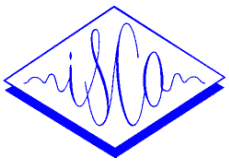

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A New Model of Intonation for use with Speech Synthesis and Recognition

Paul Taylor and Stephen Isard
Centre for Speech Technology Research
University of Edinburgh
U.K.
email: pault@cstr.ed.ac.uk

Abstract

This paper describes a synthesis from analysis scheme for producing natural sounding intonation for speech synthesis. The paper presents a new method of describing F_0 contours in terms of three basic phonetic intonation elements. Details are given of an automatic system for labelling F_0 contours, which could be used for speech recognition purposes. Current work on extracting a phonological description from this phonetic description is discussed.

1 Introduction

Our current text-to-speech system uses synthesis from analysis techniques for both the duration and the segmental components of the the system. The segmental component uses diphones spoken by a single speaker and is able to capture a good likeness of that speaker's voice quality. The duration component makes use of a database of phone and syllable durations to model a speaker's durational characteristics. These two systems provide models which link high level phonological descriptions to low level descriptions. The diphone synthesizer takes a high level phoneme description and produces a low level waveform output. The duration system uses information on stress, accentuation, phrasing etc as input and produces a lower level millisecond duration for each phone.

This synthesis from analysis paradigm is useful in many ways. First, it aims for a high degree of naturalness in that it tries to model a particular speaker's voice. To model a new speaker's voice, all one need do is collect and analyse the appropriate data from that speaker. If automatic analysis techniques are available, this allows coverage of a wide range of the speaker's voice. Both the diphone and the duration techniques avoid having human analysts specify the values of numerical parameters: the broad method is determined by the human designer, the numerical detail is supplied by computer analysis. If sufficient data is available, specific context sensitive effects can be modelled, which increases the naturalness of the system.

The aim of the current project was to use this analysis/synthesis paradigm for intonation. A database was designed that had adequate coverage of all the intonational effects of the language to be synthesized. A database was designed that would be recorded by the chosen speaker and then analysed to determine how the speaker realised high level phonologcial descriptions as F_0 contours. An automatic analysis system would save on human labelling time and provide consistency in labelling criteria. Once this intonation system had been developed, it could be used in synthesis mode: given a high level description an intonation contour could be produced.

While developing this system it was seen that the automatic labelling system could serve as part of an intonation module in a speech recognition system. Although the primary goal was to develop and automatic method of labelling F_0 contours for analysis/synthesis, the system was developed with an eye to recognition as well.

1.1 An Analysis/Synthesis System for Intonation

In our system, five levels of representation were used.

1. **Acoustic Waveform**

2. **F_0 Contour**
   Fundamental frequency (F_0) is most often measured from digitally sampled waveforms using automatic analysis algorithms.

3. **Phonetic Description**
   The phonetic description is a description of the contour as a sequence of discrete units.

4. **Normalised Phonetic Description**
   The normalised phonetic description is a discrete description which has been normalised to filter out phonemic effects. Phonemic effect include differences in F_0 depending on vowel type, and differences in accent position depending on syllable structure [1].

5. **Phonological Description**
   The phonological description is a qualitative abstract description of the intonation of an utterance.

Phonological intonation schemes that are often used include that of Pierrehumbert [2], who describes intonation in terms of high and low tones, and that of the British school [3] [4] which uses rises and falls. Phonetic descriptions include that of Ladd [5] whose phonetic model implements Pierrehumbert's phonology, and Isard and Pearson [6] who implement the British school. Others have designed phonetic descriptions that are not linked to any particular phonology, these include Fujisaki [7] and the Dutch School [8].

The aim of this project was to design a system for automatically linking these descriptions for both synthesis and analysis purposes. Automatic analysis is desirable as it allows large amounts of data to be analysed and it standardises the analysis process.

Four main systems were planned:-

1. **Signal Processing**
   Extracting F_0 from the acoustic or laryngograph waveform.

2. **Contour Analysis**
   Extracting the phonetic description. from the F_0 contour

3. **Normalisation**
   Filtering out the phonemic effects so as the contour is normalised with respect to phoneme information.

4. **Phonological Abstraction**
   Deriving the phonological description from the phonetic one.

1.2 General Aims

From the outset it was considered desirable to ensure that any of the methods or descriptions used were general enough to analyse and synthesize all the intonational effects of English. Many TTS systems deal exclusively with a neutral declarative type of intonation [5], [1]. Although our system can only produce neutral declarative phonological intonation descriptions from text analysis, it is hoped in the future to extend this to other types of utterance such as yes/no questions. Also we wanted to ensure that the model could analyse intonation in a wide variety of contexts so that any intonation contour would be analysable within the system.

10.21437/ICSLP.1992-346

It was also a key aim that the phonetic description should be as accurate as possible, and thus be able to capture detail in Fo contours. An analysis/resynthesis test can be used to test the accuracy of a model. A contour that has been analysed into its phonetic description can be resynthesised. By comparing the analysed and resynthesized versions one can assess the accuracy of the model.

Various phonetic and phonological description systems were looked at to assess their suitability for being part of an algorithmic system that had a wide coverage of English intonation and was capable of capturing detail in F_0 contours.

The following sections describe the analysis system as it stands at present.

2 Signal Processing: F_0 Extraction

F_0 extraction from waveforms is difficult and the resultant F_0 contours are seldom 100% accurate. In order to obtain more reliable F_0 contours, a laryngograph was used. This is a device which measures the impedance across the vocal folds. The waveform produced from this device can be pitch-tracked much more easily than standard acoustic waveforms. This device is not practical for every situation, but is a very useful means of extracting F_0 when recording conditions can be controlled.

3 Phonetic Descriptions and F_0 Analysis

Two existing phonetic models were examined to assess their suitability for our purpose.

3.1 The Dutch Model

The Dutch intonation model [8][9], models F_0 contour by stylisation and standardisation. Stylisation involves describing F_0 contours in terms of a small number of straight lines. The standardisation process then classifies this description in terms of a system where the F_0 contour either follows a high, mid, or low declination line, or is moving between these three lines.

The Dutch model has two major drawbacks. Firstly, the straight line approximations often result in resynthesized contours that are quite different from the original. While it is argued that the straight line approximation may be perceptually equivalent to the original, this makes automatic analysis more difficult. It is difficult to model a curve with a straight line: many different possible fits of equal distance may be found for a single curve. The second problem arises with the strict use of the three declination lines. It has been shown that speakers can easily use up to five different levels when speaking and any system that makes use of a limited set of levels may run into problems [10].

3.2 The Fujisaki Model

Fujisaki proposed an interesting model of intonation whereby F_0 contours where generated by passing step functions and impulses through second order critically damped filters. For many accents of English, this model can produce an F_0 contour that is very close to the original. Unfortunately, it was very difficult to extend the model to deal with low accents or slowly rising sections of contour. Modifications of the model were tested, but all either failed to produce acceptably accurate F_0 contours, or needed input requirements that were just "hacks" and would be very difficult to link to a phonological description.

3.3 A New Phonetic Description

As neither the Dutch model or the Fujisaki model seemed suited for our purpose, a new phonetic description was developed.

The new model describes F_0 contours in terms of a linear sequence of non-overlapping elements. Three types of element exist: the *rise element*, the *fall element* and the *connection element*. The shape of the rise and fall elements is given by a mathematical function which can be scaled on the x and y axis to fit the F_0 contour. The normalised version of the shape (described between (0, 0) and (2, 2)) has zero tangent at y = 0 and y = 2, passes through the point (1, 1) and has a variable degree of curvature. A curve which has this property is defined below.

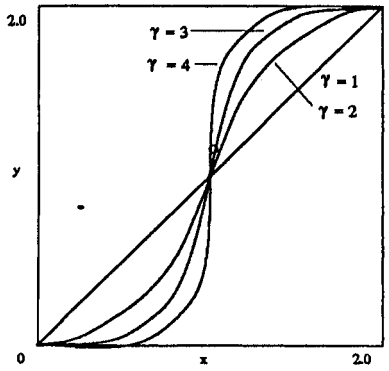

Figure 1: The polynomial functions used for rise and fall shapes. As γ increases the curve becomes less like a straight line and more curved.

$$\begin{aligned} 0 < x < 1 \quad & y = x^{\gamma} \\ 1 < x < 2 \quad & y = 2 - (2 - x)^{\gamma} \end{aligned} \tag{1}$$

Equation 1 is shown in figure 1 for different values of γ. Figure 1 shows the shape for a rise element: a fall element is the same shape but reflected in the y axis.

The *connection* elements are modelled by straight lines. Connection units can be used between pitch accents and at the starts and ends of phrases. They can be of any length and gradient.

Usage

Pitch accents are modelled by using the rise and fall elements in a variety of ways. The most commonly found pitch accent in our data was the *fall* accent which is described as **H*** or **H* + L** by the Pierrehumbert system. This accent is realised as a rise element followed by a fall element. Figure 2 shows three different fall accents.

Fall elements always occur in relation to a pitch accent. Rise elements are used in pitch accents but may be also used at the beginnings and ends of phrases.

More complex pitch accents can be modelled by using connection units and by using the rise units at the phrase boundaries.

Performance

A database of 64 F_0 contours was designed and recorded that covered a wide range of pitch accent types, a wide range of nuclear accent positions, and a variety of phrasing situations. These contours were hand labelled by dividing the utterance into sections, where each section was marked as being of a particular type of element.

A synthesis program was developed that could reconstruct F_0 contours given this labelling. The resynthesised F_0 contours were compared to the original. Figures 3 and 4 show real F_0 contours compared with the reconstructed versions. In all cases a satisfactory fit could be achieved, although sometimes the placing of the boundaries between two elements was somewhat arbitrary.

The closeness of the fit using this new model was substantially better than that of the Dutch model, and this system was able to model a greater variety of accents than the Fujisaki system.

4 Automatic F_0 Labelling

An automatic labelling system was built to label F_0 contours.

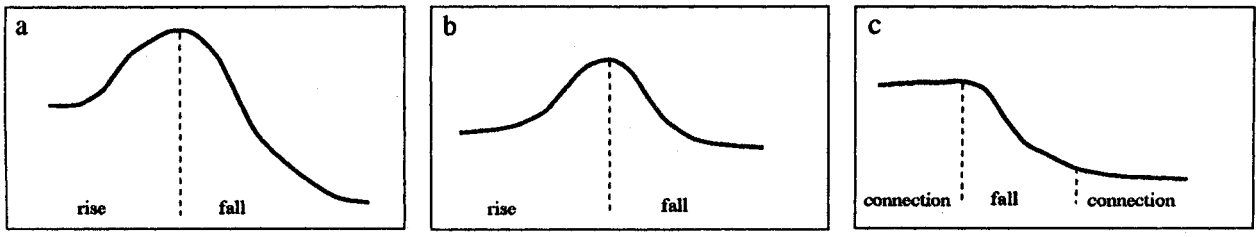

Figure 2: Three typical fall accents shown with different rise and fall combinations. (a) is a typical nuclear fall, (b) is likely to be found in a prenuclear position and (c) is often seen in sequences of downstepping accents.

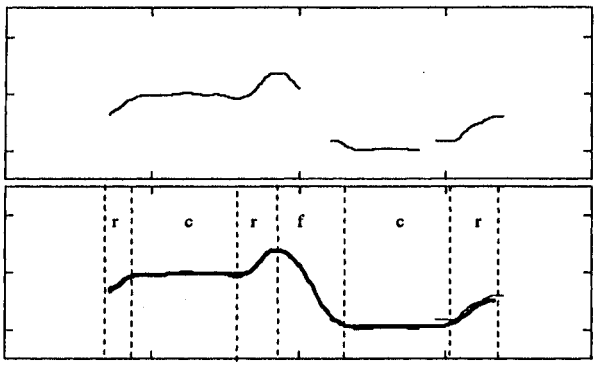

Figure 3: The top graph shows the F_0 contour, the bottom graph shows the same F_0 contour but with the reconstructed F_0 contour shown superimposed in bold. The utterance is "Do you *need to win everything?" (* denotes accented word).

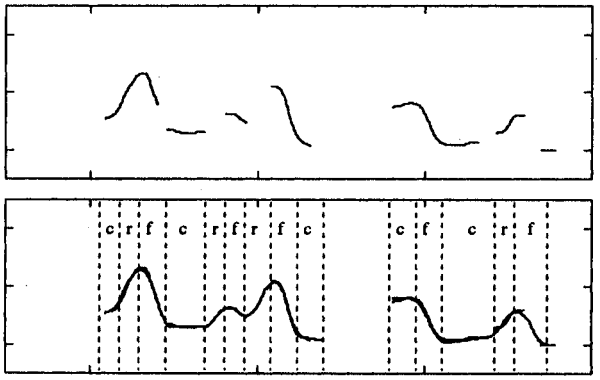

Figure 4: The top graph shows the F_0 contour, the bottom graph shows the same F_0 contour but with the reconstructed F_0 contour shown superimposed in bold. The utterance is "The large window stays closed, the small one you can open"

4.1 Contour Preparation

Intonation recognition is difficult due to problems with the F_0 contour itself. The contour is not continuous as there is no fundamental frequency in the unvoiced regions. F_0 is also dependent on segmental effects - before and after stops the contour can deviate sharply. F_0 tracking is difficult and prone to errors - and even when successful, there is a considerable amount of short term variation in the F_0 contour due to pitch perturbations.

It is only possible to perform very crude analysis with the F_0 contour in this form. To filter the intonation information from the F_0 contour smoothing is performed, using a simple median smoother. After this, straight lines are interpolated through the unvoiced regions. Finally, the contour is smoothed again to make the boundaries between the interpolated regions and the original curve continuous. Inevitably, this smoothing distorts the F_0 values of the contour somewhat, which results in the contour being slightly flatter than before. However, in the worst cases the smoothed version deviates from the original by about 10 Hz. The degree of smoothing (21 point median smoothing) is a compromise between obtaining a smooth continuous contour that is suitable for automatic analysis, and a contour that stays close to the original F_0 contour. Even with this heavy smoothing, segmental effects ("bumps") were still present in the smoothed F_0 contour.

4.2 Broad Class Frame Labelling

The smoothed contour is divided up into 100 ms frames whose value was the average frequency values of the surrounding F_0 contour. Each frame is then compared with its neighbours to calculate the rate of change of the contour at that point. Two statistically calculated thresholds are used to determine whether the contour is rising sufficiently fast to be part of a rise element, or is falling fast enough to be classed as a fall element. Each frame is labelled rise, fall or unlabelled. Any adjacent frames of the same class are grouped together. Thus the contour is divided into labelled *sections*.

These sections are then examined to see if they can be further grouped. For example two long rise sections separated by a short unlabelled section will be classified as a single larger rise section.

4.3 Optimal Matching

The broad classification divides the contour into sections which are labelled rise, fall, and unlabelled. The boundaries of the sections are very imprecise due to the size of the 100ms frames. The next stage in the labelling procedure is to optimally fit the rise and fall shapes to the designated sections and determine the precise boundaries between the sections.

If an area is marked as a fall, start and end regions are defined around the boundaries of the labelled section. The start region begins 50ms before the marked boundary and ends 25% into the section. The end region starts 75% into the section and finishes 50ms after the boundary. Every possible fall shape that fits within this region is calculated, and the one showing the smallest normalised euclidean distance from the F_0 contour is chosen. The same procedure is used for matching rise sections.

After the optimal matching process, all the sections labelled rise and fall will have their boundaries marked. The remaining unla-

belled sections are designated as connection elements. No direct matching of connection elements is done as such, but it has been found that although F_0 contours between accents may not be exactly linear, the approximation usually gives a good fit.

The output of the analysis is a list of elements with start times, durations and amplitudes.

4.4 F_0 Analysis: Assessment

By comparing the parameters of the automatic segmentation with those of the hand segmentation program, it is possible to judge how successful the recogniser is. The first experiment was to compare the hand labelling of the 64 F_0 contours with the automatically labelled versions. In this test, half the data was used to determine the rise and fall thresholds, and half the data was used for an open test.

The assessment criteria was based on two types of errors. The first was concerned with inaccuracies in labelling boundaries. For every 10 ms that the boundaries disagreed, a penalty was incurred. The second type of error concerned insertion, substitution and deletion errors. These incurred a larger penalty.

Somewhat arbitrarily, a penalty of 3.0 was used for a insertion etc error, and a penalty of 0.1 was used for each 10ms of boundary error. By using these criteria, a score was calculated for each utterance. A perfect fit would receive a score of 0.0. The worst case is more difficult to calculate, but was simulated by comparing two correct segmentations of different utterances. This gave a score of about 100. In the database of 64 contours the best received a fit of 0.0, and the worst 16.72. There seemed to be no significant difference between the results obtained from the closed and open tests.

A second database of 45 long sentences and 30 short paragraphs from a different speaker is currently being analysed, although no performance figures are available yet. However, it does seem that although good labelling can be achieved using the first speaker's rise/fall thresholds, performance improves when the thresholds are trained on that speaker. Thus the system is not totally speaker independent. Work needs to be carried out in quickly training the system on a small amount of data.

Other data which has not been recorded using a laryngograph has been analysed also. Here the performance is considerably worse, almost entirely due to F_0-tracking errors. Work needs to be carried out to improve the F_0-tracking algorithms and making the F_0 analysis procedure more robust.

5 Phonemic Normalisation and Phonological Abstraction

No work has yet been carried out on phonemic normalisation. Most phonemic effects (such as intrinsic F_0 vowel height) are small enough to be ignored for purposes of automatically determining the phonological description of a contour from the phonetic description. Where phonemic normalisation will be needed most is when the precise sizes of rise and fall sections are needed as data in the synthesis system.

As with phonetic descriptions, many phonological descriptions were looked at to see which would best describe the intonation. The most commonly used intonational phonology today is that of Pierrehumbert [2] which uses two phonological tones to describe intonation. From the experiments and analysis carried out, it was judged that the process of relating the new phonetic description to Pierrehumbert's system would be difficult. It was often clear that phrases did not end in distinct high or low tones but rather there were many phrase endings, each with subtly different meanings. Also the use of two tones to describe accents and the use of Pierrehumbert's dipping interpolation between two H* accents seemed unattractive.

Hence a new phonological description is being developed which classifies pitch accents into two main categories, high and low, but uses features to distinguish variations within these accent classes in a way similar to that of Ladd [11]. Compound accents may be realised by using connection elements and phrase final rise elements.

The phonological abstraction process seems much easier than the F_0 analysis process. Both the phonetic and phonological descriptions are discrete and often there is a one to one mapping between the symbols. A high accent is most commonly realised by a rise followed by a fall; low accents typically have fall elements preceding them.

6 Conclusion

This paper has described a framework for analysing and synthesising intonation. The work carried out so far has concentrated on the F_0 analysis process, and the initial results from the automatic labeller look very promising. Work still needs to be carried out in this area to make the system more robust against noisy F_0 contours. The phonological abstraction process is the subject of current work, but the basis of a system has been developed.

It is hoped that this system will greatly improve the intonation in the synthesizer by making using a wide range of accents and by producing contours that are similar to real F_0 contours. It is also hoped that the intonational characteristics of different speakers will be captured thus adding to the naturalness of the system.

References

[1] K. Silverman, *The Sructure and Processing of Fundamental Frequency Contours*. PhD thesis, University of Cambridge, 1987.

[2] J. B. Pierrehumbert, *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980. Published by Indiana University Linguistics Club.

[3] M. A. K. Haliday, *Intonation and Grammar in British English*. Mouton, 1967.

[4] J. D. O'Connor and G. F. Arnold, *Intonation of Colloquial English*. Longman, 2 ed., 1973.

[5] D. R. Ladd, "A model of intonational phonology for use with speech synthesis by rule," in *European Conference on Speech Technology*, ESCA, 1987.

[6] S. D. Isard and M. Pearson, "A repertoire of British English contours for speech synthesis," in *SPEECH '88, 7th FASE Symposium*, FASE, 1988.

[7] H. Fujisaki and H. Kawai, "Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation," in *Working Group on Intonation, 13th International Congress of Linguists, Tokyo*, 1982.

[8] J. t'Hart and A. Cohen, "Intonation by rule: a perceptual quest," *Journal of Phonetics*, vol. 1, pp. 309–327, 1973.

[9] N. J. Willems, "A model of standard English intonation patterns," *IPO annual Progress Report*, 1983.

[10] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language Sound Structure* (M. Aronoff and R. T. Oehrle, eds.), MIT Press, 1984.

[11] D. R. Ladd, "Phonological features of intonation peaks," *Language*, vol. 59, pp. 721–759, 1983.