APPLYING THE TONETIC STRESS MARK SYSTEM TO THE SYNTHESIS OF BRITISH ENGLISH INTONATION

Briony J. Williams and Peter R Alderson

IBM UK Scientific Centre, St Clement Street, Winchester, SO23 9DR, U.K.

Abstract

The synthesis of intonation in a text-to-speech system has long been a neglected area. Recently, work by Pierrehumbert has developed a model for synthesising American English intonation which uses a string of 'pitch accents', assigned autosegmentally. On the other hand, the 'British school' of intonation analysis has developed a representation of intonation that has been used successfully in transcribing spoken (British) English and in teaching intonation to foreign learners of English.

The work reported here is an attempt to blend the two appraoches in the context of a text-to-speech synthesis system for British English. The input text contains a linguistic representation of intonation, using units such as 'nucleus' and 'head' These units are converted to a set of abstract 'target values', restricted to a scale of one to ten. These in turn are converted to frequency values by the superimposition of a declining frequency envelope, the parameters of which are dependent both on the speaker model used and on the higher-level declination currently in force. The frequency values are added to the segmental information, and the result is output as speech.

1 'British' school of intonation analysis

Most analyses of English intonation proposed by linguists may be placed in one of two major schools of thought: the 'American' and the 'British'. The 'American' approach sees pitch levels as phonemic for intonation, and pitch contours as simply the concatenation of levels. For linguists of the 'British' school, however, the pitch contour is the primary unit of analysis and there is no attempt to segment it into its constituent levels. This approach was developed partly as a pædagogical tool for the teaching of English as a foreign language, and also for the practical transcription of the intonation of real speech. An example is the work of O'Connor and Arnold [4] or Crystal [2], who split each intonational phrase or **word group** into constituent units. A word group contains one obligatory unit, the nucleus, which falls on the most prominent word of the group. Preceding accented syllables are referred to collectively as the **head**, and any unstressed syllables before these are known as the **prehead**.

1.1 Adapted 'British' system

The work reported below makes use of a model of intonation based on that of O'Connor and Arnold, with features from Crystal's analysis but differing in some respects from both. It has been formulated to avoid some inconsistencies found in the units proposed by O'Connor and Arnold, as detailed in [8]. The system as a whole closely parallels that found in the work of other 'British school' linguists. The basic units of the system are shown in Figure 1 below.

Tone-units: A major tone-unit boundary mainly occurs at a longer pause; a minor tone-unit boundary is mostly found at a shorter pause or *filled pause*, i.e. with lengthening of the final syllable of the minor tone-unit.

Accented syllables: Five types of pitch movement are recognised for accented syllables: fall, rise, fall-rise, rise-fall, and level. If the accented syllable is followed by one or more unaccented syllables, then the pitch configuration is spread over the accented syllable and the following unaccented syllables. The five accent types apply equally to the nucleus and the head, thus simplifying the analysis considerably. For O'Connor and Arnold, as for Crystal, the types of pitch pattern found in the head are phonemically distinct from those found in the nucleus. The analysis described here makes no such rigid division, thus allowing a generalisation to be stated. The accent types may be either *high* or *low* (represented by super- and subscript symbols respectively). These terms refer to the initial pitch of the accented syllable as compared to the pitch of the preceding syllable.

Tone-unit boundaries

Major:	||	Minor:	|

Accented syllables

Fall:	ˋs, ˎs	Rise:	ˊs, ˏs
Fall-rise:	ˇs, ˬs	Rise-fall:	ˆs, ˍs
Level:	¯s, _s		

Unaccented syllables

Booster:	↑s	Drop:	↓s
	Stressed:	·s	

Figure 1. Intonational units used

Unaccented syllables: Stressed but non-pitch-prominent syllables may occur at any point in the tone-unit. They are marked with a mid-high dot. Pitch-prominent but unstressed syllables are those syllables which deviate markedly from the pitch direction so far established. They may be either much higher or much lower than the immediately preceding syllable, and are marked by up-arrow and down-arrow respectively. Unstressed and non-pitch-prominent syllables form the majority of unaccented syllables, and are notationally unmarked.

1.2 Background to the model

A 'British school' system was chosen, rather than an 'American school' system, because the former type has proved its value in the transcription of real speech. Although O'Connor and Arnold originally used only carefully-constructed examples, for pædagogical purposes, the same type of system has been used successfully in the transcription of sizeable corpora of spoken English ([2], [6], and the corpus described below). The American school' type of model has not been as extensively used for this purpose. Therefore it was felt that the former type was more likely to reflect all and only the linguistically-significant pitch movements of (British) English.

2 Spoken English Corpus

The intonational model described above is being used for the prosodic analysis of a corpus of contemporary spoken British English that is currently being compiled by researchers at the University of Lancaster, U.K., and the IBM UK Scientific Centre. This involves the recording of programmes from the radio. These are non-spontaneous monologues dealing with such subjects as current affairs (both newsreading and live reporting), financial advice, Open University lectures, dramatic narrative, religious services, and general-interest lectures.

After the initial high-quality recording of a programme, a portion is transcribed prosodically using the system outlined above. The prosodic transcribing is divided between two phoneticians: Dr. Gerry Knowles of Lancaster University, and Dr. Briony Williams of the IBM UKSC. There seem to be no serious discrepancies between the two transcribers, and there is a high degree of agreement between them on the accent types and boundary locations used. To date, approximately 33,000 words have been transcribed prosodically. The finished corpus is expected to contain 50,000 words, all prosodically transcribed.

3 Synthesising from a prosodic transcription

A few sentences were chosen at random from texts included in the Spoken English Corpus, and the (manually-assigned) prosodic transcription of these sentences was used as the basis for synthesis of the intonation. The hypothesis was that the prosodic transcription, having been made by hand from the recording, was a full and sufficient description of the linguistically-relevant pitch variation in the utterance. If a version of the utterance synthesised from the prosodic transcription then proved to be essentially indistinguishable from the (resynthesised version of) the original, this would support the view that the linguistic units chosen for annotation were necessary and sufficient for the prosodic characterisation of that utterance. With this in mind, the following sentence was arbitrarily selected as an example:

Dada did not really attempt itself to offer a consistent solution. It was enough to expose the crisis in the relevance of art. However. Dada did put forward some positive proposals.

3.1 From the prosodic transcription to 'target values'

Using the (manually-assigned) prosodic transcription shown in Figure 2 as input, syllables were then assigned **target values**. These are integer values between 1 and 10, representing an abstract scale of linguistically-relevant pitch height. These target values are similar to those in [5]. Each accented syllable had a target value, while those carrying pitch glides had two or three as appropriate. In addition, each unaccented syllable at the end of an accent contour (i.e. just before the next tonetic stress mark, or before a tone-unit boundary) was given a target value.

↑¯Dada did not _really aˋttempt | itˋself | to ¯offer a con_sistent soˋlution || it was eˋnough | to ex¯pose the ˏcrisis | in the ˋrelevance of ˋart || ↓howˊever | ·Dada ¯did put _forward _some | ·positive proˋposals ||

Figure 2. Prosodic transcription made by hand from recording

The target values, under the proposed system, are assigned according to simple rules based on the accent types marked. For example, a high (superscript) fall is assigned an initial target value that is three greater than that of the immediately preceding syllable within the same minor tone-unit, while its final target value is six less than this initial value (with a minimum value of 1 and a maximum of 10). The final value applies to the end of the syllable, if the accent is monosyllabic: otherwise, it applies to the last of the following unaccented syllables, the F0 of the intervening ones being later interpolated.

3.2 From target values to Hz frequency values

These target values are then converted into frequency values in Hz. This is done using essentially the same method as in [5]: i.e. superimposing an overall pitch envelope that incorporates declination. In this case, unlike the original method used by Pierrehumbert, the baseline represents the lowest possible limit of the speaker's pitch range, and is constant. The topline, on the other hand, declines exponentially from start to end of a minor tone-unit. The topline declination is set on a global basis, by specifying its value at the beginning and end of the (first) minor tone-unit, and interpolating exponentially between those values. At the start of any following minor tone-unit within the same major tone-unit, the initial F0 value for the topline is reset, but at a point somewhat lower than that of the corresponding point in the preceding unit; and similarly by the same proportion for the value of the topline at the end of the minor tone-unit. Thus the effect is an exponential decline in topline reset values over the course of a major tone-unit. In addition, at the start of a new 'paragraph', the topline returns to its original value.

For the purposes of the present investigation, the values for the baseline, topline start, topline end, and drop in reset value of topline, were adjusted such that the closest possible match was obtained between the output Hz values for the vowels and those of the original utterance. The aim was to match the output to the original utterance in order to form an impression of the validity of the linguistic units used.

Having set the values for the overall pitch envelope as described above, the target values were then taken as specifying proportions of this overall envelope. The program superimposing the declination envelope converted each target value to a frequency value in Hz. The recorded utterance was digitised at 10 kHz using a 4.5 kHz low-pass filter. This digitised utterance was then analysed by linear predictive coding (LPC), using a filter order of 64. The excitation coefficients were then replaced by the F0 values obtained from the process described above. Each F0 value was assigned to the vowel of the syllable, at a point in time that was 25% into the vowel's duration. It was found that this gave a more natural-sounding

The output of the above processes is shown in Figure 3, where it is plotted with the F0 of the original utterance. output than if the F0 value were assigned at the very onset of the vowel, or halfway into the vowel. Once all values had been assigned, the F0 was interpolated between them.

Finally, F0 perturbations of 15 Hz were added at the boundaries between voiced and voiceless segments. This process reflected a physiologically-determined effect occurring in real speech at such boundaries. Although no attempt was made to allow for intrinsic vowel pitch and other perturbations, it was found that this one process greatly improved the naturalness of the synthesised output. The reference form of the original utterance is not the digitised version, but the version obtained by resynthesis from the LPC coefficients for the original. This was felt to be more comparable to the experimental resynthesised version, factoring out the effects of LPC resynthesis to display only the effects of alterations in the F0. In addition, the boundary between the second and third sentences was treated as a paragraph boundary, with complete resetting of the topline to its original value at the start of the third sentence. This was felt to be justified before the accented sentence adverb *however*, which was here functioning in an introductory, paragraph-initial fashion.

The match between the rule-synthesised F0 and the resynthesised original is good. To the ear, the match is even closer: a surprising discovery was that many discrepancies seen on the F0 plot in Figure 3 were not in fact perceptually salient. These discrepancies could be heard only on careful listening and in full knowledge of what to listen for, and seemed to be related to segmental micro-effects on F0 rather than to linguistically significant intonation. This suggests that attempts to match as precisely as possible to the original F0 may be unnecessary. A more useful metric is that of the **perceptual equality** of two F0 contours, as used by some Dutch workers on intonation synthesis (e.g. [7], [3]). Their notion of perceptual equality is based on linguistic and auditory indistinguishability, rather than on acoustic identity. Since no two utterances of the same sentence are ever completely identical acoustically, the notion of perceptual equality may well prove to be of value in the assessment of synthesised intonation.

4 Discussion

The investigations reported above may have implications for the way in which the synthesis of intonation is approached. An attempt has been made to use a theoretical model which expresses just those pitch movements that are linguistically significant in British English,

and which has been used successfully for many years for the practical transcription and teaching of British English intonation patterns. The results so far support the view that the model chosen is able to account satisfactorily for the large-scale, linguistically-relevant features of pitch movement. If these movements are correctly specified, it is then possible to go on to consider segmental effects on F0, which affect the perceived naturalness of the output without contributing to the linguistic message.

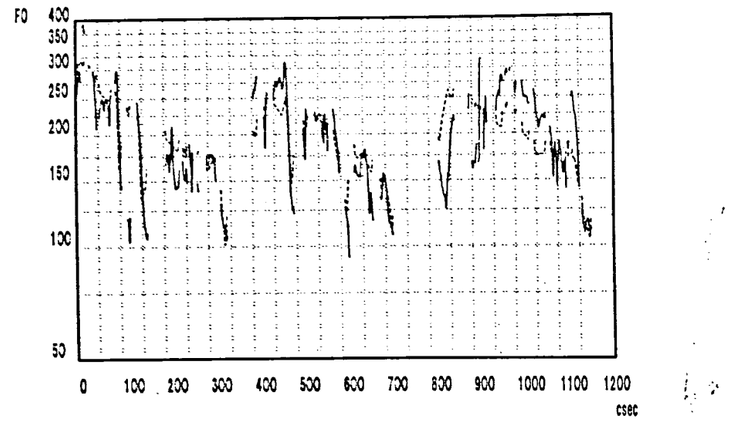

Figure 3. Original resynthesised F0 vs. F0 synthesised by rule

Solid line = F0 of resynthesised original utterance: hatched line = F0 of utterance synthesised by rule from prosodic transcription.

The assessment of intonation contours is peculiarly difficult, as it is rare for these to be definitively correct or incorrect: listeners will strive to fabricate a convincing scenario for an inappropriate intonation contour, rather than reject it out of hand. Thus it is difficult to find appropriate measures of the 'correctness' of synthesised intonation contours. As a first approximation to such a measure, we have used the F0 of the original utterance as a yardstick. However, the usefulness of this method is limited, as in no sense is the precise F0 of an original utterance to be taken as canonical. It is in this respect that the notion of perceptual equality is particularly useful. Two utterances that are perceptually equal in their intonation patterns can be said to be linguistically equivalent, carrying the same prosodic connotations. The synthesised utterances subjected to the process described in this paper seemed, on informal listening, to meet this criterion (in fact, in a few cases, the original and the rule-synthesised version were effectively indistinguishable). To establish the bounds of perceptual equality, however, more formal listening tests are required.

5 Beyond synthesis from annotated text

Having chosen a theoretical model for the representation of intonation, and having concluded that the units it provides are in fact of use in synthesising intonation, it is necessary to consider whether the model is capable of being related to other components of the grammar for the purposes of intonation synthesis from unannotated text. Bachenko et al. [1] outline a method of using the (surface) syntactic structure of an utterance to derive the prosodic representation, taking into account the syntactic constituent structure, grammatical function (head-modifier, etc.), and constituent length. In the context of a text-to-speech synthesis system, a syntactic parsing module will yield a syntactic representation giving the class of each word and the constituent structure (it is assumed that there will be no means of deriving semantic information, as the input will not be annotated in any way). The syntactic representation would be tagged with grammatical function to indicate the most likely points for intonational breaks (here interpreted as tone-unit boundaries). For Bachenko et al., there are four types of grammatical relations: these are shown below in order of strength, where the first is the most likely to cause an intonational break.

1. Sentence and adjunct: e.g. *Insert unit into correct shelf location - per detail instructions*
2. Subject and predicate: e.g. *The 48-channel module - has two di-groups*
3. Head and complement: e.g. *has - two di-groups; shows - you - how to fly your kite*
4. Head and modifier: e.g. *the echo cancelers - that are in that shelf; that are - in that shelf*

Some preliminary work has begun on specifying a prosodic representation according to criteria such as these, and the results indicate that it is indeed possible to use the type of model described above to derive intonation from syntactic structure. In this exercise, the criterion of success cannot be a match to the intonation of a particular token of that utterance, as there is no reason why the underlying prosodic representation should be the same in each case. What is required is that the intonation so derived should be at least a plausible pattern for that particular utterance, in that a listener should not need to stretch the bounds of possibility to make intonational sense of the resulting synthesised output.

At this stage, the most it is reasonable to aim for is a relatively neutral style of intonation without significant emotional colouring. Although it is debateable whether any intonation can truly be said to be 'neutral', it is a necessary idealisation in the present situation, where the relationship between syntax and prosodic structure is the least well understood aspect of intonation. In this respect the Spoken English Corpus described above is of great value, as it contains a large proportion of unemotional speech. It therefore provides data for the development of a basic intonational model which could then form the core of a more fully-specified theory of intonation that accounts also for emotional variation.

References

[1] Bachenko, J., Fitzpatrick, E. & Wright, C.E. (1986) The contribution of parsing to prosodic phrasing in an experimental text-to-speech system. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics.*

[2] Crystal, D. (1969). *Prosodic Systems and Intonation in English.* Cambridge: CUP.

[3] de Pijper, J.R. (1983). *Modelling British English Intonation.* Netherlands Phonetics Archives, Vol. III. Dordrecht: Foris Publications.

[4] O'Connor, J.D. & Arnold, G.F. (1961). *Intonation of colloquial English.* London: Longman.

[5] Pierrehumbert, J.B. (1980). The phonology and phonetics of English intonation. Unpublished Ph.D. dissertation, MIT.

[6] Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation.* Lund Studies in English, no. 56. Lund: C.W.K. Gleerup.

[7] Willems, N. (1982). *English intonation from a Dutch point of view.* Netherlands Phonetic Archives, Vol. I. Dordrecht: Foris Publications.

[8] Williams, B.J. & Alderson, P.R. (1986). Synthesising British English Intonation using a Nuclear Tone Model. IBM UKSC Report no. 154.