Synthesizing Fundamental Frequency Using Models Automatically Trained from Data

Kurt Edward Dusterhoff

Thesis submitted for the degree of Doctor of Philosophy
University of Edinburgh
2000

Declaration

All research presented in this thesis is a product of my own work, unless noted otherwise.

Acknowledgements

Firstly, I would like to thank my advisors, Steve Isard and Alice Turk, who provided invaluable direction and support throughout my time at CSTR. In particular, they helped me to remember that the details are as important as the overall picture. I would especially like to thank Alan Black, who introduced me to the complexities of intonation synthesis. He directed my initial research and set an example of the highest standard. I owe many thanks to all of the staff and students at CSTR, particularly Paul Taylor, Robert Clark, Simon King, Laurence Molloy, and Janet Hitzeman, who supplied software, instruction, and valuable discussion. This research could not have taken place without funding. I would like to thank the United States government for providing the Stafford Loan program, which funded the bulk of my fees and expenses. I would also like to thank the University of Edinburgh Deparment of Linguistics, which awarded me with a Small Grant during the final stages of research. Finally, I must thank my family and friends, without whom I never could have begun, much less completed, this research. I especially thank my wife, Nicola, who showed great forebearance when this thesis looked only a remote possibility.

Abstract

This thesis presents a methodology for use in building intonation synthesis models which are automatically trained from annotated speech data. The research investigates four subtopics: intonation synthesis, automatic intonation analysis, intonation evaluation, and interactions between intonation and speech segments (phones).

The primary goal of this research is to produce stochastic models which can be used to generate fundamental frequency contours for synthetic utterances. The models produced are binary decision trees which are used to predict a parameterized description of fundamental frequency for an utterance. These models are trained using the sort of information which is typically available to a speech synthesizer during intonation generation. For example, the speech database is annotated with information about the location of word, phrase, segment, and syllable boundaries. The decision trees ask questions about such information.

One obvious problem facing the stochastic modelling approach to intonation synthesis models is obtaining data with the appropriate intonation annotation. This thesis presents a method by which such an annotation can be automatically derived for an utterance. The method uses Hidden Markov Models to label speech with intonation event boundaries given fundamental frequency, energy, and Mel frequency cepstral coefficients. Intonation events are fundamental frequency movements which relate to constituents larger than the syllable nucleus.

Even if there is an abundance of fully labelled speech data, and the intonation synthesis models appear robust, it is important to produce an evaluation of the resulting intonation contours which allows comparison with other in-

tonation synthesis methods. Such an evaluation could be used to compare versions of the same basic methodology or completely different methodologies. The question of intonation evaluation is addressed in this thesis in terms of system development. Objective methods of evaluating intonation contours are reviewed with regard to their ability to regularly provide feedback which can be used to improve the systems being evaluated.

The fourth area investigated in this thesis is the interaction between segmental (phone) and suprasegmental (intonation) levels of speech. This investigation is not undertaken separately from the other investigations. Questions about phone-intonation interaction form a part of the research in both intonation synthesis and intonation analysis.

The research in this thesis has resulted in a methodology which can be used to automatically train and evaluate stochastic models for intonation synthesis from automatically annotated speech databases.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Intonation processing techniques have reached the point where the resulting analyses and synthetic contours are useful in computing applications. Intonation in speech synthesizers can now account for variation in intonation, rather than relying on a very limited number of possible intonation contour patterns as was once the case (e.g. [IP88]). Even the synthesizers which use a restricted "intonational vocabulary" are robust in their generative capacities (e.g. [Bea94]). However, it seems that intonation synthesis has reached a performance plateau. Recent research shows that, while both rule-based and stochastic models can produce natural sounding intonation, there is wide variability of success in different conditions [SCM+]. There is also a recognition among researchers in the field that many of the current methods for intonation processing utilize only a small portion of the available data. The aim of this research is to improve the capacity of intonation generation models to produce natural contour shapes by training intonation synthesis models from data. The models should be able to produce natural sounding fundamental frequency contours which vary according to the patterns of the training data.

Intonation, for the purposes of this thesis, refers to movements of fundamental frequency over time which relate to linguistic constituents larger than the syllable nucleus. High-level linguistic information results in what is known as macro-intonation (or, more generally, intonation). Typically, such pitch movements are judged to be large movements such as those which correspond to phrase final falls and rises or audible pitch accents. Micro-intonation is loosely defined as everything that is left over. More specifically, micro-intonation takes in the small fluctuations in fundamental frequency which are caused by the physiology of speech rather than intentional communication through pitch. Within this thesis, the scope of investigation of micro-intonation is restricted to vowel intrinsic F0 and coarticulation F0 effects on intonation events. Some interactions between micro- and macro-intonation are investigated within the context of automatically training intonation synthesis models.

This thesis presents a methodology with which fundamental frequency can be synthesized using statistical models that have been automatically trained from annotated speech data. A review of intonation modelling theories and techniques examines the possibilities for the infrastructure of the research and explains the use of the Tilt intonation model. The research has addressed the basic problems of how to acquire database annotations and how to create and use F0 synthesis models once the annotations have been acquired. An automatic annotation method which uses Hidden Markov Models to label speech with Tilt intonation labels is presented. The F0 synthesis models consist of a set of regression trees which are trained to predict parameterized descriptions of fundamental frequency contours. Within the context of these two systems, the role of segmental effects on intonation is reviewed and empirically tested. Finally, a number of methods which may be used to evaluate the output of the two systems are reviewed.

As Figure 1.1 shows, several areas must be addressed to automatically create intonation synthesis models from data. While the primary goal of the research in this thesis is to build stochastic models which can be used to synthesize fundamental frequency, the initial problem addressed is one of data availability. Statistical models generally require substantial data. Automatic intonation labelling is an alternative to manual intonation labelling, which has been the preferred method of data acquisition in the past. An approach to automatic intonation labelling is discussed in Chapter 4. Prior to the discussion of the automatic labelling models examined in the research, Chapter 2 provides a review of literature on segmental effects on fundamental frequency. This review concludes by suggesting that accounting for segmental effects on F0 will provide better automatic intonation analysis and synthesis than would be available from a system which did not account for such effects. This suggestion is tested within the research in Chapters 4 and 5.

Before the synthesis or analysis models can be built, an intonation model, or framework, has to be chosen. Chapter 3 reviews four common approaches to intonation modelling. This review includes discussions of framework-internal specifics as well as actual applications which follow each approach. The Tilt intonation model was chosen for the framework for this research, based on these discussions.

With the framework and background in place, Chapter 4 returns to the question of data acquisition. The goal of the research in this chapter is to improve on previous efforts to automatically annotate speech with intonation labels. The experiments test the suggestion that including information which could account for segmental effects on intonation would assist in this goal.

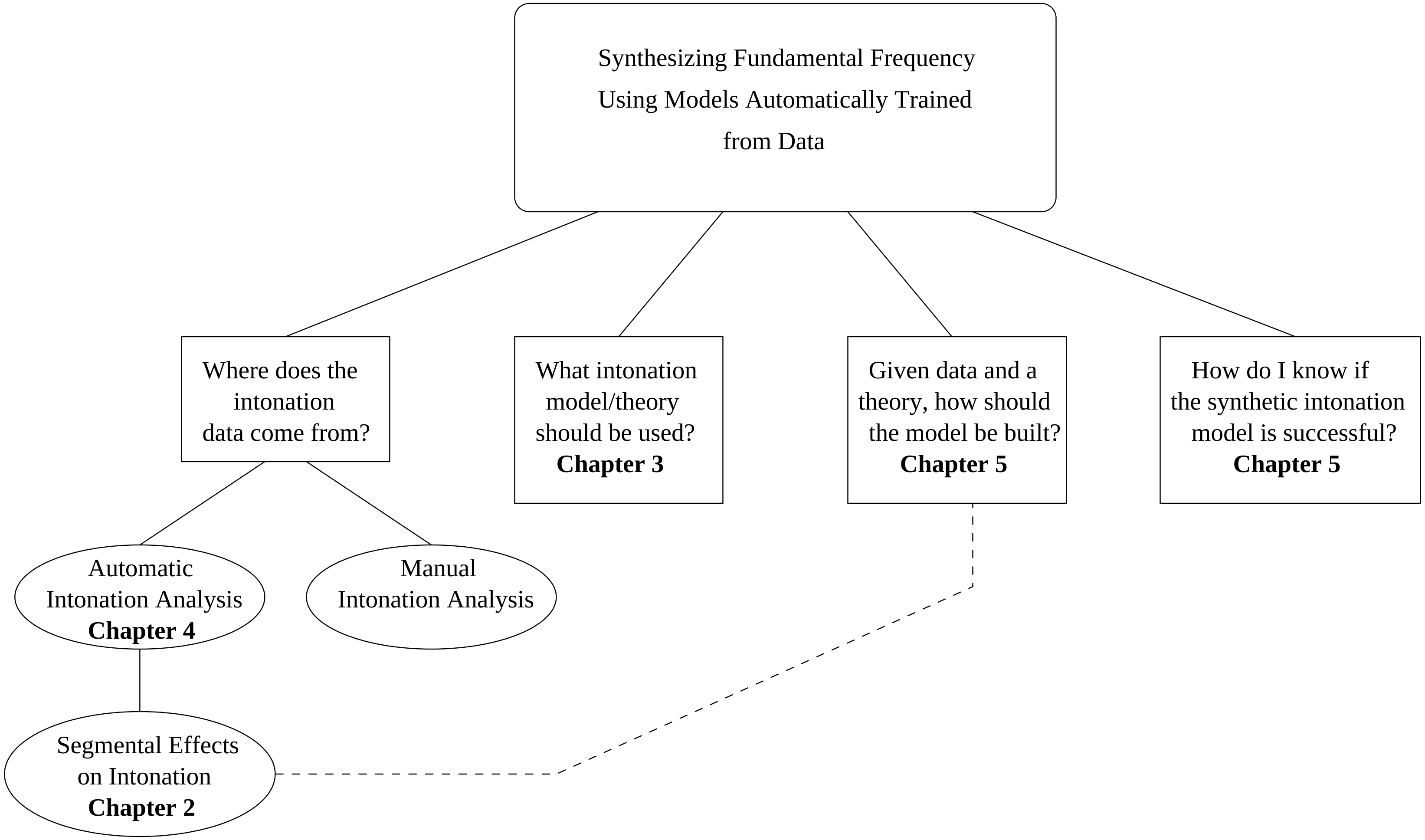


Figure 1.1: Schematic diagram of thesis structure

The results of these experiments show that including Mel Frequency Cepstral Coefficient data improves the performance of a methodology which previously used only fundamental frequency and RMSE energy measurements. The added low-level acoustic information improves the number of correctly labelled intonation events, and in some cases reduce the number of erroneous labels. However, the automatic labelling models are only successful when the database which they modelled is large enough to provide many examples of each intonation event type. Therefore, while the methodology is sound, it is more useful as an aid to manual annotation until such a time as more data is available to build more robust models.

The main body of research in this thesis concerns using annotated speech data to automatically build and train fundamental frequency synthesis models. Chapter 5 discusses the experiments which were used in the composition of a methodology to produce such models. This training consists of providing information to a regression tree building system which automatically chooses what parts of the information to use when building the trees. The trees predict parameterized descriptions of the fundamental frequency contour for the synthesized utterance. The data features used in the trees give detailed descriptions of the text which is to be synthesized. The data features which the decision trees use were selected based on the theoretical and experimental literature reviewed in Chapter 3. These data features are pared down using an automatic step-wise data reduction method. An advantage of using this type of algorithm is that only features which are necessary for training a tree are used in the training, allowing a large number of features to be tested without undue problems of inter-feature noise. The greatest advantage of the overall methodology is that it is possible to investigate the utility of features and feature classes when the training algorithm is combined with the use of the Tilt model, which has parameters for individual aspects of intonation events. This aspect of the modelling technique is an improvement on previous research, which does not present the contribution of different information types within a system.

The definition of success criteria for intonation synthesis is an extremely difficult task. Objective evaluation of intonation processing output is vital if researchers are to have immediate feedback on the progress of small changes in their systems. As section 5.5 shows, subjective methods of evaluation, which have been applied to the output of the synthesis methodology used in this thesis, do not always provide the correct information. Objective measurement techniques which are designed for use with fundamental frequency evaluations would be ideal. However, at present, the currently best supported metric is Root Mean Squared Error. When viewed in concert with knowledge of a speaker's pitch range, this metric is a useful tool for assessing the success of an intonation synthesis system. As Chapter 5 shows, the methodology developed within this thesis is as successful as previous techniques at producing intonation similar to the natural intonation on which it is based. Section 5.4 shows that training synthesis models from automatically annotated data is nearly as successful as training the models from manually labelled data. Thus, the synthesis of F0 from models automatically trained from data can be extend to include data-labelling models which are also automatically trained from data.

Chapter 2

Segments and Intonation

Until recently, research in segmental acoustics and intonation research were virtually mutually exclusive. Researchers worked very hard to separate the two as early as possible. In synthesis, those interested in the linguistic aspects of an interaction joined in work on improving the quality of synthetic speech, and enjoyed some success. This chapter examines investigations into the nature and use of micro-intonation within the context of intonation processing. A presentation of segmental interaction with intonation expands on the brief introduction into micro-intonation from the previous chapter. A discussion of uses for micro-intonation within speech applications follows, covering past and current approaches to including information about segments in suprasegmental research. This discussion provides a background for the investigations into adding aspects of micro-intonation to both the analysis and synthesis of intonation, as examined in chapters 4 and 5.

2.1 Joining Source and Filter

Looking at intonation with Fant's source/filter model ([Fan60]) in mind, macro-intonation refers to the fundamental frequency of the voice source (glottal waveform), and micro-intonation reflects interactions between the source and the vocal tract (filter). This simple interpretation of Fant's model has formed the basis of much intonation work in the past. Fundmental frequency contours are smoothed to remove any trace of micro-intonation, with the justification that the smoothing results in a clearer picture of intentional pitch movements. This justification assumes complete independence of the source from the filter. The independence of the source and filter breaks down somewhat when one considers the effects the filter has on the communicative pitch movements (intonation events). As Reinholt Petersen states:

> It is a well-established fact that the fundamental frequency (F0) of speech is not only determined by higher-level linguistic information such as sentence type, stress pattern, and tone, but also by the segments constituting an utterance. ([Pet86]:31)

This chapter is not concerned with a detailed analysis of micro-intonation, but with interactions between micro-intonation and macro-intonation. Such interactions can be investigated by examining the correlations between regions of macro-intonation (e.g. pitch accents) and categories of segments associated with those areas. Intonation synthesis research (e.g. [vSH94]) has shown that the segmental text of an utterance is highly correlated with both the timing and size of intonation events. Because the fundamental frequency is being purposely modulated during intonation events, the interaction between such events and the vocal tract is complex. The question remains as to whether correlations between event shape and linguistic text are the result of the vocal tract (filter), communicative necessity (e.g. placing a pitch movement over a consonant cluster is not an effective way of using pitch to signal meaning), or both. This thesis assumes the third option, and attempts to explore and exploit some of these correlations, without becoming lost in an exhaustive examination of the filter effects on F0 over "intonationally insignificant" text.

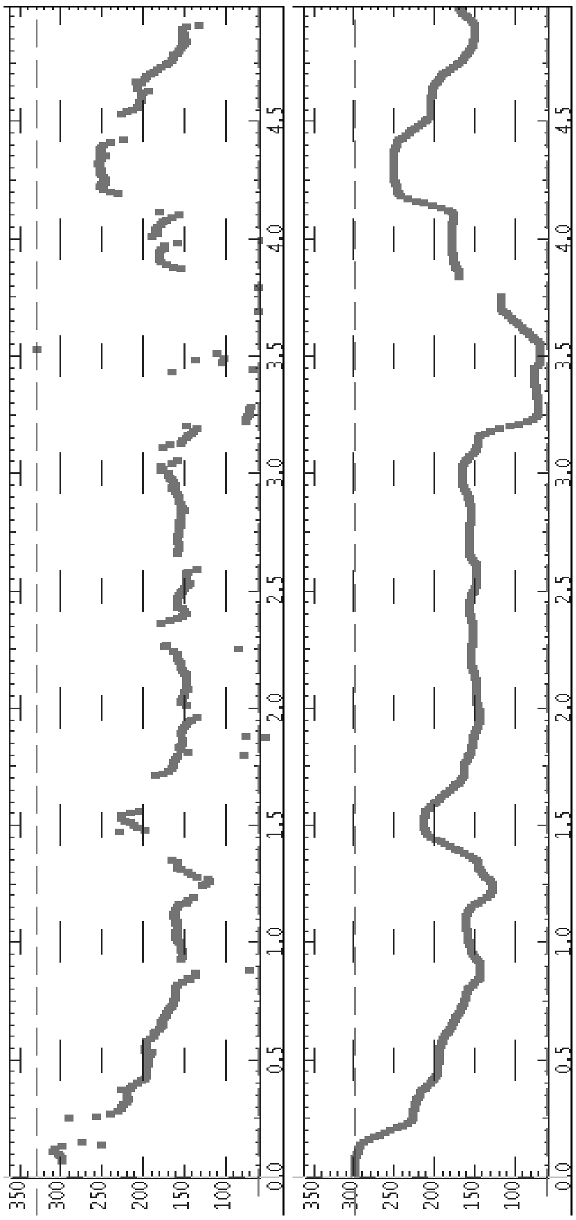
The perceptual relevance of micro-intonation has been the subject of much research and discussion. While researchers generally acknowledge that segmental interaction with F0 exists, there has been a wide disregard for its possible importance. Mertens *et al* [MBd97] go to great lengths to determine the threshold at which a pitch movement may be classified as "macro-prosodic." The reason for this effort is that microprosodic pitch excursions vary in size, with the largest movements sometimes being larger than audible pitch accents [Sil87]. If micro-intonation is capable of greater magnitude than macro-intonation, then the question of perceptual relevance becomes clouded. Do listeners notice all pitch excursions? If they do, how do they know which ones are macro-intonation and which are micro-intonation?

The research in chapters 4 and 5 is concerned with exploiting three types of interaction between linguistic text and intonation events. The first type of interaction is what generally falls under the micro-intonation banner - vowel intrinsic F0 and coarticulation F0 effects on intonation events. The second type of interaction concerns subsyllable constituents (onsets and codas) and their relationships with event peak height and timing. The final interaction is between individual segments and salient points within intonation events. The rest of this chapter discusses these interactions in sufficient detail to provide the necessary background for the new research presented in this thesis.

2.2 The Relevance and Function of Micro-Intonation

The small perturbations in F0 which are generally attributed to the shape of the vocal tract have caused confusion and difficulty for many researchers. They create difficulties for pitch tracking (e.g. [MYC91]) as well as adding to the problems that intonation researchers must contend with. Traditionally, the favored method of studying micro-intonation has been to study ways of eliminating it. Bagshaw [BHJ93] tested a number of methods for extracting fundamental frequency, finding the super-resolution pitch detection method [MYC91] superior to others he tested. Taylor [TCB98] improved on the super-resolution pitch detection (SRPD) algorithm by introducing smoothing to create an "intonation contour."

Figure 2.1 shows one F0 contour which was extracted using the SRPD method, with minimal smoothing, and the intonation contour for the same utterance, extracted using Taylor's "intonation contour detection algorithm" (ICDA). Notice that the outlying pitch readings in the SRPD contour are eradicated in the ICDA contour. ICDA contours are very useful tools for the human labeller. The lack of visual "noise" from the outlying dots, as well as the ease of a fully interpolated and continuous line, enable the labeller to speedily pick major peaks and troughs for further consideration. Pitch-tracking errors (e.g. jumps of 200Hz in 10ms as near 3.5sec) are easily observable in the isolated points of the SRPD contour. Micro-intonation remains observable as small bumps in the ICDA contour (e.g. near 2.5sec). While both of these types of F0 movement can obstruct intonation analysis, only the micro-intonation is of interest in this thesis. As Figure 2.1b shows, it is possible to minimize the pitch tracking errors while retaining micro-intonation



(a) SRPD Contour (b) ICDA Contour

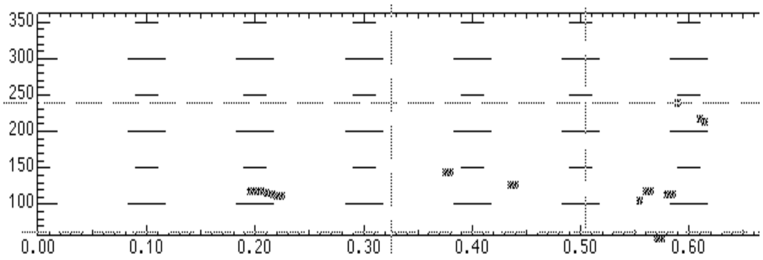
Figure 2.1: F0 contours extracted using SRPD and ICDA methods

information. In the last fifteen years, the body of micro-intonation research has been slowly moving towards understanding what causes micro-intonation, and how such knowledge might be applied to speech processing. In this thesis, micro-intonation is investigated in relation to macro-intonation. That is, only areas where macro- and micro-intonation overlap are studied. The reason for this focus is, as discussed below, that intonation processing is sufficiently poor that good micro-intonational models can be completely overshadowed by poor macro-intonation.

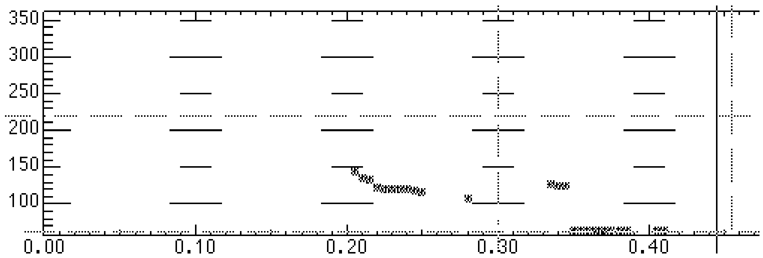
Two basic types of micro-intonation are discussed in this section - coarticulation effects on F0 (CF0) and vowel intrinsic F0 (IF0). Coarticulation effects on F0 generally consist of perturbations in F0 of varying sizes which are the result of vocal tract shape and movement during non-vocalic speech. Vowel intrinsic F0 is a variation in F0 level due to the geometry of the vocal tract and postioning of the tongue during vocalic speech. The literature suggests that both types of micro-intonation are perceived by listeners, to such an extent that they are actively removed from speech should they be likely to cause confusion with intonation (e.g. [Sil87]). Gandour and Weinberg [GW80] show that vowel intrisic F0 is produced by laryngectomized speakers with equivalent magnitudes of non-laryngectomized speakers, which also suggests that micro-intonation plays a role in speech perception. This chapter presents an introduction to micro-intonation. Silverman [Sil87] provides a thorough analysis of micro-intonation with a comprehensive review of basic literature.

Coarticulation effects on F0 often appear on raw F0 traces as sudden jumps or outlying points. These steep movements are often removed from the contour using smoothing techniques, as discussed above. Frequently, though, the perturbations are less obvious. These perturbations are the effect of a shift in F0 over a longer period of time. Effects of this type can look like a small pitch accent, and have been shown to cause low accuracy in automatic analysis methods (e.g. [MBd97]). An illustration of coarticulation effects on fundamental frequency is shown in Figure 2.2, where the F0 during /ti/ is shown to start slightly higher than during /di/. These two syllables are in the context "a X ta." The two syllables are marked by the vertical dashed lines. Though the F0 is not as clear in some longer speech sections, the contours do show a clear difference between /ti/ and /di/. The F0 during the /i/ portion of /ti/ is represented in the picture by two dots at approximately 150Hz and 125Hz. Interpolating between these points gives an indication of the F0 for the /i/ segment. Similarly, the best indication of F0 for /di/ is in the section of contour at 130Hz. Looking at these two sections of contour, it is clear that /ti/ begins roughly 20Hz higher than /di/. While it is will always be possible to find examples of /ti/ which are higher than /di/, and vice versa, Figure 2.2 illustrates the type of differences which can be expected based on experimental research such as Silverman ([Sil87]).

In an experiment to examine the main effects that consonants have on nearby F0, Silverman presented subjects with target nonsense words of the type @Cv_1Cv_1C followed by @ (where @ is schwa, and the consonant is the same) in a carrier phrase. He found that F0 falls before consonants, with a steeper fall before voiceless consonants than voiced ones. He also found that the magnitude of the perturbations increases with the amount of stress placed on a syllable. Other interesting findings from this experiment were that the F0 perturbations were the same across stops and fricatives, and that sonorant consonants also perturb F0. Silverman also found that post-consonantal perturbations do not depend on vowel height, which implies that CF0 and IF0 are indeed two separate types of micro-intonation. Figure 2.3 contains



(a) /ti/



(b) /di/

Figure 2.2: Fundamental frequency contours over two syllables (each bounded by vertical dashed lines)

examples of micro-intonation which looks similar to macro-intonation. The smoothing of the intonation contour will have removed some segmentally-induced perturbation. However, not all micro-intonational F0 movement is gone. One example of coarticulation effects on F0 is found between 28 and 28.5 seconds, during “for an.” This small bump in F0, which occurs during the transition out of the /r/, is still large enough and of the correct shape to potentially disrupt automatic analysis. On the larger end of the perturbation scale is the large F0 movement at the end of “emergency.” This steep drop occurs during an unstressed syllable, with no audible pitch change. This drop could be the result of effects of the /m/ in “meeting” on the preceding vowel, and is easily large enough to confuse an intonation analysis system.

Vowel intrinsic F0 is a natural variation in F0 level due to the type of vowel being uttered, (e.g. /i/ has a higher IF0 than /a/). Like CF0, IF0 magnitudes are often greater during stressed syllables. Silverman finds that IFO magnitudes can be of the same order as pitch accent magnitudes (sometimes over 20Hz). Therefore, vowel intrinsic F0 poses problems for intonation analysis, in that either human or computer must decide whether a perturbation is macro- or micro-intonational. Silverman and Reinholt Petersen ([RP80], [Pet86]), find that listeners adjust their perception of IFO, to the extent that, all else being equal, a high vowel is perceived as having a lower pitch than a low vowel, even when the F0 is the same. This can cause difficulties for magnitude estimation of intonation events. During synthesis, it may be necessary to increase or decrease the magnitude of an event based on the vowel associated with it. Figure 2.4 shows an example of vowel intrinsic F0 differences. These two pictures show the syllables /pi/ and /pa/ in precisely the same intonation and segmental context (within the nonsense words “apita” and “apata”). The difference between /pi/ and /pa/ is approximately 10Hz

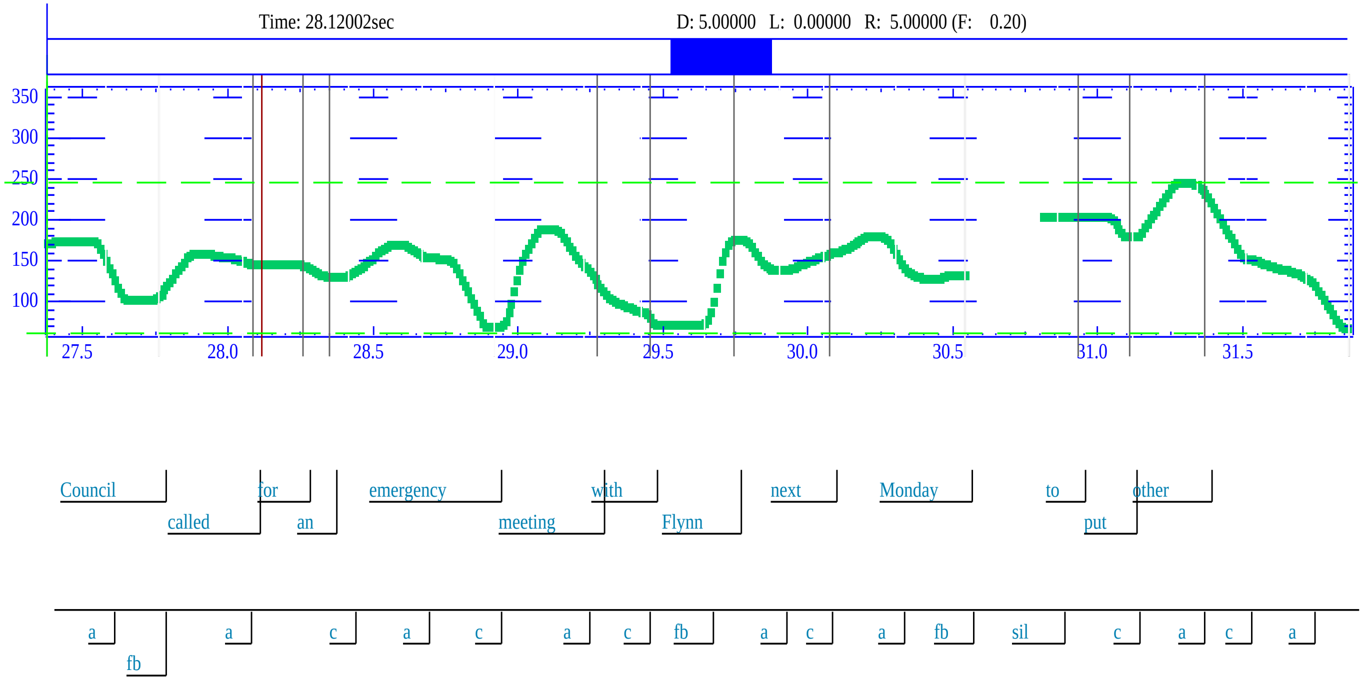
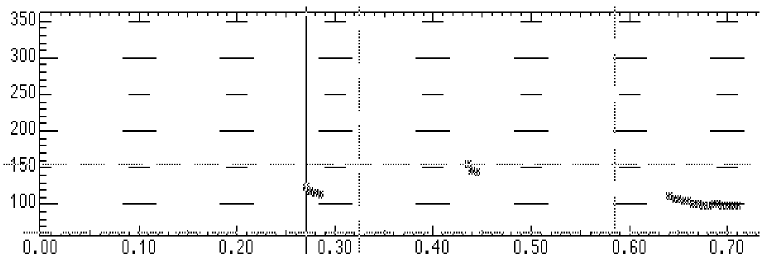
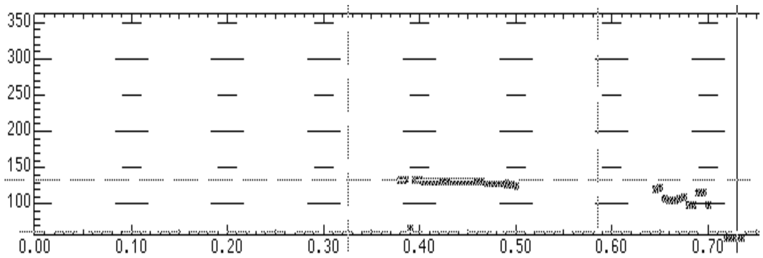


Figure 2.3: An example contour containing micro-intonational movement

in this case. Even though much of /i/ is lost in aspiration, an intonation processing system would still have to model the F0 in such situations. One would possibly want to synthesize the extra magnitude and would certainly not want to analyze the extra height as a pitch event.



(a) /pi/



(b) /pa/

Figure 2.4: An example of vowel intrinsic F0 differences between two syllables (bounded by the dashed lines)

Figure 2.3 contains an interesting example of vowel intrinsic F0. The falling boundary associated with "Monday" begins to fall, but levels off during the latter half of the dipthong. It is probable that an /a/ in the same location for this speaker would continue the sharp F0 drop, rather than levelling off. This conjunction of micro- and macro-intonation is a good example of the type of interaction which is examined further in chapters 4 and 5. In terms of intonation synthesis, it may be important to raise or lower the F0 slightly over some syllables regardless of their accentual status, simply because of the type of vowel they contain. For analysis and synthesis alike, it may be the case that accents will have different magnitudes based on the vowel type. Such effects, if present, should be accounted for, and could be quite useful for intonation processing.

If one intends to use micro-intonation within an intonation processing application, its importance to listeners should be taken into account. It becomes necessary to determine the way in which people perceive and use micro-intonation. DiCristo and Hirst [DCH86] argue that micro-intonation is both perceived and used by listeners. They note that coarticulation effects on F0 can be quite large (up to three semi-tones), and can influence the F0 level throughout the syllable in which they occur. While vowel intrinsic F0 effects are smaller (they observe a 1-2 semi-tone difference between open and close vowels), these effects also appear large enough to be noticeable. DiCristo and Hirst claim these effects may be useful for segmental identification by listeners. These findings are consistent with a number of previous studies ([LS84], [Pet86], [Sil87]).

Interestingly, DiCristo and Hirst also find that their intonation synthesis algorithm is not significantly improved perceptually by including a coarticulation effects model. Coupled with the finding that vowel intrinsic intonation effects are less in continuous speech than isolated words [LS84], one wonders what forms the basis of any belief in the importance of micro-intonation in intonation synthesis. The following paragraphs explain this basis.

Reinholt Petersen [Pet86] provides support for including micro-intonation in speech synthesis in his work on how and where listeners perceive micro-intonation. He presented listeners with a synthetic vowel continuum from /u:/ to /o:/ within a two-syllable carrier nonsense word, with F0 in one of three pitch ranges (high, middle, low). The subjects reliably judged low-pitched vowels in the middle of the continuum as /o:/, and /u:/ for the same vowels in the high-pitched condition. He concludes that while micro-intonation is used in the perception and disambiguation of segments, they are effectively "compensated" for prosodically. In other words, listeners do perceive and use micro-intonation, but they use it in the segmental domain. The effects of segments on intonation are used to assist in segmental perception, rather than intonation perception. Understanding of prosody allows listeners not to be confused.

Silverman [Sil87] includes models of micro-intonation in his rule-based intonation synthesis system. IF0 and CF0 are modelled separately, and the frequency contours are added to base intonation contours. In tests which asked subjects to rank two synthetic voices as more or less natural than the other, where the only difference was the intonation synthesis method, Silverman's method was ranked higher in five out of six cases. However, no systematic evaluation of the contributions to the individual perturbation models was performed which would decisively explain the reasons for the better results. Perhaps the improved results Silverman attained by adding micro-intonation to his synthesizer are as much to do with removing segmental ambiguities as with reducing pitch ambiguities.

If DiCristo and Hirst have examined the effect of including CF0 in their intonation synthesis and found no improvement in the perception of intonation by subjects, it may be because the effects of the inclusion are not realized in the perception of intonation. They find segmental effects from both CFO and IFO which improve overall quality, but their intonation model is not improved by the discovery. Therefore, if, as Vaino *et al* [VAKA97] assert, modelling micro-intonation will improve a synthesizer's quality, *all* *other things being equal*, then such a model may be relevant to speech synthesis. However, it seems that it is not relevant to perception of "intonationally significant" pitch movements. It seems logical, therefore, that it is more important in intonation synthesis to correctly model macro-intonation than the micro-intonation over intonational "connective tissue." Dusterhoff and Black [DB97] amend this conclusion by showing that some aspects of segment-F0 interaction are useful in improving specific aspects of the statistical models. In terms of intonation analysis, however, any information which may help distinguish between large micro-intonation (or pitch tracking errors) and small macro-intonation could be useful in automatically labelling speech for macro-intonation. Improved distinction between macro- and micro-intonational movements should result in more small intonation events receiving correct event labels and fewer micro-intonational movements incorrectly receiving event labels.

2.3 Subsyllable Units and Intonation

The consensus of the researchers discussed is that *speech* quality, rather than intonation quality, is likely to improve with the inclusion of micro-intonation. However, exploiting relationships between segments and intonation can be useful in modelling intonation event peaks for synthesis. Van Santen and Hirschberg [vSH94] and others (e.g. [PvSH95], [vSM97], [DB97] show that segmental content is associated with both the timing and height of F0 peaks during pitch accents. All of these authors utilize the categorizations of onset and coda from Van Santen and Hirschberg [vSH94] to enhance stochastic models for intonation generation.

Van Santen and Hirschberg [vSH94] examine the effects of segmental context on pitch accent height and alignment. They divide onsets and codas into the following three categories, based on the least sonorant portion of the unit.

- -V (voiceless consonants)
- +V-S (voiced obstruents, including null onset)
- +S (sonorants, including null coda)

Table 2.1 shows some examples of these categories. Null onsets were classified as +V-S due to the glottal stops which invariably preceded them in the data. Null codas were classified as +S, ostensibly to allow for heavy vowel constructions which mimic a CVV-type construction.

-V	/st/,/sl/
+V-S	/b/,/dr/
+S	/n/,/r/

Table 2.1: Examples of Onset and Coda Classificatoin

The study examined phrase final nuclear accents. Van Santen and Hirschberg observe that peak location varies systematically according to onset and coda type. They find that differences in peak location related to onset type can be partially explained in the different lengths of the three types (e.g. -V consonants averaged 173ms to 104ms for +S consonants). However, correlations between peak location and coda class could not be reduced to such simple explanations. They report that peak alignment can be accurately determined given segment durations and coda class, without reference to onset class or vowel height.

In their study, van Santen and Hirschberg also confirm the claim (by [Pet86] and [LS84], among others) that vowel intrinsic F0 has a greater effect on pitch accented syllables than on unaccented syllables. The effect can be viewed as an amplification of the accent which is larger the higher the vowel. They also find that longer syllables tend to have greater F0 height movements than shorter syllables, and that onset-induced F0 excursions tend to last only a short time ($<$ 50ms).

All of the findings in van Santen and Hirschberg (1994) have applications in modelling F0 curves over intonation events (accents and boundaries). While the initial study examined short sentences (e.g. "Now I know X"), the onset and coda classification system has been successfully adapted to other contexts and speech types (e.g. radio news broadcast) and shown noticeable improvement for F0 generation ([DB97]).

2.4 Segmental Anchor Points

Recent work by Ladd and others ([ALM98]), [LFFS99]) provide an interesting insight into intonation/segment interaction. Their studies have investigated anchor segments for intonation events (e.g. rises start at the beginning of an accented vowel and end at the beginning of the next vowel).

In a study of Greek rise (L+H*) accents, Arvaniti *et al* [ALM98] investigate regularities in L and H placements with regards to word, syllable, and segmental boundaries. Their three experiments cover accents placed on syllables in a variety of contexts. Post-accentual syllables were varied between one and five. Lexical stress location was varied in one study and remained constant (antepenult) in the other two. In the most constrained experiment (low number of sentences repeated six times), the peak occurred, on average, 10ms into the vowel (SD=14ms). The peak and the onset of the first post-accentual vowel correlated heavily (R^2=0.806). In less constrained environments, where both the sentence content and the accented syllable length were varied, the timing was similar (mean 17ms into the vowel, SD=32ms), and the correlation was lower, but still high enough to suggest a pattern (R^2=.453). In no case was the length of the post-accentual vowel shown to be correlated with the distance from the vowel onset to the peak ($R^2 < 0.010$). Their final experiment was a pilot which judged the effect of tonal crowding on the peak anchor. Tonal crowding occurs when multiple intonation events occur on a single or successive syllables. Crowded events tend to be truncated or compressed, when compared with uncrowded events. Of their five subjects, only one showed any evidence of tonal crowding effects. As a whole, these results suggest a segmental anchoring pattern.

Investigating the effects of altered speech rate on the alignment of rising accent start and peak points, Ladd *et al* asked speakers to change their speech rate while reading a short piece of text. Thirteen test words were placed in locations conducive to rise accents (**Adjective** + Noun, **Adverb** + Verb). Six subjects were tested. At a "normal" speech rate, the onset of the rise was consistently within 25ms of the beginning of the onset consonant of the accented syllable for five of the six subjects. At "fast" and "slow" rates, the rise and syllable onsets were consistently within 20ms of each other for all six subjects. This finding supports Arvaniti *et al* by showing that the rise onset's segmental anchor is not likely to be affected by speech rate.

The peak of the rise accent was measured in relation to both the offset of the accented vowel and the onset of the following vowel. The results of these experiments were less clear-cut than those for the rise onset. The slower speech rate appeared to cause difficulties due to inserted prosodic boundaries. Therefore, the slow rate data was removed for a clearer picture. In the fast and normal speech rate, rate had, at best, a marginally significant effect on the alignment of the peak in relation to accented vowel offset (p=0.047). The effect of rate on peak alignment relative to the onset of the following vowel was insignificant, supporting the claim that the peak is aligned to segments, rather than being the result of a fixed time or slope function. This result further supports Arvaniti *et al.*

Though the experimental work on this topic is still in its early stages, the concept of segmental anchoring is interesting. In a survey of the F2B database (see section 4.2.3 for details of this database), where syllabification is derived from lexical lookup, 2275 of 2700 accent peaks occurred during the final consonant of the accented syllable. These results are rather different from Ladd's and Arvaniti's. In Ladd *et al* and Arvaniti *et al* the intonation contours are not interpolated through voiceless speech, as the target words are all sonorants, while the contours used in the survey of F2B were interpolated through voiceless speech. This difference in contour type could explain why more peaks fall during consonants than in Ladd *et al.* Another possible reason for the large difference could be the relatively unconstrained nature of the F2B database, as compared with the experimental conditions under which Ladd's data was acquired. Ladd's analyses are based on a small number of syllables which included only sonorant phones (9 test words in one experiment, 14 in another). The F2B data is not constrained for segmental content.

Should segmental anchors be found systematically occurring in natural speech and speaking conditions, they could be quite useful in building rule-driven intonation synthesis systems. Segmental anchors could also prove helpful in constraining statistical models of intonation. The use of anchor points was briefly investigated within the context of the research in Chapter 5. However, the information which anchor points provided to the statistical models was already provided by other features, which resulted in anchor points not being used in any of the experiments detailed in Chapter 5.

2.5 Intonation Analysis and Segments

While researchers have noted the difficulties in analyzing intonation which micro-intonation can cause, little has been done to address the problem. The study of segmental effects on intonation for analysis purposes remains almost completely in the area of how to avoid such effects, rather than in qualitative or quantitative examinations of them. Taylor's Intonation Contour Detection Algorithm [TCB98] is one of the latest examples of the desire to smooth an F0 until almost no segmental perturbations are left. Taylor's algorithm allows the user to select the amount of F0 smoothing. For the purposes of manually labelling data, for example, a very smooth F0 may be desirable. For the experiments within this thesis, the smoothing is minimized, as is discussed in section 4.2.3. Mertens *et al* [MBd97] search for ways of stylizing F0 contours automatically, which is essentially the same process. They found that, even with a F0 smoothing algorithm, some of the large coarticulation excursions remained. These excursions generally look like small-to-medium sized pitch accents, and confuse intonation analysis.

Ljolje and Fallside [LF87] chose instead to control their dataset in their work to model "rise," "fall," "rise-fall," and "fall-rise" pitch movements over isolated words. Their analysis system, which used Hidden Markov Models to model the pitch movements, was designed to automatically recognize the above classes of movements. Rather than use smoothing algorithms, they control for coarticulation effects and vowel intrinsic pitch by using minimal pair word sets. Thus, each type of segmental effect would be modelled in the context of the type of pitch movement over a single word. While not necessarily an ideal method for use with continuous speech recognition, the research shows that there are ways of working with the segmental content of data without destroying data through excessive smoothing.

However, as Mertens and his colleagues noted, ignoring the effects of segmental context on F0 is non-trivial, and perhaps using the context in some way would be more productive. Ostendorf and Ross [OR97] use a stochastic segment model which was originally meant for phone modelling to attempt to give the segmental context (phone labels). As is discussed in Chapter 3, their accent detection is quite successful, while boundary detection is less accurate.

Chapter 4 discusses continued work on including segmental data in intonation analysis. The research examines several acoustic aspects of the segmental stream and attempts to exploit them in automatic intonation analysis.

2.6 Summary

This chapter has provided an insight into the complex nature of micro-intonation, including methods of addressing the interaction between segments and intonation. An interesting segmental anchoring hypothesis from Ladd *et al* is reviewed. Research into interactions between sub-syllable units (onsets, rhymes, codas) and intonation event peak height and location is presented, and placed into the context of intonation processing. Much of the research that forms the basis of this thesis is an extension of this movement towards incorporating segments into intonation processing.

Chapter 3

Intonation Models for Intonation Processing

Intonation plays a humanizing role in speech synthesis. By emulating the intonation patterns of human speakers, speech synthesizers can take on the characteristics of different languages, dialects, or speakers. One of the more difficult tasks in building models for speech synthesis is capturing the variety of intonation patterns that occur in natural speech. Early models of intonation (e.g. [OA61]) consisted of a small number of canonical forms. Such a system captures the basic needs of language learners, who, with increased use, are then able to increase their "intonational vocabularies." However, writing rules for all possible, or even probable, intonation contours is impossible. A compromise between modelling a small number of prototypical intonation contours and modelling all likely forms is to model as many types of intonation event as possible over a specific style of speaking. If one desires a synthesizer which will answer the phone and relay information to callers, then only a specific style of intonation is necessary. Similarly, if the task is to read news over the Internet, a different speaking style should be modelled. This approach is similar to O'Connor and Arnold [OA61] in its use of prototypical intonation contours. Both abstract away from F0 contours. O'Connor and Arnold view the prototypical contours as whole units. Unlike O'Connor and Arnold, though, the models which are discussed in this thesis approach intonation contours as sequences of events.

This chapter presents a some of the more popular intonation models which have been used for automatic intonation synthesis, analysis, or both. This chapter also includes a review of techniques for evaluating synthetic intonation and intonation analysis output. Once a model is presented, some applications which use the model are presented. The applications are discussed in terms of evaluation techniques reviewed in this chapter.

There are several different approaches to intonation event modelling, and each approach has spawned multiple variations. In some cases, the difference between models begins with a difference in the theoretical assumptions of the modellers. The AM (Autosegmental-Metrical) school represents intonation as a sequence of tone levels. IPO[1] has rejected the tonal representation, and treats intonation as a sequence of pitch movements. The superpositional models are based on a physiological model of the speech chain ([Fuj83]), combined with a hierarchical theory of prosodic phonology [NV86]. These models represent intonation as a sequence of events whose domains overlap. Continuous parameterized models attempt to interpret F0 in an acoustic domain, describing intonation in terms of F0 movement over time (e.g. rise and fall height, slope, duration). Each of these methods has its strong points. Some methods, however, may lend themselves more readily than others to the automatic intonation processing tasks which form the basis of this thesis.

Each of the modelling techniques is represented by at least one application

[1] Instituut voor Perceptie Onderzoek (Institute for Perceptual Research)

for intonation synthesis. The applications discussed here were chosen because they are among the most successful instances of the various techniques and because each application illustrates difficulties in automated intonation processing. The ToBI (an instance of the AM approach, [SBP$^+$92]) implementations are effective (e.g. [Ros94], and result in a fully automatic system for modelling intonation for speech synthesis from an annotated database. The superpositional (SP) implementation is also fully automatic, and is based on an annotated database ([Spr98]). Beaugendre's French IPO model is automatically derived from an annotated database, but the required automatic stylization system is not integrated as yet. The Tilt synthesis implementation is, like the other synthesis modelling techniques, automated and built on the basis of an annotated database.

Regardless of the methods one uses to process intonation, some type of evaluation of the method is required. Both subjective and objective assessment techniques have been applied to synthetic intonation. Intonation analysis also requires evaluation, either by comparing symbolic output with original symbolic data, or by using an analysis-by-synthesis method. Analysis-by-synthesis is an assessment method whereby the intonation analysis output becomes input to an intonation synthesis system, and the resulting synthetic output is evaluated using one of the methods for evaluating synthetic intonation. The following section introduces the basic ideas behind intonation evaluation. Specific evaluation methods are discussed in more detail together with the intonation processing applications that they assess.

3.1 Intonation Evaluation

One of the primary difficulties of analyzing synthetic intonation is how, or if, one should use subjective perceptual examinations. While it is desirable for people to listen to system output and proclaim that it sounds acceptable, the opinion of listeners is not always constructive. When naive subjects are used, it can be difficult to find out why they are giving one score or another. The naivete that makes them so useful for opinions uncoloured by theory and sensitivity to the subject matter also acts as a barrier if explanations and discussions of their opinions are required. It is important to know the reasons behind the opinion as well as the opinion itself. For example, if ten out of ten subjects rate an utterance as acceptable, or even as natural, this must be qualified. How does that result compare to previous results? What has changed since the previous test? Which changes in the system could have caused the change in results? In order to know precisely what the scores reflect, it is necessary to carry out a subjective test every time the synthesis system changes.

Similarly, while an objective metric which implies that an utterance has been produced with intonation just like the original may be useful, there is a question of how fine a judgement is available for less-than-perfect intonation. It is necessary to know how well the metric relates to perceived intonation quality. Given that no known speech synthesis system produces intonation with the controlled variation in pitch and timing of your average human, it is necessary for an objective metric to give an insight not just into how good intonation is, but also how much, and in what ways, it has improved or deteriorated.

As mentioned above, assessing intonation analyses is a necessary part of intonation processing. The benefit of an analysis-by-synthesis approach to evaluating intonation analysis methods is that it allows one to use the same evaluation methods for analysis output as for synthesis output. Therefore, the discussion of methodology above is relevant to both directions of intonation processing. Alternatively, the use of symbolic output of intonation analysis systems for evaluation creates a parallel with other speech recognition tasks. If the symbolic output matches the symbolic representation of the data, the analysis system is successful.

The parallel with speech recognition assessment gives rise to its own difficulties. In speech recognition tasks, if "I would like a large pepperoni pizza, please" is recognized as "I would like a large pepperoni pizza, please" then the analysis is generally successful (limitations to this assumption are discussed in section 4.2.5). However, unlike textual recognition, where the text is the only important output, intonation evaluation involves the time at which the symbol is "recognized" as well as the recognition of the symbol itself. Recognizing, correctly, that there are three accents and a falling boundary on an utterance is only successful if the accents and boundary are recognized as being in the correct location. With intonation analysis, it is quite possible for the symbolic representation to be the same in the analysis output and the original data, while the timing is completely different. For example, with a poor recognizer, the sequence "Accent Accent Falling Boundary" may be output on the first word of a five word phrase. Using evaluation methods which only assess whether the output string is correct could result in a score of 100%, where in truth, the output is exceptionally wrong, because the intonation events are not in the right place. Thus, the timing of the symbols is at least as important, if not more so, than what they are called. This evaluation technique is revisited in Chapter 4, in a discussion of the intonation analysis methodology which was developed for this thesis.

3.1.1 Subjective Evaluation of Synthetic Intonation

Any evaluation of synthetic speech has three basic tasks: determining the quantity of "good" speech (how much of the output is understandable?), assessing the quality of the speech as a whole (is there enough clear speech to out-weigh any errors?), and examining the ability of listeners to understand the message which the speech is meant to carry. Evaluating intonation is somewhat less straight-forward. Quantification of movements in F0 is at best a difficult task. While it is fairly simple to judge a symbolic string which is then translated into an acoustic signal (e.g. [ML90]), it is, in the long run, the quality of an acoustic signal which must be evaluated. Regardless of whether pitch accents are placed on the correct syllable, the fundamental frequency must conform to patterns of the language being used.

Perceptual evaluation of intonation comes in a variety of forms, all of which offer similar benefits and suffer similar difficulties. The primary benefit of perceptual tests is that they can provide an insight into the opinions of system users. In the end, regardless of experiments and statistics, synthetic speech which is unacceptable to its users is a failure.

The most accepted form of intonation assessment at present is some form of subjective perception test. Any improvement in intonation synthesis is immediately queried by other researchers until it has been subjected to such tests. Two of the methods widely used in intonation evaluation are pairwise comparison and acceptability ranking.

Pairwise comparisons generally compare the similarity of two utterances under a variety of conditions. A sound experiment requires a subject to recognize that two equivalent utterances (with either synthetic F0 or F0 regenerated from natural speech) are equivalent, while a *different* pair (with one synthetic and one natural) is ranked for similarity. The purpose of the first type of pair is to establish that the subjects can reliably complete the task, and to give a baseline against which the *different* pair's ratings are judged. Subjects have been successful in completing the judgement task (e.g. [dP83]), and van Bezooijen and Pols [vBP90] have confidence in the abilities of subjects to reliably judge suprasegmental quality. The biggest problem with the design of this sort of test is that it is only useful for short segments of speech, due to the difficulties with accurately remembering the details of intonation over longer utterances.

Acceptability ranking involves listening to an utterance and ranking it on a scale (typically 5 to 10 points) according to how natural the intonation sounds. Van Bezooijen and Pols [vBP90] note that such tests are usually undertaken using short utterances. The benefit of ranking over pairwise comparison is that the subject's memory limit when comparing paragraph-length passages is less likely to cause a problem. The subjects do not need to worry about the previous paragraph. They only need to decide how good the current one sounds. Isard and Pearson [IP88] illustrate the problem with not using long passages: isolated sentences may sound natural, while two or three sentences together may not, even if the same model is used to generate them. Therefore, in order to gain meaningful judgement, acceptability tests should place the type of utterance which the system is meant to generate in the context of its use (e.g. dialogue, telephone prompts, news reading).

The primary difficulty with subjective rankings is that they are dependent on the individual experiences and thoughts of each subject, as well as the quality of the speech. Monaghan and Ladd [ML90] also highlight the difficulty of evaluating subjective rankings. Individual examples of speech may score equally well (or poorly) for entirely different reasons, which are not always clear to the experimenter. Monaghan and Ladd attempted to lessen the unknowns by restricting the judgements to symbolic descriptions of intonation, rather than pitch. In this way, they eliminate noise from synthesis quality and acoustic interpretation of the symbolic descriptions of intonation which most current synthesizers use. In a similar attempt to reduce the possible reasons for poor evaluation, Dusterhoff and Black [DB97] use the original symbolic descriptions of utterances and test intonation generation based on equivalent intonation event placement and a stochastic generation model. Thus, any audible difference between an utterance synthesized using the original F0 and one using synthetic F0 is solely due to the difference between the two F0 contours. Similarly, any objectively measured difference between the original and synthetic contours could be reflected in the difference between the utterances synthesized from the contours. How important any given difference between contours is remains a difficult question which is revisited in the assessments of the applications throughout this thesis.

3.1.2 Objective Evaluations of Synthetic Intonation

While there are many subjective testing methods available, a sound test with a representative sample of subjects and examples is time consuming, and not always useful for day-to-day model development. Therefore, objective measures of F0 modelling are necessary. Two general techniques are used in developing the synthesis models in the research for this thesis. One technique evaluates how well individual aspects of the models represent the data, while the other evaluates how well the synthetic F0 of an utterance matches the natural one. The first technique is only applicable to the research in Chapter

5, and will be discussed there. The second technique is more widely used (e.g. [Ros94], [DB97]). The basic idea is to measure the difference between natural intonation for an utterance and intonation synthesized for the same utterance. Chapter 5 goes into more detail about the way this method is used. The most common objective evaluation metrics currently in use are Root Mean Squared Error (RMSE) and Pearson's Correlation ([Edw84], [GD82]). RMSE measures how far apart two intonation contours are, while Correlation shows how closely one contour (e.g the synthetic one) relates to another (e.g. the natural one) in direction and range. Root Mean Squared Error measures the distance between two contours on the time axis, such that the distance being measured at regular (e.g. 10ms) points is perpendicular to the time axis, regardless of the F0 shape. Similarly, Pearson's Correlation is calculated based on measurements every 10ms. The correlation coefficient measures the degree to which the variables are linearly related. Thus, a high correlation coefficient shows a close linear relation (which should be the case with two similar F0 contours from the same utterance), while a low coefficient shows that the linear relationship is not close: that the two lines are diverging regularly. These objective metrics cannot determine whether a variation in the synthetic contour makes the synthetic speech sound any more or less natural. They only measure how much two contours vary.

This section has provided a basic outline of the methods used to evaluate intonation processing applications. The many difficulties in such assessments are discussed as they arise in the discussions of the models and applications below.

3.2 Autosegmental-Metrical Intonation Modelling

Autosegmental descriptions of tone languages developed by Goldsmith (1976, [Gol76]) and others have been applied to intonation by proponents of the Autosegmental-Metrical school of intonation modelling. Previous attempts at describing intonation as tonal sequences (Pike, 1945) were discredited by their inability to deliver a predictable and consistent tonal inventory. Each pitch level found in a dataset was described as a new tone, resulting in four or five tone categories which did not behave in a categorical manner.

In 1980, Janet Pierrehumbert [Pie80] used an autosegmental approach to American English intonation which allowed her to describe the multiple tone categories of her predecessors with two basic tones, High and Low. This advance enabled intonation study to treat pitch contours as tonal sequences.

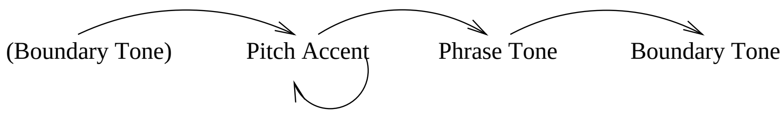


Figure 3.1: Finite state tonal grammar

Pierrehumbert's tonal inventory is used within a finite state grammar, which as shown in Figure 3.1, consists of an optional boundary tone, one or more pitch accents, a phrase tone, and a boundary tone. The grammar constrains the interaction of tone types, such that, for example, multiple final boundary tones cannot occur.

3.2.1 Tone Inventory

The basic tone inventory in Pierrehumbert's system consists of a High tone (H) and a Low tone (L). Tones are used to describe three basic types of intonation event: pitch accent, phrase tone, and boundary tone. Each event type is represented by adding a diacritic to the tone: * for pitch accents, $^-$ for phrase (floating) tones, and % for boundary tones. By combining the basic tones, Pierrehumbert arrives at a larger inventory which is used to describe rises and falls as well as level pitch. The L and H tones are combined to make L+H and H+L accents, which each have two further classes. The * diacritic is used to show which tone is aligned with an accented syllable (e.g. L+H* is an accent where the High tone aligns with the accented syllable and the pitch rises to the high from a preceding Low tone, while the L*+H aligns the Low tone with the accented syllable). The full inventory is shown in Table 3.1 (%H is sometimes used to mark an initial High boundary).

L*	L*+H	L+H*	L$^-$	L%
H*	H*+L	H+L*	H$^-$	H%

Table 3.1: Tone Inventory

3.2.2 Tonal Phonology

The use of tones as phonological units is the keystone of the AM approach to intonation modelling. In its original form, Pierrehumbert's system involved the process of downstep (a successive lowering of accent height) being an automatic result of interaction among tone types. The model also presents an interaction between tone types and phonetic realization, in rules of rightward tone spreading. In these rules, floating tones (X$^-$) spread rightward to more prominent tones. Thus, the floating tone (or phrase tone) would typically be associated with a series of syllables, resulting in, for example, a long low frequency trough. The ability to accurately describe downstep within the two-tone inventory makes Pierrehumbert's system more attractive than previous tonal systems (e.g. [Pik45]), which required a different tone for each level of downstep. Because accounting for downstep has been difficult for some approaches in the past, it is one important test of whether a model can cope with a wide range of intonation phenomena. The AM school has evolved in its approach to downstep, as is discussed below. The discussion of downstep serves not just as a presentation of an important aspect of an intonation model, but also shows how one intonation model has adapted and progressed from inception to successful application.

The original form of downstep as a phonological process [Pie80] states that, given a sequence of H L H which contains a bitonal accent, the accent following the bitonal accent will be downstepped. Thus, if the sequence is H*+L H*, the second H* is downstepped. If the sequence is H* L*+H, the next accent will be downstepped. Ladd [Lad83], suggests that downstep be treated as a binary feature [+/- downstep] which reflects a choice by the speaker to downstep accents or not. Beckman and Pierrehumbert [BP86] note a number of flaws with this suggestion, notably that a binary downstep feature makes it possible to represent impossible and non-occurring accent types. They attempt to improve the downstep rule by stating that it is triggered by L+H accents, rather than worrying about non-bitonal/bitonal sequences. Ladd [Lad90] proposes a metrical structure which acts as a constraint on downstep and pitch range. This suggestion allows Ladd to support the idea that intonation is not necessarily locally dependent on preceding tone types while retaining a phonological process to explain downstep.

In Beckman and Pierrehumbert [BP86], which applies the Pierrehumbert model to a comparative study of English and Japanese intonation, the basic nature of Pierrehumbert's system remained intact. However, some structural changes played a large part in improving the model. First, the expansion of the model for use with Japanese gave the model support as being more than an *ad hoc* description of American English. The Japanese data also brought about the inclusion of intermediate phrases in English. The intermediate phrase (*ip*) structure was an important addition to the model. The *ip* is bounded by the phrase tone (L^-, H^-), giving these units a purpose, rather than a mere existence. The *ip* was also identified as the domain for downstep. Such a domain restriction was seen to improve the ability of the model to handle pitch range resetting. Accounting for downward trends in intonation was therefore more useful, as the return to a higher pitch range was also accounted for. The use of the *ip* as a constituent in the intonational hierarchy has been further supported by work involving Bengali intonation [HL91].

The final piece of the evolution of Pierrehumbert-based intonation modelling, in terms of moving a theoretical model towards successful applications, is the ToBI [SBP+92] intonation transcription system. ToBI (**To**nes and **B**reak **I**ndices) is a system used to transcribe prosodic phrasing (Break Indices) and intonation. ToBI uses the basic Pierrehumbert tonal inventory, except that downstepped H tones are explicitly marked as !H, which allows the accent to be accounted for by an automatic phonological process, or not, as individual researchers wish. This system is one of the most widely used intonation transcription systems in the world, and has been adapted for a number of languages and dialects with moderate success.

3.2.3 Applications of the AM approach

The ToBI labelling convention is currently one of the most popular prosodic annotation systems. As a result, it has been used in developing intonation models for a wider range of uses than some of the other models. Ross and Ostendorf [Ros94], [OR97] have used a modification of ToBI in developing stochastic models for both intonation synthesis and prediction of intonation labels. Black and Hunt ([BH96]) use a similar modified-ToBI system to build a stochastic intonation synthesis model. All of these systems collapse the ToBI tone inventory into the simpler inventory of H*, !H*, and L*.

- ***F0 Synthesis using ToBI***

Ross's intonation synthesis system ([Ros94], [RO94]) predicts the location and tone type of pitch accents for each syllable in an utterance using classification trees and Markov sequences. The level of prominence is predicted using regression trees. The F0 is generated from these predicted accents using a dynamic system ([Dig92]) which models (and predicts) F0 and RMS energy jointly [RO94].

Classification and regression trees [BFO84] are fairly common in stochastic modelling of intonation, as are Markov sequences [RJ93]. Decision trees (which include both classification and regression trees) are typically binary decision trees. Tree nodes are typically questions which have "yes" and "no" branches, terminating in a class identifier or a numeric value. The trees are used to make predictions for feature vectors based on the distribution of similar vectors in a training set. Markov sequences are probabilistic models where each state in the sequence has a probability of remaining in that state and a probability of moving out of that state.

The dynamic system that Ross and Ostendorf use is a hybrid model. The classification trees in this model are used to provide probability distributions for use by the Markov sequences. Ross ([Ros94]) reasons that intonation is, in some important ways, a series of inter-related events. The lack of independence among tones in the Autosegmental-Metrical model creates, according to Ross, a difficulty for the classification tree. Because the timing of a pitch accent, for example, is affected by the proximity of other pitch accents ("accent clash"), the tree's feature vector is an inadequate description of all relevant information. He argues that a sequence of related events can be modelled better by incorporating a sequential model. For each utterance, the tone sequence is predicted by Markov sequences. Regression trees then predict the prominence (encoded in normalized peak energy and F0) of each pitch accent.

The statistical models are trained and tested on a portion of the Boston University Radio News Corpus [OPSH95]. The label prediction model assigns intonation labels from a collapsed ToBI inventory to each syllable in an utterance. The inventory is collapsed to take account of the nature of the database. Some tone types - L*+H, L*+!H, %H, H+!H*, and X*? (undecided accent by the labeller) - are rarely seen in the database, and are combined with similar tone types. Further labeller incertainty led Ross to collapse the accent tone inventory to four: unaccented, high, low, and downstepped. The boundary tones are divided into three categories: L-L%, L-H%, and H-L%. These labels are predicted for each syllable using information about the syllable, the word of which the syllable is a part, the larger prosodic phrase, the paragraph, and nearby intonation labels.

The syllable information that Ross uses includes lexical stress, from a dictionary, and vowel type (tense/lax). The word information includes part-of-speech, content/function classes, neighboring part-of-speech labels, whether the word is a part of a complex nominal, and a given/new distinction. The length of the prosodic phrase and the position within the phrase of the syllable are also used in the tone prediction model. For the boundary tone prediction, information about the location of the prosodic phrase within a sentence, the phrase and sentence within a paragraph, sentence length, and punctuation are also used. Finally, the prediction models include information about surrounding intonation labels: preceding label, number of unaccented syllables prior to the syllable, and other tones on the word. The regression trees which predict peak prominence use subsets of the above features, selected using clustering experiments. The accent prominence tree, for example, uses information about the tone label, the number of accents within the phrase, the position of the phrase in a sentence, and the previous accent label. The prominence values are used by the dynamical system to generate fundamental frequency and energy contours.

Black and Hunt ([BH96]) produced a similar, if less complex, system to Ross and Ostendorf. Unlike Ross ([Ros94]), Black and Hunt are not predicting the location and type of intonation events. They take the event type and location as given, and predict only the F0 contour. Black and Hunt use linear regression to model F0 values for the start, middle, and end of each syllable. Twenty-eight features, as follows, are modelled:

- the ToBI accent type on the syllable and the two preceeding and succeeding syllables
- the ToBI endtone type on the syllable and the two preceeding and succeeding syllables
- the lexical stress on the syllable and the two preceeding and succeeding syllables
- the number of syllables from the previous major phrase break and to the next.

- the number of stressed syllables from the previous major phrase break and to the next.
- the number of accented syllables from the previous major phrase break and to the next.
- the phrase break index (0-4) of the syllable and the two preceeding and succeeding syllables

As discussed in section 3.1, one method of assessing synthetic intonation is to objectively compare intonation contours. Both Ross and Black and Hunt evaluate their synthetic intonation using Root Mean Squared Error. Ross's comparison of the original contours with the generated ones results in a RMSE of 34.7Hz on independent test data [Ros94]. Black and Hunt report RMSE of 34.8Hz on the same database as used Ross [BH96]. In isolation, these values do not carry any weight. For some voices, a 35Hz difference in F0 can be phenomenally bad. For others, it can be barely noticeable. The speaker which these experiments used has a standard deviation in F0 of 42Hz. As section 5.3 discusses further, it is likely that producing a smaller RMSE than the standard deviation will produce generally acceptable intonation contours. This interpretation of the results is supported by a perceptual experiment performed by Ross, where subjects rated his synthetic intonation to be as natural as the original intonation from the database ([Ros94], section 5.3.2).

- ***Intonation Analysis using ToBI***

Ostendorf and Ross [OR97] use what is effectively the reverse of their synthesis model for automatic intonation labelling. The analysis system requires word and sub-word labelling, which, as with their synthesis data, is automatically labelled and hand-corrected. Using the F0, energy, and linguistic context, the system determines whether or not each syllable is accented, and gives accented syllables a tone label. This automatic analysis method resulted in 88% correct labels with 11% insertions. In terms of the evaluation metrics described in Section 4.2.5, these results translate as 88% correct, 77% accuracy, 23% error: where correct is the number of recognized accents which are correct, accuracy is correct minus recognized accents which are incorrect, and error is 100% - accuracy. These results are comparable to the accuracy of human labellers who have the same data available (e.g. syllable, segment, word, and phrase boundaries) as shown in Silverman *et al* [SBP+92] and Taylor [Tay00].

3.2.4 Summary

The drive to create a system of intonation which is related to a system for describing tone has a number of linguistic implications which are outwith the context of this thesis. Debates over whether intonation should, "theoretically," be described in terms of levels or movements are well outside the scope of this thesis. The question here is whether it is possible to model intonation automatically. The ability to describe a language's intonation in terms of two basic units encourages automatic processing. Such a minimal inventory invites systematic treatment and categorization of intonation units.

However, the reality is that systematic treatment of English intonation as a two "tone" system is not a simple matter. The ToBI system has been widely criticized, even from within the AM school (e.g. [Lad96]), for unnecessary complexity and an inability to cope with natural language use. The complex issue of whether the bitonal inventory is adequately described, or even needed, illustrates that the power of the AM models needs constraining. Because a large portion of all realized accents are simple or downstepped H^*, much work in using ToBI has abandoned over half of the possible tone inventory (e.g. [Ros94]).

The autosegmental/metrical school of intonation modelling has contributed heavily to the current state of intonation theory. The ability to break intonation contours into pitch accent units has brought intonation modelling to a stage where automatic synthesis and analysis are computationally viable. ToBI has allowed for large-scale intonation data creation, making stochastic models for synthesis ([BH96] [RO94] [Ros94]) and analysis ([OR97]) a reality.

3.3 The IPO Modelling Method

Like the AM model, the IPO model represents intonation in a series of discrete events. However, where the AM model uses an inventory of tones, the IPO model inventory consists of pitch movements. The basic premise of the IPO model is that in modelling intonation, one only need model those pitch movements which are perceptually relevant to intonation. The pitch movements which are deemed relevant to intonation are those which have been intentionally produced by the speaker. Pitch movements are seen to have some psychophysical properties (e.g. F0 of over 40Hz, as in [tHCC90]), and may be approximated as linear changes in the (log)F0/time domain ([dP83]). Based on experimental data, a basic inventory of rises and falls is created for a language (e.g. five rises and five falls for Dutch). The basic model and methodology were developed for Dutch (e.g. [tHCC90]). However, attempts to apply the model to British English [dP83] and French [Bea94] have given support to the wide applicability of the model.

3.3.1 Contour Stylization

IPO has invested a great deal of time into the study of approximating fundamental frequency with straight lines (in the log domain). The process of straight-line stylization presumes that the use of linear sequences for approximating F0 does not detract from the perceived pitch quality [tH91].

The method of creating "perceptually equivalent" approximations, or "close-copy stylizations" is interactive [tHCC90]. A listener will replace a small section of a pitch contour with a straight line, and then resynthesize the utterance. This process iterates through the contour until the minimum number of straight lines is used to approximate F0, while the perceived pitch quality has not changed.

DePijper [dP83] tested a small database of British English close copies against original pitch traces, using 64 native speakers of British English to judge the equivalence. Utterances which were synthesized with the copy were mistaken for utterances synthesized with the original F0 in over 85% of cases. Thus, IPO are able to support the use of the close-copy stylization as a tool for approximating fundamental frequency.

3.3.2 Pitch Movement Inventory

The basic unit of perceptual analysis in the IPO model is the pitch movement. The model describes movements according to direction, timing in relation to syllable boundaries, rate of change, and size. The IPO model limits the movement inventory according to the perceptual differentiation among rises and falls.

Using the stylization method described above, experimental data is examined to determine the smallest possible inventory which may be used to adequately account for all relevant pitch movements. Table 3.2 shows a feature description of the 10 pitch movements used to describe Dutch intonation [tHCC90].

	1	2	3	4	5	A	B	C	D	E
rise	+	+	+	+	+	-	-	-	-	-
early	+	-	-	-	+	-	+	-	-	+
late	-	+	-	+	-	-	-	+	+	-
spread	-	-	-	+	-	-	-	-	+	-
full	+	+	+	+	-	+	+	+	+	-

Table 3.2: Feature description of Dutch pitch movements ('t Hart et al, 1990:153)

The Dutch inventory consists of five rises and five falls, which are further distinguished in timing of the movement in relation to the syllable perceived as accented (early, late, middle), duration (spread, unspread), and height (full, half). These features describe the standardized movements which are used to account for the stylizations, rather than individual pitch movements of individual contours. They are, in effect, standardized approximations of approximated contours. The second level of approximation results in a small number of standardized pitch movements, to which it is possible to give acoustic values.

Full rises and falls take 160ms to move one octave. Half-size elements have the same slope, but take only 80ms to complete their movement. The spread feature is somewhat redundant here, as it relates to the syllabic content of the element (-spread takes place in one syllable, +spread in two or more). The choice of early, middle, and late for the elements corresponds roughly to a position in the accented syllable where the movement begins.

Similar inventories have been created for several languages, British English [dP83], French [Bea94], Russian [Ode89], and German [Adr91].

3.3.3 Configurations and Contours

Once a pitch movement inventory is determined for a language, the next step is to organize the use of the movements in synthetic speech. The IPO model uses a grammar of pitch movements to constrain the possible combinations of movements over a given domain (e.g. the clause).

Two or more pitch movements can combine to form a configuration. The possible configurations are determined by movement sequences in experimental data. For example, because the rise '1' is followed by fall 'B', but never fall 'C', there is a restriction on the number of acceptable combinations or rise '1' and fall elements based on the data being modelled.

Configurations, themselves, are classified as Root, Prefix, and Suffix configurations. Using this classification, a contour may then be made following Equation 3.1.

$$\text{Contour} \rightarrow (\text{P})^n\text{R}(\text{S}) \tag{3.1}$$

Thus, a contour is made up of one Root configuration (e.g. 3C: a full, middle, fast rise followed by a full, late, fast fall), an optional Suffix configuration (e.g. 2), and an optional Prefix configuration, which may or may not be nested. In terms of the British School [OA61], the contour consists of any number of pre-heads of the same shape, a nucleus, and an optional tail. The full grammar for Dutch is detailed in [tHCC90].

The contour grammar for a language is used by a generative rule set to arrive at a contour for individual utterances. Rules determine accent placement, prosodic phrasing, de-accentuation, and similar functions that are typical of an intonation generation system.

3.3.4 Applications Using the IPO Approach

The IPO methodology has yielded intonation models for a number of languages. De Pijper [dP83] developed a model for English, Beaugendre [Bea94] has developed an IPO-style model for French synthesis, and Mertens *et al* [MBd97] have recently begun work on an automatic F0 stylization system for French. Unlike work in the AM school, IPO modelling does not require a great deal of automatic analysis of the stylized contour. Once a contour has been stylized according to the IPO methodology, the pitch movements are categorized statistically, a process which is easily automated anyway.

The difficulty in automating the IPO method is in the stylization process. The stylization process has been justified based on its interactive nature, which allows listeners to determine how much a section of contour may be stylized. Without this interactive perceptual testing, it is difficult to maintain the "perceptual equivalence" between the natural contour the stylized contour. Therefore, if Mertens *et al* are able to automate stylization, the IPO method will become more usable.

- ***F0 Synthesis Following the IPO Approach***

French intonation is widely recognized as having two accent types (c.f. [AEFN97], [Bea94]), a primary and an optional secondary accent. The primary accent is located on the final syllable of an accent group (i.e. accent groups are right-headed). The secondary accent is similar to the English pitch accent, in that it is used to signal intentional prominence, and can fall anywhere in a phrase. Beaugendre uses a number of accentuation and de-accentuation rules to assign locations and pitch movements to the two accent types. The pitch movements, augmented by a number of connection rules, dictate the resulting F0 contour.

Beaugendre's pitch movements are classified by four features: direction, amplitude, duration, and syllable content. Unlike other intonation models, Beaugendre includes syllable content as a feature of his pitch accents. The first three rise elements (R1-R3), for example, are distinguished not only by the pitch movement. They are specifically used only on syllables of the type mentioned (e.g. R2 is only used with syllabic consonants). The ten basic pitch movements are described as follows:

R1 CV syllable, rises above the top-line, very steep

R2 C, rises to the top-line, very steep

R3 CV, rises to the top-line, steep

R4 V+CV, rises to the top-line, shallow, starts midway between top and baseline

R4+ steep version of R4

R5 C, rises to mid-line, steep

F1 gradual fall to baseline covering several syllables

F2 steep fall to baseline

F3 short, steep fall from baseline even further down

D flattish connection.

Using a combination of the seventeen accentuation, de-accentuation, and movement selection rules, the system assigns one of the ten pitch movements to each accent location. The accent locations are assigned as the system iterates through the words of the text to be synthesized. The de-accentuation rules then deal with cases of adjacent accents, removing one if they are in the same accent group. Finally, a series of rules fill in connections. The resulting F0 contours were judged by a group of native speakers as being "qualitatively similar" to the originals. However, as the goal of IPO is to provide perceptually adequate synthesis, objective evaluations of the synthetic intonation was not provided.

- ***F0 Analysis Following the IPO Approach***

Mertens *et al* [MBd97] has attempted to automate the IPO stylization process in order to make the IPO methodology more accessible for widespread use. The system takes a smoothed F0 and, using zero-crossing and energy data, segments the speech into voiced and unvoiced sections. The voiced segments are then processed to produce the stylization.

For each voiced section (typically syllable-sized), the pitch contour is divided into tonal segments. Each tonal segment is classified as a single pitch event (rise, fall, level). The tonal segment is characterized by the time and pitch at its starting and ending points. The tonal segments are classified in reference to thresholds for slope and amplitude of pitch movement.

The resynthesized contours were judged by 10 native French speakers to be "perceptually equivalent" [dP83] to both the manually stylized contours and the originals. As with other IPO evaluations of synthetic intonation, no objective measurement of the output was provided.

3.3.5 Summary

The IPO method represents intonation in a series of discrete pitch movements. IPO models are based on experimental data, from which a basic, standardized inventory of rises and falls is extracted. A model is then used in the automatic analysis and synthesis of intonation.

The IPO intonation modelling methodology is one that appears successful in its aim to adequately model macro-intonation. The rectilinear approximation techniques used to create close-copy stylizations, and the tests which support their adequacy, have opened the way for a number of stylization techniques used in more recent work (e.g. [CFHV97]).

Because of the standardized pitch movement inventory, the IPO method has been criticized for inflexibility. It has been lumped into a category of concrete inventory systems (e.g. Pike's 4 tone system [LP84]) on the grounds that it uses a strict declination system, and pitch movements occur only in certain shapes and sizes, the idea being that such a system cannot be used to account for downstep. However, no proof of this claim has been offered. Taylor [Tay92] contends that more than two layers of downstep cannot be reconstructed from an utterance, based on use of the standardization procedure described above. However, no evidence of this difficulty exists. The ability of the Dutch grammar, for instance, to produce contours of half rise, full fall type allow for a small peak followed by a steep fall, which is the pattern of multiply downstepped utterances. It is also possible for there to be no rise at all in the contour. The variability of concatenative possibilities allows the IPO system to produce peaks which are scaled down relative to previous peaks, and other factors of downtrends.

The problem which poses the greatest problem for IPO models is that they are meant to be useful for both analysis and synthesis, which has proven a difficult task. The interaction between human and computer during the stylization process requires long hours of intensive listening to and resynthesis of speech in order to arrive at a robust account of a language's pitch movements.

As Taylor [Tay92] notes, lack of formality in the creation of data for experimentation places doubt on the adequacy of IPO models. However, Mertens *et al* [MBd97] have begun to address this problem with a perceptually grounded automatic stylizer for French. While the initial results show that the result of resynthesis is a good close-copy stylization, the automatic stylizations contain far more straight line segments than the manual method. Because of the larger inventory of pitch movements, which are not as easily categorized as the hand-labelled inventory, the output requires a change in the grammar in order to be integrated into the rest of the IPO methodology.

3.4 Superpositional Intonation Modelling

Where the previous models have consisted of linearly ordered tones or pitch movements, the superpositional (SP) models are the result of interactive intonation "commands" [Fuj83], which represent syllables, accent or stress groups, phrases, and increasingly larger units. These commands are combined so that each level of the hierarchy is represented in the generated F0 contour. Fujisaki's model has been used to model Japanese, and to some extent English, intonation. Variations of his model have been used to model German [Möb95] and Danish [Grø95].

The superpositional model stems from two sources. Fujisaki's model is based on the peak and decay of F0 as produced by the glottal source [Fuj83]. Coupled with the desire to model the production mechanism is the need to account for evidence of intonation being the result of local and non-local factors. However, phenomena such as downstep, which suggest that intonational planning takes place over a large domain (e.g. Gønnum's textual domain) are difficult for the Fujisaki model and its variations to replicate [Tay92], [LP84].

3.4.1 Commands

Each command consists of a peak (or valley) followed by an exponential decay. The style of peak differs from model to model. In Fujisaki's model, the peak is the output of a filter which has been excited by a rectangular command input (as shown in Figure 3.2). Grønnum dispenses with the rectangle and simply provides peak and decay parameters.

The larger the command domain, the smaller the command peak. Thus, a textual unit [Grø95] which encompasses multiple utterances has a general decay, where the beginning frequency is approximately half an octave higher than the end. Grønnum's utterance command begins with a minimal peak, and then decays. Below this domain, the various superpositional models are very similar. The phrase command, which is the largest unit of Fujisaki's model, has a large peak relative to the larger domains of Grønnum's model. This command basically sets the intonation register for the phrase. The phrase command domain is roughly comparable to the Contour domain of the IPO models and the Intermediate Phrase from the AM models. The accent command, which is comparable to IPO's pitch movement and the AM pitch accent, is a sharp peak with a short decay time, representing F0 movement over an intonation event. Figure 3.2 shows how commands can combine to form an intonation contour.

3.4.2 Applications Using the Superpositional Approach

The superpositional modelling technique is the basis of the intonation models for a number of languages in the Bell Labs multilingual TTS synthesis system. [vSSM98] The model used by Bell Labs uses commands (curves) for minor phrases, accents, and sonorant-obstruent transitions. The system is somewhat of a hybrid with a simplified ToBI system, in that it predicts command parameters based on a predicted tone from a collapsed ToBI tone inventory. The curves are added together in the logarithmic F0 domain to result in a fundamental frequency contour.

Minor phrase curves are the result of interpolation between the beginning of the phrase and the beginning of the final accent group in the phrase and the start and end of the final accent group in the phrase. The F0 is computed by rule based on the location of the minor phrase within a major phrase. While this method of deriving the phrase curve parameters is rather simplistic, its designers admit this and intend to improve this module in future research.

The accent curves are defined with time-warped templates. The templates are basic command parameters based on the location within the phrase of the accent, and whether the accent group contains a question or continuation rise. The time-warping of the template is the result of using a number of peak alignment rules to alter the pre-defined peak location. The size of the prominence is determined by the location of the accent group in the phrase, the prominence (which is determined by rule based on accent location within

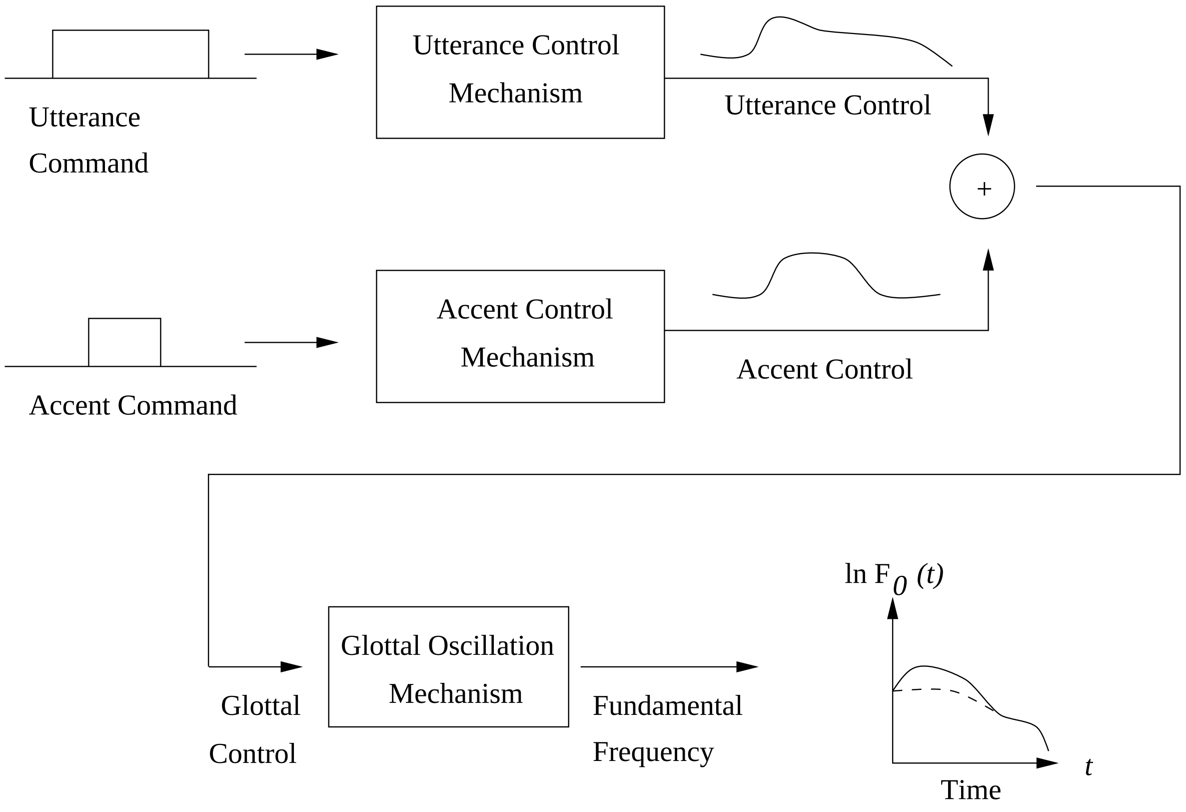


Figure 3.2: A functional model for generating F0 contours using a superpositional model. ([Fuj83]:42)

an accent group and the tone type being predicted), and the duration of the accent group.

Segmental effects on F0 are modelled by adding a small curve at consonant/vowel junctions. Where the consonant is sonorant, the curve is flat. The height of the curve rises in inverse proportion to the sonority of the consonant.

3.4.3 Summary

Taylor [Tay92] adapted the Fujisaki model for English in order to determine how well the model was able to reproduce intonation contours. While no formal results were mentioned, he asserts that a superpositional approach successfully reproduces neutral declarative utterances, but encounters difficulty with a wider representation of language. Other superpositional models have shown more promise than Fujisaki's original model. Bell Labs [Spr98] uses a similar SP model in its multilingual synthesis system. The SP approach is promising, but questions remain about its viability in a stochastic modelling domain, rather than a rule-based system.

3.5 Continuous Parameterized Models

The continuous parameterized (CP) models attempt a slightly different manner of describing F0. Taylor's Tilt model [Tay00] and Portele's prominence-based description [Por97] describe the F0 of intonation 'events.' Like the other models, the CP models generally ignore the detail of F0 in between events. Campione *et al* [CFHV97] use a different method of describing the F0 in a continuous manner: they judge target points in relation to a speaker-specific pitch range and in relation to adjacent target points, creating a continuous description of the upward and downward movement of F0.

In comparison to the previous methods of intonation modelling, the Tilt and prominence-based description (PBD) are something of a hybrid of the superpositional and IPO modelling techniques. Like the superpositional models, these two models attempt a low-level analysis of the F0. There is no standardization technique which removes information from the already-stylized F0. Both Tilt and PBD - which are quite similar to each other - exploit the IPO claims that small pitch movements are not particularly relevant to intonation perception. By modelling only intonation events (i.e. prominences, or accents and boundaries), these methods expect that interpolation of the remaining F0 contour will not lead to a degradation of intonation quality. Tilt and PBD also parallel the IPO models in that they treat both the rise portion and the fall portion of a pitch prominence as important aspects of the prominence. This differs somewhat from ToBI, where the primary emphasis is on tone, with some influence of alignment. ToBI expects that the pitch movement portions of a prominence are predictable, and therefore unimportant in the description of F0.

The INTSINT (INternational Transcription System for INTonation) model (e.g. [CFHV97]) is of an even lower phonetic level than the other two CP models. Target points are established at each point where the F0 changes direction and are described by their relationship to each other and the speaker's pitch range. This results in a serial description of the entire contour, rather than a description of the so-called 'important' segments.

Each of these models is based on the idea that a phonetic description of F0 is necessary for accurate intonation modelling. The models also have varying degrees of higher-level information, as may be noted in the discussions of the individual models.

3.5.1 Tilt

In an effort to simplify intonation analysis, Taylor [Tay00] presents a system which is based solely on the acoustic details of the speech stream. Intonation contours are divided into "phonetic phrases" delimited by silence. Rather than posit a set of phonological features or a large, complex tonal inventory, Taylor evaluates fundamental frequency contours based solely upon their shape.

In the Tilt model, there are four basic intonational units: pitch accents, boundary tones, connections, and silence. These four units can be divided into two classes. Pitch accents and boundary tones (intonation events) form one class, while connections and silences form the other. Intonation events are described by five parameters. Connections and silence by one. Pitch accents and boundary tones each contain two distinct parts, the rise and the fall. Every event contains a description for both the rise and the fall. Often, one or the other of these portions is described with parameters of zero, such that they are simple rises or simple falls. The description is the result of functions of the F0 movement over time.

Each of the descriptions includes the fundamental frequency at the start of the unit. Pitch accents and boundary tones are also described by their duration, their absolute amplitude (the sum of F0 movement over the event), the position at which the rising portion of the event stops and the fall begins (peak position), and the tilt value. Such a description would appear in the description file as in Table 3.3. Figure 3.3 shows how the parameters relate to an example pitch accent.

End Time	Event Type	Start F0	Absolute Amplitude	Duration	Tilt	Peak Pos.
1.253	a	206.5	35.2	0.302	0.012	0.124

Table 3.3: Example Tilt description

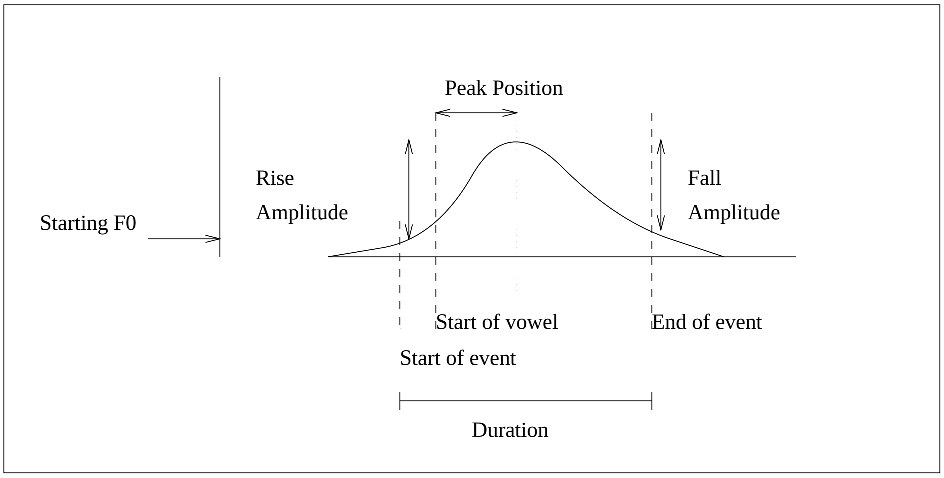


Figure 3.3: Tilt parameters

The starting F0 of an event is the anchor point for all parameters in the frequency domain. The amount of rise (in Hertz) from the starting F0 to the peak is the first portion of the absolute amplitude parameter. The second portion is the amount of fall from the peak to the end of the event. Either of these portions may be zero, if the event is a simple rise or simple fall. The two amplitude values are added together to form the absolute amplitude value.

It is obvious from the calculation of absolute amplitude that one must know the point at which the rise becomes a fall. This point is called the peak position. The peak position value is used not only to compute the absolute amplitude, but it is necessary for the computation of tilt as well. Depending on the intended use of a model, the peak position is described in terms of absolute time (i.e. the number of seconds from the start of the speech signal)

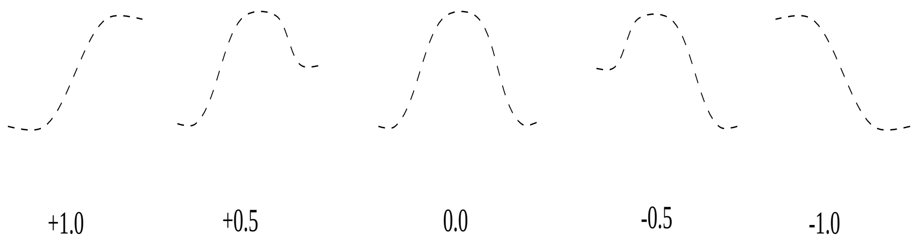


Figure 3.4: An Illustration of *Tilt* Parameter Values

or in relative time (e.g the number of seconds from the start of the accent, or from the start of the accented syllable).

Tilt is described, as seen in equation (3.2), as the difference of the amplitude portions divided by their sum [DB97]. Tilt has a range of -1 to 1, where -1 is pure fall, 1 is pure rise, and 0 contains equal portions of rise and fall. This continuum is illustrated in Figure 3.4.

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{3.2}$$

The Tilt approach to intonation provides a good phonetic model of F0. As Chapter 5 discusses, the method for audibly assessing the success of an intonation model is to synthesize an utterance with the generated F0 contour (test) and compare it to the same utterance generated using the original F0 contour (reference). Section 5.5 discusses both this subjective evaluation method and the objective assessment metrics noted below, but informal listening tests have shown that for most utterances, the test and reference F0 contours cannot be audibly distinguished. To objectively evaluate the Tilt model, the reference F0 was parameterized, and the test contour was generated from those parameters. For two large American English databases[2] a comparison of the contours resulted in each database averaging a correlation

[2] over 400 sentences spoken by a male, 45 minutes of news broadcast spoken by a female

of over 93% and a RMSE error of less than one-third the standard deviation of the speaker's natural F0 variation.

There is almost no predictive grammar for the Tilt model (as opposed to Pierrehumbert's finite state grammar). The only compulsory ordering is that no two connections should be juxtaposed. This restriction, however, has more to do with the definition of connection than any ordering or constituency constraints.

However, some higher-level information has found its way into the Tilt model. The first "phonological" distinction made in the Tilt model is to differentiate accents from boundaries. Because accent events and boundary events are described in exactly the same fashion, different labels for the two types is a concession to the phonological supposition that the events types differ on a level above phonetic analysis. In addition to the concession to "theory," concessions to users of the model have added more high-level information.

In its original instantiation, the Tilt model considered rising boundaries and accents as events, and categorized falling boundaries (as the default condition before silence) with connections and silences. However, when applying the model to labelling natural speech data, where there are numerous disfluencies to cloud the picture, the inclusion of falling boundaries into the event class became necessary as not all silences in natural speech are a result of a prosodic boundary. This alteration came about, not for model-internal reasons, but due to labeller demand. Labellers found it preferable to have a labelled distinction between the end of speech and a marked decline in fundamental frequency which marked an intonation boundary. Not all Tilt applications have included this expansion of the model (e.g. [DB97], [Tay00]).

A final concession to the use of Tilt in labelling natural speech is the inclusion of accent/boundary concatenations [Tay00]. At times, accents and boundaries occur on a syllable in such a way that they blend into each other, rather than having distinct borders (e.g. a rise-fall accent which ends with a low falling boundary tone). In the ToBI system, such a phenomenon would not be difficult to label, as the accent labels are placed at peaks and boundary labels at the end of a pitch movement. The IPO models account for such an occurrence as a typical combination of configurations. Similarly, the combination label (e.g. afb - accent + falling boundary) has been included into the Tilt model. However, one must remember that, as an event type, the combination events are described in the same manner as all other events. Therefore, phonetically, the model has remained unchanged. The "phonological" inventory has merely gained a few allophones.

3.5.2 Prominence-Based Description

Portele [Por97] presents a model very similar to the Tilt model. He describes pitch peaks in the following terms:

- peak position relative to syllable nucleus onset
- peak height relative to speaker-dependent top- and baseline
- slope of rise and fall portions of peak.

Because the peak is the anchor for both the frequency and the time domains, the prominence-based description (PBD) dispenses with the starting F0 parameter used in the Tilt model. The duration parameter from Tilt is also unnecessary in this model, as the timing information is a portion of the slope calculations. Like the Tilt model, resynthesis using the PBD results in no appreciable difference in intonation.

3.5.3 INTSINT

The INTSINT model (e.g. [CFHV97]) is a symbolic coding system which provides a low-level account of F0 movements. Target points are measured every 30ms in reference to the speaker-dependent pitch range and each other. Each target is then coded as being at the top, mean, or bottom of the range or, if none of these "anchor points," as being higher, lower, or the same as the points around it. The result is a sequence of codes which represent the heights of the target points.

A number of calculations are necessary for the translation of F0 to the coded sequence. Thresholds are set for the top (T) and bottom (B) codes such that 5% of points must be coded as T and 5% as B (assuming a normal distribution). The frequency of occurance of T codes is then calculated as the mean value of the T and B targets beyond the thresholds. The frequency range between the top and bottom thresholds is then divided into three, such that each band contains 30% of the remaining target points. The higher and lower (H and L) codes are then given values calculated by a regression of the target points in the relevant band. In essence, this systematically normalizes target points based on their relationship to a normal distribution of their frequencies combined with the relationship among adjacent points.

Resynthesis based on the normalized coding resulted in contours which were close to, but distinguishable from, the original curves. However, only in the cases of extreme distance above or below the relevant top and bottom threshold showed any perceptible differences. The total average variance from the original curves was less than 0.1, which is a promising result.

Campione *et al* treat the INTSINT model as a first level of modelling an intonation contour. The coded sequence is not meant to provide high-level information, but to act as a reflection of the acoustic realities of speech data, which may then be processed to provide a second-level model such as Tilt or the PBD, or even a purely "phonological" model like ToBI. Therefore, while INTSINT may be useful in creating intonation models, it cannot be considered in the domain of this thesis as a such a model in itself.

3.5.4 Applications Using the CP Approach

The Tilt intonation model [Tay00] is one of the intonation modules being used for F0 generation in the Festival Text-to-Speech System [BTC98], and is the object of recent research in automatic intonation analysis [Tay00], [Dus98]. Unlike the ToBI and IPO models, Tilt is a recent development, and the systems described are both new research and functional implementations.

- ***Synthesis***

Dusterhoff and Black [DB97] have developed a model for predicting the individual Tilt model parameters using regression trees [BFO84], based on the approach taken in Black and Hunt ([BH96]) above. They performed a number of experiments, using Tilt as an intermediate between ToBI labels and F0 as well as generating F0 straight from Tilt descriptions. The first experiments, using Tilt as an intermediate between ToBI and F0 included questions about the ToBI labelling in the regression trees. The latter experiments did not include any information about the ToBI labels in the feature database. The experiments which did not use ToBI label information produced the best results, which are reported below. The regression trees predict each parameter (starting F0, amplitude, duration, tilt, and peak position) for each intonation event. The experimental methodology is explained in detail in Chapter 5, as Dusterhoff and Black [DB97] forms the basis for the work in that chapter.

The experiments were undertaken using the same database as was used by Ross [Ros94] and Black and Hunt [BH96]. Dusterhoff and Black use the same contextual features listed above from Black and Hunt. Additionally, they include features related to the syllable content - sonority of onset and coda and duration of onset and rhyme.

Comparisons between the generated and original F0 contours show that the Tilt model is at least as effective for intonation synthesis as the various ToBI models on the same database. Table 3.4 shows the results of Dusterhoff and Black in reference to the studies discussed in section 3.2.3.

	Dusterhoff & Black	Black & Hunt	Ross & Ostendorf
RMSE	32.5Hz	34.8Hz	33Hz
Correlation	0.60	0.62	Not Given

Table 3.4: Comparison among Tilt and ToBI F0 generation methods

Having shown adequate results, and generating contours which, informally, have been deemed acceptable to native English speakers, this modelling technique has been included into the widely-distributed Festival system while continuing research takes place.

- ***Analysis***

Taylor ([Tay00]) trains a set of Hidden Markov Models (one for each Tilt intonation event type) to detect accents, boundaries, connections and silences.

He trains the models using fundamental frequency and RMS energy. He examines various combinations of normalized F0 and energy along with the first and second derivatives of each feature. He achieves his best results by normalizing both F0 and energy and including both derivatives. All of the tests were constrained by a bigram/unigram grammar which was built from the training corpus.

Taylor also tests the inclusion of a new event type in these experiments. A "minor" pitch accent category was included in the label inventory, and is defined as a pitch movement which a labeller believes might be an accent. A portion of Taylor's research looks into whether labeller uncertainty can be quantified by use of the minor event.

Using the best combination of data types on a speaker-independent dataset, Taylor achieves detection results of 72.7% correct and 47.7% accurate if the minor event label is considered. Without the minor label, the results are 81.9% correct and 60.7% accuracy. The auto-labelling results are not dissimilar to inter-labeller results for humans (81.6% and 60.4% with minor, 88.6% and 74.8% without) though there is a notable insertion difference.

3.5.5 Summary

The continuous parameterized models all attempt to represent the acoustic correlates of intonation. The INTSINT model is a low-level representation of F0 movement, without the trappings of pitch accent or other "phonological" classifications. It would be interpreted by a further modelling system to determine the location and quality of such classes as pitch accents or boundary accents.

The Tilt and PBD models are a step above the INTSINT model in that they only attempt to model the "important" F0 movements in speech. The theoretical assumption that these models rely on is that only some F0 movements are necessary for an accurate reflection of intonation. They assume the location of such important movements is either already known (e.g. through the use of a system which interprets an INTSINT-like sequence) or may be found with relative ease. Tilt and PBD contain varying degrees of "phonological" information, as befits their respective levels of use. Tilt is currently used as a full intonation labelling system, and requires at least enough information to allow accurate use. PBD is being used to study perceptual classification of tonal prominence, and currently only contains the location of "important" prominences.

While each of the models have uses on different levels, all maintain the ability to be used for accurate analysis and regeneration of fundamental frequency contours. This quality is of great importance for automatically building models of intonation for speech synthesis from natural or near-natural data.

3.6 Discussion

This chapter has discussed some of the prominent theories and models of intonation. The four basic categories have resulted from the different aims of intonation researchers, from theoretical linguistic concerns to acoustic-phonetic description. While no claims are made within this thesis as to a value ranking of models, there are some which fit the aim of this research better than others.

The AM model has been used for tasks in automatic intonation synthesis and labelling [Ros94], [OR97]. However, this work approached automatic labelling from the view that all other linguistic data was known (i.e syllable, phrase, prosodic boundaries already existed). While this task is valid, it does not represent the research of interest in this thesis. The AM model contains little or no acoustic (or phonetic) information. As the AM model assumes the location within speech of various high-level linguistic forms, it is unlikely that it will be useful for the differentiation of intonationally important pitch phenomena from unimportant ones.

The IPO methodology is more relevant to the task at hand, acoustically, than the AM model. The standardized pitch movements are useful for both analysis and synthesis. IPO provides psychoacoustic parameters for pitch movements relevant to intonation study. The problems with this modelling method is, as noted above, the lack of formality in the determination of a stylized F0. The IPO method, parameters, and pitch movement inventory all rely on an impressionistic interpretation of intonation. Such a basis is not a good grounding for automatic intonation analysis.

The SP models are purely acoustic in nature. Their mathmetical formality and nature are ideal for building a formal intonation model. The models themselves are reported to successfully reproduce fundamental frequency. This approach may be useful for the task of automatically building synthesis models. At present, the main difficulty of the SP approach is dividing the various pitch movements by their causes. Möbius [Möb95] maintains that the division of cause and effect is one of the advantages of SP models.

The most promising model class for both automatic synthesis and analysis in an acoustic-phonetic domain is the continuous parameterized class. These models are acoustic-phonetic representations of F0. While these models generally function on some sort of stylization, the Tilt model, in particular, has been shown to function equally well on high-quality, unstylized F0 traces. As noted above, the INTSINT model is not appropriate for developing the synthesis models desired, as it requires further interpretation to distinguish intonation events from non-event portions of the contour. The Tilt model has two advantages over the other acoustic-phonetic CP models mentioned here. First, the Tilt parameters represent salient areas of intonation. Modelling success or failure of any of the parameters can be related to specific aspects of F0 movements which are important both experimentally and theoretically. Second, all of the tools which are used in conjunction with the Tilt model were easily and readily available at the outset of this research.

Chapter 4

Sub-syllable Acoustics in Automatic Intonation Analysis

The goal of the intonation analysis research detailed in this chapter is to create a system which can automatically label speech with intonation information. As shown in Figure 1.1, the work in this chapter was designed to assist in data collection for the intonation synthesis research which is discussed in Chapter 5. As Chapter 2 discussed, the relationship between segments and intonation can be exploited in intonation processing applications. An investigation into one way of exploiting this relationship is presented in the experiments below. These experiments are designed to examine what acoustic correlates of segments can be used to improve an existing acoustic modelling method.

The chapter begins with an outline of intonation analysis. This background discussion is followed by a description of experiments that examine what sort of information may be used in the creation of such a system. The system created by Taylor ([Tay00]) acts as the basis of the system which is presented in this chapter. Taylor's system uses Hidden Markov Models to model intonation events based on fundamental frequency and RMS energy data. This chapter presents experiments designed to show whether it is possible to improve Taylor's system by augmenting F0 and RMS energy with other acoustic data. The experiments expand the acoustic data used for intonation analysis to include information about the segmental make-up of the speech, such as zero-crossing and cepstral coefficient data.

4.1 Intonation Analysis

Intonation analysis generally involves three basic tasks: event detection, event identification, and event-syllable association. Detection is the process of finding intonation events. Identification is the process of naming the detected events. In the Tilt model, for example, identification involves determining whether an event is an accent, a boundary, or perhaps a combination of both. Using the ToBI model, the process involves not only determining whether the event is an accent or boundary, but what the tones are that make up the event. The third task, association, is the act of linking an event with a portion of linguistic text (e.g. syllable nucleus, demi-syllable, syllable, word, or phrase). The choice of linguistic constituent is somewhat arbitrary. For example, it may make sense to associate accents with syllables, given the roles of lexical stress and metricality in much of the prosodic literature. However, associating boundaries to a prosodic unit above the syllable may be an interesting way of investigating parallels in phenomena such as final lowering and final lengthening.

This chapter is primarily concerned with event detection. Event identification is secondary, in that all intonation event types are treated simply as events in the detection evaluation. However, the model-building process involves first building models of individual event types, and then using all of the individual models to detect events in novel speech. This division of event types parallels the approach Ross chose ([Ros94]) in order to minimize the size of the training database. Because the events are already grouped into broad classes, the data need only be used for making the finer distinctions (e.g. early versus late peak), rather than both gross and fine distinctions.

While the detection process utilizes models of specific event types, the detection evaluation counts all different event types as being simply events, and therefore equivalent. This notion of equivalence relies on the nature of the Tilt model, where each event is described in exactly the same format. Details of this evaluation technique are discussed later in this chapter, in section 4.2.5.

Once the task is defined, questions about data come into play. What type of speech will be analyzed? What information about that speech will be used for the analysis?

The first question is easily answered: use the type of speech that is being modelled. The basis for this answer lies in the distribution of event types within speech sub-classes. In an application where many questions are asked, training models on news speech is not likely to capture the variation in question intonation forms. While the two types of speech have overlapping distributions of intonation event classes, it is not likely that they will have equivalent distributions. Both speech types will contain large pitch accents, small pitch accents, rising boundaries, and falling boundaries. However, different event types will dominate each speech type. Rising final boundaries, for example, do not often occur in a news broadcast. In fact, only 10% of the F2B database used for the experiments in this chapter received rising boundary labels. The probability of an event type being present in enough contexts, enough times, to build a robust model depends on whether that event type is suitable for frequent use in the type of speech which comprises the database. By building intonation models from the type of speech that the models will be used on, one can capture the distribution of relevant intonation event types with a smaller database than if the database contains multiple or different speech tasks.

The second question is considerably more difficult to answer. Ostendorf and Ross [OR97] have a database which contains information about words, phrases, and syllables. They decide that, as they have all of this information, they should use all of the data that is available. Taylor [Tay00] has the same annotation for his data, but wants to use intonation to improve other speech recognition tasks, so he opts to use only the data which can be derived from the acoustic signal. Because the text is not required *a priori*, Taylor's automatically detected intonation events can be used to assist in deriving the text. Prior intonation analysis can be used for word disambiguation (e.g. [Bar97]), and discourse analysis (e.g. [WT97]) to name but two applications. Both of these approaches to automatic labelling of intonation are valid, and each is suited to its application. The use of wholly acoustic data for intonation labelling avoids the difficulties of acquiring accurate word, segment, and syllable boundaries. Taylor's method looks interesting and open for improvement. Therefore, in an attempt to improve on Taylor's work, the approach to the research described in this chapter follows Taylor in building models only from acoustic information which can be readily extracted from the speech waveform.

4.2 Experimental Methodology

As mentioned at the beginning of this chapter, the purpose of the experiments presented in this chapter is to determine what, if any, acoustic features can be used to augment and improve an automatic intonation analysis system. As section 4.3 discusses, the added acoustic data is chosen because it is associated with segmental interactions with fundamental frequency.

Each experiment is designed to test whether an acoustic property of speech which relates to both segments and intonation can successfully augment Taylor's F0 and energy model. Initially, as section 4.3 will further discuss, Taylor's model was replicated to provide a baseline system. Each experiment thereafter adds to this baseline system. First, a single acoustic feature is added to the fundamental frequency and energy. Auto-correlation peak coefficients, zero-crossing values, and Mel Frequency Cepstral Coefficients are the acoustic features tested. Each of these added features is tested in a number of conditions, as is detailed below. Auto-correlation is tested because it can be used both in broad-scope phoneme classification (e.g. vowel/sonorant consononant/obstruent) and fundamental frequency tracking (e.g. Entropic's get_f0). Similarly, zero-crossing can be used to broadly classify segments into voiced/unvoiced/sibilant categories. Mel Frequency Cepstral Coefficients are regularly used in speech recognition (e.g. [YJO+96]), and are therefore likely to adequately represent segment information. They also provide a representation of the speech spectrum (a cepstrum is a transformation of the spectrum). As section 4.2.4 discusses, the spectrum has been linked experimentally to intonation events.

For each experiment, a number of Hidden Markov Models are built using fundamental frequency, energy, and one or more of the acoustic features listed above. The experiments examine how these features affect the quality of the intonation analysis.

Section 4.2.5 discusses how the output of the intonation analysis is evaluated, so that a notion of success is available. Some experiments also vary the ways in which the acoustic features are combined in the modelling process, to gain an insight into the relative power of the features within the models.

4.2.1 Hidden Markov Models

The manner in which the analysis models are built and used is essentially the same as that employed in word and phone recognition tasks. Figure 4.1 shows a diagram of the process. Hidden Markov Models are trained for all of the speech events which are to be recognized. The difference between phoneme recognition and event detection is that there are fewer than ten possible segments to recognize (accents, rising boundaries, falling boundaries, accent/boundary combinations, silence, and connections). A standard phoneme recognizer would have at least one HMM per phoneme in the language. This intonation recognizer parallels the phoneme recognizer, with one HMM per event type.

The Hidden Markov Models used in these experiments are created using Entropic's Hidden Markov Model Toolkit [YJO+96]. They are trained using Baum-Welch Re-estimation. Figure 4.2 shows a simplified example of a falling boundary HMM. This Hidden Markov Model is essentially a state sequence where each internal state (numbered circle) has a probability that the state will remain active (a) and a probability that the next state will activate (b). The intonation event models are slightly more complicated than this basic sequence model, in that some states may be skipped. As the figure

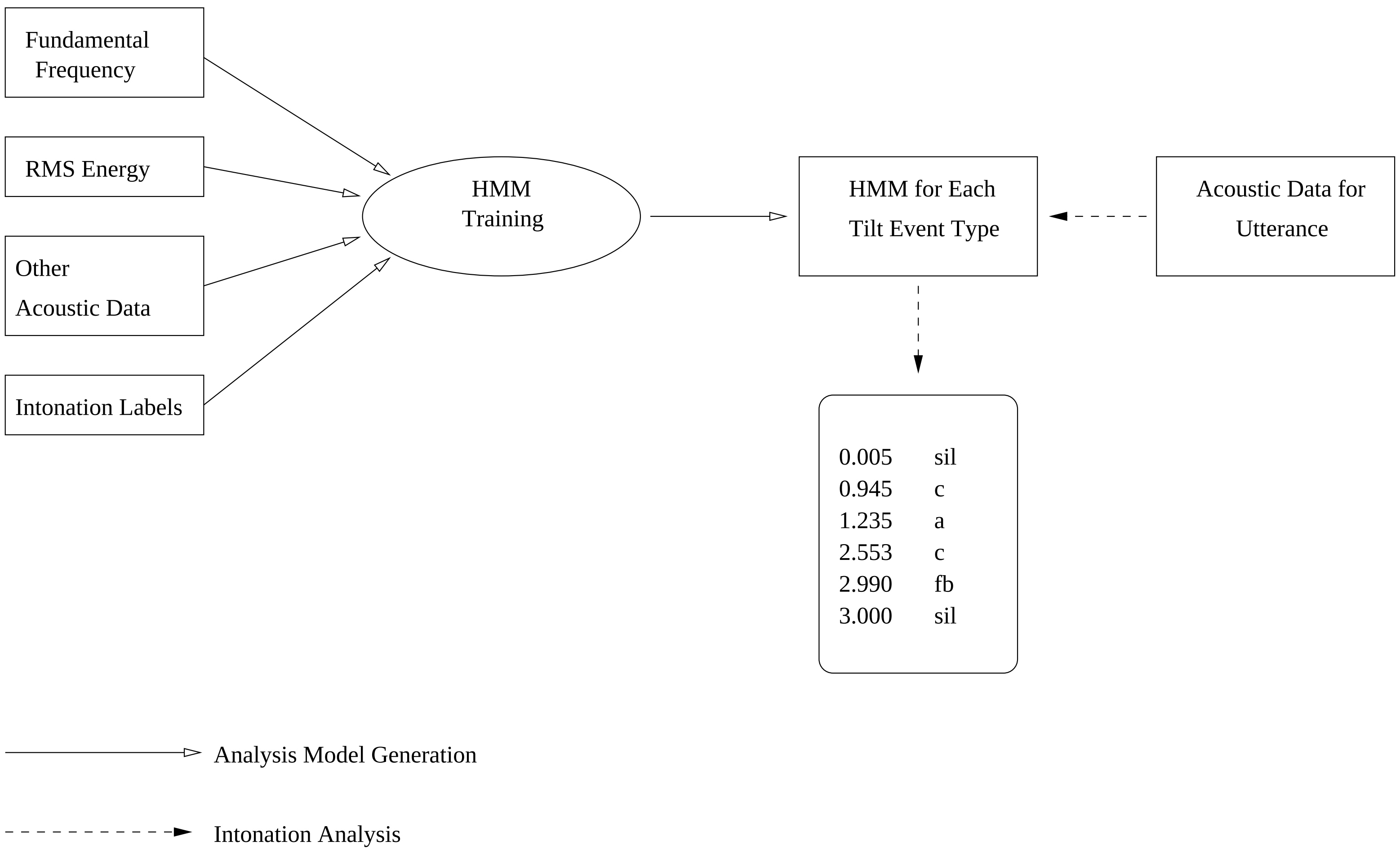


Figure 4.1: A diagram of the intonation analysis process

shows, the first internal state may be skipped in the falling boundary model, which means that, for the entry state, there is a probability (c) that the first state will be skipped and the second state activated.

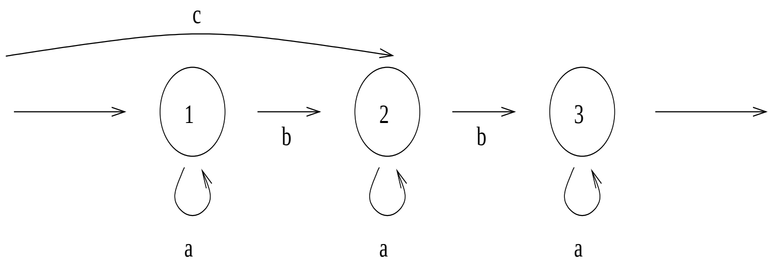


Figure 4.2: Fall event HMM example

Once each event type is represented by an HMM, the quality of the models is tested by using them to produce intonation labels for a set of test utterances. The test utterances are presented as a vector of acoustic data in the same format as the training data. The result is, as shown in Figure 4.1, a series of intonation labels similar to the original intonation labels for the test utterances.

In each case, unless otherwise noted, five-state, left-to-right HMMs are used. The states roughly represent the beginning, rise, peak, fall, and end of a pitch event. Transitions exist from state to state serially, as well as from beginning to peak and peak to end. By allowing the skipping of states, the models match a conceptual model where a pitch event is rise-fall, rise, or fall (e.g. the Tilt intonation model). One experiment with a four-state model (conceptually leaving out the peak state) was undertaken. No noticeable difference was found between the four- and five-state models, with the relative error rates separated by less than one hundredth of a percentage point.

For most of the speech databases used in the experiments, the models were trained on 70% of the database, and tested on a validation set which comprises the other 30%. The exception from this division is the case of the F2B database, where the validation set contains 20% of the data and 10% was held out for blind testing at the end of all experiments. This difference is because this database is large enough to allow the extra division, as is discussed further below.

4.2.2 Constraints

All of the tests were constrained by a bigram/unigram grammar which was built from the corpus being tested. The likelihood that a particular event will occur, as predicted from the Hidden Markov Models, is weighted by the probability of that event occuring as a part of the sequence which has already been predicted. The grammar was built using the CMU-Cambridge Statistical Language Modeling Toolkit [CR97]. For each database, initial evaluation of a grammar scaling factor was undertaken to determine the general range of productive grammar weighting values. The weights tested ranged from 3 to 20 (where 0 is no reference to the grammar at all). A similar set of tests examined the use of an external transition weighting, which can be used to globally penalize state transitions. A negative value lowers the transition probability (which reduces insertions), while a positive value raises the transition probability (which increases insertions). Values were tested from -60 to 30 at five-point intervals. Models were trained using odd-numbers of Gaussian components from 1 to 29. Scores were obtained for each set of models, under each constraint combination. Thus, for each database, for each testing condition (e.g. stream weight, feature type), 15 sets of models are examined. Seventeen possible grammar scaling factors and eighteen possible transition weights are examined. As is discussed below, some experiments also examined different internal stream weights. Not all of the scores provide immediately interesting insights. The complete set of scores is presented in Appendix A. Once a set of weights is determined for a testing condition, the possible combinations are reduced to cover only this range. The constraints are then optimized over the validation set, and where possible, tested at the end of all experimentation. Because many scores serve only to delimit the search area, only the best results of each database are reported here.

Most of the scores which are reported in this paper were achieved with constraints optimized on the validation set, as no held-out testing set was available for the smaller databases. However, the HMMs and optimized constraints which received the best scores were also used to automatically label the blind (held-out) set on the F2B database once all other experiments were complete. This score is comparable to the score received for the validation set, as is discussed below.

4.2.3 Data

This research is primarily based on 45 minutes of radio news broadcast from the Boston University Radio Corpus [OPSH95], speaker F2B (over 5000 intonation events). Other corpora examined are two databases spoken by the author (male American English speaker). Of these, one is a series of weather-related sentences (KDW - 2400+ events), and the other is a museum guide (KDS - 3200+ events). Each corpus has been hand-labelled with Tilt intonation labels. The intonation event inventory for this study is accents, rising boundaries, falling boundaries, and concatenated accents and rise/fall boundaries (this represents an extended inventory of the Tilt model).

The acoustic information was extracted using the following methods. In each case, the fundamental frequency was derived using Taylor's Intonation Contour Detection Algorithm [TCB98] which provides a smoothed, interpolated F0 trace, as discussed below. The Mel Frequency Cepstral Coefficients were calculated using the HCopy function of the Entropic HTK package. The energy, auto-correlation peak, and zero-crossing values were extracted from Entropic's get_f0 output. The F0 and energy values were normalized on a scale of -1 to 1 for each database individually, using a program within the Edinburgh Speech Tools [TCB98]. The mean and standard deviation values were calculated using only the speech portions of the database. Each F0 or energy value is then normalized by subtracting the respective mean and dividing by twice the standard deviation.

The fundamental frequency smoothing algorithm uses windows of 105ms (first pass) and 35ms (second pass) to remove outlying points, but to leave behind as much contiguous data as possible (thereby providing as much micro-intonation information as possible while removing isolated outlying F0 points). The large first window eliminates many of the short, sharp F0 movements often associated with micro-intonation, such as outlying F0 points. It does not, however, generally remove longer, more slowly varying movements, such as intonation events and micro-intonation over multiple segments, such as raised F0 over a high vowel. These movements are generally contiguous and occur in a consistent direction, unlike the 20-100Hz jumps from one 10ms frame to the next which can obscure the direction of fundamental frequency changes. The second window covers three frames, and further smoothes the rough edges. Again, though, the constant, contiguous movements which would normally cause problems for an intonation analysis system because of their similarity to intonation events retain their character, and still have a capacity for causing confusion. As mentioned previously, the confusion caused by these movements is what the experiments in this chapter aim to reduce.

4.2.4 Mel Frequency Cepstral Coefficients

The use of cepstral coefficients reflects some of the experimental findings in the literature. Spectral tilt and general formant information are represented in cepstra. Campbell and Beckman [CB97], among others (e.g. [SvHP97] [HW92]), have provided support for links between spectral tilt and the existence of pitch events. The formant values can provide useful information about the type of segments associated with a given pitch event. Such information should be useful in determining whether a F0 movement is associated with segment classes which are likely to be associated with intonation events, hopefully lowering the number of pitch movements which are incorrectly analysed as intonation events. The experiments in this chapter which use MFCCs typically use the first thirteen coefficients (one of which represents energy).

4.2.5 Evaluation

The output of the intonation analysis process is evaluated in terms of three basic measures: percent of detected events which are correct, accuracy (correct - percent of detected events which are incorrect), and error (100% - accuracy). Because timing is as important as scaling in intonation, the evaluation method requires a definition of correctness which accounts for both a symbol and its timing. As Hunt found with continuous speech recognition ([Hun88]), at low-performance levels, recognizers can provide the correct label, but at the wrong time. His solution was to require recognized words to occur in time with the reference word in order to be counted as correct. Within this thesis, a similar convention, used by Taylor ([Tay00]), is adopted, where a detected event is deemed correct when it overlaps an observed event by at least 50%. This loose definition allows for the equivalent of two human labellers disagreeing on the exact location of an accent within a syllable, or at most, within a word.

An important difference between the evaluation method used here and others which have been used is that it evaluates the intonation labels themselves, rather than the association between events and syllables. For example, Ostendorf and Ross [OR97] score their intonation labeller in terms of the number of syllables which are correctly labelled for intonation. This scoring method is a valid and useful way of assessing the success of accent association algorithms. It is also a useful way of evaluating intonation events which are essentially event peak markers. This method, though,is not useful when the intonation events have a duration. It is important, with the Tilt events, to assess the location of the whole event. The evaluation method used in this thesis looks at the timing and duration of each event, judging the success of the model on an acoustic level.

A quantitative assessment of the automatic labels is a second way of determining how well the models work. This type of evaluation shows whether the models produce labels in a similar distribution to the manual labels in a database. This type of evaluation is useful as a secondary check. If, for example, the models are failing according to the acoustic evaluation, it could prove useful to examine the distribution of the output. More importantly, though, it is necessary to examine the output distribution if the models appear to succeed according to the acoustic evaluation. Because the first evaluation accepts a match between any event types, it is necessary to show that the models are actually labelling the data in a similar fashion to the human labellers. Table 4.8 (109) shows such an analysis for the most successful model set described in this chapter.

4.3 Pilot Study

An initial pilot study investigated links between sub-segmental acoustic data and intonation. This series tested a number of low-level acoustic features which are related to both intonation and broad classification of speech sounds. Zero-crossings, auto-correlation peak, and energy are used in some fundamental frequency calculation methods (e.g. Entropic's get_f0). These features are also useful in distinguishing broad classes of speech sounds (e.g. voiced/unvoiced consonants, high/low vowels). Energy is already used in conjunction with fundamental frequency in intonation analysis algorithms (e.g. [Ros94] [Tay00]). The first experiments emulated Taylor's study [Tay00], building the HMMs using only fundamental frequency and RMS energy. The second round of experiments examined whether zero-crossings or autocorrelation could be used to improve intonation event detection. As discussed below, these experiments showed that adding either zero-crossings or autocorrelation to F0 and energy did not improve intonation event detection.

In an attempt to continue approaching the task as one of speech recognition, a further experiment was designed to include Mel Frequency Cepstral Coefficients with F0 and energy. MFCCs are used in some speech recognition systems ([YJO^{+}96]), and are accessible directly from the waveform. The initial experiment tested MFCCs 0-4 (energy and the first three coefficients) in conjunction with F0. This experiment showed promising results, and a new series of tests was designed to use all thirteen coefficients. This series forms the basis for the rest of the intonation event detection experiments described in this chapter.

4.4 Experiments

The basis of comparison for this study is a portion of Taylor's study [Tay00] which examines event detection of the F2B data. However, prior to the outset of this study, some errors in the hand-labelled events which he used were corrected. Therefore, it is expected that, while very similar, the results of the replication experiments will differ somewhat from Taylor's results.

Taylor built models of intonation event types using F0 and RMS energy. The portion of his research on F2B which relates to this study used normalized F0 and RMS energy, together with the first and second derivatives of each feature. The results of the experiments which are relevant to this chapter are 79% of detected events correct, and 59% accurate (error of 41%). Taylor's use of normalized values stems from his desire to create a speaker-independent analyzer. Both normalized and non-normalized values were investigated in the experiments discussed here.

Two forms of Taylor's study were replicated in the process of creating baseline results. First, non-normalized F0 and RMS energy were modelled, with results (Base 1) in Table 4.1 of 78% correct and 61% accuracy (error of 39%).

	Correct	Accuracy	Error
Taylor	79%	59%	41%
Base 1	78%	61%	39%
Base 2	78%	59%	41%

Table 4.1: Comparison of baseline results

As these results were reasonably close to Taylor's, normalized F0 and RMS energy were modelled in order to provide a direct comparison to [Tay00]. The results of this experiment (Base 2) were 78% correct and 59% accuracy (error of 41%). The close similarity of these results allows for a reasonable comparison between any results in this chapter and [Tay00].

4.4.1 Zero-crossings and Auto-correlation Peak

The results from the initial F0/energy experiments provide a point of departure for investigating segmental/suprasegmental interaction in intonation analysis. This investigation involved two variations on the basic method. As noted above, the basic process is to build HMMs using fundamental frequency and one or more additional features, as well as the first and second derivative of each feature. The first variation is to build the HMMs using only the additional feature data. One set of experiments follows the basic methodology. The zero-crossing or auto-correlation peak data is included when building, training, and using the hidden Markov models. This creates a database of nine-item feature vectors (F0/energy/other feature + first and second derivatives).

The second variation on the basic methodology involves splitting off the new feature data so that the feature vector contains two streams which may be weighted separately. This weighting is one of the constraints introduced in section 4.2.2. One difficulty with using weighted data is determining what the weighting should be. The experiments which involved weighted data also involved testing a number of weights for each data stream. The weighting tests consisted of holding one stream weighting at 1, while altering the other stream weighting from 1.6 to 0.6 in 0.2 step intervals. The process was repeated with the previously static stream being altered. The best weights were found to be 1 for the stream which includes F0 and either 0.8 or 0.6 for the other stream. The results reported in this section which involve weighted data include the weighting for the second stream.

Table 4.2 shows the best results from the zero-crossing experiments (using non-normalized data). It is obvious that zero-crossing data alone cannot provide any useful input into event detection. The number of insertion errors drove the error to well over 100% (accuracy of <0%). This means that any correct detections were more than cancelled out by insertions. The results of this test are borne out when zero crossing data is combined with F0 and energy in unweighted data.

Z-C Alone	With F0 and Energy
<-0%	58%

Table 4.2: Accuracy of auto-labelling with HMMS using zero-crossing data

Table 4.3 provides more hopeful results. Auto-correlation peak information was significantly more useful than zero crossing, with an error of 74% (accuracy of 26%) when used alone. When added to F0 and energy in unweighted data, the result was a reasonable 61% accuracy. The relative success of tests on unweighted data encouraged weighting the data, which resulted in further improvements

A-C Peak Alone	With F0 and Energy	Weighted with F0 and Energy
26%	61%	63% (0.8)

Table 4.3: Accuracy of auto-labelling with HMMS using auto-correlation peak data

As table 4.4 shows, this experiment yielded a relative error reduction of 4% against the baseline result. The relative error increase of 8% over the baseline result suggested that further experimentation with zero crossing data would be fruitless.

4.4.2 Experiments with MFCCs

As mentioned above, the results of auto-correlation peak and zero-crossing data encouraged experimentation on a somewhat different data type. If simple correlates of segments can affect the model output, a more complex form of information may prove more effective. There are thirteen cepstral coefficients used in leading speech speech recognition systems (e.g. [YJO+96]). These coefficients are typically used in conjunction with their first and second derivatives, in order to trace the spectral change over time. In the interest in saving computing space and time until the utility of MFCCs was determined, the initial tests only used four of the coefficients. Table 4.5 shows the top results of this series of tests.

Experiments on using Mel Frequency Cepstral Coefficients in conjunction with non-normalized F0 data show even greater error reduction than auto-correlation. The relative error reduction over Base 1 of 9% (error of 36.5%) encouraged experimentation using all coefficients. The success of weighted data in the auto-correlation peak experiments suggested that weighting would be interesting for these experiments.

The results from experiments with non-normalized F0 and all thirteen MFCC (all data for all experiments includes first and second derivatives) are very promising. As Table 4.7 illustrates, a relative error reduction of 15% (accuracy of 67%, weight: 0.6) shows that the use of MFCC data is a step forward in automatic intonation analysis.

In order to allow direct comparison between this work and previous research, [Tay00], normalized F0 was used. Table 4.6 shows the results for the two best weightings for the normalized F0 and MFCC experiments.

The relative error reduction of the MFCC experiments is encouraging, but it could also be incomplete. The manner in which error is calculated allows for an error reduction without a decrease in insertions (by improving correct detection). Therefore, an investigation of all three evaluation metrics is useful to determine whether using MFCCs to improve analysis results in increasing the number of correct identifications, decreasing the percentage of identifications which are insertions, or a combination of the two.

Table 4.7 shows a comparison of the MFCC experiments with the respective baselines and [Tay00]. As accuracy is correct minus the percentage of detections which are insertions (incorrect), it is important not only that the correct score rises, but also that the gap between correct and accuracy shrinks. The non-normalized experiment shows a rise in both correct and accuracy, resulting in a reduction of error. However, one may note that the percentage of insertions has remained the same (17%). This means that the error reduction, while welcome, is not the result of reduced insertions. The results of the normalized data, in contrast, show both an improvement in correct identification and a reduction of insertions (from 19% to 16%). Thus, while the normalized data does not show as large an improvement over Base 2 as the non-normalized data shows against Base 1, the improvement is on a wider scale.

Finally, the table shows that, on the blind set, the result pattern is upheld. This result is the true test of the system, and was only performed once the other MFCC experiments had finished. The blind set was labelled using

Acoustic Features	Error	Relative Change
F0 and Energy + Zero Crossings	42%	+8%
F0 and Energy + Weighted A-C Peak	37.5 %	-4%

Table 4.4: Error change relative to the baseline

Unweighted	Weighted (0.8)	Relative Error Reduction to Baseline
64.5%	63.5%	9%(weighted)

Table 4.5: Accuracy of auto-labelling with HMMS using four MFCCs and normalized F0

Weight	Accuracy	Relative Error to Baseline
0.8	63%	-10%
0.6	64%	-12%

Table 4.6: Error of experiments using 13 Mel Frequency Cepstral Coefficients to augment Normalized F0 and energy, with relative error

	Correct	Accuracy	Error
Base 1	78%	61%	39%
Non-normalized MFCC	84%	67%	33%
Taylor	79%	59%	41%
Base 2	78%	59%	41%
Normalized MFCC	80%	64%	36%
Blind Set	85%	66%	34%

Table 4.7: Comparison of results to baselines and Taylor

normalized data and the HMM settings which were used for the best results from the validation set. Thus, it should be compared with Base 2. The similarity of the results between the blind and validation sets serves two purposes. First, it confirms the results on the validation sets. Without this confirmation, the validation results are not necessarily indicative of results on unseen data. Secondly, the confirmation provides some strength to the results of tests of this methodology on smaller databases, which were not large enough to support dividing the utterances into three sets. As the table shows, the blind set performs as well as any of the validation sets. The blind set, reproducing the best results from the validation sets, produces results which are comparable to the 88.6% correct and 74.8% accuracy scores of human labeller comparisons using the Tilt model [Tay00]. While not quite of the same level, the results from these experiments are very promising. The work in this chapter was performed using a database of news broadcast. The human comparison from Taylor is from a few minutes of a dialogue database, with a large number of rising boundaries and large pitch excursions, which made for generally straight-forward labelling.

- ***A quantitative analysis***

As mentioned above, it is important to show that the successful models are in fact producing labels with a distribution similar to the manual labels. Table 4.8 shows how the models which produced the figures on the blind set shown in 4.7 (85% correct, 34% error) compare quantitatively with the manual labels.

The table shows that the distribution of intonation events of the automatic labels is very similar to that of the manual labels (except for "rb"). It is reasonable to conclude from this that the two label sets are similar. The

Event Type	Automatic Labels	Manual Labels
a	4225	4581
afb	236	297
arb	34	59
fb	1102	1059
rb	348	632

Table 4.8: A quantitative assessment of automatic intonation labels

notable exception of rising boundaries is related to the poor representation of rising boundaries in the data. Unlike the other event types which are not well represented (afb, arb), rising boundaries need not be large pitch excursions. The rises which have not been recognized are generally small excursions. These rises probably are not distinctive enough to train the HMM adequately.

4.4.3 Using MFCCs without the Second Derivative

The pilot experiments showed that zero-crossing data was wholly ineffectual for improving automatic intonation analysis. Auto-correlation peak information was somewhat useful, but the greatest improvement in accuracy of analysis was achieved by using normalized cepstral coefficients together with fundamental frequency to build and use hidden Markov models. This large improvement may well be the result of simply adding more data (42 data items per vector as compared with 9) than auto-correlation peak can provide.

One risk of using MFCC data, as has been alluded to already, is that the large feature vectors requires a large database for training, as the number of distinctions which may be made is larger with a 42 element feature vector than with a smaller vector. Two small experiments were used to further examine the nature of the MFCC data. One experiment adds acceleration peak to the F0 and MFCC data. The other removes the second derivative from the MFCC data.

The first experiment combines the acceleration peak and MFCC data into the second stream. Because the acceleration peak and MFCC coefficients provide overlapping data, the expectation is that the results will not be an improvement over the best results seen so far. However, there was a possibility that, as both acceleration peak and MFCC data helped to improve the simple F0/Energy model, the combination of the two acoustic feature classes would result in a further improvement. The recognition result on the blind set was 79.30% Correct and 62.31% Accuracy. This result is, as expected, lower than the best results for F0 and MFCC data.

The second experiment uses normalized F0 and MFCC data without including the second derivative for the cepstral coefficients. The second derivative was removed to reduce the vector size while hopefully retaining the ability of the model to track the change over time in the spectral shape. The MFCC weight for this experiment was 0.8. This experiment resulted in higher accuracy on the validation set than was seen in the experiments which included both derivatives, as shown in Table 4.9. The table also shows that the loss of the second derivative for the MFCC data coincides with a poorer result on the blind test set. This result suggests that the second derivative is required to minimize over-training, as it takes into account several frames of data. It also suggests that using further derivatives to take into account even longer time windows may be an interesting topic for further research.

4.5 Extension to New Databases

The level of improvement which is achieved by adding cepstral information to the intonation analysis process indicates that acoustic data which reflects the type of segmental text associated with an intonation contour is useful for intonation analysis. In order to press this claim, three databases were tested in addition to F2B. As discussed above, each database is substantially smaller than F2B. Therefore, no blind set was held out for further use, primarily because it would consist of no more than a paragraph or two. Instead, the tests rely on the assumption gained from F2B that, given a reasonable sized database, the blind set will score similarly to the general test set.

Table 4.10 shows results on KDS, the largest of the databases read by the author. The most notable aspect of these scores is that on all counts, they are considerably lower than those for F2B. The list of possible reasons for this difference is extensive. The most likely reason is that the database is 60% the size of F2B. The database must be large enough to provide enough instances of each event type for the statistical model to form generalizations. KDS is quite small, and there are some events (e.g. "rb") which are not well represented. One way to test the hypothesis that database size is affecting the quality of results is to apply the method to another small database. We would expect that performance would improve slightly with a larger database, and degrade with a smaller database. This expectation is born out by KDW,

	Correct	Accuracy	Error
Validation Set	84.22%	68.01%	31.99%
Blind Set	79.96%	59.69%	40.31%

Table 4.9: Evaluation of F0 with first and second derivatives plus MFCC with first derivative

	Correct	Accuracy
Normalized Data		
F0 + Energy	71.08%	56.21%
F0 + MFCC (weight 0.8)	77.82%	60.28%
F0 + MFCC (weight 0.6)	75.96%	59.93%
Non-Normalized		
F0 + Energy	71.31%	56.44%
F0 + MFCC (weight 0.8)	73.98%	59.12%
F0 + MFCC (weight 0.6)	74.1%	59.7%

Table 4.10: Analysis Results for Database KDS

	Correct	Accuracy
Normalized (0.8 weight)	84.16	56.19
Non-Normalized (0.6 weight)	78.47	51.23

Table 4.11: Analysis Results for Database KDW

which is smaller than KDS, as shown in Table 4.11. For this very small database, the process failed to result in HMMs capable of producing sensible label files at all. The sequence of labels produced for KDW generally followed the lines of "sil, c, rb, rb, rb, sil." Considering that only 88 of the 2400+ events in the database are "rb" events, such results are not impressive. It appears that the small continuation rises which constitute most of the "rb" events in this database are contributing to a model which incorrectly labels small rising F0 movements as pitch events. The other models would generally be expected to counter this error, with minor movements being modelled in the "c" HMM. However, the paucity of data means that none of the models is as robust as it should be for accurate labelling, and the fit of any portion of changing F0 to the correct model is essentially chance.

4.6 Discussion

This chapter has shown that it is possible to improve upon previous methods of automatic intonation event detection without relying on interpretations of acoustic data (e.g. phone, syllable, word annotation). One advantage of improving the acoustic processing rather than relying on higher-level data is that methods which do include linguistic constituents (e.g. [OR97]) in the analysis process can also use the methods discussed here, hopefully with improved results. The greatest advantage to improving acoustic intonation processing methods is the minimization of process ordering constraints. Intonation analysis can be performed before, after, or in parallel with other speech analysis tasks, without relying on a computationally expensive Viterbi search.

The most important difference between this research and previous research which uses only acoustic data is that this work presents a way of approaching the interaction between the supraglottal vocal tract and intonation. As discussed in Chapter 2, exploitation of this interaction appears to provide a significant step forward in intonation modelling in general. Advances in including sub-syllabic constituents in intonation generation rely on knowing what those constituents are at generation time. Such knowledge, though, is not always available to speech recognition and understanding systems.

The use of Mel Frequency Cepstral Coefficients within the context of intonation analysis is a novel application of general speech recognition methods. Until such a time as many hours, as opposed to many minutes, of intonationally annotated speech are available for model training, problems will hamper continued research in this area. Difficulties such as generally nonsensical models being built, as occurred with KDW, or minimal progress, as was the case with KDS, will remain in the short-term. However, the limited success this method achieved on a 45 minute speech database shows how by providing good labels which require manual correction, bootstrapping from minimally labelled databases can lead to greater data availability, similar to the growth in other areas of speech recognition fifteen years ago. As intonationally labelled data becomes more widely available, real-time applications which utilize intonation information will be able to incorporate the type of intonation analysis described in this chapter.

Chapter 5

Synthesizing Intonation

This chapter describes original work in predicting intonation from data labelled for textual and intonation information. Each chapter up to this point has acted as a building block for some part of the work described below. As discussed in Chapter 3, the research uses the Tilt intonation model [Tay00]. This model is a parameterized description of fundamental frequency contours. The intonation synthesis models described below are designed to predict the parameters which form Tilt descriptions of F0 contours. Chapter 2 discusses ways that segments interact with intonation. Part of the research discussed in this chapter investigates some effects that exploiting these interactions can have on the intonation synthesis models. As we saw in Chapter 4, stochastic modelling techniques can be data-intensive. Section 5.4 shows how the synthesis techniques described in this chapter can be used to build models from the automatically labelled data from Chapter 4.

Some F0 prediction methods ([Ros94], [BH96]) predict the fundamental frequency on each syllable. The work described here is based on the view that intonation forms a separate level from the text, and should be treated accordingly. Rather than predicting F0 values for each syllable, the shape of each accent is predicted and anchored in the time and frequency domains, in accordance with the Tilt intonation model. Therefore, generating F0 contours is a two-step process. First, the Tilt description parameters must be predicted. Then the parameters are translated into F0 contours. The research described in this chapter is only concerned with the first step of this translation.

The intonation synthesis models which are described below predict the parameters which describe intonation contours under the Tilt intonation model. This chapter presents experiments which examine whether intonation contours generated using these models are adequately similar to natural intonation contours from the same utterances. Along the way, the analyses of the models provide insights into more detailed aspects of intonation synthesis. We examine whether, as Chapter 4 suggests, some interactions between segments and fundamental frequency can be used to improve intonation synthesis. We also look at how different types of input to the models affect the prediction of the Tilt parameters. The input types range from segmental categorization (e.g. sonorant/obstruent), prosodic phrasing, and intonation context. The methodology described in this chapter is designed to provide transparent results, so that we can understand which types of input are useful in modelling different aspects of intonation. Unlike the work in the previous chapter, there is no constraint on using only acoustic data to train models. Primarily, this is because the synthesizer is creating the acoustics, rather than analyzing them. The bias towards acoustics in this chapter is that the end product of the synthesis models is a low-level description of the timing and frequency of the intonation events, which can be directly translated into a fundamental frequency contour. The only constraint on the type of input to the models is that it be readily and sensibly available at F0 generation time in a speech synthesis system. “Sensibly available” is a weak constraint, but in the context of this discussion it means that the synthesizer need not suffer any re-ordering of processes in order to use the models.

All of the original research presented in this chapter was undertaken using the Festival Speech Synthesis System [BTC98] and the Edinburgh Speech Tools speech processing package [TCB98], which are widely accessible. While I have had some part to play in the development of these packages for intonation processing, the vast majority of the infrastructure they provide is the work of staff at the Centre for Speech Technology Research, University of Edinburgh.

5.1 Methodology

The intonation prediction experiments consist of a basic four step process, as shown in detail in Figure 5.1. First, information about each utterance in a database is extracted. In the figure, this step encompasses the top three rows. As discussed in 5.2.2, the Tilt descriptions form a part of the context feature set, as the features being modelled. Therefore, these descriptions must be made available from a Tilt analysis step, (see section 3.5.1 and [Tay00]).

Regression trees are built for each Tilt parameter of each Tilt event type, shown in Figure 5.1 as the Wagon step (for details of the Tilt model, see section 3.5.1). As discussed above (3.2.3) and below (5.2.1), regression trees are a type of decision tree. In this case, the regression trees are binary decision trees which predict a value for each Tilt parameter for each Tilt event given its context in the utterance to by synthesized. These models are then used to generate Tilt descriptions of the fundamental frequency for each utterance. The F0 contours that result from these descriptions are then

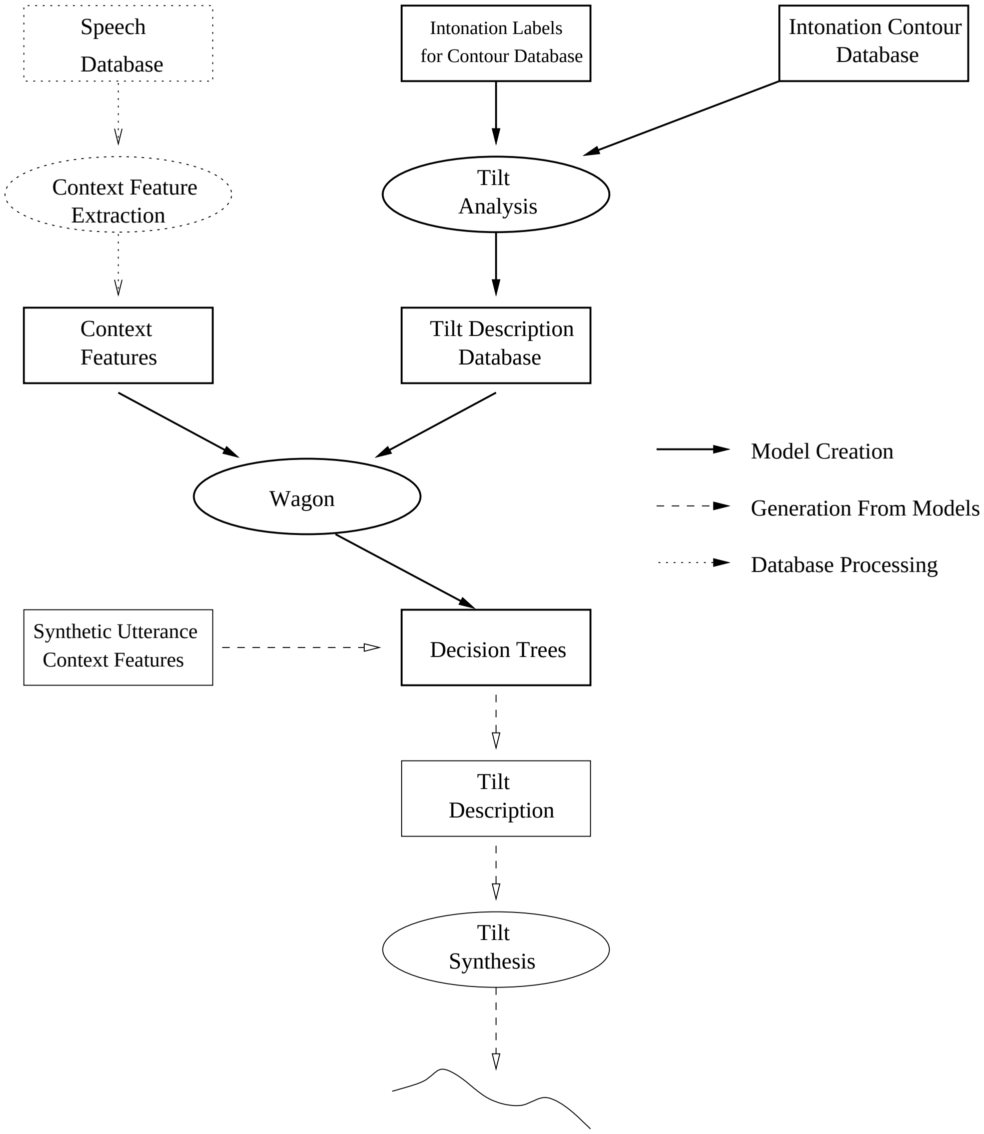


Figure 5.1: Creating and Using F0 Generation Models

scored against the original contours (see section 4.2.5).

Page 120 shows a regression tree for the "accent/falling boundary" class, to illustrate the process of modelling the Tilt parameters and then using these models to generate an F0 contour. Figure 5.2 shows an accent/falling boundary ("afb") event spoken by FHL (see the next section for database information). The top of Table 5.1 shows the original *peak position* value for the pictured event, together with values for those extracted features (section 5.2.2) which are relevant to the *peak position* prediction tree for "afb" events. Below this list is the regression tree which is used to predict the *peak position* values for "afb" events. Each branching leaf of the tree (those with feature names) contains a feature name and a query about that feature. The non-branching leaves of the tree (those with two numbers) give the standard deviation and mean values for all examples of "afb" in the database which are described by the conditions leading to a given leaf.

The first question in the tree in Table 5.1 (line 1) asks how many syllable have passed since the previous accent. If less than 3.5 syllables have passed (effectively three or fewer syllables), then the second question is asked (line 2). If four or more syllables have passed, then the leaf on line 13 of the tree is selected for the prediction. The second question asks how many syllables remain before the next major phrase boundary. If there are five or fewer syllables, then question three is asked (line 3). If six or more syllables remain, then the leaf shown on line 12 is selected. Question three asks if there are fewer than three syllables remaining before the next major phrase boundary. If so, then question four is asked (line 4). If not, then the leaf on line 11 is selected. Question four asks if the syllable associated with the "afb" event is word final in a multisyllabic word. If so, then the leaf on line 5 is selected. If not, then question five (line 6) is asked. The fifth question asks if the onset

Original Peak Position	-0.108
Syllable.last_accent	1
Syllable.ssyl_out	8
Syllable.position_type	final
Syllable.lisp_get_onset_length	0.28
1 ((Syllable.last_accent < 3.5)	
2 ((Syllable.ssyl_out < 5.5)	
3 ((Syllable.ssyl_out < 2.5)	
4 ((Syllable.position_type is final)	
5 ((0.0772446 -0.111))	
6 ((Syllable.lisp_get_onset_length < 0.16)	
7 ((0.0692528 0.028619))	
8 ((Syllable.lisp_get_onset_length < 0.208)	
9 ((0.070402 -0.0154167))	
10 ((0.101499 -0.0567143)))))	
11 ((0.0673618 -0.00727273)))	
12 ((0.0585719 -0.055)))	
13 ((0.0646414 -0.00821053)))	
Predicted Leaf	((0.0585719 -0.055)))
Synthetic Peak Position	-0.055

Table 5.1: Relevant extracted features, afb peak position regression tree, and peak position value predicted for this afb using the regression tree

of the syllable associated with the event is less than 160ms. If the answer is yes, then the leaf on line 7 is selected. If not, then question six asks if the onset is shorter than 208ms. A yes answer to this question results in the leaf on line 9 being selected for the prediction. If the answer is no, then the leaf on line 10 is selected.

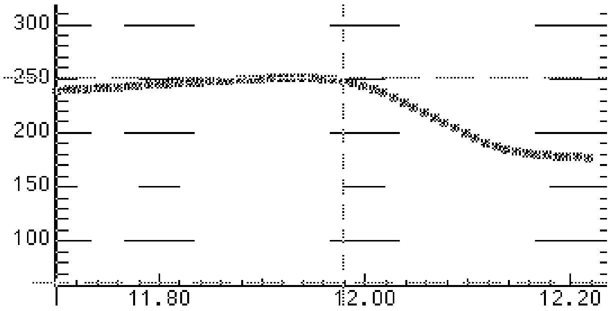


Figure 5.2: Original F0 contour

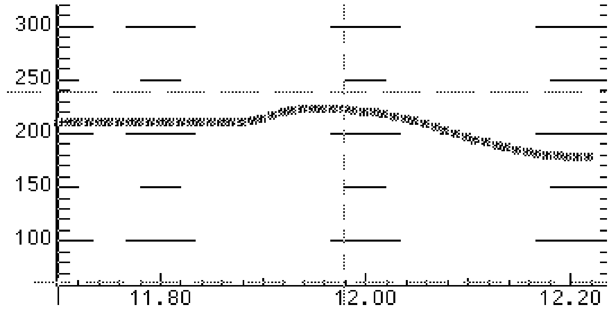


Figure 5.3: Synthetic F0 contour

The bottom of the table shows the leaf which is selected from the tree if the feature values at the top are plugged into the tree. This leaf is selected because the answer to question one is yes, and the answer to question two is no. Therefore, the leaf shown on line 12 is selected for the predicted value. The second number in this leaf is the mean of all "afb" *peak position* values in the training data for the database which fit the criteria for the leaf. This mean is the predicted *peak position* value which forms part of the Tilt description, as shown in Figure 5.1. As shown in the figure, and discussed in section 5.2.3, when processed with the other predicted parameter values, is used to generate the F0 contour shown in Figure 5.3.

5.2 Data

The data used in the intonation synthesis experiments covers three distinct speech types: news commentary, isolated sentences, and instructional text. The same experiments were carried out on each of the databases described below.

The news commentary database is a portion of the Boston University Radio News Corpus [OPSH95], speaker F2B. This database is the same database used in the previous chapter. The database consists of 114 paragraphs of news commentary (approximately 45 minutes) as delivered by a female speaker of American English. The database is labelled with segment, syllable, and word boundaries, and includes lexical stress markings. It is also labelled with intonation labels based on the Tilt intonation model [Tay00]. The labellers not only provide intonation labels, but also provide an association between each event and a syllable. This association is determined on the basis of both visual evidence (where the peaks and troughs occur) and audio evidence (which syllable sounds accented). The F2B database is used for a more examinations than the other databases, due principally to its size and availability.[1]

The isolated sentence database (KDT) is a set of 450 (TIMIT-style) phonetically balanced sentences, of which ten percent are questions. These sentences are spoken by a male American English speaker, and are annotated in the same manner as the F2B database.

[1] Tests were run on F2B a year before any of the other databases were available.

The instructional text database consists of forty-three paragraphs of computer-generated text which describes exhibits in a museum. This database has one set of the paragraphs spoken by a female Scottish English speaker and one set spoken by the same speaker as the isolated sentence database. This database is labelled with word boundaries and intonation labels. Syllable boundaries are estimated, and segmental boundaries are not used. Lexical stress is taken from dictionary entries, and is therefore approximate.

Each of the databases is tested in isolation. Cross-data training was avoided so that each individual speaker and style could be modelled without the difficulties caused by the different data types.

5.2.1 Building Regression Trees

As mentioned above, regression trees are decision trees which can be used to divide data according to a series of questions. The result of dividing the data is to arrive at a subset of the data whose mean value can, in the best case, be used in place of the original value if the tree is used in reverse, to generate values. The trees consist of questions about features which are used to predict a particular parameter. Each node of the tree consists of a question, a sub-tree for "yes" answers, and a sub-tree for "no" answers. The leaves of the trees contain mean and standard deviation values for the data points which are classified by the answer path required to reach a given leaf.

For each tree needed (one for each parameter for each accent type), a tree is begun by finding the feature that partitions the data such that the standard deviation is lowest within the two partitions. The tree is then used to generate values for the parameter for each instance of the relevant event type in a held out set. The tree is grown by continuing such question selection until a specified minimum number of data points is reached. The algorithm is greedy, in that it selects the best partition and question at a given time, rather than testing all possible combinations, which is computationally prohibitive. The tree receives a "score" which consists of the Root Mean Squared Error and a Pearson's Correlation Coefficient, relating the generated and actual values on this held out set.

As noted above, the data is divided by both accent and parameter type. For example, the tree for the peak position parameter for accents is separate from the peak position tree for rising boundaries. The parameter separation is possible because the Tilt parameters have been shown to be statistically independent [Tay00]. Thus, each parameter can be predicted in isolation, as it should not be statistically tied to the other parameters. The separation is also useful in minimizing the database size requirements, as with the HMM modelling discussed in Chapter 4 Separately modelling the different event types also allows one to examine how different features affect the various aspects of each event type.

Previous experiments which have used this technique ([DB97] [BDTss]) have included minor hand-optimising of the feature set for noise reduction. However, as discussed in section 5.5, it is unclear whether the resulting, nominal improvement in correlation (2 percentage points) over a large corpus has any real effect on any particular intonation contour. Therefore, the results of experiments below, unless otherwise noted, do not include any hand-optimisation of the feature set.

The regression trees used for the experiments were built using the Wagon classification and regression tree tool [TCB98] which uses standard cart techniques [BFO84].

5.2.2 Feature Extraction

The most difficult portion of this procedure is the first step, which is as much intuitive as computational. For each utterance, a variety of information is extracted which may assist in modelling F0. The difficulty of this step is not in extracting the feature information, but in determining which features to use. A list of the features extracted for these experiments is below. Most of the features are directly related to the literature which is reviewed in earlier chapters. Some features were developed specifically to address specific questions. A discussion of the which features were useful in particular trees follows the results of the experiments. Figure 5.4 illustrates the two different ways in which the data is approached. For any given intonation event (e.g. the circled *a* in the figure), features are extracted in terms of a syllable window and an intonation window. It is important to note that, regardless of which window is being used, connections between the syllables and the intonation labels are preserved. For example, one feature looks two syllables ahead of the circled syllable in order to find out if there is an intonation label associated with it. For the purposes of analysis, the features are described in terms of five classes, which are presented below.

The lexical stress (0 or 1) of a given syllable and the two syllables on either side make up the first feature class. The second class concerns the position of a given syllable within a phrase. The features extracted in the first two classes follow [BH96], who generate F0 values for each syllable in a synthetic utterance (see section 3.2.3). The second set of features are:

- The distance in syllables from the previous event (i.e. accent or bound-

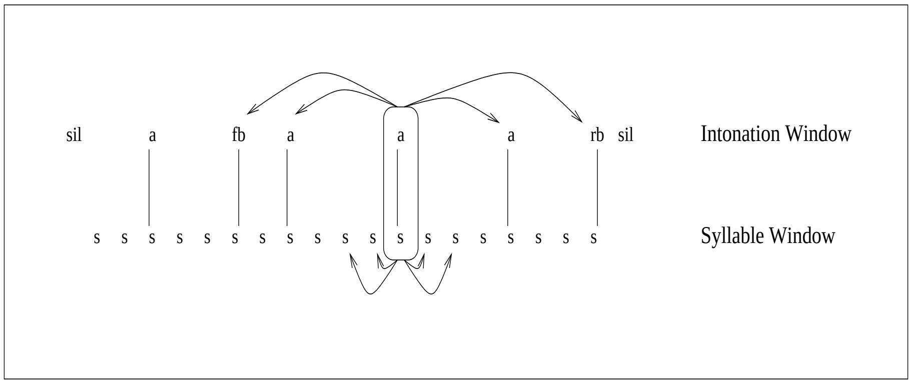


Figure 5.4: An illustration of the feature extraction windows

ary) and to the next.

- The distance in syllables from the previous major phrase break and to the next.
- The distance in lexically stressed syllables from the previous major phrase break and to the next.
- The distance in accented syllables from the previous major phrase break and to the next.
- The phrase break index (0-4) of the syllable in a window of two before and two after.

The third feature class contains information about the composition of the syllable and its place in a word (referred to in this thesis as sub-syllable features). The composition-related features are the length of the syllable onset and rhyme and a classification of syllable onset and coda, following [vSH94] and [PvSH95] (see section 2.3.

The fourth class is similar to the lexical stress category, but relates to intonation events. Two features are used here, one each for accent and boundary.

A value is extracted (0 or 1) if the syllable is associated with an accent or a boundary.

The final class is more suprasegmental in nature than the other classes. Rather than being based on syllables, the features in this class are the types of event associated with a syllable, and the two events on either side, regardless of their location in terms of syllables. This class uses the intonation window shown in Figure 5.4. This view of the data was necessary because intonation events do not occur on every syllable, and a syllable-based window will not always contain information about any events. If only a syllable window is used, some intonation events are viewed in isolation from other events, as the previous and next events may be three or more syllables away.

5.2.3 Generating F0 Contours

As noted at the outset of this chapter, the shape of each accent is predicted and anchored in the time and frequency domains. Generating F0 contours is a two-step process. First, the regression trees just described are used to predict the Tilt description parameters. Then the parameters are translated into F0 contours. The research described in this chapter is only concerned with the first step of this translation. However, a brief look at the way the Tilt descriptions are translated into F0 contours is included so that the reader is aware of the whole process whereby the F0 contours are generated.

The translation process uses software available in the Edinburgh Speech Tools package [TCB98]. The Tilt synthesis tool takes the parameterized description, and in accordance with the equations detailed in section 3.5.1, gives a shape to each intonation event. Each event is described by the fundamental frequency at its beginning, the time at which the peak in F0 occurs, and the gross amplitude of the F0 movement over the accent. It is mathematically trivial to plug the values into the equations and arrive at an event shape anchored in both time (by the peak position value) and frequency (by the starting F0 value) domains. The spaces in between events are then filled by straight lines.

It is important to note that the straight line connections make an exact replica of an intonation contour impossible to achieve. However, it has been argued repeatedly that such a stylization of non-event material has no audible effect on an utterance's intonation pattern ([dP83], [tHCC90], [tH91], [Tay00]). Therefore, as the next section discusses further, the target for F0 generation cannot be 100% correlation with the original, but between 85 and 95% correlation. This range reflects the success that would be acheived over the various databases if the Tilt descriptions were exactly reproduced by the decision trees. As section 3.5.1 mentions, the best results for F0 regeneration from Tilt descriptions fall within this range. As such success is yet to be reached by this or any other research, it is reasonable to assume that approaching such a target would constitute a considerable achievement.

5.3 Results

Because generating fundamental frequency contours using the methods described in this section requires two steps, two levels of evaluation are available. As section 5.3.1 discusses, each regression tree is given a score (RMSE and correlation) based on its ability to produce the correct values in a held-out validation set. These scores are used to determine whether individual Tilt parameters are being modelled as expected. The second level of evaluation gives an idea of how well entire intonation contours relate to the original data, again using RMSE and correlation. As noted in section 5.5, the evaluation of fundamental frequency differences does not easily lend itself to fine judgements about important questions such as whether the duration of accents is being modelled well. The results discussed in the remainder of the chapter should be viewed with this two-tiered assessment system in mind. Each database is evaluated primarily by the individual tree scores, with the gross measurements over whole contours playing a support role. This departure from discussing intonation synthesis in terms of intonation contours is a result of both the difficulty in objectively evaluating F0 and the lack of useful information which such an evaluation can provide.

5.3.1 Decision Tree Assessment

The basic measure of success in all of the intonation synthesis experiments is the scores of the Tilt parameter prediction decision trees. These scores provide an insight into which aspects of intonation are modelled well, and which are modelled poorly. They also provide and instant assessment of whether a new feature or feature set results in better models.

The minimal success requirement is obtained by comparing the RMSE value for a tree to the standard deviation of the same parameter's values in the training set. While the RMSE and standard deviation are not completely comparable, it is logical to presume that, if the dataset which is being tested has the same distribution as the training set, then predicting the mean on the validation set would result in an RMSE similar to the standard deviation of the training set. This presumption results in the weak test that, if the RMSE score for the tree is not near or lower than the standard deviation of the training set, the tree is ineffectual - producing values which are no better than chance, and one would be better off just predicting the mean throughout. For example, Table 5.2 shows that the standard deviation of accent peak position is 101 milliseconds. If the mean value were generated for each accent peak, the peak would occur 70ms after the start of each vowel associated with an accent event. The resulting RMS error would be 101ms. If the tree which is trained to predict accent peak position results in predictions with an RMSE of 100 milliseconds against the correct values of the validation set, then it is likely that using the tree will not result in better intonation than predicting the mean peak position. As section 5.3.3 will show, such trees produce wholely inadequate intonation.

The trees' correlation scores have a somewhat different function. As with the F0 contour comparison, they show whether the relationship between the actual and predicted values is holding (e.g. are *amplitude* values consistently in the right range). An example of where this is useful is in predicting the *tilt* parameter. As Section 3.5.1 discusses, the *tilt* parameter has a trimodal distribution for accents, with a small number of instances in the ranges between the value concentrations. A high correlation for the *tilt* tree would suggest that this distribution is being accurately modelled, regardless of the RMSE scores. Tilt values are roughly divided into three ranges: high (full rises), low (full falls), and middle (rise-falls). If the inter-relationship among these three accent types is maintained, for example, by predicting values of -0.8, 0, and 0.8, then the predicted values are within the same range as the correct values, and correlation will remain high.

While Appendix B contains detailed tables of results for various databases and experiments, this section gives a more digestible account in the context of modelling the intonation of the KDT database. A number of different experiments were undertaken on each database, but those discussed here for KDT were performed on all of the databases. The baseline scores came not from parameter prediction trees, but from replacing the trees with the mean and standard deviation for the parameters. This baseline is used for the following reasons. First, it is relatively straight-forward to calculate these values for each parameter. If these values result in adequate output, then there is no need for the use of regression trees or training algorithms. Secondly, a pair of numbers is substantially easier to store and is faster to use than a long decision tree. As mentioned above, there is also a loose relationship between the standard deviation and RMSE values, for a parameter and tree respectively, which is a useful point of departure for process development.

Table 5.2 shows the mean and standard deviation values for each parameter of some intonation event types (accents: *a*, rising boundaries: *rb*, and falling boundaries: *fb*). As mentioned above, the five parameters of the tilt description are *start F0*, the F0 value at the beginning of the event; *amplitude*, the measure of how much the F0 rises and falls; *duration*, the length of the event (given in seconds); *tilt*, a description of overall event shape; and *peak position*, the time, relative to the start of the vowel of the associated syllable, at which the event peak occurs.

	start F0	amplitude	duration	tilt	peak position
a	**218.23**/22.02	**32.96**/23.38	**0.26**/0.08	**0.16**/0.53	**0.07**/0.10
rb	**129.16**/28.2	**27.18**/22.85	**0.18**/0.04	**0.28**/0.88	**0.09**/0.12
fb	**200.00**/23.05	**37.84**/32.75	**0.20**/0.07	**-0.26**/0.45	**-0.03**/0.12

Table 5.2: Database KDT: Mean Values and Standard Deviation for Tilt Parameters of some Intonation Event Types (Mean in bold)

As is mentioned above, the initial goal of using decision trees is to achieve better results than those given by the mean and standard deviation. Therefore, Table 5.2 provides a useful reference point when viewing the tables below.

Table 5.3 shows the tree scores for the accent and falling boundary event types using all of the features discussed above. There were very few rising boundary events in this dataset, which results in use of the mean value instead of a predicted value. The accent and falling RMSE values shown are certainly lower than their respective standard deviation values, which loosely suggests that the values which are predicted using these trees resemble the original values against which they are tested.

	start F0	amplitude	duration	tilt	peak position
a	**9.89**/0.61	**11.28**/0.40	**0.05**/0.40	**0.49**/0.33	**0.06**/0.42
fb	**6.65**/0.55	**11.01**/0.40	**0.06**/0.63	**0.37**/0.43	**0.07**/0.5

Table 5.3: Database KDT:**RMSE**/Correlation scores for accent and falling boundary trees

A tree-by-tree analysis of these two event types provides the sort of insights which are useful for intonation generation research. It is immediately obvious from Table 5.3 that the *start F0* parameter is the best predicted parameter for accents. The correlation score suggests that the distribution of parameter values follows the correct pattern (in this case, higher values at the beginning of phrases, lower values as the phrases continue). The RMSE score for *start F0* shows that in addition to the correct pattern, the predicted values are also similar to their targets. The *amplitude* scores show that the magnitude of accents which will be generated is somewhat large, averaging 10Hz. In a rise fall accent (*tilt* around 0.0), this *amplitude* average error would correspond to a 5Hz error in each of the rise and fall portions of the generated accent. In a pure rise or fall accent, the error would be the full 10Hz. In this speaker's range (mean 127Hz, sd 42Hz) such magnitude differences are minor. At the top of the range, a 10Hz difference would be nearly imperceptible. At the bottom of the range, the difference might be noticeable, but it is difficult to determine whether any difference in meaning would result. Further research which links these objective measures with perception would be very useful in assisting in such a determination.

The most interesting poor result for the accent trees is in the *tilt* prediction. The correlation score suggests that the typical trimodal distribution for *tilt* values (to give basic rise, fall, and rise-fall shapes) is not being retained in the predicted data. The RMSE value corroborates this conclusion, as an error of 0.5 on the *tilt* continuum could mean the difference between a fall and a rise-fall shape (see Figure 3.4, page 78). These scores for the *tilt* parameter suggest that the tree is not partitioning the data in a way that follows the typical *tilt* distribution. A survey of *tilt* prediction trees built in different conditions shows that the predicted values fall in the range of 0.3 to -0.6, which essentially negates the pure rise and pure fall types of accents. Only a continuum of rise-fall accents is predicted. While the features used to predict *tilt* values are successful in many cases (in that a range of values along the continuum is predicted), they are not asking the right questions to capture the full distribution. The *tilt* values at the ends of the scale are not being modelled adequately. One probable reason for the difficulty in modelling the ends of the tri-modal distribution is that, as Figure 5.5 shows there are considerably more accents in the middle of the range than on the ends. However, the figure also shows that the ends of the range are distinct from the middle range values. One possible reason for the inability of the decision tree to successfully capture the value distribution is that more data is needed. The uneven representation in the data could mean that much more data is required for all parts of the tri-modal distribution to be accurately modelled.

Another strong possibility for the failure to capture the full distribution is that the choice of rise-only or fall-only accents is related to some construct

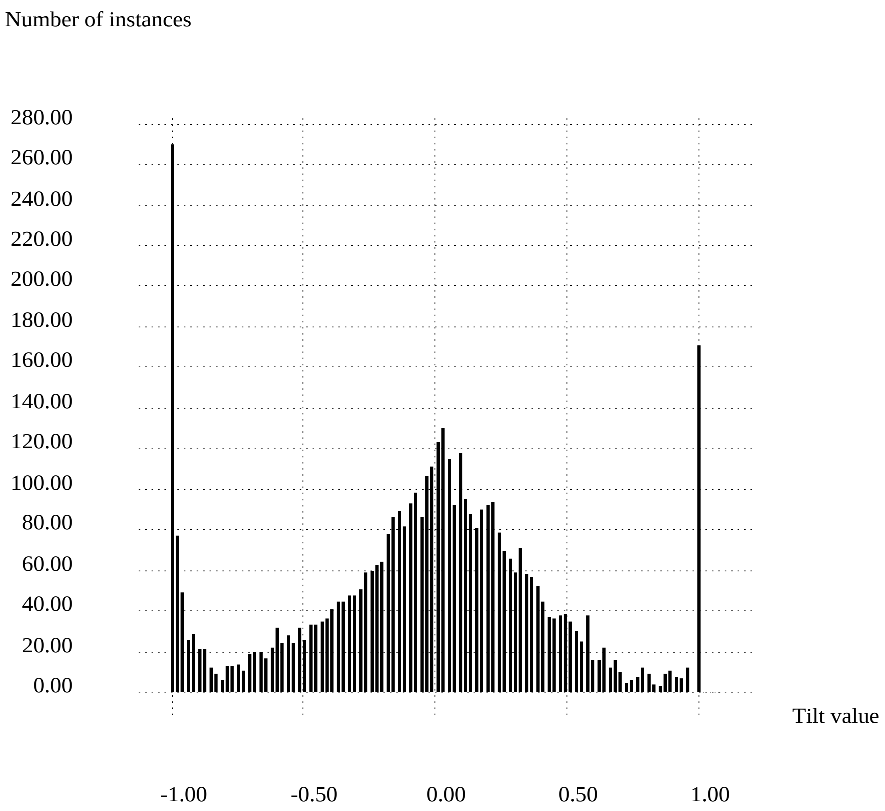


Figure 5.5: Distribution of accent *tilt* parameter values

which is not part of the extracted feature set. The choice could be related to the wider discourse or semantics. Alternatively, the ends of the range could represent purposeful variation in intonation for minimizing monotony. As these areas are not covered by the feature set used for the experiments, it is impossible to say whether their addition would result in a better distribution of the predicted values.

The falling boundary trees also provide some interesting results. The *duration* prediction tree shows an RMSE of about 50ms, or one quarter the mean falling boundary length. Another way to look at this value is that it is similar to the mean syllable onset length for KDT (50ms) and less than the mean coda and syllable durations (143ms, 193ms). Thus, the *duration* prediction both follows the desired value distribution and keeps the average error to within a small portion of a syllable length. Also interesting in the timing of falling boundaries is the *peak position* prediction. Here, as with accents, the correlation is not particularly good (0.5). The error in peak timing is larger than that for *duration*. This is not surprising, as a falling boundary need not preclude the option of a short rise in F0 before the fall takes place. The interest in this tree is that a 70ms error in peak timing could, in a categorical framework like ToBI, change the category of the event. Thus, while the error is only a fraction of syllable length, it is potentially audible.

It is easy to suppose, given the above examples, that when one score is good, both are good (or equally one bad score means two bad scores). An example of where this is not the case is in the prediction of *start F0* for silences (the final F0 value for a section of speech). The correlation score for this tree is 0.97, suggesting that the distribution is being very effectively modelled. Thus, the predicted value is low for speech ending in a falling boundary, high for speech ending in a rising boundary, and in the middle of the possible F0 range where no intonational boundary occurs at the end of speech (a mid-phrase pause, for example). However, the RMSE score for this tree is 10.97. Therefore, while the basic range is correct, the error is still in the region of 10Hz. Thus, while generally RMSE and correlation reflect the success of the trees in the same manner (both are good or both are bad), the relationship does not always hold. Therefore, for the rest of this section, the correlation score is only presented where RMSE is not providing an adequate picture.

The previous examples highlight once again the benefits of evaluating individual aspects of intonation generation. As the tables and relevant discussion below will show, this individual assessment allows one to examine the effects of particular features and feature sets on different aspects of intonation.

5.3.2 The Contribution of Sub-Syllable Features

In order to determine what effect, if any, the sub-syllable features have on Tilt parameter prediction, the decision trees were built using the same methods listed above, but excluding all sub-syllable features from the tree-growing algorithm. Table 5.4 shows the scores for these trees for the KDT database.

	start F0	amplitude	duration	tilt	peak position
a	10.22Hz	11.79Hz	0.05sec	0.50	0.07sec
fb	6.85Hz	11.36Hz	0.06sec	0.38	0.07sec

Table 5.4: RMSE scores for trees with no sub-syllable information

In comparison with Table 5.3, the RMSE values are the same or slightly higher than those which resulted from the full feature-set being available. The obvious conclusion to draw from this comparison is that the sub-syllable features are only helping the prediction in some cases. To test this conclusion, a similar comparison was made for speaker FHL, as shown in Table 5.5. The FHL RMSE values support the conclusion that the sub-syllable features are being used, but do not greatly improve the trees.

	start F0	amplitude	duration	tilt	peak position
with	18.41Hz	23.65Hz	0.07sec	0.54	0.09sec
without	19.26Hz	24.45Hz	0.08sec	0.54	0.1sec

Table 5.5: RMSE scores for FHL accent (a) trees with sub-syllable and without sub-syllable information

The small RMSE changes tend to occur more in the F0 domain (*start F0* and *amplitude*) than in the time domain (*duration* and *peak position*) parameters. Interestingly, the correlation values tell a different story. In the case of FHL, the correlation for the peak position parameter is 0.46 with sub-syllable information, and 0.33 without. KDT shows similar results with values of 0.42 and 0.29 (with and without sub-syllable features). It is also worth noting that these changes in accent peak position correlation scores go against the tendency of most other trees for little or no effect by the sub-syllable features (see Appendix B for full result tables). However, as can be seen in the tables above, it is clear that these higher correlations are not likely to result in audibly different peak locations (a 10ms difference in RMSE of *peak position*). Therefore, it may be that the timing is already so bad that the higher correlation will not significantly affect the outcome. Or, more kindly, the improvements brought about by including sub-syllable features are still overshadowed by the many difficulties faced in modelling peak timing in general. The only real way to tell if these small differences can result in noticeable improvements in the F0 contour is to compare the generated F0 contours with those which they are meant to reproduce.

5.3.3 Fundamental Frequency Comparisons

Tables 5.7 through 5.10 show how the results of the intonation generation method described in the previous sections compare with the original intonation of the databases. Each of the results shows a target and at least one experimental result. The targets result from comparing the smoothed F0 contours from which the original Tilt parameters are extracted with the F0 contours generated by the original Tilt parameters. In other words, this error would be given if the Tilt parameter prediction were 100% correct on all counts. Because the databases represent different voice types, dialects, and genders, it has been useful to consider the the correlation, as well as the RMSE results in terms of their relation to the standard deviation of F0 in order to compare them with each other. Thus, a 34Hz RMSE may *look* like a large error, but if it is achieved on a voice with a large standard deviation (e.g. 53Hz), the error is relatively low. For the female speakers, the target RMSE score is roughly one-third of the standard deviation of F0. For the male speaker, the target RMSE is approximately one-seventh of the standard deviation (see table 5.6).

Speaker	Mean F0	σ F0	Target RMSE	Target Correlation
F2B	163.5	42.2	14.5	0.93
KDT	126.9	27.9	3.9	0.94
FHL	210.5	31.8	12.5	0.87

Table 5.6: F0 and Target Value Information for Three Speakers

As table 5.6 shows, the three voices modelled cover a wide range of mean F0 as well as pitch ranges. For any single method to be successful in a useful application, it must work adequately on a variety of speaker types. These three speakers offer a fair test of that ability.

Table 5.7 shows comparison results for the F2B database. Intonation contours were generated for F2B under three different conditions: 1) using mean and standard deviation parameter values, rather than predicting parameter values using a decision tree, 2) training the decision trees using the methods described above, and 3) manually optimizing the decision trees. The manual optimization process involved examining the features which were automatically chosen in the training process, and attempting to re-build the tree without each feature in turn. This optimization process aims to reduce one possible disadvantage of the greedy algorithm being used: in some cases, a single feature could perform well in dividing the data, but prevent other potentially relevant features from playing a part. As Table 5.7 shows, the manual optimization did give some improvement, but not enough to warrant such optimization for all databases.

The target figures for F2B are somewhat daunting. A correlation of 93% results in an audibly identical F0 contour for most utterances. As discussed above, mean F0 and standard deviation are relatively simple to acquire for numerical data. A lower acceptability limit for F0 contour comparison was determined by replacing each tree for the Tilt parameters with a single leaf containing the mean and standard deviation for all data points which would be used to build that tree. An unexpected problem with the use of the mean values for each parameter to generate F0 contours was uncovered in determining this lower limit. The parameter-to-F0 process shows that the Tilt parameters are not all independent of the text associated with an event. Remembering the 201ms *duration* mean value from Table 5.2, it is conceivable that an anomoly would occur if this peak location were assigned to an utterance final syllable of short duration. In fact, such over-shoot of event timing causes a failure in the translation which results in an error message, rather than a contour, being produced. As table 5.7 shows, the result is a RMS error of 39Hz and a correlation of 0.36 for the roughly 50% of paragraphs which did not fail to process. These failures are most attributable to poor timing prediction, which resulted in intonation events being predicted outside the bounds of the utterance upon which the contour was imposed. Errors of this kind caused a failure within the synthesis process. In other words, half of the Tilt descriptions generated by predicting mean values were so bad that they resulted in nonsense and failed to result in an F0, while the other half were only moving in the right direction a third of the time, albeit with an F0 value that was likely to be within the right pitch range.

	Target F0	M/STD F0	Base F0	Hand-tuned F0
RMSE	14.5	39.0	34.7	34.3
Correlation	0.93	0.36	0.58	0.6
% failed	0	50	<10	<10

Table 5.7: F0 Comparison Results for F2B

Now that a suitable range of results is delimited, the discussion of the rest of the results is possible. A base result, arrived at using the methods described in section 5.1 approximates the results already shown in table 3.4. After hand-optimizing the feature set, the results improve slightly to approximate Dusterhoff and Black's results even more closely. While these results are not very close to the target values, they are considerably better than the lower limit.

As table 5.8 shows, it is easier to predict the intonation of a database when it is for mostly declarative, isolated sentences that are spoken by a male speaker with little F0 movement. The KDT results are interesting in a number of areas. First, the target for KDT is similar to the target for F2B, in terms of correlation. Therefore, we know that the Tilt descriptions and the related tools handle the male and female voices equally well.

	Target F0	M/STD F0	Predicted F0
RMSE	3.9	11.6	9.1
Correlation	0.94	0.45	0.74

Table 5.8: F0 Comparison results for KDT (isolated sentences)

The difference in RMSE targets reflects the difference in the speakers' pitch ranges. For KDT, who has less natural variation in F0, it is necessary to prevent large variations in the generated contours, as they will likely sound out of place. This restriction was not true of F2B, who had a naturally larger range of possible F0 values. Therefore, one might assume that using the mean values instead of full trees would provide a better result for KDT than it did for F2B. The lower limit does appear to be higher for KDT, in terms of correlation, than it was for F2B. However, when the RMSE portion of the evaluation is taken into account, one sees that, while the mean may provide the correct range and basic shape of contour more readily for KDT than for F2B (generally declining F0), the contour itself is nowhere near the correct F0 values. In fact, reviewing table 5.6 reveals that the maximum RMSE is over twice the standard deviation of KDT's F0 values. Thus, while the contour generated using mean, rather than predicted parameter values, may look similar to the original, the F0 scaling is unlikely to be correct. One would certainly expect that a better result could be achieved with a more complex model. Table 5.8 reveals that the more complex model does in fact perform much better than than the simple mean value model.

Two other comparisons were performed on the KDT database which provide insight into the use of sub-syllabic features in intonation generation. As mentioned above, small differences in individual tree performance should be

	Avg. RMSE	Avg. Correlation
No Sub-syllabic features	9.2	0.72
Only Sub-syllabic features	11.3	0.55
All Features	9.1	0.74

Table 5.9: F0 Comparison of based on trees built 1) with no sub-syllabic features, 2) with only sub-syllabic features, and 3) with all features allowed

examined within the context of the F0 contour as a whole. Trees were built for the KDT database which included only sub-syllabic features or excluded all sub-syllabic features. These trees resulted in intonation contours which compared with the originals as shown in Table 5.9.

As Table 5.9 shows, the small improvements in the tree scores which resulted from the inclusion of sub-syllabic features is reflected in small improvements in the overall F0 comparisons. Interestingly, the trees which include only sub-syllabic features produce moderate results, suggesting that some of these features may be accounting for the same data as some of the other features.

Table 5.10 shows that some databases are more difficult than others to model. The target correlation is noticeably lower than that of the other two speakers. This suggests that perhaps the labels are not of as high a quality, or perhaps that there is more movement in the non-event (connection) portions of the original F0, lowering the correlation score even if the events are accurately regenerated. Regardless of the cause for the lower target, it is important to recognize that a lower target will likely correspond to a lower F0 comparison. Therefore, the results for FHL, while lower than for F2B and KDT, are comparable to F2B's results. The resulting RMSE is less than twice the target (as compared with almost 2.5 times for F2B) and only slightly more than one-third σ F0. While these comparisons do not have any inherent meaning in themselves, they show that the FHL results are in the same range of success as the F2B results, while remaining lower than the KDT results.

	Target F0	Predicted F0
RMSE	12.5	21.1
Correlation	0.87	0.53

Table 5.10: F0 Comparison results for FHL

Finally, in order to place this work in the context of recent, similar research, Table 5.11 shows three previous studies, all using F2B[2], and the three studies discussed in this section.

Study	RMSE	Correlation
Dusterhoff & Black	32.5	0.60
Ross & Ostendorf	33	Not Given
Black & Hunt	34.8	0.62
F2B	34.3	0.60
KDT	9.1	0.74
FHL	21.1	0.53

Table 5.11: Comparison of F0 generation research

While the RMSE results are only comparable across the studies using F2B, research in F0 evaluation suggests that correlation can be used to compare all of these studies [Her98]. The F2B results are all similar, suggesting that the methods used in this thesis result in valid and reasonably successful intonation generation. The difference between Dusterhoff and Black and the F2B research carried out in this thesis is attributable to a different set of F0

[2]Many changes to the required systems and feature architecture occurred between the work with F2B found in Dusterhoff and Black and the availability of KDT and FHL. It was necessary to re-build the F2B models so that the same system was used for F2B, KDT, and FHL. Therefore, Dusterhoff and Black's results are comparable to the F2B results in this chapter, but are one step removed from a valid comparison with KDT and FHL.

contours. In order to better understand the role of sub-syllabic features, the research in this thesis used smaller contour smoothing windows than were used in Dusterhoff and Black (see section 4.2.3 for window details). Otherwise, the only difference in method is that in Dusterhoff and Black, the feature selection was entirely by hand, which explains the use of the manually optimized results for comparison. Because the level of smoothing of the F0 contours can explain a difference in RMSE of almost 2Hz, it would be unwise to rank the four F2B results, given their similarity. The results, viewed as a whole, suggest that a small body of very consistent data (KDT) is easier to model than a large body of data with more variation. As is expected, a small database with a lot of variation (FHL) is the most difficult to model.

5.3.4 Context Features

An analysis of how the features and feature classes are used to achieve these results shows how this research can be used to advance the field. The parameter prediction trees for F2B provide the clearest picture of the three databases, as it has the most varied content both phonetically (important when using segmental information) and intonationally (there are enough intonation events of each type for a clear picture). Therefore, this analysis is restricted to the F2B database.

As discussed in section 5.2.2, the features used for parameter prediction are divided into five broad classes. The first class is local lexical stress. The second class describes the phrasal position of a given syllable. The third category contains syllabic constituency information. The fourth class is intonation information on local syllables. The fifth class is intonation information along an intonation tier. These classes are included based on expectations from experimental linguistics literature, as reviewed in chapters

3 and 2.

The lexical stress information is included in the feature set to help model "stress clash" (basically, where an accent peak is moved left in the presence of a following stressed syllable). This phenomenon is widely reported and accepted, leading to the inclusion of similar features by Ross ([Ros94]) and Black and Hunt ([BH96]). The features in this class which should play the largest role are the right-hand context features. I expect that the accent *peak position* tree would make use of these features, and that perhaps the "afb" and "fb" trees will use one of the left-hand context features where syllable lengthening and tonal crowding occur (introduced in section 2.4).

The prosodic phrasing class is likely to contain the most used features. This class should contribute to modelling tonal crowding, phrase boundary effects (e.g. lengthening, pitch resets). One important area where this class is expected to be important is in modelling the frequency of events (*start F0*). As Clark [Cla99] shows, the phrase initial events start consistently higher than phrase medial events. He also shows that initial events in F2B have a greater magnitude, which is represented by the *amplitude* parameter. *Amplitude* parameter trees should include members of this feature class. Minimally, the accent *amplitude* tree should incorporate information about where in a phrase an accent is.

The sub-syllable features, as discussed in section 5.3.2 and Chapter 2 should be useful in predicting *peak position* parameter values. They may also help in *amplitude* prediction, provided that vowel-intrinsic pitch is perceptible in this database. Because the sub-syllable features relate to the content and duration of sub-syllable constituents, it is also likely that these features will appear in *duration* trees.

The fourth class, which views intonation events through a five-syllable window, is expected to contribute to modelling of tonal crowding, event magnitude, and downstep. For example, an accent which contains an accent within the right and left context for this feature class is probably within a downstepped sequence, and will consequently have a lower *amplitude* than it would have if there were no accent events within the window.

The final class, which represents intonation events in sequence without reference to any other context, is expected to play a very important role in *start F0* prediction for all events. This feature set provides information about the surrounding events, regardless of their distance from the event whose parameters are being predicted. If, for example, an accent is preceded by a falling boundary, it will probably start higher than if preceded by an accent (given Clark's findings).

The features are expected to contribute to some, but not all trees. All of the features are included in the model building process, for completeness. As discussed above, each class was included in the process for a specific reason. In some cases, the features had less of a role than was expected. Some features were more useful than expected. Because of the potential for inter-feature noise, as showed by the necessity for a feature-reduction algorithm, it is unwise to state that any single feature accounts for any specific phenomenon. Additionally, it is probable that the features will contribute differently given a different database. The claims and discussion below are therefore approached as generalizations of the effects seen on the F2B database.

The lexical stress information has played a very small role in accent parameter prediction and a slightly larger role in boundary parameter prediction. For accents, the lexical stress of the syllable following the accented syllable is used for *peak position* prediction. This is presumably to account for "stress-clash" conditions. For the boundaries (falling, rising, with, or without accents), the local stress features played a role in the *duration* and *tilt* trees. The preceding stress information was important here, where it may be related to syllable lengthening phenomena. Support for this supposition comes from the syllable and sub-syllable features which play an important role in all of the timing domain parameters of the various boundary types, as is discussed below.

Predictably, the phrasal positioning features were important in almost all of the trees. In the F2B database, the location of a pitch event within a phrase is very closely related to the pitch range of that event. These features can also provide an approximation of whether an event is in "nuclear" position, and whether it is likely to be near the end of a major or minor phrase. Such information would be helpful in an event's magnitude and basic shape. Interestingly, only one of these features is used to predict *peak position*, and that only in the falling boundary tree. This feature appears to be used to determine whether the falling boundary is internal or final, in the context of a larger prosodic constituent.

The reason the third class, containing information about the syllable constituents, was included in this research at all is that claims in the literature (see Chapter 3) suggest that *peak position* prediction would be improved. While the improvement is not great there is no doubt that the features are being used, as they are present in several of the decision trees. The accent tree includes vowel height classification and onset classification information. Each of the boundary trees include at least two of the nine available features in this class. In addition to the *peak position* trees, the *duration* and *tilt* trees also use the onset and rhyme features (categorization and duration).

The use of these features in the *duration* and *tilt* trees supports the suggestion above that the sub-syllable features may be related to final lengthening phenomena.

The fourth category, which provides information about intonation events within the five-syllable window, proved useful in most trees. For example, the accent *start F0* tree relies on information about whether there is an accent or boundary associated with the prior syllable. This feature use may relate to the lowering of fundamental frequency through an utterance. It could also relate to downstep, where a succession of closely placed accents may be a sign of the sort of environments where downstep occurs. Alternatively, the lack of space between two accents may prevent the starting frequency of the accent from having lowered excessively. In another example, the falling boundary *peak position* tree looks for accents on earlier syllables and the syllable associated with the boundary. Such feature use is not surprising, but does not follow a specific enough pattern to make any strong claims.

The final feature class, which views intonation labels on their own tier without reference to the syllable tier, is generally the most important class. Only two trees (accent *peak position* and *amplitude*) do not select a feature from this class. In the other trees, at least one feature from this class is included in the first three features selected. These features are especially important at the edges of phrases. For example, an accent preceded by a falling boundary will be involved in a pitch reset, where an accent preceded by silence will fall into that class of higher scaled events discussed by Clark [Cla99]. Similarly, there is an expectation that falling boundaries which are followed by accents (and are therefore medial in a larger intonational unit) will fall less than those followed by silence.

The distribution of features in the prediction trees does not throw up any surprises. In general, they follow the expectations which the experimental linguistic literature predicts. The importance of the fifth class, as shown by the location of the features used in the top three features selected, is predictable given most intonation literature written in the last twenty years. One surprise, though, is the number of modelling and synthesis techniques described in Chapter 3 which neglect this approach to intonation.

It is important to remember that certain features being used for certain predictions does not mean that the resulting values are particularly good. The results discussed above show that while the features used in this study are providing important improvements, there are many improvements yet to be made. Future research into new features and methods of presenting the feature data could prove useful in getting more out of the advances made by this work.

5.4 Building Models from Auto-labelled Data

Because the F2B database is also large enough to be useful for the auto-labelling work discussed in Chapter 4, one experiment was undertaken to build intonation synthesis models using the methods described in this chapter using automatic intonation labels for F2B. In order to examine the potential for future work in this area, the best of all possible HMM sets and external constraints was used to automatically label the database for intonation. These conditions are discussed in Chapter 4.

In addition to the problems with intonation parameter modelling already discussed, this experiment also required a step which associates the automatically derived labels with the syllable labels. The method used for manually labelled data requires the human labeller to decide which syllable each intonation event should be associated with (i.e. the perceptually accented syllable). Rather than emulate this process, which is an extremely difficult task, it is possible to achieve an adequate simulation of this process by associating each intonation event to the syllable during which the peak occurs. This type of linking is systematic, if not theoretically motivated.

Another difficulty faced by using the automatic labels is that the quality is not easy to judge. Therefore, two initial tests were undertaken to determine whether using the auto-labels was feasible. As Chapter 4 shows, a quantitative comparison of event types shows whether the distribution of auto-labelled events is similar to the manually labelled events. Table 4.8 shows that, for the F2B database, with the best labeller, the distribution of event types is similar. The second test examines the means and standard deviations of the Tilt parameter values of the auto-labels. This test is designed to show whether the properties of the auto-labelled events which will be modelled for synthesis are similar to the properties of the manually labelled events.

Table 5.12 shows the means and standard deviations of the Tilt parameters in the manual and automatic labels. This table shows that the two label sets are reasonably similar, but there are some interesting differences.

The auto-label *start F0* values are very similar to those for the manual labels. This suggests that the events for the two label sets fall in the same basic places from utterance to utterance. However, the *amplitude, duration*, and *tilt* values show some very large differences. The mean values for *amplitude* show that the auto-labelled events fall more in a “categorical” range

	start F0	amplitude	duration	tilt
a	164.5/42.1/	71.9/49.5	0.307/0.069	0.040/0.480
a	*166.9/40.5*	*80.8/53.9*	*0.347/0.084*	*-0.040/0.466*
arb	134.6/32.0	55.1/28.3	0.380/0.073	0.377/0.579
arb	*134.6/30.3*	*58.5/29.9*	*0.413/0.061*	*0.266/0.616*
afb	144.4/31.8	84.4/46.5	0.362/0.057	-0.206/0.342
afb	*146.2/30.1*	*77.7/41.1*	*0.396/0.062*	*-0.292/0.380*
rb	145.7/39.1	45.9/31.5	0.243/0.084	0.302/0.563
rb	*158.8/42.5*	*49.2/30.4*	*0.343/0.084*	*0.020/0.724*
fb	145.9/34.2	63.3/52.5	0.204/0.073	-0.088/0.544
fb	*157.0/34.8*	*52.2/36.1*	*0.323/0.091*	*-0.631/0.651*

Table 5.12: Comparison of Mean and Standard Deviation of event parameters (automatic event details in italics)

(higher for accents and rising boundaries, lower for the falling boundaries). The *duration* means are longer in each case for the auto-labels. Again, for all event types, the *tilt* means are lower for the auto-labels, and in most cases the standard deviations are larger. These comparisons lead to the conclusion that the automatic labelled events are examples of typical instances of each event type. Certain variations of each event type are picked up (e.g. falling boundaries with minimal early rise). One event type need not be restricted to a single typical representation (e.g. the *rb* values suggest that the auto-labeller is picking flattish continuation rises and sharp final rises, but not those in between).

The expectation for the synthetic models developed from this data is that there will be less natural variation accounted for by the parameter prediction trees, but much of the variation which is retained in the automatic labels will be accounted for. In terms of the evaluation techniques used on the intonation generation research already discussed, this means that the assessment of the trees will show that the trees successfully model the appropriate parameter values, while the resulting F0 contour will be less like the smoothed original than the contours generated from manual label models.

Tables 5.13 and 5.14 show the tree scores for the two label sets. The automatic label trees consistently have higher correlations and generally have lower RMSE scores, as was predicted. Table 5.15 shows how these tree scores translate into F0 contours. The average scores for the contours generated from the manual label models, using the exact same validation, training, and blind test sets (held out for very occasional use), are 37.09 (RMSE) and 0.564 (correlation). These scores are better than those achieved by the models derived from automatic labels, but they are in the same range, suggesting that this process is worthy of future research.

5.5 Perception of Synthetic F0

This section places the assessment methods which have been used in the previous sections into the wider context of intonation evaluation. So far, only objective methods of evaluating the synthetic intonation contours have been discussed. This section reviews two experiments in which the contours generated using the methods described in this chapter have been subjectively evaluated. First, we re-visit some basic concepts of intonation evaluation, as introduced in section 3.1. Following this review, we discuss an experiment which subjectively assessed synthetic intonation developed using the basic methodology described above. Finally, an experiment which attempts to link objective and subjective assessment methods is presented.

Subjective tests require human subjects, who judge the synthetic intonation in relation to some standard. Synthetic intonation for an utterance may also be compared against natural intonation for the same utterance us-

	start F0	amplitude	duration	tilt
a	31.59/0.62	49.89/0.326	0.074/0.429	0.413/0.406
arb	26.75/0.567	18.23/0.359	0.08/0.609	0.527/0.360
afb	19.32/0.695	32.75/0.528	0.053/0.488	0.266/0.776
rb	27.09/0.76	26.75/0.387	0.084/0.379	0.476/0.774
fb	25.34/0.59	29.99/0.447	0.081/0.418	0.459/0.760

Table 5.13: Decision tree evaluation for automatic intonation labels (RMSE/Correlation)

	start F0	amplitude	duration	tilt
a	33.04/0.618	46.58/0.25	0.057/0.519	0.427/0.354
arb	30.03/0.540	22.76/0.397	0.118/0.113	0.531/0.502
afb	28.63/0.471	39.82/0.484	0.06/0.461	0.278/0.447
rb	26.18/0.499	28.42/0.485	0.057/0.733	0.489/0.380
fb	25.66/0.494	43.59/0.345	0.06/0.524	0.486/0.400

Table 5.14: Decision tree evaluation for manual intonation labels (RMSE/Correlation)

Files	Average RMSE	Average Correlation	No. of files
All Auto-labels	39.23	0.5271	108
Blind Auto-labels	40.32	0.4807	9
Validation Auto-labels	38.16	0.5545	30
Train Auto-labels	39.57	0.5212	69
All Manual Labels	*37.09*	*0.564*	*108*

Table 5.15: Comparison of F0 contours generated from models developed from automatically derived intonation labels and the smoothed original F0 contour for the same utterance. (Manual label figures in italics for comparison)

ing objective measurement techniques. Two common subjective assessment techniques used with synthesized intonation are pairwise comparison and acceptability ranking. Pairwise comparison tests, such as those used by de Pijper [dP83], present subjects with paired synthetic utterances. The subjects are asked to judge which utterance sounds more natural, or which one sounds like natural speech, as opposed to synthetic speech. The basic success criterion is when synthetic and natural intonation are often judged as equally natural. A second common method for subjectively assessing synthetic intonation is acceptability ranking ([vBP90]). Subjects are presented with synthesized utterances, which have either synthetic or natural intonation, and are asked to rank the utterance on a scale of how natural, or acceptable the utterance sounds. An example of such a test, described in detail in [SMD+98], is discussed in section 5.5.1.

Objective evaluations do not require human subjects, but attempt to assess intonation in a way that relates to what subjects would perceive. As mentioned above, the current objective evaluation of synthetic intonation involves use of Root Mean Squared Error and Correlation. These two measurement techniques have been used for a number of intonation evaluation (e.g. [Ros94], [DBT99]), and were shown by Hermes [Her98] to be better reflections of perceived intonation differences than other available metrics. The standard RMSE provides a measure of how close two contours are to each other. The correlation coefficient measures the degree to which the variables are linearly related. Thus, a high correlation coefficient shows a close linear relation (which should be the case with two similar F0 contours from the same utterance), while a low coefficient shows that the linear relationship is not close: that the two lines are diverging regularly.

5.5.1 Acceptability Judgements

As mentioned above, a common way of evaluating synthetic intonation is to solicit opinions from subjects about the naturalness of a piece of speech generated with a synthetic intonation contour. One such test has been used which includes intonation models built as described in this chapter. Syrdal et al [SMD+98] assesses three different intonation modelling techniques by asking subjects to rate 144 test utterances (12 texts by 12 synthesis conditions) on a five-point scale (5=excellent, 1=bad). The synthesis conditions varied synthesizer (two options) and intonation model (six options, including natural intonation). Forty-three subjects were used in the test. All were adult native speakers of American English (the test language). The test lasted approximately one hour, including instructions and time to practice using the scale and touch screen with which the rating was recorded. The six intonation conditions were: 1) natural intonation, 2) a rule-based model [JMD], 3) an implementation of the Tilt model which is based on the methods described in this chapter, and more specifically follows [DB97], and 4) three variations of the PaIntE model (Parametric Intonation Event, [MC98]) a vector quantized intonation model similar to Tilt, which predicts its parameters as a vector, rather than individually.

Unsurprisingly, the natural prosody model received the highest ratings (3.6260) of the six models. A variation of PaIntE scored second best (3.3430), with two other variations on PaIntE and the rule-based model scoring in the 3.22 to 3.24 range. The Tilt condition scored lowest, (3.143). However, while there is clearly a difference of ratings, and natural intonation is obviously better than any of the other models, Syrdal *et al* do not provide any evidence of significance between models. Therefore, while there is significance among the models as a group, it is unclear if any synthetic intonation model is significantly better than any other.

This evaluation method has provided some interesting results, but nothing exceptionally useful. It is obvious to the researchers involved that the synthetic intonation is less natural than natural intonation. However, a useful result is not one which states that natural speech is significantly more natural than synthetic speech, but one which is able to highlight specific failings and successes of the synthetic speech. If the test could have shown how the intonation produced by the best PaIntE variation is better than the other models, then researchers would be able target specific areas of deficiency. Ideally, one wishes to understand what aspects of each model are successful and what aspects require further work. For example, had this evaluation shown that the alignment of the peak in the Tilt-synthesized contours causes unnatural sounding intonation, then this experiment would have been developmentally useful. The objective evaluations of the individual decision trees, as discussed in section 5.3.1, provide the type of specific quality assesment that researchers can use. Subjective experiments which mirror this lower-level evaluation could be quite useful. Experiments such as this, which give only gross results, without specific comparisons of different aspects of intonation, provide a reasonable overview of completed systems, but do very little for ongoing research.

5.5.2 Linking Subjective and Objective Assessment

Because it can be very difficult to obtain the specific comparisons one may need for research, it is desirable to have an objective measure which is closely related to subjective evaluation. Sections 5.3.3 and 5.4 discuss currently used objective methods for evaluating whole intonation contours: RMSE and Pearson's Correlation. These sections show how different experimental conditions result in improvements in the comparisons which use these metrics.

However, it is unclear how large an improvement in either metric must be before it is reflected in perceptual improvement. It is also unclear how detailed an analysis these metrics provide. Therefore, two other metrics are presented here, both of which are similar to a basic RMSE measurement. These metrics were developed with Clark [CD99], and measure F0 using the Hertz scale. All four metrics are compared to a perceptual examination in order to relate the objective measures to perceived differences between contours. An introduction to the new metrics as they compare with RMSE is followed by a discussion of the perceptual experiment.

The basic RMSE measures the distance between two contours on the time axis, such that the distance being measured (the dotted line) at regular (e.g. 10ms) points is perpendicular to the time axis, regardless of the F0 shape (Figure 5.6a).

- ***The Tangential Estimation Method***

The tangential estimation method computes an RMSE measurement between contours at regular intervals. The difference between this method and the basic RMSE method is that the measurement takes place on a line normal to the tangent of the contour at each interval on the reference contour. The assumption underlying this new approach is that a similar rate of F0 change between contours will be reflected by including an aspect of the time domain with the frequency domain. By removing the restriction of measuring distances perpendicular to the time axis, time and frequency can be measured using a single metric.

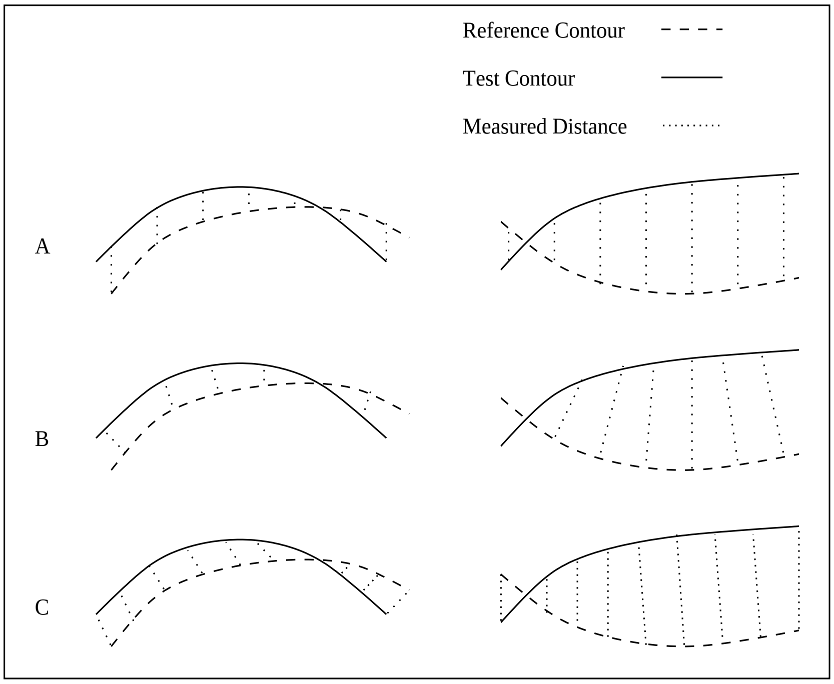


Figure 5.6: Three Objective Evaluation Metrics: A) RMSE, B) Tangential Method, C) Warping Method

$$d^2 = (r_x - c_x)^2 + (r_y - c_y)^2. \tag{5.1}$$

where:

$$\begin{aligned} c_x &= \frac{1}{2}\frac{[f_1(t+1) - f_1(t)][f_1(t+1) + f_1(t) - 2f_2(t)] + 1}{[f_1(t+1) - f_1(t)][f_2(t+1) - f_2(t)] + 1} \\ c_y &= \frac{1}{2}\frac{f_2(t+1)[f_1(t+1)^2 - f_1(t)^2 + 1]}{[f_2(t+1) - f_2(t)][f_1(t+1) - f_1(t)] + 1} \\ &\quad - \frac{1}{2}\frac{f_2(t)[f_1(t+1)^2 - f_1(t)^2 - 1]}{[f_2(t+1) - f_2(t)][f_1(t+1) - f_1(t)] + 1} \end{aligned} \tag{5.2}$$

and:

$$\begin{aligned} r_x &= 0.5 \\ r_y &= \frac{1}{2}[f_1(t) + f_1(t+1)] \end{aligned} \tag{5.3}$$

The basic RMSE typically takes measurements on a frame-by-frame basis, (e.g. calculating distances every 10ms). The tangential method also works on a frame-by-frame basis. The difference between the two is not limited to the angle of the measurement line (dotted on Figure 5.6). The reference points on the two contours which make up the ends of the measurement line are also different, which is a result of the different angle of measurement. On the reference contour (dashed on Figure 5.6), the midpoint between frames is the point from which the measurement line begins. On the test contour (solid on Figure 5.6), the intersection between the contour and the measurement line forms the end point of the measurement line. The measured distances are then combined in the same way as the basic RMSE metric.

Figure 5.7 shows the process for a single measurement between contours. If $f_1(t)$ is the reference line, and $f_2(t)$ the test line, taken in the frame from times t to $t+1$, then the squared distance (d^2) is calculated as shown in

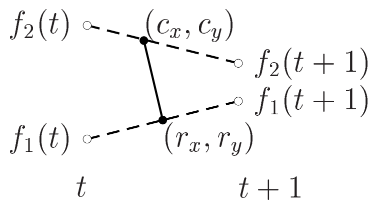


Figure 5.7: Computing the tangental metric

Equation 5.1 (by translating the time axis to the origin).[3]

One difficulty with this metric is that if the reference and test contours are swapped, the measurement value will also change. There are two ways to combat this difficulty. One is to take a measurement, swap the two contours (so that the reference from the first measurement becomes the test in the second), and average the scores. Another way is to always make sure that one type of contour (e.g. the natural contour) is the reference. The second method was used in the tests carried out here.

- ***The Warping Method***

The warping method is an attempt to measure the area between two contours. This area is the difference between the contours. As shown in Figure 5.6, the tangential method can leave some areas unmeasured, when there is no intersection between the test contour and the line normal to the reference contour. Basic RMSE has the same problem if there is no F0 at a point in one contour where there is in the other. Rather than be constrained in any way by the time axis, the warping method begins with the assumption that equivalent contours will begin and end at the same times. If they do not, it is likely that the overshoot will not be caught by either of the other methods.

[3]Figure 5.7 and related equations taken from [CD99].

The warping method ensures that such F0 mis-alignments are punished in the evaluation.

Unlike the other metrics, the warping method is designed for use on small sections of contours. Label files can be supplied which allow the evaluation to take place on pre-defined contour sections (e.g. falling boundaries or rise-fall accents), or the algorithm may be used on arbitrary or successive contour sections, as the user prefers.

Once a contour section is isolated, the length of the section is calculated as a sum of the distances between frames (taking both time and frequency into account). This effectively measures the contour section as if it were a road on a map, following the line of the curve, rather than only one axis or the other. The sum of the section distances is then divided into a fixed number of segments (10% of the distance of the reference contour segment). The RMSE distance calculation is computed at the segment boundaries.

5.5.3 Perception of F0 Difference

The subjective evaluation of intonation difference consisted of asking novice subjects (first and second year undergraduate students) to rate twenty-four utterance pairs on a five-point scale (0-4). Appendix D contains the list of utterances used for this experiment. The pairs, bar a control set, consisted of one stimulus synthesized with the F0 which was extracted from the original utterance (and smoothed) and one synthesized from an F0 generated using the method described above. The control set of four pairs contained only one or the other for both stimuli. The stimuli were generated using LPC resynthesis of the original waveform and the imposed F0. The utterances were presented to the subjects via a web interface over Sennheiser headphones using standard audio software on Sun Ultra workstations. The subjects participated in the experiment in a quiet, closed computing laboratory.

The data was designed, for lack of a better continuum, to cover a range of RMSE scores. This design assumed that, even if RMSE is not the best measure of perceived difference, it is available and should reflect some sort of difference. Therefore, pairs ranging from 35Hz RMSE to 50Hz RMSE were chosen, with the ends of the scale less represented than the middle. The uneven distribution resulted from a desire to find out how much of a difference subjects could hear in contours on the middle of the RMSE scale, as this is where most generated contours rest.

Nineteen native speakers of British English took part in the test. The subjects had a general understanding of intonation when questioned, were provided with written instructions, and were able to practice using the interface and ask questions prior to beginning the test. The subjects were instructed to listen to each utterance of a pair as many times as they liked, and then rank the pair according to how different the intonation of two utterances in the pair sounded to them. As noted above, the ranks ranged from 0 (no difference) to 4 (completely different).

The web-style interface consisted of four pages, each with six stimulus pairs and written details about how to use the rating scale and an introduction page with full instructions on their task and seven example pairs. The example pairs also included example scores to illustrate a rough guide to the audible differences which the subjects might hear and as an example of how the web interface worked.

For each stimulus pair, two buttons could be clicked by the subjects - one to play each stimulus. Five ranking buttons were lined up to the right of the stimulus buttons, with the ranking value heading each column. The subjects could listen to each stimulus as many times as they wanted before choosing a ranking on the scale. Having made the decision, the ranking was selected by clicking on the appropriate ranking button. Figure 5.8 shows a sample page.

Subjects who were able to accurately place the control pairs in the 0 or 1 ratings and used at least four of the five possible rating choices were included in the correlation with the objective measures. Those who did not consistently place the control pairs in the 'most similar' range, or who placed all utterances in the 0-3 range were not included, as they did not show a potential to make the sort of fine distinctions that the objective measures are being tested for.

After the subjects who were unable to meet the criteria mentioned above were omitted, a Friedman test [GD82] was carried out to determine whether the subjects were making distinctions between of different F0 shapes ($\chi^2 = 170.32$, $df = 23$, and $p < 0.0001$). The test refutes the null hypothesis (that subjects randomly scored stimulus pairs). This result suggests that the subjects do make some sort of distinction between F0 differences, and that the subjective results can be compared to the objective measures. Below, this comparison is presented in order to determine whether the distinctions the subjects make are the same distinctions that the objective metrics measure.

In order to ensure that the inter-subject scoring was consistent, the raw scores for each subject were rescaled onto the original 0-4 scale. This allows the use of scores from subjects who used only four of the possible five ranks to be included with the other subjects. The scores for each participant were rescaled by mapping the maximum and minimum scores to the ends of the

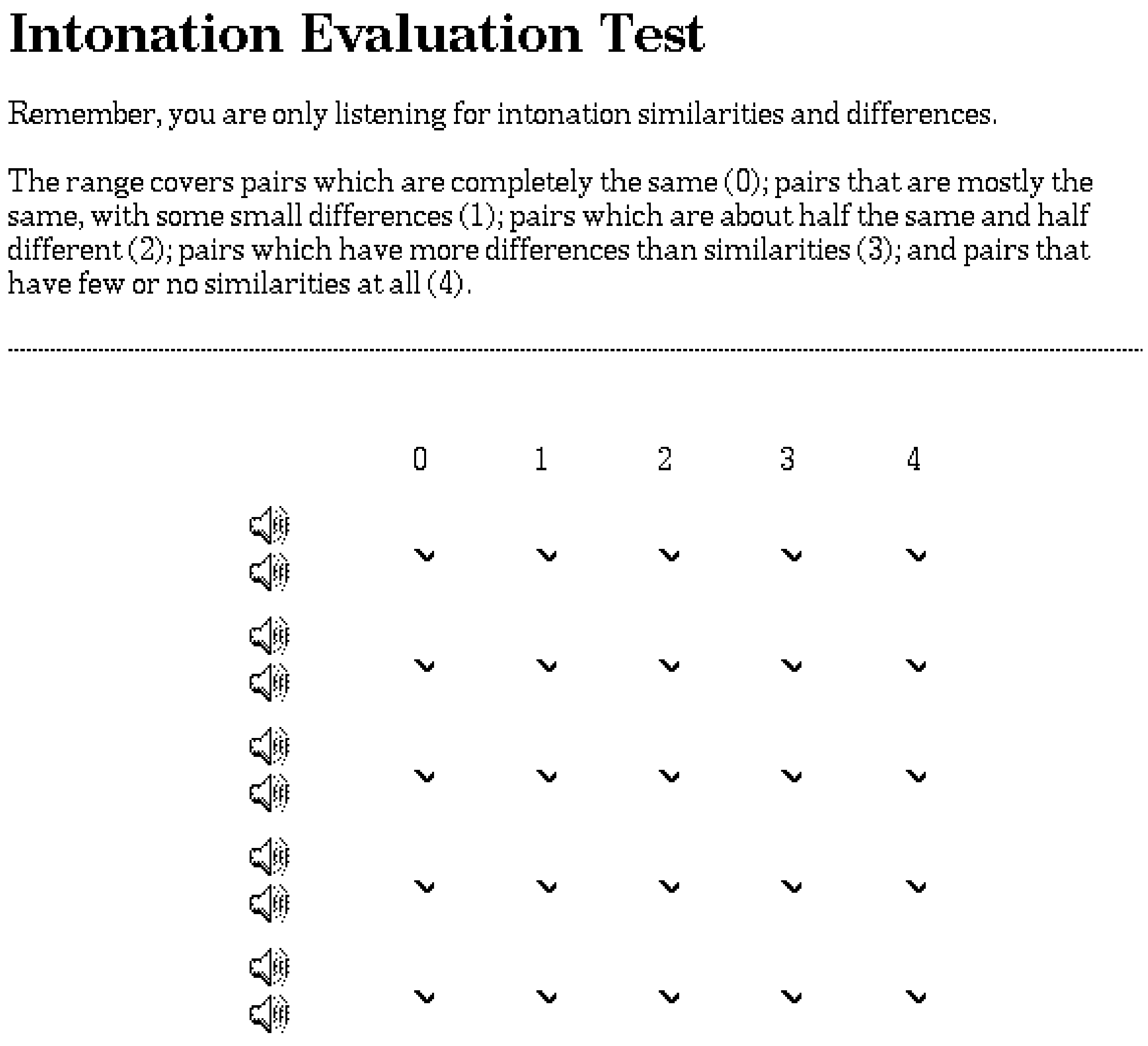


Figure 5.8: Sample page of perceptual experiment interface

scale (4 and 0) and linearly rescaling the intermediate scores. With all of the subjects' scores now on the same scale, it is possible to calculate an average score for each stimulus pair. This score can then be compared with the objective metrics.

Table 5.16 shows the Pearson correlation coefficients between the averaged perceptual scores and the objective metrics. While the segmentation of the contours into intonation events for objective scoring was originally intended only for use with the warping method, the segmentation was carried out on all of the RMSE methods in order to provide an accurate comparison. These scores are shown in the "Events Only" portion of the table.

Whole Contour	RMSE	corr.	tangent	warping
perceptual score	.6441	-.5497	.6150	.6003
	p=.000	p=.005	p=.001	p=.002
Events Only	RMSE	corr.	tangent	warp
perceptual score	.6534	–	.5878	.6499
	p=.001	–	p=.003	p=.001

Table 5.16: Correlation of perceptual scores and F0 contour distance metrics.

One important aspect of the comparison between subjective and objective scores is that all of the objective scoring methods are significantly correlated to the perceptual scores. Over entire contours, RMSE clearly provides the best correlation. Pearson Correlation, though, gives the lowest correlation. When only the intonation events are evaluated, the difference between RMSE and the warping method is negligible, but the difference between RMSE and the tangential estimation method is noticeably larger.

The comparison between objective and subjective scores results in fairly low scores in all cases. The highest correlation, $R = 0.6534$, results in accounting for only 42% of the variation among subjective scores ($R^2 = .4269$).

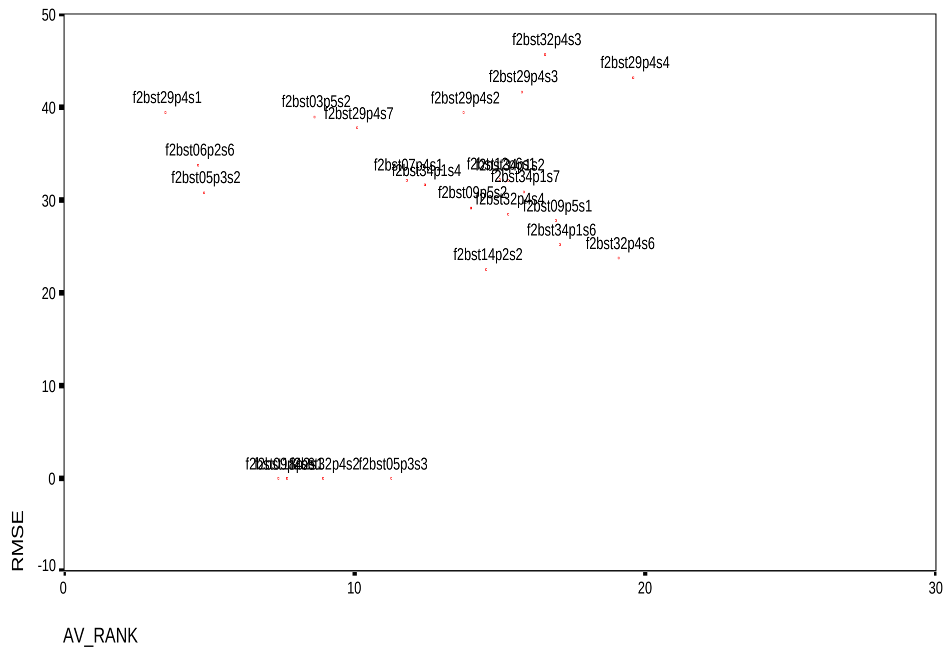


Figure 5.9: Scatter plot of subjective score and RMSE

Clark and Dusterhoff [CD99] offer one possible explanation for this low coverage: the data does not relate to perceived differences. A second possibility for the low correlation between subjective and objective scores is that the objective scores are not measuring on the correct scale. It is possible that measuring the contours on a logarithmic scale would turn up different results.

As Figure 5.9[4] shows, the consistency of judgements across listeners in the lower end of the RMSE scale was almost non-existent. Each point in the graph represents one of the stimulus pairs. The y-axis shows the RMSE of the pair. The x-axis shows the average ranking which the pair received. The subjective scores (0-4) were translated into a ranking for each subject. Ideally, this results in those pairs with an RMSE of 0 being ranked 0-4. Those pairs scored as 1 would get ranked as 5-8, for example. At the bottom of the

[4] Graph by R. Clark from the research leading up to Clark and Dusterhoff, 1999

RMSE scale (RMSE=0), the four pairs which were equivalent were all ranked in the ten most similar utterances. However, also in that top ten list are five pairs near the top of the similarity scale (RMSE $>$ 30). If RMSE were a clear indicator of perceived difference, a regression line from the origin to the upper right would be evident in this graph. As the graph shows, there is no grouping of any kind which shows a regression line from the origin to the upper right. As RMSE correlates most highly with the subjective rankings, it seems clear that none of the objective measures capture the variation that the subjects are hearing.

5.6 Conclusions

Perhaps the most important aspect of intonation synthesis to note is the inherent difficulty in measuring success. An evaluation of synthetic intonation which relies on subjective judgements must also account for the basis of those judgements. Simply determining that an utterance's intonation sounds natural, or sounds just like another utterance's intonation, is only so useful as its ability to be translated into improvements in an application. Similarly, an objective metric which shows that the current synthetic F0 contour is 15% more like the original than last week's synthetic contour is only useful if one knows whether that 15% results in an audible improvement.

The experiments discussed in this chapter show that it is possible to model intonation using the Tilt model and regression trees. These experiments have also shown that the synthetic intonation which results from these trees is similar to the natural intonation. Measured objectively, the synthetic intonation approximates the similarity to natural intonation which has been shown by other studies (e.g. [Ros94], [BH96]).

The main benefit of generating intonation using the methods described in this chapter is that individual aspects of intonation modelling may be investigated. Thus, it is possible to discover both which aspects of intonation events are being modelled, and what types of features are useful for modelling each event and parameter type. In the most common event types (accents and falling boundaries), the RMS error is generally small for each tree. The distribution of predicted paramenter values, which correlation reflects, is also acceptable in most cases. The overall comparison of natural and synthetic intonation contours shows that the research discussed in this chapter results in intonation which is comparable to previous work using different methods. Finally, the methodology described in this chapter has been applied to both manually and automatically annotated databases, and has resulted in reasonably natural sounding intonation contours in both contexts. As the final section of this chapter shows, assessing synthetic intonation in finer terms is difficult.

Small improvements in individual tree evaluations are shown to carry through to the overall comparison between natural and synthetic intonation contours. One conclusion that can be drawn from the individual tree assessment is that using sub-syllabic feature data to predict Tilt parameter values improves synthetic intonation. The level of improvement over intonation which was generated from trees which did not include sub-syllabic features was small. Given that the intonation contours from which the trees were trained were smoothed, a large change in results due to sub-syllabic features would have been unusual. The fact that these features still play a part in modelling smoothed F0 contours shows that sub-syllabic feature data are useful in intonation models. The research in this chapter attempts to determine the utility of feature classes in intonation modelling, so that the methodology is applicable to other feature classes as well. If features describing focus, contrast, or topic structure were to be included in a similar investigation, the methodology described above would provide a basis for determining how each feature type affects timing and scaling factors in intonation modelling. The limits which were reached in the models discussed above, especially those involving the *tilt* parameter, may be overcome if such higher-level information were included in the model.

Chapter 6

Conclusions

This thesis has discussed a methodology with which fundamental frequency can be synthesized using statistical models that have been automatically trained from annotated speech data. The thesis began with an introduction to the research contained in this volume. This introduction was followed by a discussion of segmental interactions with intonation. This discussion provided the background for an important area of investigation within the automatic intonation labelling work which was presented in Chapter 4. Chapter 3 completed the background in intonation theory and experimentation that is necessary to place the research in chapters 4 and 5 in context.

The research in Chapter 4 addressed basic problems of how to acquire database annotations. The goal of the research in this chapter was to improve on previous attempts at modelling F0 using acoustic information. Based on the literature presented in Chapter 2, experiments were performed which examined acoustic information which is related to both segmental classification and F0 movement. These experiments showed that including Mel Frequency Cepstral Coefficients improved the performance of an HMM modelling methodology which had previously used only F0 and RMS Energy data.

The addition of MFCC information increased the number of correctly identified intonation events, and in some cases reduced the number of inserted event labels. The experiments also showed that the accuracy and reliability of the system suffered with smaller databases. Therefore, this methodology, while sound, is more useful as an aid to manual annotation until a database is large enough to create reliable models than as an initial database labelling technique.

The main body of research in this thesis concerned generating parameterized descriptions of F0 contours using decision trees which have been automatically trained from annotated speech data. These descriptions were then automatically converted into F0 contours. Chapter 5 discussed the experiments which were used in the composition of a methodology to produce such models. This methodology consists of providing information about the text being synthesized to a regression tree building system and using the resulting trees to predict parameterized descriptions of the fundamental frequency contour for the synthesized utterance. The use of a single regression tree for each parameter of each intonation event type allowed an analysis of how the data features affect the prediction of each parameter. This aspect of the modelling technique is an improvement on previous research, which does not present the contribution of different information types within a system. Chapter 5 also presents evaluations of F0 contours which have been generated from the parameters predicted by the decision trees. Evaluations of these contours were shown to compare favourably with similar evaluations of previous research using the same database. Two different perceptual tests involving intonation contours generated using systems based on the methodology described in section 5.5 were reviewed. These tests showed that 1) the synthetic contours are not sufficiently good to be adjudged as sounding natural as often as the natural contours for the same utterances, and 2) perception of intonation does not necessarily agree with commonly used objective metrics in fine grain comparisons.

The research in Chapter 5 showed that 1) position within a phrase is important in most aspects of parameter prediction, 2) sub-syllable features are useful in predicting event timing, both the peak and the duration, and 3) viewing intonation events in relation to an intonation tier is much more useful in parameter prediction than viewing the events in relation to a syllable tier. All of these findings are supported by the theoretical and experimental literature. The fundamental frequency contours generated from the parameters predicted by the regression trees are comparable to those produced by previous research, and sound reasonably natural. The synthesis process was also carried out using automatic intonation labels which were acquired using the methods described in Chapter 4. The trees and contours produced in this experiment were similar to those produced using the manually derived labels, although slightly less successful.

6.1 Limitations of the Research

The research discussed in this thesis is faced with large obstacles, some of which have only been partially overcome. The greatest limitation on this research has been the availability of appropriately annotated data. A wide variety of data was available for this research, but individual speakers' databases tended to be quite small. This limitation hampered the ability to train both the synthesis and analysis models to work with the different speech styles and voices. This limitation, though, can be overcome with the provision of a great deal more data and time to ensure its proper preparation.

A number of related limitations of this research involve the data features used in model building. Within the intonation analysis task, only a small number of features were tested. It is quite possible that more features, or different combinations of the existing features would yield better results. However, it is difficult to say what other features should be tested.

On the intonation synthesis front, over fifty individual data items are used in the tree building process. These data items cover more than twenty features and the extensions of those features to adjacent constituents (e.g. lexical stress is applied to five syllables for each intonation event, giving five data items). With such wide coverage, it is difficult to imagine that there are not enough features being used. In this case, the problem is that each of the features is used for at least one tree, and therefore should not be removed from the feature set. Therefore, the obstacle is not that there are too few features, but that there may be too many features which can be used in the same way. This problem is compounded by the lack of high-level features. While, in many ways, this thesis is concerned only with phonetic aspects of fundamental frequency and low-level, sub-syllabic influences thereon, intonation is related to structures on a range of levels. Features which relate to focus, reference, contrast, and discourse structure do not play a role in the model training. Partially, this omission is due to the lack of such annotation in the data. The availability of a proven experimental model which would lend itself to such annotation would help alleviate the problem.

Perhaps the most perplexing obstacle that this research faces is a way of objectively evaluating the fundamental frequency contours which are created by the synthesis system. Within this thesis, the task was taken as one of determining how closely the model output matches the model training input. To a degree, this is a successful way of judging the success of the model. This method was chosen in order to evaluate the overall methodology, the feature set, and individual aspects of intonation prediction. Evaluating these pieces of the intonation synthesis system in a way that provides feedback for the researcher is of more use to research than deciding that thirty of forty-five listeners prefer model X. Although it may be difficult to determine if the methods used in this thesis produce intonation which 67% of listeners prefer over some other method, it is at least clear what causes improvements and where the methodology could be improved.

6.2 Future Work

Each of the areas researched in this thesis could be enhanced by future experimentation. Obviously, use of the same methodology on different and larger databases would prove useful in supporting or amending the current findings. Within the intonation analysis work, continued investigations of different types and combinations of acoustic information would be interesting. As mentioned in the discussion above of the limitations of the synthesis research, including high-level information in the feature set could improve the system further. Future work which replicates the research in this thesis using a perception-based pitch scale (in logarithmic domain, rather than linear) would prove interesting. Often such scales are supported in the literature, but they are not always supported in system infrastructure. A comparison where only the intonation measurement scale differed would provide an idea of precisely what differences the two methods produce, and how those differences affect the intonation models. Section 5.5 discusses some recent collaborative research into the correlation between objective and subjective evaluation of intonation. Future work in this area would be very beneficial, either in developing new metrics or performing more detailed and controlled experiments. The benefit of objective measurements is that researchers can easily determine the limitations and future directions of their work. The difficulty at present is that such measurements are not always reflected in what people hear when the intonation is part of synthesized speech.

Appendix A

Autolabelling Results

Tables contain results with information as follows:

- Type: One or two stream, number of mixture components in parentheses.
- Correct: Percent of autolabels which are correct. (see Chapter 4 for scoring method)
- Accuracy: Correct minus the percent of autolabels which are insertions.
- S: Grammar scaling factor
- P: Transition penalty (only noted where not 0)

As noted in the main text, each experiment uses a bigram/unigram grammar which has been trained on the database.

In cases where there are two streams, there are stream weights of 1.0 and 0.8 on all states, unless otherwise noted. In the two stream experiments, the F0 and energy data are in the first stream, and other accoustic data is in the second stream.

A.1 Speaker F2B

A.1.1 Unnormalized Data

The tables in this section relate to unnormalized F0 and Energy data.

Type	Correct	Accuracy	S
1 Stream (3 components)	77.15	56.37	8
1 Stream (5 components)	73.19	57.03	10
1 Stream (7 components)	74.12	60.06	12
1 Stream (9 components)	75.80	60.12	10
1 Stream (11 components)	77.10	60.33	8
1 Stream (13 components)	78.35	61.04	6
1 Stream (15 components)	77.59	59.63	6
1 Stream (17 components)	77.54	59.85	6
1 Stream (19 components)	74.39	60.17	12
1 Stream (21 components)	77.97	60.44	6

Table A.1: F0 and energy + delta + acceleration

Type	Correct	Accuracy	S
1 Stream (3 components)	69.45	44.81	11
1 Stream (5 components)	77.54	53.61	7
1 Stream (7 components)	78.19	54.96	6
1 Stream (9 components)	79.71	58.17	6
1 Stream (11 components)	78.19	57.56	7
1 Stream (13 components)	78.3	58.06	7
1 Stream (15 components)	73.36	56.59	12
1 Stream (17 components)	72.98	57.95	14
1 Stream (19 components)	75.15	52.63	10
1 Stream (21 components)	78.13	58.17	7

Table A.2: F0, energy, and zero-crossing (with delta & acceleration)

Type	Correct	Accuracy	S
1 Stream (3 components)	76.56	52.74	15
1 Stream (5 components)	79.16	58.54	10
1 Stream (7 components)	78.46	60.61	11
1 Stream (9 components)	79.81	59.41	8
1 Stream (11 components)	77.16	61.47	14
1 Stream (13 components)	80.68	61.1	6
1 Stream (15 components)	76.72	60.99	14
1 Stream (17 components)	79.22	53.06	3.0
1 Stream (19 components)	72.98	49.10	11
1 Stream (21 components)	76.23	57.82	13

Table A.3: F0, energy, and auto-correlation peak (with delta & acceleration)

Type	Correct	Accuracy	S
1 Stream (3 components)	47.42	22.30	6
1 Stream (5 components)	49.97	20.13	3
1 Stream (7 components)	49.86	21.49	4
1 Stream (9 components)	60.61	21.59	1
1 Stream (11 components)	56.05	24.25	2
1 Stream (13 components)	52.63	24.31	3
1 Stream (15 components)	50.89	24.09	4
1 Stream (17 components)	52.85	25.66	3
1 Stream (19 components)	49.81	25.18	5
1 Stream (21 components)	48.72	22.03	4

Table A.4: Auto-correlation peak only (with delta & acceleration)

Type	Correct	Accuracy	S
2 Stream (3 components)	76.45	57.95	13
2 Stream (5 components)	77.37	58.65	12
2 Stream (7 components)	77.59	58.49	11
2 Stream (9 components)	82.29	60.22	6
2 Stream (11 components)	77.26	61.20	13
2 Stream (13 components)	76.88	59.95	13
2 Stream (15 components)	76.72	61.20	13
2 Stream (17 components)	79.16	61.53	9
2 Stream (19 components)	77.86	62.13	12
2 Stream (21 components)	77.21	62.07	13

Table A.5: F0, energy, and auto-correlation peak (with delta & acceleration)

Type	Correct	Accuracy	S
2 Stream (3 components)	77.26	56.58	11
2 Stream (5 components)	76.83	58.54	12
2 Stream (7 components)	76.02	58.22	13
2 Stream (9 components)	77.16	60.28	12
2 Stream (11 components)	76.23	60.82	14
2 Stream (13 components)	79.11	59.90	8
2 Stream (15 components)	77.54	60.23	10
2 Stream (17 components)	76.94	61.20	12
2 Stream (19 components)	77.48	62.13	12
2 Stream (21 components)	78.35	62.4	10
2 Stream (23 components)	76.83	62.89	13

Table A.6: F0, energy, and auto-correlation peak (with delta & acceleration), stream weights of 1.0 and 0.6

Type	Correct	Accuracy	S
1 Stream (3 components)	80.25	57.24	7
1 Stream (5 components)	80.47	61.31	9
1 Stream (7 components)	77.97	60.22	12
1 Stream (9 components)	82.31	61.04	6
1 Stream (11 components)	80.68	61.04	8
1 Stream (13 components)	80.47	61.85	8
1 Stream (15 components)	81.39	60.44	6
1 Stream (17 components)	82.04	60.88	6
1 Stream (19 components)	79.00	64.02	13
1 Stream (21 components)	78.97	64.08	13
1 Stream (23 components)	77.75	63.86	14
1 Stream (25 components)	78.35	64.3	13
1 Stream (27 components)	81.61	64.51	8

Table A.7: F0, energy, and MFCC[0-3] (with delta & acceleration)

Type	Correct	Accuracy	S
2 Stream (3 components)	77.97	59.08	11
2 Stream (5 components)	81.71	60.66	6
2 Stream (7 components)	80.41	62.13	7
2 Stream (9 components)	79.16	61.96	8
2 Stream (11 components)	79.76	62.07	8
2 Stream (13 components)	80.85	63.10	7
2 Stream (15 components)	78.51	61.80	10
2 Stream (17 components)	80.47	63.48	8
2 Stream (19 components)	81.06	61.96	6
2 Stream (21 components)	80.57	62.4	7

Table A.8: F0, energy, and MFCC[0-3] (with delta & acceleration)

Type	Correct	Accuracy	S
2 Stream (3 components)	92.91	59.90	11
2 Stream (5 components)	83.99	64.51	13
2 Stream (7 components)	83.78	65.98	14
2 Stream (9 components)	84.21	67.55	15
2 Stream (11 components)	84.48	67.39	15
2 Stream (13 components)	83.99	66.84	15
2 Stream (15 components)	86.92	65.87	10
2 Stream (17 components)	85.89	65.11	11
2 Stream (19 components)	86.20	66.93	12.0
2 Stream (21 components)	85.13	61.01	13

Table A.9: F0 and MFCC[all 13] (with delta & acceleration)

A.1.2 Normalized Data

This section relates to experiments which use normalized F0 and Energy data.

Type	Correct	Accuracy	S
1 Stream (3 components)	70.7	51.06	12.0
1 Stream (5 components)	72.00	53.87	11.0
1 Stream (7 components)	70.59	51.11	11.0
1 Stream (9 components)	74.55	54.53	7.0
1 Stream (11 components)	75.64	55.34	5.0
1 Stream (13 components)	71.89	54.80	10.0
1 Stream (15 components)	75.42	56.37	7.0
1 Stream (17 components)	75.69	56.70	8.0
1 Stream (19 components)	76.07	57.51	6.0
1 Stream (21 components)	78.19	59.09	6.0
1 Stream (23 components)	73.57	58.06	7.0
1 Stream (25 components)	74.61	57.19	6.0
1 Stream (27 components)	74.33	57.35	6.0
1 Stream (29 components)	74.28	56.92	6.0
1 Stream (31 components)	76.02	58.17	6.0

Table A.10: F0 and energy (with delta & acceleration)

Type	Correct	Accuracy	S
2 Stream (3 components)	79.65	54.64	12
2 Stream (5 components)	77.43	57.35	17
2 Stream (7 components)	77.10	59.62	18
2 Stream (9 components)	80.90	60.5	12
2 Stream (11 components)	82.96	60.93	9
2 Stream (13 components)	79.87	59.36	14
2 Stream (15 components)	78.40	59.30	19
2 Stream (17 components)	81.82	62.50	14
2 Stream (19 components)	83.4	63.54	11
2 Stream (21 components)	81.66	63.37	9

Table A.11: F0 and MFCC[all 13] (with delta & acceleration)

Type	Correct	Accuracy	S
2 Stream (3 components)	74.88	51.65	9
2 Stream (5 components)	79.33	57.46	7
2 Stream (7 components)	78.62	58.82	8
2 Stream (9 components)	77.32	60.82	14
2 Stream (11 components)	82.20	60.28	6
2 Stream (13 components)	77.64	58.87	13
2 Stream (15 components)	76.50	60.12	17
2 Stream (17 components)	80.14	62.50	12
2 Stream (19 components)	81.77	63.81	10
2 Stream (21 components)	79.38	63.48	15
2 Stream (23 components)	80.24	64.29	15
2 Stream (25 components)	79.27	63.70	16
2 Stream (27 components)	82.64	63.64	19
2 Stream (29 components)	79.87	63.70	14

Table A.12: F0 and MFCC[all 13] (with delta & acceleration), stream weights of 1.0 and 0.6

Type	Correct	Accuracy	S
2 Stream (3 components)	71.49	52.82	20
2 Stream (5 components)	77.23	55.46	14
2 Stream (7 components)	78.43	61.00	13
2 Stream (9 components)	80.08	62.14	11
2 Stream (11 components)	78.61	62.63	14
2 Stream (13 components)	80.71	63.46	13
2 Stream (15 components)	82.82	63.9	11
2 Stream (17 components)	83.10	65.34	12
2 Stream (19 components)	81.94	65.26	13
2 Stream (21 components)	80.56	65.67	16
2 Stream (23 components)	82.37	66.48	13
2 Stream (25 components)	83.64	67.73	12
2 Stream (27 components)	84.22	68.01	11
2 Stream (29 components)	81.91	67.53	14
2 Stream (31 components)	83.82	67.89	10

Table A.13: F0 and MFCC[all 13] (with delta, no acceleration)

A.1.3 Blind Results

There are nine news story paragraphs in the blind set. Table A.14 shows three conditions, as discussed in Chapter 4.

Condition	Correct	Accuracy
Best F0/MFCC Combination	85.51	65.89
Best F0/MFCC (without acceleration)	79.96	59.69
Best F0/MFCC-AC Combination	79.30	62.31

Table A.14: Three test conditions using the blind data set

A.2 Speaker KDS

In the KDS database, the Correct and Accuracy scores are divided into All/Major scores (e.g. 53.77/53.59).

A.2.1 Normalized Data

This section relates to experiments which use normalized F0 and Energy data.

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	53.77/53.59	31.94/35.77	10	0
2 Stream (3 components)	64.81/65.50	44.95/45.64	15	0
2 Stream (5 components)	71.66/71.89	50.98/51.33	6	0
2 Stream (7 components)	74.45/74.8	50.17/52.15	4	0
2 Stream (9 components)	72.82/73.17	52.26/52.73	7	0
2 Stream (11 components)	71.78/72.12	53.66/54.12	8	0
2 Stream (13 components)	68.41/68.76	54.70/55.05	20	0
2 Stream (15 components)	71.08/71.43	54.82/55.17	12	0
2 Stream (17 components)	71.31/71.66	55.86/56.21	12	0
2 Stream (19 components)	71.43/71.78	55.52/55.86	11	0
2 Stream (21 components)	72.24/72.59	55.05/55.40	8	0
2 Stream (23 components)	71.54/71.89	55.75/56.1	10	0
2 Stream (25 components)	70.73/71.08	55.86/56.21	13	0
2 Stream (27 components)	69.45/69.80	54.94/55.28	14	0
2 Stream (29 components)				

Table A.15: F0 and energy (with delta & acceleration)

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	73.05/73.86	45.53/50.17	9	10
2 Stream (3 components)	77.82/78.4	59.35/60.16	17	10
2 Stream (5 components)	80.49/80.95	54.24/56.79	5	15
2 Stream (7 components)	75.49/76.07	55.86/56.79	10	15
2 Stream (9 components)	76.19/76.77	53.77/54.59	7	15
2 Stream (11 components)	73.64/74.22	54.70/55.40	11	15
2 Stream (13 components)	76.19/76.89	55.52/56.21	16	10
2 Stream (15 components)	75.61/76.19	57.61/58.19	19	10
2 Stream (17 components)	76.77/77.49	55.86/56.56	8	15
2 Stream (19 components)	82.23/82.69	56.91/57.49	6	10
2 Stream (21 components)	73.63/74.22	57.61/58.19	16	15
2 Stream (23 components)	77.82/78.51	57.37/58.07	6	15
2 Stream (25 components)	73.98/74.68	57.72/58.42	7	20
2 Stream (27 components)	77.23/77.82	57.61/58.19	7	15
2 Stream (29 components)	77.23/77.82	59.7/60.28	9	15

Table A.16: F0 and 13 MFCC (with delta & acceleration)

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	72.47/73.29	45.53/50.06	9	0
2 Stream (3 components)	79.09/79.67	55.98/60.16	5	10
2 Stream (5 components)	78.63/79.21	54.70/56.91	5	10
2 Stream (7 components)	73.52/74.1	56.91/57.84	5	15
2 Stream (9 components)	77.58/78.16	56.1/56.79	5	10
2 Stream (11 components)	72.71/73.29	55.52/56.1	18	5
2 Stream (13 components)	71.66/72.36	55.52/56.33	6	15
2 Stream (15 components)	74.1/74.68	57.03/57.61	10	10
2 Stream (17 components)	73.75/74.45	57.26/57.95	6	15
2 Stream (19 components)	74.22/74.8	59.00/59.58	5	15
2 Stream (21 components)	75.03/75.72	58.42/59.12	10	10
2 Stream (23 components)	77.92/78.51	58.65/59.35	4	10
2 Stream (25 components)	75.38/76.07	58.30/59.00	3	15
2 Stream (27 components)	72.12/72.82	57.14/57.84	7	15
2 Stream (29 components)	75.26/75.96	59.23/59.93	10	10

Table A.17: F0 and 13 MFCC (with delta & acceleration) stream weights of 1.0 and 0.6

A.2.2 Unnormalized Data

This section relates to experiments which use unnormalized F0 and Energy data.

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	53.77/54.59	31.93/35.77	10	0
2 Stream (3 components)	64.81/65.50	44.95/45.64	15	0
2 Stream (5 components)	69.22/69.69	51.22/51.68	11	0
2 Stream (7 components)	69.57/69.92	52.03/52.38	13	0
2 Stream (9 components)	69.69/70.03	53.31/53.66	17	0
2 Stream (11 components)	71.43/71.78	53.77/54.12	11	0
2 Stream (13 components)	68.99/69.34	54.70/55.05	18	0
2 Stream (15 components)	70.61/70.96	54.82/55.17	13	0
2 Stream (17 components)	70.96/71.31	56.1/56.44	12	0
2 Stream (19 components)	69.57/69.92	55.28/55.63	14	0
2 Stream (21 components)	72.82/73.17	55.63/55.98	8	0
2 Stream (23 components)	72.47/72.94	55.63/56.1	9	0
2 Stream (25 components)	70.5/70.96	55.63/56.1	13	0
2 Stream (27 components)	70.03/70.5	54.82/55.28	13	0
2 Stream (29 components)	68.52/68.87	54.70/55.05	17	0

Table A.18: F0 and energy (with delta & acceleration)

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	73.05/73.87	45.53/50.17	9	10
2 Stream (3 components)	78.98/79.56	57.03/60.63	7	15
2 Stream (5 components)	79.56/80.02	54.82/56.91	6	15
2 Stream (7 components)	75.49/76.07	55.86/56.79	10	15
2 Stream (9 components)	73.40/73.98	54.24/55.05	6	20
2 Stream (11 components)	75.26/75.82	54.12/54.82	8	15
2 Stream (13 components)	76.19/76.65	56.1/56.56	17	10
2 Stream (15 components)	75.03/75.61	56.56/57.26	12	15
2 Stream (17 components)	72.24/72.94	55.52/56.21	16	15
2 Stream (19 components)	74.33/75.03	55.75/56.45	20	10
2 Stream (21 components)	72.94/73.63	56.91/57.61	18	15
2 Stream (23 components)	72.59/73.29	55.40/56.1	18	15
2 Stream (25 components)	75.03/75.61	57.72/58.30	14	15
2 Stream (27 components)	73.29/73.98	57.42/59.12	11	20
2 Stream (29 components)	77.82/78.51	57.49/58.19	4	15

Table A.19: F0 and 13 MFCC (with delta & acceleration)

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	72.47/73.29	45.53/50.06	9	0
2 Stream (3 components)	78.98/79.56	55.86/60.05	5	10
2 Stream (5 components)	78.74/79.33	54.82/57.14	5	10
2 Stream (7 components)	73.29/73.75	54.59/55.86	4	10
2 Stream (9 components)	76.77/77.35	56.79/57.49	5	10
2 Stream (11 components)	74.1/74.68	56.21/56.91	8	10
2 Stream (13 components)	78.16/78.74	55.98/56.68	10	5
2 Stream (15 components)	75.26/75.84	56.68/57.37	8	10
2 Stream (17 components)	71.89/72.59	56.79/57.49	7	15
2 Stream (19 components)	73.98/74.68	57.49/58.19	5	15
2 Stream (21 components)	74.68/75.37	57.37/58.07	4	15
2 Stream (23 components)	73.40/74.1	57.72/58.42	6	15
2 Stream (25 components)	73.52/74.22	58.19/58.88	7	15
2 Stream (27 components)	79.63/79.33	58.88/59.58	4	10
2 Stream (29 components)	73.40/74.1	59.00/59.7	6	15

Table A.20: F0 and 13 MFCC (with delta & acceleration) stream weights of 1.0 and 0.6

A.3 Speaker KDW

Like KDS, KDW uses All/Major scores for Correct and Accuracy. Some experiments using the KDW database used a bigram/unigram grammar trained on the F2B database. Unlike the other databases, this database gave the best results with stream weights of 1.0 and 0.6. All of the tables below assume these stream weights. As was noted in the main text, some experiments used 4-state HMMs, rather than the typical 5-state. This is noted in the table caption where applicable.

A.3.1 Normalized Data

This section relates to experiments which use normalized F0 and Energy data.

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	72.77/74.26	34.90/38.12	5	10
2 Stream (3 components)	71.29/73.27	42.57/45.05	11	10
2 Stream (5 components)	75.74/77.72	48.02/51.73	6	10
2 Stream (7 components)	77.23/78.96	46.78/48.76	12	5
2 Stream (9 components)	75.25/77.23	48.76/52.23	6	10
2 Stream (11 components)	80.2/82.53	49.75/53.46	8	5
2 Stream (13 components)	75.99/77.97	49.75/52.23	7	10
2 Stream (15 components)	75.99/77.72	51.73/53.46	9	10
2 Stream (17 components)	77.47/78.96	51.49/53.21	6	10
2 Stream (19 components)	83.17/84.16	54.70/56.19	10	5
2 Stream (21 components)	80.45/81.93	53.96/55.45	13	5
2 Stream (23 components)	75.49/76.98	53.46/55.2	10	10
2 Stream (25 components)	75.25/76.98	54.45/56.44	11	10
2 Stream (27 components)	75.74/77.47	54.70/56.68	10	10
2 Stream (29 components)	77.72/79.45	51.73/55.45	6	10

Table A.21: F0 and MFCC[all 13] (with delta & acceleration) 4-state hmm F2B Grammar

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	72.77/74.26	35.15/38.37	5	10
2 Stream (3 components)	71.78/73.76	42.57/45.3	10	10
2 Stream (5 components)	75.74/77.72	48.12/51.73	6	10
2 Stream (7 components)	77.23/78.96	46.78/48.76	12	5
2 Stream (9 components)	75.25/77.23	48.76/52.23	6	10
2 Stream (11 components)	80.2/72.43	49.75/53.46	8	5
2 Stream (13 components)	75.99/77.97	49.75/52.23	7	10
2 Stream (15 components)	76.24/77.97	52.48/54.21	9	10
2 Stream (17 components)	80.69/81.68	52.48/53.46	13	5
2 Stream (19 components)	83.17/84.16	54.70/56.19	10	5
2 Stream (21 components)	80.45/81.83	53.96/55.45	13	5
2 Stream (23 components)	75.5/76.98	53.47/55.2	10	10
2 Stream (25 components)	75.25/76.98	54.46/56.43	11	10
2 Stream (27 components)	75.74/77.48	54.70/56.68	10	10
2 Stream (29 components)	78.96/80.45	51.73/53.22	14	5

Table A.22: F0 and MFCC[all 13] (with delta & acceleration) 5-state hmm F2B Grammar

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)	83.91/84.65	32.92/35.64	6	0
2 Stream (3 components)	72.77/74.75	40.59/45.54	7	10
2 Stream (5 components)	75.49/77.72	46.53/51.24	6	10
2 Stream (7 components)	82.92/84.41	46.53/53.46	5	5
2 Stream (9 components)	76.73/78.71	47.52/50.25	9	5
2 Stream (11 components)	79.21/81.43	49.01/52.72	8	5
2 Stream (13 components)	76.48/78.46	48.76/52.23	6	10
2 Stream (15 components)	74.75/76.73	51.73/53.71	11	10
2 Stream (17 components)	81.93/82.92	51.98/53.71	9	5
2 Stream (19 components)	79.95/81.43	53.22/55.2	12	5
2 Stream (21 components)	74.50/76.48	52.72/55.2	11	10
2 Stream (23 components)	80.44/82.42	52.23/54.70	10	5
2 Stream (25 components)	73.76/75.49	51.73/53.71	11	10
2 Stream (27 components)	80.94/82.67	52.72/55.94	10	5
2 Stream (29 components)	80.69/82.18	51.73/54.21	10	5

Table A.23: F0 and MFCC[all 13] (with delta & acceleration) 5-state hmm KDW Grammar

A.3.2 Unnormalised Data

This section relates to experiments which use unnormalized F0 and Energy data.

Type	Corr/Maj	Acc/Maj	S	P
2 Stream (1 component)				
2 Stream (3 components)	73.02/75.25	36.88/39.12	13	0
2 Stream (5 components)	86.39/86.88	45.54/46.04	9	0
2 Stream (7 components)	81.93/83.66	44.55/46.29	11	0
2 Stream (9 components)	86.14/87.38	44.31/46.53	7	0
2 Stream (11 components)	76.49/77.97	45.05/46.53	16	0
2 Stream (13 components)	77.48/79.70	48.76/50.99	19	0
2 Stream (15 components)	78.71/80.69	44.06/46.04	14	0
2 Stream (17 components)	74.26/76.24	44.06/48.02	6	5
2 Stream (19 components)	75.99/78.47	47.03/51.23	5	5
2 Stream (21 components)	78.71/81.93	44.55/47.77	17	0
2 Stream (23 components)	79.95/82.18	48.02/50.25	16	0
2 Stream (25 components)	81.44/83.17	49.01/50.74	12	0
2 Stream (27 components)	80.2/82.43	50/52.23	13	0
2 Stream (29 components)	77.48/79.95	50/52.48	16	0

Table A.24: F0 and MFCC[all 13] (with delta & acceleration) 4-state hmm F2B Grammar

Appendix B

Synthesis Decision Tree Tables

	start_f0	amplitude	duration	tilt	peak_pos
a	164.52/42.08	71.94/49.46	0.307/0.07	0.040/0.48	0.052/0.73
arb	134.59/32.01	55.07/28.28	0.380/0.07	0.377/0.57	0.163/0.11
afb	144.42/31.82	84.36/46.53	0.362/0.05	-0.206/0.34	0.025/0.06
rb	145.72/39.06	45.92/31.54	0.243/0.08	0.302/0.56	0.119/0.10
fb	145.96/34.22	63.27/52.52	0.204/0.07	-0.088/0.54	0.086/0.86
c	155.48/43.17				
sil	113.53/50.53				

Table B.1: Individual Event/Parameter Results for F2B Mean and Standard Deviation Values (Entries are MEAN/STD)

	start_f0	amplitude	duration	tilt	peak_pos
a	33.04/0.61	46.58/0.25	0.057/0.51	0.427/0.35	0.08/0.50
arb	30.03/0.54	22.76/0.39	0.118/0.11	0.531/0.50	0.073/0.53
afb	28.63/0.47	39.82/0.48	0.06/0.46	0.278/0.44	0.597/0.33
rb	26.18/0.49	28.42/0.48	0.057/0.73	0.489/0.38	0.074/0.67
fb	25.66/0.49	43.59/0.34	0.06/0.52	0.486/0.40	0.073/0.67
c	34.17/0.60				
sil	28.96/0.80				

Table B.2: Individual Event/Parameter Results for F2B (Entries are RMSE/Correlation)

	start_f0	amplitude	duration	tilt	peak_pos
a	33.79/0.59	46.58/0.24	0.058/0.49	0.424/0.36	0.072/0.31
arb	27.15/0.65	20.43/0.50	0.119/0.15	0.467/0.58	0.062/0.57
afb	28.14/0.47	42.84/0.34	0.058/0.48	0.281/0.41	0.056/0.43
rb	26.34/0.50	28.03/0.49	0.057/0.71	0.491/0.37	0.073/0.67
fb	25.57/0.50	43.78/0.32	0.060/0.50	0.480/0.44	0.075/0.64
c	34.18/0.60				
sil	28.97/0.79				

Table B.3: Individual Event/Parameter Results for F2B - Hand Tuned (Entries are RMSE/Correlation)

	start_f0	amplitude	duration	tilt	peak_pos
a	32.82/0.62	45.92/0.29	0.057/0.50	0.425/0.36	0.072/0.33
arb	27.15/0.65	20.43/0.50	0.123/0.33	0.444/0.67	0.061/0.57
afb	28.11/0.45	41.68/0.43	0.056/0.54	0.281/0.41	0.056/0.43
rb	26.34/0.50	28.13/0.49	0.057/0.72	0.491/0.37	0.073/0.67
fb	25.52/0.50	43.74/0.32	0.060/0.50	0.480/0.44	0.075/0.64
c	34.18/0.60				
sil	28.97/0.79				

Table B.4: Individual Event/Parameter Results for F2B Further Hand Tuned (Entries are RMSE/Correlation)

	start_f0	amplitude	duration	tilt	peak_pos
a	33.54/0.58	51.03/0.22	0.085/0.26	0.421/0.36	9.668/0.60
afb	24.33/0.39	35.82/0.44	0.067/0.51	0.243/0.51	10.177/0.59
rb	19.97/0.72	22.90/0.46	0.095/0.37	0.407/0.60	10.583/0.60
fb	25.52/0.48	42.19/0.27	0.088/0.30	0.537/0.32	9.691/0.60
c	29.39/0.68				
sil	27.43/0.82				

Table B.5: Individual Event/Parameter Results for F2B Auto-labels (Entries are RMSE/Correlation)

	start_f0	amplitude	duration	tilt	peak_pos
a	18.41/0.52	23.65/0.42	0.07/0.55	0.54/0.36	0.09/0.46
afb	12.26/0.32	23.56/0.50	0.06/0.52	0.21/0.44	0.05/0.68
rb	18.44/0.78	25.67/0.44	0.06/0.59	0.53/0.77	0.16/0.64
fb	15.98/0.74	30.38/0.59	0.07/0.6	0.48/0.27	0.09/0.56
m	21.10/0.57	12.90/0.58	0.07/0.57	0.64/0.37	0.07/0.52
c	22.5/0.59				
sil	14.89/0.96				

Table B.6: Individual Event/Parameter Results for FHL (Entries are RMSE/Correlation)

	start_f0	amplitude	duration	tilt	peak_pos
a	19.26/0.46	24.45/0.34	0.08/0.29	0.54/0.35	0.1/0.33
afb	12.26/0.32	23.56/0.50	0.06/0.54	0.21/0.45	0.06/0.53
rb	18.44/0.78	25.67/0.44	0.06/0.59	0.53/0.77	0.16/0.64
fb	16.20/0.74	28.32/0.64	0.07/0.61	0.49/0.26	0.09/0.60
m	21.10/0.57	11.86/0.60	0.07/0.57	0.64/0.37	0.07/0.52
c	22.5/0.59				
sil	14.89/0.96				

Table B.7: Individual Event/Parameter Results for FHL without sub-syllable features (Entries are RMSE/Correlation)

	ev_f0	amplitude	duration	tilt	peak_pos
a	9.49/0.64	11.44/0.39	0.05/0.55	0.44/0.49	0.08/0.45
afb	3.73/0.77	6.85/0.81	.04/0.76	0.17/0.78	0.02/0.91
fb	6.32/0.65	9.95/0.72	0.04/0.77	0.49/0.59	0.06/0.44
m	3.55/0.9	14.94/0.83	0.03/0.88	0.84/0.43	0.06/0.87
c	11.57/0.61				
sil	10.31/0.98				

Table B.8: Individual Event/Parameter Results for KDT (Entries are RMSE/Correlation)

	start_f0	amplitude	duration	tilt	peak_pos
a	218.23/22.02	32.96/23.38	0.263/0.084	0.160/0.53	0.070/0.10
rb	129.16/28.2	27.18/22.85	0.180/0.040	0.284/0.88	0.094/0.12
m	211.26/17.73	17.56/14.09	0.229/0.073	-0.039/0.52	0.049/0.08
afb	209.76/17.55	56.62/29.76	0.346/0.080	-0.374/0.28	-0.034/0.08
fb	200.00/23.05	37.84/32.75	0.201/0.077	-0.267/0.45	-0.031/0.12
c	109.81/14.12				
sil	48.73/49.24				

Table B.9: Individual Event/Parameter Results for KDT Mean and Standard Deviation Values (Entries are MEAN/STD)

	start_f0	amplitude	duration	tilt	peak_pos
a	11.87/0.32	11.93/0.24	0.06/0.28	0.50/0.30	0.07/0.17
afb	4.94/0.38	7.70/0.45	0.05/0.25	0.23/0.36	0.05/0.43
fb	7.69/0.22	11.51/0.26	0.06/0.52	0.41/0.08	0.07/0.39
m	8.77/0.53	7.20/0.19	0.05/0.21	0.60/0.1	0.08/0.23
c	No Tree				
sil	No Tree				

Table B.10: Individual Event/Parameter Results for KDT without sub-syllable features (Entries are RMSE/Correlation)

Appendix C

Tilt Parameter Prediction Trees (see disk)

The trees which are associated with experiments and tables within this thesis are located on the accompanying computer disk. Each database is represented by a directory ("f2b," "fhl," and "kdt"). Within those directories, the trees are named according to the following convention:

accents.txt Prediction trees for all accent parameters

fb.txt Prediction trees for all falling boundary parameters

rb.txt Prediction trees for all rising boundary parameters

afb.txt Accent/falling boundary trees

arb.txt Accent/rising boundary trees

minora.txt Minor accent trees

sil_c.txt Silence and connection trees.

C.1 F2B Trees

The trees in this section correspond to Table B.2 from Appendix B. These trees are found on the accompanying disk, in directory “f2b.”

C.2 FHL Trees

This section contains two sets of decision trees. The first set corresponds to the standard feature set, as shown in Table B.6, Appendix B. The second set of trees was trained without the subsyllable features, as shown in Table B. The standard feature trees are located in directory “fhl,” subdirectory “standard.” The trees trained without subsyllable features are in “fhl,” subdirectory “nosubsyl.”

C.3 KDT Trees

This section contains KDT parameter prediction trees which corresponds with Table B.8 in Appendix B. These trees are found in directory “kdt.”

Appendix D

Stimuli for Synthesis Perception Experiment

All stimuli for this experiment come from the Boston University Radio News Corpus [OPSH95]. A label precedes the transcript of each utterance. This label notes the story number (f2bst##), paragraph number (p#), and section number (s#).

Introduction Page Utterances:

f2st03p5s1 : Boston is already divided, says Boscan, in terms of class, race and ethnicity.

f2bst06p2s2 : But by the time the drug shows up in Boston...

f2bst09p4s4 : ... should the state have to buy out some land owners.

f2bst32p4s1 : Several prominent Democrats in the environmental movement...

f2bst14p3s1 : Attorney General James Shannon.

f2bst07p4s2 : ... but not at Barney Frank, who paid for sex when he was still in the closet.

Test Stimuli:

f2bst29p4s4 : For WBUR, I'm Margo Melnicove.

f2bst05p3s2 : ... while they await a state sanitarian who may never show.

f2bst12p6s2 : ... Joseph Ierna, is another longtime Ballaga observer.

f2bst29p4s3 : The same amount, says UNICEF, that the worlds nations spend on weapons each day.

f2bst05p3s3 : Director of the Division of Healthcare Quality, Priscilla Plato ...

f2bst09p5s2 : ... it will take at least a year to finalize regulations ...

f2bst09p5s1 : If the measure wins the legislature's final approval before this session ends ...

f2bst34p1s7 : The operation would also be good for Marshall's profit margin

f2bst34p1s2 : Marshall says it would take about ten million dollars to equip the building for glass making ...

f2bst29p4s1 : Meyere says every chief of state will profess his or her love for children ...

f2bst29p4s2 : ... but their budgets primarily reflect a love affair with the arms race.

f2bst32p4s2 : ... including former U.S. Senator Paul Tsongas who are backing Silber, agree.

f2bst03p5s2 : And in a fragmented city he says, one finds a great deal of violence.

f2bst34p1s4 : Marshall says the venture would be good for Grafton ...

f2bst32p4s3 : Attorney Douglas McDonald, a specialist in environmental law

f2bst06p2s6 : The hospital, the Mayor's Office, the school system, police and others ...

f2bst14p2s2 : the criminal justice system's state of total crisis.

f2bst18p2s1 : Jonie says the old ways are also important.

f2bst32p4s6 : ... and says the quote beaver thing was unfortunate.

f2bst09p4s6 : ... and the state may be hit with a series of costly lawsuits.

f2bst32p4s4 : ... says Silber's got his vote because of his anti-C.L.T. stance.

f2bst34p1s6 : And Marshall says, it would be good for the environment ...

f2bst29p4s7 : The same amount, says UNICEF, that the worlds nations spend on weapons each day.

f2bst07p4s1 : Boston city counselor David Scondras who's gay is plenty angry ...

Bibliography

[Adr91] L. Adriæns. *Ein Modell deutscher Intonation*. PhD thesis, Technical University Eindhoven, 1991.

[AEFN97] C. Astesano, R. Espesser, E. Flachaire, and P. Nicolas. Tonal configurations of prominence in French. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 29–32, Athens, September 1997. ESCA.

[ALM98] A. Arvaniti, B. Ladd, and I. Mennen. Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, 26:3–25, 1998.

[Bar97] K. Bartkova. Some experiments about the use of prosodic parameters in speech recognition system. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 33–36, Athens, September 1997. ESCA.

[BDTss] A.W. Black, K. Dusterhoff, and P. Taylor. *Using the Tilt intonation model for speech synthesis: a data driven approach.* in press.

[Bea94] F. Beaugendre. *Une étude perceptive de l'intonation du fran cais.* PhD thesis, l'Universitè Paris XI, 1994.

[BFO84] L. Breiman, J. Friedman, and R. Olshen. *Classification and Regression Trees.* Wadsworth and Brooks, Pacific Grove, CA, 1984.

[BH96] A. Black and A. Hunt. Generating F_0 contours from ToBI labels using linear regression. In *ICSLP 96*, Philadelphia, Penn., 1996.

[BHJ93] P.C. Bagshaw, S.M. Hiller, and M.A. Jack. Enhanced pitch tracking and the processing of F_0 contours for computer aided intonation teaching. In *Proc. Eurospeech '93*, volume 2, pages 1003–1006, 1993.

[BP86] M. Beckman and J. Pierrehumbert. Intonational structure in Japanese and English. In C. Ewan and J. Anderson, editors, *Phonology Yearbook 3*, pages 255–309. Cambridge University Press, 1986.

[BTC98] A.W. Black, P. Taylor, and R. Caley. *The Festival Speech Synthesis System: system documentation.* The Centre for Speech Technology Research, University of Edinburgh, 1.3 edition, 1998. http://www.cstr.ed.ac.uk/projects/festival/manual-1.3.0/festival_toc.html.

[CB97] N. Campbell and M. Beckman. Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 67–70. Athens, September 1997.

[CD99] R.A.J. Clark and K.E. Dusterhoff. Objective methods for evaluating synthetic intonation. In *Proceedings Eurospeech 1999*, 1999.

[CFHV97] E. Campione, E. Flachaire, D. Hirst, and J. Veronis. Stylisation and symbolic coding of F_0: a quantitative model. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 71–74. Athens, September 1997.

[Cla99] R.A.J. Clark. Using prosodic structure to improve pitch range variation in text to speech synthesis. In *Proceedings, ICPhS 1999*, 1999.

[CR97] P. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings, Eurospeech 1997*, 1997.

[DB97] K. Dusterhoff and A. Black. Generating F_0 contours for speech synthesis using the Tilt intonation theory. In *Proceedings of ESCA Workshop on Intonation*, Athens, Greece, 1997.

[DBT99] K. E. Dusterhoff, A.W. Black, and P. Taylor. Using decision trees within the Tilt intonation model to predict F_0 contours. In *Proceedings Eurospeech 99*, 1999.

[DCH86] A. Di Cristo and D.J. Hirst. Modelling French micromelody. *Phonetica*, 3(1):11–30, 1986.

[Dig92] V. Digilakis. *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition.* PhD thesis, Boston University, 1992.

[dP83] J.R. de Pijper. *Modelling British English Intonation.* Foris, Dordrecht, 1983.

[Dus98] K. Dusterhoff. An investigation into the effectiveness of sub-syllable acoustics in automatic intonation analysis. In H. King and C. Whincop, editors, *Proc. University of Edinburgh Linguistics/Applied Linguistics Departments 1998 Joint Postgraduate Conference*, http://www.ling.ed.ac.uk/research/, 1998.

[Edw84] A.L. Edwards. *An introduction to linear regression and correlation*. W.H. Freeman and Company, second edition, 1984.

[Fan60] G. Fant. *Acoustic Theory of Speech Production*. Mouton, 1960.

[Fuj83] H. Fujisaki. *The Production of Speech*, chapter Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing, pages 39–56. Springer-Verlag, 1983.

[GD82] J. Greene and M. D'Olivera. *Learning to use statistical tests in psychology*. Open University Press, 1982.

[Gol76] J. Goldsmith. *Autosegmental Phonology*. PhD thesis, MIT, 1976. Published 1979 by Garland Press.

[Grø95] N. Grønum. Superposition and subordination in intonation: a non-linear approach. In *Proceedings of ICPhS*, volume 2, pages 124–131, Stockholm, 1995.

[GW80] J. Gandour and B. Weinberg. On the relationship between vowel height and fundamental frequency: Evidence for esophageal speech. *Phonetica*, (37):344–354, 1980.

[Her98] D. J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41:73–82, February 1998.

[HL91] B. Hayes and A. Lahiri. Bengali intonational phonology. *Natural Language and Linguistic Theory*, 9:47–96, 1991.

[Hun88] M.J. Hunt. Evaluating the performance of connected-word speech recognition systems. In *Proceedings, International Conference on Acoustics, Speech, and Signal Processing, 1988*, 1988.

[HW92] J. Hirschberg and G. Ward. The influence of pitch range, duration, amplitude, and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, 20(2):241–251, 1992.

[IP88] S.D. Isard and M. Pearson. A repertoire of British English intonation contours for speech synthesis. In *Proceedings of Speech 1988*, volume 4, pages 1233–1240, 1988.

[JMD] M. Jilka, G. Mohler, and G. Dogil. Rules for the generation of ToBI-based American English intonation.

[Lad83] D.R. Ladd. Phonological features of intonational peaks. *Language*, 59:721–759, 1983.

[Lad90] D.R. Ladd. Metrical representation of pitch register. In J. Kingston and M. Beckman, editors, *Between the Grammar and the Physics of Speech*, number 1 in Papers in Laboratory Phonology, pages 35–57. Cambridge University Press, 1990.

[Lad96] D.R. Ladd. *Intonational Phonology*. Cambridge University Press, 1996.

[LF87] A. Ljolje and F. Fallsade. Recognition of isolated prosodic patterns using Hidden Markov Models. *Computer Speech and Language*, 2:27–33, 1987.

[LFFS99] D.R. Ladd, D. Faulkner, H. Faulkner, and A. Schepman. Constant "segmental anchoring" of F_0 movements under changes in speech rate. 1999.

[LP84] M. Liberman and J. Pierrehumbert. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oerhle, editors, *Language sound structure*, pages 157–233. MIT Press, 1984.

[LS84] D.R. Ladd and K.E.A. Silverman. Vowel intrinsic pitch in connected speech. *Phonetica*, 41:31–40, 1984.

[MBd97] P. Mertens, F. Beaugendre, and C. d'Alessandro. Comparing approaches to pitch contour stylization for speech synthesis. In *Progress in Speech Synthesis*, pages 347–364. Springer, 1997.

[MC98] G. Möhler and A. Conkie. Parametric modeling of intonation using vector quantization. In *Proceedings Third International Workshop on Speech Synthesis*, 1998.

[ML90] A.I.C. Monaghan and D.R. Ladd. Symbolic output as the basis for evaluating intonation in text-to-speech systems. *Speech Communication*, pages 305–314, 1990.

[Möb95] B. Möbius. Components of a quantitative model of German intonation. In *Proceedings ICPhS 95*, volume 2, pages 108–115, 1995.

[MYC91] Y. Medan, Eyal Y., and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Transactions on Signal Processing*, 39(1):40–48, January 1991.

[NV86] M. Nespor and I. Vogel. *Prosodic Phonology.* Foris Publications, 1986.

[OA61] J.D. O'Connor and J.F. Arnold. *Intonation of colloquial English.* Longman, London, 1961.

[Ode89] C. Ode. *Russian Intonation: a Perceptual Description.* Rodopi, 1989.

[OPSH95] M. Ostendorf, P. Price, and S. Shattuck-Huffnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, 1995.

[OR97] M. Ostendorf and K. Ross. A multi-level model for recognition of intonation labels. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody*, pages 291–308. Springer, 1997.

[Pet86] N.R. Petersen. Perceptual compensation for segmentally conditioned fundamental frequency perturbation. *Phonetica*, 43(1):31–42, 1986.

[Pie80] Janet B. Pierrehumbet. *The Phonology and Phonetics of English Intonation.* PhD thesis, MIT, 1980.

[Pik45] K. Pike. *The Intonation of American English.* University of Michigan Press, 1945.

[Por97] T. Portele. Perceptual evidence for accent categories: preliminaries and first results. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 271–274. Athens, September 1997.

[PvSH95] P. Prieto, J. van Santen, and J. Hirschberg. Tonal alignment patterns in Spanish. *Journal of Phonetics*, 23(4):429–451, 1995.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech signal processing*. Prentice Hall, 1993.

[RO94] K. Ross and M. Ostendorf. A dynamical system model for generating F_0 for synthesis. In *Proc. ESCA Workshop On Speech Synthesis*, pages 131–134, Mohonk, NY, 1994.

[Ros94] K. Ross. *Modeling of intonation for speech synthesis*. PhD thesis, Boston University, College of Engineering, 1994.

[RP80] N.R. Reinholt Petersen. The effect of consonant type on fundamental frequency and larynx height in danish. In *Annual Report of the Institute of Phonetics*, number 14, pages 317–354. University of Copenhagen, 1980.

[SBP+92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labelling English prosody. In *Proc. ICSLP*, pages 867–870, 1992.

[SCM+] A.K. Syrdal, A. Conkie, G. Möhler, M. Jilka, K. Dusterhoff, and A.W. Black. *Three Methods of Intonation Modeling*.

[Sil87] K.E.A. Silverman. *The structure and processing of fundamental frequency contours*. PhD thesis, Cambridge University, 1987.

[SMD+98] A. Syrdal, G. Möhler, K. Dusterhoff, A. Conkie, and A.W. Black. Three methods of intonation modeling. In *Proceedings 3rd ESCA Workshop on Speech Synthesis*, pages 305–310, 1998.

[Spr98] R. Sproat, editor. *Multilingual Text-to-Speech Synthesis*. Kluwer, 1998.

[SvHP97] A.M.C. Sluijter, V.J. van Heuven, and J.J.A. Pacilly. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1):503–513, January 1997.

[Tay92] P. Taylor. *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh, 1992.

[Tay00] P. Taylor. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 2000.

[TCB98] P. Taylor, R. Caley, and A.W. Black. *The Edinburgh Speech Tools Library*. The Centre for Speech Technology Research, University of Edinburgh, 1.0.1 edition, 1998. http://www.cstr.ed.ac.uk/projects/speechtools.html.

[tH91] J. 't Hart. F_0 stylization in speech: Straight lines versus parabolas. *Journal of the Acoustic Society of America*, 90(6):3368–3370, 1991.

[tHCC90] J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: An experimental phonetic approach to speech melody*. Cambridge University Press, 1990.

[VAKA97] M. Vainio, T. Altosaar, M. Karjalainen, and R. Aulanko. Modeling Finnish microprosody for speech synthesis. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 309–312. Athens, September 1997.

[vBP90] R. van Bezooijen and L.C.W. Pols. Evaluating text-to-speech synthesis: some methodological aspects. *Speech Communication*, pages 263–270, 1990.

[vSH94] J.P.H. van Santen and J. Hirschberg. Segmental effects on timing and height of pitch contours. In *ICSLP*, volume 2, pages 719–722, Yokohama, 1994.

[vSM97] J. van Santen and B. Möbius. Modeling pitch accent curves. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 321–324. Athens, September 1997.

[vSSM98] J. van Santen, C. Shih, and B. Möbius. *Multilingual Text-to-Speech Synthesis*, chapter 6, pages 141–190. Kluwer, 1998.

[WT97] H. Wright and P. Taylor. Modelling intonation using Hidden Markov Models. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, models and applications*, pages 333–336. Athens, September 1997.

[YJO+96] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *HTK manual*. Entropic, 1996.