Proceedings of the IRCS Workshop on Prosody in Natural Speech, August 5-12 1992

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

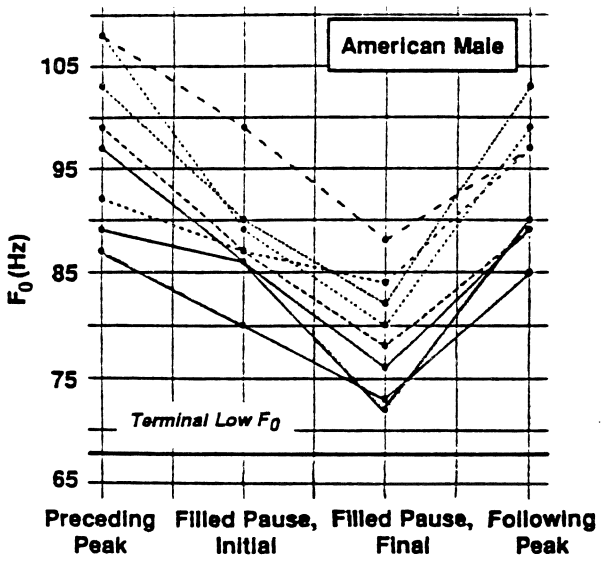

Figure 1: Peak and Filled-Pause F0 for American Male

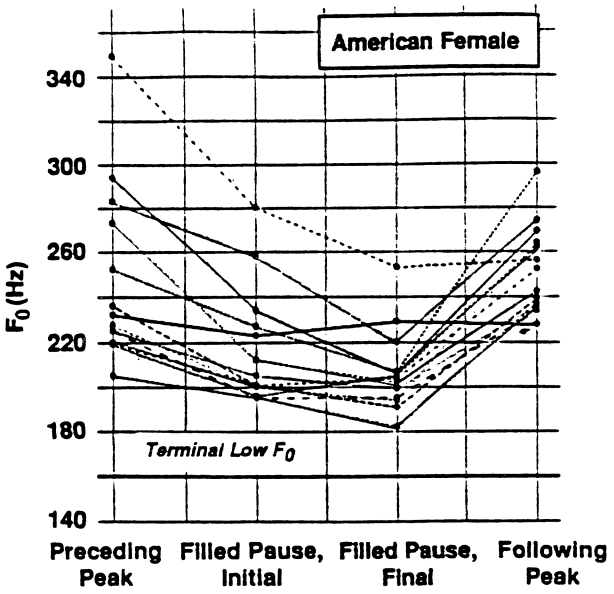

Figure 3: Peak and Filled-Pause F0 for American Female

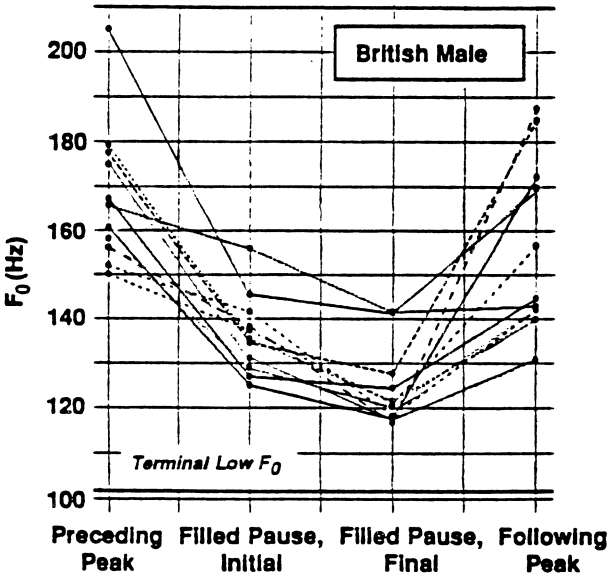

Figure 2: Peak and Filled-Pause F0 for British Male

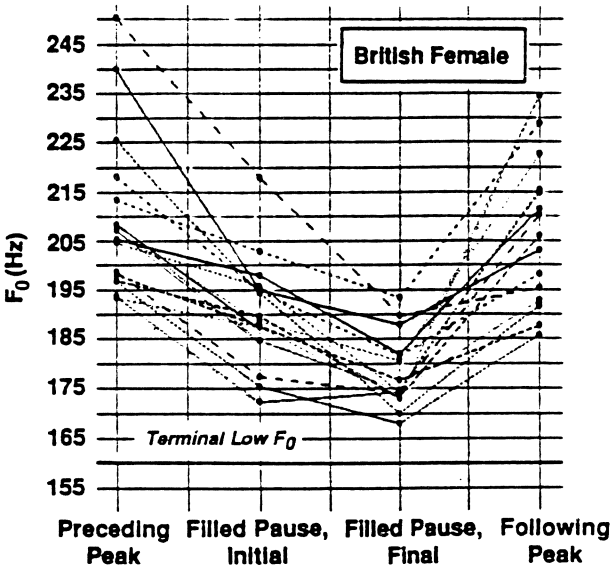

Figure 4: Peak and Filled-Pause F0 for British Female

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.

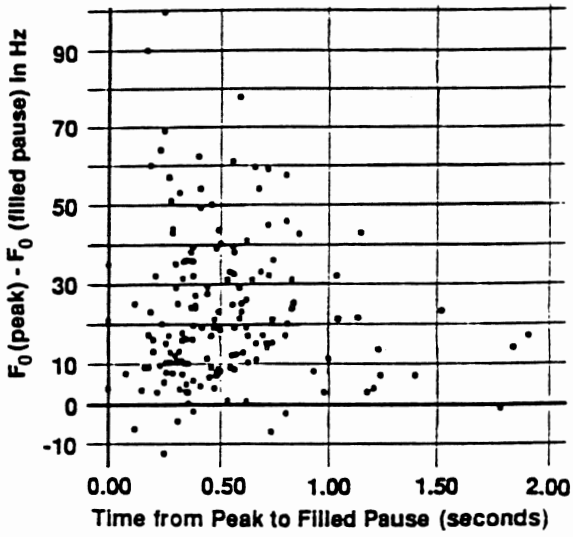

Figure 5: Effect of Time from Peak on F0 Drop

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.

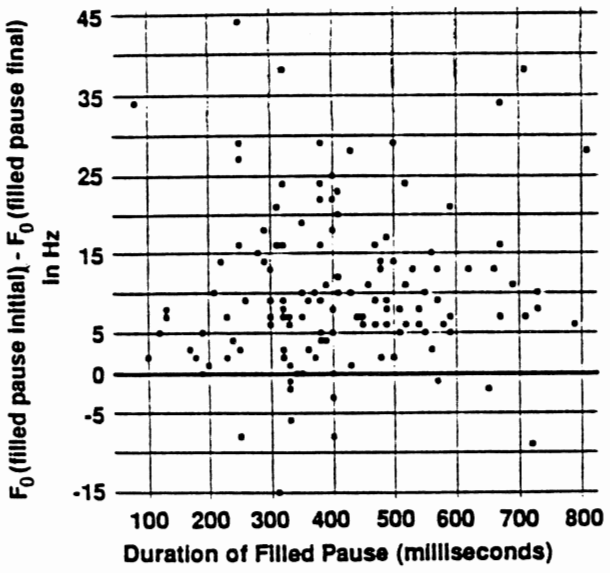

Figure 6: Effect of Filled-Pause Duration on Filled-Pause Fall

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