UTTERANCE-INTERNAL PROSODIC BOUNDARIES

Kim E. A. Silverman

Department of Linguistics Research
AT&T Bell Laboratories

ABSTRACT — This paper investigates minor prosodic boundaries that often occur in fluent speech, and yet are not well understood. A corpus was collected of utterances with a range of segmental structures, where each utterance was spoken both with and without such an internal boundary. Acoustic measurements of the utterances were then related to perceptual ratings of the salience of the boundaries. Results showed that the F0 fall from the preceding pitch accent is much steeper before a boundary, and while these boundaries do not contain pauses they do alter the temporal structure of the speech. The segmental material is lengthened, and the preceding F0 accent occurs considerably earlier relative to its accent-bearing syllable.

One of the more fundamental and recalcitrant problems in modelling speech intonation is the lack of agreement concerning prosodic phrasing. Some major phrase boundaries, such as those that would often be transcribed as full stops or commas, are so obvious as to appear in all descriptions of English intonation. They typically are accompanied by a pause, by lengthening of the segmental material, and by characteristic patterns in the pitch contour. However many utterances seem to also contain more minor boundaries which are less disruptive to the flow of speech. Linguistic models of English intonation differ markedly on how to classify these utterance-internal boundaries (cf de Pijper, 1983; Beckman and Pierrehumbert, 1986; Ladd, 1986), and their detailed phonetic form is not well understood. Yet their presence or absence can be shown to convey important distinctions in meaning (e.g. Hirschberg and Pierrehumbert, 1986). Indeed the speech synthesis community has found they make a crucially important contribution to the quality and naturalness of synthetic speech (Silverman, 1987). It is these minor boundaries that are the topic of this paper.

Figure 1a shows the digitised waveform and extracted fundamental frequency (F0) contour for the sentence *My Mama lives in Memphis* as it might be spoken in answer to the question *How did you decide to take your vacation in Tennessee?* The words *Mama* and *Memphis* both have local F0 maxima associated with them — in the terminology of Pierrehumbert (1980) these are **H*** pitch accents — which serve to give the words a particular salience and discourse function. Figure 1b shows the same sentence as it might be spoken in answer to a question like *Are you going to visit your Mama while you're in Nashville?* Again the same two words have **H*** pitch accents, but this time the fall from the first accent peak is much gentler, being an almost linear interpolation along to the start of the rise to the second accent peak. The difference between these two utterances is that the first of them contains an utterance-internal prosodic boundary at the right-hand edge of *Mama*, while the second does not.

AIMS

Because of the practical importance as well as the theoretical significance of these boundaries, the current research aimed to [1] gather a controlled corpus of utterances in which the syllable structure, segmental context, and word structure were systematically varied, and in which each utterance was spoken both with and without an internal minor prosodic boundary; [2] investigate whether listeners would agree concerning the presence or absence of a minor boundary within each utterance, and [3] develop a quantitative model of the acoustic/phonetic characteristics of these boundaries. In the light of some recent research results concerning the coordination between words and melody in speech (Steele, 1986; Silverman and Pierrehumbert, 1988), the possible influence of minor boundaries on the temporal alignment between F0 and the segmental structure was of particular interest.

METHOD

Two adult male native speakers of English (KB, American; KS, Australian) each recorded multiple versions of the sentence *My _____ lives in Memphis*, where the second word was *Ma, Mum, Mama,*

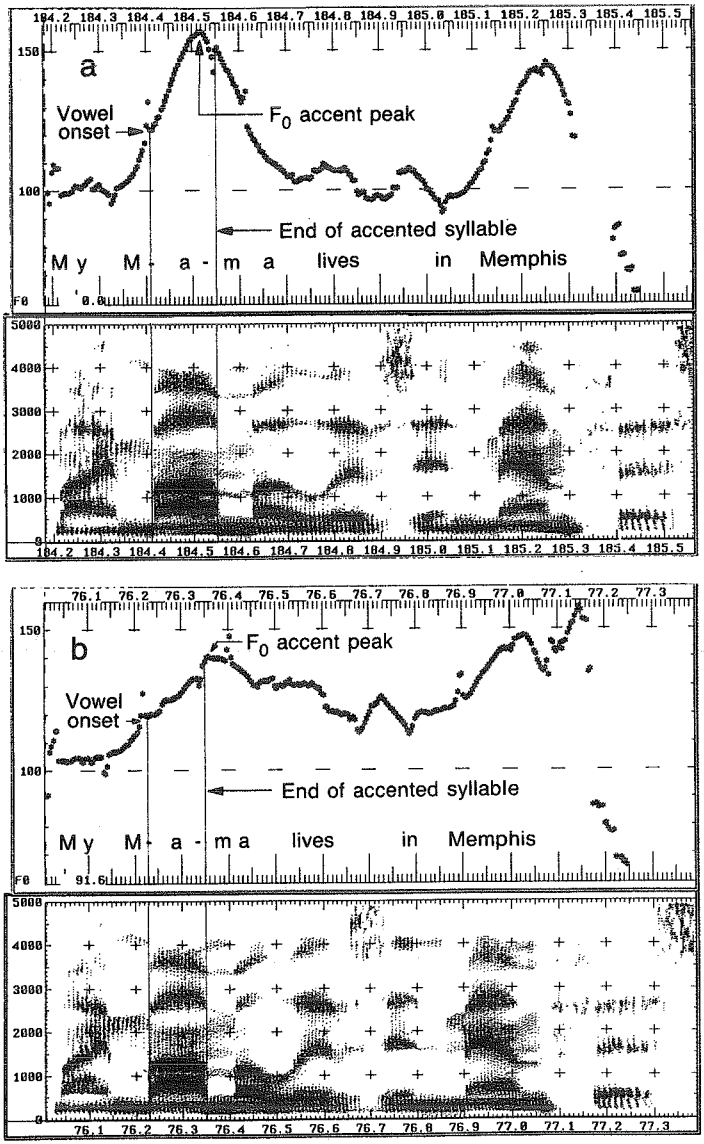

Figure 1. F_0 contour and wide-band spectrogram for *My Mama lives in Memphis* spoken with (1a, top) and without (1b, bottom) a minor prosodic boundary after *Mama*. Time is in seconds, F0 in Hertz.

Pa, Pop, and Papa. (In American English, all six words have the same vowel in the stressed syllable). Speaking rate (slow, normal, fast) and overall pitch range (high, normal, low) were varied in order to reveal invariant relationships more clearly, yielding a total of 432 utterances. The first three target words (containing /m/) were chosen to minimise segmental perturbations of the F0 contour, while the second three (containing /p/) were included to test the generality of the results in the presence of different consonant types. Comparison of *Ma* and *Pa* with *Mum* and *Pop* would show the difference between open and closed syllables, and comparison of these four words with *Mama* and *Papa* would yield information about the effect of an intervening unstressed syllable between the pre-boundary pitch accent and the boundary itself. The current paper reports on the initial analyses for the 214 utterances spoken by speaker KB.

For each utterance, the segmental durations and F0 contour were measured with the aid of interactive high-resolution graphics displays of the digitised acoustic waveform, automatically-extracted F0 values, and wide-band spectrograms.. In parallel with this, 15 listeners were asked to rate on a 5-point scale whether or not each utterance contained a prosodic boundary after the first noun phrase. They were played examples of clear cases of the contrast on sentences containing *Mama* and *Papa* in each voice range (Figures 1a and 1b are the normal-range examples), and were also given the two alternative preceding contexts.

RESULTS AND DISCUSSION

An intercorrelation matrix showed that all of the listeners correlated positively and significantly with each other across the utterances; the lowest correlation coefficient (r) between any two listeners was 0.213, and the highest was 0.794 (to be statistically significant at $p < 0.01$ on a two-tailed test, $r_{(212)}$ must be greater than 0.176). In other words, the listeners agreed with each other to a very high degree. A principal components analysis of the judgements confirmed this agreement, and showed that there was a single main dimension underlying the judgements which explained 58% of the total variance. All listeners had positive and almost equal weighting on this dimension. The second-largest dimension only explained a further 7% of the variance, and subsequent dimensions even less. On the basis of this result, each utterance's score on the main dimension was used as the best estimate of the strength of its prosodic boundary.

An attempt was then made to predict the perceived boundary strength score for each utterance on the basis of the measured segmental durations and F0 contour, via the use of stepwise multiple regression. The first thing to point out is that not one of the utterance-internal boundaries was accompanied by a pause; all of the utterances were continuous and fluent. But a number of aspects of the temporal and pitch structure of the utterances did systematically vary in relation to the perceived boundary strength. Perhaps not surprisingly, the single variable which carried the most predictive power by itself was the extent of the F0 fall after the first accent peak ($r_{(212)} = 0.895$, $p \cong 0.0$). This variable, henceforth called **F0 drop**, accounted for 80% of the variance in the perceived boundary strengths. It was calculated as (height of first F0 peak - F0 in the middle of *lives*)/(height of first F0 peak + 163). The addition of 163 in the denominator was arrived at empirically — it serves to give extra weight to the F0 fall in the higher voice ranges. For the utterance in Figure 1a, the value of F0 drop is 0.160, while for 1b it is only 0.035.

Although **F0 drop** by itself was such a good predictor of the perceptual scores, the prediction could be still further improved on the basis of other information about the acoustic structure of the utterances. Two more variables each significantly increased the amount of variance explained. The first of these, **peak proportion**, alone accounted for 28% of the total variance ($r_{(212)} = 0.895$, $p. \cong 0.0$), and it accounted for 27% of the residual variance unexplained by **F0 drop** ($t_{(211)} = -8.742$, $p. \cong 0.0$). This variable represents how the F0 peak of each **H*** pitch accent is aligned relative to the syllable bearing that accent, and it is calculated by dividing the time from the vowel onset to the accent peak by the time from the vowel onset to the end of the syllable (see Figure 1). In the utterance in Figure 1a, **peak proportion** has a value of 0.79, while in 1b it is 1.05. What this means is that F0 peaks are normally located near the end of their associated syllable — in fact a little past the end of it when the syllable is not word-final, as in 1b. But when there is a prosodic boundary to the right of the accented word, as in 1a, the accent peak occurs proportionally earlier in the syllable. It is as if the boundary "pushes" the preceding pitch accent somewhat to the left.

The second variable that improved the prediction of the perceptual boundary scores, **duration**, was calculated by dividing the duration of the target word (*Ma, Mum* etc) by the duration from the start of *lives* to the end of the *Mem* in *Memphis*. It can be thought of as the length of the relevant word, normalised for speaking rate. In Figure 1a, it has a value of 0.646, while in 1b it is only 0.514. Thus words are longer when they are followed by a prosodic boundary. By itself, however, **duration** is not particularly meaningful, because bisyllabic words (*Mama* and *Papa*) are longer than monosyllabic words regardless of whether or not they precede an utterance-internal boundary(1). It is possible to control for this in the regression, though, with a binary variable that was set to 1 for utterances containing *Mama* and *Papa*, and to 0 for all others. In conjunction with this variable, **duration** accounted for 23% of the total variance in the judgements ($r_{(212)} = 0.482$, $p \equiv 0.0$), and for 36% of the residual variance unexplained by **F0 drop** ($F_{(2,210)} = 38.131$, $p \equiv 0.0$).

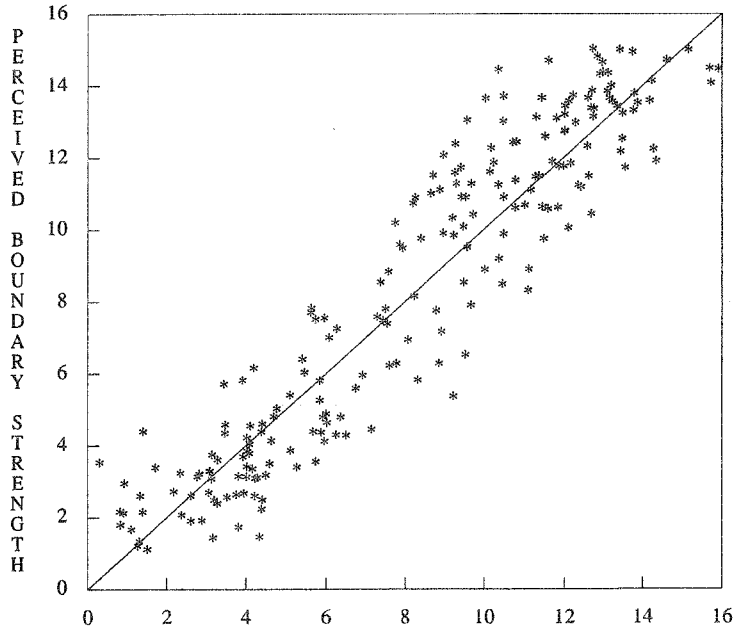

Figure 2

The actual boundary strength scores are plotted in Figure 2 against those predicted from a linear combination of all of the variables so far mentioned. The fit is quite good (the predicted scores account for 88% of the total variance), but this plot can be misleading because of what it does *not* reveal. Although both **peak proportion** and **duration** each significantly improve the prediction of perceived boundary strength, their contributions are not independent. Adding either of them to **F0 drop** increases the amount of variance we can explain, but adding both of them is hardly better than either one of them individually. This indicates that they both are carrying information about the same underlying property of the utterances, a property which itself is not well represented by **F0 drop**. This property is a temporal re-organisation of the segmental material and of how the F0 contour is aligned with it. The pre-boundary word is lengthened by about 10%, and the F0 peak of the pre-

boundary accent is located about 50% earlier relative to the rime of its associated syllable (ie relative to the distance between the vowel onset and the end of the syllable). Figure 3 illustrates part of the temporal re-organisation, by plotting the distance from the vowel onset to the end of the accented syllable on one axis, and the distance from the vowel onset to the F0 peak on the other.

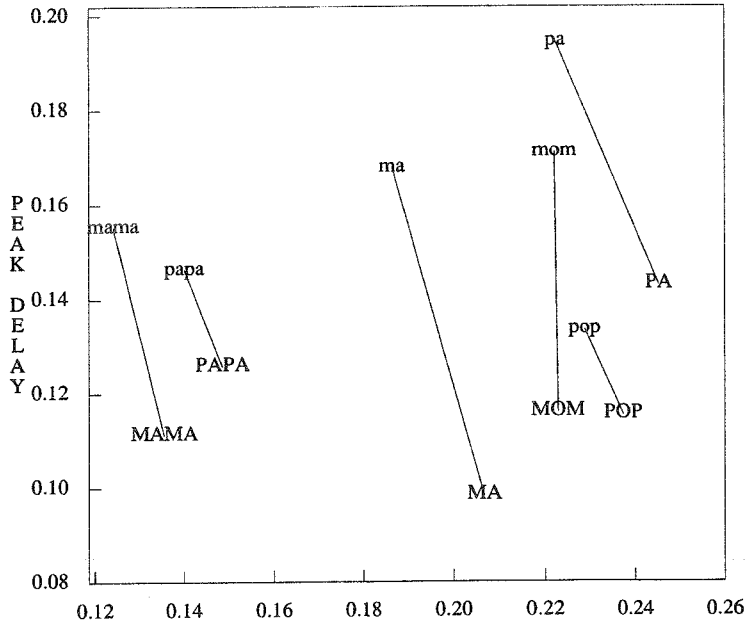

Figure 3

Each lower-case point (e.g. "ma") represents the mean of the corresponding subset of utterances with the weakest perceived boundaries (scores < 4.5); each upper-case point (e.g. "MA") represents the subset of utterances with the strongest boundaries (> 11.5). Clearly, when a word precedes a boundary the F0 peaks are earlier while at the same time the accented syllables tend to be longer. This temporal re-organisation contributes to the perception of a prosodic boundary, over and above the signalling power of a sharp fall in F0. In the current corpus it covaries to some extent with **F0 drop**; indeed **peak proportion** and **duration** (in conjunction with the encoded word structure) jointly accounted for 29% of the variance in **F0 drop** itself. But despite this covariance there is some independence, so that the temporal re-organisation makes its own additive contribution to the perception of the boundary. Most of the variation in **F0 drop** was *not* accompanied by joint variation in the temporal structure; similarly the majority of the variation in temporal organisation was *not* accompanied by joint variation in the extent of the F0 fall.

This leads to an important question for further study, namely the extent to which minor utterance-internal prosodic boundaries necessarily consist of associated changes in pitch and temporal structure both at once, in a tightly-linked way. Most qualitative descriptions, and also most synthesis-by-rule systems, consider prosodic boundaries to always be accompanied by both pitch and durational adjustments. In contrast to this view, the results here showed that even in a corpus containing only one type of boundary, there was some independence between the pitch and durational structure.

CONCLUSION

Concerning the boundaries in the current study, there are three points to make: [1] that boundaries are accompanied by changes in both the F0 contour and the temporal structure; [2] the changes to the temporal structure are more than a mere lengthening of the segmental material, there is also a concomitant change in the coordination between segmental timing and production of the melody; and [3] there is evidence that the F0 patterns associated with a boundary and the temporal re-organisation can vary independently of each other. In the richer structures of normal day-to-day discourse, speakers may even more independently manipulate the melodic and durational components of prosody

More generally, there results argue against approaching a corpus with the question "What is the most important single acoustic cue to such-and-such a category?" Speakers don't seem to manipulate single cues any more than listeners single-mindedly attend to only one aspect of an utterance. Rather, we should be seeking to discover the complex of jointly-varying dimensions of a signal, and using them on the one hand to elucidate the mechanisms by which underlying phonological categories are translated into phonetic detail, and on the other hand to further our understanding of just what the underlying phonological representations themselves are.

NOTES

(1) This is not always the case. Steele (1986), for example, found that the last word of *I gave it to **Nana*** was often *shorter* than the last word of *I gave it to **Nan***. One obvious difference between their experiment and the current one is that in their case the words preceded a much stronger, utterance-final boundary rather than a minor utterance-internal one.

REFERENCES

Beckman, M.E.& Pierrehumbert, J.B. (1986) *Intonational Structure in Japanese and English.* Phonology Yearbook 3, 255-309.

Hirschberg, J. & Pierrehumbert, J.B. (1986) *The Intonational Structuring of Discourse.* Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, 136-144.

Ladd, D.R. (1986) *Intonational Phrasing: The Case for Recursive Prosodic Structure.* Phonology Yearbook 3, 311-340.

Pierrehumbert, J.B. (1980) *The Phonology and Phonetics of English Intonation.* PhD Dissertation, Massachussetts Institute of Technology.

de Pijper, J.R. (1983) *Modelling British English Intonation.* (Foris: Dordrecht).

Silverman, K.E.A. (1987) *The Structure and Processing of Fundamental Frequency Contours.* PhD Dissertation, University of Cambridge.

Silverman, K.E.A. & Pierrehumbert, J.B. (1988) *The Timing of Prenuclear High Accents in English.* in M.E. Beckman and J. Kingston (eds),*Laboratory Phonology I: Between the Grammar and the Physics of Speech*, (Cambridge University Press: Cambridge).

Steele , S (1986) *Nuclear Accent F0 Location: Effects of Rate, Vowel, and Number of Following Syllables.* Journal of the Acoustical Society of America, 80, supplement 1, s51.