Proceedings of The Institute of Acoustics

AN ACCENT-UNIT MODEL OF INTONATION FOR TEXT-TO-SPEECH SYNTHESIS

Michael Johnson and Jill House

Dept. of Phonetics and Linguistics, University College, London.

INTRODUCTION

The prosodic component of a text-to-speech synthesis system cannot replicate all the variation in fundamental frequency of an intonationally rich language such as English, given the incomplete predictive models currently prevailing, and the high processing overheads that are entailed by the development of a system with 'a mind of its own'. A compromise approach is to model in detail some subset of the intonational system. Usually, bearing in mind the likely short-term applications of a text-to-speech system, prosodic components have been developed which generate acceptable intonation in a declarative style of discourse (e.g. Pierrehumbert [1]). A standard pattern is applied to all declarative sentences, with the result that an extended text may become intonationally repetitive (e.g. the current JSRU text-to-speech system, see Edward [2]).

To avoid this, we have built indeterminacy into our model. One reason for adopting this approach is our belief that the control mechanisms for speech are less tightly constrained than a fully specified model, such as Liberman and Pierrehumbert [3] aim for, would predict. Evidence to support this belief is found in our detailed analysis of natural speech, described below. Phonologically equivalent intonation contours may have a number of possible phonetic realisations, not all of them attributable to linguistic choice, range variation, or microprosodic fluctuation. For instance, the simplified 'head' patterns proposed by O'Connor and Arnold [4], have many 'allotonic' variants in practice (see Nolan [5]). Some such variation may be characteristic of a lect or idiolect, though an individual speaker uttering an extended text will not choose allotonically identical patterns throughout. Variability of this type may not always be perceptually salient, but to ignore it is to risk contour homogeneity which itself may be perceived as unnatural.

DESCRIPTIVE PRIMITIVES

Utterances may usefully be divided into breath groups comprising an ordered set of tone-units. Each tone-unit comprises one or more accent-units, plus any syllables preceding the first accent-unit. An accent-unit consists of an initial accented syllable followed by zero or more unaccented syllables. An accented syllable is rhythmically prominent (stressed), and will therefore always coincide with the beginning of a rhythmic foot. Not all feet are separate accent units, however, since a criterion of 'pitch prominence' must also be met: if the pitch pattern obtaining over a whole foot (mono- or polysyllabic) can be perceived as a linear extension of a preceding foot, it is considered to form part of the same accent-unit, its stressed syllable being unaccented. Otherwise it will fulfil the conditions for pitch prominence and constitute a separate accent-unit headed by an accented syllable. Tone-group-initial stressed syllables are deemed unaccented only if they are uttered at or

very close to a speaker's excitation modal frequency. This is a quantitative variable; otherwise, the descriptive basis of the model is auditory, hence our use of the term 'pitch'.

The nucleus is defined as the last accented syllable in the tone-unit. Nuclear syllables initiate nuclear accent units; this paper leaves open the exact relationship between pre-nuclear, or 'head', accent-units and nuclear accent-units, which are at present separately specified.

Accent-unit categories

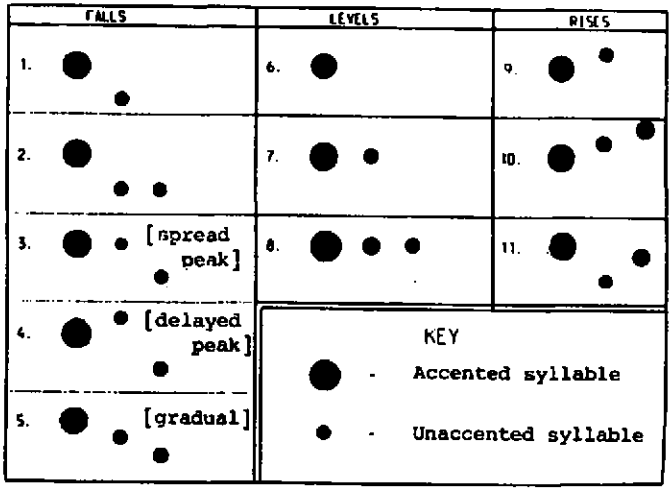
The various accent-unit configurations were derived by auditory analysis and transcription from recordings of two short texts (information bulletins) made by six RP or near-RP speakers. Examples of four basic nuclear tones were identified in this corpus: 'Fall', 'Fall-Rise', 'Rise' and 'Level'.

The analysis yielded three sets of head accent-unit configurations, represented schematically in Fig. 1, each blob representing one syllable. An accent-unit transcription of the recorded texts, using this notation, may be mapped on to Fx traces derived from the same recordings in a straightforward way, though the implications of a few discrepancies need to be explored. Auditory percepts were felt to be appropriate for determining local pitch patterns, while analysis of the Fx traces has yielded preliminary probabilities for the height of successive accents.

Head accent-units appear to be formally distinct over a maximum of three syllables, starting with the accented syllable. It is hypothesised that certain configurations are merely variants within a natural class: any configurations in which the final syllable is at a lower pitch than the accented syllable are grouped together, as are those ending at a higher pitch, and those with all syllables at the same pitch. This results in a grouping into falls, rises and levels. Single syllable accent-units are treated in this paper as levels, although auditory analysis does reveal a fair degree of pitch movement on some of these.

Of those head accent-units which were polysyllabic (605 in all), some 57% were level, 41% falling and 2% rising. There is no evidence to suggest that particular concatenations of units are impossible; the few gaps so far discovered may be accidental. However, we noted certain sequential probabilities which need to be incorporated in the model. The most striking was a bias towards falling head units occurring immediately before a Fall-Rise nucleus: 59% of the accent-units in this position were falling, compared with only 21% immediately preceding Fall nuclei.

Figure 1
Head Accent-unit configurations



Exact proportions, and any derived probabilities, reflect this small corpus only. Looking at individual speakers, one finds apparently idiosyncratic preferences for configurations and combinations.

Configurational variants

Following the general approach outlined in Ladd [6], the different falling configurations found in head accent-units are generated by reference to a small set of contour-shaping features. We propose three features to account for varieties of falling unit: unit 3 of Fig. 1 is produced if the feature [spread peak] is applied to the accented syllable; unit 4 is produced by the feature [delayed peak]; and unit 5 by the feature [gradual].

MODEL VARIABLES AND THEORETICAL ASSUMPTIONS

Declination

Declination is not explicitly modelled; instead, there is an increasing probability through the course of a tone-unit that the frequency of an accented syllable will not occur outside a restricted range of frequencies, defined in relation to the modal frequency. Furthermore, the range of fundamental frequency excursion within an accent-unit is reduced as the head syllable approaches the mode. However, it is possible, though not probable, for a head-accent to occur at a higher frequency than a preceding one in the same tone-unit.

The possibility of the first accented syllable occurring at the modal frequency is precluded by definition, since it would then merely be stressed. In fact, it

is probable that the first accented syllable in a breath-group will occur within one standard deviation of the mean of a normal distribution which tails off on one side towards the mode, and on the other towards the upper end of the speaker's range.

Declination is also implied across tone-units in a breath-group. Just as there is not a high probability that the fundamental frequency of a head-accent will be higher than that of a preceding accent within the tone-unit, so it is less than probable that the first accent in a tone-unit will be higher than the first accent of a preceding tone-unit within a breath group.

Range specifications

The topline is the top end of a speaker's range, the baseline the bottom. A reference value is calculated by subtracting a set fraction of the total range from the modal excitation frequency, which is a value considered appropriate for pre-accentual syllables. For real speakers, the modal value corresponds to the 'preferred' frequency observed in Fx distribution plots, made over a sample of speech at a given time.

OPERATION OF THE MODEL

The text is assumed marked into accent-units. Computation of the salient fundamental frequency (SF) is then performed. (The salient fundamental frequency is a frequency which, when extended throughout the length of a syllable in resynthesized running speech, provides the same gross auditory impression as a natural contour over that syllable. It is not always the same as the peak excitation frequency of a syllable, after consonantal transitional values have been ignored).

Firstly, an initial standard deviation is computed by dividing the total range by a value specifiable before operation of the model. This is for use in deriving SF values for accented syllables according to a normal distribution, where the mean is calculated as stated below. This parameter changes in proportion to the mean as it varies across the frequency range, so that the SF value for an accented syllable is less variable , within a local range, the closer the local mean is to the overall mode.

Then, for each tone-unit in a breath-group, SF is calculated for individual syllables, as follows:

i) Pre-accentual syllables (i.e. those occurring before the first accent-unit) are assigned the modal frequency.

ii) For each accented syllable, a putative mean SF value is first computed, according to a formula containing the following variables:

(a) number of head accent-units in tone-unit.
(b) position of accent-unit in tone-unit.
(c) position of tone-unit in breath-group.
(d) total Fx range (in Hz.).
(e) reference level (in Hz.)

Then, SF is computed by the formula

$$SF = a_m + sz$$

where a_m = computed mean SF; s = standard deviation, computed initially as as R / c (c being an arbitrary constant) and subsequently scaled according to distance from modal Fx; and z = randomly generated z-value for the standard normal distribution.

(iii) For unstressed syllables, SF depends on the type of accent-unit in which the syllable appears, any configurational features obtaining, the position of the syllable within the accent-unit, and the distance of the SF value of the immediately previous accented syllable from the top of the range and from the reference level.

If the head accent-unit is not immediately pre-nuclear, its type is determined from a transition probability matrix for pre-nuclear occurrences of high, level and falling accent-units (derived from the transcribed corpus), along with the probabilities for the three types in accent-unit initial position. If it is immediately pre-nuclear, a second matrix is used, which relates specifically to the transition probabilities between the four nuclear types and the three classes of head accent-unit. Table 1 shows the first matrix, Table 2 the second.

Table 1.
Transitional probability matrix for head accent-units. accent-units.

	L	F	R
L	0.70	0.28	0.02
F	0.58	0.36	0.06
R	0.33	0.67	0.00
IP	0.52	0.46	0.02

Table 2.
Transitional probability matrix for nuclear accent-unit and immediately preceding head accent-unit.

	L	F	R
F	0.73	0.21	0.06
FR	0.40	0.59	0.01
R	0.57	0.43	0.00
L	0.40	0.60	0.00

(L - Level accent-unit, F - Falling accent-unit, R - Rising accent-unit. IP - Tone-unit-initial probability. F - Falling nuclear accent-unit, FR - Fall-Rise nuclear accent-unit, R - Rising nuclear accent-unit, L - Level nuclear accent-unit).

In determining the distribution of features obtaining in falling head-units which are not immediately pre-nuclear, the stationary probabilities for unit types 2-5 in such positions are used (see Table 3). If the head-unit is immediately pre-nuclear, the transitional probabilities between the four nuclear types and the four pre-nuclear falling variants are used (see Table 4). Note that any two-syllable falling accent-unit cannot be feature-modified in the current static-syllable model. This constraint will be relaxed upon introduction of SF movement within pre-nuclear syllables.

Table 3.
Stationary probabilities for varieties of falling head accent unit (not immediately pre-nuclear).

[sp]	0.38
[gr]	0.18
[rp]	0.11
[pr]	0.33

Table 4.
Transitional probability matrix for nuclear accent-unit and type of immediately pre-nuclear falling head accent-unit.

	[sp]	[gr]	[rp]	[pr]
F	0.00	0.14	0.07	0.79
FR	0.43	0.24	0.04	0.29
R	0.20	0.50	0.00	0.30
L	0.33	0.00	0.00	0.67

([sp] - spread peak, [gr] - gradual, [rp] raised peak, [pr] - precipitous fall).

The type and feature specification of an accent-unit having been determined, the SF value of an unstressed syllable is calculated by case:

1. LEVEL

$SF_i = SF_{i-1}$ (where SF_i is the SF of the ith syllable, and SF_{i-1} the SF of the immediately previous syllable).

2. RISE

$SF_i = SF_{i-1} + ((H - SF_{i-1}) * g)$; (where H is the topline value, and g is a constant fraction expressing a gradual step).

3. FALL

3.1 First unstressed syllable

$SF_i = SF_{i-1} + (del * ((H - SF_{i-1}) * g)) - (grad * og) - (prec * op))$ (where del, grad and prec are binary variables taking a value of 0 or 1 and corresponding to the configurational features [delayed peak], [gradual] and the inverse of [spread peak], i.e. 'falling precipitously'; p is a constant fraction expressing a precipitous step, and o is the difference between SF and the reference value).

3.2 Second unstressed syllable

$SF_i = SF_{i-2} - o_2 p$ (where o_2 is the difference between SF_{i-2} and the reference value).

3.3 Subsequent unstressed syllables

$SF_i = SF_{i-1} - grad * (og)$

(iv) For post-accentual stressed unaccented syllables, SF is computed according to the following formula:

$SF_i = S_{i-1} + (d_1 * T(d_1, d_2))$ (where d_1 is the difference between the SF of the head accent and that of the first unstressed syllable of the most recent polysyllabic accent-unit, and d_2 the difference between the latter SF value and that of the second unstressed syllable of the same, within a breath-group. T is a truth-function with a value of 1 iff $d_1 = d_2$).

DISCUSSION

The model has been devised to exploit the natural variablilty of pre-nuclear contours used by speakers of RP English. In refining the model we shall consider the following questions:
1. The relationship between nuclear and other accent units. This is an early priority.
2. The role played by factors other than first order transition probabilities in determining concatenations of accent units (e.g. intonational idiom).
3. The best way of characterising a variety of English (averaging from different speakers, or using a single speaker model?).
4. The influence of textual factors (e.g. syllabic or morphological structure) on the application of configurational features.

ACKNOWLEDGMENTS

We would like to thank our sponsors in the Speech Research Unit at R.S.R.E. for their support in this project. Colleagues in the Department of Phonetics and Linguistics have provided much helpful advice, both in the preparation of this paper and in theoretical discussion. Geoff Lindsey and Mark Huckvale deserve special thanks in this respect.

REFERENCES

[1] Pierrehumbert, J., 'Synthesizing Intonation', J.A.S.A. 70, no.4, 985-995, (1981).

[2] Edward, J.A., 'Rules for Synthesising the Prosodic Features of Speech', JSRU Research Report No.1015, (1982).

[3] Liberman, M. & Pierrehumbert, J., 'Intonational Invariance under changes in Pitch Range and Length', in Aronoff, M. & Oehrle, R, (eds), Language Sound Structure, M.I.T. Press, (1984).

[4] O'Connor, J.D. & Arnold, G.F., Intonation of Colloquial English, 2nd edition, Longman, (1973).

[5] Nolan, F., 'Auditory and Instrumental Analysis of Intonation', Cambridge Papers in Phonetics and Experimental Linguistics, (1984).

[6] Ladd, D.R., 'Phonological Features of Intonational Peaks', Language 59, no.4, 721-759, (1983).

Figure 2

Two example contours produced by the model for the sentence "Here is the British Telecom Traveline Bulletin, prepared by the BBC Motoring and Travel Unit, for motorways." This comprises one breath group, containing three tone-units bounded by the punctuation. Thin line depicts the natural contour, bold line the synthetic, except during nuclear accent-units, where synthetic contours have not been generated. Segment durations are synthetic.