The Rise/Fall/Connection Model of Intonation

Paul Taylor
Human Communication Research Centre

Human Communication Research Centre,
University of Edinburgh,
2 Buccleuch Place,
Edinburgh EH8 9LW

Note: This work was carried out while the author was employed at the Centre for Speech Technology Research, University of Edinburgh, and at ATR Interpreting Telecommunication Laboratories, Kyoto.

email
pault@cogsci.ed.ac.uk

Abstract

This paper describes a new model of intonation for English. The paper proposes that intonation can be described using a sequence of rise, fall and connection elements. Pitch accents and boundary rises are described using rise and fall elements, and connection elements are used to describe everything else. Equations can be used to synthesize fundamental frequency (F_0) contours from these elements. An automatic labelling system is described which can derive a rise/fall/connection description from any utterance without using prior knowledge or top-down processing. Synthesis and analysis experiments are described using utterances from six speakers of various English accents. An analysis/resynthesis experiment is described which shows that the contours produced by the model are similar to within 3.6 to 7.3 Hz of the originals. An assessment of the automatic labeller shows 72% to 92% agreement between automatic and hand labels. The paper concludes with a comparison between this model and others, and a discussion of the practical applications of the model.

Résumé

Nous présentons dans cet article un nouveau modèle d'intonation pour l'anglais, selon lequel celle-ci est décrite en termes "descentes"(rise), "montées" (fall) et "lignes de connection" (connection). Les accents et les montées de continuation sont représentés par les éléments de "montée" et de "descente" ; les lignes de connection sont utilisées partout ailleurs. Un système d'équations permet de reconstruire le contour de fréquence fondamentale à partir de ces éléments. Nous décrirons ensuite un système d'étiquetage automatique qui associe à toute phrase une représentation descentes/montées/connections sans connaissance a priori, ni analyse descendante. Une série d'expériences portant sur un corpus de phrases prononcées par 6 locuteurs de langue anglaise et d'accents variés valide ce modèle : l'erreur de synthèse est comprise entre 3.6 et 7.3 Hz. Par ailleurs, une comparaison des étiquetages automatiques et manuels montre une correspondance de 72% a 92%. Nous conclurons par une comparaison du modèle proposé et des modèles existants et en discuterons les applications pratiques.

Zusammenfassung

Wir stellen ein neues Model für englische Intonation vor. Wir gehen davon aus daß Intonation sich durch eine Reihe von Steig-, Fall- und Verbindungselementen beschreiben läßt. Pitchakzente und Grenzsteigerungen werden durch Steig- und Fallelemente beschrieben und Verbindungselemente werden benutzt, um alle anderen Phänomene zu beschreiben. Die Fundamentalfrequenzkonturen können mittels Gleichungen von diesen Elementen berechnet werden. Ein automatisches Markiersystem wird beschrieben mit dem die Steig/Fall/Verbindungselementkette automatisch ohne vorheriges Wissen oder top-down Berechnung erstellt werden kann. Synthese- und Analyseexperimente sind beschrieben für Äußerungen sechs verschiedener Sprecher unterschiedlicher Akzente. Ein Analyse-Resyntheseexperiment ist beschrieben, welches zeigt, daß die Konturen um zwischen 3.6 bis 7.3 Hz vom Original abweichen. Eine Bewertung des automatischen Markiersystems zeigt, daß etwa 72% bis 92% Übereinstimmung zwischen den automatischen und von Hand geschriebenen Markierungen besteht. Ein Vergleich mit anderen Modellen und eine Beschreibung wie das Modell benutzt werden kann beenden dieses Paper.

1 Introduction

This paper describes a new model of intonation which accurately synthesizes fundamental frequency (F_0) contours from a linguistically relevant description. An automatic labelling system is described which can derive this linguistically relevant description from any F_0 contour.

Fundamental frequency synthesis algorithms are often based on the principle of describing a set of prototypical F_0 patterns which represent different intonational tunes of the language in question (**?**), (**?**), (**?**), (**?**), (**?**). However, what these models often fail to describe is the *range* of acoustic patterns associated with an intonational tune. For example, for a given type of pitch accent, there can be a very large variation in the acoustic patterns associated with this accent, even when factors such as pitch range are taken into account. Thus prototypical pattern algorithms simply describe a notional centroid in the distribution of the acoustic patterns associated with the pitch accent type, and give little or no explanation of how or why the acoustic patterns for that pitch accent vary.

For speech synthesis purposes, this is often adequate, as a single acoustic pattern is often sufficient for each pitch accent type. Given the fact that only a single pattern is needed, it makes sense to use a prototypical pattern for each pitch accent type. For speech analysis purposes, this paradigm poses problems because the distribution of a class may be complex and it is not a simple matter to identify the pitch accent type of a section of F_0 contour by matching observed patterns to prototypical patterns. The use of this paradigm is made even more difficult because the prototypical patterns given in the literature are often described in ad-hoc ways: no principled (i.e experimental) reasons are given for why the patterns are as they are, and no account is given for what variations may be expected between speakers etc.

An alternative approach is to use a low-level acoustic-phonetic model. Fujisaki (**?**), t'Hart & Cohen (**?**) and Hirst (**?**) report work in developing models of this kind. Models of this kind do not (in the first instance) synthesize prototypical acoustic patterns, rather they are general devices for describing F_0 contours, and as such attempt to synthesize any F_0 pattern that a speaker would be expected to produce. With suitable constraints, these models can also be prevented from synthesizing contours which are impossible for speakers to produce. A price must be paid for this useful flexibility, and these models typically do not take phonological descriptions as input, rather some sort of numerical, phonetic description.

These models take low-level phonetic input and produce F_0 contours, and thus when used in reverse for analysis purposes, can be used to derive low-level phonetic information from F_0 contours. These low-level phonetic descriptions are useful as they help eliminate linguistically irrelevant information in the F_0 contour, giving a compact, linguistically meaningful description for a further system to operate on. The key notions in designing models such as these are to make the model accurate such that it produces all and only the F_0 contours that speaker's produce, and also to try and make the low-level phonetic description as useful as possible for any further processing. Accurate models facilitate simple automatic analysis because the analysis algorithm must only derive the low-level phonetic description for a given F_0 contour, and does not have to deal with the complex mapping to a prototypical pattern. Accurate models are useful for synthesis purposes as they can be relied upon to generate any naturally occurring contour. Finally, the more linguistically relevant the low-level phonetic description is, the more useful the model will be when linked to a phonological system.

This paper describes a new model of this type. Section 2 describes the model itself, and section 3 describes an automatic labelling system which can derive the model's parameters from F_0 contours. Section 4 describes the data we used to test the new model with, and sections 5 and 6 discuss the synthesis and analysis capabilities of the model. Section 7 discusses the need for flexibility when modelling F_0 and compares the new model with previous work.

2 The RFC Model

The model presented here, termed the *Rise/Fall/Connection* Model (RFC for short), was not developed with a particular phonological theory in mind. However, as it is the end goal of this research to eventually provide a complete mapping between acoustics and phonology, it is necessary to show that the RFC descriptions produced by the model do in fact make a substantial contribution to the solution of the overall problem. One can see from a survey of the literature that phonological theories of intonation largely agree on what phenomena need to be described, even if there are differences in the details of classification. Most contemporary theories agree that phonological descriptions of English intonation are primarily concerned with describing behavior of *pitch accents* on stressed syllables and *boundary tunes* at prosodic phrases edges. These theories also state that there are distinct classes of pitch accent and boundary tunes which have distinctive F_0 patterns. These two basic types of phenomena operate with respect to different prosodic units: pitch accents are associated with syllables and boundary tunes are associated with the beginnings and ends of prosodic phrases. This view is more or less consistent with the work proposed by Pierrehumbert (**?**), O'Connor and Arnold (**?**), Halliday (**?**), Ladd (**?**) and others[1]. Grice (**?**) reviews and compares these and other phonological theories.

[1] Pierrehumbert also includes the "phrase accent" as a fundamental unit of intonation. This is not problematic to our baseline notion in that the phrase accent can be included in the general classification of boundary tune, as phrase accents are realised only after the nuclear accent. O'Connor and Arnold make a four way classification of fundamental units. Following from Pierrehumbert and others, we include "head" and "nucleus" to be of the same basic type (pitch accents), and also classify "tail" and "pre-head" as being describable by a mechanisms of boundary tune.

2.1 Modelling F_0 Contours with Equations

From examination of the large number of pitch accents in our data (described in section **??**), it was clear that there were two basic types of F_0 patterns associated with pitch accents. The first type were "peak" or "high" accents, in which the accented syllable was associated with a peak in the F_0 contour. The other type were "trough" or "low" accents, where the converse effect was obvious - i.e. the accented syllable was associated with a trough in the F_0 contour. Peak accents roughly correspond with the H*, H*+L, and L+H* classes of Pierrehumbert (**?**), the H and HL classes of Ladd (**?**) and the fall and rise-fall class of O'Connor and Arnold (**?**). Trough accents are roughly equivalent to the L*, L* +H and H+L* classes of Pierrehumbert, the L class of Ladd and the low-rise and high-rise class of O'Connor and Arnold. This two way peak/trough classification may seem a little unusual to those familiar with the sophisticated phonological intonation inventories just mentioned. However, this division is only intended as a rough classification of the surface acoustic patterns observed in F_0 contours, and is not intended to represent a phonological classification.

Peak Accents

Using a empirical trial and error study, an equation was devised that can accurately synthesize any pitch accent in our data.

Peak accents are modelled by describing the rise and fall parts of the accent separately. Equation **??** (termed the *monomial* equation) is used to model both rise and fall. The form of the equation shown here describes the F_0 of the fall part of a peak accent: when this curve is reflected in the y axis it describes the rise part. In our data there was a large variation in the amplitudes and durations of the rises and falls (both absolutely and relative to one another) and there was also a wide variation in the gradients of these rises and falls. To model this variation, two scaling factors are applied to the equation, allowing the equation to be stretched in the x and y dimensions. Equation **??** describes the formula with the scaling variables A and D, and figure 1 shows a plot of this function.

Although the rise and fall parts are given by the same equation, these parts still have to be modelled separately as there was no observable simple relation between the amplitude, duration or gradient of the rise and fall parts of an accent.

$$\begin{array}{ll} y = 1 - 2.x^2 & 0 < x < 0.5 \\ y = 2.(1-x)^2 & 0.5 < x < 1.0 \end{array} \tag{1}$$

$$\begin{array}{ll} f_0 = A - 2.A.(t/D)^2 & 0 < t < D/2 \\ f_0 = 2.A.(1-t/D)^2 & D/2 < t < D \end{array} \tag{2}$$

where A is element amplitude and D is element duration.

These rises and falls are termed *elements*, thus peak accents are modelled as a *rise element* followed by a *fall element*. The amplitude and duration scaling factors are called the element *parameters*. Often in a peak accent no rise is present, so the pitch accent is modelled using only a fall element. The phenomenon known as "flat hat" accents in the Dutch school (**?**) could either be interpreted as two accents, one with a single rise element, the other with a single fall element; or as a single accent, where the rise and fall elements were separated by a straight section of F_0 contour.

Trough Accents

Trough accents are also modelled by the monomial function. The basic pattern for these accents is the reverse of the peak accent, i.e. a fall element followed by a rise element. Often a trough accent only needs a single rise or fall element.

Boundary Rises

Sharp rises are commonly found at both phrase beginnings and ends, and these are modelled using rise elements. Phrase initial rises are often termed *declination resets* (**?**) as they serve to make sure the starting F_0 of the phrase is higher than the F_0 values of the previous phrase. Phrase final rises usually either give the impression of a continuation rise, indicating that more information is to follow, or of being part of a complex nucleus/tail relationship, where they often give the precept of a question.

Between Rise and Fall Elements

Not every part of an F_0 contour is comprised of pitch accents and boundary rises. Often there are parts where nothing of intonational interest is occurring. In these areas a straight line is used to model the F_0 contours. This element is termed the *connection element*. The connection element also has variable duration and amplitude parameters, thus the straight line need not be horizontal or level.

2.2 Well-Formedness Conditions

Thus in our model, F_0 contours are modelled with a linear sequence of rise, fall and connection elements, each of which can have any amplitude or duration. A set of well-formedness conditions are used to constrain this system:

- Pitch accents are modelled using at most a single rise and a single fall element.
- Boundary rises are modelled using a single rise element.
- Connection elements are used to model everything else in the F_0 contour, thus rise and fall elements may only be used for modelling pitch accents and boundary rises.
- Only one connection element is allowed between rise and fall elements.

An example RFC description as hand labelled from an F_0 contour in the test data is given in table 1, and the synthesized F_0 contour from this description is shown later in figure 3.

2.3 Deriving Phonological Information from RFC Descriptions

RFC descriptions contain much of the information needed to distinguish pitch accent types. For example the "high-rise" and "low-rise" pitch accents (**?**) are distinguished by the amplitude parameter of the RFC elements. The amplitude of the rise is typically higher for high-rise than for low-rise, and the starting F_0 value for the rise (which can be calculated from the amplitudes of the preceding elements) is also higher.

Some (e.g Crystal (**?**)) have proposed a phonological class of "level" pitch accents. Level pitch accents with no observable F_0 patterns are not given explicit RFC descriptions, as the RFC model only attempts to describe surface patterns. It would be inappropriate for the model to describe these patterns in much the same way as it is inappropriate for a narrow phonetic transcription to mark segments which are present underlyingly, but are not physically produced. This type of accent was very rare in our data (see section **??**) with only one occurrence.

When the RFC description is aligned with the phonetic segmentation of the utterance, differences in alignment can be used to distinguish other classes of pitch accent. For example, a Pierrehumbert H*+L accent can have the same RFC fall description as for a L* accent. The difference is that the fall in the H*+L accent occurs much later with respect to the vowel of the accented syllable.

3 Automatic Labelling of F_0 contours using the RFC model

The system described below automatically labels F_0 contours in terms of the rise, fall and connection elements. It works on unconstrained input: any F_0 contours can be labelled by this system without prior knowledge of the content of the utterance. Although there are trainable parameters in the system which give better performance when adapted for particular speakers, this system can label any type and any length of F_0 contour in our data.

3.1 F_0 Extraction

F_0 contours are usually extracted from speech waveforms using *pitch detection algorithms* (PDAs). The PDA used here was the super resolution pitch detection (SRPD) algorithm originally developed by Medan et al. (**?**) and implemented by Bagshaw et al. (**?**).

In normal usage the PDA produces F_0 contours which are influenced by segmental effects such that the immediate segmental environment affects the local shape of the F_0 contour. *Unvoiced segments* are the most noticeable effect: during such segments there is no fundamental frequency at all. *Obstruent perturbations* cause sudden spikes

and glitches, often at the boundary between a voiced and unvoiced part of the F_0 contour. *Intrinsic vowel pitch* is another source of segmental influence in the F_0 contour, whereby high vowels cause slightly higher F_0 values than low vowels (**?**). It is desirable to normalise for these affects wherever possible because they are purely a result of the segmental environment and play no role in determining the underlying intonational tune of the utterance. To put it another way, two utterances with the same intonational tune can have apparently different F_0 contours solely due to these utterances having different segmental content.

For our purposes, the obstruent perturbation effects were the most problematic as the labelling modules described below might confuse the perturbations with rise and fall elements. They were dealt with by adding a post-processing module to the PDA that converts the normal output into F_0 contours which are as free from segmental influence as possible, while maintaining exactly the same underlying intonational content. Intrinsic vowel pitch was not dealt with as this caused no problems for the labelling modules. The amplitudes of the RFC elements are therefore affected by intrinsic vowel pitch, which makes direct comparison of element amplitudes from different vowels inadvisable.

In all the experiments described here, F_0 contours were specified at regular 5ms intervals. The first stage in the post-processing is to perform a 15 point median smoothing on the 5ms contour. This removes most of the obstruent perturbations and also the small scale effects of pitch perturbation or jitter that arise from the normal behaviour of the glottis. Next, the unvoiced regions are filled using straight line interpolation. Finally, a further 7 point smoothing is used to remove the occasional sharp edges at the boundary between the existing and interpolated parts of the F_0 contour. This processing is shown in figure 2.

The smoothing is effective due to the difference in the durations of segmental perturbations and the rise and fall elements. In data set C (see section **??**), the average duration of the perturbations was 18 ms, whereas the average duration of rise and fall elements was 158ms and 200ms respectively. However, there was a small but significant number of longer segmental perturbations which were not removed by the smoothing. The deletion process described in the next section was used to deal with these.

3.2 Broad Class Labelling

The *broad class labelling module* takes the post-processed F_0 contours and locates rise, fall, and connection elements. This module distinguishes rises from falls by the principle that all rises must have positive gradient and all falls must have a negative gradient. The rise and fall elements are distinguished from connection elements on the basis that these elements have steeper gradients than those of connection elements.

The F_0 contour is re-sampled at 50ms intervals as it is at those sorts of distances and above that pitch accent rises and falls are realised. Next, the F_0 of each frame is compared with the F_0 of the previous frame, and if this exceeds a threshold, the frame is labelled a rise, and if it is below another (negative) threshold, the frame is labelled as a fall. All frames between these two thresholds are left unlabelled. These thresholds (termed the *rise gradient threshold* and *fall gradient threshold* respectively) are trained by the method described in section **??**. At this stage, every frame is labelled either rise or fall, or is left unlabelled. Frames which have the same labels as their neighbours are grouped together, dividing the F_0 contour into approximately marked rise and fall elements.

The labelling produced by this method identifies most of the rise and fall elements in the F_0 contours. As mentioned above, the F_0 post-processor does not remove all obstruent perturbations, and occasionally these are mislabelled as rise or fall elements. These spurious labels have characteristically short durations. Using this fact, a *deletion module* was developed, whereby elements below a certain minimum duration are deleted. This *deletion threshold* is set by the training method described in section **??**. This module greatly reduces the number of spurious elements.

The labelling produced by this system has divided the F_0 contour into rise, fall and unlabeled elements, and by implication, the pitch accents and boundary rises have been located. However, the boundaries between these elements are only accurate to the size of the frame, i.e. 50ms. It was desirable to go further and find the exact boundaries of the elements as these are crucial in distinguishing different types of pitch accent. Precise boundary adjustment is performed by the *optimal matching module*.

3.3 Optimal Matching

Using the labels produced by the above procedure, the original 5ms sampled F_0 contour is analysed to determine the precise boundaries of the rise and fall elements.

Around each approximate boundary, a *search region* is defined. This extends an absolute amount, typically 0.15 seconds outside the approximately marked boundary, and a percentage distance, typically 20%, inside the approximately marked boundary. Thus for a fall element, a region is defined starting 0.15 seconds before the start of the fall and extending 20% into the fall, and another region is defined starting 20% from the end of the fall and extending 0.15 seconds after the end of the fall. To determine the precise boundaries, every possible fall shape that can be defined as starting and ending in these areas is synthesized and compared to the F_0 contour. The shape showing the lowest euclidean distance between itself and the F_0 contour is chosen as the correct shape, and its start and end position determines the precise boundaries of the fall element. Likewise for rise elements.

In principle, the size and position of the search regions could have been trained, but it was found that in all cases a search region corresponding to 0.15 seconds outside the approximately marked element and 20% inside the element is large enough to insure that the precise element boundaries will be found. Our experience in developing the system was that so long as some variation from the approximately labelled boundaries was allowed, the system always chose the same start and end positions. This gives reason for confidence that the start and end positions are not arbitrarily marked, as the same positions are consistently chosen by the system independently of where the search regions are defined.

With the optimal matching process complete, all remaining unlabelled sections are labelled as connection elements. Thus the entire F_0 contour is now labelled in terms of the RFC system.

4 Data

The speech recognition community has benefited greatly by the use of standard databases. These databases are publicly available and have been hand labelled to a common standard. For prosodic work, there are no such standard databases as yet, although recently there have been moves towards promoting prosodic labelling standards (**?**). Until such times as suitable databases are widely available, and have been labelled to a common standard, it is unfortunately the case that tests will have to carried out on individual data which makes direct comparison between systems difficult.

Six sets of data were used here from American, English and Irish male and female native speakers of English were used to test the model. The data was hand labelled by the author using the labelling criteria described in section **??**.

Data set A from a male speaker (Northern Ireland accent) comprised of 64 carefully spoken sentences that cover the major tune types described in O'Connor and Arnold (**?**). Data set B was from a male speaker (southern English RP accent) and comprised of 45 utterances from email and Unix news articles. Data sets C to F were from the ATR-CMU conference registration database, which simulated conversations between receptionists and guests at conferences. This data was spoken by 2 male (C and E) and two female speakers who were born and raised in Pittsburgh in the United States of America[2].

This data totalled 231 utterances, within which there were 1654 rise and fall elements and 533 intonational phrases, as hand labelled. The utterances varied in length from a single word to 40 words in length. There was a wide variation of intonational tune types due to the conversational nature of much of the speech.

5 Synthesis Assessment and Results

An analysis/resynthesis test was used to measure synthesis accuracy. For each utterance in the database, its F_0 contour was analysed (automatically and by the hand labelling method described in section **??**) and an RFC description was produced. From this, an F_0 contour was synthesized and compared to the original.

Subjective tests are sometimes used to assess the intonation component of speech synthesis systems. In these tests, listeners are played re-synthesized pairs of utterances, one with the original F_0 contour and one with the synthesized F_0 contour, and if listeners cannot distinguish the two, then the synthesized F_0 contours are deemed to be perceptually equivalent to the originals.

It is not clear that this sort of subjective evaluation is the most suitable way of assessing the synthesis accuracy of a model such as ours. Here a different approach was taken and *objective* tests were used, whereby an algorithm assesses the differences between the original and synthesized F_0 contours. This was because:

[2] Data sets A and B are described fully in Taylor (**?**), and the ATR-CMU data is described in Wood (**?**).

- It is unclear how meaningful yes/no decisions are on the equivalence of synthesized F_0 contours. F_0 contours are specified in the continuous frequency domain, and as with any continuous variable, it is somewhat nonsensical to compare for direct equivalence. When comparing any real numbers, it makes sense to state how close they are, rather than give a yes/no decision of equivalence.

- As noted by Hirst (**?**), "a model which seemed perfectly adequate for LPC diphone synthesis may appear less satisfactory when used with very high quality speech synthesis such as that provided by PSOLA". Thus different waveform synthesis techniques can influence the judgment of listeners. We would not want to run the risk of claiming that our model can produce acceptable F_0 contours to have a later more sophisticated waveform synthesis technique to prove otherwise. The assessment of F_0 synthesis should be independent from waveform synthesis techniques.

- Objective tests are very practical in that a large amount of data can be easily tested. Other researchers can easily compare the synthesis accuracy of their model with that of ours.

Thus the synthesis assessment method that was used was one of direct comparison between original and synthesized F_0 contours. Each point in the original F_0 contour is compared with the point at the equivalent time frame in the synthesized version, and the euclidean distance is measured. A score is produced by summing the values for an utterance and dividing by the total number of frames for that utterance. This score gives a measure of the average differences between two F_0 contours, or to put it another way, at any given point the expected difference between the F_0 contours is given by this score. Table 2 gives the scores for the 6 data sets.

These results clearly show that the synthesized F_0 contours are very similar to the originals. Many of the bad scores are due to PDA problems such as pitch doubling, which causes a large comparison error. Thus the scores would be better if a more accurate PDA was used.

To put the figures in table 2 in perspective, it is worth making some points on the accuracy of PDAs and the difference limens of F_0 changes in humans. Hess (**?**) gives a review of both these issues and describes several experiments on how sensitive humans are to F_0 changes in speech. Figures of between 0.3-0.5% for pure synthetic vowels and 4-5% for natural speech are given. The 4% figure measured at 195Hz, giving an absolute value of 7.8Hz, and the 5% figure was measured at 150Hz, giving 7.5Hz. Thus our synthesis results are close to the difference limen for natural speech.

Rabiner et al. (**?**) discuss the assessment and accuracy of PDAs, but perhaps the most useful and up to date study is that of Bagshaw et al. (**?**). Their experiment examined the output of 6 PDAs, including the super resolution pitch detection algorithm used here, and compared them to F_0 contours directly measured by a laryngograph. The average error of these algorithms varied from 1.78 Hz to 3.25 Hz for male speech, and from 4.14 Hz to 6.39Hz for female speech, with the SRPD algorithm giving the best results[3]. The results for the 4 males speakers and female speaker E are close to a few Hertz of the accuracy of the SRPD algorithm, and the results for the female speaker F are within the margin of error.

6 Automatic Labeller Assessment and Results

6.1 Hand Labelling and Analysis Assessment

The main problem with analysis testing is not so much in the exact method of testing but rather the difficulty in determining the correct labelling for an F_0 contour. The only practical way of assessing the automatic labeller's performance is to compare the labels it produces with those of a human labeller. The problem with such a method is that it relies on the ability of the human labeller. Human labelling is always prone to some inconsistency and arbitrariness no matter how expert the labeller. However, it is clear that some human made decisions are consistently more reliable than others, for instance, it is often much easier to determine the *location* of a pitch accent than trying to decide what the *type* of the pitch accent is.

The database of F_0 contours was hand labelled according to the following criteria. First every syllable was marked as to whether or not it was associated with a pitch accent. This was done by both examining the F_0 contour and listening to the speech. Next, all phrase beginnings and ends were marked as to whether or not they had a

[3] Bagshaw et al. go on to explain improvements to the SRPD which give better results than the original described in Medan et al. (**?**). This improved version was not available for the experiments described here.

boundary rise. This was usually a straightforward process. The only real difficulty occurred with a small number of cases where the size of the pitch accent or boundary rise was very small. Although the different types of pitch accent would naturally be expressed by their different rise and fall labellings, no explicit labelling of pitch accent class was performed.

Once each accent or boundary rise had been identified, it was necessary to decide what RFC descriptions they should be given. This was done by chosing appropriate elements and hand marking a series of start and end positions for each element (the end position of one element is the start position of the next). Each start and end position has an associated start and end F_0 value, which is the F_0 value of the contour at the point where the start or end is located. The element duration is found by subtracting the start from the end position, and the element amplitude is found by subtracting the start F_0 value from the end F_0 value.

Each boundary rise was assigned a rise element. For a phrase initial rise, the start position of the element was marked as the start of the F_0 contour for that phrase, and the start amplitude was marked as being the F_0 value at that point. The end position of the rise was determined by a trial and error procedure of keeping the start position fixed, and systematically trying rises of different duration. For each provisional end position, a partial F_0 contour corresponding to the element was synthesized and compared to the original contour. A computer graphics program was used to superimpose the synthesized element contour onto the original contour and judge the fit. The end position was chosen as being the position where the synthesized element contour fitted best. It usually required two or three attempts to find a good fit. Phrase final boundary rises were marked in much the same way, expect that the end position was kept constant and the start position was varied.

Some pitch accents had a rise and a fall and some had only a fall element. For those with only a fall element, a similar procedure to that outlined above was used, but both the start and end positions were varied. Pitch accents with both a rise and a fall were slightly more complicated in that the end of the rise had to exactly match the start of the fall. However, the principle was the same and by adjusting start and end positions and superimposing element contours on the original, a good fit was found.

Once all the boundary rises and pitch accent rises and falls had been marked, connection elements were used to fill in any gaps. The durations and amplitudes of the elements were then calculated from the start and end positions and F_0 values.

The first part of the labelling process, which marked pitch accents and boundary rises, involves linguistic, impressionistic judgement as to whether or not a syllable is accented or not. The second part of the process was particular to this model and was really just a case of adjusting the boundaries until a good fit was achieved.

6.2 Analysis Assessment

The analysis assessment method compares the rise and fall elements from the automatic labeller with the rise and fall elements from the hand labeller and calculates the percentage of rises and falls correctly identified. Thus the system simply counts the number of insertion and deletion errors, and divides this by the total number of tokens, resulting in a percentage recognition score.

6.3 Training the Automatic Labeller

A number of adjustable thresholds operate in the analysis system. The most important are the rise and fall *gradient* thresholds in the broad class labeller. These are used to determine whether a 50ms frame is to be labelled as a rise, fall or left unlabelled. In addition there are two *deletion* thresholds which specify the minimum allowable size for an element. Any element identified by the broad class labeller which is below this duration is deleted. Although the rise and fall thresholds operate independently of each other, the gradient thresholds interact with the deletion thresholds.

The training procedure operates by systematically adjusting the thresholds until the optimal set is obtained. Taking the case of the fall thresholds, a 2 dimensional table is built with one dimension representing the gradient threshold and the other representing the deletion threshold. 10 values are used in each dimension. The gradient threshold is varied on a logarithmic scale from 20Hz/second to 500 Hz/second, and the deletion threshold is varied linearly from 0.025 seconds to 0.475 seconds. Using each set of thresholds, the system is run over a set of data, the transcriptions produced are compared with the hand labelled versions, and the recognition score for that set of thresholds is recorded. After all possibilities have been tried, the set of thresholds giving the best recognition score is chosen as being the optimal set.

This technique can be used to train on any amount of data, but 10 utterances are sufficient to ensure safe training as recognition scores do not significantly improve with more training data. In all data sets, a deletion threshold of 0.075 or 0.125 seconds was chosen as best[4]. The gradient thresholds varied more between speakers, from 70Hz/second to 120Hz/second. These thresholds do not give an indication of how steep the rise and fall elements actually are; rather, they are the optimal thresholds that distinguish legitimate rises and falls from connection elements and obstruent perturbations.

6.4 Analysis Results

Table 3 shows the results of the automatic labeller for the six speakers. The overall accuracy rate in the high 70s leaves room for improvement, but considering the fact that the system is working in an unconstrained fashion with no top-down processing, these results are promising. The errors were examined to discover their source, and the results from this study showed that the overall picture is much better than the above results might lead one to believe.

Four sources of error were identified. These were:

F_0 errors A small number of the errors were due to the PDA making a mistake such as pitch doubling. This causes a sudden jump in the F_0 contour which can be mistakenly interpreted as a rise or fall element. As Hess (**?**) (page 66) notes, these gross F_0 errors can often be compensated for when hand labelling, as the eye is able ignore the erroneous values and detect the underlying pattern.

Obstruent Perturbations The F_0 post-processor and deletion module did not account for all obstruent perturbations and a number of errors arose from the labeller mistaking glitches or spikes for small rise or fall elements. The insertion errors, where a perturbation is mistaken for a rise or fall, can easily be eliminated by increasing the gradient thresholds, but this has the effect of causing deletion errors as small genuine elements are not detected. So far the system has worked using F_0 as input alone: it might be necessary in future implementations to use additional information. For instance, a phonetic segmentation would give information on where obstruent perturbations occur, and heavier smoothing could be used in these areas. If better post-processing was used fewer perturbations would be classed as rises or falls, and the gradient thresholds could be lowered so as to accept more of the genuine rises and falls.

Data set A had a considerably higher recognition rate than the other sets and this was mainly due to the speech in set A being mostly voiced and being freer from obstruent perturbations than the other sets.

Algorithmic Problems Some errors arose from straightforward mistakes by the labelling algorithm arising from phenomena not foreseen in the initial design. A common mistake was for downstepping pitch accents on successive syllables to be labelled as a single large fall. Here the system would detect a long falling section of F_0 and assume this to be a single accent. A simple modification to the optimal matching module allowing more than one fall shape to be fitted to elements that have been labelled as falls by the broad class labeller should help solve this problem.

Labelling Problems A small number of errors arose from situations where is was not clear that the hand labelling was correct. Nearly always these cases involved small phrase initial boundary rises or small pre-nuclear pitch accents. The arbitrary nature of the hand labelling in these cases was not such much of an inherent problem in the RFC model, as in any system it is often difficult to decide whether certain stressed syllables should be marked as having pitch accents or not.

The majority of errors arose from problems with small pitch accents or phrase initial boundary rises. However, it is a general feature of the intonational system of English that perceptually important accents have on average larger F_0 excursions than normal accents, all other things being equal[5], and therefore it is the case that the system recognised important accents best. To confirm this, an additional test was performed which measured the accuracy of the labeller on rises and falls which were part of nuclear accents (defined as the last pitch accent in the

[4] 0.075s corresponds a single 50 ms frame being deleted (50ms ¡ 75 ms) and 0.125s corresponds to two 50 ms frames being deleted (2 x 50ms ¡ 0.125).

[5] Many factors such as pitch range and declination need to be taken into account, but it does seem generally true that accents which are important, e.g. nuclear accents are consistently larger than those which are not.

phrase). The results given in table 4 clearly demonstrate that the automatic labeller is very successful at finding and classifying nuclear accents.

As mentioned previously, some types of pitch accent are distinguished more by their alignment rather than by their F_0 pattern. Thus it is important for the recogniser to accurately locate element boundaries as well as finding the elements themselves. In data sets C and D, the average difference between hand and automatically labelled boundaries was 40ms. This is substantially less than the average durations of the elements (158ms for rise durations and 200ms for fall durations in data set C).

A single labeller was used throughout. Additional labellers would have been useful in that the consistency of the labelling criteria could have been measured. The problem with using only a single labeller is that we have to compare the automatic systems labels against human labels without knowing how accurate the human labelling is. One would expect some inconsistency between labellers and this would be useful in putting into perspective the inconsistencies given here between the automatic and hand labels.

Graphs of F_0 synthesis and Analysis

Figures 3 and 4 show F_0 contours from two utterances in set A. In each figure, graph (a) is the original F_0 contour as produced by the post-processing module. Graphs (b) and (c) show the original contour, the labels from an RFC description and a synthesized contour from this RFC description. Graph (b) shows a hand labelled RFC description, and graph (c) shows an automatically labelled RFC description.

These figures demonstrate a number of points. By superimposing the synthesized F_0 contours on the originals it is possible to gain a subjective impression of the model's synthesis accuracy. While slight differences between the F_0 contours can be detected, it is clear that the synthesized versions are close to the originals, supporting the objective evidence in table 2.

The (c) graphs show some typical errors from the automatic labeller. The second pitch accent of the first phrase in figure 3 is mislabelled as the automatic system failed to recognise the small fall element of this accent. In figure 4, an insertion error occurs as an obstruent perturbation was mislabelled as a fall element. These graphs support the evidence in table 3 which shows that the automatic labeller labels large pitch accents more accurately than smaller ones.

6.5 Other Automatic Analysis Systems

A number of automatic intonation analysis systems have recently been proposed, but it is difficult to make comparisons owing to the different nature of the tasks attempted by these systems and the lack of standard databases. Jensen et al. (**?**) describe a system that recognises O'Connor and Arnold tune types from F_0 contours. They report a 71% accuracy in classifying nuclear accents, which is worse than the results reported in table 4, but as they are attempting a full phonological description, their basic task is much harder. Geoffrois (**?**) describes a recognition scheme for Japanese speech using the Fujisaki model where he correctly recognises 91% of accent commands. Although the step and impulse functions of the Fujisaki model are roughly equivalent in terms of level to the rises and falls of the RFC model, direct comparison is again difficult due to the fundamentally different nature of English and Japanese intonation.

7 Discussion

7.1 Flexibility and Variance in the RFC model

It is simple to devise a model which can synthesize F_0 contours with a high accuracy: any sort of unconstrained target system can do this. What is much more difficult is to design a model with high synthesis accuracy that generates these F_0 contours from a *linguistically useful description*. It should be clear from the well-formedness conditions given in section **??** that the RFC descriptions *are* linguistically. These conditions state that rise and fall elements may only be used to model phonologically significant events, and phonologically significant events can only be described using rises and falls. Therefore all phonological events are readily detectable in an RFC description.

The only debatable point is whether the flexibility in the RFC description is justified. The free parameters in the model are the durations and amplitudes of the rise and fall elements. A typical pitch accent with a rise and fall

element thus needs four parameters. In our data the amount of variance in pitch accents is considerable. A previous study of data set A showed that rise element amplitudes vary from 10Hz to 96Hz and fall element amplitudes vary from 11Hz to 140Hz (**?**). This is not controversial as most systems have some way to gradiently scale accent height, e.g. Liberman and Pierrehumbert (**?**) (we are not of course claiming that our scaling factors are directly equivalent to theirs). However the RFC model also allows variability in the slope gradients of the rise and fall elements.

Table 5 shows statistics derived from the full set of hand RFC labels for speaker C. The large variance of the amplitudes, durations and gradients should make it clear that these parameters do indeed need to be flexible to account for the data. Any model failing to acknowledge this variance will have considerably worse synthesis accuracy than the RFC model.

Informal investigation into the causes of the variance shows relationships between element duration, gradience and the sonority of syllables, ie. syllables with short voiced regions have shorter durations and steeper gradients than those with long voiced regions. Nuclear accents occur earlier with respect to syllable boundaries than pre-nuclear accents, which is in line with the more thorough study of Silverman and Pierrehumbert (**?**). The wide amount of variance is therefore probably due to differences in the sonority of the syllables, the position of accents in the phrase and the phonological class of the pitch accent. Thus the flexibility in the model is required if it is to synthesize F_0 contours with high accuracy. The strength of the RFC model lies in its ability to do this while still making the RFC descriptions easily amenable to phonological analysis.

7.2 Comparison with other Models

Fujisaki

Fujisaki's model of Japanese intonation uses two critically damped second order filters to generate F_0 contours. The *phrase component* uses impulses as input and models long term phenomena such as downdrift and resets. The *accent component* uses step functions as input and models pitch accents. Separate time constants control the rate of rise and fall of each component, and in the classical definition of the model, these constants are invariant for a speaker.

The Fujisaki accent component only requires two numbers, amplitude and duration, to model pitch accents compared with the four mentioned above for the RFC model. Thus the Fujisaki model is more constrained than the RFC model. However, the accent component cannot synthesize the pitch accents in our data with the accuracy of the RFC model. Amendments can be made, such as allowing a negative step in the later half to account for downstepping accents, but this would add two extra parameters (the amplitude and the position of the polarity change). Furthermore, the Fujisaki model predicts that pitch accent gradience is invariant for a speaker, and as table 5 shows, the gradient of rises and falls vary considerably for a speaker. Therefore the Fujisaki model would have to allow the time constant to vary also, adding even more parameters.

The phrase component is problematic as it cannot model long continuously rising sections of F_0 contour, such as commonly observed after L* accents (as seen in figure 2 of Beckman and Pierrehumbert (**?**)), or in the pre-nuclear position in the surprise redundancy contour (as seen figure 11 in Ladd (**?**)). Figure 4 also demonstrates this effect. We found it difficult to postulate any amendment to the model which would account for such phenomena.

It must be noted that the intonational system of Japanese is quite different from that of English, and in particular it has a much smaller inventory of pitch accents (Beckman and Pierrehumbert (**?**), like Fujisaki, use only one). Thus it is to expected that a model developed with the Japanese language in mind would be restricted in pitch accent shape.

Dutch

In the Dutch school (**?**), (**?**), (**?**), F_0 contours are analysed in two stages. The *close-copy stylization* describes the F_0 contour in terms of a number of straight lines. These close-copy stylizations are claimed to be perceptually indistinguishable from the original F_0 contours, as demonstrated by re-synthesis tests (**?**). This is essentially a data reduction exercise. The second stage is termed *standardization* whereby the close-copy contour is further coded into a series of standard rise and fall patterns.

Although the close-copy process may produce F_0 contours with a synthesis accuracy near that of the RFC model, the standardization process makes the Dutch model more like the prototypical systems mentioned in the

introduction. Therefore the aim is not to accurately synthesize any F_0 contour, but to discover a set of F_0 patterns which are within the "linguistic tolerance" of a listener (**?**).

Hirst

The RFC model is similar in some ways to the model proposed by Hirst (**?**) where an attempt is made to derive the phonological description of an utterance from its F_0 contour. The first stage of his system uses a spline fitting algorithm that reduces an F_0 to a number of target points. Later these target points are classified in terms of a surface phonological description. Although we are not aware of any exact performance figures for Hirst's system, it would be likely that his spline model's synthesis accuracy should be least as accurate as the RFC model, as his target points are not as constrained as the RFC model. The RFC model lies somewhere between Hirst's spline level and the surface phonological level in that the location of the pitch accents and the boundary rises are explicitly marked in the RFC description whereas the spline description needs further processing to extract this information.

7.3 Practical Applications

The RFC model forms the basis of the intonation component in the speech synthesis system being developed at ATR (**?**). In this system, intonational tune is described by a system of intonational elements (H, L etc) and features (delayed, downstep etc), which is similar to the intonational tune phonology of Ladd (**?**). F_0 contours from the ATR-CMU database (including data sets C to F) have been labelled using the RFC model and the tune phonology. From the RFC descriptions of a speaker's utterances, it has been possible to collect statistics on the amplitudes and timing characteristics of pitch accents. These statistics are used in the intonation component of the speech synthesis system to ensure that the synthesized intonation has a good likeness to the original speaker. It is much easier to derive these statistics from an RFC description than directly from an F_0 contour, as the parts we are most interested in (e.g. pitch accents) are explicitly marked in an RFC description in a regular manner. A description of the tune phonology system and its relation to Ladd's is given in Taylor (**?**), and a description of the intonational component of the speech synthesis system is given in Taylor (**?**), (**?**), and in Black and Taylor (**?**),

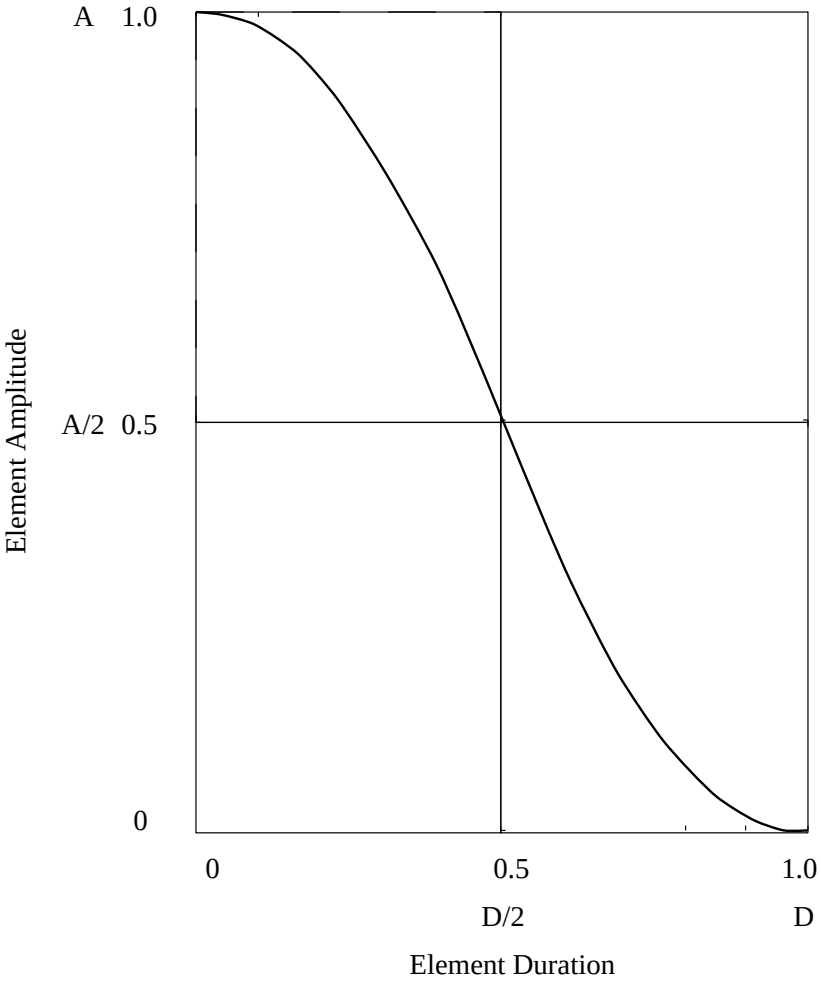
We are currently working on integrating the automatic labeller into the ATR speech recognition system. A system already exists which uses the RFC description from the automatic labeller to derive the intonational tune of the utterance (**?**). Work is underway to use this tune description to help derive the speech act (question/statement/greeting etc) of the utterance.

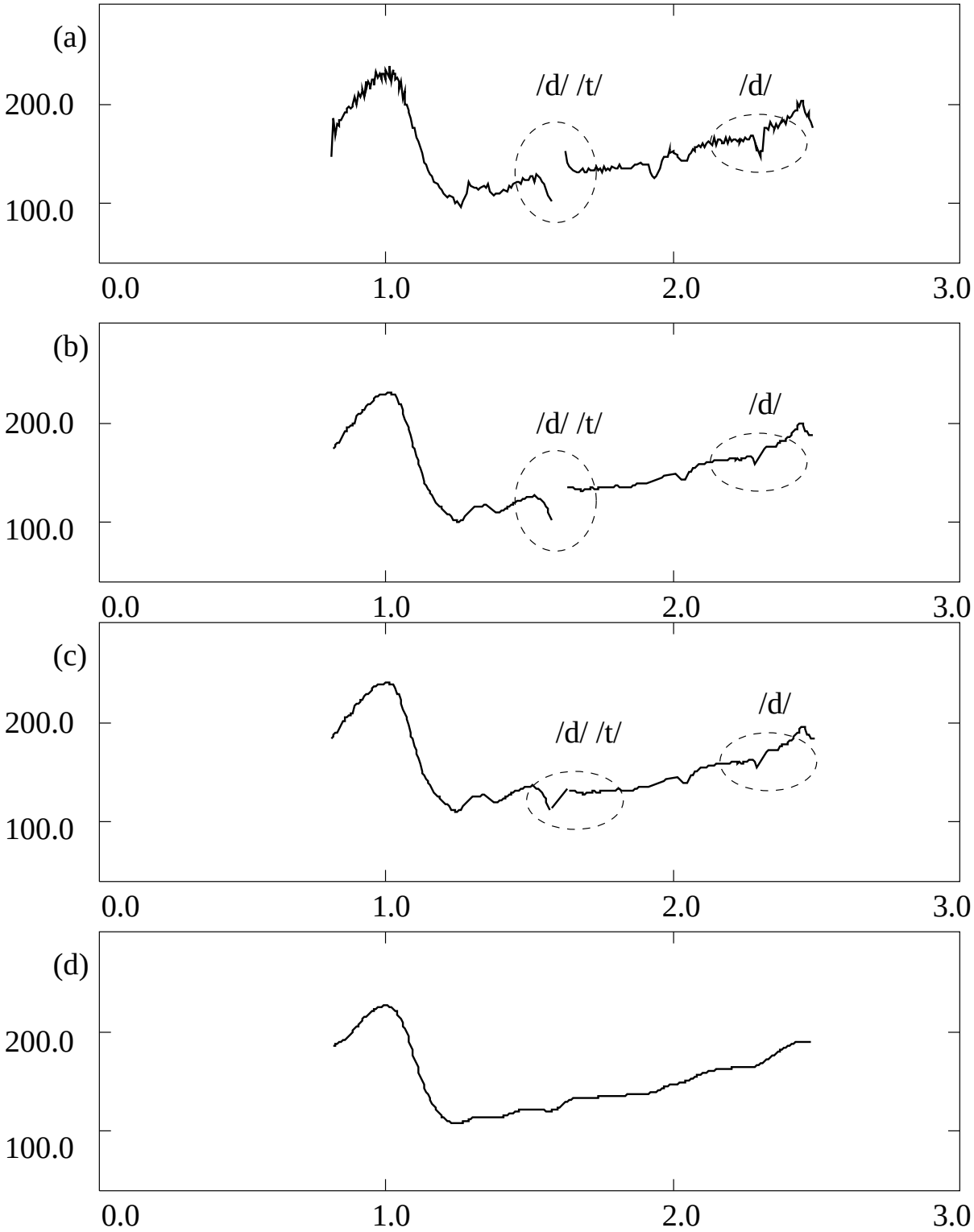
The synthesis of F_0 contours from an RFC description is computationally trivial. When running on a workstation (Sun sparcstation LX, approx. 20 Mips), the automatic labeller takes about 0.3 seconds of processing time for every 1.0 second of speech. Faster performance figures should be possible as no speed optimisation has been attempted on these programs.

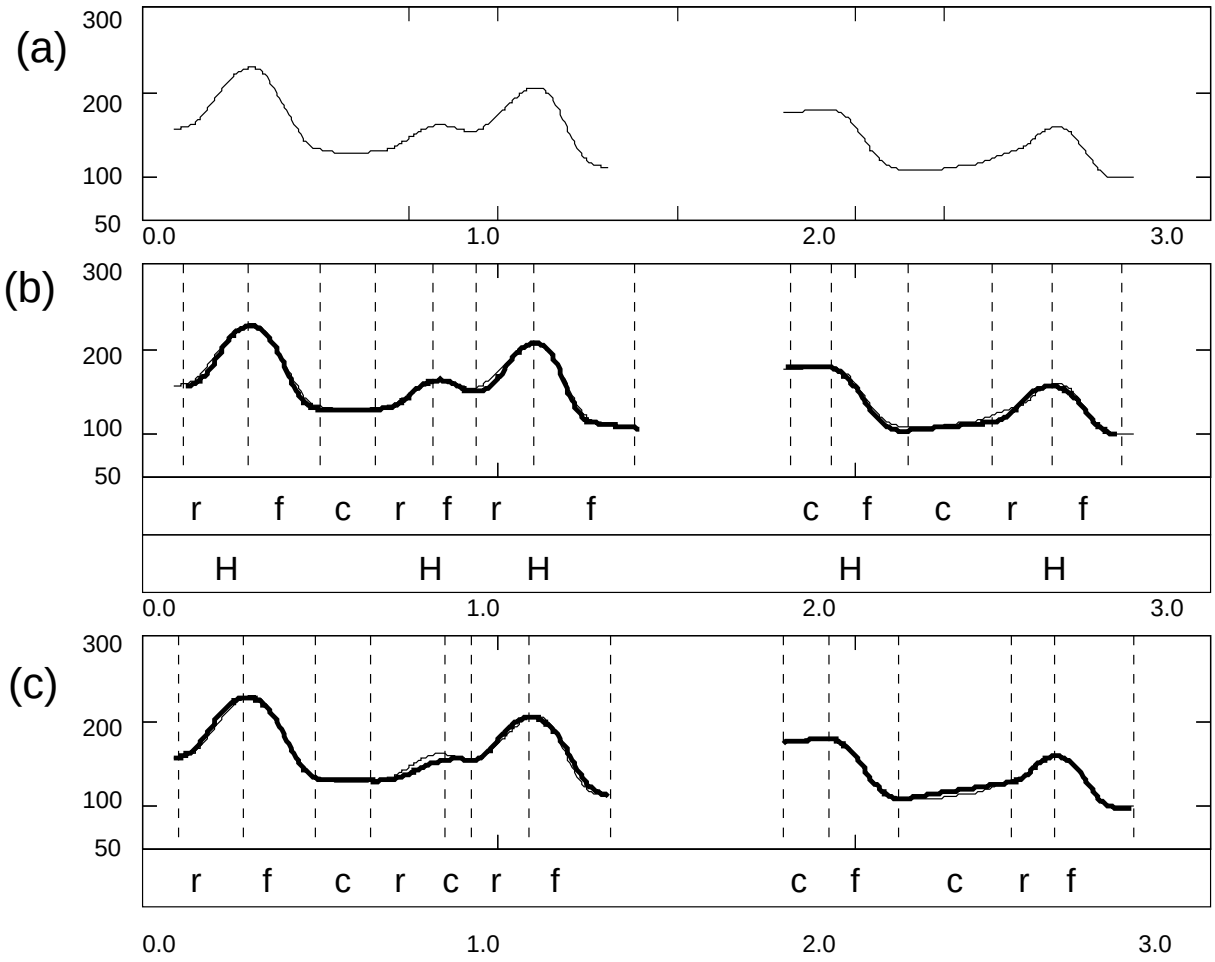
7.4 Conclusion

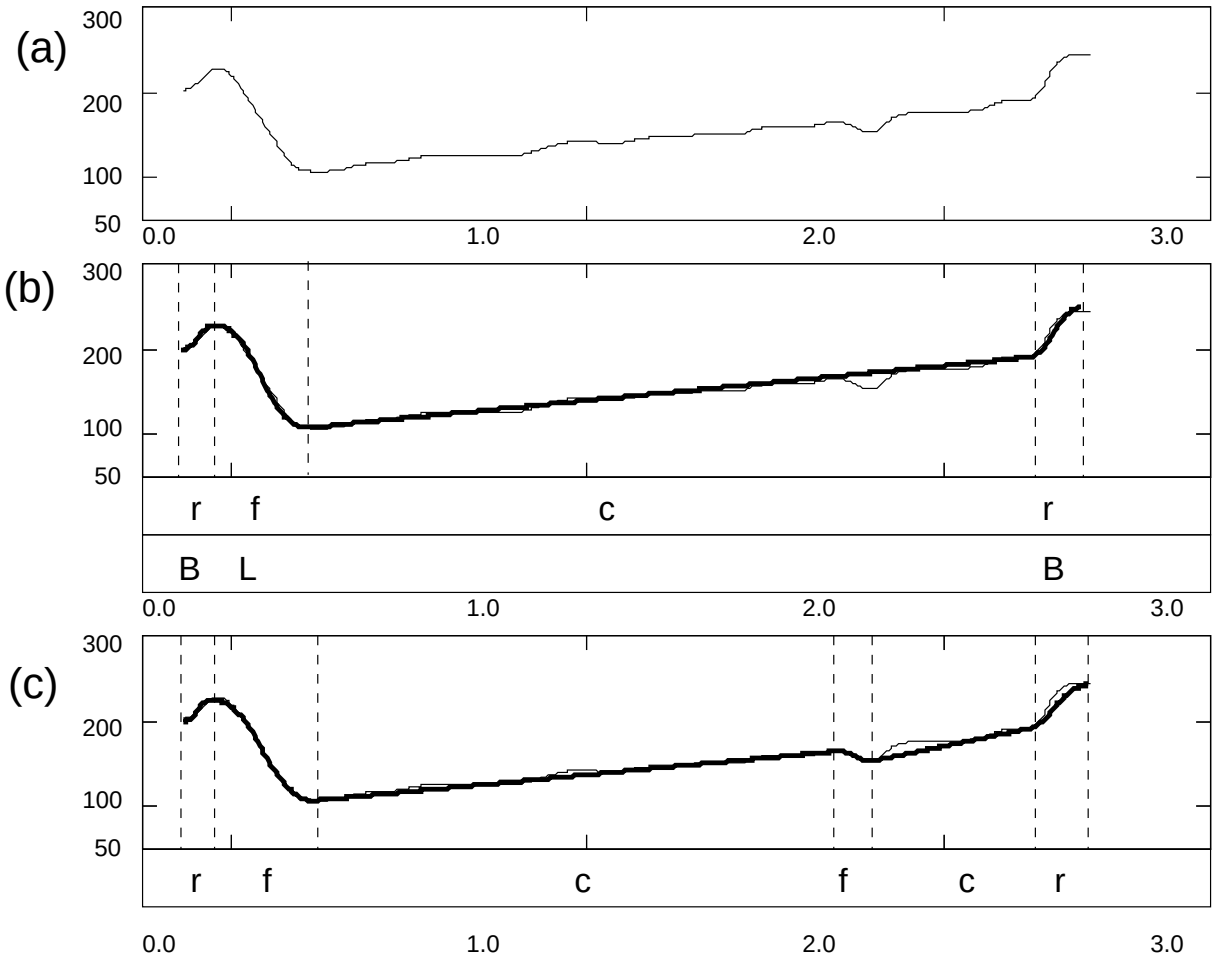
As far as extending the model to produce a higher level analysis based on an RFC description is concerned, two points can be made. Firstly, due to the fact we can reconstruct a contour very similar to the original one, no information present in the original F_0 has been lost; rather it has been converted to a form more amenable to further analysis. Secondly, in an RFC description, pitch accents and boundary rises are identified, and further analysis need only concentrate on classifying the rise and fall descriptions into separate phonological classes of pitch accent and boundary tune.

In conclusion, describing intonation with rise, fall and connection elements is very useful in that this description is relevant to both the phonological and acoustic descriptions of intonation. Pitch accent and other intonational tune information can be derived from an RFC description, and F_0 contours can be accurately synthesized from the RFC description.


A 1.0
A/2 0.5
0
Element Amplitude
0
0.5
1.0
D/2
D
Element Duration


(a)
200.0
100.0
/d/ /t/
/d/
0.0
1.0
2.0
3.0
(b)
200.0
100.0
/d/ /t/
/d/
0.0
1.0
2.0
3.0
(c)
200.0
100.0
/d/ /t/
/d/
0.0
1.0
2.0
3.0
(d)
200.0
100.0
0.0
1.0
2.0
3.0


(a)
300
200
100
50
0.0
1.0
2.0
3.0
(b)
300
200
100
50
r f c r f r f
c f c r f
H H H H H
0.0
1.0
2.0
3.0
(c)
300
200
100
50
r f c r c r f
c f c r f
0.0
1.0
2.0
3.0


(a)
300
200
100
50
0.0
1.0
2.0
3.0
(b)
300
200
100
50
r
f
c
r
B
L
B
0.0
1.0
2.0
3.0
(c)
300
200
100
50
r
f
c
f
c
r
0.0
1.0
2.0
3.0

Type	Duration (seconds)	Amplitude (Hz)
rise	0.187	70
fall	0.187	-97
conn	0.175	0
rise	0.165	34
fall	0.100	-14
rise	0.171	57
fall	0.159	-93
conn	0.135	-7
silence	0.405	73
conn	0.105	0
fall	0.225	-76
conn	0.240	10
rise	0.175	43
fall	0.191	-57

Data set	Number of utterances	From hand labels	From automatic labels
A	64	4.9 Hz	4.7 Hz
B	45	7.3 Hz	5.4 Hz
C	55	3.6 Hz	4.2 Hz
D	19	4.1 Hz	4.3 Hz
E	21	3.7 Hz	3.8 Hz
F	17	4.9 Hz	3.9 Hz

Data set	Number of utterances	Numbers of elements	% of rise and falls correct
A	64	352	92
B	45	589	86
C	55	332	75
D	19	125	72
E	21	138	74
F	17	109	72

Data set	Number of utterances	Number of nuclear accents	% Correct
A	64	136	98.5
B	45	156	95.4
C	55	139	97.6
D	19	34	94.1
E	21	39	94.8
F	17	29	96.5

Parameter	Mean	Standard Deviation
rise gradient	18 st/sec	15
fall gradient	-26 st/sec	15
rise duration	157 ms	92
fall duration	165 ms	82
rise amplitude	2.57 st	1.93
fall amplitude	-4.18 st	2.49

Figure 1

The quadratic monomial function. The plot is shown in the x and y space, and also with the axes marked for duration and F_0 amplitude.

Figure 2

Contour (a) is the normal output of the PDA. Contour (b) shows the result after 15 point smoothing. Contour (c) shows the interpolation through the unvoiced regions and contour (d) shows the final output of the modified PDA.

Figure 3

Graphs of the utterance "The large window stays closed: the small one you can open". Graph (a) shows the original F_0 contour. In graphs (b) and (c) the original F_0 contour is shown by the thin line and the synthesized F_0 contour is show by the thick line. Rise, fall and connection elements are labelled "r","f" and "c". In the second label box of the graph (b), the "H" symbol indicates the presence of a high or peak pitch accent.

Figure 4

Graphs of the utterance "Must you use so many large unusual words when you argue". The same labelling conventions apply as for figure 3 with the addition that "L" indicates a low or valley accent and "B" indicates a boundary rise.

Note: I recommend that figures 3 and 4 should be printed using the full width of the page.

Table 1

Example of RFC description. These labels were derived using the criteria explained in section **??**

Table 2

Average distances in Hertz between synthesized and original F_0 contours for hand and automatic labels.

Table 3

Percentage recognition scores for the automatic labeller on the six sets of data.

Table 4

Percentage recognition scores for the automatic labeller on the nuclear accents of the six sets of data.

Table 5

Mean and standard deviations for hand labelled element parameters for speaker C. "st" stands for semi-tones, "sec" for seconds and "ms" for milli-seconds