# Linguistic Variation
## Models and Methods

Edited by
*David Sankoff*
*Center for Mathematical Research*
*University of Montreal*
*Montreal, Quebec, Canada*

ACADEMIC PRESS New York San Francisco London
A Subsidiary of Harcourt Brace Jovanovich, Publishers

switching rates in sequences and in unconstrained successive pairs for a number of speakers. For this analysis, we deliberately chose speakers who showed at least a minimal degree of switching behavior in their sequence data. Nevertheless, as Table 1 shows, the sequencing context is much more restrictive than just simple proximity between variables. Indeed, the switching rates for unconstrained pairs are not clearly different from the overall proportion of the variants, indicating little or no effect of stylistic homogeneity. This also explains why an exhaustive examination of switching in unconstrained pairs for all speakers is of little interest.

Turning to the other variables, we again encounter a data problem. There are too few sequence-constrained contexts for switching from *ils* to *on* in our corpus to permit analysis. However, almost 30 speakers do show both enough variation in this variable and enough sequence-constrained pairs for *on* to *ils* switches to warrant the graphical representation in Figure 3. Although less dramatic than the previous two, the tendency is for the same effect to be present, restricting switching in the sequence environment.

The case of *on/nous* is somewhat more clear-cut. The data problem here is that there are less than 200 tokens of *nous* as a subject clitic in our corpus. These are, however, concentrated among a small number of conservative speakers and we can portray their switching behavior as in Figure 4. Once again the sequencing constraint is unequivocally present, though it cuts out perhaps only half of prospective switches rather than the two-thirds suggested in the case of *on/tu–vous*.

## REFERENCES

Berdan, R. The necessity of variable rules. In R. W. Fasold & R. W. Shuy (Eds.), *Analyzing variation in language*. Washington, D.C.: Georgetown University Press, 1975, 11–26.

Laberge, S. The changing distribution of indeterminate pronouns in discourse. In R. W. Shuy & A. Schnukal (Eds.), *Language use and the uses of language*. Washington, D.C.: Georgetown University Press, forthcoming.

Laberge, S. Etude de la variation des pronoms sujets définis et indéfinis dans le français parlé à Montréal. Unpublished Ph.D. dissertation, Université de Montréal, 1977.

Labov, W. The study of language in its social context. *Studium Generale 23,* 1970, 30–87. Reprinted in W. Labov, *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press, 1972, 183–259.

9

# *Modeling of Duration Patterns in Reiterant Speech*

Mark Liberman

## INTRODUCTION

This chapter reports on some preliminary attempts at quantitative modeling of duration patterns in English. The intent is to demonstrate the feasibility and interest of such modeling, rather than to present a particular model as a finished product. I will begin by suggesting why duration data is an especially appealing candidate for quantitative modeling, and will describe then the particular body of data that was used, the modeling method, and some of the results.

### Why Speech Timing Is Interesting

Patterns of duration in speech are quite reproducible. If the same speaker repeats the same utterance a number of times, without changing the stress pattern or the intonation, the durations of comparable segments or syllables (for those aspects of the speech signal which have conveniently measurable durations) typically have quite small variances: Standard deviations of 10 and 15 msec are fairly typical. This remains true, in my experience at least, even when there are fairly long periods of time between repetitions. Large variations in overall rate increase the variance somewhat, and variations across speakers also blur the sharpness of the patterns, but there remains substantial agreement.

An understanding of such duration patterns, and of the processes which generate them, is of great interest both practically and theoretically. Practical applications include speech synthesis by rule and speech recognition, for which any source of information describing reliable connections between language and sound is useful. The theoretical interest of this area of study has several aspects, of which I will mention three. First, linguistic constructs such as stress pattern, syllabic structure, surface constituent structure, and so on, affect timing, and an understanding of these relationships provides evidence which can help to choose among alternative linguistic theories. The relative stability of duration patterns, previously mentioned, means that in principle a substantial amount of information is available for such efforts. Second, there is reason to suppose that much of the information present in patterns of timing is used by the perceptual system. This raises, in a very pointed way, the perceptual problem posed by the dynamic aspects of speech, the problem of how our perceptions orient themselves amid the incoming stream of acoustic information, not only "normalizing" the variable dynamics of acoustic cues, but actually using this variation to provide information of value to the decoding process. Third, speech rhythms (exemplified for present purposes in patterns of duration) can be studied in relation to the problem of rhythmic organization in other human activities, notably music.

## Some Problems

Although the study of speech timing is interesting and important for the reasons just mentioned, it faces a number of very substantial difficulties. Two classes of these difficulties are especially worthy of note: the number and nature of factors influencing duration, and the arbitrariness of duration measurements.

There are a large number of factors known to influence duration in speech. A nonexhaustive list includes the nature of the segment in question, the local segmental environment, syllabification, stress pattern, constituent structure, and intonation contour. Each of these variables has a large number of possible values. It is clearly not feasible to vary all of them orthogonally—life is too short. Indeed, many of them are nonorthogonal by nature. The usual practice is to pick some feature (e.g., position in the word) and vary it systematically, while trying to obtain some reasonable sampling of values for other variables. A lot has been learned by such techniques; however, in averaging across categories, much of the precision of temporal control is thrown away.

It has long been known that linguistic phonetic elements do not correspond to discrete portions of the acoustic signal, but rather produce a complex pattern of overlapping acoustic effects. Thus it is in some very real sense meaningless to talk about the **duration** of phonetic elements in speech. When phoneticians use this somewhat loose way of talking, they refer to the fact that in many cases there are local discontinuities in the acoustic signal which can be taken to specify the boundaries of **something;** generally these are points of closure and release, points of voicing onset or offset, the beginning and end of turbulence, and so on. Such points can often be measured with an accuracy of 5 msec or better; it is by reference to such points that we discuss **segment durations** or **syllable durations.** But sometimes there are several such measurable points in close succession—in the case of aspirated stops, for example, the interval of aspiration could be assigned to the stop, to the vowel, or to neither (with three different results for the body of data a theory of speech timing is asked to describe). Furthermore, we cannot even be sure that it is by reference to these apparent discontinuities in the acoustic signal that human beings reckon duration. In fact, there is good reason to suppose that this may not be the case—segments such as [y] or [r] do not create any such discontinuities, but duration cues do not seem to be obscured in utterances that happen to contain such segments. So from a theoretical point of view there is a very serious amount of arbitrariness in any decision about how to interpret duration measurements, even for those cases where measurements are possible.

## A Solution

In order to get around such difficulties, phoneticians have traditionally resorted to nonsense. The advantages of nonsense syllable strings over (more or less) natural speech are obvious—segmental variables can be eliminated from the model, to whatever extent is desired, and the necessary arbitrariness of measurement criteria can be minimized by reducing the number of boundary types to one or two. Of course, the usual sort of nonsense **words** will not do for the study of phrase-level prosodic effects, so Lynn Streeter and I developed[1] the idea of mimicking a natural utterance while substituting some nonsense syllable, such as *ma,* for each syllable of the original.

This technique, for which Nakatani and Schaffer (1976) have suggested

[1] Actually, we redeveloped this idea, which was previously used by various Swedish researchers and reported in (apparently) unpublished manuscripts referred to in Lindblom and Rapp (1973).

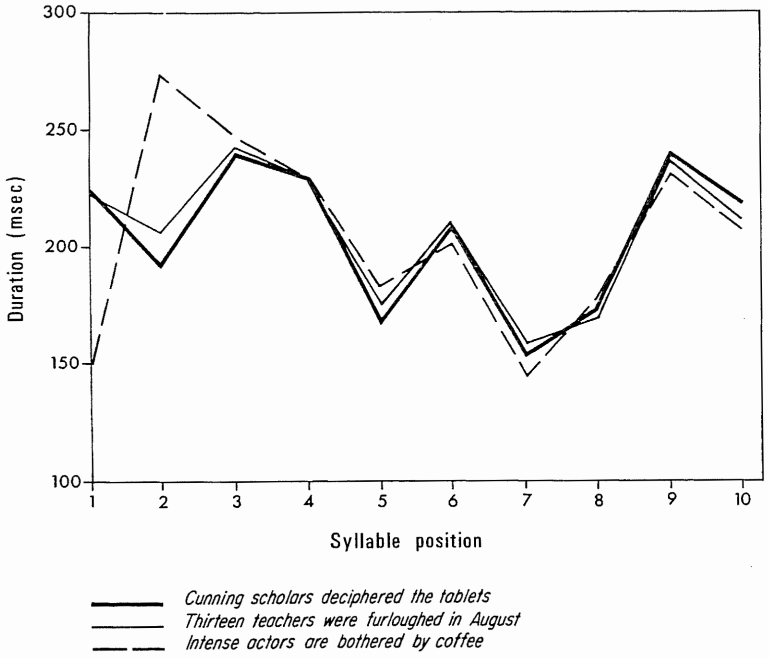


**Figure 1.** Syllable duration in reiterant speech imitations of three sentences.

the term **reiterant speech (RS)**, is described and justified in some detail in Liberman and Streeter (1976). The relevant findings of that study can be summarized as follows: (*a*) The reproducibility of durations in reiterant speech is comparable to what is found in natural speech; and (*b*) utterances with the same stress pattern and constituent structure produce nearly indistinguishable reiterant speech durational patterns, even when the durations of the originals are very dissimilar due to segmental effects.

Figure 1, taken from Liberman and Streeter (1976), presents syllable[2] durations in *mama* imitations of three target sentences:

(1) *Cunning scholars deciphered the tablets.*

(2) *Thirteen teachers were furloughed in August.*

(3) *Intense actors are bothered by coffee.*

[2] For simplicity of exposition, we will use syllable (rather than segment) durations throughout this paper. In some ways segment durations are more interesting, but the issues involved are not relevant to this paper.

Note that the first two cases, in which stress and constituent structure are nearly the same, have very similar duration patterns, while the third case, in which the first word has a different location of main word stress, differs greatly in the first two positions, but remains quite similar to the previous cases in positions 2 through 10.

## MATERIALS AND METHODS

It appears that different utterances with the same prosodic structure (stress and constituent structure) have the same reiterant speech timing pattern. It follows that we should, in principle, be able to predict such RS timing patterns as a function of the prosodic structure of the target utterance. In order to attempt such a prediction, we need three things: (*a*) RS duration data for some set of utterances; (*b*) a precise definition of the notion **prosodic structure** for those utterances; and (*c*) some assumption about the function which maps prosodic structure, as defined in (*b*), into durations, as measured in (*a*).

### Data

The modeling described in this chapter is based on a body of data collected by Lynn Streeter and myself for other purposes. It consists of 20 utterances (and their RS imitations) spoken at least 10 times each by one speaker, and 17 utterances (and their RS imitations) spoken at least 10 times by a second speaker. The data from the second speaker was collected in two sessions about 6 months apart; the first speaker's data was collected in three sessions. (Some of this same material was used in Liberman & Streeter, 1976.) In each recording session, the speaker being recorded read a target sentence from a card, using a normal speaking rate and intonation pattern, and then after a suitable pause, imitated the target sentence by substituting a [ma] for each of its syllables, while attempting to preserve the original rhythm and intonation. After all the utterances in the experimental set had been produced, the cards were shuffled and the process repeated a total of 10 times to obtain the 10 tokens of each target sentence and each *mama* imitation to be averaged.

Durations in the target utterances were measured by means of a computer wave-form editor. Duration of the RS versions were measured automatically by a computer pattern-recognition technique, based on the voice/unvoiced/silence decision algorithm described in Atal and Rabiner (1976), and modified to decide among the three categories [m], [a], and silence.

### Prosodic Feature Set

Nine binary features were chosen as a means of encoding prosodic structure. These features cover three general areas: stress, boundary location, and rhythmic grouping. Binary features were used in order to permit the model to take a maximally simple form, as described in the following subsection.

Before being coded in terms of these features, an utterance is divided into **feet,** which generally run from main word stress to main word stress. Normally, then, foot boundaries are inserted in front of the main stress of every content word, and nowhere else. There are two exceptions: (*a*) In sequences of monosyllables such as *new blue boat,* where the stress pattern would be classically described as 2 3 1, the medial monosyllable is not given as a separate foot boundary, resulting in the division | *new blue* |*boat;* and (*b*) a function word which is a stress maximum is taken to begin a foot, as in *the* |*cat is* |*on the* |*mat*. Also, it is assumed that feet are interrupted by major phrase boundaries.

Our nine prosodic features can now be described as follows:

1. *Stress*. This feature is assigned to every stressed syllable.
2. *Main foot stress*. This feature is assigned to the stressed syllable at the beginning of each foot.
3. *Main phrase stress*. This feature is assigned to stressed syllables which have a major pitch accent. Typically there is one such syllable in each phrase. Obviously, every main phrase stress must also be a main foot stress.
4. *End of word*. This feature is assigned to the last syllable of each lexical word.
5. *End of phrase*. This feature is assigned to the last syllable of each major phrase (where there is a noticeable pause or pseudopause). All but one of the utterances in the data set used for this chapter consisted of exactly two phrases.
6. *Start of trochee*. This feature is assigned to those main foot stresses which are followed by at least one syllable within the same foot. Since two or more following syllables are also consistent with this feature, the word trochee must be taken in a loose sense.
7. *Start of dactyl*. This feature is assigned to those main foot stresses which are followed by at least two syllables within the same foot.
8. *Second position in dactyl*. This feature is assigned to the syllable immediately following the main foot stress of a dactylic foot.
9. *Third position in dactyl*. This feature is assigned to the last syllable in a dactylic foot.

Features 1–3 encode information about stress level; features 4 and 5 encode boundary information; features 6–9 provide information about rhythmic grouping. These features were chosen with two ends in view: (*a*) to include some representation of factors known or alleged to influence duration; and (*b*) to be definable with minimum opportunity for disagreement regarding their values in a particular case. Note that no theoretical or practical validity is being claimed for the details of this feature set—it is simply a conveniently definable set of features, which is likely to be highly correlated with whatever the "true" feature set is.

### Assumptions of the Model

For a first attempt, I adopted a rather simple view of the function mapping prosodic feature specifications into durations. We assume $n$ well-defined prosodic features. Each element in the data set (syllable, segment, or whatever) is marked either + or − for each such feature. Then the predicted duration for a given syllable or segment is determined by adding, to a fixed base duration, a fixed quantity (which can be positive or negative) for each prosodic feature which is present.[3] Thus the influences of the various features are assumed to be additive and independent. Symbolically, we assume that

$$D = B + a_1 f_1 + a_2 f_2 + \cdots + a_n f_n \qquad (1)$$

where $D$ is the predicted total duration of a given segment or syllable; $B$ is the (invariant) base duration; $a_i$ is 1 if the $i$th feature is present, 0 if the $i$th feature is absent; and $f_i$ is the durational increment (plus or minus) attributed to the $i$th feature.

Since we have a list of observed durations, and specification of a prosodic feature vector for each one, multiple linear regression will give us values for the base duration and feature-associated durations which minimize the (squared) prediction error.

## RESULTS

In order to see how successful this procedure is at predicting novel data, the data set for each of the two subjects was divided into two subsets

[3] Of course, when different phonetic elements are examined, a different base duration and set of feature-associated durations is assumed in each case. One would like the feature-associated duration values to be predictable, at least for a given type of phonetic element, either by being invariant or by being some function of that element's base duration. Since the only phonetic element considered in this chapter is the syllable [ma], we avoid such questions for the present.

of 10 utterances each in the case of the first speaker, and of 8 and 9 utterances in the case of the second speaker. Since the average number of syllables per utterance was 10, each subset contained about 100 syllables. The parameter values resulting from regression on one of the data subsets for each of the speakers are given below:

(2)

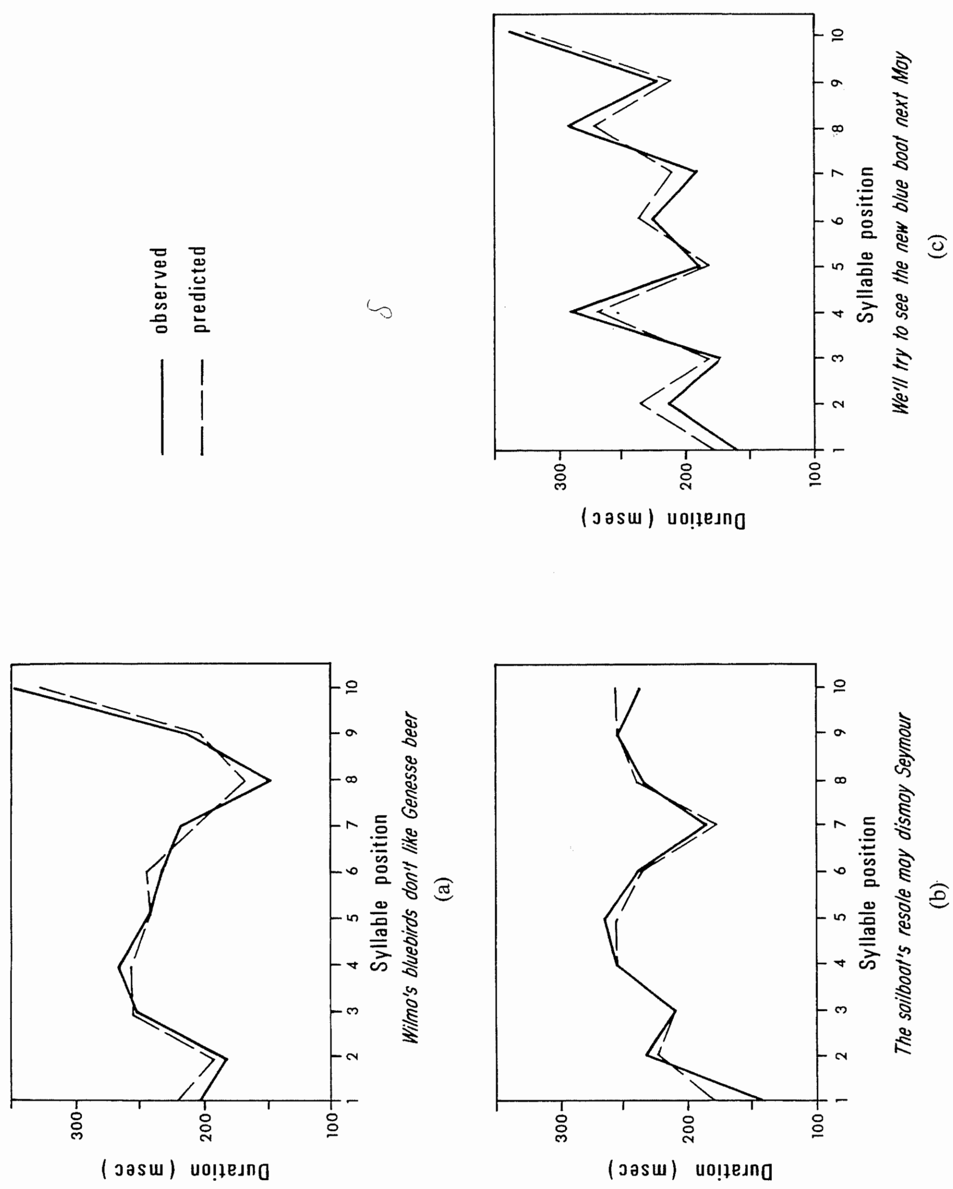
| Features: | Base | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Subject A: | 179 | 18 | 33 | 34 | 13 | 48 | −7 | −15 | −12 | 10 |
| Subject B: | 174 | 27 | 30 | 22 | 16 | 55 | 5 | −22 | −21 | 8 |

Figure 2 shows some superimposed plots of predicted versus observed durations for three utterances, using the parameter values given for Subject A in (2) and taking the observed durations from the same data subset that was used to generate those parameters. These cases were chosen as fairly typical of their kind. For the data subset in question, the percentage of variance accounted for by the model was 88%, and the mean absolute deviation (average of the absolute value of prediction versus observation) was 11 msec. Percentage of variance accounted for is not a very meaningful measure in this kind of situation, but it will serve for rough comparison with the other conditions described below.

In the data subset for which Subject B's parameters were obtained, the percentage of variance accounted for was 92%, and the mean absolute deviation was 9 msec. Because of the nonorthogonality of the feature set, it is difficult to compute the significance of these results, but we can offer for comparison the results of running the same regression, with the same prosodic feature matrix, using durations from the original target utterances, or using the set of RS durations randomly permuted. Using the target utterance durations, 63–68% of the variance was accounted for by the model, with mean absolute deviations of 38–44 msec. Using randomly permuted RS durations, approximately 5–15% of the variance was accounted for.

Figure 3 shows a similar set of plots for the case in which parameters derived from the first subset of data were employed to predict durations in the second subset. For the second subset as a whole, duration prediction on the basis of parameters derived from the first subset accounted for 85% of the variance in the case of one subject, and 82% in the case of the second subject. The mean absolute deviations were 12 and 18 msec, respectively.

It is worth noting that the data subsets were not selected randomly, but rather tended to respect the boundaries of different data collection ses-

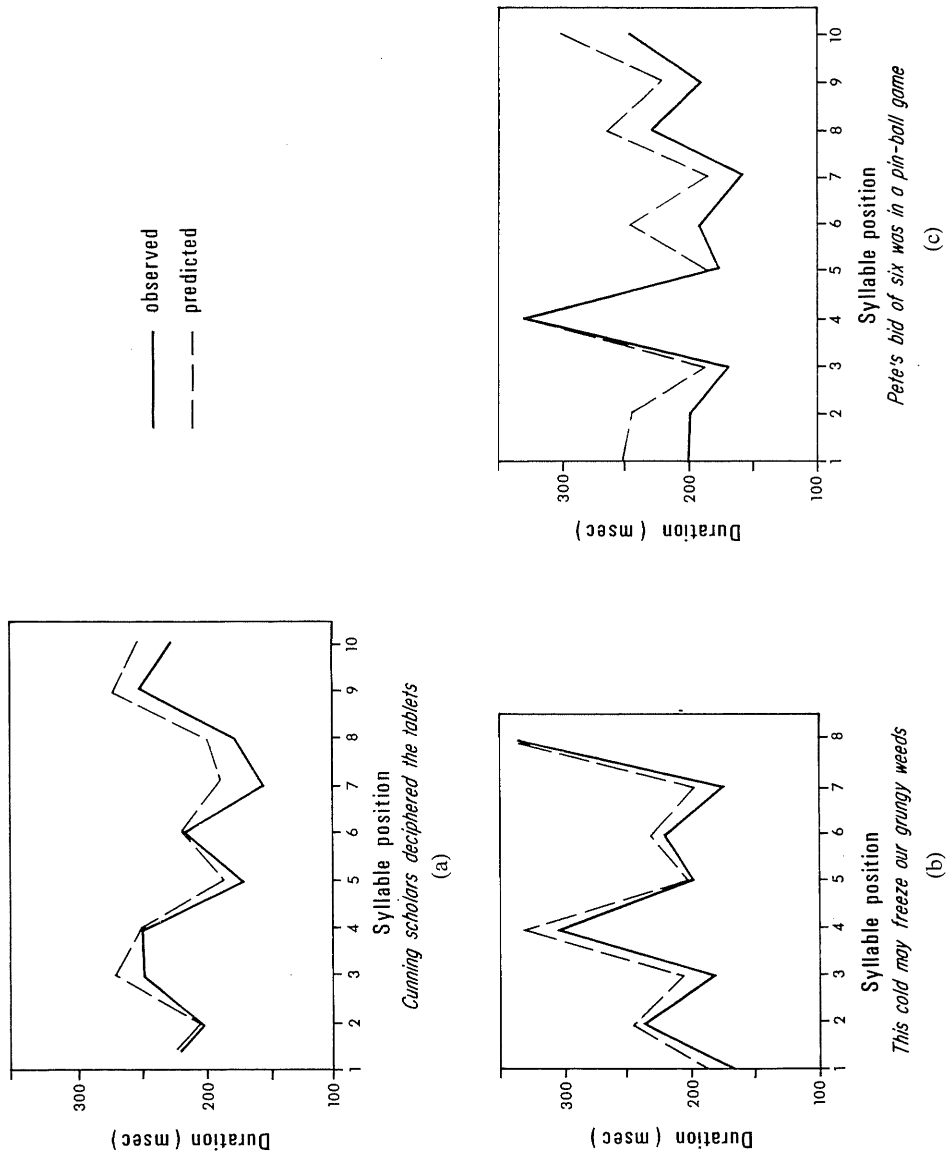


*Wilmo's bluebirds don't like Genesse beer*
(a)

*The sailboat's resole may dismay Seymour*
(b)

*We'll try to see the new blue boat next May*
(c)

Figure 2. Prediction within a data set



*Cunning scholars deciphered the tablets*

(a)

*This cold may freeze our grungy weeds*

(b)

*Pete's bid of six was in a pin-ball game*

(c)

**Figure 3.** Prediction across data sets.

sions. In the case of the data represented in Figure 3, the speaker in question spoke somewhat faster in the session represented in the second data subset, an effect which is especially noticeable in Figure 3(c). This last case has the highest mean absolute deviation of any of the across subset predictions (nearly 30 msec).

To help clarify the modeling method, the duration prediction for the utterance in Figure 3(a), *cunning scholars deciphered the tablets,* is given in Table 1.

## DISCUSSION

When I began this study, I viewed it as a *reductio ad absurdum,* a trial of a model which was so simpleminded that it had no hope of any substantial success, but which might provide some lessons in the details of its failure. I remain convinced that the feature set employed is inadequate for the general case. Specifically, it seems unlikely that only two degrees of boundary strength are sufficient, and perhaps in general one should regard prosodic features (stress, boundary, and rhythmic group-

TABLE 1

Duration Prediction for *Cunning Scholars Deciphered the Tablets*

| | Syllable position | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parameters of the model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Base | 179 | 179 | 179 | 179 | 179 | 179 | 179 | 179 | 179 | 179 |
| 1 | 18 | | 18 | | | 18 | | | 18 | |
| 2 | 33 | | 33 | | | 33 | | | 33 | |
| 3 | | | 34 | | | | | | 34 | |
| 4 | -7 | | -7 | | | -7 | | | -7 | |
| 5 | | 13 | | 13 | | | 13 | | | 13 |
| 6 | | | | 48 | | | | | | 48 |
| 7 | | | | | | -15 | | | | |
| 8 | | | | | | | -12 | | | |
| 9 | | | | | | | | 10 | | |
| Total | 223 | 192 | 257 | 240 | 179 | 208 | 180 | 189 | 257 | 240 |
| Observed | 224 | 192 | 240 | 238 | 167 | 209 | 151 | 172 | 241 | 217 |

ing) as being hierarchically defined, along the lines suggested in Liberman (1975) and Liberman and Prince (1977).

Although the present study shows little evidence of any interaction among the various prosodic features employed (e.g., different values of phrase boundary for stressed and unstressed syllables), it is hard to believe that such interactions do not exist. Furthermore, when different segments are mixed together, as in natural speech, interactions between segmental and prosodic effects may well arise which would complicate the model even for the treatment of a specific syllable such as [ma]. Finally, it is possible that reiterant speech itself produces a prosodically unnatural (though lawful) style of speech, whose spurious regularities are lacking in more normal linguistic activity. Even if this were true, however, the laws governing reiterant speech would retain some theoretical interest, since they must somehow be generated by more ordinary linguistic knowledge and skills.

There are various other ways of looking at the results of this modeling attempt, and many other questions about such modeling in general, which will not be discussed here. The main point of this chapter is simply that the prognosis is favorable—even the simplest kind of model, assuming that it embodies a sensible set of features, appears to have substantial predictive value.

## REFERENCES

Atal, B. S., & Rabiner, L. R. A pattern-recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* (Vol. ASSP-24), 1976, 201–212.

Liberman, M. The intonational system of English. Unpublished Ph.D. dissertation, Massachusetts Institute of Technology, 1975.

Liberman, M., & Prince, A. On stress and linguistic rhythm. *Linguistic Inquiry* 1977, *8*, (2), 249–336.

Liberman, M., & Streeter, L. Use of nonsense-syllable mimicry in the study of prosodic phenomena. Talk given at the 92nd meeting of the Acoustical Society of America, San Diego, November 1976.

Lindblom, B., & Rapp, K. *Some temporal regularities of spoken Swedish.* Publication No. 21. Institute of Linguistics, University of Stockholm, 1973.

Nakatani, L. H., & Schaffer, J. A. Hearing "words" without words: Speech prosody and word perception. Talk given at the 92nd meeting of the Acoustical Society of America, San Diego, November 1976.

# 10

# *Cross-Language Study of Tone Perception*

Jackson T. Gandour / Richard Harshman

## INTRODUCTION

One of the aims of modern linguistic theory is to develop a set of linguistic-phonetic features that are universally applicable to all languages of the world. The precise number and nature of the features that should be included in this universal set are still very much in dispute. The present study represents an attempt to bring fresh experimental data to bear on the number and nature of phonetic features or dimensions related to **tone** (see Wang, 1967); to determine the dimensions underlying the perception of tone, and also the degree to which an individual's language background influences his tonal perception.

Multidimensional scaling turns out to be a useful tool for measuring human perception in the tonal domain. Briefly, multidimensional scaling procedures spatially represent the underlying structure of a matrix of data values that generally correspond to subjective distances between stimulus objects (**stimulus space**), based on judgments of different individuals. Individual differences multidimensional scaling procedures (PARAFAC: Harshman, 1970; INDSCAL: Carroll & Chang, 1970) additionally provide information about the relative importance of each dimension to every individual (**subject space**). This information about the weights of the di-