Proceedings of The Institute of Acoustics

SYNTHESISING BRITISH ENGLISH INTONATION USING TONETIC STRESS MARKS

B.J. Williams and P.R. Alderson

Speech Research, IBM UK Scientific Centre,
Athelstan House, St Clement Street, WINCHESTER, SO23 9DR

Abstract

A recent popular method of intonation synthesis-by-rule (Pierrehumbert 1980, Liberman & Pierrehumbert 1984) makes use of linguistic units (pitch accents) derived from the 'American' school of intonation analysis. An alternative intonation synthesis-by-rule method is described that makes use of the framework of the 'British' school of intonation analysis. The input to the method comprises long and fluently-spoken natural utterances, intonationally transcribed. Rules convert the intonational units to 'target values' on a ten-level scale. A declining topline and level baseline are then superimposed. The resultant F0 contour is compared with the original. The close match obtained suggests that this theoretical model is a valid starting-point for the synthesis of F0 variation.

1 Introduction

Much work has been done to date on the synthesis of the segmental features of speech. However, comparatively little work has concentrated on the synthesis of the suprasegmental features of speech, in particular intonation. This paper addresses the problem of intonation synthesis in the context of a text-to-speech system. The work reported here is mainly intonation synthesis by rule - that is, from prosodically-annotated input text. At a future stage, it is hoped to integrate this work into the overall system, such that the initial prosodic representation itself will be derived from the (surface) syntactic parse, plus information on illocutionary force, grammatical function, etc.

It is necessary to specify pitch variation correctly, where 'correctly' refers to a linguistically appropriate representation. It is possible, using simple resynthesis techniques, to reproduce the pitch variation of a natural utterance in such a way that the output is barely distinguishable from the original. Thus the equipment exists for generating natural-sounding intonation: it is a question of formulating the phonological representation of intonation in a way that accounts both for the (relevant) phonetic detail, and the linguistic organisation of intonation. It is to this question that the work reported here addresses itself.

2 Two schools 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'. A brief summary of each will be given here: for fuller details, see Ladd, chapter 1 [1].

2.1 'American' school:

The characteristic 'American' approach to suprasegmental analysis may be exemplified by the work of Trager and Smith [2], where pitch variation is accounted for in terms of four

pitch phonemes, or levels. Thus, the sequence of pitch phonemes *3 1 2* is for them phonemically distinct from the sequence *4 2 3*, even though both describe a falling-rising pattern. This is because the contours formed by movement between the four pitch levels are seen as merely allophonic. A modern development of this use of pitch levels is seen in the work of Liberman [3], who recognises four pitch phonemes: High, High-Mid, Low-Mid and Low. This approach has been taken further in the work of Pierrehumbert [4], who recognises just two pitch levels: High and Low.

2.2 'British' school:

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 (the pitch level approach, whilst aiming at greater theoretical sophistication, is a cumbersome tool for transcription). This pitch contour-based approach may be exemplified by the work of O'Connor and Arnold [5], 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***. Any post-nuclear syllables are referred to collectively as the ***tail***. This approach emphasises functional relevance, perhaps at the expense of phonetic explicitness in terms of levels: the reverse is the case for the 'American' approach. A later development of this approach is that described by Crystal [6]. This system has been used by Crystal for the transcription of a corpus of approximately 30,000 words of educated British English, the transcriptions being checked by two linguists. This analysis, framed overwhelmingly in formal rather than functional terms, takes the basic units of intonational variation to be contours rather than levels, as in the work of O'Connor and Arnold.

2.3 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 Williams and Alderson [7]. The system as a whole closely parallels that found in the work of other 'British school' linguists. It is in the detailed designation of the units that points of difference emerge. The basic units of the system are set out in Figure 1, together with the symbols used to transcribe them.

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. If there are no following unaccented syllables, then the pitch movement is realised as a pitch glide. 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

		Tone-unit boundaries			
Major:	‖	Minor:	|		
		Accented syllables			
Fall:	ˋs, ˎs	Rise:	ˊs, ˏs	Level:	¯s, ˍs
Fall-rise:	ˇs, ˬs	Rise-fall:	ˆs, ‸s		
		Unaccented syllables			
Booster:	↑s	Drop:	↓s	Stressed:	·s

Figure 1. Intonational units used

division, thus allowing a generalisation to be stated in terms of the five accent types shown in Figure 1. 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 immediately preceding syllable.

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.

A 'British school' system was chosen, rather than an 'American school' system, because the former type of system 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 (Crystal, Svartvik *et al.* [8], and the corpus described below). The 'American school' type of system has not been as extensively used for this purpose. Therefore it was felt that the former type of system was more likely to reflect all and only the linguistically-significant pitch movements of (British) English. This type of system was used also by Young and Fallside [9], but in an abbreviated and less flexible form. They report that the quality of the output prosody is 'quite good'. This finding would support the view that the type of intonational representation used is capable of accounting for the linguistically significant features of suprasegmental variation.

3 Spoken English Corpus

The intonational model described above is being used for the prosodic analysis of a corpus of contemporary spoken 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 extracted from it and transcribed orthographically. This transcript, which contains no punctuation marks of any kind, is then 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, who have together formulated the intonational system described above. Consistency is monitored by regular checking of short independently-transcribed identical passages, and by comparison of these with a plot of the fundamental frequency (F0) in the case of discrepancies. 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 (end of September 1986) approximately 30,000 words have been transcribed prosodically. The finished corpus is expected to contain 50,000 words, all prosodically transcribed. It is hoped to analyse the relationship between the intonational transcription and the (surface) syntax of the texts.

4 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: ***Every morning, long queues are forming outside courts three one eight and five o nine.***

4.1 From the prosodic transcription to 'target values'

Using the (manually-assigned) prosodic transcription shown in Figure 2 as input, each syllable was then assigned one or more ***target values***. These are integer values between 1 and 10, representing an abstract scale of linguistically-relevant pitch height (i.e. before the effects of declination are seen in the pitch curve). These target values are similar to those in Pierrehumbert [4], which are decimal values to one place of decimals. The difference lies in the fact that Pierrehumbert's target values are ***a)*** assigned only to accented syllables: unaccented syllables are allowed to drift down or up as appropriate; and ***b)*** a direct reflection of the metrical prominence of the accented syllable, as determined from the metrical tree of the utterance.

ˏEvery ˏmorning | ˍlong ˇqueues are ·forming | ˍout·side ·courts
ˇthree one eight | and ˍfive o ˋnine ||

Figure 2. 'Every morning...': prosodic transcription

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-rise is assigned an initial target value that is three greater than that of the immediately preceding syllable within the same minor tone-unit, and a subsequent target value that is up to six less than the initial value, but with a minimum value of 1, while its final target value is three greater than the second value. 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 intervening ones being interpolated. The target values assigned to the example sentence are shown in Figure 3.

```
ˏEvery   ˏmorning | _long ˇqueues are    ·forming | _out·side·courts
 2  5     1     4    2      5         1    1      4    2    2    2

ˇthree one eight |and _five o    ˋnine ||
 5       1   4      4    1    1   4, 1
```

Figure 3. 'Every morning...': target values

4.2 *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 that in Pierrehumbert [4]: i.e. superimposing an overall pitch envelope that incorporates declination. In this case, 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 for the value of the topline at the end of the minor tone-unit. This 'somewhat lower' is a constant proportion used also for a third and subsequent minor tone-units within the same major tone-unit. Thus the effect is an exponential decline in topline reset values over the course of a major tone-unit.

For the purposes of the present experiment, 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, as in Pierrehumbert [4]. Since the envelope declined over time, the 'graph paper' for the target values decreased progressively in range. The program superimposing the declination envelope converted each target value to a frequency value in Hz. The frequency values assigned to the example utterance are shown in Figure 4.

Every	morning	long	queues	are	forming	outside	courts
148, 198	126, 166	143	182	123	122, 156	140, 137	134

three	one	eight	and	five	o	nine
163	120	149	158	121	120	144, 113

Figure 4. 'Every morning...': frequency values 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. Using this many LPC coefficients gave synthesised output of a very high quality.

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 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 output of the above processes is shown in Figure 5, where it is plotted against the F0 of the original utterance after LPC resynthesis. The rule-synthesised F0 is shown after the application of the F0 perturbations mentioned above.

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 the discrepancies seen on the F0 plot in Figure 5 were not in fact perceptually salient. These discrepancies could be heard only on careful listening and in full knowledge of what to listen for. This suggests that the attempts by Liberman & Pierrehumbert [10] to match as precisely as possible to the original F0 may perhaps be misplaced. It seems that a more useful metric is that of the ***perceptual equality*** of two F0 contours, as used by the 'Dutch school' of workers on intonation synthesis (e.g. Willems [11], de Pijper [12]). Their 'perceptual equality' is based on linguistic and auditory indistinguishability, rather than on the acoustic identity sought by Liberman & Pierrehumbert. Since no two utterances of the same sentence are ever completely identical acoustically, the notion of perceptual equality may well prove to be of great value in the assessment of synthesised speech.

4.3 In longer utterances

A few other sentences were subjected to the same process. Two of these were approximately nine seconds long. The additional sentences were taken from different texts, using different speakers. It will be noted that the length of the utterances experimented with - between about six and nine seconds - reflects the length of utterances found in real speech. These are not short, laboratory-manufactured sentences. The use of longer stretches of speech provides a more rigorous test of the method. If the method is successful for data of this kind, then there is reason to believe that it will also be successful over an unrestricted range of input text in the final text-to-speech system.

The match between rule-synthesised and original resynthesised F0 was found to be quite good. As in the first example above, the match was better when heard by the ear than when seen by the eye. In fact, a listener hearing the original and rule-synthesised versions for the first time, with no access to the F0 plot, would find it most difficult to tell them apart. This suggests that the two types of F0 contour are perceptually equivalent, and thus that the linguistic units forming the basis for the synthesis by rule are in fact valid ones.

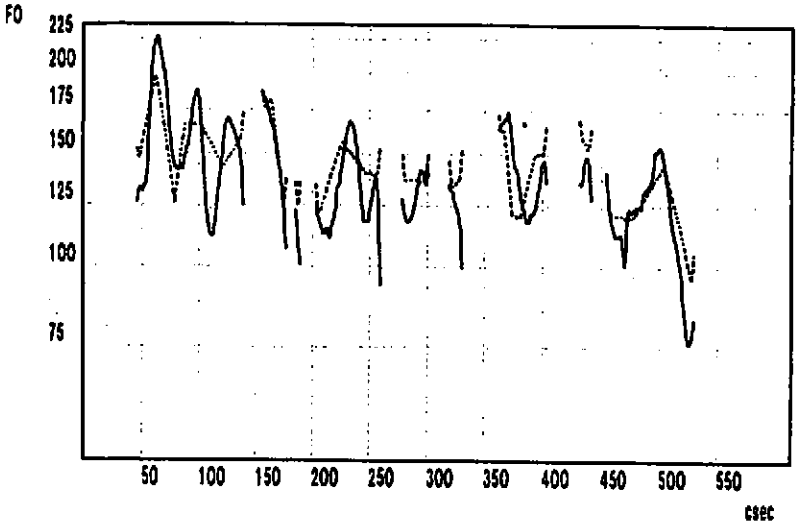

Figure 5. **'Every morning...': original resynthesised F0 vs. F0 synthesised by rule:** Solid line = F0 of resynthesised original utterance: hatched line = F0 of utterance synthesised by rule from prosodic transcription.

5 Discussion

The investigations reported above have implications for the theory of intonation in English as well as for intonation synthesis. An attempt has been made to use a system which is capable of relating pitch movement to linguistic function in a transparent way. The results so far support the view that the system chosen is able to account for all and only the linguistically-relevant features of pitch movement.

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. As every utterance of the same sentence will be acoustically different, there is little point in attempting to match exactly to a particular token, and lack of such an exact match is not to be taken as an error. It is in this respect that the notion of ***perceptual equality*** (see discussion in section 4.2) is particularly useful. It has been used by workers in the 'Dutch school' of intonation analysis to validate their synthesised intonation contours as valid descriptions (see Willems [11], de Pijper [12]). 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 described in the work reported above seem, on informal listening, to meet this criterion. This is because many of them meet the more stringent criterion of effective indistinguishability: they can only be distinguished given careful listening on the part of a listener who is aware of what to listen for. Since these listening conditions are in no way characteristic of real speech, it can safely be said that the utterances in question are perceptually equal. However, a controlled series of perceptual experiments is necessary, to investigate the nature of perceptual equality in relation to utterances synthesised using the pitch contour model described above. It is hoped shortly to carry out such experiments.

References

[1] D.R.Ladd, 'The Structure of Intonational Meaning', Bloomington: Indiana University Press, (1980).

[2] G.L. Trager, & H.L. Smith, 'Outline of English Structure', Norman, Okla.: Battenburg Press, (1951).

[3] M. Liberman, 'The Intonational System of English'. Ph.D. dissertation, MIT. Distributed by Indiana University Linguistics Club, Bloomington, Indiana, USA, (1978).

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

[5] J.D. O'Connor & G.F. Arnold, 'Intonation of colloquial English', London: Longman, (1961, 2nd. edition 1973).

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

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

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

[9] S.J. Young, & F. Fallside, 'Synthesis by rule of prosodic features in Word Concatenation Synthesis', International Journal of Man-Machine Studies, Vol. 12, 241-258, (1980).

[10] M. Liberman & J. Pierrehumbert, 'Intonational Invariance under Changes in Pitch Range and Length', In: M. Aronoff & R.T. Oehrle (eds.), 'Language Sound Structure', Cambridge, Mass. and London: MIT Press, (1984).

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

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

Acknowledgements

We should like to thank Dr. Gerry Knowles and Mrs. Lita Taylor of Lancaster University for their invaluable collaboration in the Spoken English Corpus project.