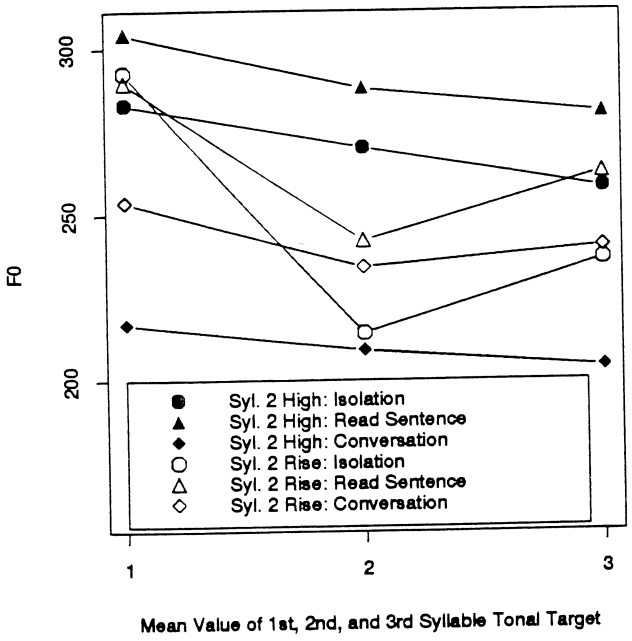


Figure 3: Tonal Trajectories of 3 Syllable Words

conditions: roughly speaking there is a drop of 15 to 25 Hz between the first and the third syllables, but otherwise the decline is almost linear. We assume that the significantly lower pitch value in the "Conversational" condition is due to many of these utterances having occurred fairly late in their contextual sentences, so that what we are seeing here is due to the accumulation of lowering effects such as declination and downstep.

The 1 2 1 words show a very different pattern. In the isolation context, represented by open circles, there is a drop of more than 75 Hz between the first syllable's tone 1 (H) and the second syllable's L tone. The third syllable's tone 1 only rises about 20 Hz, possibly due to the combined effects of downstep and final lowering. (See [10] for a discussion of these effects and their interaction.) In the read sentence context the drop between the first two syllables is less but still significant (47 Hz); since these words are not utterance-final, the final lowering effect is removed and the third syllable's tone 1 has a much higher value than in the isolation context. Most striking, however, is the conversational context where the drop between syllable one and two is minuscule (less than 20 Hz), and the overall pattern for the three syllables strongly resembles a sequence of three high tones. So, in the 1 2 1 contexts, we have what appears to be a gradient effect from the "Isolation" context, through the "Read Sentence" context, through the "Conversation" context.

That this behavior is indeed gradient is shown by the plot in Figure 4. Here we plot the F0 value for syllable 1 on the x-axis against the F0 value for syllable 2 on the y-axis, for both 1 1 1 and 1 2 1 words under all 3 conditions.

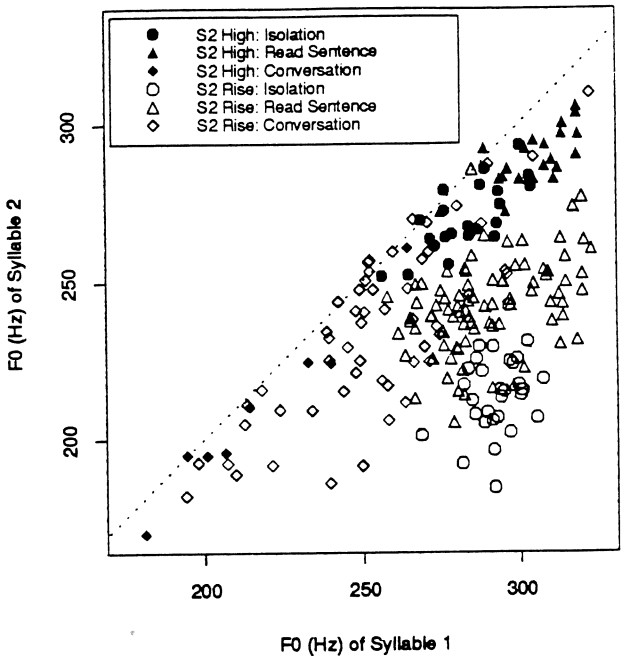


Figure 4: F0 of Syllable 1 vs. Syllable 2

The plotting symbols are the same as the previous plot: black symbols for the 1 1 1 cases; open symbols for the 1 2 1 cases; circles for the "Isolation" context; triangles for the "Read Sentence" context; and diamonds for the "Conversation" context.

The dotted line is the line x=y, and the fact that most points lie below that line, even for the 1 1 1 cases, is consistent with the overall declination effect (i.e., syllable 2 generally has a slightly lower pitch value than syllable 1, even if both are high tones). As can be seen in this plot all of the high tone cases cluster fairly close to the x=y line, as expected. The 1 2 1 cases, however, range from having syllable 2's F0 much lower than syllable 1's F0 -- e.g., the "Isolation" cases in open circles -- all the way to having the two syllables have roughly equal tone values -- e.g., many of the "Conversation" cases in open diamonds -- and thus being effectively indistinguishable from a sequence of three lexical high tones.

The gradient implementation of L target can be further supported by performing t-tests comparing the pitch difference between syllable 1 and syllable 2 for, on the one hand, all of the 1 1 1 cases, and on the other each of "Isolation", "Read Sentence" and "Conversation" for the 1 2 1 cases. These results are given in Table 2.

t-test: all 111 vs. 121		
	t(df)	p
121 Isolation	28.2 (82)	< .001
121 Read Sentence	13.4 (136)	< .001
121 Conversation	2.2 (105)	< .05

Table 2: T-Test Scores

For both the "Isolation" and "Read Sentence" cases we can reject the hypothesis that they are the same as the 1 1 1 cases, at the .001 level; note however that the actual t score for "Read Sentence" is less than that for the "Isolation", consistent with the observation that the "Read Sentence" cases have a smaller drop between syllable 1 and syllable 2 than the "Isolation" cases. For the "Conversation" cases, there is still a significant difference between them and the 1 1 1 set, but only at the .05 level. This is interesting for two reasons. First of all, it suggests that under conversational conditions the difference between the 1 2 1 and 1 1 1 cases is indeed reduced, consistent with Chao's rule. On the other hand, the fact that even under conversational conditions the 1 2 1 set cannot be treated as identical to the 1 1 1 set suggests that we must be dealing here with a gradient phonetic implementation rule, rather than a categorical phonological rule whose effect is to rewrite a tone 2 as a tone 1.

3.2 Duration

Given the rather strong evidence that the 2TS must therefore be a phonetic implementation rule, the next question is what the conditioning factors might be. One obvious possibility is speech rate or durational effects. When the duration is short, the speaker may not have enough time to reach the intended target, leading to the reduced pitch difference between a L target and the preceding H tone, and the apparent assimilation effect on the surface. To test this hypothesis, we checked the amount of pitch drop in 1 2 1 words against a number of duration measurements, such as the duration of each segment, the duration of each syllable, the duration of two syllables, and the duration of the whole word. We show one of these plots in Figure 5, the whole word duration against the amount of pitch drop. The correlation is among the best, and the pattern of the distribution is very representative of the cases where other duration measurements are used.

In Figure 5, the duration (in seconds) of 3-syllable 1 2 1 words is plotted on the x-axis, and the amount of pitch drop (in Hz) from the first syllable (H target) to the second syllable (L target) is plotted on the y-axis. Open circle, black circle, and open diamond represent the "Isolation" context, the "Read Sentence" context, and the "Conversation" context respectively. The dotted line represents the least-squares fit regression line for all the

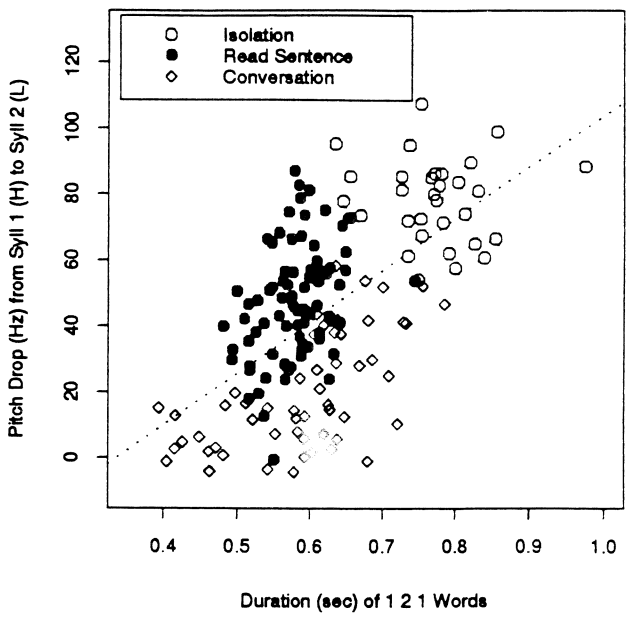


Figure 5: Duration vs. Pitch Drop in 1 2 1 Words

data.

As can be seen from the plot, there is some correlation between the duration variable and the pitch variable. For example, the "Isolation" context tend to be associated with longer duration (as might be expected) and they are also the cases with the largest pitch drop; contrariwise, some of the conversational cases have shorter durations and they also show the least pitch drop. However, the correlation is not good: the R-squared value for this correlation is only 0.36 (sample size is 170).

A striking feature of Figure 5 is the separation of the three contexts: On the one hand, the "Isolation" context do not overlap much in their time-axis distribution with the "Read Sentence" and "Conversation" contexts, but it overlaps considerably along the pitch-axis with the "Read Sentence" context. On the other hand, the "Read Sentence" context and the "Conversation" context show a very significant overlap in the time-axis, but are clearly separated in the pitch-axis. The complication of styles or test contexts makes it more difficult to generalize the relation between duration and pitch drop.

More importantly, there is no correlation between duration and pitch drop within the "Isolation" context, and the correlation is poor for the "Read Sentences" context (R-squared value is 0.13, sample size is 85), suggesting that the amount of pitch drop cannot be reliably predicted from duration for members within each of these two groups. The "Conversation" context has better correlation (R-squared value is 0.36, sample size is 54), but it is also

clear that the distribution is triangle shaped: samples with short duration, especially those shorter than 0.5 second, have less than 20 Hz of pitch drop, but many samples with much longer duration still fall within the same range It seems that only the extremely short duration is responsible for the undershoot of the target, represented by the 10 samples at the lower left corner of the plot. For the rest of the data, although duration or speech rate is surely a factor, there appears at the present time to be an additional irreducible effect of speech style in the phonetic implementation of the 2TS.

3.3 Stress

Another possible factor is stress. Indeed, we have hypothesized with Yip and Zhang that stress is the real reason behind Chao's "three syllable condition": only when a tone 2 occurs after a H target on a preceding syllable, and before some other non-neutral-toned syllable can it both be in the right tonal environment AND be in a metrically weak position. In order to evaluate the effect of stress, we turn now to the four-syllable cases. In Mandarin, in a four syllable sequence ABCD, the B syllable is generally the weakest prosodically (see summary in Zhang [4]). Thus, if stress is a factor in the implementation of 2TS, we would expect to find difference in behavior between the tone 2's in 1 2 1 1, where the tone 2 is on the weakest syllable, and 1 1 2 1, where it is on a stronger syllable. Assuming that stronger stress correlates with a more pronounced rendition of a tonal target, we would expect the tone 2 in 1 2 1 1 to show more 2TS effects than 1 1 2 1. This expectation is confirmed. In Figure 6 we plot data from "Conversation" and "Read Sentences" contexts the F0 value for the syllable preceding the tone 2 against the F0 value for the tone 2 itself. The 1 2 1 1 cases are represented by open triangle, and the 1 1 2 1 cases by black circle. The dash line is the x=y line. As can be seen, the value for tone 2 is significantly further away from the x=y line for the 1 1 2 1 cases than for the 1 2 1 1 cases, consistent with the stronger stress for the tone 2 in 1 1 2 1.

Both cases presented in Figure 6 are relatively weak, comparing to some strong positions that are expected to receive primary word stress, such as a monosyllabic word, the tone-carrying final syllable of a word, or the syllable before a neutral tone. If stress is a factor in the implementation of the 2TS, we would expect even more pitch drop in the strong cases then in 1 1 2 1. This prediction is again bourn out. The difference between the 1 1 2 1 case and other cases with even stronger stress is a good evidence against the phonological analysis of Zhang [4], where it is claimed that 2TS does not apply in the 1 1 2 1 cases because there is some level of stress dominating the tone 2 syllable.

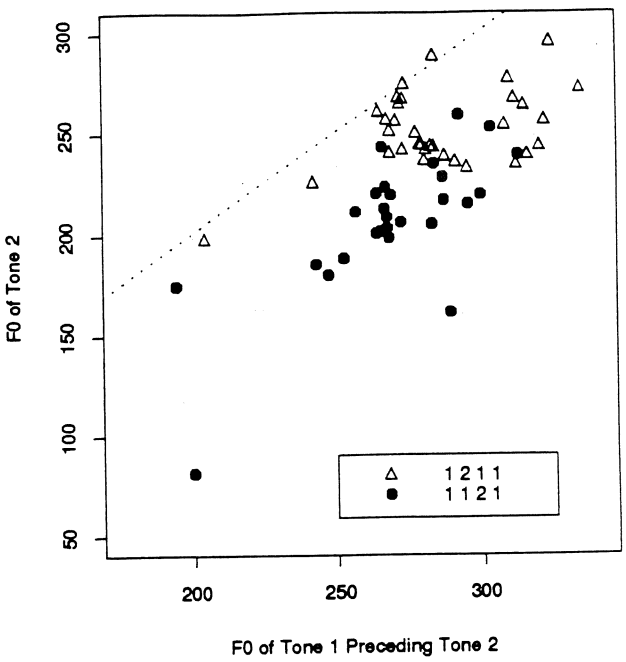


Figure 6: Pitch Drop in 1211 and 1121

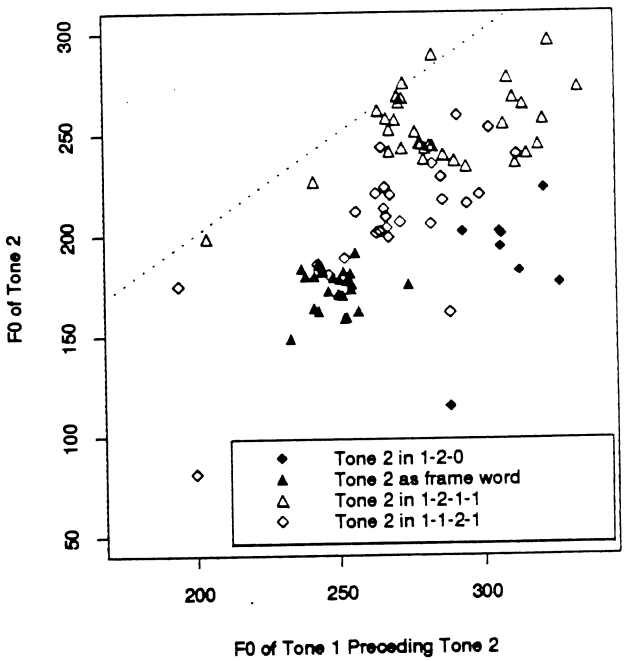


Figure 7: Pitch Drop in Weak Positions (1211, 1121) and Strong Positions (120, 2)

Figure 7 adds two other cases where tone 2 is expected to be strong (in black), contrasting to the tokens in Figure 6 (in white). One of the strong cases we used is the tone 2 in tonal combination 1 2 0, where the third syllable has a neutral tone (tone 0), and cannot receive stress. The middle tone 2 will be assigned primary stress instead because

it is now the last tone/stress carrying member of the word. In this case we plot the value of the preceding high tone (x-axis) against the value of the tone 2 (y-axis) in black diamonds. Another presumably strong case is tone 2 of a monosyllabic word. We plot the value of tone 1's (on x-axis) from the frame word in Read sentence condition (IIb) against the tone 2 (on y-axis) from the frame word in Read sentence condition (IIc). (Given a stimulus word ABC, with tonal combination 1 2 1, we plot A (tone 1) from the sentence "**A** of ABC" against B (tone 2) from the sentence "**B** of ABC".) Black triangle is used as the plotting symbol. It is clear that all of the expected strong tone 2's have comparatively lower F0 value for their L targets.

Figure 7 gives strong support to the notion that the value of L target in H tone context is sensitive to stress: L target in stronger positions is implemented with lower F0 value, with bigger excursion away from the surrounding environment; L target in weaker positions is implemented with higher F0 value, which is more similar to the H tone environment.

4. Conclusion

To summarize, we have shown that Chao's 2TS rule is best viewed as a phonetic implementation rule that raises the pitch value of the L excursion in a tone 2 when that tone 2 occurs after a H tone. The tone 2 must be followed by a non-neutral toned syllable, which provides the minimum environment in which the tone 2 can be metrically weak. We have shown that stress is indeed an important factor by showing a robust difference among the implementation of the tone 2 in 1 2 1 1, 1 1 2 1, 2, and 1 2 0 environments. Moreover, in both 1 2 1 1 and 1 1 2 1 cases the tone 2 is in a metrically relatively weak syllable, but in 1 1 2 1 it is nonetheless in a somewhat stronger position than in 1 2 1 1. Tonal implementation is sensitive to the difference. Note that this difference does not follow from Chao's statement of the rule, and that the stress condition thus has the nice property of both explaining an odd feature of Chao's rule (the three syllable condition) and actually offering a better account of the data than his rule. We have also tested whether duration or speech rate may be a factor in the implementation of the L target, and show that there are some problem in interpreting the correlation of duration and the amount of pitch drop. Furthermore, consistent with Chao's original description, there seems also to be an irreducible effect of style.

5. References

[1] Chao, Y. R. *A grammar of spoken Chinese*. Berkeley: University of California Press, 1968.

[2] Cheng, C. C. *A synchronic phonology of Mandarin Chinese*. The Hague: Mouton, 1973.

[3] Yip, M. *The tonal phonology of Chinese*. Ph.D. dissertation, MIT, 1980.

[4] Zhang, Z. S. *Tone and tone sandhi in Chinese*. Ph.D. dissertation, Ohio State University, 1988.

[5] Hoa, M. *L'accentuation en Pekinois*. Paris: Centre de Recherches Linguistiques sur l'Asie Orientale, Editions Langages Croises, 1983.

[6] Lin, M., Yan, J. and G. Sun. The stress pattern and its acoustic correlates in Beijing Mandarin. pp. 504-514, in *Proceedings of the 10th International Congress of Phonetic Sciences*, 1983.

[7] Liberman M. and A. Prince. On stress and linguistic rhythm, pp. 249-336, in *Linguistic Inquiry* 8.2, 1977.

[8] Shih, C. L. Mandarin third tone sandhi and prosodic structure. In. J. Wang and N. Smith eds, *Studies in Chinese phonology*, Foris Publication, in press.

[9] Pierrehumbert J. and M. Beckman. *Japanese tone structure*, Cambridge: MIT Press, 1988.

[10] Liberman M. and J. Pierrehumbert. Intonational invariance under changes in pitch range and length, pp. 157-233, in M. Aronoff and R. Oehrle eds. *Language Sound Structure*, Cambridge: MIT Press, 1984.

The Relationship of Filled-Pause F0 to Prosodic Context

Elizabeth E. Shriberg

SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025 USA
and Department of Psychology, University of California at Berkeley

Robin J. Lickley

Centre for Speech Technology Research and Department of Linguistics
University of Edinburgh, 80, South Bridge, Edinburgh, EH11HN UK

1. ABSTRACT

Filled pauses in spontaneous speech present problems for models of speech understanding and automatic speech recognition. A potentially important cue to their recognition by both humans and machines is their typically low F0 [9, 7]. The current paper discusses results of a study [10] which sought to determine whether the F0 of filled pauses is relative to, or independent of, the F0 of surrounding lexical material. Clause-internal filled pauses and preceding peak F0 values for speakers of American and British English were examined. Higher peaks were found to be systematically associated with higher filled-pause values within speakers, supporting the "relative" hypothesis. In modeling this relationship it was found that a linear model, in which filled-pause F0 was expressed as an invariant (over speakers) proportion of the distance between the preceding peak F0 and a speaker-dependent terminal low F0, produced results nearly identical to those of a two-parameter model in which the coefficients of peak and terminal low F0 were allowed to vary freely. Analyses of additional variables showed the model to be less appropriate for filled pauses after sentence-initial peaks, but unaffected by temporal variables. These results suggest that clause-internal filled pauses, while lower in F0 than words in the message stream, nevertheless preserve information about the local prosodic context. Implications for psycholinguistics, speech recognition, and linguistic theory are discussed.

2. INTRODUCTION

Phenomena exhibited in spontaneous speech present new challenges for researchers in psychology, speech technology, and linguistics as the object of study shifts from carefully prepared "laboratory speech" to natural conversation. An important difference between spontaneous speech and speech that is read or rehearsed is that spontaneous speech is characterized by relatively high rates of hesitation pauses, repetitions and reformulations [3]. This paper examines one of the most common types of hesitation phenomena: the filled pause, usually realized orthographically as "um" or "uh."

Filled pauses can present problems for models of human language understanding and automatic speech recognition. In the case of human perception, what is remarkable is the extent to which filled pauses are "filtered out" in comprehension. Those familiar with the task of transcribing spontaneous speech will note that filled pauses are often missed in first passes at transcription; laboratory experiments [e.g., 5] have shown that listeners have difficulty locating filled pauses when monitoring for sentence content. In the case of speech recognition, filled pauses are problematic in that they are often misrecognized as words having similar phonetic features, such as "a", "an' or "and," or as syllables of longer words [1, 7, 9].

One source of information that is likely to be important in the successful perception and processing of spontaneous speech in general [see, for example, 6] and speech containing filled pauses in particular, is prosody. Recent work has contributed to our knowledge of the prosodic features of filled pauses. Studies of hesitations in a database of human-computer dialog [4, 11] show that filled pauses tend to occur in the lower region of a speaker's F0 range and have a level or falling tone [7], and, more specifically, that their F0 is typically lower than that of both accented and unaccented neighboring syllables [9].

For human perception, these findings may provide an account for the apparent perceptual separation of filled pauses from the message stream. The low F0 of filled pauses could aid automatic recognizers in distinguishing filled pauses from real words. In addition, linguists may be concerned with how to best represent these predictably low-F0 units in prosodic descriptions of spontaneous speech.

A question relevant to each of these areas concerns the nature of the relationship between the low F0 of filled pauses and the intonation of surrounding material. There are three possible relationships: 1) filled pauses may be produced at an absolute, speaker-specific F0 value regardless of their position within the sentence; 2) the F0 of filled pauses may vary within speaker, but the variation may be unpredictable; or 3) the F0 of filled pauses for a particular speaker may be predictable at better than chance, given knowledge about the prosodic context.

A study previously reported in [10] investigated the relationship between filled-pause F0 and intonational context; the current paper discusses results of that study in further detail. Since the question of interest concerned prosodic context, the relevant filled pauses to examine would be those that interrupt a prosodic phrase, as opposed to those that initiate a speaker's turn or occur between intonation phrases. The task of choosing filled pauses that occur within a prosodic phrase poses difficulties, however, in that: (1) it would be unclear how to label the data prosodically, since existing prosodic theories are not tailored to the description of material surrounding hesitation phenomena; (2) it is not clear what level of prosodic structure would be appropriate to use as the relevant unit for "interruption;" (3) choosing filled pauses on the basis of the prosody of surrounding material is potentially circular in that hesitations may themselves influence the prosody of that material; and (4) prosodic labeling requires listening to utterances and is time-consuming.

The scheme adopted was to study filled pauses that occurred within a syntactic clause. Filled pauses were considered to be "within-clause" if lexical material preceding the filled pause was syntactically incomplete, and strongly predicted continuation of the utterance after the filled pause. The value of the closest F0 peak preceding the filled pause was used as a measure of prosodic context, and the initial F0 value of the filled pause was used as a measure of filled-pause F0.

Within-clause filled pauses from speakers of American and speakers of British English, in two different discourse contexts, were examined to evaluate the three alternative hypotheses. The "absolute" hypothesis predicted that filled pauses would occur at a constant, speaker-dependent F0 value regardless of the value of the preceding peak F0. The "random" hypothesis predicted that filled-pause F0 values from a particular speaker would vary in a manner uncorrelated with preceding peak F0 values. The "relative" hypothesis predicted some form of systematic relationship between the peak and corresponding filled-pause F0 values.

3. METHOD

3.1. Subjects

Two quite different sets of data were analyzed. The first was a set of 120 clause-internal filled pauses from digitized utterances from 29 speakers (14 male, 15 female) of American English making air travel plans by speaking to a computer. The multi-site database is described in detail in [4]. The majority of examples came from "Wizard-of-Oz" systems, in which a human interpreted and responded to requests and thus "recognition" was perfect; a small number came from interaction with a Spoken Language System [11]. The number of clause-internal filled pauses per speaker used in the analyses ranged from 2 to 13; 82 of the examples came from 12 speakers (6 male, 6 female) having 5 or more examples each.

The second set consisted of 87 filled pauses taken from a corpus of six dialogues recorded digitally at the Department of Linguistics at the University of Edinburgh. Dialogues involved the second author and a colleague or acquaintance; they were natural, spontaneous conversations on various topics, with no set task. The subjects were 3 male and 3 female speakers of British English, without strong regional accents, who were unaware of the purpose of recording the conversations. The number of clause-internal filled pauses per speaker used in the analyses ranged from 6 to 28.

3.2. Filled Pauses

The goal of the study was to examine filled pauses that were likely to interrupt a prosodic phrase; however, because it would have been difficult and time-consuming to label the data sets prosodically in order to select the desired filled pauses, a method based largely on syntax was used. In general, the filled pauses selected for analysis were those that directly followed lexical material that would have been syntactically incomplete if the utterance had not continued after the filled pause. It was felt that this would be an efficient, straightforward, and easy-to-replicate method for capturing many of the filled pauses that did interrupt prosodic phrases, while avoiding the complex and time-consuming task of prosodic labeling. Some examples from the American data set are listed in Table 1.

Table 1: Examples of Clause-Internal Filled Pauses

Incomplete	"Looking for"	Example
NP	N	...the lowest [uh] fare...
VP (trans)	NP	...book [uh] the flight...
PP	NP	...leave at [um] noon...
AUX	S	Does [uh] Delta fly...

The researchers tried to determine whether or not a listener would feel it was possible that the speaker could have ended an utterance before the filled pause, based on a transcription alone, but taking semantic and pragmatic information into account. For example, filled pauses in utterances such as:

Show me flights flying [uh] from Boston.

in which material before the filled pause is not necessarily syntactically incomplete, but which would seem incomplete to a listener given the discourse context, were included in the analyses.

Conversely, some utterances which could be viewed as meeting the syntactic expectancy requirement were not included in the analyses. These were cases in which the only item preceding the filled pause in the same clause was a conjunction such as "and" or "but.," a lexical filler such as "well" or "okay," or another filled pause. Such cases were excluded because of the higher likelihood of a prosodic boundary immediately preceding the filled pause.

3.3. Apparatus

The digitized waveforms were sampled at 8 or 16 kHz and all waveforms and pitch tracks were examined using the Entropic ESPS/Waves+ software on a Sun 4 workstation.

3.4. Procedure

The American and British data were coded independently by the first and second authors, respectively. For each within-clause filled pause having reliable pitch tracks, the researcher recorded five F0 values, four measures of duration, and values for four additional variables.

The F0 of each filled pause was measured at both the beginning and end of the filled pause. These values describe the F0 of filled pauses well, since most fall fairly linearly. Analyses in the present work used the initial filled-pause F0 as a measure of filled-pause F0. F0 was also recorded at the F0 peaks most closely preceding and following the filled pause; results reported here used only the preceding peak as a measure of prosodic context. Alternative measures of context (for example topline, or preceding low accents) could also be used, but could be more difficult to measure and locate than F0 peaks. Peak values were restricted to occur on words within the clause containing the filled pause. In most cases, the peak was marked on a syllable perceived to be accented; in a few cases no accented syllable was available and the highest preceding F0 value was used.

A fifth F0 value, which will be referred to as the "terminal low F0," was measured after final lowering in a manner similar to that described in [2]; i.e. for utterances containing a terminal fall, F0 was measured at the lowest point in the fall, disregarding regions associated with errors in pitch tracking or vocal fry. The purpose of this measure was to provide a single, stable, speaker-dependent F0 value for each speaker. The underlying assumption in the present work was that this value should correspond to a speaker's lowest possible F0, as opposed to the lowest F0 realized in any particular utterance, since the former would be the more stable value given the inherently positively skewed distribution of terminal low F0 values. Therefore, terminal low F0 values were obtained for all utterances for a particular speaker that contained a terminal fall. The lowest of these values was then used as the estimate of the speaker's terminal low F0 for all speech tokens from that speaker in the analyses. Care was taken to assure that the lowest terminal F0 value did not appear to be an outlier when compared with the other terminal F0 values obtained for the same speaker.

Four measures of duration were recorded, including the duration of the filled pause, that of preceding and following silent hesitation pauses (if any), and that of the time (and also the number of syllables) between the preceding peak and the beginning of the filled pause.

Values for additional variables of interest were also recorded, including the sex of the speaker, whether or not the filled pause preceded a repetition, repair, or fresh start, whether or not the preceding peak was marked on a sentence-initial accent, and whether the filled pause was "um" or "uh."

4. RESULTS

Figures 1-4 show data for a male or female speaker from each of the data sets (American and British). Time-normalized F0 values are shown for the preceding peak F0, initial filled-pause F0, final filled-pause F0, and following peak F0 in multiple examples of filled pauses for the particular speaker. Each speaker's estimated terminal low F0 is also indicated.

4.1. Testing the Hypotheses: Sign Test

The first thing to note about the plots is that, in general, the drop to the filled pause from the preceding peak scales with the peak values, so that higher peaks tend to have higher following filled pauses. This simple assumption was tested using data from all 35 speakers. The highest and lowest preceding peak F0 values over all examples from a particular speaker were extracted and the associated filled pause values compared in a Sign test. In 34/35 cases, the higher preceding peak value was associated with a higher filled pause value, $p < .0001$. This highly significant result is consistent with the relative hypothesis and inconsistent with the absolute and random hypotheses.

4.2. Modeling the Relationship

A second observation about Figs. 1-4 is that there appears to be a lower bound of F0: filled pauses do not seem to go below the terminal F0. This suggests that filled-pause F0 cannot be expressed as a simple subtractive function of

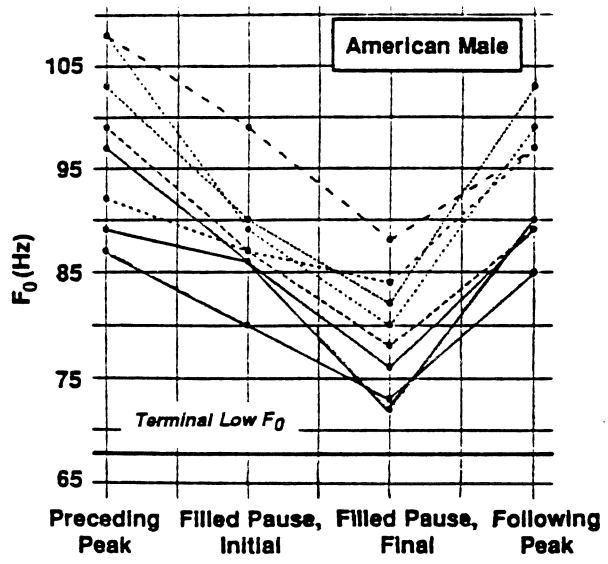


Figure 1: Peak and Filled-Pause F0 for American Male

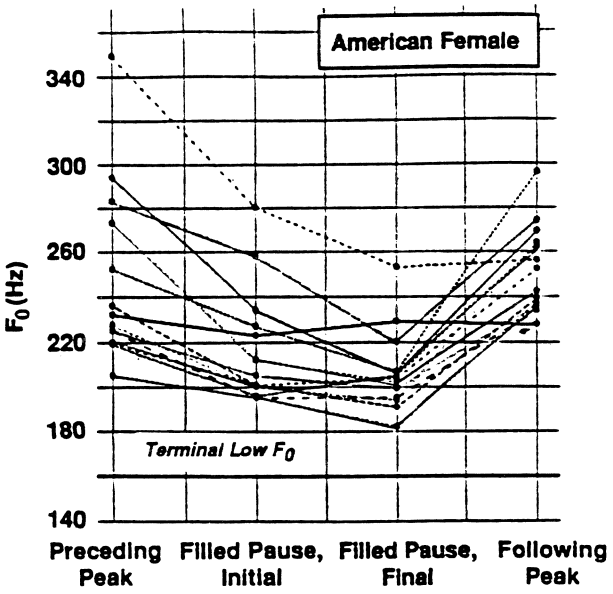


Figure 3: Peak and Filled-Pause F0 for American Female

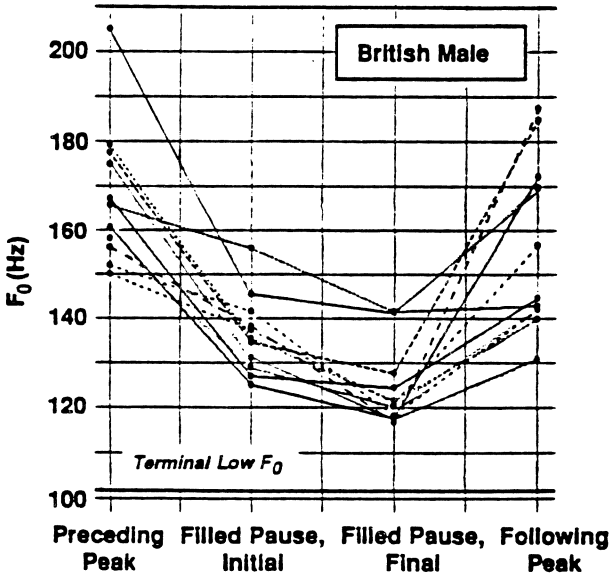


Figure 2: Peak and Filled-Pause F0 for British Male

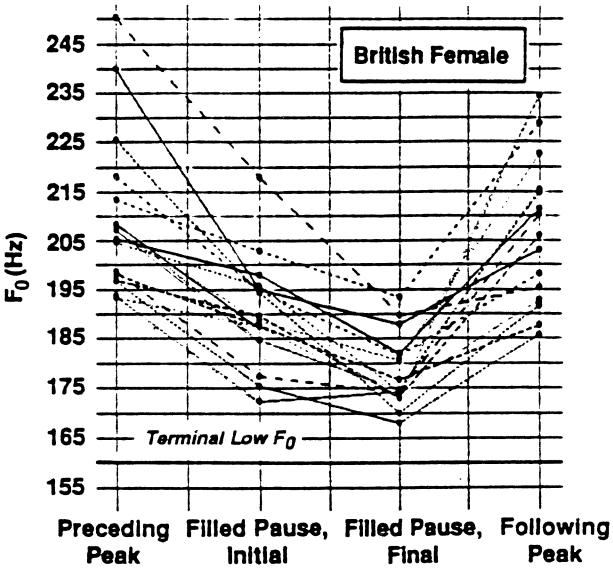


Figure 4: Peak and Filled-Pause F0 for British Female

peak F0. A third observation is that there seems to be a compressive effect for peaks closer to the terminal F0, with lower peaks producing less of a drop to the filled pause than higher ones. This observation suggests that filled-pause F0 cannot be expressed as a simple multiplicative function of peak F0, since such a function would predict parallel curves. Exceptions to this trend are the filled pauses following the very highest peak examples in Figs. 1, 2, and 4, which do not drop as far as expected. However, these examples form a special class; they correspond to filled pauses following peaks marked on sentence-initial accented syllables which, as discussed later, appear to behave differently from other clause-internal filled pauses.

Based on these observations, we proposed a simple linear model, in which filled-pause F0 (F0 fp) is the F0 value occurring at a fixed proportion of the distance between the peak F0 (F0 peak) and the terminal low F0 (F0 min):

$$F_{0\,fp} = r\,(F_{0\,peak} - F_{0\,min}) + F_{0\,min}$$

This is a single-parameter model, since the coefficients of peak F0 and terminal low F0 are both determined by r.

We determined the value of r empirically for each filled pause token from the set of American and British speakers with five or more examples each (18 subjects, 169 filled

pauses.) Means for tokens broken down by American/British and male/female are shown in Table 2.

Table 2: Values of r

Subject	# of speakers	# of tokens	Mean r	s.d. of r
American				
male	6	39	.596	.214
female	6	43	.626	.158
British				
male	3	55	.607	.240
female	3	32	.636	.242

Because results for the American and British data were remarkably similar, data were pooled for all further analyses. Although the value of r appears to be slightly higher for women in both groups, the differences are nonsignificant (as can be seen by comparing them to the magnitude of the standard deviations.)

A linear regression with the constant term suppressed, performed using the raw data from subjects represented in Table 2, and using the mean r determined over the entire set (0.62), yielded a standard error in prediction of 15.41 Hz. A comparison of this model to two other linear models is shown in Table 3. Investigation of higher-order models was not warranted given the lack of evidence for a nonlinear relationship, and the potential danger of over-fitting the small data set at hand. The proposed model was clearly better than one in which only the peak was used to predict the filled pause F0. It was also remarkably close in prediction accuracy to results produced by a two-parameter model which allowed the coefficients of peak and terminal low F0 to vary freely.

Table 3: Comparison of Models

Variables	# of Parameters	RMS error (Hz)
peak, terminal low F0	1	15.41
peak	1	19.58
peak, terminal low F0	2	15.25

4.3. Optimal Reference F0

An issue addressed was whether, given the proposed model, the estimated terminal low F0 values used corresponded to the optimal reference F0 values for prediction. Ideally, regressions solving for the optimal r and constant for each speaker would allow for comparison of these results to those obtained using the observed terminal low values; however, to be meaningful such analyses require more data per speaker. Nevertheless, analyses performed for a subset (N=6) of the 18 subjects who had the largest numbers of examples revealed that in each case the optimal reference F0 was higher than the observed terminal low F0. Therefore a number of modifications of the observed values in the 18-speaker data set were computed. For each modification, r was redetermined using the new terminal low values, and filled pauses were predicted using the new, overall average r and new low F0 values. It was found that the minimum standard error (15.16 Hz, as opposed to 15.41 Hz for the original terminal low values) was produced when observed terminal low values were increased by roughly 10%.

4.4. Effect of Duration

There was no correlation between the time or the number of syllables from the peak to the filled pause and the drop size. As shown in Figure 5, the drop in F0 from the preceding peak to the filled pause did not seem to depend on the amount of time elapsed between these two points.

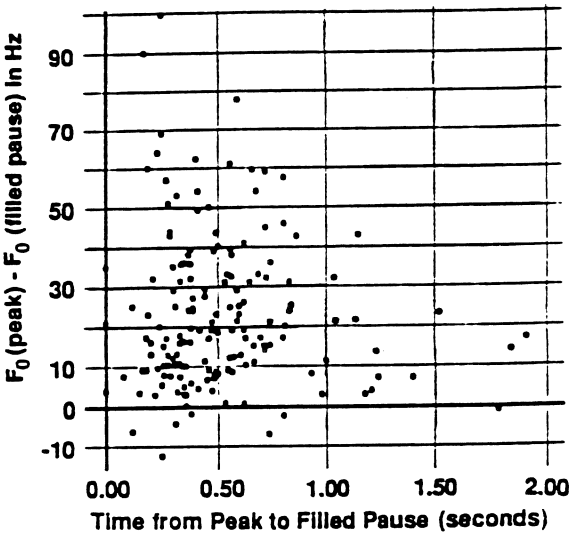


Figure 5: Effect of Time from Peak on F0 Drop

In addition, there did not seem to be any relationship between the duration of the filled pause itself and the size of the fall in F0 over the course of the filled pause, as shown in Figure 6.

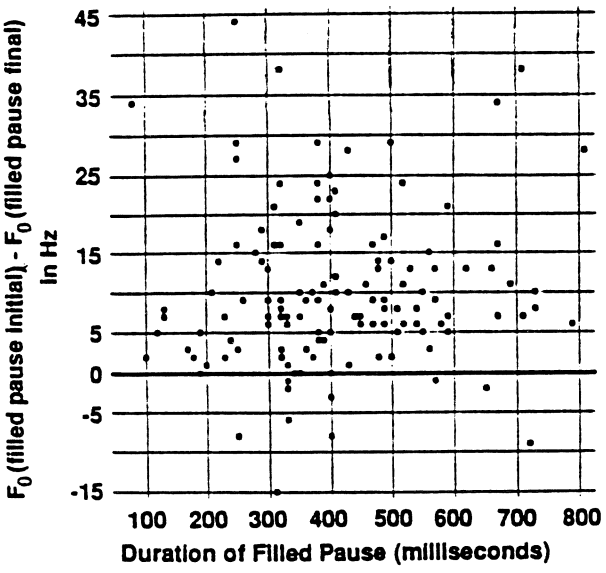


Figure 6: Effect of Filled-Pause Duration on Filled-Pause Fall

4.5. Effect of Additional Variables

Results of regressions performed using the observed terminal low F0 values and selecting independently for values of additional variables are shown in Table 4.

Table 4: Effect of Additional Variables

Data in Analysis	RMS error (Hz)	# of tokens
all data	15.41	169
male speaker	12.36	94
female speaker	18.42	75
peak on sentence-initial accent	30.30	26
peak not on sentence-initial accent	10.90	143
no other disfluency present	14.36	141
filled pause precedes repetition	23.90	11
filled pause precedes replacement	13.09	7
filled pause precedes fresh start	17.90	9
filled pause is "um"	15.29	86
filled pause is "uh"	15.20	83

As can be seen, the factor most influencing prediction accuracy was whether or not the preceding peak was marked on a sentence-initial accented syllable. Although conclusions cannot be drawn given the small number of tokens of this type, it is worth noting that the error in prediction was always in the same direction, with the actual filled pause occurring at a higher F0 value than predicted by the model.- Tokens not involving disfluencies had a lower standard error than that observed overall;, however, results for the different types of disfluencies were inconclusive due to small sample size. Prediction error was not affected by whether the filled pause was "um" or "uh" (although "um" tokens were significantly longer in duration than "uh" tokens, and it should be borne in mind that the present model predicted only the initial F0 of the filled pause.) Prediction accuracy was also not affected by the sex of the speaker; that females had a higher standard error than males was expected given the roughly 50% higher terminal low F0 values for the females.

5. DISCUSSION

5.1. Evaluation of Hypotheses

Two different sets of spontaneous speech data were examined to explore the relationship between the F0 of clause-internal filled pauses and their surrounding context. Results show that the initial F0 of clause-internal filled pauses scales with the F0 of preceding peaks, strongly supporting the "relative" hypothesis.

5.2. Modeling the relationship

Inspection of data from individual subjects revealed that in addition to the scaling of filled pause F0 with preceding peak F0, there was also a lower bound of filled-pause F0 values, and a compressive effect on the size of the drop from the preceding peak to the filled pause as peaks approached the lower portion of a speaker's range.

A model of filled-pause F0 was proposed to reflect these observations. The model was not necessarily intended to have any theoretical interpretation, but rather simply to predict the value of filled-pause F0 using other accessible values of F0. Filled-pause F0 was expressed as a function of three values: (1) a speaker-dependent fixed terminal low F0 value (representing the speaker); (2) the value of the preceding peak F0 (representing the particular prosodic context); and (3) a fixed, speaker-independent scaling factor, r (to express the relationship between the two previous values and filled-pause F0). This is an extremely constrained model, with only one free parameter (r). In addition, the constant term in the model corresponds to a speaker's empirically measured terminal low F0, as opposed to some

F0 value unrelated to prosodic phenomena (for example one outside the speaker's range). Clearly, the current model could also be rewritten to be expressed using coordinates related to a different model (for example, a declination model); the present model is at least as parsimonious as any alternative model in which the functions rewriting peak and terminal low F0 in terms of other variables are linear.

One certainly cannot draw conclusions about the appropriateness of models based on examination of the limited set of data used in the present study. Nevertheless, it is impressive how well the proposed model was able to predict the data. Of possible linear models (there was no evidence for a nonlinear relationship when data from individual subjects were examined) the present model performed extremely well, producing results only very slightly less accurate than a linear model with an additional parameter (in which the coefficients of peak and terminal low F0 were allowed to vary freely.) Real evidence in support of a model such as the present one, however, will probably have to come from comparison of r in the present model to scaling factors proposed in studies of other prosodic phenomena, for example low-tone scaling or the scaling of parentheticals.

5.3. Values of r

It was found that the average value of the parameter r, which expresses the proportion of the distance from terminal low F0 to peak F0 at which filled-pause F0 occurs, did not differ across the American and British data sets. This suggests that the intonation of clause-internal filled pauses, at least as measured by the relationship between preceding peak F0 and initial filled-pause F0, may be independent of factors such as dialect and discourse setting. Mean r values also did not differ across sex. Since speaker sex is highly correlated with the terminal low F0, this lack of a difference in r between sexes is consistent with the appropriateness of a linear model.

5.4. Optimal Reference F0

The value of terminal low F0, a speaker-dependent variable corresponding to the lowest observed F0 value produced after a terminal fall, was found to be slightly lower than the value which optimized prediction. The overall standard error over the data set was slightly decreased when the value of terminal low F0 was raised by 10% for each speaker. A larger data set, with more tokens per speaker, is needed in order to further investigate this finding; it suggests, however, that the value used to scale pitch over the course of an utterance is higher than the F0 measured after final lowering. This is consistent with proposals in the literature [e.g., 8], although it does not distinguish between a declination model and one in which F0 falls abruptly at the end of an utterance. It should be noted that the decision to use the lowest observed terminal low F0, as opposed to other possible values (for example, the mean of all observations) was made because the aim was to get a stable estimate for each speaker, given a positively skewed distribution of low F0 values. Using values such as the mean would therefore be inappropriate. That is, by using mean low F0, one cannot improve results in a principled way, whereas by using a stable estimate such as minimum low F0 (assuming however that there are enough observations available to adequately estimate this value), one can examine the relationship between minimum low F0 and the F0 that optimizes prediction. For exploratory purposes, however, an analysis using mean low F0 values was performed post hoc on the present data set. Results showed a marked reduction in prediction accuracy, and a distribution of r values with much higher standard deviations. Nevertheless, it is conceivable that an analysis using mean low F0 values on a different set of data could produce better results than an analysis using minimum F0 values; such a result would not be meaningful, however, but would rather be due to the fact that mean low F0, like optimal reference F0, is higher than minimum low F0.

5.5. Effect of Duration

Results also suggest that the intonation of filled pauses may be independent of temporal variables. As shown in Fig. 5, there was no correlation between the size of the drop in F0 from the preceding peak to the filled pause and the distance (in time or syllables) between these points; i.e. filled-pause F0 was unrelated to whether or not words and/or silent pauses intervened between the preceding peak and the filled pause. Also, rather surprisingly, there was no correlation between the duration of the filled pause and how far in F0 it fell, as shown in Fig. 6. Most clause-internal filled pauses have a slight linear fall; the fact that longer filled pauses do not fall to a lower F0 than shorter filled pauses implies that the longer tokens either start out with a shallower falling slope, or that they level off in F0 once they reach a point that is "too low" for the local prosodic range. It is also possible that for long hesitations, speakers may stop the filled pause completely and use a silent pause when they have dropped too far. Future work will attempt to examine these issues more closely. These results add further support to the notion that clause-internal filled pauses are in some sense "well-formed" since the range of F0 values for a filled pause is determined by the local prosodic context. In addition, these findings suggest that prosodic regularities in filled pauses may be found more in F0 than in duration measures; this possibility seems reasonable because hesitations, by definition, interrupt the temporal course of production.

5.6. Effect of Sentence-Initial Peaks

As shown in Table 4, prediction error of the proposed model was much greater for filled pauses following peaks marked on sentence-initial accents than for filled pauses elsewhere. In each case following a sentence-initial peak, the prediction of the model for filled-pause F0 was lower than the

observed value; when this relatively small set of tokens was removed from the analyses, the overall error in prediction was reduced substantially. This finding is consistent with the notion that the F0 of filled pauses preserves information about the current prosodic context: filled pauses after peaks corresponding to extra-high sentence-initial accents are themselves extra-high.

5.7. Implications for Areas of Research

The finding that the F0 of filled pauses is relative to prosodic context has implications for models of human speech perception, automatic speech recognition, and for theoretical and descriptive studies of prosody.

The low F0 of filled pauses may help explain why listeners have trouble locating them with respect to words in the message stream; low F0 may also contribute to listeners' ability to filter out filled pauses in comprehension. Experiments designed to test these hypotheses, by using resynthesis to "lift" filled pauses up to the F0 of the region of the lexical material in an utterance, will be conducted in future work. These tests predict that raising the F0 of filled pauses will facilitate listeners' ability to locate them, and also possibly impair comprehension. The finding that the F0 of filled pauses is relative to prosodic context suggests that speakers may attempt to preserve the current prosodic range when hesitating, possibly to inform the listener that they intend to continue where they left off, rather than to abandon a portion of the utterance preceding the filled pause. Thus, a question to be pursued in further work is whether there is a difference between filled pauses that interrupt otherwise fluent clauses, and those that occur at the interruption point of a repair or before a fresh start, since in the latter cases the speaker is abandoning previous material. There were not enough examples of filled pauses in repairs or fresh starts in the present data set to address this question; however preliminary results of additional data suggest that very brief filled pauses, which fall rapidly in F0, often mark a repair (but these are not necessary features for the marking of a repair), and that an unexpectedly high F0 on a filled pause seems to be a very good indicator of a fresh start (essentially an F0 "reset" to begin a new utterance after the filled pause).

Speech recognition systems may be able to take advantage of predictably low F0 in spotting filled pauses. In order to do so successfully however, at least in the case of filled pauses within a clause, these systems will need to take into account the intonation of the local context, rather than using absolute speaker-specific F0 values. Spoken language systems may also benefit from knowing more about prosodic differences between filled pauses in different syntactic environments. Preliminary analyses suggest that whereas clause-internal filled pauses nearly always have a low and falling F0, filled pauses that occur turn-initially or between sentences often have a higher and level or even slightly rising F0. Such information should aid attempts to recognize filled pauses; in addition the recognition of filled pauses having these different prosodic characteristics could contribute information about sentence structure for natural language processing.

As linguists move from the study of read or rehearsed speech to spontaneous discourse, it should become increasingly important for them to consider the prosody of disfluencies, since as shown in the present study, some phenomena considered to be disfluent may exhibit prosodic regularities. This work also suggests that in the case of clause-internal filled pauses, F0, rather than duration, may be the most important prosodic feature to explore. It should prove useful for linguists to include methods for annotating disfluencies in systems developed for the prosodic labeling of spontaneous speech.

6. CONCLUSION

This work has shown that the F0 of one type of speech disfluency, the clause-internal filled pause, is related to the intonation of surrounding material in the message stream. Further work in this area could enhance our knowledge of the production and processing of spontaneous speech, help us learn how to apply these findings to aid speech recognition, and encourage the consideration of hesitations and other disfluencies in theoretical and descriptive work on prosody.

ACKNOWLEDGMENTS

We wish to thank Mark Anderson for helpful discussions on the modeling of F0, and John Bear and Beth Ann Hockey for suggestions regarding syntactic-based principles for categorizing filled pauses. The research of the first author was supported by the Defense Advanced Research Projects Agency under Contract ONR N00014-90-C-0085 with the Office of Naval Research, and also by NSF Grant IRI-890529 from the National Science Foundation. The second author was supported by Award number 87310722 from the UK Science and Engineering Research Council. The opinions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

1. Butzberger, J., H. Murveit, E. Shriberg, & P. Price, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

2. Liberman, M. & J. Pierrehumbert, "Intonational Invariance under Changes in Pitch Range and Length," *Language Sound Structure*, M. Aronoff and R. Oehrle (eds.), MIT Press, 1984.

3. Maclay, H. & C. Osgood, "Hesitation Phenomena in Spontaneous English Speech," *Word*, 15, pp. 19-44, 1959.

4. MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

5. Martin, J., and W. Strange, "The Perception of Hesitation in Spontaneous Speech," *Perception & Psychophysics*, 3, pp. 427-38, 1968.

6. Nooteboom, S., P. Brokx & J. De Rooij, "Contributions of Prosody to Speech Perception," *Studies in the Perception of Language*, W. Levelt and F. D'Arcais (eds.), John Wiley and Sons, 1978.

7. O'Shaughnessy, D., "Recognition of Hesitations in Spontaneous Speech." *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 521-524, 1992.

8. Pierrehumbert, J., "The Phonology and Phonetics of English Intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

9. Shriberg, E., "Intonation of Filled Pauses in Spontaneous Speech." Paper presented at the Conference on Grammatical Foundations of Prosody and Discourse, July 5-6, Santa Cruz, 1991.

10. Shriberg, E. & Lickley, R. "Intonation of Clause-Internal Filled Pauses. *Proceedings of the International Conference on Spoken Language Processing*, 1992.

11. Shriberg, E., E. Wade & P. Price, "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992

THE PHONETICS OF DISCOURSE: STRONG SYLLABLE POSITIONS IN MEXICAN SPANISH AND BRAZILIAN PORTUGUESE

Antônio R. M. Simões

The University of Kansas
Department of Spanish and Portuguese
Lawrence KS 66045

ABSTRACT†

The main objective of this investigation is the analysis of segment duration in discourse in order to provide a basis for the study of lexical and intonational stress in Spanish and Portuguese. The present study is limited to the study of lexical stress in discourse. Duration of syllable nuclei is studied acoustically and perceptually from spontaneous speech recordings.

In the acoustical study, measurements using Sona-Graph™ were taken from speech samples recorded from Mexican and Brazilian television broadcasts, and from dialogues between four subjects in interviews led by the researcher. The results indicate certain "areas of prominence" at the discourse level that can be contrasted in both Mexican Spanish (MSpn) and Brazilian Portuguese (BP). These can be found at any linguistic level: segment, syllable, word, sentence or discourse. Due to "areas of strength" there are syllables that will preserve most of their quality at those "strong points" and lose quality in "weak points". The acoustical results in this study indicate that syllables are strong in MSpn at the stressed and poststressed positions, whereas in BP the strong position is at the stressed syllable position only.

In addition to these results, measurements of duration in non-prepausal position, show that vowels can be longer in pretonic and postonic word position. It was not attempted, however, to determine which lenghtening factors operate in the languages studied.

In the perceptual study, subjects were able to point out which vowels were deleted and replaced by noise. In a group of twelve speakers of American English, Spanish and Brazilian Portuguese, American subjects performed better, indicating that probably it was due to their awareness of forms in a second language.

An additional finding is the presence of recurring intonational phrase in Spanish and BP. In Spanish, these "phonetic continuums" are often characterized by a relatively flat prefinal melodic curve, with an overall varying durations of 1.2 to 2.0 seconds. In BP, such continuums have durations of .9 to 1.2 seconds, and varying melodic curves. Such results may profitably be linked to metrics in poetry, and consequently to the study of metrics and any other area of research in phonetics at the sentence and discourse domains.

† This study was supported by the General Research Fund of the University of Kansas.

1. INTRODUCTION

This is a study of stress which forms part of a long-term project to find the principles of production and perception in Spanish and Portuguese, within the realm of experimental phonetics. Any scholar who has studied linguistic stress, whether lexical or intonational, knows how problematic it can be to find an explanation or definition of "stress" (see Bolinger: 1961; Lehiste: 1970; Ladefoged: 1982). Therefore, it is not clear in the general theory of phonetics what "stress" means. I find it helpful though, to use Ladefoged's (1975) remarks on this topic which states that the major components of stress are fundamental frequency, intensity and duration. It is extremely difficult to prove or disprove this claim empirically. Even to prove the existence of truly minimal pairs in discourse, in terms of lexical stress, is problematic. First, words like the noun "record" and the verb "record" cannot be considered minimal pairs since in English the vowels in these words do not have the same quality. Truly minimal pairs in terms of contrastive lexical stress can only be found in very few words like the noun "rebel" and the verb "rebel" in certain varieties of English. Although in BP and Spanish discourse, it is not easy to find minimal pairs contrasting stress, it is not as

difficult as it is in English. There are also serious obstacles to extract these minimal pairs properly from spontaneous speech for experimental work with them. Brazilian Portuguese, for example, is somewhat similar to English in terms of the constantly changing quality of vowels. As an illustration, suppose one attempts to extract minimal pairs in BP, using word pairs such as "evitar" (to avoid) and "evita" (s/he, it avoids), hoping that the speakers will produced expected minimal pairs [e.vi.tá] (to avoid) and [e.ví.ta], (s/he, it avoids). In other words, Brazilian linguists know that these are common forms of pronunciation in discourse, but they also know that the written "r" may surface in spontaneous speech, the vowel [e] in the first syllable may raise to an [i], and that the postonic [a] may show some degree of centralization. Therefore, such variations make extraction of minimal pairs difficult because a great amount of speech recordings is necessary and most of the data will not be used for this particular goal.

At this stage of the process, one may either hypothesize that there is a functional or contrastive entity "stress" or that there is not. If there is such an entity, its fundamental components also have to be defined and then we need to verify whether or not these fundamental components exist and to what extent. In the case of BP and MSpn, the hypothesis here is that there is contrastive stress. Again, only after one has gathered valid data from spontaneous speech in contrast to any artificial elicitation of minimal pairs, can the existence of contrastive stress be empirically shown. On the basis of linguistic intuition, we make the assumption that contrastive stress exists in MSpn and BPand attempt in this paper to find to what extent duration can be considered a major component of stress in BP and MSpn.

Although the present investigation deals only with recordings of Latin American Spanish and Brazilian Portuguese, the long-term goal is to include all varieties of these languages. The term "strong position" refers to any highlighted or prominent syllable nucleus in the discourse. "Strong positions" may be found in any linguistic domain: for example, in the segment domain at syllable nuclei filled by inherently strong vowels such as, in Spanish and BP, the [a] vowel relative to vowels such as [e] or [o], at stressed position within a word, in diferent areas of a sentence, in the intonational phrase, etc. Thus, these "strong positions" include "stress" in the word domain and "accent" in the intonational domain. In sum, "strong positions" are recurrent patterns of better defined acoustical images at any point in the discourse. My goal is to find out what factors influence the occurrence of recurrent patterns of clear acoustical images, and then how to interpret them at the perceptual level.

To close these introductory remarks, it is important to mention that results from this study coincide in part with studies of metrics the Spanish and Portuguese poetic traditions. It is hoped, that by using the experience accumulated in a different area of language analysis, we can find different ways to look into discourse analysis, and by doing so, explain various linguistic phenomena from this perspective.

2. Relevance of Studies in Duration

Although there is no phonemic contrast between long and short segments in MSpn and BP, there are many reasons for analyzing duration. Fant (1970, 224) had already observed that "the simple and fundamental cue of duration deserves greater attention than is conventionally paid to it." In addition, I would like to note that acoustics and perception of speech, namely functional integration of all linguistic components, cannot be dissociated from the time axis.

Investigators have dealt with temporal organization of speech sounds in a variety of of ways. A brief look into applied and theoretical areas of speech analysis such as building of speech models, or in descriptive works dealing with rhythmic patterns, intrinsic duration, semantics, and syntax, will show continuing efforts to understand the temporal patterns of speech sounds. One of the major obstacles in studies of duration is to decide precisely where a sound segment begins and ends. In the case of a vowel, its duration may be the partion that goes from the onset of the formant structure to its offset. The terms onset and offset are used in the same sense they are used in Lehiste and Peterson (1961). Or, vowel duration may extend beyhond these boundaries, as seen in some studies (e.g. Parker and Diehl, 1984). In order to eliminate inconsistencies, one has to decide where segmental boundaries are located, and develop procedures for finding them.

One of the first works that dealt with phonological and phonetic factors in terms of temporal organization of sound segments is the work of A. Martinet (1949) which points to a universal tendency in languages to shorten vowels following tense consonants, lengthen vowels that follow lax consonants. Jakobson and al. (1952) observed that tense consonants, i.e. [f,s,š,p,t,k], are longer than lax consonants, i.e. [v,z,ž, b,d,g]. A number of other studies have dealt with duration in the phonological and phonetic domain, such as Fry's (1955) where duration is shown to

be a more effective cue for judgments of stress, the study by Miller and Nicely (1955) that proposes the acoustico-physiological feature "duration" to distinguish [s,š,z,ž] from twelve other consonants, and Peterson and Lehiste's (1960) study which noted that duration is affected by the nature of the consonant after a syllable nucleus, namely, a syllable becomes longer when followed by a voiced consonant and shorter when followed by a voiceless consonant with the longest syllable nucleus occurring before a voiced fricative. Other languages show some difference in how segment duration is patterned. Chafcouloff and al (1976) observed that in French all constrictives, *viz.* fricatives, especially in bisyllabic words, become longer if followed by [i,y,u], but [š] becomes shorter in contact with the rounded, mid, front [œ]. In Italian, Ferrero et al. (1979) showed that shortening of frication duration in Italian unvoiced fricatives [s,š] does not bias perception toward the corresponding [z,ž] as English does (Cole and Cooper, 1975), but instead toward the unvoiced affricates [ts,tš]. Major (1981) and Nobre and Ingemann (1987) concluded that duration is a major component of stress in BP. In Spanish, Delattre (1966) finds that closed *syllables* are longer than open ones, whereas in Navarro Tomás (1967) we find a study of vowel duration where *vowels* in closed syllables are shorter than vowels in open syllables. Navarro Tomás also observes the kinds of consonants that affect differently the vowel duration.

Rhythmic factors may also affect word duration as Pike (1945) has argued. On the other hand, studies by Kozhevnikov and Chistovich (1965) and Noteboom (1972) propose that rhythm is independent of word duration, viz. subjects are much more aware of duration if a monotonous picth is used than if normal pitches are used. Simões (1987a) notes the inadequacy of relating word duration to rhythm in BP, and Kelm (1989) also sees no results that makes such a correlation in both MSpn and BP.

Studies at the word level by Lehiste (1970), Raphael (1972), Umeda and al. (1973), Simões (1980) indicate that duration is linked to position within word, length of word, word boundary and the interaction of these factors. Major (1981, 1985) observed that in BP any phonological process has to take place first at postonic position, then pretonic and then in stress position.

In term of syntactic factors, it is important to mention the work of Gaitenby (1965, mentioned in Klatt (1976)) that shows duration as a parameter that delimits syntactic units. Other investigations followed confirming duration as a cue that signals syntactic boundaries: Harris and Umeda (1974), Lindblom (1978), and Schreiber and Read (1982), to mention some.

These results have helped researchers in the design of speech models, especially models that attempt to include characteristics related to semantics (Umeda:1975; Klatt:1976): emphasis, contrastive stress, topicalization (focus), and word novelty. Schreiber and Read (1984) gave extra duration to words at syntactic boundaries to improve listening comprehension in children. However, they did not observe improvements in the listening comprehension of adults under the same conditions.

Therefore, as we study lexical stress in discourse, other factors have to be considered whenever a given duration becomes relatively shortened or lenghthened. These factors include as mood or physical condition of the speaker (extralinguistic), emphasis or word novelty (semantic consideration), prepausal lengthening (sentence domain), intrinsic features of speech sounds, and so forth as set forth in Klatt (1976). It is not in the scope of the present study to observe which and how these factor operate in Spanish and Portuguese.

Some works that used duration in the design of models are those of Klatt (1976), for English, which suggests that recurrent rules to shorten or lengthen inherent durations from smaller units (phonemes), i.e. locally, up to higher units (sentences), operating cyclically; Lindblom and Rapp (1973), for Swedish, suggest the inverse operation, viz. from higher units into smaller units, down to the word domain, only. Later, Lindblom and al. (1981) extended Klatt's (1976) formula and developed several hypotheses about the psychology of speech timing. In their extension of Klatt's formula, Lindblom and al. (1981) applied duration rules cyclically from smaller to higher units. Although Klatt's (1976) model does not necessarily reflect speech production processes, Simões (1987) suggested a redefinition of inherent duration of a sound in BP in terms of median duration in case one attempts to extend the model in terms of speech production processes. By the same token, since there is no mininal duration at discourse domain, namely a syllable nucleus can be completely deleted, Klatt's (1976) linear equation can be simplified from $D_O = K * (D_i - D_{min}) + D_{min}$ *to* $D_O = K * D_i$.

As a general synthesis of these results we may say that duration at the phonological and phonetic level is affected by language specifics and articulatory effort. When a minimal effort is required, there is a reduction of the targeted sound segment, whereas under maximum articulatory effort there is a lengthening of the targeted

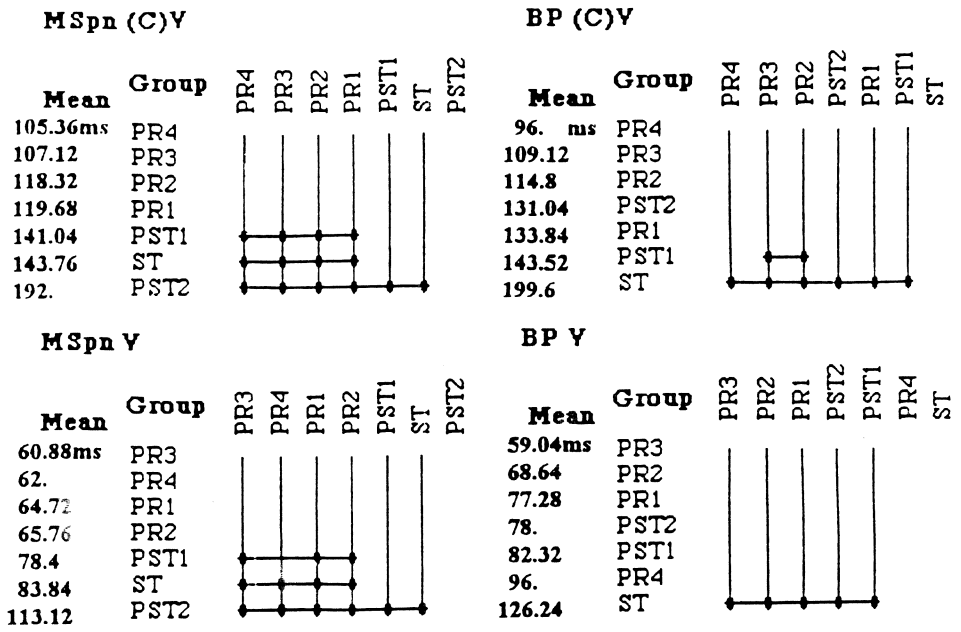
segment. The sounds adjacent to the targeted sound are affected differently in each language. Articulatory effort will be greater during vocal cords vibration, stressing of a sound-segment, or in articulatory displacement (distance) during changes in articulatory gestures. As Lehiste (1970) remarks, the greater the displacement the longer the sound-segment.

3. Results and Discussion

The experimental procedures for first stage of this study was as follows. Initially, television broadcasts via satellite, from Mexico and Brazil, were recorded, first into a regular video cassette and then from the video cassette into audio cassettes. From these audio cassette recordings, both the syllable nuclei, and sequences of consonant plus a syllable nuclei were measured. Segmentation procedures followed those described in Simões (1987) with modifications made as needed.

More than four hundred measurements of these durations were taken, several statistical tests were made. Results of an ANOVA multiple range test of the relationship between all syllables measured are shown in Figure 1. Since there is no significant difference between measurements taken of the syllable nuclei and measurements of a consonant followed by a syllable nucleus, only the measurements of the syllable nuclei will be discussed. The abbreviations in Figure 1 mean the following: ST, stressed; PR1, prestressed, *viz.* one syllable before the stressed one; and similarly PR2, PR3, PR4; PST1, poststressed, *viz.* one syllable after the stressed syllable; and similarly PST2, PST3. The results indicate that all relationships are linked to lexically stressed and poststressed syllables in MSpn and to stressed syllable in BP. Such results are probably showing that these prominent syllables are significant points of reference in discourse in terms of production.

Figure 1: ANOVA results of multiple range test. The intersection of lines denotes pairs that are significantly different at the .05 level. Method-1 is indicated by (C)V and method-2 by V.



Following the results from the acoustical analysis as seen in Figure 1, a perceptual test was designed. In this perceptual study, sentences from the same recordings, as the acoustical analysis described above, were modified in the DSP5500 Sona-Graph™. These modifications were made in terms of sentence and word domains (see Appendix). At the sentence domain modifications took place near the beginning or the end of the sentences; then, at those points in the sentence, vowels in prestressed, stressed, and postressed syllables were edited out and replaced by background noise of equal duration from portions of the recordings in which there was no

speech. The whole formant structures of vowels were replaced from the onset from the consonant release to the offglide of each vowel, using, in general, the second formant as reference (Lehiste and Peterson:1961; Simões:1987). The vowels replaced are shown in the Appendix.

Native speakers of Spanish listened to the Spanish tapes, native speakers of Brazilian Portuguese listened to Brazilian Portuguese tapes and native speakers of American English with near-native speaker proficiency of Spanish and/or Brazilian Portuguese were also asked to listen to the tapes. They totalled twelve subjects, four for each group. These subjects were asked to perform two tasks in listening to these modified tapes. First, they were told that these tapes were recorded via satellite and that possibly the tapes lost some information during the transmission or maybe no information was lost. They were asked to comment after each sentence. After having listened to one sentence, their impressions on any aspect of the task (comprehension, speakers accent, their judgement about the test itself, and so forth) were discussed. The intention was to conduct a less structured, less controlled experiment, so that unpredicted results could appear. Thus, instead of the usual procedure in this type of study, namely to give subjects a set of possible answers to choose from, subjects were interviewed and asked to comment on each sentence they listened to, and as they talked I took notes.

After all sentences were played and discussed, I explained to them that, in fact, some words in the sentences had been modified. In the second task, I asked them to listen again to each sentence, and decide: if they thought a word was modified and to point out the word. If they identified a word, I asked them to tell me what the change was and where in the word the change took place.

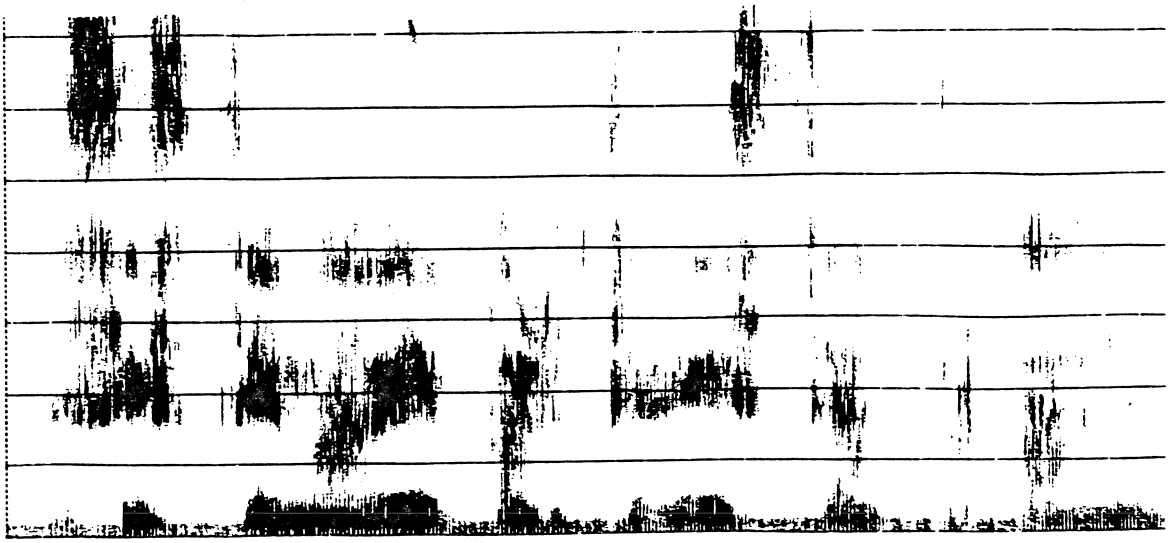
The general results I have gathered, so far, from their reactions, can be summarized as follows. Some of the subjects in this experiment had more difficulty in pointing out the modification at the beginning of a sentence than at the end of a sentence. This is surprising, considering that perceptual tests have shown that subjects tend to perform poorly in specifying where information was deleted or added. Furthermore, the Spanish speakers in this experiment had more difficulty in pointing out where the modification took place and very often did not hear any change. The American subjects performed very well in pointing out in Spanish and Portuguese, where the modifications took place which vowel was missing. The Brazilians were good at that task, too, but they did not identify the correct place as often as the Americans. There was no control of the environments where these syllable nuclei were deleted, since the sentences were selected randomly.

After having examined the speech samples and interpreted the acoustical and perceptual results in the preceding television broadcasts, two interviews in Spanish and Portuguese were made. The speakers were two native speakers of Spanish, one female university student from Colombia and one male university student from Mexico; then, two native speakers of BP, one female university student from the city of São Paulo and one male student from the city of Rio de Janeiro. The four speakers are between 20 and 30 years old.

The duration of syllable nuclei and larger speech units, that is "phonetic continuums" or intonational phrases, was acoustically analyzed in these interviews. The goal here is to observe duration patterns in dialogues, and thus to what extent it played a role as a correlate of lexical stress. In terms of the relevance of duration to stress at the word level, duration can be considered as one of its major components. In discourse, however, factors such as word or phrase final lengthening may cause unstressed vowels to be longer than stressed vowels. This applies to both Spanish and BP. One example from Spanish is shown in Figure 2. Although it is not difficult to find in non-prepausal position stressed syllable nuclei shorter than unstressed ones in Spanish and BP, it is more common to find these occurrences in Spanish.

In larger units, phonetic continuums, the equivalent of intonational phrases, were observed in this study to fall within the intervals of .9 to 1.2 seconds in BP and 1.2 to 2.0 in Spanish in the recordings made for the dialogues. These continuums are often characterized by an overall flat prefinal intonation as depicted in Figure 3. This of course does not mean that all Spanish intonation is like this example. Such continuums, in terms of duration, seem to be equivalent to the eight-syllable verse in Spanish and the seven-syllable verse in Portuguese, which coincide somewhat with the English iambic tetrameter, under the English tradition of the syllabic foot.

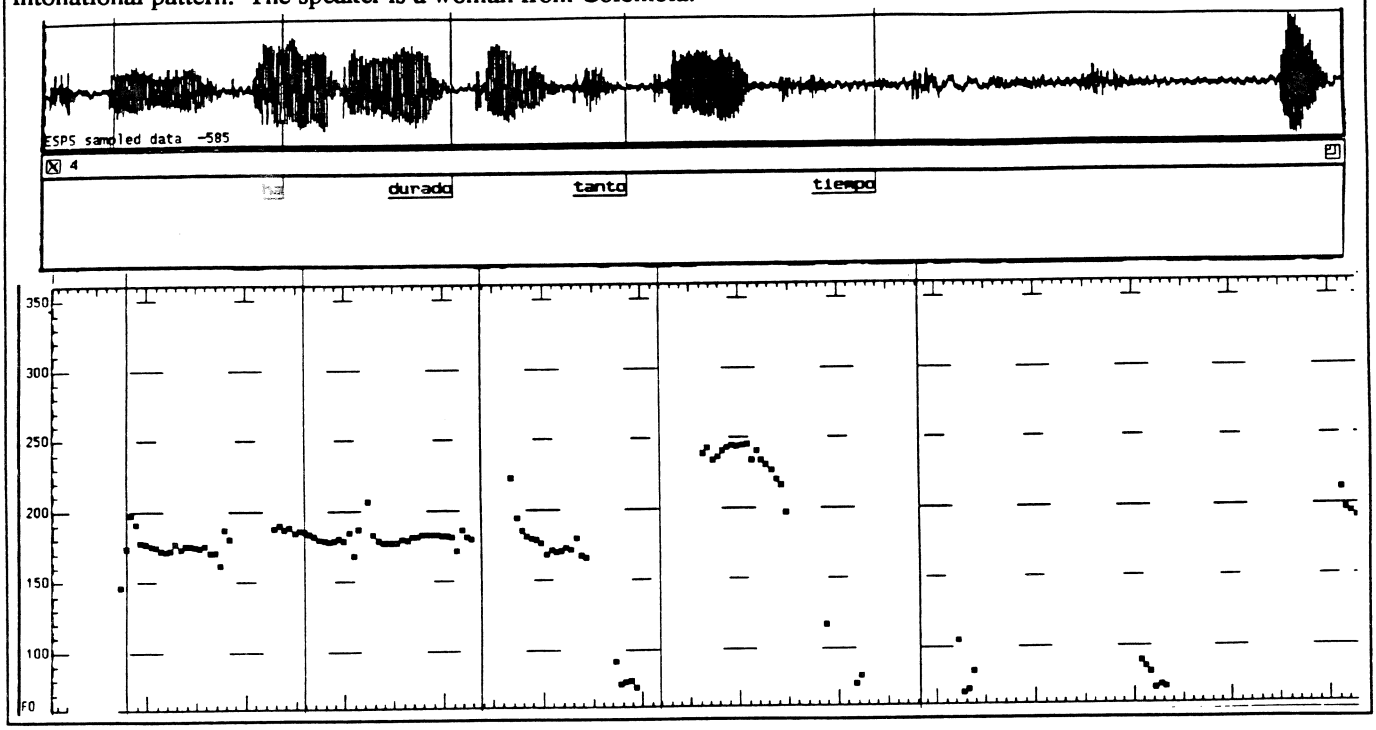
Figure 2: Speech sample "... sistema bipartidista ..." (ing. "bipartisan system") from Spanish, spoken by a Colombian woman. The word "bipartidista" is in prepausal position. Lexically stressed syllables are "-te-" in "sistema," and "-dis-" in "bipartidista." In [sis.t̪é.ma] the postressed [a] is longer than the stressed [é], and in [β̞i.par.t̪i.d̪ís.t̪a], all prestressed vowels are about the same or longer than the stressed [í].



[s i s t̪ é m a β̞ i p a r t̪ i d̪ í s t̪ a]

time in seconds .075 .088 .084 .078 .1 .075

Figure 3: An example of recurring phonetic continuum in Spanish extracted from spontaneous dialogues: a flat prefinal intonational pattern. The speaker is a woman from Colombia.



Analysis of the data has brought to my attention a possible link between the present results and techniques for metrics in poetry. If one looks into versification in Spanish and Portuguese, it will be observed that syllable counting in these languages reflects what has been observed here. In other words, Spanish metrics, similar to Italian, counts the number of syllables in a verse by adding one count after the last stressed syllable, regardless of the presence or absence of a poststressed syllable. Portuguese had the same syllable counting until the last century. Since then, starting with Castilho (1908) in 1850, and confirmed in Spina (1971), metrics in Portuguese is done by counting up to the last stressed syllable in a verse, and no longer adds another count. French and Provençal are known to count syllables similarly to Portuguese.

The length of what I have been calling a "phonetic continuum" in the dialogues analyzed varies as expected, but there are some generalization that can be drawn. As mentioned in the results, these phonetic continuums fall into different patterns in Portuguese and Spanish. If we refer again to poetic metrics, it will be observed that, similarly to what one finds in spoken language, although a verse in poetry can vary from one to usually twelve syllables, poets, and we may infer native speakers who listen to poetry, have their preferences to convey their messages. In BP, a verse with seven syllables is the most used by poets, and the most popular type of verse for both reading and singing; in Spanish, the preference is for eight-syllable verses in singing and eleven-syllable verses in reading. This preference can, in my opinion, indicate what patterns to look for in discourse.

Examination of spectrograms in the present study shows that signal amplitude of Spanish words is clearly maintained not only on stressed syllables but also on poststressed syllables in physical and structural (i.e. expected) prepausal positions, creating a prolonged intonation at these points in the discourse. In BP, in such positions, poststressed syllables tend to disappear at the discourse level, creating a damping, namely a rapid decrease in signal amplitude after the stressed syllable.

4. Conclusion

According to Ladefoged (1982:104) "The most reliable thing for a listener to detect is that a stressed syllable frequently has a longer vowel." This seems to be in fact the case for lexical stress, and especially in English. In Spanish and Brazilian Portuguese *discourse*, especially in Spanish, although duration is one of the major components of stress, as we go into spontaneous speech, in non prepausal position, duration is not the most reliable correlate of lexical stress. Perhaps, in the case of Spanish and Portuguese, amplitude may play a more reliable role for in the identification of stress in discourse.

To summarize the major points of the present study, *acoustically* (1) in discourse, in prepausal and non-prepausal, lexically stressed syllable nuclei are often shorter than unstressed syllable nuclei; (2) in Mexican Spanish, lexically stressed and poststressed syllables are the most significant syllables in spontaneous speech, whereas in Brazilian Portuguese only the stressed ones have been observed as most significant syllables. This paper calls such syllables as strong syllables; (3) again, still in acoustical terms, recurring phonetic continuums were observed to measure between 1.2 to 2.0 seconds in MSpn and .9 to 1.2 seconds in BP. These phonetic continuums coincide with studies in metrics in the poetic tradition in Spanish and Portuguese. Finally, in a *perceptual* study (4) subjects were able to identify place and quality of vowels that were replaced by noise in sentences extracted from television broadcasts, and that subjects performed better in these tasks in a second language.

It seems to me that a theory of stress has to be elaborated in terms of spontanceous speech, and in terms of physiological, acoustical, and perceptual correlates.

WORKS CONSULTED

Abaurre-Gnerre, M.B. (1979). *Phonostylistic aspects of a Brazilian Portuguese dialect: implications for syllable structure constraints*, unpub. diss. Buffalo: State University of New York.

Abaurre-Gnerre, M.B. (1981). Processos fonológicos segmentais como índice de padrões prosódicos diversos nos estilos formal e casual do português do Brasil. *Cadernos de Estudos Linguísticos*, 2, 23-44.

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Beckman, M.E. (1986). *Stress and non-stress accent*. Dordrecht: Foris Publications.

Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language*, 37, 83-96.

Câmara, J.M. (1970). *Estrutura da língua portuguesa*. Petrópolis, Brazil: Editora Vozes.

Câmara, J.M. (1977). *Para o estudo da fonêmica portuguesa*. Rio de Janeiro: Padrão.

Castilho, A.F. de (1908). *Tratado de Metrificação Portugueza*, In *Obras completas de A.F. de Castilho*, LVI, 5.ª

edição, vol 1. Lisboa, Portugual: Empreza da História de Portugal, Livraria Moderna.

Chafcouloff, M. et al. (1976). Effets de la coarticulation sur les caracteristiques acoustiques des contoïdes fricatives du français. In *Travaux de l'Institut de phonétique d'Aix*, 3, 61-113.

Cole, R.A. and W.E. Cooper (1975). Perception of voicing in English affricates and fricatives. In *Journal of the Acoustical Society of America*, 58, 1280-87.

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics*, 4, 183-98.

Fant, C.G.M. (1970). Acoustic theory of speech production, 2nd ed. The Hague: Mounton.

Ferrero, F.E. and al (1979). Fricative duration of Italian unvoiced fricatives. In *Frontiers of speech communication research*, B. Lindblom and S. Ohman, eds. New York: Academic Press, 159-65.

Fry, F.B. (1955). Duration and intensity as physical correlates of linguistic stress. In *Journal of the Acoustical Society of America*, 27, 4.

Flege J.E. & O.-S. Bohn (1989) An Instrumental Study of Vowel Reduction and Stress Placement in Spanish-Accented English. *SSLA*, 11, 35-62.

Godínez, M. (1978). A survey of Spanish and Portuguese phonetics. *UCLA Working Papers in Phonetics*.

Green, J.N. (1988). *Spanish. The romance languages*, M. Harris & N. Vincent, eds. New York: Oxford University Press.

Harris, J.W. (1983). *Syllable structure and stress in Spanish.* A nonlinear analysis. Cambridge, MA.: MIT Press.

Harris, M.S. and N. Umeda (1974). Effects of speaking mode on temporal factors in speech. In *Journal of the Acoustical Society of America*, 56, 1016-18.

Jakobson, R., G. Fant, and M. Halle (1969). *Preliminaries to speech analysis*, 8th edition. Cambridge, Ma.: MIT.

Kelm, O.R. (1989). *Temporal aspects of speech rhythm which distinguish Mexican Spanish and Brazilian Portuguese*, unpub. diss. Berkeley, Ca.: University of California.

Klatt, D.H. (1976). Segmental duration in English. *Journal of the Acoustical Society of America*, 59, 1208-21.

Klatt, D.H.& L.C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87 (2), February, 820-57.

Kozhevnikov, V.A. and L.A. Chistovich (1965). Speech articulation and perception, *JPRS* 30. Washington, DC.

Ladefoged, P. (1982). *A course in Phonetics*, 2nd. edition. San Diego: Harcourt Brace Jovanovich.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. & G.E. Peterson (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, 33, 268-77.

Lindblom, B. (1978). Final lengthening in speech and music. In *Travaux de l'Institut de phonétique de Lund*, 13.

Lindblom, B. and K. Rapp (1973). Some temporal regularities of spoken Swedish. Publication no. 21. Stockholm: Institute of Linguistics of the University of Stockholm (unpub.).

Lindblom, R. et al (1981). *Durational patterns of Swedish phonology: do they affect short-term motor memory processes?* Bloomington, Indiana: Indiana University Linguistic Club.

Major, R. (1981). Stress-timing in Brazilian Portuguese. *Journal of Phonetics*,9,343-51.

Major, R. (1985). Stress and rhythm in Brazilian Portuguese. *Language*, 61, 2, 259-89.

Martinet, A. (1949). Phonology as functional phonetics. *Publications of the Philological Society*, no. 15. London: Oxford University Press, 1-27.

Miller, G.A. and P.E. Nicely (1955). An analysis of perceptual confusion among some English consonants. In *Readings in acoustics phonetics*, 301-15

Navarro Tomás, T. (1967). *Manual de pronunciacón española*, 6ª edición. Madrid: Consejo Superior de Investigaciones Científicas.

Navarro Tomás, T. (1983). *Métrica española*, 6ª edición. Barcelona: Editorial Labor.

Nobre, M.A. and F. Ingemann (1987). Oral vowel reduction in Brazilian Portuguese. In R. Channon and L. Shockey, eds., *In Honor of Ilse Lehiste*. Providence, USA: Foris Publication.

Noteboom, S.G. (1972). Temporal patterns in Dutch. In *Proceedings of the 7th international congress of phonetic sciences*, 984-89.

Parker, E.M. and R.L. Diehl (1984). Identifying vowels in CVC syllables: effects of inserting silence and noise. In *Perception & Psychophysics*, 36 (4), 369-80.

Parkinson, S. (1988). Portuguese. *The romance languages*, M. Harris & N. Vincent, eds. New York: Oxford University Press.

Peterson, G.E. and I. Lehiste (1960). Duration of syllable nuclei in English. In *Journal of the Acoustical Society of America*, 32, 693-703.

Pike, K.L. (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

Quilis, A. & J.A. Fernández (1982). *Curso de fonética y fonología españolas para estudiantes anglo-americanos*, 10ª edición. Madrid: Consejo Superior de

Investigaciones Científicas, Instituto Miguel de Cervantes.

Raphael, L. (1972). Preceding vowel duration is a cue to the perception of the voicing characteristics of word final consonant in American English. In *Journal of the Acoustical Society of America*, 51, 1296-1303.

Schreiber, P.A. and W.Ch. Read (1982). Why short subjects are harder to find than long ones. In *Language acquisition: the state of the art*, Wanner and Gleitman, ed. Cambridtge, Ma.: Cambridge University Press.

Simões, A.R.M. (1980). *Étude acoustique des consonnes [s] et [z] du Brésilien.* Mémoire de D.E.A., ms. Aix-en-Provence, France: Institut de phonétique.

Simões, A.R.M. (1987) *Temporal organization of Brazilian Portuguese vowels in continuous speech: an acoustical study*, unpub. diss. Austin, TX: University of Texas.

Simões, A.R.M. (1987a). Brazilian Portuguese rhythm: stress-timed, syllable-timed, or samba? Paper presented at the *University of Texas Colloquium on Hispanic and Luso-Brazilian Literatures, and Romance Linguistics*, October.

Spina, S. (1971). *Manual de versificação românica medieval.* Rio de Janeiro: Ed. Gernasa.

Umeda, N. et al (1973). Fricative--the physical properties and allophones. In *Journal of the Acoustical Society of America*,53,373(A).

APPENDIX

Sentences for the perceptual test in the first stage of the experiment. These sentences are from television broadcasts from Brazil and Mexico. Sentences were chosen randomly and modifications were made in the first and last content words of each sentence. Vowels of any type were deleted and replaced by background noise from the DSP 5500 Sona-Graph of same duration, in prestressed, stressed, and poststressed positions. Replaced vowels are in bold types, in parentheses, with its phonetic symbol and syllable position indicated at the end of the sentence.

Brazilian Portuguese:

Sentence 1: O g(**e**)neral Agenor Homem de Carvalho esteve hoje no Rio acompanhando a visita do Presidente Collor de Mello. PR2 [**e**] replaced by noise.
Sentence 2: Janine sempre conviveu comigo no Brasil, isso não resta dúvidas, os documentos são suficientes pra provar isso... Mesmo em português... Não resta dúvida de que... eu vou levar Janine, vou provar a verd(**a**)de. ST [**a**] replaced by noise.
Sentence 3: O Min(**i**)stro da Justiça e o Chefe do Gabinete Militar da Presidência são convidados pela Polícia Federal a depor sobre as denúncias contra o ex-ministro Antônio Rogério Magri. ST [**i**] replaced by noise.
Sentence 4: O código de defesa do consumidor completa um ano e cria novos hábit(**o**)s. PST2 [**u**] replaced by noise.
Sentence 5: As mulher(**e**)s foram exigir a regulamentação da aposentadoria aos cincüenta e cinco anos para a mulher do campo. PST1 [**i**] replaced by noise.
Sentence 6: A temperatura chega a trinta graus em São Paulo, trinta e oito no Rio de Janeiro, vinte e nove em Belo Horizonte, vinte e seis em Brasília e trinta e dois em F(**o**)rtaleza. PR2 [**o**] replaced by noise.

Mexican Spanish:

Sentence 1: El d(**e**)stituido mandatario recordó que la Organización de Estados Americanos, la OEA, efectuó gestiones para resolver la crisis haitiana mediante la firma de un protocolo firmado el pasado veinticinco de febrero, en Washington. PR2 [**e**] replaced by noise.
Sentence 2: La falta de medicamentos sigue golpeando a los pacientes de la Caja Costarricense de Seguro S(**o**)cial. PR1 [**o**] replaced by noise.
Sentence 3: En Guatem(**a**)la, el coronel Mario Rolando Terraza Pinto, comandante de la zona militar número veinte, con sede en Santa Cruz del Quiché, dijo que la guerrilla fue derrotada en el departamento occidental de esa región. ST [**a**] replaced by noise.
Sentence 4: El sistema educativo de Costa Rica podría experimentar una huelga debido al malestar que existe entre los docentes por los nuevos programas de est(**u**)dio. ST [**u**] replaced by noise.
Sentence 5: Una band(**a**) de traficantes de niños que pagava trescientos cincuenta dólares por cada menor robado a famílias pobres y luego los ofrecía en adopción a extranjeros, fue desarticulada en Honduras. PST1 [**a**] replaced by noise.
Sentence 6: Un comunicado oficial expedido en la víspera, sostiene que el secuestro y muerte del ex-ministro Durán, resaltan la actitud criminal de quienes conforman la autodenominada Coordinadora Guerrillera, y constituye una muestra aberrante de violación de los derechos humanos, que no tiene ni puede tener justificación algun(**a**). PST1 [**a**] replaced by noise.

THE PROSODIC STRUCTURING OF INFORMATION FLOW IN SPOKEN DISCOURSE

Marc Swerts & Ronald Geluykens

Institute for Perception Research (IPO)
P.O. Box 513, 5600 MB Eindhoven
The Netherlands (*)

ABSTRACT

This paper describes a study on the prosodic demarcation of larger-scale topical units in spontaneous discourse, in terms of various melodic variables and pause structure. The research reported upon centers on a specific kind of spontaneous language use, viz. so-called instruction monologues of three different Dutch speakers. These monologues are such that macro-units can easily be specified on the basis of criteria which are independent of supra-segmental information. It was found that, in order to indicate which stretches of discourse constitute meaningful units, the three speakers indeed exploit both melodic variables (boundary tones, variable height of Fo-maxima, overall downward tendency in pitch over the course of a topic) and pause structure (important points in the flow of information are marked with long pauses, the lenghts of which depend on the deepness of the boundary). However, we also observed some speaker variation, in that not each of our informants appeared to use each of the prosodic demarcation devices to the same extent.

1. INTRODUCTION

This study tries to establish to what extent the topical structure of a spontaneous monologue is reflected in its prosodic make-up. The investigation stems from the general assumption that discourse is more than just the sum of isolated sentences. Indeed, in both spoken and written texts, one can often distinguish homogeneous sequences of sentences that somehow 'belong together'. As they together express one coherent information unit of a speaker or a writer, such groups of utterances can be said to form larger-scale discourse units (which we will label 'topics', cf. infra).

Language users have at their disposal various means with which they can bring out discourse structure. There is some evidence that particular morpho-syntactic devices exist to regulate information flow. Geluykens (1992a) argues, for instance, that the phenomenon of left-dislocation is an important syntactic mechanism to introduce specific topics into the discourse. Many languages have specific particles that introduce or close a paragraph (cf. Schiffrin 1987). Also, the use of pronouns is frequently explained on the basis of the maintenance of coherence in a text (Geluykens 1989, in press).

In a written text, the different supra-sential units can easily be visualized by orthographic means. Macro-units in spoken language, on their part, may have specific prosodic correlates. It has already been reported by e.g. Lehiste (1975) and Thorsen (1985) that read-aloud paragraphs possess a characteristic melodic supra-structure. Lehiste (1975) observes that a speaker can signal by temporal means whether a sentence is paragraph-final or not. Moreover, both researchers found strong indications that these prosodic properties are perceptually relevant: listeners appeared to use such suprasegmental information to decide correctly where in a paragraph an individually presented sentence had to be situated. The latter results, however, were obtained from analyses of read-aloud, small-sized paragraphs; it remains to be seen to what extent they are still applicable to spontaneous speech, which is, after all, a more common, less restricted spoken language use. Spontaneous language use generally involves only little pre-planning, which may have its repercussions on the prosodic properties of macro-units (see Levelt 1989 for a more thorough discussion.

2. SPEECH MATERIALS

However, the study of larger-scale units in spontaneous discourse faces some considerable methodological problems (Swerts & Collier, in press). One serious difficulty is to find an operational definition of a macro-unit: to avoid running into circularity, one is in need of a manageable criterion to specify such discourse units that is independent of the prosodic characteristics of the speech studied. As a solution, this study proposes to investigate a specific kind of descriptive language use, namely the instruction monologues that were used by Terken (1984) to test

specific hypotheses on the distribution of pitch accents (for the details of the total experimental design (subjects, recordings, elicitation procedure, materials), see Terken 1984). These monologues consisted of a series of instructions from a speaker to a listener to assemble the front view of a house from a set of ready-made pieces of cardboard (e.g. a roof, a front door, etc...); as such, it is clear that they have some internal organization, which reflects the different instructions. Moreover, on a purely linguistic level, they exhibited a clear topical structure, as will be shown below. An excerpt of such a monologue is presented below [accented syllables are underscored; English glosses are approximations rather than literal translations]:

(1)

1. dan hebben we het zwarte vierkant
then we have the black square

2. daar gaan we nu een dak opzetten
now let's put a roof on it

dat is het groene driehoek
that is the green triangle

de grote groene driehoek
the large green triangle

die zetten we er boven op
we place that on top of it

3. dan pakken we het woonkamerraam
then we take the living room window

dat draaien we met de kleurzijde om
we turn its colored side up

en leggen het links onderin
and lay it bottom left / leaving

met wat ruimte eronder
some space underneath it

zodat de lange kant evenwijdig ligt aan de
so that the long side is parallel to the

onderkant van het huis
bottom of the house

4. dan pakken we de voordeur
then we take the front door

en die zetten we een eindje rechts van het raam
and we put that a little to the right of the window

met de korte zijde naar onder
with the short side down

We have already noted that the various instructions can be seen as meaningful units, as they consist of semantically coherent utterances dealing with the same building block of the house, i.e. with the same instruction. There is independent textual motivation for our structural analysis, however.

Instructions in the monologues generally are of the following form: a referent (e.g. het zwarte vierkant (the black square)) is introduced, which constitutes the core of the instruction, and on which some action has to be performed. These referents will be labelled discourse topics, following Geluykens (1992b), as they are both 'non-recoverable' and 'persistent'. First of all, they are highly irrecoverable: they concern information which cannot be retrieved, directly or via inferences, from the preceding discourse record (Geluykens 1988a). A good example is het woonkamerraam (the living room window) in the third instruction of (1) above. Secondly, they have a strong degree of persistence (after Givón 1983): they recur in various surface forms in the subsequent utterances that belong to the same instruction. This persistence (see Geluykens 1991, 1992b) can be either direct, through recurrence of the same referent in the subsequent discourse, often pronominalized (e.g. dat (that), het (it) in the same instruction above), or indirect, through mention of a semantically closely related referent (e.g. de kleurzijde (the colored side), de lange kant (the long end)). Such an analysis shows that there is an independent, 'information flow' (after Chafe 1987) motivation for indeed regarding the different instructions as separate topical units. On top of this, there are some other signals which indicate the structuring of the discourse by the speaker (e.g. the use of dan (then) at the start of each new instruction. All this provides us with an informational analysis which is independent from prosodic considerations.

From Terken's (1984) original eleven recorded monologues, three speakers (one male (HZ) and two females (SK, NE)) were selected for further analysis. This choice was determined by the fact that these speakers appeared to have the least problems with the experimental task and thus did not reveal much irrelevant stretches of speech. Moreover, they produced monologues of which the topical structure in terms of the relevant instructions was fairly easy to specify. These monologues were fed into the computer with a 10 kHz sampling frequency at 12 bits. The speech was LPC-analyzed and the fundamental frequency (Fo) was determined by means of a method of subharmonic summation (Hermes 1986).

In this study, two prosodic variables are investigated with regard to the supra-structure of the discourse: (i) the intonation or speech melody (melodic boundary markers, scaling of Fo-maxima and mean Fo of subsequent clauses) and (ii) the temporal structure, more specifically distribution and relative duration of pauses.

3. SPEECH MELODY

3.1. Boundary tones (table 1)

Brown et al. (1980), among others, have already argued that intonation is often exploited to signal topic-continuity or -finality by the use of different melodic boundary tones. Their claim is that so-called 'low terminals' are regularly associated with the end of a topic, whereas 'not-low terminals' would serve to indicate that there is more to come on the same topic. It was checked whether some more evidence for these statements could be found in the three instruction monologues. At the end of each clause in the monologues, the course of Fo was examined from the last accent till the beginning of the next clause; the latter was operationally defined as a syntactic entity containing a finite verb. A classification was made into low-ending contours and high-ending contours. The distribution of these is depicted in Table 1 (all tables can be found in the appendix).

These findings seem to confirm the earlier claims by Brown et al (1980). There is indeed a correspondence between the topical structure of the discourse and the use of low versus high boundary tones. The majority of the low-ending contours are located at the end of the final clauses of the various instructions: their function seems to be to signal that an informational unit has been rounded off. Most of the non-low contours occur within instructions: they signal that there is still more to come on the same topic. However, our data also reveal some exceptions to this general tendency. From the class of low boundaries that do not coincide with the end of an instruction, some can still be argued to be dependent on the topical structure of the discourse (especially in the monologue of speaker SK). That is, they occur at points where the information conveyed is sufficient enough to enable a listener to successfully execute the instruction of the speaker. However, in a sort of afterthought, the subsequent clause provides some details that are redundant from a purely informational point of view or that are so obvious or deducible from the previous discourse that they are strictly speaking not necessary to be communicated. Of course, this non-essential information may facilitate the communication between speaker and hearer, as it confirms what has been said in earlier utterances. In (2), which is instruction 3 of speaker SK, an example is presented an an illustration of what is meant (H% symbolizes a high boundary tone, L% a low boundary tone):

2) dan pakken we de voordeur (H%)
then we take the front door

en die zetten we rechts in het zwarte vierkant (H%)
and we put that right in the black square

rechtsonder (H%) zodat de smalle kant van de voordeur
bottom right so that the small side of the front door

tegen de onderkant van het zwarte vierkant aanzit (H%)
sits on the bottom side of the black square

en een klein stukje een centimeter of twee
and a little bit about two centimeters

vanaf de rechterzijkant van het zwarte vierkant (L%)
from the right side of the black square

dus de onderkant van de voordeur loopt gelijk
so the bottom side of the front door runs parallel

met de onderkant van het zwarte vierkant
to the bottom side of the black square

van de voorgevel (L%)
of the front view

In (2), it can be observed that the first occurring low boundary is located at a position where the instruction is informationally complete. The subsequent utterance only paraphrases what has already been said in the previous part of the instruction. In the three monologues, 5 instances were found of low boundary tones that occurred at the end of an informationally (quasi-)complete unit, that did not coincide with the end of the total instruction.

3.2. Fo maxima (table 2)

Another melodic variable is the location of the Fo maxima in accent-lending pitch movements (Fo maximum being defined as the end of a rise or the beginning of a fall). There were some difficult cases, namely the abrupt pitch rises that occurred relatively late in one-syllable words and that seemed to consist of an accent-lending and a non-accent-lending part. It was difficult to locate the exact transition point between these two parts, as they present themselves as one fluent, complete Fo movement. Therefore, the Fo maximum of the entire movement is taken as a measure point. These 'exaggerated' values are italicized in Table 2. Other Fo maxima are also depicted in Table 2.

Terken (1984) has already observed in these speech materials that noun phrases introducing a new topic into the discourse were always accented. This finding was

interpreted to mean that by accentuation the speaker gives an indication of the degree of availability of the information conveyed. As the referents introduced are always irrecoverable, they always get an accent. Moreover, these new items constitute the topics of the subsequent discourse and are therefore made prominent in order to signal that the referent must be given preferential status in the listener's discourse model. The data on Fo maxima seem to give further support to these ideas as these accents also appear to differ qualitatively from other accents in the discourse: the accents on the referent-introducing noun phrases are very conspicuous since they are located very high in the speaker's register: such very prominent accents may function as 'warning signals' from the speaker to the listener that a new topical unit has been started. However, the latter interpretation only seems to hold for two of the three speakers. Indeed, NE does not consistently provide the topic-introducing noun phrase with the highest Fo maximum as do HZ and SK.

It has to be noted here that the Fo maxima on referent-introductions (Table 2) are not always easy to identify, as some introduction are accompanied by elaborative material, making the introducing phrase quite long (e.g. the second instruction in (1) above: het groene driehoek, de grote groene driehoek). This may account for some of the discrepancies between referent-introductions proper and Fo maxima in Table 2, the Fo falling on the elaborative material.

3.3. mean Fo of subsequent clauses (table 3)

The measurement of the boundary tones and the Fo maxima were relatively local. In addition, some more global calculations of the fundamental frequency were also performed. The mean Fo was determined over the range of one clause (see section 3.1 for an operational definition of clauses). If there was a large pause (i.e. longer than 1000 ms), the clause was split up into two separate units. The measured means of Fo of subsequent clauses are shown in Figure 1.

It can be seen that the larger-scale informational units of speakers SK and HZ appear to exhibit a global phonetic characteristic: the instructions are provided with a superordinate melodic structure. In their data, the Fo is, on the average, relatively high at the beginning of a unit, and it then slowly decreases over the course of the instruction; at the beginning of a new topical entity, the Fo is again shifted up. The macro-units of speaker NE, however, do not have this general prosodic feature.

At this point, it is not clear yet how the global decrease in the mean Fo (that is, in the data of HZ and SK) must be interpreted. Though exact measurements are not yet performed, it is our impression that it is the composite result of two mechanisms present: a general decline in Fo register and a global decrease in the excursion size of the movements. In any case, it suggests that relatively global correlates can be observed in spontaneous discourse (a claim which has been questioned a few times in the literature), provided that the speaker is enabled to pre-plan much of his speaking unit. Of course, this does not mean that this finding also holds for conversational, non-monitored speaking style.

4. PAUSE STRUCTURE

A second major prosodic dimension which speakers may manipulate to structure their information flow is the temporal one, more particularly the use of pauses in discourse. (Pauses are operationally defined as periods of silence, equal to or longer than 100 ms; they were measured directly on the digitized speech waveform.)

4.1. Distribution

A first look at the distribution of the pauses in the three monologues brings to light that many of them occur at the end of a clause or a phrase. However, it would be a mistake to conclude from this that pausal structure can be explained purely in syntactic terms. Apart from the fact that some of the pauses are present at relatively shallow structural breaks (e.g., in between an article and a noun), it is also the case that not all clause boundaries are marked by a period of silence. This is clearly exemplified by (3) and (4) below, where major clause boundaries do not coincide with the presence of a pause, respectively after daar gaan we nu een dak opzetten in (3) and dan heb ik nog een groen frotje over in (4):

(3) dan hebben we het zwarte vierkant (0.33)
then we have the black square

daar gaan we nu een dak opzetten (no pause)
on that we are going to put a roof now

dat is het groene driehoek (0.17)
that is the green triangle

(4) dan heb ik nog een klein groen frotje over (no pause)
then I have another small green thingie left

dat zal wel een bloempotje zijn (0.43)
that will probably be a flowerpot

dat zetten we bij het voorraam onder (7.18)
that we put by the front window below

The likelihood of occurrence of a clause-final pause appears to be very high at two important discourse locations. Firstly, pauses are present at all transitions between instructions, i.e. between all topical units. Secondly, pauses consistently occur right after the topic-introducing phrase or clause. These two locations constitute crucial information flow positions, as will be shown below.

4.2. Duration (table 4; see appendix)

The picture becomes even more interesting if one looks at the respective lengths of the pauses in various discourse locations, i.e. (i) in between instructions, (ii) after the clause or phrase introducing a new referent and (iii) at other positions. Results can be found in Table 3.

This table shows that, in these three monologues, pause duration is dependent on the topical structure of the discourse: the longest silence intervals are found in between instructions; within a topical unit, the pauses in post-referent-introducing position are consistently longer than in other locations. Though the three speakers differ considerably in their absolute pause lengths, they all share this same pattern of varying pause duration as a function of discourse location.

Having established that there is a strong correlation between pause structure and the topical organization of the discourse, it remains to be explained what cognitive or communicative factors might account for this regularity. In the following, a few tentative solutions will be presented that need further experimental verification. First of all, one can argue that pausal structure is a result of cognitive processing by the speaker. In this view, the silence intervals in between instructions could reflect the planning carried out by the speaker before s/he embarks on the next instruction. Similarly, the pause after the referent-introduction could be caused by the speaker's planning as regards how to develop the newly introduced topic in the subsequent discourse. The subsequent utterances within the same topical unit would then require less processing, and pauses are consequently shorter. However, since the experimental task seems to be a very simple one, which does not need considerable mental effort, the cognitive explanation does not appear to be very likely as the sole factor governing pause length.

An alternative hypothesis (which is not mutually exclusive with the previous one) concerns the communicative goals of the speaker, and his need to be cooperative towards the hearer. By manipulating pause length, one could argue, the speaker is trying to make it easier for his interlocutor to process discourse structure. In such a way, he is not only marking which major chunks of the discourse (the instructions) belong together, but is also drawing attention to the newly introduced referents by making them prosodically more salient, and hence more easily identifiable as new discourse topics. This outcome is very compatible with the finding that two out of three speakers generally provided the topic-introducing NPs with the highest Fo maxima (see section 2).

This latter view (and, especially, the discussion about the pauses in post-referent-introduction position) is also compatible with a third, more interactional explanation. Although the speech materials under investigation consist of monologues, the test setting really was a communicative one: speakers had to give instructions to hearers who were physically present. There is thus an interactive dimension to this discourse, despite the fact that no verbal or visual feedback was possible, as is the case in real conversational data (see Terken 1984 for the experimental design). In Geluykens (1991, 1992b), it is argued that referents in conversation are introduced in a collaborative manner, through a three-stage interaction between a speaker and a hearer. These stages are, respectively, introduction by the speaker of a new referent, acknowledgement by the hearer of this new referent, and establishment of the new referent by the speaker, by developing it as a discourse topic (see (5)).

(5) C: Prof. Worth asked me to get some books for him
B: oh yes yes
C: I've just arranged for those to be sent over by taxi
(simplified from Geluykens 1991)

Acknowledgement, it is claimed, can be either verbal, as in (5), (usually through a short acceptance signal such as yeah, mhm, and the like), or implicit, without an overt linguistic signal. In the latter case, the speaker pauses to give the hearer the opportunity to take in the new referent cognitively, but also to enable him to reject the new referent if s/he should feel that way inclined. Given the normal politeness principles operative in conversation (see Brown & Levinson, 1987), such rejection is not very likely. In the vast majority of the corpus-data analyzed in Geluykens (1991, 1992b), it was found that referent-introductions were followed by an overt acknowledgement signal, but also often by a pause.

The long post-referent-introduction pauses in our data could therefore be argued to reflect this interactive dimension, giving the hearer the chance to process the new referent, but also, theoretically, giving him the opportunity to intervene if necessary, either to request more information or to short-circuit the referent. In other words, some of these pauses would thus be quasi-conversational and

essentially interactive. As it is impossible to verify this unequivocally in the data, such a statement needs further experimental support.

5. GENERAL DISCUSSION AND CONCLUSION

This study has shown that speakers may enrich spontaneous discourse through prosodic structure in a variety of ways. It was found that the topical make-up of three monologues could be clarified by speech melody (use of various melodic boundary markers, the scaling of Fo maxima in accented words, the average Fo calculated over a clause) and by the variable duration of pauses. However, another important observation was that there was some speaker-variation, since not every speaker exploits each of the above prosodic structuring devices. Obviously, prosody is just one of a variety of means (lexical, syntactic, perhaps even non-linguistic, such as visual) to indicate the structure of the discourse. Therefore, as there could be some trade-off between these various mechanisms, some liberty in the use of melodic and temporal signals may be allowed without dramatic consequences for the understanding of the spoken text. Further experimentation is obviously necessary here.

This work has to be extended in two directions. First of all, it needs to be explored whether a speaker's prosodic structuring is also important from a listener's point of view. The assumed prosodic structuring devices can only be communicatively relevant if they are in some way meaningful to a listener. To gain more insight into this problem, we are currently conducting a series of perception experiments, both with filtered versions of the monologues studied here (Swerts, Geluykens & Terken, in press) and with utterances in which speech melody was systematically manipulated (Swerts, Bouwhuis & Collier 1992).

Secondly, the current paper has only been concerned with monologues. Therefore, it remains to be seen to what extent these findings can be extrapolated to discourse situations which are more interactional, in the sense that hearer-feedback is made possible. Both corpus-based (Geluykens 1992a, 1992b, in press) and experimental (Clark & Wilkes-Gibbs 1986) research suggests that there is a very outspoken collaborative dimension in the way new discourse topics are introduced in dialogue situations. It seems logical to assume that this collaborative dimension will also be reflected prosodically, in the way information is structured both on a melodic and a temporal level. It would be interesting to put this hypothesis to the test in conditions which are somewhat controlled and yet permit spontaneous, unplanned interaction. It is our intention (Geluykens & Swerts 1992) to further explore functions of prosody in (spontaneous) discourse, both from the point of view of interaction (in relation to the turn-taking mechanism) and of information flow regulation.

(*) Acknowledgments

The authors are also affiliated to the Belgian National Science Foundation (NFWO) and to the University of Antwerp (UIA and UFSIA, respectively). Thanks are due to Jacques Terken for allowing us to make use of the speech material employed in Terken (1984).

REFERENCES

Brown, G., K. Currie & J. Kenworthy (1980). *Questions of intonation*. London: Croom Helm.

Brown, P. & S.C. Levinson (1987). *Politeness: Some universals in language usage*. Cambridge: CUP.

Chafe, W.L. (1987). Cognitive constraints on information flow. In: R. Tomlin (ed.), *Coherence and grounding in discourse*. Amsterdam: Benjamins, 21-55.

Clark, H.H. & D. Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition* 22: 1-39.

Geluykens, R. (1988a). Five types of clefting in English discourse. *Linguistics* 26: 823-841.

Geluykens, R. (1988b). The interactional nature of referent-introduction. *CLS* 24 (1): 141-154.

Geluykens, R. (1989). Referent-tracking and cooperation in conversation: Evidence from repair. *CLS* 25 (2): 65-76.

Geluykens, R. (1991). Topic management in conversational discourse: The collaborative dimension. *CLS* 27 (1).

Geluykens, R. (1992a). *From discourse process to grammatical construction: On left-dislocation in English*. Amsterdam/Philadelphia: Benjamins.

Geluykens, R. (1992b). Topics in English conversation: On topic-introduction in conversational discourse. Ms.

Geluykens, R. (in press). *The pragmatics of discourse anaphora in English: Evidence from conversational repair*. Berlin: Mouton de Gruyter.

Geluykens, R. & M. Swerts (1992). Prosodic topic- and turn-finality cues and their transcription. *Paper presented at the Workshop on Prosody in Natural Speech Data*, University of Pennsylvania, August 1992.

Givón, T. (ed.) (1983). *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam/Philadelphia: Benjamins.

Hermes, D.J. (1986). Measurement of pitch by subharmonic summation. *Journal of the Acoustic Society of America* 83: 257-264.

Lehiste, I. (1975). The phonetic structure of paragraphs. In: A. Cohen & S. Nooteboom (eds.), *Structure and process in speech perception*. Berlin: Springer Verlag, 195-206.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: CUP.

Swerts, M., D. Bouwhuis & R. Collier (1992). End(s) of intonation: A perceptual study of melodic cues to finality. Ms.

Swerts, M. & R. Collier (in press). On the controlled elicitation of spontaneous speech. To appear in *Speech Communication*.

Swerts, M., R. Geluykens & J. Terken (in press). Prosodic correlates of discourse units in spontaneous speech. *To appear in Proceedings of the ICSLP*, Banff, Canada, October 1992.

Terken, J.M.B. (1984). The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech* 27: 269-289.

Thorsen, N. (1985). Intonation and text in Standard Danish. *Journal of the Acoustic Society of America* 27: 1205-1216.

Table 1. Distribution of low and not-low boundary tones as a function of discourse position

	SK		HZ		NE	
	End of Instr.		End of Instr.		End of Instr.	
	yes	no	yes	no	yes	no
Low	12	5	9	7	9	3
Not low	1	38	3	18	3	12

Table 3. Mean pause durations (in sec.) at various discourse locations

	SK	HZ	NE	Overall
inter-topic pauses	2.89	1.82	7.59	4.10
within topics				
- ref-intro pauses	2.21	1.03	2.38	1.87
- other pauses	0.92	0.70	0.67	0.76

Table 2. Successive Fo-maxima (expressed in Hz) within instructions: ☐ represent the value of the major accent in the topic-introducing phrase, **boldtype** represents the highest Fo-maximum within an instruction. The italicized numbers represent the 'exaggerated' measurements (further explanations in the text).

SK

Inst.	Successive Fo-maxima within instructions
1	303 [303] 250 238 *323*
2	[313] 256 238 256 263 256 213
3	286 [323] 303 278 233 256 286 270 263 213 238 204 263 222 217 204 222 200
4	[303] 250 303 256 294 250 244 244 294 263 233 227 263 294 217 227 233
5	270 [*400*] 238 286 250 233 244 278 227 222 244 *227* 270 238 250 227 270 192
6	[*400*] 233 204 233 250 213 *323* 233 196
7	244 [278] *270* 222 263 222 256 ***294*** 263 217 244 204
8	[278] 263 270 286 222 286 233 294 213 217 208 182
9	[286] 222 ***333*** 244 244 286 238 263 244 250 233 244 196
10	244 [345] 244 238 233 238 244 222 222 217 175
11	227 [357] 233 213 208
12	[*385*] *286* 238 244 263 233 357 256 233 233 217
13	[357] 233 244 222 233 208 213 185

NE

1	[370] 345 313
2	323 [370] 345 313 323 313
3	[333] 313 278 303 313 323 303 286
4	[313] 303 323 323 345 313 294 323 323 303 256
5	294 [303] 286 303 303 303 278 ***370*** 313 278 244
6	[313] 286 222 222 286 303 303 286 286 250
7	[286] 278 256 303 278 323 303 303 286 286 333 250
8	286 [294] 263 256 270 270
9	[270] 270 294 313 278 263
10	256 [263] 278 303 294 238

HZ

1	143 172 [217]
2	[192] 192 189 175 189 149 120
3	[196] 139 164 *256* 137 152 ***303*** 137 167 145 167 141 137 103
4	[222] 139 141 147 119 132 112
5	[179] 167 139 204 164 106
6	[189] 179 147 120
7	130 [152] 137 104 154 133 111 95 120 108
8	[182] 154 143
9	145 147 [159] 123 122 112
10	154 [175] 164 ***200*** 128 116 145 127
11	[167] 159 122 133 109
12	179 [200] 139 115 118

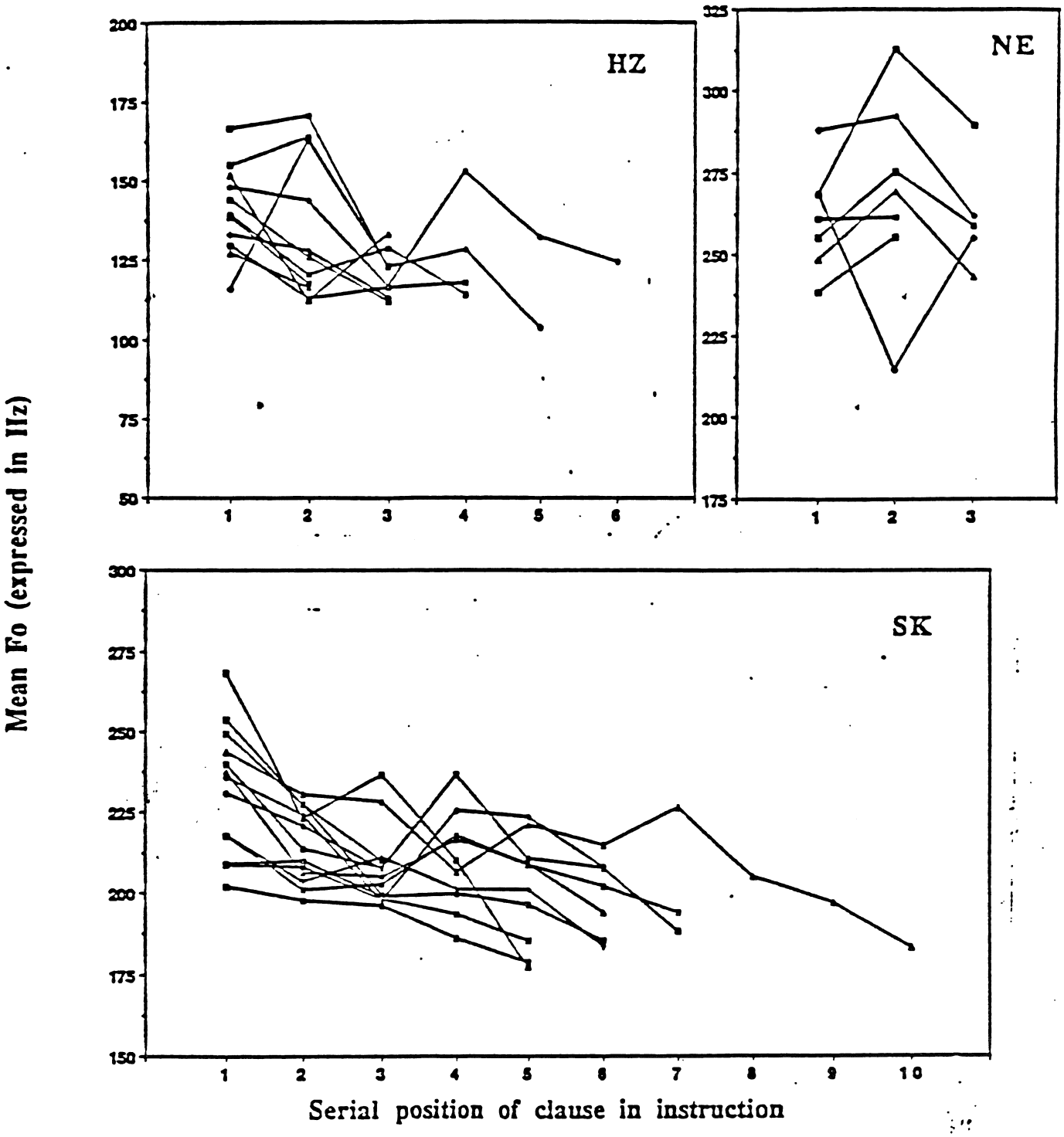


Figure 1. Mean Fo of successive clauses of the instructions. Identical symbols connected through lines represent the means within the same instruction

When Given Information is Accented: Repetition, Paraphrase and Inference in Dialogue

Marilyn A. Walker

Computer and Information Science
University of Pennsylvania
200 S. 33rd St.
Philadelphia, Pa. 19104

ABSTRACT

A classic function of intonation is to indicate the distribution of given and new information in an utterance. This paper defines given in two ways: known and salient. It then examines 63 utterances from a radio talk show corpus to determine whether either definition of given is predictive of the intonational contours found in the corpus. Given as salient is found to reliably predict one class of contour: the sustained tones.

1. Introduction

A classic function of intonation is to indicate the distribution of given and new information in an utterance[5, 7]. The pitch accent on new information indicates the information focus, the discourse entity about which a predication is being made, whereas given information typically occurs without a pitch accent. The traditional view consists of three basic claims: (1) Each phrase contains an item marked by the main pitch accent as the information focus; (2) The remainder of the phrase is given information, the ground; (3) There are a limited number of special cases in which given information may be accented.

Pitch accents may function to draw attention to or to increase the amount of processing devoted to the information focus[4]. A complementary viewpoint is that deaccenting plays a functional role as well; it indicates to the hearer that the deaccented item is currently salient in the discourse[20, 17]. The combination of these two factors allows the distribution of pitch accents to guide the hearer's processing.

Other researchers have claimed that given information may be accented in special cases such as when it is thematic, contrastive, or exclamative, as well as when the speaker echoes part of a previous utterance with surprise or incredulity or denies a presupposition in the previous utterance[3, 13, 18].

A less traditional view is that given information can occur with a pitch accent, but the type of pitch accent is qualitatively different than that on new information[15]. P&H claim that the complex bitonal accent L*+H marks information that is known but not currently salient, whereas the bitonal H+L accents mark the propositional content of an utterance as being inferable. The H+L* accent indicates that 'the desired instantiation of a salient open proposition is already among the mutual beliefs' of the conversants. The H*+L differs from H* in conveying that the hearer 'should locate an inference path supporting the predication'([15], sect 5.4).

In order to investigate some of these claims, this work develops independent logical criteria for classifying utterances as consisting of given or new information, and then examines whether in fact the intonational realization of given information corresponds with the predicted intonational patterns. This paper examines utterances that consist wholly of given information, e.g. repetitions of previous utterances. A definition of this class of utterances will be provided in section 2. I will call these INFORMATIONALLY REDUNDANT utterances, IRU's[22, 21].[1] Since the classical view is that each utterance has at least one item of new information, and since IRU's provide no new information, they potentially have an anomalous intonational realization.

The data consist of 63 IRU's from a corpus of naturally occurring dialogues, from a radio talk show for financial advice.[2] IRU's constitute about 12% of the utterances in this corpus. The instances of IRU's that have been analyzed intonationally demonstrate cases of pitch accents on given information that do not seem to fit the special cases described in previous work.

Section 2 describes the independent criteria used to classify utterances as consisting of given information, the types of prosodic realization found in the corpus, and a number of distributional parameters used to classify the

[1] These utterances are not however communicatively redundant, and yet they provide no new information.

[2] This corpus was initially transcribed by Hirschberg and Pollack from tapes of a live radio broadcast of a talk show called *Speaking of Your Money* on WCAU in Philadelphia[16]. I am grateful to Julia Hirschberg for generously providing me with the tapes of the original broadcast. Digitizing, pitch tracking, and transcription of the original broadcast was done with WAVES and additional programs generously supplied by Mark Liberman. There are some problems with this corpus, mainly being that there is some overlapping speech and the dialogues are taped in single track.

utterances in the corpus. The following sections examine particular subsets of the corpus defined by certain distributional properties, and finally section 7 proposes some issues for future research.

2. Informational Redundancy

The term INFORMATIONALLY REDUNDANT utterances (IRU's) describes utterances that consist wholly of given information. In what follows, it will be useful to have a term to refer to the utterance(s) that originally added the propositional content of the IRU to the discourse situation. This the IRU's ANTECEDENT.[3]

A definition of when an utterance counts as informationally redundant is given below[6]:[4]

Definition of Informational Redundancy
An utterance u_i is INFORMATIONALLY REDUNDANT in a discourse situation $\mathcal{S}$

1. if u_i has already been said in $\mathcal{S}$
2. if u_i expresses a proposition p_i, and another utterance u_j that entails p_i has already been said in $\mathcal{S}$
3. if u_i expresses a proposition p_i, and another utterance u_j that presupposes or implicates p_i has already been said in $\mathcal{S}$ either non-adjacent to u_i or by another speaker

Condition (1) of the definition means that saying an utterance in a discourse situation adds the propositional content of that utterance to the discourse situation. Condition (2) depends on identifying what is entailed from what is said; it relies on concepts such as paraphrase and logical inference.[5] For conditions (1) and (2), a diagnostic of whether the propositional content of an IRU is defeasible can be used to test whether the information is already available in the discourse situation[19]. This diagnostic cannot be used for cases defined by condition (3) since some of these inferences are defeasible.

Thus there are 4 logical types of IRU's defined by their relation to their antecedent. An IRU may be a: (1) repetition, (2) paraphrase, (3) entailment, or (4) non-logical inference from its antecedent(s). This defines given information based on purely semantic properties.

I will also examine the interaction of salience with the semantic definition of given as informationally redundant that is provided above. The term given has been used to mean both semantically given as well as 'in the hearer's consciousness' or salient[2, 17]. In fact, Brown argues that only when 'given' means 'salient' does it have relevance for intonational realization[1].

In the remainder of this section I will first describe the way that the IRU's in the corpus can be prosodically characterized, and then define a number of distributional parameters to use to determine whether it is possible to predict the different intonational realizations.

2.1. Intonational Description

I will use the system for intonational description proposed by Pierrehumbert[14], with two modifications. First, I will use the diacritic [ds] to indicate downstep[8], replacing the abstract L in the H*+L contour that was the trigger for downstep in Pierrehumbert's original system. Second, I will adopt the parameter of ':' for sustained tones, from McLemore[12]. This parameter indicates that a tone is sustained until the next tone.[6].

Most of the IRU's examined here, (48 of them), are roughly categorized into three intonational patterns, all of which end in falls; the difference between them is in the relationship between the two or more high pitch accents (H*) that each pattern contains. I will call these (1) sustained tones, e.g. H*: H* L L%, (2) downstepped H, e.g. H* H*[ds] LL%[8], and (3) upstepped H, e.g. H* H* L L%[11, 9]. Figure 1 shows a sustained tone. Figure 2 gives an example of downstep and Figure 3 gives an example of upstep.[7] I have limited the cases I examine here to IRU's that are realized with final falls or levels. Some have a downstepped phrase accent, or final Mid[9, 10].

For utterances that fit in these three main classes, there is often very little juncture between pitch accents in their realization. This means that the whole utterance seems to be treated as a unit since no single sub-part of the utterance is selected as focal. This is interesting due to the potential anomaly referred to earlier; theories that say that given information is de-accented predict that

[3] Actually I use the term antecedent to refer to both the prior utterance and the proposition realized by that prior utterance, but this should not cause any confusion.

[4] An utterance is defined as a clause, or a phrase in cases when there is no finite verb in an utterance.

[5] Other information is commonly included in the discourse situation such as that which is evoked by the physical situation or by common-sense or plausible inference[17]. However, I will only look at a subset of the 'available' information.

[6] Neither the system presented in Pierrehumbert's dissertation nor the recently proposed 'standard' transcription system seems adequate to transcribe this contour.

[7] The terms downstep and upstep are used to refer to precisely defined phenomena in African tone languages; here, I am using them simply as descriptive terms to refer to a relationship between adjacent H* tones.

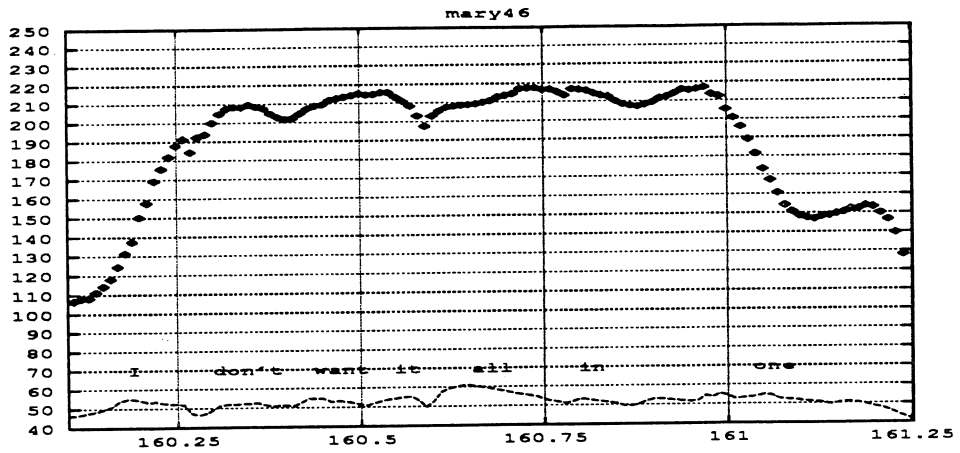


Figure 1: Mary 46. Salient paraphrase, Sustained Tone

the whole utterance would be de-accented, and yet this is in conflict with the assumption that at least one item in an utterance is always accented. Realizing the utterance with a sustained H* or with broad focus makes sense if every item has the same information status. In this case, the whole utterance consists of given information.

There are 7 IRU's in the corpus that are ambiguous between the three patterns described above because there is only one pitch accent in the utterance. Clearly one cannot distinguish sustaining a tone, downstepping from a tone, or stepping up to a tone when only one tone is realized. These will be called one-tone and will be discussed in more detail in section 4. Additionally, there are 15 tokens that do not fit into these three patterns and which I will briefly discuss in section 6. I should also note that there are cases in which it is difficult to distinguish a downstepped tone from a sustained tonal value; these are where the values of two adjacent tonal targets seem subject to a non-categorical kind of gradual decay, ie. there is very little difference between the two adjacent tones. I depended on the way the utterance sounds to make this distinction.

The following section discusses the distributional parameters used to classify the corpus and presents some initial distributional results. These results will then be discussed in the remainder of the paper.

2.2. Distributional Description

One of the main distributional parameters is the logical type of the IRU as defined above, whether it is related to its antecedent as a repetition, a paraphrase, an entailment or a non-logical inference. Of the 63 tokens of IRU's examined here, 13 are repetitions, 30 are paraphrases, 13 are entailments, and 7 are non-logical inferences.[8]

The second main distributional parameter is salience. An IRU may have an antecedent that is currently salient in the discourse context, i.e. just said by the other speaker or within the same turn of the current speaker. An IRU may also have an antecedent that is not currently salient. Its antecedent has been DISPLACED by an intervening change in topic[1]. Of the 63 tokens examined here, 42 have salient antecedents, and 21 have displaced antecedents.

The distribution of the corpus according to these parameters is presented below. Figure 4 shows the distribution of the three main contour types, presented in section 2.1, with respect to whether or not their antecedent is salient or displaced. Figure 4 also includes the 7 tokens that are called One-Tone, those with only one pitch accent and thus could fit in any of the sustained tone, downstep and upstep categories. It also includes the set of tokens classified as Other; these Others typically have an item realized with narrow focus somewhere in the middle of the phrase, or have an atypical syntactic structure such as topicalization. These will be discussed in more detail in section 6.

As figure 4 shows, salience is a predictor of sustained tones($\chi^2 = 5.600, p < 0.02$, for comparing salience as a predictor of sustained tones vs. downstep + upstep + other). Furthermore **all** the tokens that are difficult to classify because they only have one pitch accent, ie. the one-tone category, have a salient antecedent. Salience is a predictor of one-tone as well($p < 0.05$). The one-tone

[8]However, some examples of paraphrases seem closer to inferences based on axioms in lexical semantics.

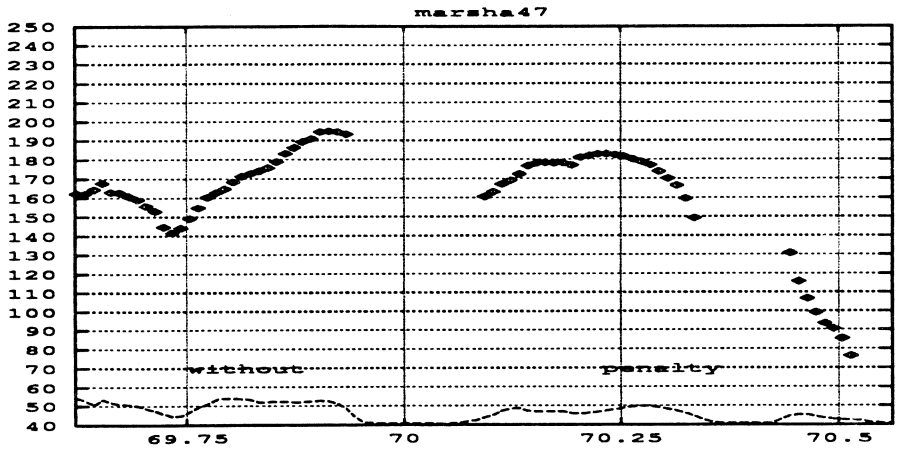


Figure 2: Marsha 26. Displaced paraphrase, Downstep

	SustT	DownS	UpS	One-T	Other
Salient	8	13	4	7	10
Displaced	0	12	4	0	5

Figure 4: Salience as a Predictor of Contour Distribution

tokens pattern like sustained tones in other respects as well; over half of them are repetitions. If the one-tone contours were classified as sustained tones, the relationship between salience and sustained tones would be even stronger ($p < 0.01$). Some one-tone contours will be examined in section 4.

Both the downstep and upstep contours are equally likely to have a salient antecedent as a displaced antecedent. In section 3, I will compare examples of downstep and sustained tones that occur in similar discourse situations.

	SustT	DownS	UpS	One-T	Other
Repeat	5	2	1	4	1
Paraphrase	2	18	5	0	5
Entailment	0	5	1	2	5
Non-Logical	1	0	1	1	4

Figure 5: Logical Type as a Predictor of Contour

Figure 5 examines the distribution of the various logical types of IRU's with respect to the contour categories. This figure shows that paraphrases are more likely to be realized with a series of downstepping tones ($\chi^2 = 9.877, p < 0.01$, as compared to the other logical types and other contour types).

Figure 5 also shows that repetitions are more likely to be sustained tones than any other logical type($p < .01$). However this could be due to the fact that repetitions are more likely to have a salient antecedent ($p < .02$). See figure 6.

	SustT	DownS	UpS	One-T	Other
Salient	5	2	0	4	1
Displaced	0	0	1	0	0

Figure 6: Repetitions: Salient vs. Displaced Antecedents

Finally, a comparison of IRU's inferable from their antecedents, ie. logical and non-logical inferences, with repetitions and paraphrases, shows that inferables are less likely to be realized with one of the three main patterns discussed($p < .01$).[9] See Figure .

Thus it seems that there is much more variability in the way the inferential IRU's are realized; they are neither realized consistently with the downstepping tones predicted by P&H nor with the stylized contours that were documented by Ladd and McLemore[15, 12, 8]. The following sections will discuss particular examples of the contours discussed here.

[9] A comparison of repetitions with all the other logical types is not significant ($p < .10$). However a comparison of repetitions with inferences alone is significant ($p < .05$), as is a comparison of paraphrases with inferences alone ($p < .05$).

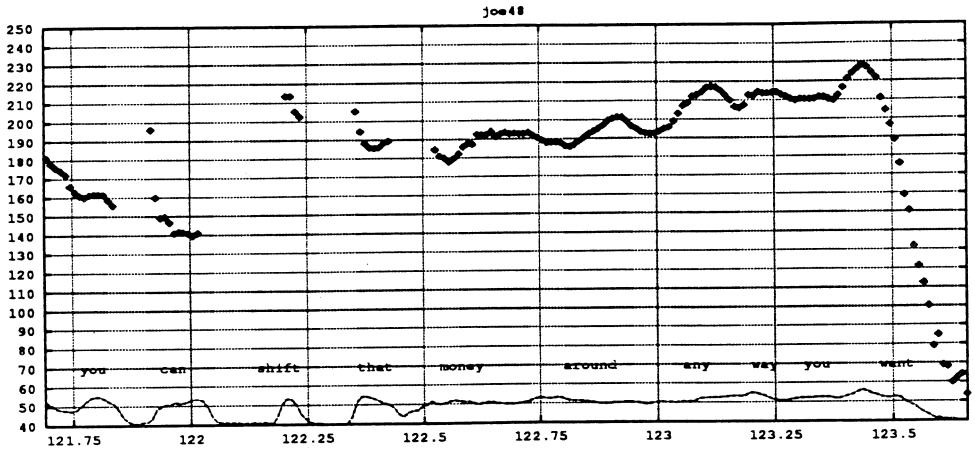


Figure 3: Joe 48. Salient paraphrase, Upstep

	Other	Not-Other
Repetitions and Paraphrases	6	38
Inferables	9	11

Figure 7: Inferables are more likely to be Other

3. Sustained Tones vs. Downstep

According to figure 4, the sustained tone contours are predicted by the salience of the antecedent. However why is it that there are so many downstepped contours with salient antecedents? The dialogue segment in 1, from (56) to (58) provides three examples of IRU's. In the dialogue excerpts given here, IRU's will be marked with CAPS whereas their antecedents will be given in *italics*.

(1) (52) h. and they will maintain their value approximately because they are variable rate funds
(53) m. I see
(54) h. Ok
(55) m. Fine
(56) h. ***and but separate it,***
I DON'T WANT IT ALL IN ONE
(57) m. TWO DIFFERENT ONES
(58) h. TWO DIFFERENT ONES, three would be even better....

The IRU in (56) is shown in figure 1. In 1-56, the speaker, (h), has paraphrased his own utterance from the just previous clause. The lexical item *separate* in (56) entails a division into at least two separate parts. As shown by plot of f0 in figure 1, this utterance is realized with a high sustained tone, followed by a mid-level final value (cf. [9, 10]), H*: H* H[ds] L%. This is an example of stylized intonation[8, 12]. Stylized intonation makes sense in these contexts since the information has just been said, it is certainly predictable. However the sense of predictability may be carried by sustained tones or a sequence of downsteps without necessarily depending on the final mid-level[15].

When we compare 1-56 with the paraphrase of it that Mary (m) produces in 1-57, we find that this utterance, in the same context is realized with a downstepping contour, rather than with the sustained tone. See figure 8. However there is a third example of an IRU in example 1-58, where Harry repeats *two different ones*. This is shown in figure 9. This utterance is counted as a sustained tone because of the difference between it and the downstep seen in figure 8. However the f0 for this utterance does go down slightly as it nears the end of the phrase. It is also realized with a phrase-final level since Harry intends to continue his turn[12, 15].

An almost identical context occurs in the following excerpt:

(2) (24) h. that is correct, it could be moved around so that each of you have 2000
(25) m. I
(26) h. ***without penalty***
(27) m. WITHOUT PENALTY
(28) h. right
(29) m. and the fact that I have a an account of my own ...

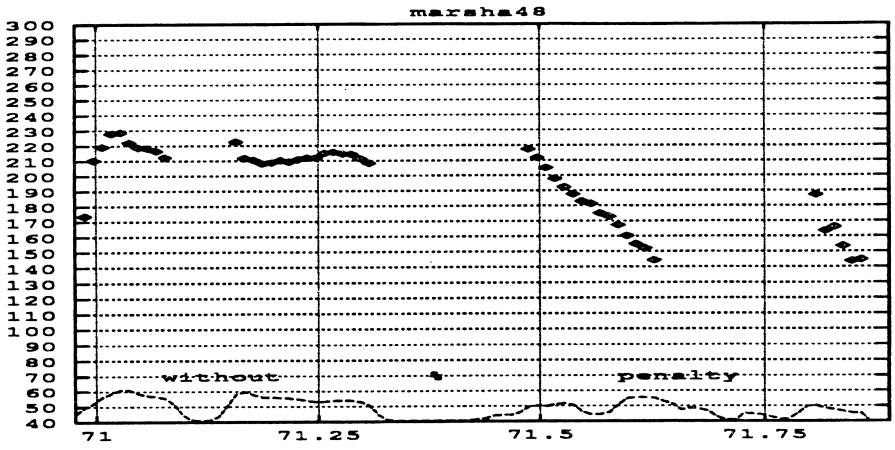


Figure 10: Marsha 27. Salient repetition, Sustained Tone

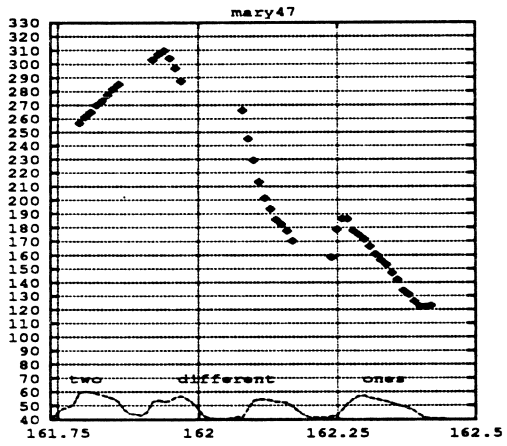


Figure 8: Mary 57. Salient paraphrase, Downstep

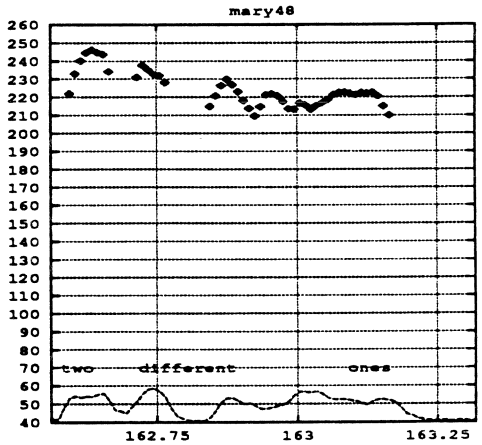


Figure 9: Mary 58. Salient repetition, Sustained Tone

The f0 for 2-(26) is shown in figure 2. This case differs from 1-56 in that even though (26) is an IRU, its antecedent is displaced, being some 10 plus utterances back in the dialogue. The salient repetition in 2-(27)(Figure 10), is realized as a sustained tone and is clearly distinct from downstepping f0 in (26). A similar example is given in 3-34 below. This is also a sustained tone, shown in figure 11:

(3) (33) h. well the amount that you have, the excess amount, the twenty eight hundred
(33.1)r. okay
(33.2)r. the amount that that was not your own contribution, *you rollover*
(34) r. YOU ROLLOVER
(35) h. right. but not your own contribution

In example 4, there is no lexical repetition but the information in (20) is a paraphrase of that in (18) and (19):

(4) (18) h. I see. *Are there any other children beside your wife*?
(19) d. *No*
(20) h. YOUR WIFE IS AN ONLY CHILD
(21) d. Right. And uh wants to give her some security .....

The corresponding contour, shown in figure 12 is classified as a sustained tone because all the main accents on *wife, only, child* are at the same f0 value.

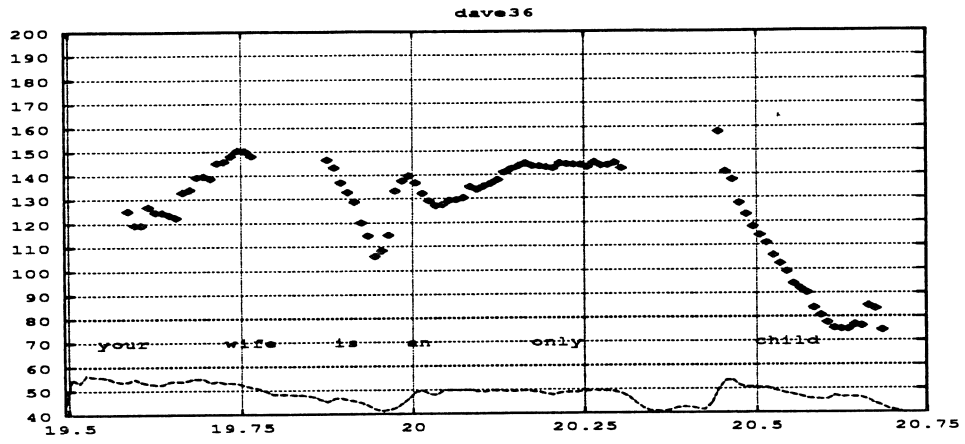


Figure 12: Dave 36. Salient paraphrase, Sustained Tone

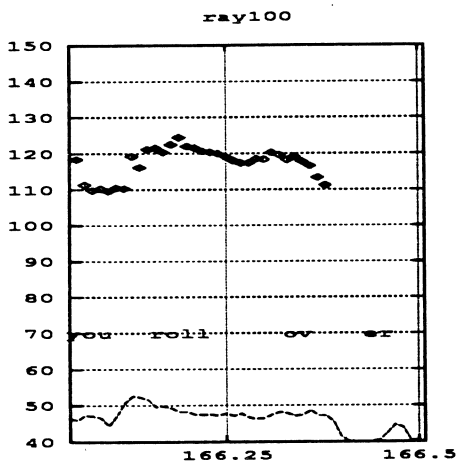


Figure 11: Ray 34: Salient repetition, Sustained Tone

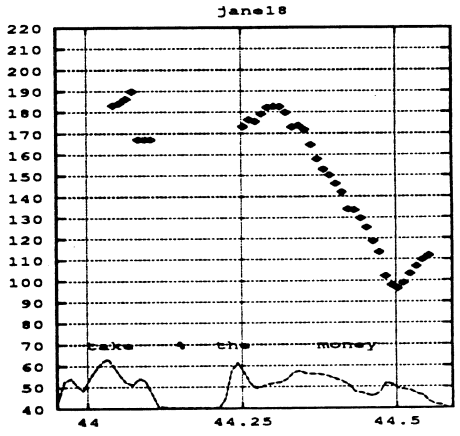


Figure 13: Jane 9: Salient repetition, Downstep

An example of downstep that is very similar to that given in figure 8 is the repetition in example 5-9. As shown in figure 13, the utterance in 4-9 is realized as a series of downstepped highs, with a pitch accent first on *take* and then *money* realized as an H*[ds].

(5) (8) h. you can stop right there: *take your money*
(9) j. TAKE THE MONEY
(10)h. absolutely.....

Note that while all the sustained tones have a salient antecedent, there are other IRU's that have a salient antecedent and yet are not realized with sustained tones. Thus I currently cannot predict when the sustained tones should occur, only when they should not.

4. One Tone Contours

As noted in section 2.2 there are some tokens which can't be classified as either a sustained tone, a downstep or an upstep because they only have one pitch accent. For example 6-8 in the excerpt below, and shown in figure 14.

(6) (6) r. Uh 2 tax questions. one: since April of 81 we have had an 85 year old mother living with us. Her only income has been social security plus approximately 3000 dollars from a certificate of deposit and I wonder what's the situation as as far as claiming her as a dependent or does that income from the certificate of deposit rule her out as a dependent?
(7) h. *Yes it does*
(8) r. IT DOES

(9) h. Yup that knocks her out.

These tend to be elliptical repetitions such as the one shown here. Figure 4 showed that these one-tone contours pattern distributionally like the sustained tones. However they cannot uncontroversially be collapsed with the sustained tones.

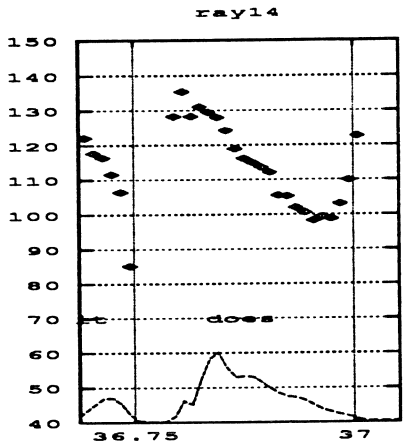


Figure 14: Ray 8, Salient repetition, One-Tone

A similar example is given below in 7 and shown in figure 15.

(7) (26) h. first of all with that forty one thousand and that's your pension alone
(27) m. yes
(28) h. completely taxable
(29) m. yes
(30) h. ok
(31) m. so we're in a
(32) h. *you're in a pretty healthy tax bracket*
(33) m. YES WE ARE
(34) h. as a result i'm not sure that I would want any of that hundred twenty thousand in any more treasury notes

Some of the contours classified as one-tone also share the phrase-final Mid with contours classified as sustained tone. For example consider the excerpt below and the corresponding f0 in figure 16.

(8) (22) b. Are there ah .. I don't think *the ah brokerage charge* will be ah that excessive
(23) h. No *they*'re not excessive but THERE ARE CHARGES

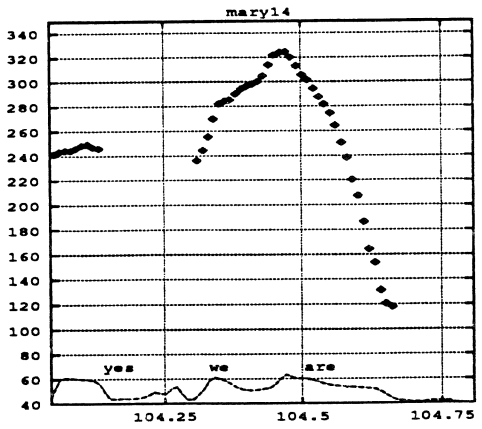


Figure 15: Mary 33, Salient repetition, One-Tone

Although this cannot be classified as a sustained tone since there is only one major pitch accent, it is also an example of 'stylized' intonation[9, 12, 8].

5. Upstepping Contours

One example of an upstepping contour was given in figure 3. Another example is given below in excerpt 9, and is shown in figure 17.

(9) (8) j. and uh i'd like to start out an I R A for myself and *my wife, she doesn't work*
(9) h. well how about last year?
....
(Intervening dialogue about eligibility for 81)
....
(17) h. ahh that then then you're not eligible for

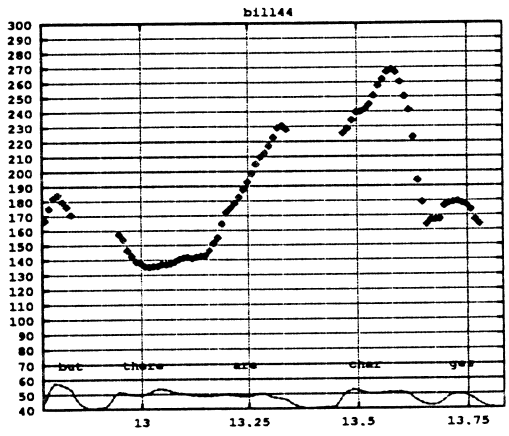


Figure 16: Bill 23, Salient Non-Logical, One-Tone with Mid

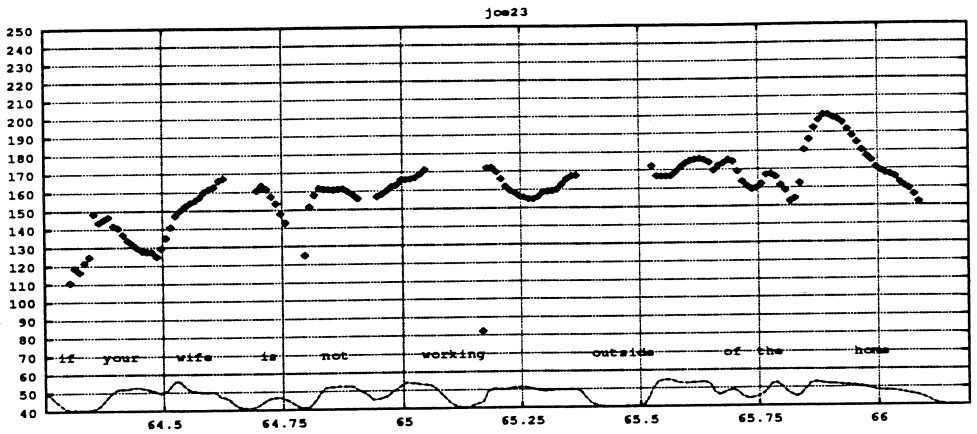


Figure 17: Joe 19: Displaced Paraphrase, Upstep

eighty one
(18) j. I see, but I am for eighty two
(19) h. You said it. You're eligible for twenty two fifty IF YOUR WIFE IS NOT WORKING OUTSIDE OF THE HOME

These upstepping contours sound as though the speaker is trying to be enthusiastic. The phrase-final tone is also Mid in this particular example. Additional tokens and further distributional analyses of this contour type would be necessary for formulating a more precise characterization.

6. Other Contours

There are 20 IRU's whose antecedents are inferable by logical inference or linguistically licensed inferences such as scalar implicatures or presuppositions(Bridge91,Hirschberg85). As discussed in section 2.2 that these are more likely to be realized as Other type contours. An example is shown in the excerpt below, where Harry (h) makes an entailment explicit from information provided in 10-7 by Jane (j).

(10) (7) j. .... and i'm entitled to a lump sum settlement which would be between 16,800 and 17,800 or a lesser life annuity. and *the choices of the annuity um would be $125.45 per month.* that would be the maximum with no beneficiaries
(8) h. you can stop right there: take your money
(9) j. take the money.
(10) h. absolutely. YOU'RE ONLY GETTING 1500 A YEAR.

Utterance 10-10 is shown in figure 18. This utterance shares the high final pitch accent with the upstepping contours and shares the note of enthusiasm with those contours.

7. Discussion

This paper has examined a number of cases where given information is not deaccented as in the classical view. I have examined the interaction of a semantic definition of given with discourse salience. Independent of whether given information is salient or displaced, given information is realized with a pitch accent.

The work presented here should be extended to actually test whether downstep is correlated with given information as P&H proposed[15]. The fact that 25 out of 63 IRU's are realized with a downstepping contour could be taken as weak support for their claims. However, I have not compared these tokens against a sample of non-redundant tokens to test whether downstep appears on these tokens just as frequently. Furthermore there are a number of other contours that these IRU's are realized with, such as the Sustained Tone, Upstep and Other contours that would not be predicted on P&H's account.

However, this paper has argued that salience is a predictor of one class of contour, the sustained tones. Other accounts have not distinguished salient and displaced mutual beliefs in terms of intonational realization[15] or have suggested that salient is the only relevant notion of given information[1]. I have shown here that, in this type of dialogue, the sustained tone contour is correlated with salience. However discourse salience is not predictive of downstepping contours. In addition, I have shown that IRU's that are inferable from their antecedents are more likely to be realized with some item in narrow focus than IRU's classified as repetitions or paraphrases.

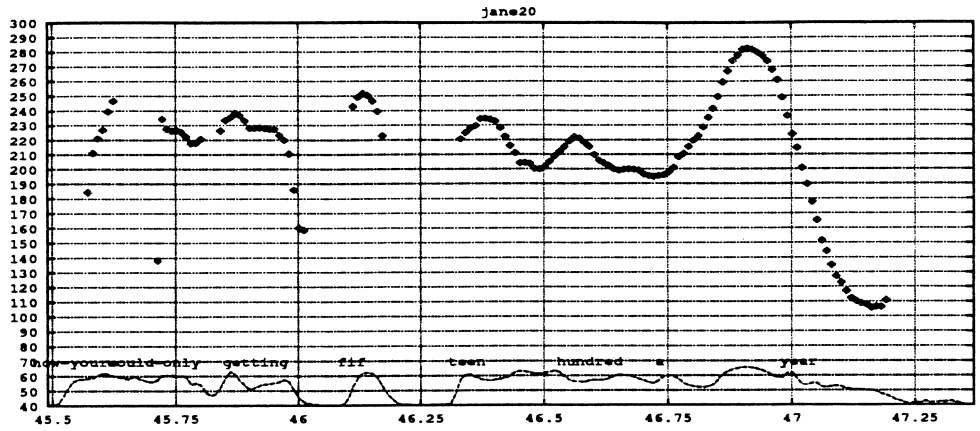


Figure 18: Jane 10: Displaced Inference, Other

I have not examined the influence of boundary tones on the pragmatic use of these contours[12]. Phrase-final Mid, characteristic of 'stylized' intonation[8], cuts across the contour classifications used here. IRU's frequently have these phrase-final Mid's, but not always. This must be examined more closely in future work. Future research should also include examination of these types of utterances in other types of dialogue in order to provide a more general account of the use of the contours described here.

References

1. Gillian Brown. Prosodic structure and the given/new distinction. In A. Cutler and D.R. Ladd, editors, *Prosody Models and Measurements*, pages 67–77. Springer-Verlag, 1983.
2. Wallace L. Chafe. Givenness, contrastiveness, definiteness, subjects, topics and points of view. In Charles N. Li, editor, *Subject and Topic*, pages 27–55. Academic Press, 1976.
3. Cruttenden. *Intonation*. Cambridge University Press, 1986.
4. Anne Cutler. Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, 20:55–60, 1976.
5. M. A. K. Halliday. Notes on contrastivity and theme ii. *Journal of Linguistics*, 3:199–244, 1967.
6. Julia Hirschberg. *A Theory of Scalar Implicature*. PhD thesis, University of Pennsylvania, Computer and Information Science, 1985.
7. Ray S. Jackendoff. *Semantic interpretation in generative grammar*. MIT Press, 1972.
8. Robert Ladd. *The Structure of Intonational Meaning: Evidence from English*. University of Indiana Press, 1980. A Cornell University Dissertation.
9. Mark Y. Liberman. *The Intonational System of English*. PhD thesis, MIT, 1975. (published by Garland Press, NY, 1979).
10. Mark Y. Liberman and Cynthia A. McLemore. The structure and intonation of business telephone openings. *The Penn Review of Linguistics*, 16:68–83, 1992.
11. Mark Y. Liberman and Ivan A. Sag. Prosodic form and discourse function. In *CLS 10*, pages 416–27, 1974.
12. Cynthia A. McLemore. *The pragmatic Interpretation of English Intonation: Sorority Speech*. PhD thesis, University of Texas, Austin, 1991.
13. S. G. Nooteboom and J. G. Kruyt. Accents, focus distribution, and the perceived distribution of given and new information: an experiment. Technical Report 538/V, IPO, Netherlands, 1987.
14. Janet Pierrehumbert. *The Phonetics and Phonology of English Intonation*. PhD thesis, MIT, 1980.
15. Janet Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In *Cohen, Morgan and Pollack, eds. Intentions in Communication, MIT Press, Cambridge, MA.*, 1990.
16. Martha Pollack, Julia Hirschberg, and Bonnie Webber. User participation in the reasoning process of expert systems. In *Proc. National Conference on Artificial Intelligence*, 1982.
17. Ellen F. Prince. Toward a taxonomy of given-new information. In *Radical Pragmatics*. Academic Press, 1981.
18. Susan F. Schmerling. *Aspects of English Sentence Stress*. PhD thesis, University of Texas, 1976. Published by UT Press: Austin.
19. Robert C. Stalnaker. Assertion. In Peter Cole, editor, *Syntax and Semantics, Volume 9: Pragmatics*, pages 315–332. Academic Press, 1978.
20. J. M. B. Terken and S. G. Nooteboom. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, pages 145–163, 1987.
21. Marilyn A. Walker. A model of redundant information in dialogue: the role of resource bounds. Technical Report MS-CIS-30-92, University of Pennsylvania Computer Science, 1992. Dissertation Proposal.
22. Marilyn A. Walker. Redundancy in collaborative dialogue. In *Proceedings of the fourteenth International Conference on Computational Linguistics*, 1992.

PROSODIC ELEMENTS AND PROSODIC STRUCTURES IN NATURAL DISCOURSE

Anthony C. Woodbury

Dept. of Linguistics
University of Texas at Austin
Austin, TX 78712

ABSTRACT

Although usually taken for granted, it is anything but clear that prosodic elements are organized into autonomous prosodic structures such as intonational phrases. A framework is outlined within which the structural and communicative organization of prosodic elements in samples of natural discourse might be discovered inductively. The framework assumes that the structural organization of a stretch of speech consists of the set of recurrent patterns it contains (including prosodic patterns), and that such patterns are recognizable to speakers. It is further hypothesized that in the normal or usual case, logically independent patterns (e.g., the placement of pauses vs. the placement of intonational cadences) will converge or unify; and that if they do not unify, speakers may draw special pragmatic inferences from this fact. Three samples of natural speech are analyzed in order to present the approach and demonstrate three key properties of the prosodic structure that it uncovers: (a) the potential independence of prosodic patterns and thematic structure; (b) the potential for bundles of prosodic elements to recur as prosodic 'macrostructures,' often associated by speakers with particular styles, contexts, and social personas; (c) the potential for prosodic patterns (and elements) to carry meaning that is iconic in character, but regulated by culturally specific conventions and practices.

0. PROSODIC ELEMENTS

When we speak of prosody, we are concerned with phonological and phonetic elements such as the following. Let us call them PROSODIC ELEMENTS:

- Pausing. (Including structural pauses after whole utterances, rhetorical pauses, micropauses, apparent hesitations and disfluencies.)
- Other durational modulations. (Including final lengthening, anacrusis, 'rhetorical lengthening,' local rhythmic and arrhythmic patterning.)
- Stress and related features.
- Pitch targets. (H,L, possibly rises and falls.)
- Pitch alignments (to prominent syllables, word edges, and other sites, cf. Pierrehumbert and Beckman 1988).
- Pitch scaling. (Including initialization, initial raising, final lowering, downstep, and catathesis (Hirschberg & Pierrehumbert 1986).)
- Segmental sandhi processes. (E.g., American English Flapping, French Liaison (Nespor & Vogel 1986).)
- Prosodic reshaping. (Including 'Rhythm rules' (Liberman and Prince 1977), phrasal truncation (Sapir 1949), refooting rules (Woodbury 1987b, 1992).)
- Voice quality modifications. (Including falsetto, breathy and creaky voice, pharyngealization, vibrato (Miller 1992), phrase-final devoicing (Michelson 1991).)
- Others?

Perhaps the three leading questions about prosodic elements are:

- THE IDENTITY QUESTION. How are they to be identified and described, and how are they perceived by humans?
- THE DISTRIBUTION QUESTION. How are they distributed in discourse, and what cognitive faculties govern those distributions?
- THE MEANING QUESTION. What do they mean, or do, in discourse?

1. THE PROSODIC HIERARCHY

An answer to the Distribution Question is offered by PROSODIC HIERARCHY THEORY (Selkirk 1980, Nespor & Vogel 1986, Hayes 1989, Inkelas and Zec 1990), and to some extent by any approach assigning autonomy to

prosodic phrases, intonational phrases, breath groups, and the like (Halliday 1967, Beckman and Pierrehumbert 1988, Chafe 1980). It also gives, implicitly, a partial answer to the Meaning Question.

Prosodic Hierarchy Theory claims that prosodic elements refer for their distribution to a hierarchy of abstract, autonomous, discrete units like that in Figure 1. Furthermore, it claims that these abstract units—rather than individual prosodic elements—map in certain ways to syntax, pragmatics, and thematic discourse structure. Thus it answers the Distribution Question by asserting massive coordination among a variety of aspects of grammar and speech. And it answers the Meaning Question with the implicit assertion that abstract phrase breaks of the hierarchy, rather than individual prosodic elements such as pauses, intonational cadences, and the like, will bearers of (pragmatic) meaning.

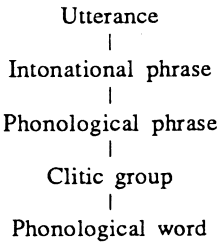


Figure 1: The Prosodic Hierarchy according to Hayes 1989.

2. CRITIQUE

Two colleagues and I have criticized this approach on empirical grounds (Liberman, McLemore, & Woodbury 1991). Our critique has several main points.

First, (most) effects of prosodic constituency are gradient with respect to junctural strength, and thus do not provide any clear evidence for qualitative constituent types.

Second, the hypothesized prosodic structure is extremely ambiguous in practice, so that determinate, intersubjectively valid descriptions are not generally possible if the hypothesis of qualitatively distinct prosodic levels is maintained.

Third, even when a given prosodic effect is nongradient with respect to junctural strength (i.e., either present, or absent), it is still generally difficult to assign it successfully to just one constituent type (Woodbury 1992).

Fourth, (most) clear "prosodic" constituents seem to correspond to independently-needed units, generally from the domain of information structure.

And fifth, (some of the) key phonetic correlates of prosodic structure can be seen as natural solutions to the problems of presenting the (non-prosodic) structure of messages and of managing communicative interaction (McLemore 1991).

In the face of these doubts, we considered it reasonable to approach prosody anew with a NULL HYPOTHESIS, according to which the distributions of prosodic element refer directly to functions and structures that are outside of prosody and that are independently known to be part of discourse, including syntax, pragmatics, and thematic structure.

In principle, this null hypothesis implies no coordination at all among prosodic elements—let alone the massive coordination implied by the Prosodic Hierarchy—since it expects only that each prosodic element will bear some relationship to something elsewhere in discourse. Nevertheless, it is crucial to determine whether any such coordination obtains, even if less rigid or necessary than that projected by the Prosodic Hierarchy; that is, whether prosodic elements can assemble themselves into PROSODIC STRUCTURES of any kind. The question is no less one of meaning (or communicative function) than it is of distribution, since if such structures do exist, then on any theory of prosody at all they too should be expected to relate to syntax, pragmatics, or some other aspect of discourse.

The problem of prosodic structure is my focus in this paper. Section 3 argues that this structure should be sought inductively through observation of natural discourse. Section 4 outlines a framework for finding prosodic (and other discourse) structure which builds on work in poetics. The last three sections raise further issues while applying the framework to three discourse samples.

3. OBSERVING NATURAL DISCOURSE PROSODY

We know so little about the distribution and meaning of prosodic elements that we must first observe and describe them in natural discourse. This is not to say that experimentation, modeling, and even introspective study do not have their place. But they work best when they rest on an idea of the variety and diversity of speech prosody.

The term NATURAL DISCOURSE needs elaboration, for investigators seem to use it in at least two quite distinct senses. On the one hand, it is used to refer to any extemporaneous—not scripted—speech. That is perhaps the grammarian's sense, since the speaker is generating the forms using his/her own grammar. On the other hand, the term is also used to designate speech that is real—not simulated—social action. And perhaps that is

the anthropologist's sense of it, since anthropologists have often observed that when people simulate behavior, they do so with reference to stereotypes or IDEOLOGIES of social action, rather than the tacit models they rely on when performing or responding to social action in real life (see Silverstein 1979 for a careful review). As a result, simulated behavior is often recognizably different. To take just one example, consider the skill an actor must have in order to perform scripted dialog believably and effectively: such special skill would be unnecessary if there were no inherent gulf between real and simulated social action.

I would suggest that by following the anthropologists' lead in connecting natural discourse to social action, we can best appreciate its diversity. From that perspective, it is not enough to sample natural discourse simply by turning on the radio, for it spans every possible facet of social life. As workers in the ethnography of speaking and sociolinguistic pragmatics have emphasized (see Hymes 1974, Bauman & Sherzer 1974, Gumperz 1982, Levinson 1983, and the journals *Language in Society* and the *Journal of Linguistic Anthropology*), it encompasses narrative, conversation, and oratory. It includes ceremonial, ritual, formal, and institutional speech, in societies both with, and without, highly diversified institutional structures. It includes magical and religious speech. It includes prose, poetry, chanting, and singing. It includes speech with different purposes, from exhortation, to instruction, to description, to elicitation. And it includes scripted speech, whether read or recited from memory, along with extemporaneous speech.

But is it really necessary to sample so many kinds of speech? Is the prosody of a language not more or less uniform, regardless of the use to which it is put? The inductive perspective advocated here lets us see for ourselves. Even the small amount of description I have done convinces me that unlike syntax, morphology, and lexical phonology, prosody seems to vary not a little, but fundamentally, across genres, varieties, uses, and the like, even within a single language. If this is so, then it certainly is worth it to pursue diversity and to generalize only cautiously about the intonational systems of entire languages.

4. FRAMEWORK

To make consistent, useful observations, it is necessary to do so within an explicit framework making at least some basic theoretical assumptions. Such a framework should allow:

- 'Thick' description of the form, distribution, and meaning of prosody in individual texts
- Comparability across descriptions of different kinds of natural discourse, to allow for appropriate inductive generalization.

If our critique of it is justified, Prosodic Hierarchy theory is not such a framework: it is less than ideal for thick description since it may deflect attention from those aspects of prosody not crucially relevant to the hierarchy; and by focusing on the abstract units rather than individual prosodic elements, it may at times overstate some similarities across descriptions while missing others. Its problem in short is that it checks for a certain kind of all-encompassing order and coordination among prosodic elements, rather than gauging structure in whatever shape or form it may take.

4.1. Jakobson's Poetics

How then is structure to be gauged? Let us begin with a particularly useful notion of discourse structure from poetics. Despite its poetic origins, it can be extended beyond what we may wish to designate as 'poetry,' or value as verbal art. The basic idea, due to Roman Jakobson, is that there is a poetics to ALL discourse; and it is a fundamental key to discourse understanding. He gives his idea the following quite pungent formulation (Jakobson 1960:358):

> The poetic function projects the principle of equivalence from the axis of selection into the axis of combination. Equivalence is promoted to the constitutive device of the sequence. In poetry one syllable is equalized with any other syllable of the same sequence; word stress is assumed to equal word stress, as unstress equals unstress; prosodic long is matched with long, and short with short; word boundary equals word boundary, no boundary equals no boundary; syntactic pause equals syntactic pause, no pause equals no pause. Syllables are converted into units of measure and so are morae or stresses.

Essentially Jakobson is proposing a principle of recurrence, and claiming that it creates or reinforces structural equations (which then invite inference about content, given speakers' expectation that form can diagram content). The strongest instance of it is simple repetition. Illustrating from written poetry:

> *and miles to go before I sleep;*
> *and miles to go before I sleep.*

A weaker instance is in parallelism, a form of partial recurrence:

He called for his pipe
and he called for his bowl
and he called for his fiddlers three.

The recurrent elements or units may be of any kind: phonological, syntactic, lexical, morphological, thematic, and so on. All the examples above involved syntactic units. But phonological units recur in rime, alliteration, and meter. And morphological units recur in grammatical parallelism.

Recurrence establishes PATTERNS of various kinds. The most elementary is simple alternation (as, e.g., in trochaic meter); a very elaborate pattern is that of the Shakespearean sonnet, shown in Figure 2, where the pattern is global (it pervades the whole poem), it involves a hierarchy of fixed depth (i.e., a fixed number of qualitatively distinct levels), and it has counted parts (e.g., five feet to a line, two lines to a couplet). Obviously, different kinds of discourse are likely to show different degrees of pattern elaboration.

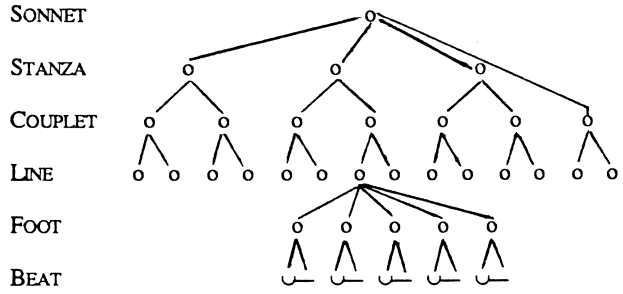


Figure 2: Shakespearean sonnet form.

4.2. The framework

Jakobson's principle can form the basis of a heuristic framework or method for determining how, and how much, a given instance of natural discourse is structured, and as part of that, the extent to which the prosodic elements within it are structured.[1] At minimum, any discourse instance should have some kind of thematic recurrence; a sequence of sentences (even if not parallelistic in any internal respects); regularly recurrent pitch cadences; and an alternation between pausing and silence. Since these different patterns need not all be congruent, our framework is most useful if it assumes at the outset that they form distinct parts the of overall discourse organization:

> ***Assumption 1.*** Natural discourse is organized as a set of components, where a component is any well-defined patterning of recurrent elements that is present in a stretch of discourse and distinct from other such patternings.

If this is to be of any interest, we must also assume:

> ***Assumption 2.*** Speakers can recognize well-defined patterning of recurrent elements.

If true, this may hold by virtue of a general human capacity for pattern recognition, rather than any specifically linguistic faculty.

The following then is a likely minimal set of components:

- Thematic patterning
- Syntactic patterning
- Pitch patterning (pitch accents, cadences, scaling)
- Pause patterning

Of course, a sample of discourse would require more if other elements were also at work creating recurrent patterns. For example, if distinct prosodic elements establish distinct patterns, then these patterns must count as separate components. The notion of component, like the notion of recurrence, is entirely empirical and heuristic.

The above assumptions may approach the minimum needed for gauging structure in discourse. But the following theoretical claims might be added as assumptions for heuristic purposes:

> ***Assumption 3.*** In the usual or default case, elements and structures of different components should converge or unify, standing in one to one correspondence, in the simplest cases (perhaps on general iconic grounds, rather than by virtue of any specifically linguistic faculty).

> ***Assumption 4.*** Deviations from such convergence may be salient to speakers and may lead them to draw special pragmatic inferences.

These assumptions do not exclude any distributional configurations: rather they predict (rightly or wrongly) how certain distributional configurations will be processed and interpreted. For example, in many instances of discourse the general pattern is for sentence breaks, intonational cadences, and pauses to coincide regularly. This would constitute a set of defaults among three logically separate components, *i.e.*, syntactic, pitch, and pause patterning. Departures from this convergence—enjambment of sentences by pause suppression, rhetorical pauses and pitch falls within sentences, and the like—would then stand out as special

[1] Hymes 1981 and Silverstein 1984 are some other explicit efforts to extend Jakobson's poetics to the analysis of natural discourse of various kinds.

and invite certain special interpretations. On these assumptions, structuredness is a matter of degree that must be gauged, rather than only a quality that must be specified. Furthermore—at least by hypothesis—it serves as a norm in terms of which potentially significant departures are measured, rather than simply a defining characteristic of well-formed speech (as in grammar).

In the following sections, all four assumptions are used to track prosodic elements, gauge prosodic structure, and evaluate their communicative roles, in three samples of natural speech. Each analysis is focused on a key property or characteristic of the structure being sought. The first and third samples come from myth performances by Central Alaskan Yupik Eskimo elders; the second is a pair of routines from an American television comedian.

5. DOES PROSODY REFLECT THEMATIC PATTERNING?

According to a widely held view, prosody is there to reinforce the preexisting thematic or content patterning of speech. In terms of the framework discussed above, prosodic elements and structures would then cue the boundaries or the internal 'high points' of such units. But workers in sociolinguistic pragmatics (e.g., Silverstein 1976:33-35, Gumperz 1982:100ff) have argued that pragmatic markers, including prosodic cues, need not only reflect or reinforce preexisting elements of context—they can affect, shape, and create participants' constructions of context. Accordingly, a subtly different view of prosody and thematic patterning might hold that although prosodic elements may reinforce independently recognizable thematic units, they may also at times be used to propose or create novel constructions of thematic patterning that are not independently inferable, or that alloy or conflate thematic patterning with other considerations, or that are present only as one of many possible 'takes' on thematic patterning.

In connection with this last point, it is important to make clear just how varied thematic patterning can be. In narrative, it can involve patterns of character foregrounding, or of tense/aspect shifts, or of scene changes, or of parallelistic, recurrent episodes. In conversation (Levinson 1983), it can be based on adjacency pairs (like question and answer, offer and acceptance, request and denial), or conversational activity types (greetings, leave-takings, 'pre-sequences,' etc.), and can be highly ritualized, as in verbal dueling (Labov 1972). In oratory, it can involve parallelistic figures that frame a rhetorical progression or transformation; or by the speaker taking the parts of alternating participants in a simulated argument or conversation. In ceremonial speech, it can involve distinct sections corresponding to different stages of a ceremony, or the progression of special speech act types, or the alternation of fixed texts and impromptu speech. At very least, this range and variety should caution that empirical result gotten for one kind of thematic patterning may well fail to predict results for other kinds of patterning; and it should indicate that for any one given text, there are many ways to conceive of thematic patterning. (In the terms of Assumption 1 above, a given discourse could have several orthogonal thematic (sub)components).

Figure 3 presents the opening of a myth performance in Central Alaskan Yupik Eskimo (CAY). Each line of the Figure shows a single word (or, when apparently sharing a single intonational contour, two words); its intonational profile, consisting of the initial pitch (if preceded by a pause), which is generally a low point; the pitch peak, which generally occurs on the first or second stressed syllable (boldface); and the pitch at the end, which generally is another low point. Following that is an indication of following pause length (0.0 if there is no pause), and then an English translation.

The myth is particularly interesting because it has a highly elaborate pattern of counted-out thematic parallelism among episodes. In the story, an orphan grandson paddles upriver in his kayak, meeting five successively more fearsome creatures and then taking them into his magical power. He returns home to his grandmother, denying having done anything special when she asks him. Then he goes into the communal men's house where he is harassed by bored, cruel shamans who want him to try his hand at conjuring. Under duress, he gives in and conjures, in turn, each of the five animals he had charmed, eliciting a successively greater reaction among his tormentors. Specifically, the whole myth divides into five Parts (Figure 4a). Of these, Parts II and IV then divide further into five Episodes, corresponding to the animals the Grandson charms (Figure 4b). Finally, each Episode divides into three Divisions according to the logic of the action (Figure 4c).[2]

Obviously, not all myths have such robust and formally elaborate patterns of thematic recurrence. Therefore, this myth presents a special opportunity to ascertain thematic patterning independently and gauge its relationship to prosody.

As it turns out, two prosodic elements, initial and final low pitch, mark off major units (e.g., Episodes in II, Divisions in IV). These units either begin with lower pitch, or end with lower pitch, or both. Thus in Figure 3, Parts I and II both begin below 80 Hz, while Part II ends somewhere below 90. (Sporadically, both of these elements also occur elsewhere.) They clearly reinforce independently recognizable thematic patterning. Even so, they still do not do this in the strictest possible way: for

[2]The terms Part, Episode, and Division are descriptively useful but theoretically arbitrary.

in Part II they mark Episodes, while in Part IV they mark Divisions. The explanation for this is that Episodes in II, and Divisions in IV, are all units of roughly the same length. Thus the distribution of low initial and final pitch is also partly rhythmic, or time dependent. By reflecting a particular mix of thematic and rhythmic factors, initial and final low pitch show a modicum of autonomy. Moreover, because they do so together, they operate not only as individual prosodic elements, but as prosodic STRUCTURES, and constitute, in the terms of Assumption 1 above, a COMPONENT in this stretch.

More sharply autonomous is a pattern involving the pitch peaks. Observe that within each sentence, the heights of successive pitch peaks generally increase, giving a culminative effect within the sentence. Moreover, peak heights also build from the beginning of the story to the end of Division A in Episode i, where 392 Hz is reached with the help of an astonishing narrative falsetto Mezak used in this and in many other of his performances. Thematically, the building pitch seems to mark the development of the action up to the point where the boy must perform masterfully. After

CAY		Init L	Peak	Final L	Pause	English
[I. INTRODUCTION]						
	Nunat ^ukut	77	116	86	<0.0>	There was a village [of people]
	uitaura'rqelriit		140	87	<0.0>	who lived
	kuigem ^ceniini.		150	84	<2.8>	on a riverbank.
	Tutgara'urluqelriigneg	104	140	97	<0.0>	A grandchild and grandparent
	ilaluteng.		152	95	<0.0>	were among them.
	Tutgara'urlurlua ^im',		146	113	<1.5>	And that grandchild,
	tan'ga'urlull'rauluni;	82	162	112	<0.0>	was a boy;
	Angutnguluni=w'.		150	<116	<2.5>	It was a male.
[II. GRANDSON GOES UPRIVER]						
	Tua=i=ll'=am	78	156	128	<0.0>	Well
	caqerluni,		151	120	<2.3>	one time
	kuigkun ^e:--	142	227	186	<0.0>	*on the river--*
	kuimegteggun=am		294	222	<0.0>	*on their river*
	asgurtuq..		333	256	<1.5>	*he [paddled] upstream.*
[i. ENCOUNTERS PTARMIGANS]						
A	Tua::=ih,	136	136	93	<1.4>	Well
	asgurturaqerluni,	109	184	136	<3.1>	as he went upstream
	qang:qiiregnek,	142	285	136	<1.7>	a pair of ptarmigan
	callul:riignek	357	392	368	<0.0>	*who were fighting*
	tekituq.		363	—	<1.8>	*he encountered.*
B	Tua=i=ll'	285	285	123	<0.0>	Well,
	ketairamikek		270	125	<0.0>	when he passed alongside them
	cenami	[83]	226	117	<0.0>	on shore
	piagnek:		133	80	<1.8>	he said to them:
	"Aah!	140	140	113	<1.7>	"Hey!
	Tua:=i!	113	156	123	<2.4>	Enough!
	Pisqekumtek ^taugaam	128	217	120	<0.0>	But when I tell you
	piniartutek!		262	127?	<1.8>	you shall [carry on some more]!"
	Aa tua=i	151	204	180	<0.0>	Hey enough now
	uter— utertek!"		214	192?	<1.6>	go— go home!"
C	Aren im*kug=am*	123	263	222	<0.0>	My and sure enough the two
	*qang*qiirek		294	144	<0.0>	the ptarmigans
	niilluteg		184	125	<0.0>	they obeyed
	ayagtuk.		126	<99	<1.7>	and went away.

Figure 3: Opening of a Central Alaskan Yupik Eskimo myth performance by Evon Mezak of Nunapitchuk, Alaska, recorded in about 1972. The text, and a detailed analysis of it, appear in Woodbury 1987a.

Myth
- I. Introduction
- II. Grandson goes upriver
- III. Grandson goes home
- IV. Grandson goes to men's house
- V. Closing

Figure 4a: Expansion of the Myth into five Parts.

Parts II and IV
- i. Grandson meets/calls ptarmigans
- ii. Grandson meets/calls dunlins
- iii. Grandson meets/calls cranes
- iv. Grandson meets/calls caribou
- v. Grandson meets/calls wolves

Figure 4b: Expansion of Parts II and IV into five Episodes.

Episode
- A. Grandson comes, and is challenged
- B. Grandson reacts masterfully (in quoted speech or song)
- C. Response to Grandson's reaction

Figure 4c: Expansion of Episodes into three Divisions.

this point there is more or less a denouement. The same building pattern then occurs again in following episodes.

On Assumption 1, these pitch fluctuations represent a unique pattern and hence call for a separate component. To be sure, the component correlates with an aspect of thematic patterning, but by picking out a climax in narrative development, it does so in a wholly different way than the low initial and final pitches. Indeed, it can even be seen as a device by which the narrator *proposes* to his hearers an interpretation of narrative development in this section.

A final interesting pattern is presented by pausing. As in much Native American discourse, the pauses are long and come at regular intervals, adding salience to pausing as a poetic feature (on Assumption 2, but also on the grounds of experience: see Tedlock 1983). This is best seen if we consider a reformatted version of the translation, Figure 5, in which line-breaks correspond to pauses. The default pattern for the whole text (Assumption 3) is for pauses to occur at sentence breaks, and, occasionally, for somewhat shorter pauses to occur at points in between. (Scrupulous observance of this default is one way that boring-sounding prose can be achieved in CAY!) Here however, the default is upset twice in the second line, where 'And the grandchild' is enjambed with the preceding sentence with no intervening pause, while simultaneously set off by a pause from it's sequel, 'was a boy'. Thematically

English	Pause
[I. INTRODUCTION]	
There was a village [of people] who lived on a riverbank.	<2.8>
A grandchild and grandparent were among them. And the grandchild	<1.5>
was a boy; He was a male.	<2.5>
[II. GRANDSON GOES UPRIVER]	
Well one time	<2.3>
on the river--on their river he [paddled] upstream.	<1.5>
[i. ENCOUNTERS PTARMIGANS]	
A Well	<1.4>
as he went upstream	<3.1>
a pair of ptarmigan	<1.7>
who were fighting he encountered.	<1.8>
B Well, when he passed alongside them on shore he said to them:	<1.8>
"Hey!	<1.7>
Enough!	<2.4>
But when I tell you you shall [carry on some more]!"	<1.8>
Hey enough now go— go home!	<1.6>
C My and sure enough the two ptarmigans they obeyed and went away.	<1.7>

Figure 5: English translation from Figure 3, reformatted so that line-breaks correspond to pauses.

speaking, the pause here marks the division between old and new information, adjoining 'And the grandchild' to the sentence that first introduced the grandchild.

Likewise, the default is upset when very long pauses occur in mid sentence, as in the first and fourth lines of Part II in Figure 5. Very long pauses are unexpected, and hence (on Assumption 4) invite special interpretation. As in all languages/speech communities with which I am familiar the unexpected delay at this point heightens narrative suspense. This effect presumably follows from basic, essentially non-linguistic strategies that all people have for dealing with expectations that fail (momentarily) to materialize.

In this section we have seen where prosodic elements reflect thematic patterning but still alter it slightly on rhythmic grounds; where prosodic elements not only reflect thematic patterning, but propose an interpretation of it; and where distinct prosodic elements are coordinated loosely, in terms of a default. Several conclusions may be drawn. First, even when thematic patterning is independently recognizable, prosodic elements need not simply reflect it. Therefore, it is not safe to assume—as many investigators seem to do—that prosody will provide a perfect diagram of some fixed (abstract) thematic structure of speech in cases where thematic patterns are not patent or overt in any other way. Second, there are interesting distributional and functional relationships not only between individual prosodic elements and thematic structure, but among prosodic elements. That is, prosody can still be said to have structure, albeit of a far more diffuse and complex type than has generally been assumed.

6. PROSODIC MACROSTRUCTURES

Although individual languages and individual speakers have broad prosodic resources, it is striking how few are actually used in many natural instances of speaking. This is illustrated by two short routines by Jay Leno, a television comic (Figure 6).

Thematically, each routine has two major parts: one where the comedian recites something heard or seen elsewhere, taking on the voice of the source; and then one where he parries in his own voice with the punch line (and then feigns nonchalance when the applause comes). The transcript is broken into lines representing putative intonational phrases, i.e., domains implicated by prominent final pitch cadences and (usually) final rallentando or lengthening. Shown in boldface are those syllables having salient pitch accents (transcribed below in Pierrehumbert's 1980 notation), and in small caps, the one among them with the highest pitch. At right is an indication of the highest pitch peak; the pitch at the final boundary (two pitches for 'continuation rises,' corresponding to the trough and the boundary); and an indication of pause time, if any.

Of all the prosodic resources or options that Pierrehumbert describes for English, only a relative few appear here: nearly all pitch accents are LH* (rises to high); they occur densely, i.e., several to an intonational phrase; the heights of pitch peaks fluctuate considerably, showing great range (with the highest marking both new topics, and very salient foci); and nearly all the cadences are falls to low. Further, pausing is highly facultative (Figure 7): the first pause phrase of the first routine is enormous (four sentences, five intonational phrases), while later on, and in the second routine, they are extremely short.

Why just these resources? Leno's choices may at first seem determined solely by content and communicative purpose. LH* intonation is often associated with new information (Hirschberg and Pierrehumbert 1991), and that is appropriate since he casts himself here as a bringer of news. His wide pitch range and high pitch accent density add punch and vividness. And his syncopated use of pause seems the essence of comedic timing. Yet this is not the whole story, for many of these effects can be approximated with a different set of resources: Why should he not draw on them too from time to time?

I would suggest that Leno has constructed what might be called a PROSODIC MACROSTRUCTURE—a small set of resources that then become the material out of which the patterns of Assumption 1 largely are crafted. It certainly is true that Leno's macrostructure is well-suited to his task. But beyond that, it becomes associated through use partly with him and his comedic style in the minds of his audience.

Macrostructures can range greatly in their conventionality, from those that are traditionally tied to particular genres, to those which have become habitual for individual in particular settings, to those composed quite on the spur of the moment. Queen (1992) raises this issue in her discussion of the oratory of Martin Luther King. She shows that King made a highly distinctive set of prosodic choices in his oratory, and demonstrates that these choices partly continue traditions of African-American preaching, and partly constitute a unique personal style. In the case of Jay Leno, his macrostructure may to some extent follow a tradition in American stand-up comedy and in part be his own construct. In any case it is interesting that it is not identical to the prosodic choices of other stand-up comedians, nor to the choices he himself makes in other speech settings (e.g., interviewing guests).

To the extent that prosodic macrostructuring turns out to be a significant fact of natural speech, several points can be made about it. First, it represents another way in which prosodic elements coordinate with each other to form prosodic structures, albeit a kind of structuring quite different even from that suggested by the Prosodic Hierarchy. Second, it means that natural discourse may be more orderly, and therefore more amenable to systematic

[ROUTINE I]			
An' **HERE'S** somp'm I got ou' th' **paper** today LH* LH* H*	285	131	<0.0>
a **MA**jor **New** York-- **new**spaper LH* H* LH*	158	109	<0.0>
I **THINK** i wz the **New** York **Po:st.** LH* LH* LH*	178	114	<0.0>
Now **THESE** a' **their** statisti:cs. LH* LH*	243	102	<0.0>
NOT mine:. LH*	238	89-147	<0.3>
They said **TWEn'y five** percent: LH* HL*	322	123	<0.0>
of the homeless are alco**HOL**i:cs, LH*	128	79	<0.7>
TWEn'y five percen' are **drug** a:ddicts, (L)H* HL* LH*	228	106-217	<0.4>
and **THIR**ty percen' LH*	316	116	<0.0>
'ave been insti**TU**tionalized LH*	232	97	<0.0>
at **ONE** time 'r another LH*	149	94-119	<0.0>
for **mental** disa**BI**lities. LH* (L-) LH*	232	100	<1.0>
Now I **know** that **seems** like a **pr**etty high per**CEN**'age. LH* LH* LH* LH*	200	109	<0.0>
But ya **KNOW,** LH* H*	166	166	<0.1>
when **ya** com**PARE** it to **co:ngress** LH* LH* LH*	312	111	<0.0>
gee it's **NO:T tha:t high: really.** LH* LH* (L)H* LH* L- L*	370	105	<1.0>
Yknow.		--	
Applause, etc.			<...>
[ROUTINE II]			
You **KNOW** what's **grea:t:**? LH* HL*	400	--	<0.0>
SEE, H*	250	250	<0.5>
SEE, H*	243	198	<0.2>
I **AL**ways **like** to watch **politicians::** LH* LH* LH* LH*	400	106	<0.0>
try to **JU**Stify:: H* LH*	185	102	<0.0>
their **JO:BS.** LH*	133	94	<0.0>
I **SAW** a **se**nator LH* LH*	200	96	<0.0>
on **ONE** a' those: LH*	116	(101)	<0.6>
sunday morning **TALK** shows the other day. H* LH*	163	86	<0.0>
An' **'E said** LH* H*	149	89	<0.0>
th't the **AC**tions of the **se:**nate LH* LH*	270	87	<0.5>
have cre**A**ted **alot** a **jo:bs:** LH* LH* LH*	164	104-107	<0.3>
for **alot** a **CI**tizens:. L* LH*	208	86	<0.0>
YEAH but: H* L- L*	133	90	<1.0>
LETS fa:ce it. H* LH*	208	95	<0.0>
YOU can't make a **career** out o'**jury** duty:. LH* LH* LH*	185	83	<0.2>
Ya **KNOW**? LH*	153	153	<1.0>
You know what I mean?	--	--	<0.0>
That--<0.5--<obs> **THAT 'S five bucks** a **day:** LH* LH* LH* LH*	169	97	<...>

Figure 6: Two American English standup comedy routines performed Jay Leno on NBC television, March 1990. The two routines occurred in immediate succession.

An' HERE'S somp'm I got ou' th' paper today % a MAjor New York-- newspaper % I THINK i wz the New York Po:st. % Now THESE a' their statisti:cs. % NOT mine:.%	<0.3>
They said TWEn'y five percent: % of the homeless are alcoHOLi:cs,%	<0.7>
TWEn'y five percen' are drug a:ddicts,%	<0.4>
and THIRty percen' % 'ave been instiTUtionalized % at ONE time 'r another % for mental disaBIlities.%	<1.0>
Now I know that seems like a pretty high perCEN'age % But ya KNOW,%	<0.1>
when ya comPARE it to co:ngress % gee it's NO:T tha:t high: really.%	<1.0>
Yknow. % *Applause, etc.*	<...>
You KNOW what's grea:t:? % SEE, %	<0.5>
SEE, %	<0.2>
I ALways like to watch politicians:: % try to JUStify:: % their JO:BS. % I SAW a senator % on ONE a' those: %	<0.6>
sunday morning TALK shows the other day. % An' 'E said % th't the ACtions of the se:nate %	<0.5>
have creAted alot a jo:bs: %	<0.3>
for alot a CItizens:. % YEAH but: %	<1.0>
LETS fa:ce it. % YOU can't make a career out o'jury duty:. %	<0.2>
Ya KNOW? %	<1.0>
You know what I mean? %	<0.0>
That--	<0.5>
<obs> THAT 'S five bucks a day: %	<...>

Figure 7: Transcript of Figure 6, reformated so that line-breaks correspond to pauses. '%' marks cadence locations.

study, than often thought. Third, it points up the need for 'thick' description in natural discourse study, since it is the particular mix of elements that gives a macrostructure its cultural and stylistic associations, rather than the abstract patterns they instantiate. At the same time, of course, such variation makes it clear that an appropriate notion of pattern must be formulated abstractly enough so as not to be locked to particular phonetic forms or choices. However practical they may be in some respects, this in fact is a weakness of many transcription oriented approaches to prosody (e.g., Hirschberg and Beckman 1992) since they tend to emphasize just certain elements, without regard to their centrality in forming salient distributional patterns or in conveying discourse meaning in the speech sample in question.

7. NATURAL AND CONVENTIONAL ASPECTS OF PROSODY

We last consider natural and conventional aspects of the distribution and interpretation of prosodic elements. The case in point is a stylized downtrending or deaccenting phenomenon that is salient in the narrative prose of many, but not all, speakers of the CAY dialect of Chevak and Hooper Bay, Alaska. This dialect is moderately different form that discussed in Sec. 5.

Shown in Figure 8 is a section from a myth told by Thomas Moses of Chevak. On thematic grounds, it constitutes a single episodic unit. Each line is a word (or two where intonation is continuous). On a pattern that is seldom violated by Chevakers, the unit on each line shows a clear pitch trough initially (on the first stress of the word), followed by a peak. The peak is at the end of the word, unless a low tone occurs there (moving the peak back to the last stress). Pragmatically, the final low marks disjunction—the lower the tone, the greater the implied break. Generally these breaks correspond well, in both placement and degree, to syntactic constituency breaks (Woodbury 1989). Lower case 'w' indicates 'whisper'; these whispers count as very low tones and arise when the tone heads below a certain threshold. Timing of pauses, if any, follows next; and last is the English translation.

The phenomenon of interest, which I call ATTENUATION, is characterized by significantly lowered H pitch peaks and reduced amplitude. Attenuated sequences are marked in Figure 8 by plain type, while nonattenuated sequences are boldface. The high pitch values of the attenuated items are generally noticeably lower. I am not in a position to assert that the attenuated/nonattenuated distinction is categorical (rather than gradient), even though it usually sounds and looks quite distinctive. Categorical or not, it counts as a prosodic structure, rather than a prosodic element, since it involves a cluster of prosodic elements (pitch scaling and amplitude) which pattern together.

A first observation about attenuation in Figure 8 is that it accompanies all postposed constituents (which are underlined). Because CAY has very rich inflectional morphology, these postposed constituents can usually count as supplements to already-fully-formed sentences. In terms of Assumption 3, there is an apparently exceptionless default pattern holding between two components (syntactic patterning, and the patterning of prosodic attenuation):

> Postposed constituents are prosodically attenuated.

This still leaves instances where nonpostposed material is attenuated. When the distribution of a prosodic structure

CAY	Init L	Peak	Final L	Pause	English
Piuraqerluni^taw'	84	135	96	<3.5>	**Then once**
caller'e'mini,	82	108		<1.3>	**when she was doing things,**
taun'^ar'e'naq,	81	92	76	<1.5>	that woman,
angutmeng^uumeng	103	149	120	<0.0>	**this man**
tang'elliuq:	89	105		<2.0>	appeared to her:
Kanaqliit=gguq=gg'^atk'ekui!	76	105	w	<6.0>	**He had a parka all of muskrat!**
Piluku^taw'	81	104		<3.5>	**He said to her**
amatngurrvakaami^taw'	100	140	126	<0.0>	**that because he was so grateful**
nuliq—nuliq—nuliqnaluk'	93	113		<0.0>	**he had come to ask her**
ullagyaaqniluku.	84	104	w	<5.5>	**to be his wife.**
Taw—Taw—Tawaten,	78	88		<1.4>	S—S—So [he spoke],
anautellermineng	81	105		<0.0>	because he was so grateful
amatngurpakaami.	84	92	w	<4.0>	that she had rescued him.
Tawa=ggur^taum	70	102		<0.0>	**And so that one**
civunran^taw',	84	102		<0.5>	**who stood before him,**
tupekluku,	74	84		<0.5>	accepted him,
uing—uing—uingyunrilami.	84/68	84	w	<5.5>	be—be—because she had no husband yet.
Tang—Tangnerrayauluni	76	138		<0.0>	**He—He seemed a stranger**
Tangnerrauluni=gguq^taw'	81	128		<0.0>	**He looked a bit strange**
taun'^angun.	78	85	w	<5.5>	that man.
Cuna=ggur^un'^taw'	86	94		<0.0>	So it was for this [woman]
nuliqluku^taum'^taw'	86	94	81	<0.0>	he married her this [man]
nuliqsagulluku.	79	93	w	<1.7>	He had her as his wife.
Piculliniluni=ggur^taun'	95	101	93	<0.0>	He was good at getting things this [man],
pissuraqami^tawaam.	82	99	w	<2.5>	when he hunted.
Maklagculuni=llu.	80	91	w	<3.5>	And good at getting bearded seal.

Figure 8: Opening of a Central Alaskan Yupik Eskimo myth performance by Thomas Moses of Chevak, Alaska, recorded in 1978 (text in Woodbury 1984).

(or element) appears not to depend on the distribution of something else, one must suspect (on the null hypothesis) that the fact of its placement alone may contribute new information. In light of the postposed attenuation cases, and certain features of the nonpostposed cases, I suspect that attenuation contributes the following pragmatic information:

> Attenuation defocuses syntactic/prosodic constituents and labels them as clarifications or SUPPLEMENTS to the interpretation that the speaker expects the audience to have constructed or deduced from the talk so far.

It is fairly clear how this applies to postposed constituents. However, it may still be reasonable to maintain the exceptionless default posited above in order to enforce the link between prosody and syntax explicitly (rather than suppose that all instances of postposing will function in context as supplements).

The pragmatic account also applies in the nonpostposed cases. An interesting one is the attenuation of nonpostposed *tang'elliuq* 'he appeared (to her)' in the fifth line of Figure 8. While 'appeared to her' cannot felicitously be cut from the English translation, it happens that CAY speakers routinely not only attenuate, but sometimes completely delete, verbs of seeing, saying, apparition, and the like. It is therefore plausible to treat *tang'elliuq* as defocused. Indeed, even for a related class of English verbs, there is a tendency place nuclear stress on the subject in preference to the verb (e.g., *THE COPS came*, *CLINTON spoke*, *TRUMAN died*, etc.) The parallel is quite striking.

Another interesting set of cases are the attenuated sentences near the bottom. Each of them contributes information which Native hearers could plausibly infer. For example, the last two sentences ascribe to the man certain abilities. Yet these are already inferable by the following logic. In CAY myth, animal transformers wear their fur or feathers as parkas; since the man's parka is all of muskrat, he is a transformed muskrat (and we learn that for sure later in the story); therefore, as a muskrat, he should be a good aquatic hunter. The

importance of these examples is that they show that the notion of supplementation described above holds not only within, but between, sentences. Because of that, it is most plausible to view supplementation as a discourse category that happens to have a conventional relationship to syntax (via the exceptionless default).

The description so far simply assumes that the pragmatic account of attenuation is a matter of convention in the relevant speech community. But this would hardly explain its similarities in both form and function to such English phenomena as postnuclear 'deaccenting' and tag intonation. In work on the form and meaning of cadences and related pitch figures in the speech of some University of Texas sorority members, McLemore (1991) has argued that natural, iconic, principles constrain the interpretation of prosodic elements. Accordingly, an iconic account for attenuation might then run as follows:

> Attenuated sequences are less prominent than nonattenuated sequences. When they follow nonattenuated sequences with otherwise similar pitch patterns, they become less prominent replicas. These properties make available the defocusing and supplementing functions, which in turn 'invite' certain postposed constituents.

At the same time, McLemore emphasizes that while iconic principles may direct interpretation, they cannot strictly determine it: a role is needed for cultural conventions. Applying this in the present case, it happens that attenuation (and the pragmatic category described as supplementation) are important tropes in much Chevak narrative. In other localities they are not. For example, the narrative in Sec. 5 contains just a few clear case of attenuation and supplementation. In yet other CAY communities, supplementation is frequent but it is not marked by attenuation (Woodbury 1992). Details like these are too particular ever to follow solely from iconic principles: they must continue historical patterns of actual use and interpretation.

8. CONCLUSION

We have made some basic assumptions about discourse structure and used them to gauge the extent to which prosodic elements in individual samples of natural discourse show structure and convey meaning. The advantage at very least is their generality and broad applicability to natural discourse prosody. Prosodic Hierarchy Theory presupposes too much structural convergence, while descriptively-oriented transcription systems presuppose the importance of particular prosodic elements regardless of their importance to the distributional or pragmatic structure of the particular speech in question.

Furthermore, I hope to have shown that investigation in these terms quickly points up interesting phenomena. Most importantly, we have been able to find prosodic structuring in natural discourse, albeit of a more diffuse and ramified type than that posited by Prosodic Hierarchy Theory. Also revealed were a complex relationship of prosodic structures and elements to thematic structure; the phenomenon of prosodic macrostructuring; and the influences of iconic principles and cultural conventions on the use of prosodic structures. Once noticed and formulated in general terms (however tentatively), such findings serve as guides in further investigation.

While I have claimed some progress on the Distribution and Meaning questions noted at the beginning of this paper, the Identity question has been left nearly untouched. One part of what is needed is simply to continue to improve our understanding of the phonetics and phonology of all potential prosodic elements. Another aspect—consonant with the natural discourse oriented approach described here—is to characterize as precisely as possible the phonetic (or phonological) correlates of the significant elements of distribution and meaning in samples of natural discourse prosody, regardless of their simplicity or complexity. That is, it is necessary to know in what ways logically separate phonetic elements of prosody might be bundled together and treated as single elements of prosodic distribution or prosodic meaning; which such bundles are common in particular cultures, languages, speech event types, or idiolects; and whether their meanings show family resemblances regardless of where they occur.

REFERENCES

1. Bauman, Richard, and Joel Sherzer. 1974. Explorations in the ethnography of speaking. Cambridge University Press.
2. Chafe, Wallace L. 1980. The deployment of consciousness in the production of a narrative. In Chafe, W.L. (ed.), The pear stories: cognitive, cultural and linguistic aspects of narrative production. Norwood, NJ: Ablex.
3. Gumperz, John J. 1982. Discourse strategies. Cambridge, UK: Cambridge University Press.
4. Halliday, M.A.K. 1967. Intonation and grammar in British English. The Hague: Mouton.
5. Hayes, Bruce. 1989. The prosodic hierarchy in meter. In Kiparsky, P. & G. Youmans eds., Phonetics and phonology, Vol I: Rhythm and meter. New York: Academic Press. 201-260.
6. Hirschberg, Julia and Mary Beckman. 1992. Draft report on the TOBI conventions for prosodic labeling. Bell Labs MS.
7. Hirschberg, Julia and Janet Pierrehumbert. 1986. The intonational structuring of discourse.

Proceedings of the 24th Annual Meeting of the ACL. Morristown, NJ: ACL. 136-144.

8. Hymes, Dell. 1974. Foundations in sociolinguistics. U Pennsylvania Press.

9. Hymes, Dell. 1981. In vain I tried to tell you: Essays in Native American ethnopoetics. Philadelphia: U of Pennsylvania Press.

10. Inkelas, Sharon and Draga Zec. 1990. The phonology-syntax connection. Chicago: U of Chicago Press.

11. Jakobson, Roman. 1960. 'Concluding statement: linguistics and poetics. In T. A. Sebeok ed., Style in language. Cambridge: MIT P.

12. Labov, William. 1972. Rules for ritual insults. In W. Labov, Language in the inner city. Philadelphia: U Pennsylvania Press. 297-353.

13. Levinson, Stephen C. 1983. Pragmatics. Cambridge, UK: Cambridge University Press.

14. Liberman, Mark Y., Cynthia A. McLemore, & Anthony C. Woodbury. 1991. On the nature of prosodic phrasing. Grammatical foundations of prosody and discourse: the conference. LSA Linguistic Institute, Santa Cruz.

15. Liberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. Linguistic Inquiry 89:249-336.

16. McLemore, Cynthia. 1991. The pragmatic interpretation of English intonation: sorority speech. Doctoral Dissertation. University of Texas, Austin.

17. Michelson, Karin. 1991. Semantic and discourse factors in Oneida utterance-final phonology. Ms.

18. Miller, Corey. 1992. Prosodic aspects of M. L. King's "I have a dream today" speech. This volume.

19. Nespor, Marina, & Irene Vogel. 1986. Prosodic phonology. Dordrecht: Foris.

20. Pierrehumbert, Janet B. 1980. The phonology and phonetics of English intonation. M.I.T. doctoral dissertation.

21. Pierrehumbert, Janet B., and Mary Beckman. 1988. Japanese tone structure. Cambridge: MIT Press.

22. Pierrehumbert, Janet and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack eds., Intentions in communication. Cambridge: MIT Press. 271-311.

23. Queen, Robin. 1992. Prosodic organization in the speeches of Martin Luther King. This volume.

24. Sapir, Edward. 1949. Male and female forms of speech in Yana. In Mandelbaum, David, ed., Selected writings of Edward Sapir. Berkeley: U of California Press. 206-212.

25. Selkirk, Elisabeth O. 1980. The role of prosodic categories in English word stress. Linguistic Inquiry 11:563-605.

26. Silverstein, Michael. 1976. Shifters, linguistic categories, and cultural description. In Basso, Keith and Henry Selby, Meaning in anthropology. Albuquerque: U New Mexico Press. 11-56.

27. Silverstein, Michael. 1979. Language structure and linguistic ideology. In Paul R. Clyne et. al., eds., The elements: A parasession on linguistic units and levels. Chicago: Chicago Linguistic Society. 193-247.

28. Silverstein, Michael. 1984. On the pragmatic 'poetry' of prose: parallelism, repetition and cohesive structure in the time course of dyadic conversation. In Schiffrin, Deborah, ed., Meaning, form, and use in context: linguistic applications. Washington: Georgetown U Press. 181-199.

29. Tedlock, Dennis. 1983. The spoken word and the work of interpretation. Philadelphia: University of Pennsylvania Press.

30. Woodbury, Anthony C. 1984. Cev'armiut qanemciit qulirait=llu: Narratives and tales from Chevak, Alaska. Fairbanks: University of Alaska Press.

31. Woodbury, Anthony C. 1987a. Rhetorical structure in a Central Alaskan Yupik Eskimo traditional tale. In Sherzer and Woodbury eds., Native American discourse. Cambridge, UK: Cambridge U Press.

32. Woodbury, Anthony C. 1987b. Meaningful phonological processes. Language 63:685-740.

33. Woodbury, Anthony C. 1989. Phrasing and intonational tonology in Central Alaskan Yupik Eskimo: some implications for linguistics in the field. In John Dunn, ed. 1988 Mid-America Linguistics Conference papers. Norman: U. of Oklahoma. 3-40.

34. Woodbury, Anthony C. 1992. Utterance-final phonology and the prosodic hierarchy: a case from the Nunivak dialect of Central Alaskan Yupik Eskimo. Paper given at Linguistic Society of America Annual Meeting, Philadelphia.

It's not what she says, it's the way that she says it: the influence of speaker-sex on pitch and intonational patterns

Nicola Woods[1]

Abstract

In this paper I discuss the relationship between speaker-sex and pitch and intonational features of language. I examine the spontaneous speech of male and female adults and children and pay specific attention to (i) pitch movements on nuclear syllables (what Halliday (1966) refers to as *tone*); (ii) pitch range; and (iii) maximum pitch. Results show that particular patterns of tone and pitch are characteristic of male and female speech.

1. Introduction

Over the past fifteen or twenty years considerable attention has been paid to the linguistic features which distinguish between the speech of men and women: research has detailed the segmental (phonological), lexical and syntactic features which characterise the linguistic behaviour of males and females. However, far less attention has been paid to the non-segmental - specifically, pitch and intonational - features which characterise men's and women's speech styles. Furthermore, it seems fair to say that the little information which is to be found in the literature is largely of an anecdotal rather than empirical nature[2].

The aim of the research which informs this paper was to investigate empirically the relationship between speaker-sex and the use of particular non-segmental linguistic features. I concentrate on proposing and providing evidence for three points. Firstly, and most fundamentally, I show that non-segmental features vary systematically according to speaker-sex in a way similar to that which has been observed for other levels of language use. Secondly, by detailing how children display similar patterns of sex-related variation as adults, I highlight the important role of non-segmental linguistic features in children's acquisition of sex-appropriate speech styles. Finally, I give evidence of patterns of stylistic shifting which strongly suggest that pitch of voice cannot be fully explained by reference only to physiological factors, but is also conditioned by the social factors which influence linguistic behaviour.

2. Methodology

The spontaneous speech produced by male and female adults and children was recorded: 5 men, 5 women, 5 boys and 5 girls. The adults were work colleagues aged between 27 and 32 years. All had attended private school, were university educated, and were speakers of Southern British English. The children were classmates (private fee-paying school) aged between 6.6 and 7.6 years. All were native Southern British English speakers of British English-speaking parents who were of a similar social class[3].

The investigation was designed in such a way as to allow the elicitation of the spontaneous speech used by informants within two different social settings: a casual conversation with a friend, and a formal interview.

2.1 Recording adults

Recordings of adults (men and women) were made in the informants' place of work. The two settings were created in the following way:

Conversation: two informants were asked to arrive for an interview at precisely the same time. They were then ostensibly "kept waiting" outside the interview room. In every case the two informants engaged in conversation. The conversations were surreptitiously recorded[4].

Interview: individual informants were brought into the interviewer's office where they were first asked a number of casual/social questions. The informants were then told that the interview "proper" was about to begin and a video camera was focused upon them. Questions in the interview were on the topic of the informants' previous employment and current work projects. Subjects' responses to these formal questions were openly recorded on the video soundtrack.

2.2 Recording children

Recordings of children were made in their classroom and in an adjoining music room. As with adults, two samples of speech were elicited from each child. The two settings were created in the following way:

Conversation: recordings were made of pairs of children conversing while making a collage. In the course of this activity children chatted together both about the picture they were making, and about other topics: e.g. the ballet they were due to perform, green "mutant" turtles, and world cup football. These conversations were surreptitiously recorded.

Interview: each child was brought into a small room adjoining their classroom where they were shown the tape-recorder and told how it worked and was used. Each child had a turn at recording and hearing a play-back of their voices before the interview started. The children were then shown the tape recorder being switched on and running. During the interview I asked each child questions about school, home, friends and family.

[1]Contact address: Linacre College, Oxford OX1 3JA.
[2]Notable exceptions to this include the work of Local (1978, 1982), Pellowe and Jones (1978) and Graddol and Swann (1983, 1989).
[3]By using these methods of selection, it was hoped that adults and children could be justifiably said to be speakers of the same or at least a similar variety (discussed at length in Woods 1992).
[4]After the conversational data had been elicted, informants were advised about the surreptitious recordings and were given the opportunity to refuse permission for the tapes to be used. No such refusals occurred.

Once a sample of each informant's speech in each situation had been collected the task of transcribing and analysing the data was begun (both impressionistic and instrumental techniques of transcription were employed). Amongst other non-segmental features, the use of pitch movements on nuclear tones, pitch range and maximum pitch were detailed. Repeated-measures analyses of variance were employed (using the Statistical Package for the Social Sciences - SPSS) in order to assess the significance of male-female differences in the use of pitch and tonal linguistic features. Significance was set at $p<.05$.

3. Results

The following presentation of results is divided into four parts. In the first section I detail the statistical results of the analysis of male and female speech (results refer to both adults and children). In the second section I discuss the pitch and intonational features used by adult men and women. Thirdly, I examine children's use of non-segmental linguistic features. And finally, in the fourth section, I present results which indicate that aspects of pitch of voice are socially as well as physically conditioned.

3.1 Statistical analysis of male-female differences

3.1.1 Tone

Results showed that the use of three tones distinguished between the speech of men and women: a 3-factor repeated-measures analysis of variance (sex x age x situation) showed highly significant between-subjects effects of speaker-sex on the use of (i) rising tones ($F = 8.98$; $df = 1,16$; $p = .009$); (ii) high fall tones ($F = 12.09$; $df = 1,16$; $p = .003$); and (iii) level tones ($F = 9.68$; $df = 1,16$; $p = .007$). These statistics reveal that females use more rising and high falling tones than males, and males more level tones than females.

No interactions were observed in the use of rising and high fall tones. This indicates that the influence of speaker-sex was consistent in both age groups in both social situations (women used more rise and high fall tones than men, and girls used these tones significantly more frequently than boys in both the conversational and more formal interview speech settings). An interaction between speaker-sex, situation and the use of level tones was observed. This interaction showed that as well as males using more level tones than female speakers, both sexes use more level tones in the formal context than in the informal/casual conversational setting (interaction: $F = 9.68$; $df = 1,16$; $p = .007$).

3.1.2 Pitch

Instrumental analysis showed that females typically use a greater fundamental frequency (Fo) range than males[1]. A 3-factor repeated-measures analysis of variance (sex x age x situation) showed a highly significant between-subjects effect of speaker-sex on informants' Fo range characteristics ($F = 13.29$; $df = 1,16$; $p = .002$). No interactions between speaker-sex and the other independent variables - age and social situation - were observed. The influence of speaker-sex on Fo range was thus shown to be consistent for both adult and child informants in both social settings: in both the conversational and interview situations females used a significantly wider range of fundamental frequencies than males.

3.2 The pitch and intonational patterns of adult men and women

> "It is generally thought that ... there are some intonation patterns, impressionistically the "whining, questioning, helpless" patterns, which are used predominantly by women". (Eble, C. 1972: 246)

(i) Rise tones

In support of previous research[2], results showed that women used more rising tones than men. This was a finding consistent in both social situations studied. In the first instance, this result seems to provide empirical evidence for anecdotal observations on male-female differences. For example, Brend (1975), in a study of American English, claims that women are the "sole" users of simple rising tones; tones which she argues reflect women's "polite" and "cheerful" natures. And Lakoff (1975) as well as claiming that women use more rise tones than men, also argues that women use rises in, if not a grammatically, then at least a pragmatically ill-formed way. That is, although the use of rise tones has been linked to the use of interrogative structures[3], Lakoff claims that women use rises when **answering** questions; a tendency which Lakoff attributes to women's "insecurity" and "lack of confidence".

However, when the category of rise tones was broken down into its component parts - complex fall-rise[4] tones and simple rise tones - I found that women's greater use of rises was a consequence of their significantly more frequent use of the former (complex fall-rise), rather than the latter (simple rise) tones. Figures 1 and 2[5] detail men's and women's use of complex fall-rise tones in the conversational and interview speech encounters. In the case of simple rises no significant difference (in either context studied) between men and women was observed. This result from British English is thus in contrast to the proposals of Brend and Lakoff, both of whom have claimed that it is the use of simple rises which characterises the speech of women.

[1] In this paper I equate Fo and pitch. However, in my thesis I discuss in detail the complex relationship which holds between these two.
[2] See Pellowe and Jones' (1978) research into the speech patterns used on Tyneside, and Elyan's (1978) study of students in Bristol.
[3] Although analysis of my spontaneous speech data showed, in line with many others (e.g. Kenworthy 1978, Gelykens 1988), that the relationship between rise tones and interrogative syntax was not predictable.
[4] I use this term to refer both to compound fall+rise and complex fall-rise tones. I acknowledge that in a phonological description of English it is necessary to distinguish between these two.
[5] Figures show the use of particular tones as a percentage of informants' total use of tones. Mean scores are detailed.

FIGURE 1

Fall-rise tones
conversation

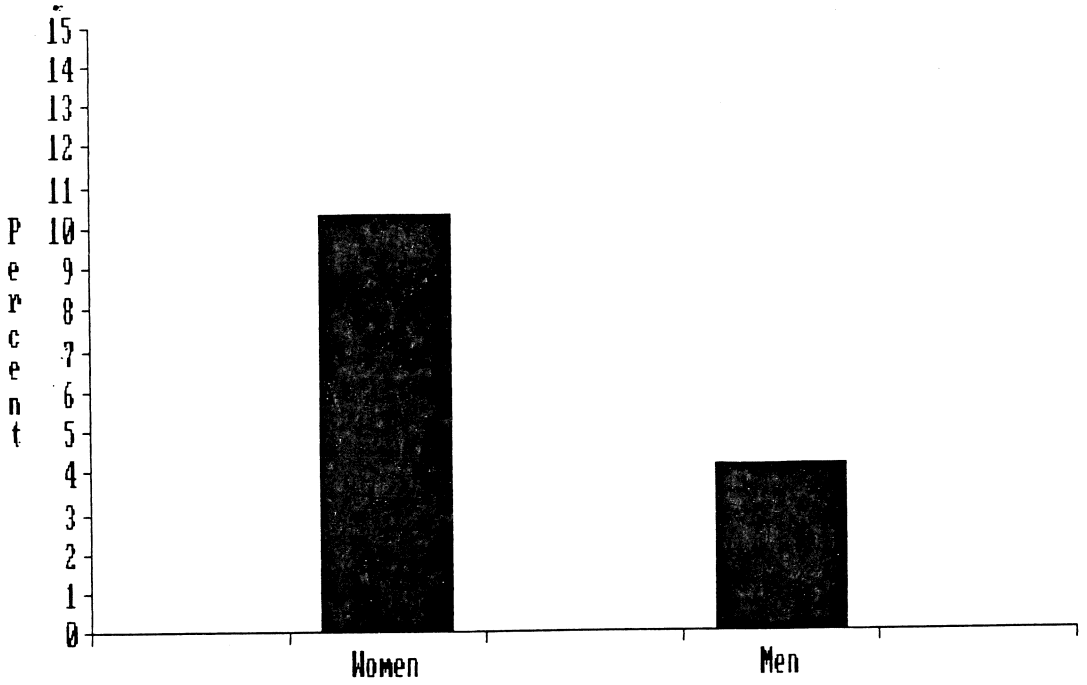
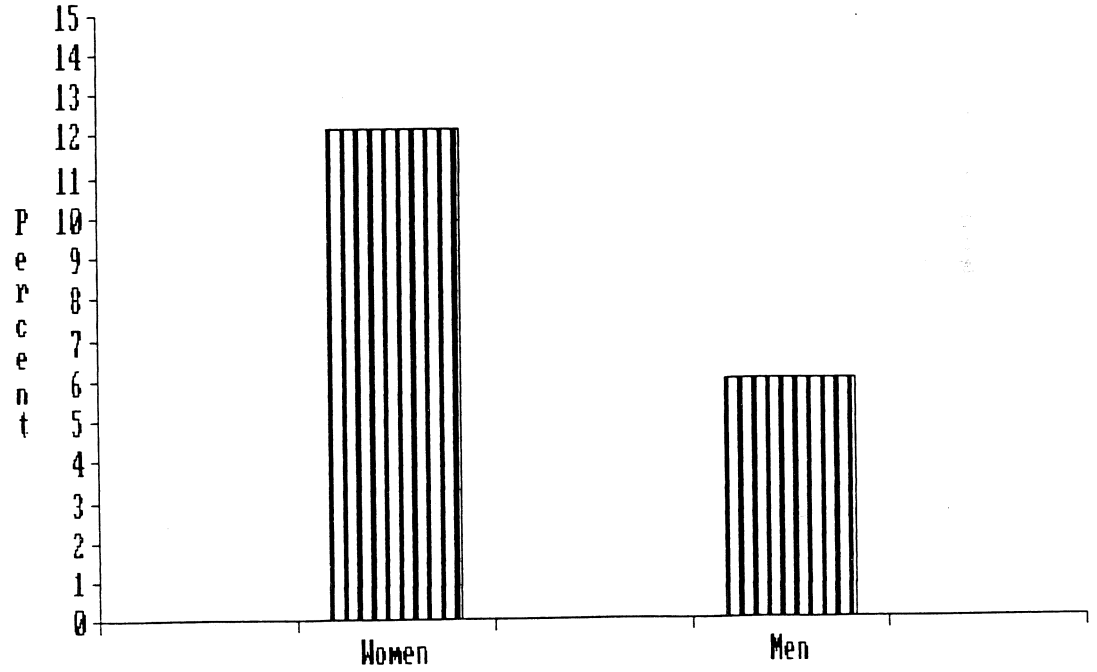


FIGURE 2

Fall-rise tones
interview



(ii) High fall tones

My research reveals that high fall tones also distinguish between men's and women's spontaneous speech styles: in both social situations studied, women used significantly more high fall tones than men. See figures 3 and 4.

(iii) Level tones

Finally, men were noted to use significantly more level tones than women. Again, as shown in figures 5 and 6, this was a finding consistent in both the conversational and interview speech encounters.

Significant differences were thus observed in men's and women's use of three types of nuclear tone; this clearly suggests that the intonational feature of tone is socially indexical of speaker-sex. However, before drawing this conclusion it is worth considering alternative hypotheses which may explain the patterns of variation. Particularly, it is important to consider whether any of the approaches which have been traditionally taken towards intonational function - grammatical, attitudinal and discourse-based - propose accounts which explain the apparently sex-related variation. A full consideration of each of these approaches is provided in my doctoral thesis (Woods, forthcoming 1992). Here I will just give an example of a hypothesis which seems to offer a persuasive explanation of the patterns of tonal variation observed.

The grammatical approach to intonation suggests that the use of tones functions to characterise various syntactic types: specifically, as I mentioned above, it is suggested that falling tones mark declarative structures, and that rising tones characterise the use of polar interrogatives. Since rises have been linked to interrogative syntax, then a possible explanation of women's greater use of rise tones could be that women ask more questions than men. Indeed, precisely this characterisation of women's speech has been proposed by a number of researchers (e.g. Lakoff 1975). If this were the case, then the difference noted between men's and women's use of rise tones would have its basis in essentially syntactic rather than tonal variation.

Given this possible interpretation, an analysis of the syntactic structures used by informants was carried out. This analysis showed that women did not use more interrogative forms than men. In fact, in the interview setting, men were observed to ask more questions than women. Added to this, I also found that 48% of the polar interrogative forms which occurred in the data did not carry any type of rising tone. Thus women's frequent use of fall-rising tones can not be attributed to their (alleged) preference for using interrogative syntax.

It thus seems fair to conclude that in certain settings particular tones are characteristic of men's and women's speech. Specifically, the result that the same patterns of sex-related variation are displayed in the data elicited from two entirely different contexts of interaction at two different times may be taken as strong evidence to suggest the significant influence of speaker-sex on the use of the intonational feature of tone.

3.2.2 Pitch

> "Her voice was ever soft, gentle and low, an excellent thing in woman"
> (*King Lear*, V.iii)

The average pitch of voice of men and women is clearly very different. This is largely, although as I show in section 3.4 not solely, a consequence of their different laryngeal anatomy: men have longer and thicker vocal folds which vibrate at lower frequencies than those of women.

However, it is not nearly so clear that pitch range is determined by larynx anatomy. It is therefore significant that my research showed differences in the range of frequencies used by men and women. Essentially, women were found to employ a far greater part of their potential pitch range than men[1]. As figures 7 and 8 show, this was a finding consistent in both the conversational and interview settings.

Thus pitch range, like the feature of tone, distinguished between men's and women's linguistic behaviour in two different speech encounters. This provides further evidence for the claim that non-segmental aspects of speech are socially indexical of speaker-sex.

Having noted these differences in men's and women's speech, the further question of whether male and female children use different patterns of pitch and intonation was addressed[2].

3.3 The pitch and intonational characteristics of boys and girls.

3.3.1 Tone

Just as three tones distinguished the speech of men and women, similarly the differential use of three tonal contours characterised the speech of girls and boys. Most significantly, the tones which showed sex-related variation in children's speech were essentially the same as those which were observed as indexical of speaker-sex in adults: that is, rise, high fall and level tones. In similar patterns of sex-related variation to that observed in adults, I found that girls used rise and high fall tones significantly more frequently than boys, and boys used level tones significantly more frequently than girls.

[1]A result which supports the incidental findings of Johns-Lewis (1986) and Graddol (1986), both of whom were concerned primarily with charting variation in aspects of Fo according to changes in discourse mode.
[2]Local (1982) observes the acquistion of dialect specific tonal forms in children as young as 4 & 5 years of age. Also, Local suggests that two rather different varieties are acquired by male and female children.

FIGURE 3

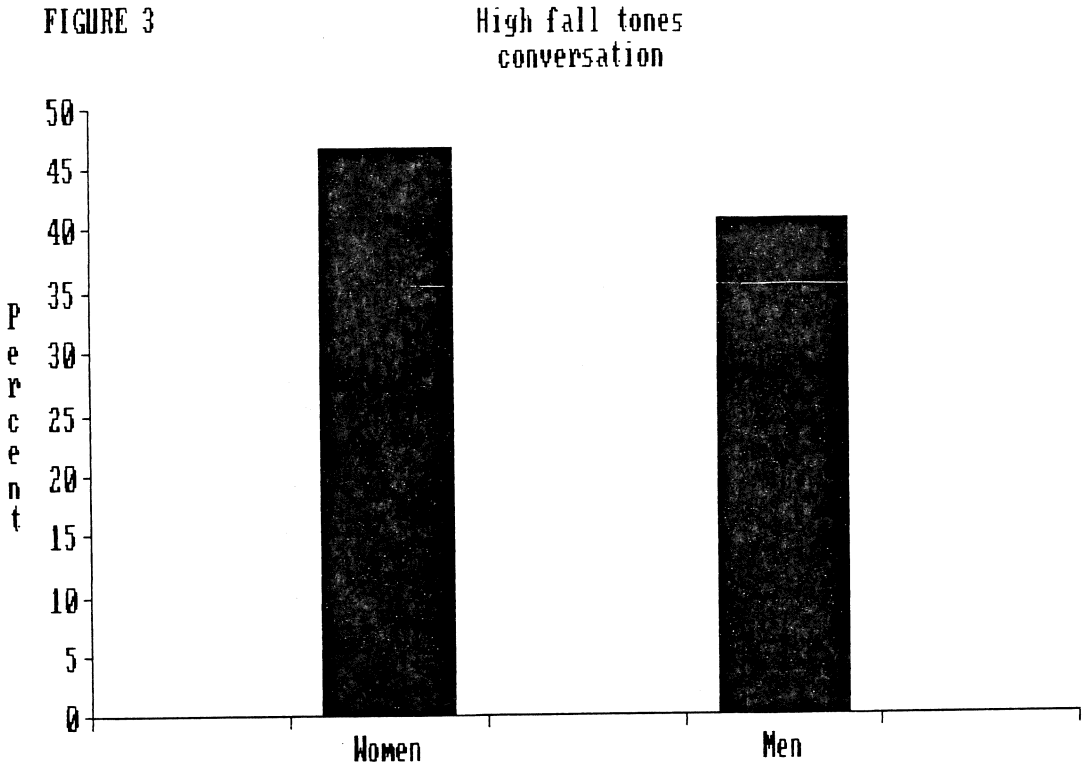


FIGURE 4

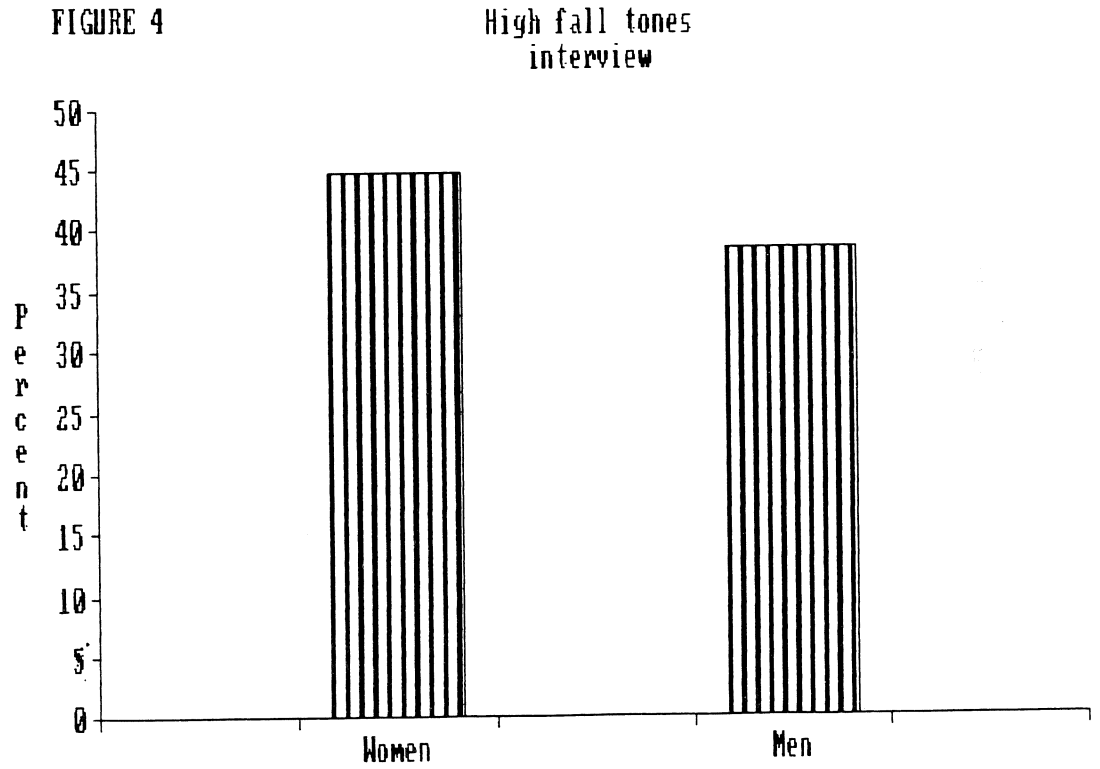


FIGURE 5

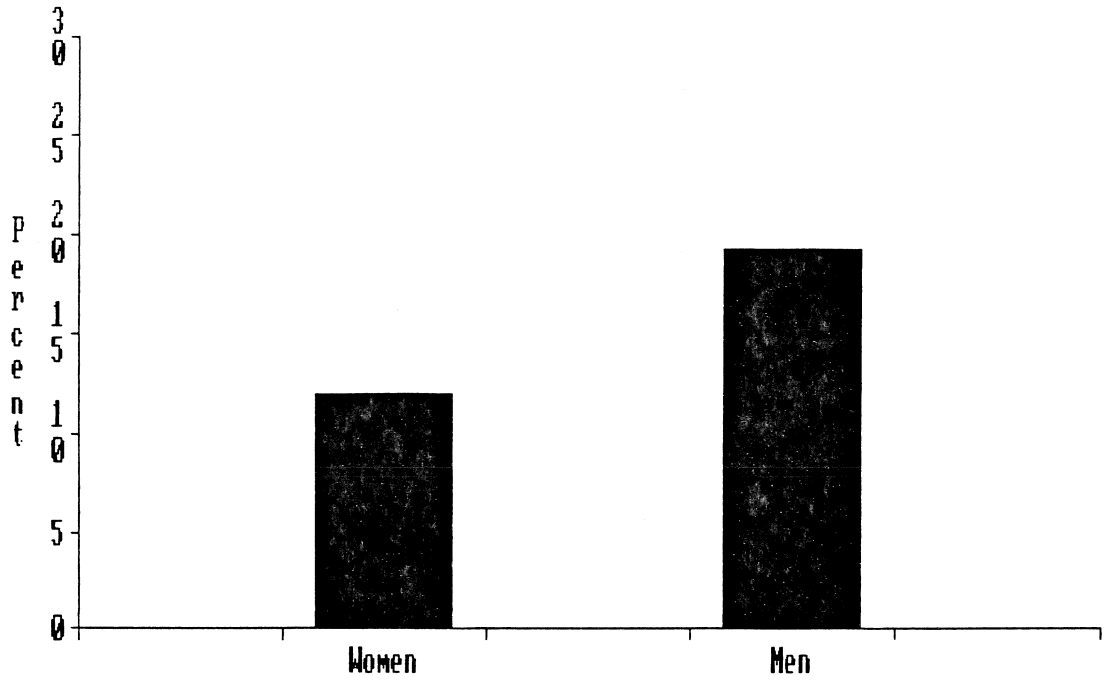


FIGURE 6

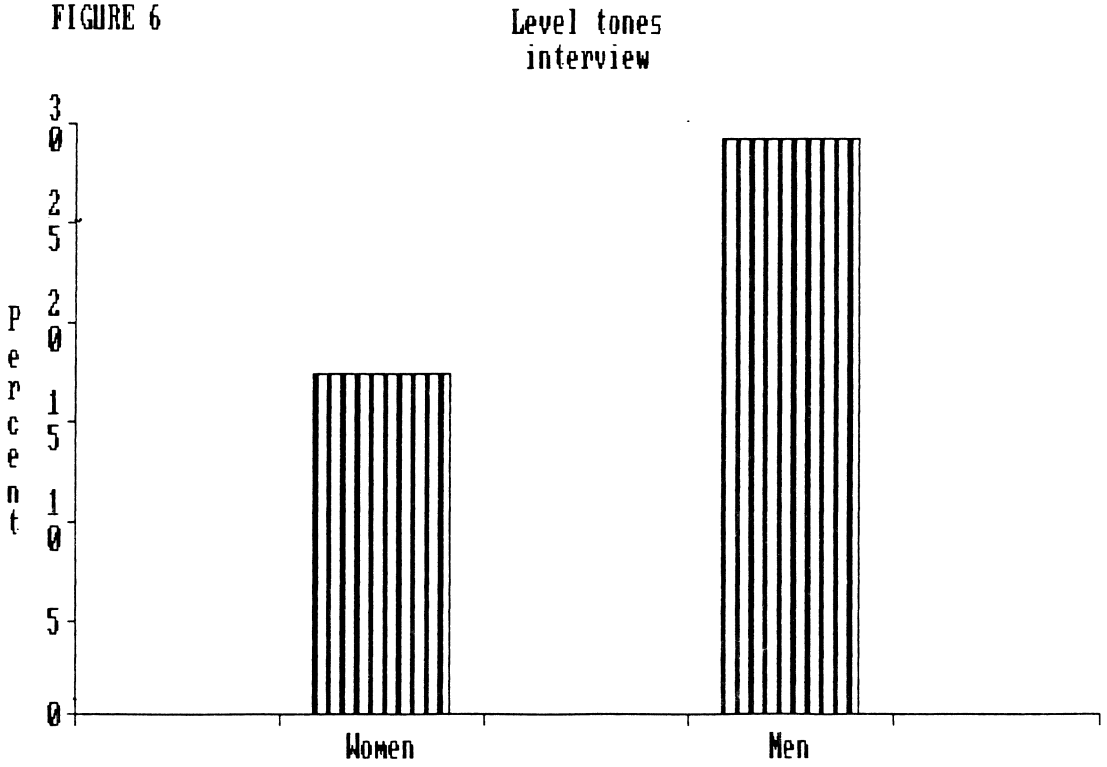


FIGURE 7

Pitch range
conversation

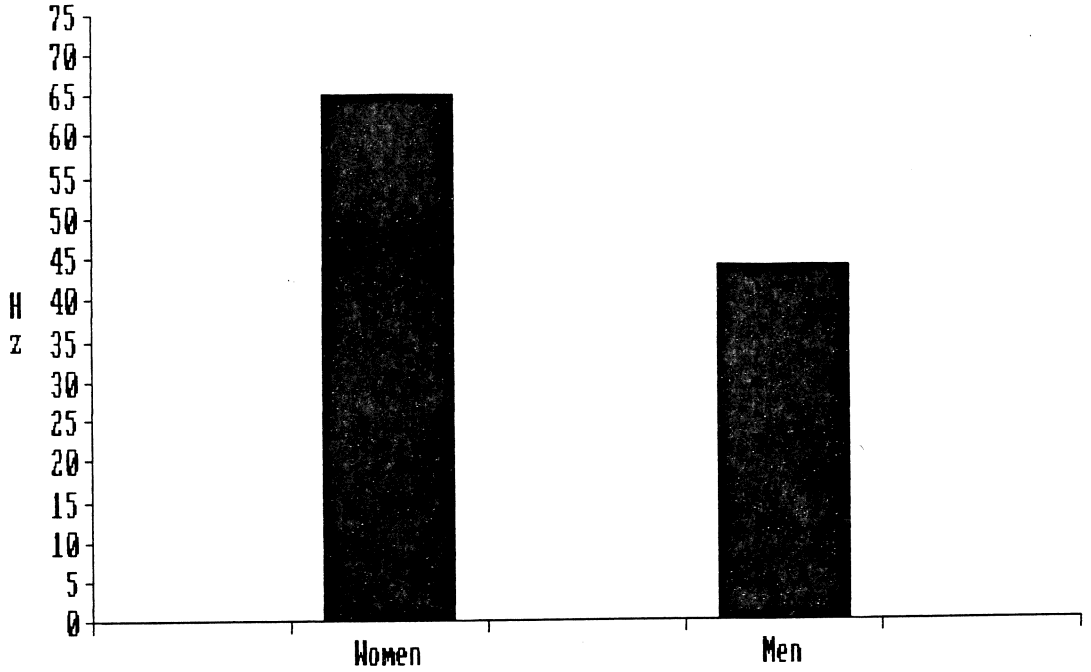
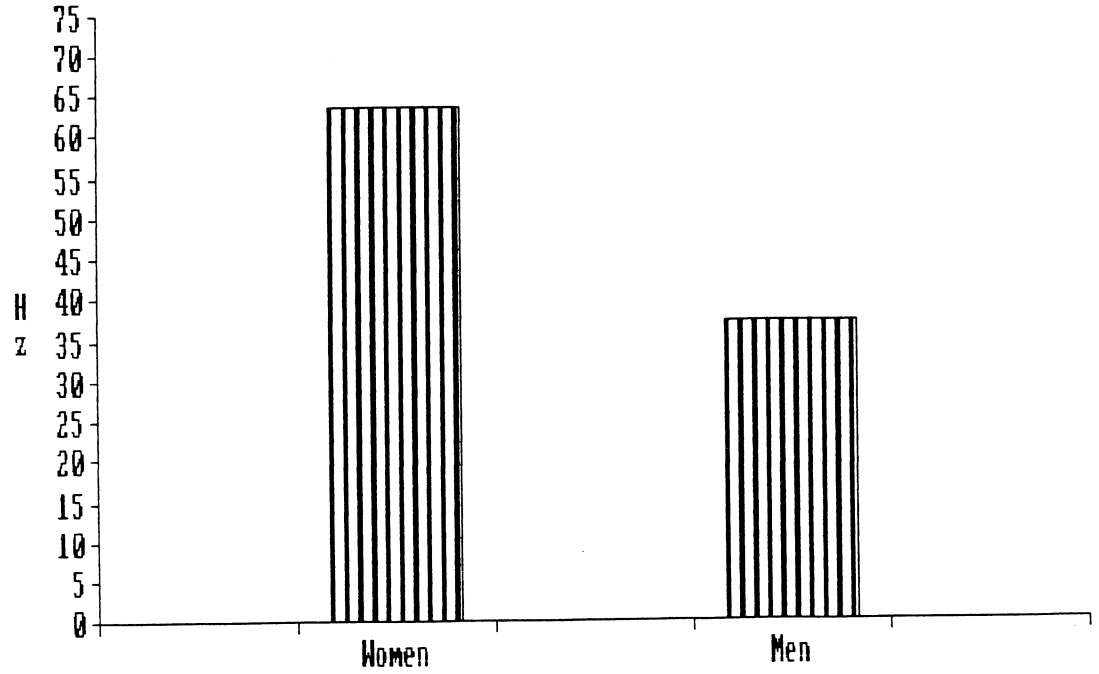


FIGURE 8

Pitch range
interview



Detailed analysis of the data showed, however, that differences in girls' and boys' use of tones was only significant in one context: that is, the interview setting. A tendency towards the same patterns of sex-related variation was observed in the conversational data, but these differences were not significant at the level set[1].

Another feature noted in children's speech was that although, in a similar sex-related pattern to that observed in adults' speech, girls used more rising tones than boys, in the case of children it was **simple** rather than complex rise tones which distinguished males from females. A hypothesis which may explain this result emerges by taking into account Local's (1982) observation that children use complex tones significantly less frequently than adults. That is, it can be suggested that girls use simple rises because they are children and female (in contrast to women's use of complex rises (adult and female)).

In sum, although patterns of sex-related variation observed in children's speech are not precisely the same as those shown by adults, nevertheless the fact that the three tones showing sex-related patterns of variation in children's speech were of the same type as those used by adults strongly suggests that the system of tone is important in children's acquisition of sex-appropriate speech styles. Further evidence which supports this hypothesis emerges from a consideration of the pitch range features which characterise children's speech.

3.3.2 Pitch

> "Boy monsters are brave and gruff. Girl monsters are high-pitched and timid" (Pogrebin on "Sesame Street", 1972)[2]

Unlike men and women, boys and girls, given approximate similarity in height, weight and body-build, have the same or at least similar larynx physiology. However, despite the lack of relevant physical differences between young girls and boys, nevertheless research has found that hearers can identify children as male and female on the presentation of voice cues alone (cf. Sachs, Lieberman and Erikson (1973), Weinberg and Bennet (1971)). The actual linguistic features which hearers use as cues to identification may not, however, be concerned with fundamental frequency. Indeed, Sacks *et al* found that although hearers showed a high success rate in the identification of children's sex from samples of their speech, the fundamental frequencies used by girls were actually lower than those employed by boys. This is, of course, the opposite of what would be expected if Fo were used as a cue in the identification of children's gender identity.

My results (on children's **production** of Fo) suggest that one of the cues which may be used in the identification of children's sex is pitch range. That is, results showed that just as women use a significantly greater part of their potential pitch range than men, girls, in the conversational encounter at least, use a far wider range of fundamental frequencies than boys[3]. Thus aspects of pitch of voice as well as tone may play an important role in children's acquisition of sex-appropriate speech styles.

The observation that features of pitch vary in the speech of children (where no significant physical differences distinguish between males and females) may suggest the social indexicality of an aspect of language use which has generally been considered to be solely a consequence of anatomy[4].

3.4 The social conditioning of pitch of voice

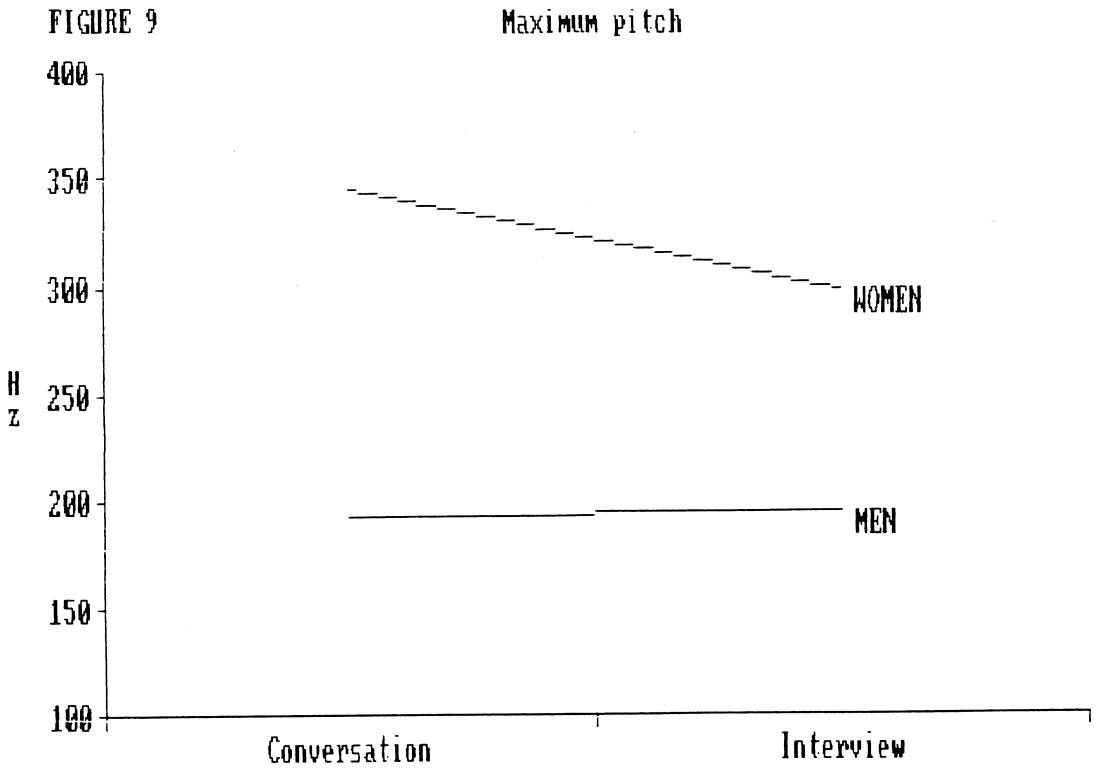
As noted earlier, pitch of voice is considered to be a product or consequence of rate of vocal fold vibration which, in turn, is conditioned by features such as the length, thickness and tenseness of the vocal folds. However, a review of studies which, though from disparate areas of research, have all addressed the question of pitch, suggests that larynx physiology fails to fully explain patterns of pitch variation. For example, in a contrastive study of pitch in Polish and English, Majewski *et al* (1972) found that Polish men spoke with a consistently higher pitch than American males. Further, Mattingly (1966) observed that the formant frequencies used by men and women could not be explained by reference to male-female differences in larynx physiology alone. Commenting on Mattingly's findings, Sacks *et al* (1973) observe that men speak as though they are bigger than they really are and women as though they are smaller.

Added to this, the hypothesis that pitch of voice is determined by physical factors cannot explain Mount and Salmon's (1988) observation that an informant who had undergone male to female sex-reassignment surgery was able to make and maintain dramatic increases in his pitch of voice: as a result of training exercises the informant achieved a raising in Fo of 85 Hz (for the vowel /i/), 100 Hz (for /a/), and 100 Hz (for /u/); the informant also achieved similar raising in formant frequency levels. Finally, an explanation of pitch of voice purely in terms of laryngeal anatomy cannot explain how Margaret Thatcher was able to decrease her pitch of voice (Atkinson reports that Thatcher achieved a decrease in her average fundamental frequency of 46 Hz) at a time in her life when physical or developmental factors would predict an increase in Fo (cf. Dordaine et al, 1969[5]).

Thus it seems fair to conclude that if, as previous research suggests, pitch of voice is language specific, and if it is possible to change one's pitch of voice at will, then it seems highly likely that pitch is conditioned by more than just larynx physiology. Other explanations of variation in pitch thus need to be explored.

[1] In the analysis of pitch range the opposite pattern was observed: differences between boys and girls were only significant in their conversational speech (discussed in detail in my doctoral thesis).
[2] Cited in Kramer (1975).
[3] Girls also used a greater part of their potential pitch range in the interview setting, but the difference between girls and boys was not significant at the level set.
[4] Although, as mentioned above, it is average fundamental frequency rather than the range of frequencies used by a speaker which is most clearly linked to larynx anatomy.
[5] Cited in Laver and Trudgill (1979).

FIGURE 9



In attempting to explain pitch more adequately a number of insights can be gained by asking the question as to why Margaret Thatcher felt the need to take training sessions designed specifically to lower her pitch of voice. An answer to this may be that, traditionally, high pitched voices have been considered to have very low prestige. As Henessee notes,

> "Higher pitched voices are still associated with unpleasantness, evoking nagging mothers or wives, waspish schoolteachers, and acerbic librarians" (1974: 243)

Further, and perhaps more significantly, high pitched voices are often considered as inappropriate for the consideration and expression of serious issues; as Mannes (1969) points out, "people don't like to hear women's voices telling them serious things"[1]. That this is not just an attitude of the past is shown by Graddol and Swann's (1989) comment that, through personal experience of dealing with media personnel, they have found that "producers are notoriously circumspect about using women's voices for "serious" work" (p.39).

Given the considered inappropriacy of women's voices for expressing serious topics, the question arises as to whether women change aspects of their voice when they wish or need to consider serious issues. I have already mentioned that the procedure used in my research was to gain access to two different styles of speech: an informal conversational encounter and a formal interview in which, by definition (in my research at least), more serious issues were the topic of discussion. A number of differences in pitch and intonational patterning were noted across these two speech encounters (see also Woods 1991). In comparing specifically the speech of women across these two situations (in order to provide some answer to the question posed above), it was initially observed that women showed no stylistic shifting in their average pitch of voice, and similarly, no situation-related variation was noted in women's use of pitch range. However, in examining the use of maximum pitches employed by women[2] (the aspect of pitch which has been most overtly and consistently stigmatized) patterns of stylistic shifting were observed. As shown in figure 9, women used significantly lower maximum Fos in the interview than in the conversational setting: specifically, women lower their maximum pitches by an average of 53 Hz (mean figure) in the interview speech encounter. Notably, no similar patterns of stylistic shifting were displayed in the speech of men. It might therefore be concluded that because of the pervasive stereotype of high pitched voices, in certain formal speech encounters where serious issues are the topic of discussion, women supress the high pitches which they use in more casual conversational styles.

This result points to a further difference in men's and women's pitch characteristics: that is, their differing tendencies towards stylistic shifting in pitch range. Furthermore, it also provides evidence to suggest that the use of certain aspects of pitch are not determined solely by physical factors, but rather are also conditioned by social influences: in this case, the stigmatization and considered inappropriacy of high pitch for discussing serious topics.

[1]Cited in Kramer (1975).
[2]This refers to the maximum pitches used by women in their ordinary speaking voices; not to the most extreme high pitches which women are able to produce.

4. Conclusion

The speech of men and women is characterised by different frequencies of intonational (specifically tonal) and pitch features: rise, level and high fall tones, as well as range of pitch, are socially indexical of speaker-sex. Further, and perhaps more significantly, the same sex-differentiating patterns observed in adults' speech are also shown in children's use of pitch and intonational features. This demonstrates the importance of non-segmental features in children's development of sex-appropriate speech styles. That pre-adolescent children show sex-marking in certain aspects of their pitch of voice may suggest a social as well as physical conditioning of Fo. This conclusion is supported by the observation that, because of the social stereotype of high pitched voices, women systematically lower their use of high frequencies in formal styles in which serious topics are discussed. It may thus be concluded that an adequate description and explanation of pitch (as well as aspects of intonation) can only be achieved by making reference to the social as well as physical factors which influence language use.

Bibliography

Atkinson, M. 1984 *Our master's voices: the language and body language of politics.* London: Methuen.
Brend, R. 1975 Male-female intonation patterns in American English. In B. Thorne and N. Henley (eds.) *Language and sex: difference and dominance.* Massachusetts: Newbury House Publishers.
Eble, C. 1972. How the speech of some is more equal than others. Annotated bibliography in Thorne and Henley, op cit.
Elyan, O. 1978 Sex differences in speech style. *Women speaking* **4**, 4-8.
Geluykens, R. 1988 On the myth of rising intonation in polar questions. *Journal of pragmatics* 12, 467-485.
Graddol, D. 1986 Discourse specific pitch behaviour. In C. Johns-Lewis, (ed.) *Intonation in discourse.* London and Sydney: Croom Helm.
Graddol, D. and Swann, J. 1983 Speaking fundamental frequency: some social and physical correlates. *Language and speech* **26**, **4**, 351 - 366.
Graddol, D. and Swann, J. 1989 *Gender voices.* Oxford: Basil Blackwell.
Halliday, M. 1966 Intonation systems in English. In A. MacIntosh and M. Halliday (eds.) *Patterns of language: papers in general, descriptive and applied linguistics.* London: Longmans.
Hennessee, J. 1974 *"Some news is good news".* Annotated bibliography in Thorne and Henley, op cit.
Johns-Lewis, C. 1986 Prosodic differentiation of Discourse Modes. In Johns-Lewis, op cit.
Kenworthy, J. 1978 The intonation of questions in one variety of Scottish English. *Lingua* **44**, 267-282.
Kramer, C. 1975 Women's speech: separate but unequal. In Thorne and Henley, op cit.
Lakoff, R. 1975 *Language and woman's place.* New York: Harper Colophon Books.
Laver, J. and Trudgill, P. 1979 Phonetic and linguistic markers in speech. In K. Scherer and H. Giles *Social Markers in Speech.* Cambridge: University press.
Local, J. 1978 Studies towards a description of the development and functioning of children's awareness of linguistic variability. Unpublished Ph D Thesis, Newcastle-upon-Tyne.
Local, J. 1982 Modelling intonational variability in Children's speech. In S. Romaine (ed.) *Sociolinguistic variation in speech communities.* London: Arnold.
Majewski, W., Hollien, H. and Zalewski, J. 1972 Speaking fundamental frequencies of Polish adult males. *Phonetica* **25**, 119-125.
Mattingly, I. 1966 Speaker variation and vocal tract size. *J. Acoust. Soc. Am.* **39**, 1219.
Mount, K., and Salmon, S. 1988 Changing the vocal characteristics of a postoperative transsexual patient: a longitudinal study. *Journal of communication disorders* **21**, 229-238.
Pellowe, J. and Jones, V. 1978 On intonational variability in Tyneside speech. In P. Trudgill (ed.) *Sociolinguistic patterns in British English.* Baltimore: University Park Press.
Sachs, J. Lieberman, P. and Erikson, D. 1973 Anatomical and cultural determinants of male and female speech. In R. Shuy and R. Fasold (eds.) *Language attitudes: current trends and perspectives.* Washington: Georgetown University Press.
Weinberg, B. and Bennett, S. 1971 Speaker sex recognition of 5 and 6 year old children's speech. *J.A.S.A.* 50, 1210-1213.
Woods, N. 1991 The Effect of Social Setting on Intonational Patterning. *Progress Reports from Oxford Phonetics* **4**, 92-101.
Woods, N. 1992 *Sociolinguistic patterns in English pitch and intonation.* Forthcoming 1992.