Abstract

The synthesis of intonation in a text-to-speech system has long been a neglected area. Recently, work by Pennebaker 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 rep- resentation of intonation that has been used successfully in train- ing English as a second language.

The work reported here is an attempt to blend the two approaches in the presentation of intonation that has been used successfully in trans-cribing spoken (British) English and in teaching intonation to foreign learners of English.

The work reported here is an attempt to blend the two approaches in the presentation of intonation that has been used successfully in trans-cribing spoken (British) English and in teaching intonation to foreign learners of English.

The input text contains a linguistic representation of intonation, using units known as 'pitch accents'. These units are converted to a list 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 declination curve, the parameters of which are de- pendent both on the speaker model used and on the higher-level destination currently in force. The frequency values are added to the segmental information, and the result is output as speech.

1.1 Adapted 'British' system

The work reported below makes use of a model of intonation based on the work of other 'British school' linguists. The basic units of the system are shown in Figure 1 below. The intonational model described above is being used for the

2. Spoken English Corpus

The intonational model described above is being used for the

3. From the prosodic transcription to 'target values'

Using the manually-annotated prosodic transcription shown in Figures 1 and 2 in input, syllables were then assigned target values. These values are between 1 and 10, representing an abstract scale of linguistically-relevant pitch height. These target values are similar to those in [5] and are used for the final contour. Each unaccented syllable had a target value, while those carrying pitch glides possessed two or three as appropriate. In addition, each stressed syllable and accented syllable is introduced with a (symbol) (i.e. just before the next tone-unit stress mark, or before a tone-unit boundary) giving a target value.

4. Discussion

The investigations reported above have implications for the way in which the synthesis of intonation is approached. An attempt is made to use a theoretical model which expresses those prosodic 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.

The assessment of intonation contours is peculiarly difficult, as it is rare for these to be definitely 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 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 case 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 information. 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 versions 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 useful 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 un annotated 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 backshift-squeeze synthesis system, a syntactic parsing module would 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 output 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 (there 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 adverb: e.g. Insert unit into correct shelf location - per dealer 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 cancellers - that are in that shell: that are - in that shell.

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 spoken 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 synthesized 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 possible that neutral intonation can truly be said to be 'neutral', if it is a necessary idealisation in the present situation, where the relationship between syntax and prosodic structure is not as fully understood as aspect of intonation. In this respect the British 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


