Synthesis by rule of English intonation patterns

Mark Anderson, Janet Pierrehumbert, Mark Liberman

Publication date: 1984/3/19

Conference name: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.

Volume 9, Pages 77-80

Publisher: IEEE

DOI
10.1109/ICASSP.1984.1172427

Synthesis by rule of English intonation patterns.

Mark D. Anderson
Janet B. Pierrehumbert
Mark Y. Liberman

Bell Laboratories
Murray Hill, New Jersey 07974

1. Introduction

This paper reports work on synthesizing English F0 contours. One motivation for this work is to improve the naturalness and liveliness of the prosody in speech synthesis systems. However, our main goal is to develop a theory of the dimensions of variation controlling intonation, and of their interaction. The fundamental frequency contours in Figures 1 through 3 illustrate three such dimensions: melody, stress, and pitch range. In Figure 1, the same word, "Anne", is produced with several different melodies. The different melodies carry different pragmatic implications: the first indicates that the speaker is providing information, the second, that he is requesting information, the third, that he is incredulous. In Figure 2, the same melody takes different forms depending on the location of the main stress of the phrase. The peak stays on the stress, while the remainder of the pattern covers the space from there to the end. Thus we view the stress pattern as affecting the F0 contour by its influence on the location (but not the type) of melodic elements. Figure 3 shows that the same melody can be produced in many different overall pitch ranges, reflecting something like the speaker's level of excitement. Overall pitch range seems to be continuously variable, whereas

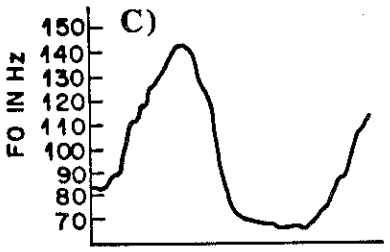


Figure 2

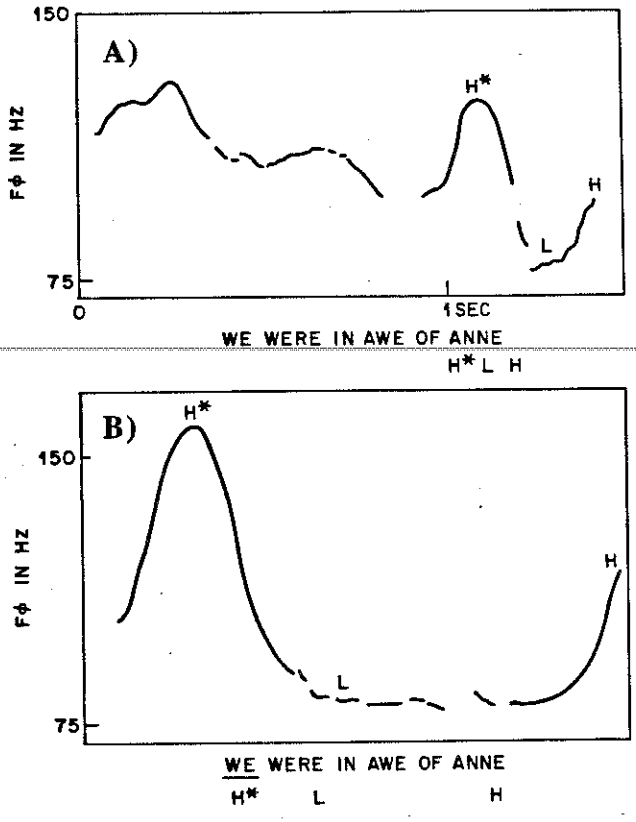
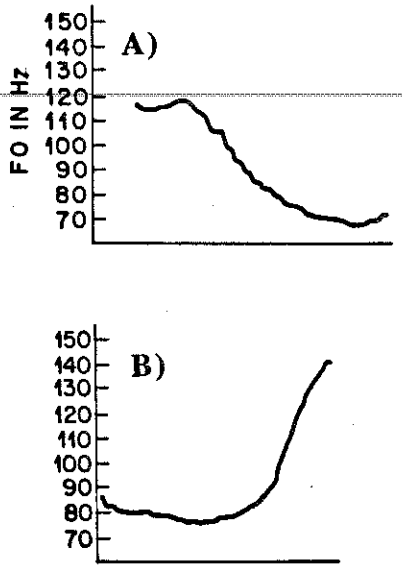


Figure 1



different melodies appear to be qualitatively different, in the same way that different words are. We note also that local emphasis within the phrase can affect the scaling of particular subparts of the F0 contour.

Synthesis provides a uniquely useful methodology for investigating this system. It enables us to assess the perceptual relevance of various features of the F0 contour. It allows us to construct controlled materials, whose acceptability or interpretation can be compared against theoretical predictions. Since speakers have poor conscious control of intonation, there is actually no other method of obtaining such judgments.

We take from Pierrehumbert [1] hypotheses about the abstract characterization of English melody; these are summarized below. We take from Liberman and Pierrehumbert [2] hypotheses about how these abstract elements are realized under variation in pitch range and utterance length. One of our aims is to determine the appropriateness of these hypotheses for a broader range of materials. Earlier work does not provide a good theory about the about the detailed shape of F0 contour for the different melodies, or about the timing of crucial points in them relative to the speech segments. As a result, addressing this issue is our primary concern.

2. The abstract characterization of melody

Some melodies are possible for English and others are not. An F0 contour which was possible for one text may be impossible for a text with a different stress pattern. Or it may count as an instance of a different melody. Just as we might hope to explicate the idea of "possible word " by positing an inventory of phonemes, rules of syllable structure, and so on, we hope to explicate the idea of "possible melody" by positing an inventory of primitive elements and rules for combining them.

In the theory proposed in Pierrehumbert [1], the primitive elements are two tones, Low (L) and High (H). The distinction between L and H is paradigmatic; L is lower than H would be in the same context. Each stressed syllable may carry a pitch accent, which consists of either a single tone or a sequence of two, with one singled out to fall right on the stress and the other leading or trailing. For example, the pattern in Figure 1B has a L on the stress with a trailing H; we will use the notation L*+H, where the * indicates alignment with the stress. A stressed syllable may lack an accent if the word is presupposed in the discourse or in a rhythmically weak position.

In addition to the pitch accents, there are two extra tones at the end of the phrase. Each can be either L or H. The first controls the F0 between the last accent and the end of the phrase, while the other controls the F0 right at the offset. If the last accent is on the last syllable, the other tones are crowded onto the same syllable. For example, in Figure 2A, a H, a L, and then a H, are all executed on the same syllable. In Figure 2B, the same melody is spread over more material. Note in Figure 6 that a H tone can also be used distinctively at the phrasal onset.

3. The shape of the F0 contour

3.1 Overall shape

Looking at an F0 contour often gives the impression that the local features ride on some global trend. Some theories reify these trends, proposing that the contour is described by the superposition of a phrasal F0 contour and the reflexes of local tonal elements. Examples of such theories include Fujisaki, Hirose and Ohta [3], Lea [4], and Cohen, Collier, and 't Hart [5].

However our recent research [2] indicates that the overall shape of the contour arises through the interaction of several different factors. First, different choices of tonal elements can in themselves create different overall shapes. A simple case is that of a L* accent followed by a H* accent; a rise results, which contrasts with the plateau which a H* H* sequence would create. Patterns involving downstep, illustrated in Figure 4, illustrate more subtly the same basic point. Each accent in this pattern causes the following one to be lowered by a constant fraction (under the right representation of scaling). Iterative application of this lowering makes a long chain of such accents come out as an exponential decay. A downtrend covering only part of the phrase will occur if such accents are mixed with accents of other types.

A second factor affecting overall shape is the relative prominence of the accents. The peak corresponding to a H* accent will be higher if the material it is attached to is more semantically or rhythmically important. Increasing the relative prominence of a L* accent seems to lower it.

After controlling for these effects, we find little evidence for declination, defined narrowly as a phrasal contribution to the F0 contour which decreases gradually as a function of time. Our results are consistent with a declination on the order of 10 Hz per sentence for a male voice, and this is also what Poser

Figure 3

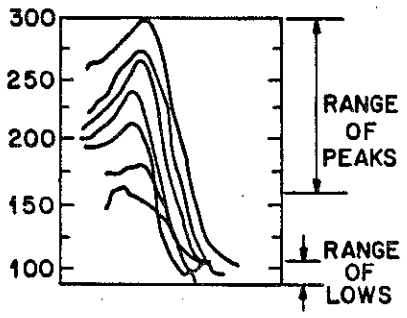
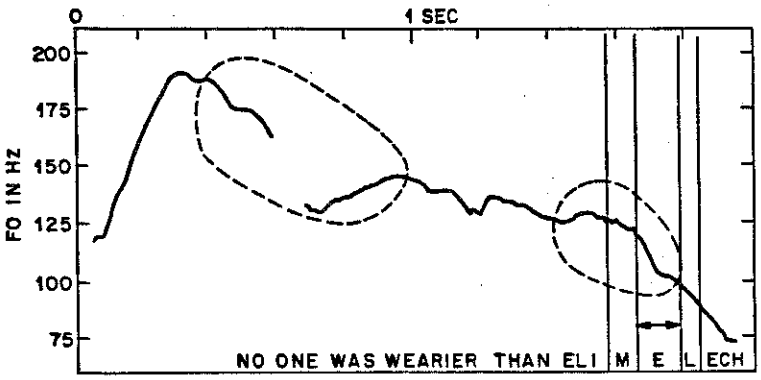


Figure 4: Downstepping accents are circled.



[6] reports for a similar analysis of F0 in Japanese. We do find that the last pitch accent in declarative sentences is lowered relative to all previous ones. We do not yet know if this final lowering applies to all pitch accent types or in all contexts.

3.2 Overall pitch range

As we noted above, a given melody can be produced on the same words, with the same stress pattern, in many different overall pitch ranges. What does it really mean to change the pitch range?

Our experiments on the results of "speaking up" indicate that the bottom of the pitch range, or baseline, remains unaffected over tremendous variability in the peak values. Speaking up changes the value of a mid position, which we call the reference line. The reference line is not revealed directly in the F0 contour, but only indirectly through its effects on the scaling of tonal features. Specifically, downstepping accents asymptote to it and H* accents scale upwards from it depending on their prominence. We conjecture that L* accents are scaled downwards from the reference line, but constrained to remain above the baseline.

In adopting this position, we have discarded two plausible alternatives. One, advanced in [1], is that overall pitch range has no theoretical status, but is just a general impression induced by the realizations of the particular tones present. This theory fails to provide a variable asymptote for downstepping contours. A second theory, suggested in Pierrehumbert [7] and Garding [8], is that the pitch range is an envelope defined by the baseline and a "topline" established through high points in the F0 contour. This point of view is not very coherent if the topline is constrained to go through high points, which can arise in many different ways and need not fall on a line or monotonic curve.

4. Local shape

Our main aim is to explicate the local features of the F0 contours by providing a rules for relating them to an abstract characterization. There are two major approaches to this problem in the literature.

Under an approach proposed in Ohman [9] for Swedish and further developed by Fujisaki et al. [3] for Japanese, the local features of the F0 contour are explained as the result of smoothing a discrete signal with a linear filter. In [3], accents are described abstractly as paired upward and downward step functions. An accent is defined by its onset, amplitude, and duration. Accents are smoothed by a critically damped low pass filter, and ride on an exponentially decaying phrase level component.

An alternative approach, exemplified in Bruce [10] and Pierrehumbert [7], relates tonal elements to target points in the F0 contour. The local shape around a target point is held to arise from the transitions between that target and adjacent ones. In the simplest case, the transitions computed are linear, but [7] suggested the use of nonmonotonic transitional functions to achieve a more sparse and phonologically motivated representation.

In comparing these approaches, one must distinguish between the essential characteristics of the existing models and the incidental consequences of additional assumptions. The treatment of pitch range in [7] is incorrect, but it would be easy to construct a target-transition model which incorporated the idea of a reference line. As noted elsewhere ([2], [6]) the Fujisaki model fails to generate a fixed terminal low point, incorrectly characterizes the context for downstep, and provides no mechanism for raising the asymptote for downsteps with overall pitch range. However, a filter model can be designed which corrects these defects.

In general, the relatively constrained nature and the mathematical tractability of filter models make them attractive. Also, they have some qualitative properties which appear to be correct and which target-transition models must build in by stipulation. For example, it is typical for a H* accent to be implemented as a rise on the stressed syllable. This would follow if the upwards excursion for the accent were timed on the syllable onset at an abstract level, but then subjected to causal filtering. [1] suggests that the delay between the L* and H in a L*+H is relatively constant, regardless of the segmental material. This would follow from an abstract characterization of the L*+H as a step, with the delay resulting from smoothing. Figure 5 shows a F0 contour in which the F0 dips between two H* accents. This dipping is common, and its admissable extent appears to depend on the separation of the accents. The filter model gives a natural and continuously behaved account of this behaviour, because an abstract dip between two targets is smoothed out as its duration becomes short relative to the width of the smoother. As a result of such observations, our current model is a filter model which makes several important changes from previous ones.

Figure 5

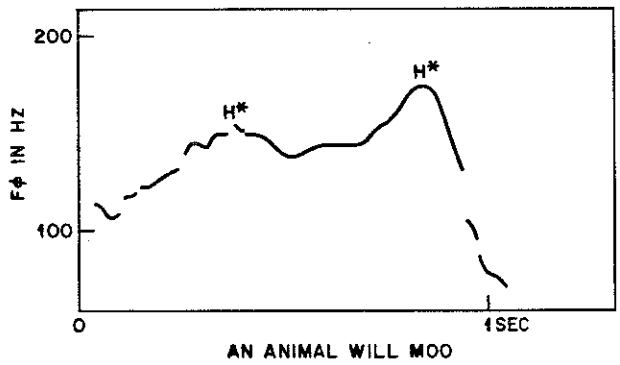
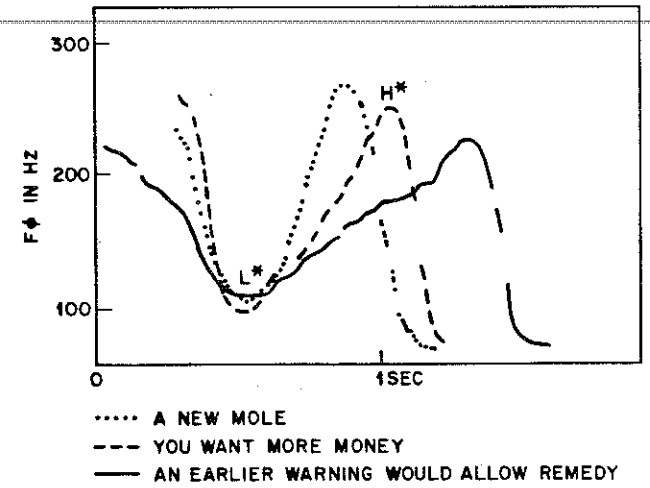


Figure 6



One set of changes is motivated by our results on pitch range and downstep. We introduce a reference line, with some tonal elements scaling upwards from there and others scaling downwards but constrained to remain above the baseline. Downstep is handled by manipulating the amplitude of the tonal realizations, rather than by superimposing them on a phrasal component decaying in time.

A second change is motivated by an intrinsic difficulty with a classic filter model: The shape of the smoother is responsible both for the F0 shape in the immediate vicinity of the accent, and for the amount of anticipation or perseveration between accents. As a result, these two issues are treated as inseparable. The F0 contours shown in Figure 6 indicate, however, that the slope leaving a L* accent depends on the location of a following H* accent over a quite wide range of separation of the targets. The smoother width needed to make such transitions look smooth, rather than displaying a rise-plateau-rise, appears to be too great to account for the contour within the stressed syllable. A target model deals easily with this regularity, but suffers from a complementary problem: The shape in the immediate vicinity of a tone is completely determined by the transitions to adjacent tones. What actually seems to be the case is that the immediate shape of the H* is quite stable; for accents which are somewhat separated, the same shape can actually be used whether the preceding accent is L* or H*. But the transition spanning the unaccented syllables varies depending on the time frequency separation of the targets. These observations led us to the hybrid model schematized in Figure 7.

A central characteristic of filter models is that they smooth out events of short duration, regardless of their type. If particular tonal features are especially resistant to being smoothed out, this must be handled by always assigning them a duration which is great enough for the smoother width used. The unstable features are the ones with variable durations. Any particular F0 contour can be well fit under this approach by making the right choice of durations. However, it remains to be seen whether this linkage of duration to stability stands in the way of developing coherent duration rules. We conjecture in particular that durations are related to the durations of the tone bearing segments; this hypothesis imposes enough additional constraint that there is a chance of failure.

5. Physiology

We would like to conclude by making a few remarks about the physiological motivation for the filter approach. Target-transition models are frankly empirical approximations to observed patterns. It is inconceivable that the motor system computes time functions centisecond by centisecond. Filter models appear to improve on this situation, since the smoothing used is plausibly related to the sluggishness of the articulators. However, neurological signals as high up as they can be traced are continuously varying time functions, rather than the discrete elements of our abstract representation. As a result, the smoothers used are (perhaps less frank) empirical approximations to the conflation of several levels.

6. References

[1] Pierrehumbert, J., (1980), *The Phonology and Phonetics of English Intonation*, Ph.D dissertation, MIT.

[2] Liberman, M. and J. Pierrehumbert, (1984) "Intonational Invariants under Changes in Pitch Range and Length," Aronoff and Oehrle, eds., *Language Sound Structure*, MIT Press, Cambridge.

[3] Fujisaki, H., Hirose and Ohta (1979) "Acoustic Features of the Fundamental Frequency Contours of Declarative Sentences in Japanese," Annual Bulletin #3, Research Institute of Logopedics and Phoniatrics, University of Tokyo.

[4] Lea, W. (1973) "Segmental and Suprasegmental Influences on Fundamental Frequency Contours," in L. Hyman, ed., *Consonant Types and Tones*, University of Southern California, Los Angeles, 15-70.

[5] Cohen, A., R. Collier, and J. 't Hart, (1982), "Declination: Construct or Intrinsic Feature of Speech Pitch?", *Phonetica*, 39:254-273.

[6] Poser, W. (1983), "On the mechanism of F0 downdrift in Japanese", *J. Acoust. Soc. Am.*, 74 Suppl. 1, S89.

[7] Pierrehumbert, J. (1981), "Synthesizing Intonation," *J. Acoust. Soc. Am.*,70:985-995.

[8] Garding, E (1982) "Swedish Prosody: Summary of a Project", *Phonetica*, 39:288-301.

[9] Ohman, S. (1967) "Word and Sentence Intonation: a Quantitative Model," *Speech Transmission Laboratory Quarterly Progress and Status Report 2-3*, 1967, 1-7, Royal Institute of Technology, Stockholm.

[10] Bruce, G. (1977) *Swedish Word Accents in Sentence Perspective*, Travaux d L'Institut de Linguistique de Lund, CWK Gleerup, Malmo.

Figure 7