SYNTHESIS OF ENGLISH INTONATION USING EXPLICIT MODELS OF READING AND SPONTANEOUS SPEECH

M. E. Johnson

Department of Linguistics
University of Edinburgh
Edinburgh, U.K.

ABSTRACT

A model of English intonation is presented in which a variety of intonation contours can be generated from a quantitative prominence labelling of stressed syllables. In one style of speech production, spontaneous speech, a short-lookahead model can generate a variety of contours from the same quantitative prominence labelling. For another style, reading aloud, a long-lookahead model determines the types of intonation patterns associated with texts. These typically come out as sequences of downward-stepping contours, given appropriate initial conditions (though there is no explicit downstep constant in the model). In both styles, the intonation contours are generated on the basis of a quantitative model of contour pitch prominence, in which the pitch prominence of the contour segments which make up the accent contours (and thence the intonation contours) is computed as a nonlinear function of the duration of the contour segment, the ratio of the F0 value at one end of the segment to that at the other, and a rhythm constant.

1. INTRODUCTION

It is quite common in models of intonation for accentual variation to occur in conjunction with some declination component, whether it have more concrete manifestation [14], [3] or less [12], [1], [9], [5]. It has been rare for the explicit declination component to be considered as dispensable [10], but common for local contributions to F0 downtrends (typically, 'downstep' and 'final lowering'), to be included as constants in the intonation model [12], [1], [10], [9]. At the same time, it has been observed that in spontaneous speech, an F0 downtrend does not always appear [15].

Many of these treatments of intonation approach modelling the variation of F0 from a phonological point of view, in which case it is natural to use categorical markers for local F0 downtrend triggers. Any global declining baselines and/or toplines can then be attributed to a lower-level productive physiological component, which is given greater [14] or lesser [1], [9] weight. However, if intonation production is taken to occur hand in hand with a control mechanism involving auditory feedback [8] (that is, a *closed-loop* mechanism), it is natural to adopt a different approach to modelling it. In one form of intonation production, downward trends may or may not occur, depending on local decisions about the F0 values needed to elicit a particular level of prominence on the accented syllables, made on the basis of what has been uttered and what is still to be uttered. This implies a *short-term* closed-loop control mechanism, and this is conjectured to be more common in spontaneous speech. In another form, downward trends occur naturally as a result of a *long-term* closed-loop control mechanism; this mechanism is conjectured to be more common in read speech. This paper outlines a model of intonation production which can utilise both mechanisms, and makes some passing consideration of its use for English text-to-speech synthesis.

2. INTONATION MODEL

2.1. Pitch Prominence of Contour Segments

The core of the model of intonation is a function yielding a pitch prominence value for a *contour segment* C, which is an ordered pair of F0-time coordinates; that is, start and end times $t1_C$ and $t2_C$ and the two F0 values sampled at those times:

$$C =< (f_0(t1_C), t1_C), (f_0(t2_C), t2_C) >$$

with the special condition that time $t1_C$ can occur later than time $t2_C$ (see subsection 2.2, Figure 3). The F0 value is taken to be some correlate of pitch at some quite high degree of abstraction, that is, after some unknown amount of neurophysiological processing has taken place in the auditory system. In fact, the model allows some samples of F0 to be taken at points when there is no vocal fold vibration, by means of interpolation (for contours passing through utterance-internal voiceless segments) and extrapolation (for contours respectively starting before and ending after vocal-fold vibration at the extremes of the utterance). In addition, the effects of segmental coarticulation are taken to have been factored out, and the contour modelled is much like the type of model abstracted in [13]; the straight line interpolation between points could also be replaced by parabolic interpolation, as in [6], without there being any change in interpretation of

the model.

There are three constants in the model, one fundamental and physiologically motivated, one less central and psychophysically motivated, and one directly empirical. The first is a rhythm constant, intended to be interpreted as the reciprocal of the the theta-rhythm frequency observed in various structures of the brain [2]. In the current implementation, it is set to 0.16. Alteration of this constant has direct effects on the intonation contour, specifically in the relative F0 values of accented syllables. The other two constants are described below.

Given the following:

$$f_C = f_0(t1_C)/f_0(t2_C)$$

$$d_C = |t2_C - t1_C|$$

then the function $P()$ which yields the pitch-prominence value of contour segment C is as follows:

$$P(C) = f_C(d_C/R)^E \quad (for \quad d_C \geq 0.03)$$

and

$$P(C) = 0 \qquad (otherwise)$$

where $E = log_B(f_C) \times (2R/(R + d_C) - 1)$ and where R is the Rhythm constant (typically 0.16). The heuristic constant B is e if $f_C >= 0$ and 0.667 otherwise (but see end of subsection 2.2). The lower limit of non-zero evaluation of the function is constant at 30ms (which is taken as the minimum duration required for perception of the pitch of a complex tone whose fundamental is in the range of F0 ([14], [4])).

The function is intended to give maximum pitch prominence to a contour segment whose duration is equal to R. When $f_C = 1$, the function evaluates to 1 for all values greater than 30ms. Inter alia, this means that a contour segment of 0.16 secs in duration has a pitch prominence of 1 pitch prominence unit, or **1 ppu** (Fig. 1a). When $f_C = 2$ (a distance of one octave), the function peaks at 2 ppu at 0.16 secs (Fig. 1b). The function is plotted for three values of f_C in Fig. 2.

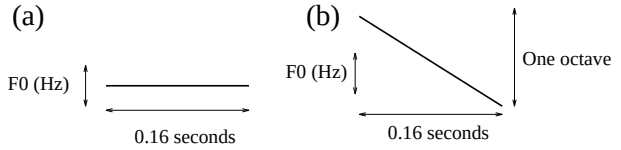

Figure 1: (a) Shape of a 1ppu contour segment lasting 0.16 secs (b) Shape of a 2ppu contour segment lasting 0.16 secs.

2.2. Pitch Prominence of Pitch-Accents

A pitch-accent is constructed from two *peak* contour segments and two *contextual* contour segments, one on either side of the peak contour segments. The contextual contour segments can be within the same syllable as the peak contour

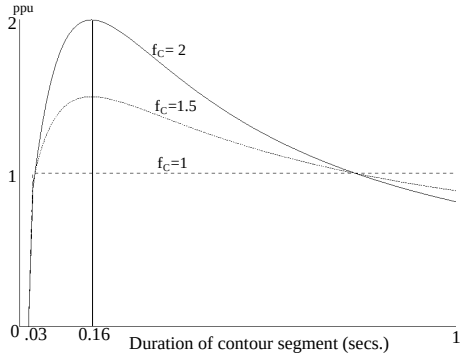

Figure 2: Pitch prominence function for three values of parameter f_C, which is the ratio of the F0 values of the endpoints of a contour segment.

segments; sometimes, they can be of zero duration. More typically, they comprise unaccented syllables either side of the accented syllable. The pitch prominence of the pitch-accent is computed as the sum of the pitch prominences of the contour segments to the left of the accent-peak (which are called the *pre-peak* contour segments, comprising one contextual and one peak contour segment) and those of the contour segments to the right of the accent-peak (*post-peak*, also comprising one peak and one contextual contour segment). The former two, though, are *reversed in time*. The prominence of a pitch-accent is thus evaluated from the peak position only, such that more recent points in preceding contour segments of the same pitch accent are interpreted as earlier points in the evaluated contour segment (Fig. 3).

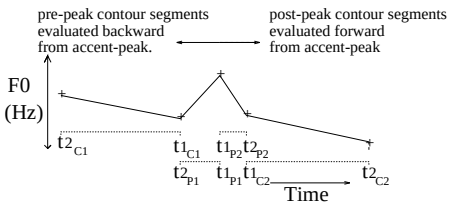

Figure 3: Evaluation of the pitch prominence of a pitch accent in running speech.

The perception of pitch prominence thus occurs as a simple additive function of peak (P) and context (C) segments at prominent peaks in the intonation contour, as follows (for pitch accent i, with $C1_i$ and $P1_i$ the pre-peak and $P2_i$ and $C2_i$ the post-peak contour segments):

$$P_{pa_i}(C1_i, P1_i, P2_i, C2_i) = P(C1_i) + P(P1_i) + P(P2_i) + P(C2_i)$$

The same formula for pitch-accent prominence can be used in a model of the production of intonation contours with an implied auditory feedback component. At any point in the production of an intonation contour, the pitch value of the upcoming accent is computed as a function of past sampled contour points in the most recent pitch accent and the contextual contour points of the upcoming accent (Figure 4).

For *equal prominence* peaks, this function is derived by simple algebraic manipulation of the equation

$$P_{pa_2}(C1_2, P1_2, P2_2, C2_2) = P_{pa_1}(C1_1, P1_1, P2_1, C2_1)$$

The quantity $f_0(t1_{P1})$ $(= f_0(t1_{P2}))$ for pa_2 (call this $f_0^{p_2}$) can be extracted straightforwardly following expansion of the basic pitch prominence function for the constituent contour segments of both pitch accents. For *different prominence* peaks, a scaling factor can be applied to the peak contour segments, and incorporated into the equation for the pitch value of the second peak. The algebraic manipulation yields two components (which we'll refer to as the *previous cycle factor* and the *current cycle factor*) derived from the formulae for the pitch-prominence of the two peak segments, and the four components relating to the pitch prominence of the contextual contour elements of each pitch accent. These components are used as follows in determining the peak F0 value of an upcoming accent (pa_2):

$$\begin{aligned} f_0^{p_2} = \quad & current\_cycle\_factor \\ & \times(S \times previous\_cycle\_factor \\ & \quad + P(C1_1) + P(C2_1) - P(C1_2) - P(C2_2)) \end{aligned}$$

where S is the peak pitch prominence scaling factor ($= 1$ for equal pitch prominence).

One important aspect of this approach is that interaccentual stretches of speech are typically taken to overlap (see Fig. 4). If the logarithmic bases B in the basic pitch prominence formula for the cases where $f_0(t1_C) < f_0(t2_C)$ and $f_0(t1_C) \geq f_0(t2_C)$ were reciprocals of each other, this would mean that the prominence of the post-peak contextual contour segment of the earlier accent and that of the pre-peak contextual contour segment of the later accent were identical. They would then cancel each other out in the formula for $f_0^{p_2}$ above. This would then make *prehead* and *tail* contour segments, to adopt the terminology in [11], the only contributors to accent peak variation, other than the peak contour segments of predecessor pitch accents. Indeed, in the situation where inter-accentual contour segments are equal in prominence anyway (when $f_0(t1_{C2_i})/f_0(t2_{C2_i}) = f_0(t1_{C1_{i+1}})/f_0(t2_{C1_{i+1}}) = 1$, where $C2_i$ is the post-peak contextual contour segment of the first pitch accent and $C1_{i+1}$ is the pre-peak contextual contour segment of the second), the F0 values of prehead and tail have a strong influence on the shape of the total intonation contour (this is often the case for the model of read speech; see section 4).

3. SHORT LOOK-AHEAD MODEL FOR SPONTANEOUS SPEECH

The approach to computing pitch-accent peak values on the basis of the immediate unaccented context and the prominence of the preceding pitch-accent is taken to be typical of spontaneous speech. There is no requirement that the unaccented contour segments form a declining baseline, though

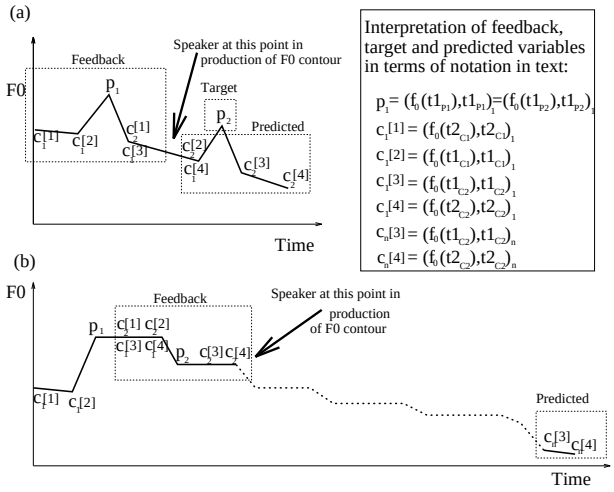

Figure 4: Calculation of peak F0 values in contour for two types of speech production. (a) Short-lookahead (spontaneous speech) (b) Long-lookahead (read speech).

this will often be the case. In Fig. 4a is illustrated the computation scenario in the production of a spontaneously spoken intonation contour. In Fig. 5 are illustrated different intonation contours (left panel, A and B) for an utterance whose accent peaks are taken to be of equal prominence. Thus, an assignment of prominence to the accented syllables (e.g. for text-to-speech synthesis) for both of these contours, would simply be the ordered pair $< 1, 1 >$. We can call this the *pitch-prominence vector*. In contour C of the left panel, the second accent is 20% more prominent than that in contour B. The pitch-prominence vector is thus $< 1, 1.2 >$. This quantitative assignment of prominence maps directly into a pitch value for each of the accent peaks, once an initial set of F0 values is given to the first pitch-accent in the utterance, and the (fixed) final F0 value for the utterance is determined. (Note that for final rising pitch, this final F0 value occurs as a trough R seconds after the end of voicing.) It is natural to expect many utterances to use a declining baseline as a reference for scaling pitch-accents. This is the case in Contour A. It is implicit in this model that although a declining baseline might be used to predict the F0 values in the tail in Contours B and C, the only parts of such a putative reference baseline that are factors in determining the F0 of the second accent peak are those which actually appear in the contours; in this case, the prehead and tail (see Ch. 4 in [8]).

4. LONG LOOK-AHEAD MODEL FOR READ SPEECH

In reading aloud from text, a different strategy is taken to be employed. Instead of using the predicted immediate post-accentual context for scaling the F0 value of an upcoming accent, the predicted post-accentual context for the final accent in the utterance is used throughout (Fig. 4b). In conjunction with the use of step-accents (which can be constructed

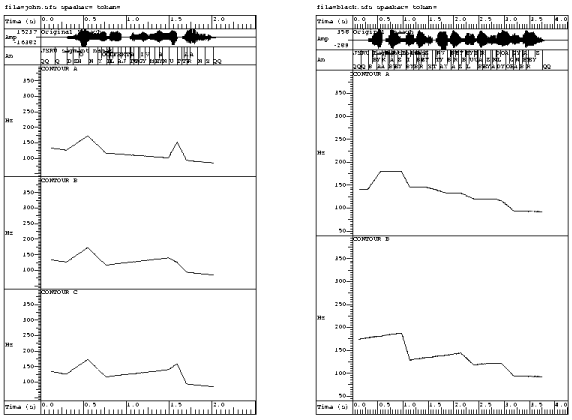

Figure 5: Left panel (Spontaneous speech: "JOHN would like to give me an UTTerance"):- Synthetic ((J)SRU text-to-speech) Speech and labels;Contour A [SOUND A677S01.WAV]: F0 contour for control strategy utilising declining baseline (Prominence vector = $< 1, 1 >$);Contour B [SOUND A677S02.WAV]: F0 contour with different interaccentual strategy (though baseline still cued) (Prominence vector = $< 1, 1 >$);Contour C [SOUND A677S03.WAV]: As B, but Prominence vector $< 1, 1.2 >$. **Right panel (Read speech: "BLACKberries, BILberries, STRAWberries, BLUEberries and LOganberries"):-** Synthetic speech and labels; Contour A [SOUND A677S04.WAV]: F0 contour with long lookahead strategy, mid-prehead (Prominence vector = $< 1, 1, 1, 1, 1 >$);Contour B [SOUND A677S05.WAV]: Same control strategy and prominence vector, but high prehead and slightly rising post-peak context.

from two contextual and two peak contour segments by, in this case, giving the first peak contour segment a negative gradient and keeping the second peak contour segment relatively flat) this results in intonation contours which have downstepping shapes. The post-peak contexts of non-final accents are not predicted by the utterance control strategy, and are simply proportional copies of the post-peak context of the initial accent in the utterance. This results in contours of the sort seen as Contour A in the *right* panel of Fig. 5 (this is an equal-prominence contour). In Contour B, the effects of changing the value of the prehead contour can be seen: another, different, downstepping contour, for the same pitch-prominence vector. This method of changing the whole shape of an intonation contour by altering initial conditions can allow for intonational variation (perhaps by stochastic variation of the initial conditions) of a more constrained, principled nature than in earlier work [7].

5. CONCLUSION

The basic behaviour of the pitch-prominence function used as a basis for intonation in this paper is plausible, and acceptable intonation contours result from its use, though work remains to justify its precise form. In addition, its physiological basis needs to be further investigated. Work is continuing in devising appropriate control strategies for text-to-speech synthesis, specifically in the case of spontaneous speech. Finally, it is hoped that the *performance* model of intonation outlined in this paper can be linked with a model of intonational *competence*.

6. REFERENCES

1. M. E. Beckman and J. B. Pierrehumbert. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:255–309, 1986.
2. B. H. Bland. The physiology and pharmacology of hippocampal theta rhythms. *Progress in Neurobiology*, 26:1–54, 1986.
3. J. R. De Pijper. *Modelling British English Intonation*. Foris, Dordrecht, 1983.
4. J. M. Doughty and W. R. Garner. Pitch characteristics of short tones. *Journal of Experimental Psychology*, 37:351–365, 1947.
5. H. Fujisaki. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency. *Annual Bull. RILP*, 21:165–175, 1987.
6. D. J. Hirst and R. Espessere. Automatic modelling of fundamental frequency. *Travaux de l'Institute de Phonetique, Aix-en-Provence*, 15, 1991.
7. J. House and M. Johnson. Enlivening the intonation in text-to-speech synthesis: An 'accent-unit' model. In *Proc. XIth Cong. of Phonetic Sciences*, 1987.
8. M. E. Johnson. *The Form and Auditory Control of Downward Trends in Intonation*. PhD thesis, University of London, 1993.
9. D. R. Ladd. Metrical representation of pitch register. In Kingston and Beckman, editors, *Papers in Laboratory Phonology*. Cambridge University Press, 1990.
10. M. Liberman and J. B. Pierrehumbert. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, editors, *Language Sound Structure*. MIT Press, Cambridge (MA), 1984.
11. J. D. O'Connor and G. F. Arnold. *Intonation of Colloquial English*. Longman, London, 1973.
12. J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980.
13. K. E. A. Silverman. *The Structure and Processing of Fundamental Frequency Contours*. PhD thesis, University of Cambridge, 1987.
14. J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge, 1990.
15. N. Umeda. 'F0 declination' is situation dependent. *Journal of Phonetics*, 10:279–290, 1982.