ENLIVENING THE INTONATION IN TEXT-TO-SPEECH SYNTHESIS: AN 'ACCENT-UNIT' MODEL

JILL HOUSE MICHAEL JOHNSON

Dept. of Phonetics and Linguistics
University College London
Gower Street, London WC1E 6BT

ABSTRACT

A new model of intonation for text-to-speech synthesis exploits natural variability within phonological constraints. Patterns are determined with reference to those preferred by an individual speaker.

INTRODUCTION

The output of a text-to-speech synthesis system needs to be intelligible, reasonably natural, and acceptable to the listener. A successful model of intonation will contribute to intelligibility, by clarifying the information structure of the text, and to naturalness, by using F0 contours characteristic of the target speech, aligned to the segmental structure of the text in a phonetically principled manner. To be acceptable to the listener, the output must combine intelligibility with whatever degree of naturalness is necessary to make the act of listening a comfortable, undemanding experience.

For the synthesis of isolated sentences, patterns may be readily specified which are plausible and 'easy on the ear'; but the use of these same patterns over longer texts, of a paragraph or more, leads to repetitiveness which the listener may find tedious: so, acceptability declines. We propose that enhanced acceptability during sustained listening may be achieved by exploiting a further aspect of naturalness: the variability to be found in the intonation patterns of natural speech.

THEORETICAL BACKGROUND

In natural speech, the choice of intonation contour for a text involves a number of separate phonological choices, some of which carry a higher functional load than others. These choices significantly constrain the degree of allowable variability, but within these constraints there is no one single 'correct' intonation pattern applicable to a given text spoken in a given context.

In developing an intonation model for synthesis-by-rule, an early priority must be to identify the sub-systems within which choices are made. For example:
(1) the division of the text into intonational phrases, or 'tone-groups';
(2) the allocation of accents (rhythmically stressed syllables which are also pitch-prominent, in the sense that they interrupt an established pitch contour);
(3) the relative prominence of accented syllables;
(4) the selection of the pitch contour whose starting-point coincides with the final accented syllable (the 'nucleus') of the tone-group -- the 'nuclear tone';
(5) the selection of pitch contour over any remaining (pre-nuclear) syllables.

These sub-systems imply a contour-based analysis which owes much to the 'British school' of intonation, notably /1/ and /2/. We believe that this approach is well motivated at the phonological level.

A theoretically sound synthesis model must allow for those formal differences for which a functional account can be given; ideally it should also model observed formal variations where no functional motivation may be apparent.

While lexical, syntactic and semantic factors play their part, the unifying principle determining intonation assignment is surely a pragmatic one -- the tailoring of an utterance to its context. In a synthesis system using unrestricted text input, any semantic or pragmatic knowledge is bound to be very limited. The rules must exploit any lexical or grammatical knowledge available, but occasional inappropriate choices will inevitably risk lowering the acceptability of the output (cf. /3/). The adverse effect of such errors may be minimised by an output which is otherwise natural-sounding and easy to listen to.

This paper does not directly address the problem of improving the syntactic, semantic or pragmatic knowledge-base. The model described assumes that the input text has been converted to a transcription on which tone-group boundaries and accented syllables are explicitly marked.

THE MODEL

Foundations: auditory analysis of a corpus

The model's phonological units and probabilistic rules were based on close auditory analysis and prosodic transcription of a short corpus of recorded texts. Four texts of 150-250 words each were derived from information bulletins -- reports on road conditions and weather forecasts -- issued over the telephone, using a declarative English style. Recordings of the original speakers (3 male, 1 female) were transcribed orthographically, using suitable punctuation, and presented as written texts to five experienced readers (3f, 1m), who in turn recorded the texts on to PCM tape in an anechoic chamber. A laryngograph signal (Lx), from which subsequent excitation frequency (Fx) analyses were made, was recorded simultaneously. All speakers used a (near) RP variety of English.

The recorded speech was transcribed prosodically, on an auditory basis, using a syllable-by-syllable interlinear notation. Comparison with the derived Fx traces indicated a reasonable match in terms of contour shape and relative pitch levels. No attempt was made to transcribe durational variation.

There was no one preferred reading for any of the texts, with respect to *any* of the sub-systems outlined above. A contour-based interpretation in terms of tone-groups and nuclear tones seemed well motivated, with falling, falling-rising, rising and level patterns all perceptually salient at the ends of intonational phrases. A consistent finding was a high degree of variability in contour-shape in pre-nuclear position. The contours were not readily interpretable in terms of fixed-pattern 'heads' (cf /1/); nor were sequences of accented syllables linked by any kind of automatic contour interpolation (cf /4/). This variability reflected a succession of choices between possible formal patterns. Their functional motivation was unclear, unless it was simply a strategy to avoid monotony. There was some evidence that individual speakers had preferred options among these patterns.

The inventory: units, contours and features

Units. The basic phonological unit chosen for the model is the *accent-unit* (AU) (cf /5/). This consists of an initial accented syllable together with any unaccented syllables following it. The unit is bounded on the right by the next accented syllable or by a tone-group boundary. Minimally, it will contain just the accented syllable; there is no theoretical upper limit, but units may contain more than one rhythmic foot, since some stressed syllables are not pitch-prominent, and are therefore deemed unaccented.

Within a paragraph of text, the largest unit recognised by the model is the *breath-group* (BG). This is normally equivalent to a grammatical sentence, since it corresponds in practice to a stretch of text bounded by /./, /!/ or /?/. A breath-group may be subdivided into *punctuation-groups* (PG) (bounded by /,/, /;/ or /:/), which in turn may contain more than one *tone-group* (TG). Tone-groups are composed of one or more accent-units, together with an optional *prehead* (PH), corresponding to any unaccented syllables preceding the first accent in the group. The final accent of a tone-group is the *nucleus*; preceding (optional) accent-units make up the *head*. All BG and PG boundaries are also TG boundaries. The hierarchical structure linking groups and units is demonstrated below:

(1) $[[[["No._{AU}]_{TG}]_{PG}]_{BG}]$ (= minimal breath-group)

(2) $[[[[$There are$_{PH}][$"more lane 'closures$_{AU}]_{TG}]$

$[[$between 'junctions$_{PH}][$"thirty$_{AU}][$"two and

'thirty$_{AU}][$"three,$_{AU}]_{TG}]_{PG}][[[$"north of$_{AU}]$

$[$"Preston.$_{AU}]_{TG}]_{PG}]_{BG}]$

Key: " accented syllable; ' stressed syllable.

Contours and features. Accent-units are characterised by contour. Nuclear unit contours in the corpus represented four basic *nuclear tones*: *falls*, *fall-rises*, *rises* and *levels*. (This formal classification does not distinguish between 'fall-rise' and 'fall + rise'.) Head unit contours fell into three natural classes: *levels*, *falls* and *rises*. An earlier version of this model /6/ treated nuclear and head units separately; the revised version considers both types to belong to the same underlying formal classes. Nuclear units typically involve more salient F0 movement than head units, and are more predictable in their alignment to syllables. Head unit contours showed much variation in this respect, depending not only on the number of syllables in the unit, but on features affecting, for example, the timing of the start and end points of the characterising contour. Stylisations of the basic contour shapes perceived, subsequently adapted for implementation in the model, are shown in Fig. 1.

The use of features owes its inspiration, but not its detail, to Ladd /7/. The unmarked forms in Fig. 1 were those characteristically found in nuclear position, where the marked forms were less common; both marked and unmarked varieties occurred freely in head units, though fall-rises as a class were rare in this position. In practice, there is normally a brief sustention of F0 at the start of the contour, which is accounted for in our implementation; contours are only considered to display the feature [+delayed start] when such a sustention continues into the second syllable. Perceptually level contours may, in fact, follow a shallow declination line; this too is accounted for in the implementation.

Distribution of AU contours in the recorded corpus

Nuclear. Between them, falls and fall-rises accounted for around 75% of all the nuclear tones. All BG boundaries ended in /./, and all were associated with a falling tone (with one exception: the sentence 'Thank you for calling.'). Any of the tones could be found at other PG and TG boundaries. Statistically, fall-rises were most probable here, particularly in the case of BG-initial tone-groups.

Head. Over 90% of the units had either level or falling patterns; rising units were rare. There were few, if any, positional constraints on these contours. However, when they were analysed in relation to the nuclear tone which followed them in the tone-group, certain tendencies came to light. Though levels constituted around 57% of head units overall, this proportion dropped to around 40% when immediately preceding a fall-rise nuclear tone, where they were overtaken by falls. Most of the few rising head-units occurred in tone-groups with a falling nuclear tone. There were no obvious constraints on the juxtaposition of different accent-units, but estimates of the probabilities of certain collocations could be derived from the corpus distributions. There was no clearly identifiable pattern governing the application of the features to these head contours.

The probability values quoted above are derived by adding together the scores for several RP-style

speakers, a procedure which allows us to make some generalisations about intonation units used in this variety of English for this discourse style, but which obscures the preferences of the individual speakers. Probabilities based on averaged values, or, preferably, on those appropriate to a particular speaker, may be adopted to implement the model in a text-to-speech system (see below).

Relative prominence of accents

The derived Fx traces relating to the corpus recordings allowed us to make an accurate assessment of the actual and relative peak frequencies of accented syllables. Within a tone-group, there was a marked tendency for the Fx of successive head accents to show some sort of decline. There were no fixed target values associated with accents in any position, but it was possible to define a frequency range within which accents were likely to occur. The position of the tone-group within the breath-group, and of the breath-group within the paragraph, were relevant in determining the height of TG-initial accents. The starting-point of nuclear accents was more varied. a step up from a preceding accent typically reflected a linguistic need to highlight the item in question.

Implementation

This section looks further at the principles guiding the rules in our implementation, rather than at the detailed algorithms, which are still subject to revision. An earlier version of the implementation is described in more detail in /6/; a report on the revised implementation is in preparation.

The model is implemented on the JSRU text-to-speech synthesiser, a system modelled on male RP speech, replacing or adapting the prosodic model of Edward /8/. Rules relate to F0 values only; all other aspects of the system are unchanged.

Reference frequencies. The overall range is based on that of a particular (male) speaker. Values are derived from frequency distribution (Dx) histograms made from the Lx signal recorded with the reading of the texts. The extremes of the range ('HiFx' and 'LoFx') are taken from the 1st order distribution, as is the 'mode' (preferred frequency) value. An additional reference value used in computing the synthetic F0 is 'LoFx2', the lower limit of the range as measured on a 2nd order distribution.

Selection of AU contours. Nuclear contours (nuclear tones) are assigned on the basis of punctuation. For instance, boundaries associated with full stops invariably generate falls; commas and unpunctuated boundaries are associated by convention with the fall-rise, but an algorithm converts this to a rising or level tone in certain circumstances. Head-unit contours are assigned according to tables of probabilities derived from analysis of the speech to be modelled. The tables take account of the transitional probabilities associated with the collocation of different contour types (see /6/ for specific examples). Feature application makes use of tables of stationary probabilitles similarly derived from the corpus.

Contour to F0 conversion. Accent-unit contours are broken down into their constituent levels: H (high) and L (low) for falls and rises, with additional constituents H' and L' to deal with the more complex and marked forms. Level contours consist of H and H' only. The features [delay] and [raised] apply to H or L as appropriate.

F0 values for H and L are calculated on separate criteria. The first H in a breath-group is plotted in relation to the mean value used for such accents by the reference speaker. The H value in subsequent accents will be at a point which is a fixed proportion of the distance between the mode and the previous H. An adjustment upwards is made in a new punctuation-group for the first H, which is related to the previous PG-initial accent. There is a degree of allowable deviation from the computed mean values. Declination between accented syllables is a *derived* effect.

In nuclear units, the value of L coincides with LoFx2, unless it is BG-final, in which case it coincides with LoFx. In head units, L is calculated as a proportion of the distance between its associated H and LoFx2.

Prehead syllables are by default clustered around the modal value.

DISCUSSION

Many of this model's algorithms are still being modified, but even at this early stage, we believe that the output has much to recommend it. It is closely based on observations of natural speech; it allows the distribution of patterns to be modelled on a particular speaker; it exploits natural intonational variability. The inventory could be readily expanded, and the tables modified, to suit different discourse-styles and lects.

In due course, the model will be integrated with improved higher-level rules for phrasing and accent-placement; and at the lower level, a set of micro-prosodic rules will adjust the essentially straight-line contours now generated (Fig 2) to enhance the phonetic naturalness of the output.

Meanwhile, the model avoids the intonational repetitiveness often associated with synthetic speech. In its present implementation, there is in fact no way of predicting precisely which set of contours will be applied to a given text. A further planned development will be a facility whereby particular patterns may be specified explicitly if required.

ACKNOWLEDGEMENTS

Thanks are due to our sponsors at the Speech Research Unit, RSRE, and to numerous colleagues at UCL for their comments and support.

REFERENCES

/1/ J.D. O'Connor & G.F. Arnold, *Intonation of Colloquial English*, Longman, 1961, 2nd ed 1973.

/2/ D. Crystal, *Prosodic Systems and Intonation in English*, Cambridge University Press, 1969.

/3/ G. Akers & M. Lennig, 'Intonation in text-to-speech synthesis: evaluation of algorithms', *JASA* 77, 2157-2165, 1985.

/4/ J. Pierrehumbert, 'Synthesising intonation', *JASA* 70, 985-995, 1981.

/5/ F. Nolan, 'Auditory and instrumental analysis of intonation', *Cambridge Papers in Phonetics and Experimental Linguistics* 3, Dept. of Linguistics, University of Cambridge, 1985.

/6/ M. Johnson & J. House, 'An accent-unit model of intonation for text-to-speech synthesis', *Proc. IOA: Speech and Hearing* 8, part 7, 409-416, 1986.

/7/ D.R. Ladd, 'Phonological features of intonational peaks', *Language* 59, 721-759, 1983.

/8/ J.A. Edward, "Rules for synthesising the prosodic features of speech', *JSRU Research Report* 1015, 1982.

Fig. 1: Schematised accent-unit contours

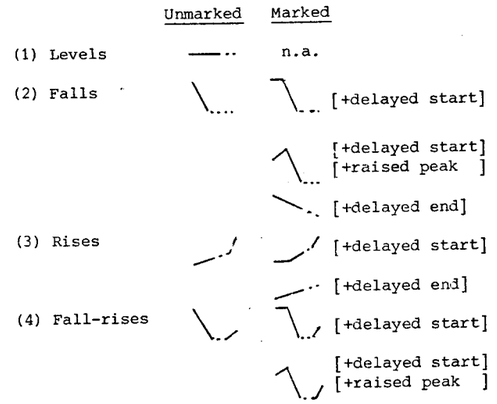
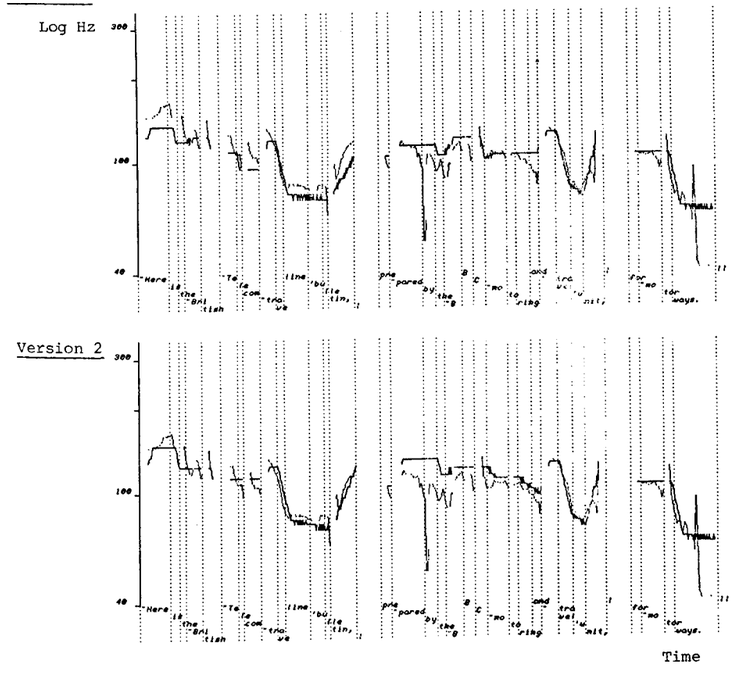


Fig. 2: Comparison between accent-unit contours and an F0 contour derived from natural speech

The solid line in each version is the accent-unit contour; the broken line is the contour derived from natural speech, aligned with the JSRU synthetic segmental durations:
"Here is the "British "Telecom "Traveline 'bulletin, pre"pared by the "BBC "Motoring and "Travel 'Unit, for "motorways.



In the contour generated by rule, H and L values are calculated in JSRU pitch levels. Interpolation between them is according to a 'moving-target' algorithm to prevent 'steppiness' in synthesis with a 100Hz frame rate.