Generating Synthetic Pitch Contours Using Prosodic Structure

Robert A. J. Clark

A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy

to
The Department of Linguistics,
University of Edinburgh

April 2003

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Robert A. J. Clark

Abstract

This thesis addresses the problem of generating a range of natural sounding pitch contours for speech synthesis to convey the specific meanings of different intonation patterns.

Where other models can synthesise intonation adequately for short sentences, longer sentences often sound unnatural as phrasing is only really considered at the sentence level. We build models within a framework of prosodic structure derived from the linguistic analysis of a corpus of speech. We show that the use of appropriate prosodic structure allows us to produce better contours for longer sentences and allows us to capture the original style of the corpus. The resulting model is also sufficiently flexible to be adapted to suitable styles for use in other domains.

To convey specific meanings we need to be able to generate different accent types. We find that the infrequency of some accent and boundary types makes them hard to model from the corpus alone. We address this issue by developing a model which allows us to isolate the parameters which control specific accent type shapes, so that we can reestimate these parameters based on other data.

Acknowledgements

Thanks to Andy Breen and Julian Page who started this project during their time at BT Labs. Many many thanks to my supervisor Steve Isard for his enduring patience and abundance of knowledge on all subjects. Thanks to everyone at CSTR, TAAL and HCRC, particular thanks to those that have kept me employed over the years and to those who have kept me sane. Thanks to Alan Black for continued advice on Festival, and thanks to the people involved with the M-PIRO and Magicster projects for providing the need and motivation to improve the generation of intonation. Thanks to Simon King and Simon Kirby, for staying when everyone else seems to be leaving. Thanks to my parents for always encouraging and never discouraging, and finally thanks to Cass for everything.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The primary goal of the work described in this thesis is to improve the generation of synthetic intonation for speech synthesis. However, we will go about this in such a way as to meet two secondary goals: The first of them is to understand how linguistic theory can be appropriately exploited to improve the generation of synthetic intonation. The second is to provide a flexible and robust intonation model. This model must be able to generate different intonation patterns for the same input text to match different intended meanings of the text. These goals will become more specific as the thesis proceeds.

1.1 Intonation Modelling

Intonation modelling for speech synthesis is now one of the big issues facing speech synthesis systems. The quality of synthesised phonetic material has progressed sufficiently that "what is being said" is sufficiently clear that "how is it being said" is a question which is raised in the mind of the listener.

From the perspective of speech synthesis we are addressing the question: "How do we best model intonation?" This leads us to first ask: "What is intonation?".

As we shall see definitions of intonation vary. Ladd (1996, p. 6) says: "Intonation, as I will use the term, refers to the use of *suprasegmental* phonetic features to convey 'post-lexical' or *sentence-level* pragmatic meanings in a *linguistically structured* way." whereas Cruttenden (1997, p. 7) says: "Intonation involves the occurrence of recurring pitch patterns, each of which is used with a set of relatively consistent meanings, either on single words or on groups of words of varying length."

We will address this issue of definitions in section 1.2.2, where we will make it clear what we mean when we talk about intonation.

The problem of intonation modelling for speech synthesis is summed up by the following quote regarding segmental effects on pitch:

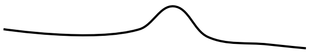
> However, we believe that our understanding of perception of pitch in fluent, meaningful speech is currently not sufficient to make strong claims about the *im*perceptibilily of any aspect of speech, so currently we have no other option but to model any effect on any acoustic feature that can be clearly demonstrated in natural speech.
>
> (van Santen & Hirschberg 1994)

We have some basic intuitive ideas about what natural pitch should sound like, but we just don't understand enough to know how the pitch associated with a specific segment, in a specific syllable with a specific accent, in a specific word in a specific phrase with a specific phrase type, in a specific context, spoken by a specific speaker, should behave.

To actually generate intonation for speech synthesis we need:

1. A formal description of the intonation in terms of a given intonation theory, be it ToBI (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg 1992), Tilt (Taylor 1994), IPO ('t Hart, Collier & Cohen 1990) or something else. Description systems are discussed further in the next chapter. A description is used to specify the type of intonation we require. We may be given this information explicitly, for example in a concept to speech system. We may be given hints to what this should be like in terms of some mark-up, or we may be given only the text and expected to derive this information from it.
2. A way of converting the above intonation description into an appropriate *pitch contour*, or at least a set of target points which represent a pitch contour and possibly some timing modifications to apply to segments.

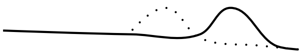
Consider the simple sentence "It was the *best* butter." (Carroll 2000, p. 71). Carroll here uses italics on the word 'best' to signify emphasis. We can consider this a very simple formal description of intonation. The combination of the given emphasis on the word 'best', and the full stop at the end of the utterance strongly suggests , assuming British Standard English, that the pitch contour that we would want to associate with the utterances would look something like:



It was the *best* butter. (1)

The pitch contour describes the change in perceived pitch as the utterance is spoken. For now this change can be considered relative; later on we will need to specify absolute frequency values when we come to actually generate speech.

The point is that we have derived what we expect the pitch, or f0, contour to look like solely from the extra-segmental material that the author usefully provided for us. A *text*-to-speech system would more likely be presented with “It was the best butter.” and default stress rules contained within the synthesiser would probably be applied, placing the the phrasal stress, along with an appropriate *accent* on the final noun, and giving us something like:



It was the best butter. (2)

There are also pitch movements associated with phrase boundaries that need to be considered, but these are usually easier to place as they only occur at the end of phrases, which are generally easier to find than accents, using clues from punctuation or syntax.

So far we have implicitly assumed that an accent is some sort of obvious ‘bump’ in pitch. What these bumps should actually look like turns out to be quite a complex problem, and is one of the questions that this thesis attempts to answer.

There are actually two things we need to model here. We need to model how the pitch range changes over the utterance as a whole and we then need to be able to overlay pitch events successfully onto this range to produce a resulting contour.

Modelling pitch range would not be a major problem for the above examples as the utterances are so short. A pitch contour that ends slightly lower than it starts with the pitch going up and back down again in the right place for the accent would be acceptable.

Pitch range becomes more of an issue in more complex utterances, for example (Carroll 2000, p. 79):

> A large rose-tree stood near the entrance of the garden: the roses growing on it were white, but there were three gardeners at it, busily painting them red. (3)

This utterance consists of a number of phrases, separated by punctuation, but the relationship between adjacent phrases is not necessarily the same. The pitch range will change between adjacent phrases, but not in the same way every time. Once we have identified the phrases, we need a way to express the pitch range of that phrase so that we can independently add appropriate pitch movements within that range. To discuss these issues in more detail we first formalise some of the ideas presented above.

1.2 Definitions of Terms

Speech synthesis and *intonation*, although both seemingly innocent terms, have in common in that you can tell someone on the street that you study them and they can have absolutely no idea what you are talking about. Furthermore, we have already seen that in the case of *intonation*, it is difficult to pin down exactly what intonation is and what it is not. Different experts give slightly different definitions. As the reader may be well acquainted with one field but not the other we will formally introduce both intonation and speech synthesis.

1.2.1 Speech synthesis

Speech synthesis is somewhat easier to pin down than intonation. In the most general sense speech synthesis is the production of 'synthetic' speech using a personal computer or other computing device. By this we mean producing an electronic signal which when played through a speaker or similar transducing device resembles human speech enough for the human brain to interpret it as such. More technically, this means that the signal must contain a reasonable representation of the voicing and the different harmonic resonances associated with the the underlying formants in the vocal tract.

There are currently three general methods that can produce acceptable synthetic speech to varying degrees.

The first method is waveform synthesis. This type of synthesis concentrates on producing an acoustic signal which resembles speech. A source/filter model is generally employed, the source representing the glottal source and the filter representing the effect the vocal tract has on the source (Holmes, Mattingly & Shearme 1964). Systems include: MITalk (Allen 1987), YorkTalk(Coleman 1990) and PAT (Lawrence 1953)

The second method of producing synthetic speech is articulatory speech synthesis. Here an attempt is made to model the articulatory processes that produce speech, by modelling the articulators themselves. Where waveform synthesis models the *effect* of speech, articulatory synthesis models the *cause*. Articulatory synthesis is a hard task, partly due to the difficulty in measuring the real articulatory processes as real speech is produced and partly due to the mathematical and computational complexity needed in the resulting models.

Progress is being made in this measuring of the articulatory processes with various techniques such as electropalletography, x-ray microbeam and EMA. As technology improves and our ability to model such processes progresses articulatory synthesis will become more widespread, but currently articulatory systems are impractical for general use.

The third and currently most popular form of speech synthesis is concatenative speech synthesis which involves chopping up pre-recorded real speech and gluing pieces of it back together to produce a suitable result. Signal processing techniques may be used here to modify the pieces to provide more suitable transitions between pieces. The size of the pieces can vary from sentences or words right down to syllable or phoneme sized pieces. Generally the more signal processing that is required to produce pieces that consistently join together the worse the quality of the resulting speech is. The currently most commonly used pieces are diphones, which are units which start at the steady state centre of one segmental phone and end at the steady state centre of the next. The theory is that these units are easier to concatenate than individual phones due to the steady state at each end. Current research is aiming to enable the use of more general pieces of varying sizes, as this can reduce the amount of signal processing required to join the units together. The term unit selection is usually used to describe this type of synthesis. Examples of concatenative speech synthesis systems include: Festival (Taylor, Black & Caley 1998), rVoice (Rhetorical Systems Ltd. 2000), Laureate (Page & Breen 1996), AT&T NextGen (Syrdal, Wightman, Conkie, Stylianous, Beutnagel, Schroeter, Strom, Lee & Makashay 2000) and DEMOSTHeNES (Xydas & Kouroupetroglou 2001).

A completely separate issue in speech synthesis concerns the type of input to a system. Levelt (1989) considers the human speech act to consist of a chain of events which starts by an idea being formed in the brain. This idea is then formalised as a sentence in a language and then it is spoken. Speech synthesis systems can be considered to follow at least part of this chain.

Traditionally speech synthesis systems have been *Text-To-Speech* (TTS) systems, where the input is a string of typed words equivalent to the formalised sentence in Levelt's terms. The system then converts these into a stream of phonemes, and synthesises them by one of the above methods. However, research is currently shifting towards building systems which perform *Concept-To-Speech*, where the speech synthesiser is tied to a language generation system, the 'brain' in Levelt's terms. These systems have the advantage that they can provide the synthesiser with more information to synthesise from. As well as the words to be spoken, syntactic, semantic, and prosodic information can be given that otherwise the synthesiser would have to derive or predict for itself. This thesis deals primarily with diphone based concatenative synthesis, both from a text-to-speech and a concept-to-speech point of view. As we are primarily concerned with the generation of an appropriate pitch pattern for a given context rather than figuring out a likely context from the word string, will we lean towards a concept-to-speech approach, to avoid some of the complications that having to decide upon suitable intonation from text alone would introduce.

1.2.2 Intonation

As already stated we need to be clear about what we consider intonation to include. There are in fact a number of terms we need to be familiar with: *intonation*, *prosody*, *suprasegmental*, *pitch accent* and *tune*.

We consider *intonation* to be the part of the acoustic speech signal that cannot be accounted for from the segmental structure of an utterance alone. This view fits somewhere in between the views of Ladd and Cruttenden which we saw in section 1.1. This component of speech is described as being *suprasegmental*. Intonation then is a quantifiable entity, unlike *prosody* which we consider to describe the way in which something is being spoken, rather than actually being part of what is being spoken. With this view, intonation can be thought of as the manifestation of the underlying prosody, and the terms can be used interchangeably when this is not a need to differentiate between the abstract and the physical.

For example, back in (1) the prosody dictates that there should be an accent on the word 'best' and an overall decline in pitch across the utterance. This is realised by the illustrated intonation contour.

Part of the difficulty with these terms is that the properties that they describe are often very difficult to tease apart from the segmental structure. For example *stress* is a lexical property: certain syllables in certain words are stressed and certain others are not, but stress and prosody are by no means independent. Emphasis is placed more on some words that others in a spoken utterance and this *prominence* tends to manifest itself in prosodic changes to the stressed syllable in the words. We call this *accenting*. Pitch variation is one particular way of accenting words and syllables, duration and intensity variation are other ways of accenting which can be used along with pitch variation or on their own.

Sometimes the term intonation is just used to refer to the pitch patterns brought about by the prosody. Cruttenden (1997) tends to take this approach. Others tend to use the term intonation to refer to both pitch patterns and the underlying prosody, de Pijper (1983) for example, whilst others, use both prosody and intonation either interchangeably or in the way described above (Beckman & Pierrehumbert 1986, Ladd 1996, Taylor 1994). In this thesis we try to use the term intonation when refering to pitch patterns, and prosody when refering to the underlying structure which governs them.

Intonational Phonology and the *Autosegmental-Metrical* (AM) approach (Liberman 1975, Bruce 1977, Pierrehumbert 1980, Ladd 1996) cover a group of theories of particular interest that take the view that the intonation is comprised of a series of discrete pitch movements. The theories study the underlying structure of prosody and the relationship between this and the resulting intonation.

The AM approach considers the intonation in languages such as English to be made up of sequences of *pitch accents* and *boundary tones* (which may be refered to as *phrase tones* or *edge tones*). Pitch accents are pitch movements which associate with stressed syllables in particular words of phrases, whilst boundary tones associate with the end of phrases.

At the lowest level intonation can be considered to result from a series of prosodic pitch targets. Theories of prosody and intonation try to explain the placement of these targets and the changes of pitch between them. Some theories take the targets (highs and lows) themselves as the main component of intonation (Pierrehumbert 1980) while others take the changes between target levels (rises and falls) as the components of intonation ('t Hart et al. 1990). There is no real evidence to suggest that one method is more correct than the other, or even that they are not mathematically equivalent, and it is probably the viewer's perspective or the use to which the ideas are being put which makes one approach more appropriate than the other.

As our concern is speech synthesis, where the application of a theory is easiest where it can be treated as a sequence of distinct ordered processes, we choose to concentrate our attention on models and theories which are compatable with such an approach. We are interested in two specific models, primarily a model that is based on Pierrehumbert's thesis and later work (Pierrehumbert 1980, Beckman & Pierrehumbert 1986, Pierrehumbert & Hirschberg 1990) which has led to the ToBI (Silverman et al. 1992) annotation scheme. These are discussed in sections 2.1 and 2.3. We will also consider the Tilt model (Taylor 1994, Taylor 1998) (see section 2.7) which provides a model more suited towards automatic analysis and synthesis.

Tune

Tune describes the gross melodic pattern of intonation and captures a notion of the regularities and repetitions found in intonation. For example, in English, questions and statements have a different tunes associated with them, and this difference is one of the ways in which a question is clearly heard as a question.

The choice of tune is related to the role of the utterance in the discourse. Statements of fact often employ *neutral declarative* contours where questions may use *interrogative* contours. That is not to say a given discourse role automatically determines contour type, but there will a subset of tunes appropriate for any given role, which the speaker can employ. (Pierrehumbert & Hirschberg 1990) push the interpretation of tune further as we will see in section 2.1

The pitch pattern associated with a given tune on different phrases will be similar but will not necessarily be identical, as both the amount of segmental material and the number and location and choice of the individual pitch accents themselves will also have an effect on the resulting contour. In the autosegmental-metrical framework, tune defines, or possibly restricts, the type of pitch accents used and through them controls the overall shape of the contour. For example a statement tune may restrict pitch accents to H accents and the final boundary to an L (see section 2.1 for the explanation of H and L).

Pitch range

In simplest terms *pitch range* is just that, the range of pitch employed by a particular speaker at a particular time and can be specified by a minimum and maximum pitch. More complex representations attempt to capture more information about the distribution of the pitch points in frequency space.

Difficulties arise when trying to relate these descriptions to abstract linguistic ideas about pitch range, for example when making comparisons between two instances of pitch range, or in our case modelling pitch range.

The first problem regards the theoretical aspect of what pitch range actually is. For descriptive analyses it is generally assumed to be the particular range of pitch employed by a given speaker. From a production point of view however it could be regarded as the potential range that could be employed for production by a given speaker. Are these the same thing? For example, given some speech from a particular speaker, how do we know if they have used their full pitch range? All we can really do is assume that if we take enough speech, then the speaker will have used their full pitch range.

There are many statistics which can be used to describe pitch range, most of which are suitable for some purposes, but are not particularly suited to ours. Absolute minimum and maximum pitch, for example does describe pitch range, but does not say much about the distribution of points. *Level* and *span* (Ladd 1996) are effectively the same measurements with level being the absolute minimum and span being the difference between maximum and minimum. Variations on this theme include the idea of a *topline* and a *baseline* (Bruce & Gårding 1978) which are effectively maxima and minima which change over the course of the utterance. This idea is theoretically attractive, but difficult to deal with for production purposes as decisions need to be made about how exactly these lines change over the course of an utterance. Liberman & Pierrehumbert (1984) take this notion further in their study of intonational invariance across pitch range where they propose the use of a fixed baseline with a moving *reference line* and ways to control pitch range without whole-phrase preplanning.

All of these maximum and minimum descriptions have their uses. However, they tend not to be specified in an algorithmic way guaranteed foolproof for speech synthesis: for example, the way in which declination is specified may cause problems in a very long utterance and result in very artificial sounding intonation. The other main issue with pitch, and pitch range in particular, is on what scale to measure it: the linear Hertz scale, a logarithmic semitone scale, or one of the more complex scales like Mels or Barks, designed to relate to human perception of pitch. The logarithmic scales are particularly favoured when directly comparing male and female pitch ranges as on a linear scale the span of a female pitch range is about twice that of a male pitch range, whereas on a logarithmic scale they are more or less the same. This may be important in descriptive systems to ease comparison between male and female voices, but is not such a concern for speech synthesis as the computational methods employed allow us to process and analysis pitch range in more complex ways. Here we ususally treat a pitch contour as a series of pitch points sampled at, say, every ten milliseconds. Calculating the mean and standard deviation and other such statistics of these points is reasonably trivial, and using one scale or another to measure pitch rarely make a difference. Explicitly treating pitch range as a distribution also allows us to normalise pitch range and easily compare different pitch ranges, and pitch points within different pitch ranges. (See section 7.3.1.)

1.3 The Road Ahead

We continue in the next chapter by taking a closer look at the literature which discusses intonation, prosody and speech synthesis. In Chapter 3 we consider the tools and resources that we need to carry out the work described here. Chapters 4–6 concentrate on the analysis of the corpus that we have chosen to work with, first looking at phrase structure and then at the accents and boundaries within phrases.

Chapter 7 then describes the development of a framework based around prosodic structure and models built within this framework. Chapter 8 deals with the evaluation of these models and Chapter 9 discusses the improvement of one particular model to be more suitable for other styles of speech. Finally Chapter 10 discusses the outstanding issues we have come across and draws our conclusions.

CHAPTER 2

The Literature

2.1 Pierrehumbert's Theory of Intonation

Pierrehumbert (1980) and later work (Beckman & Pierrehumbert 1986, Pierrehumbert & Hirschberg 1990) has made a great impact on the theory of intonation. The Pierrehumbert theory provides a phonological description of observed pitch range phenomena in a way which is particularly appropriate for use in speech technology. As it is the primary intonation theory used in the work carried out here, it makes a good starting point and comparison for other systems.

In this theory the tune of an utterance is specified as a sequence of tones which form *pitch accents*, *phrase accents* and *boundary tones*. Pitch accents mark prominences. There are two pitch accents made up of single tones: H* and L*, and four made from pairs of tones: L*+H, L+H*, H*+L and H+L*. The diacritic '*' marks the alignment of the tone; that is the starred tone in either a simple pitch accent or a complex one aligns with the stressed word—or some constituent of a stressed word. The unstarred tones in the complex pitch accents lead or trail the starred tone, but it is the starred tone which determines the pitch accent alignment, and which categorises a complex tone as high or low.

Complex utterances are divided into two levels of phrase. An *intermediate phrase* consists of at least one pitch accent followed by a simple H or L (sometimes H- and L-) *phrase accent* or *phrase tone*. An intonational phrase is made up of one or more intermediate phrases followed by an additional tone, referred to as a *boundary tone*, marked by a '%', i.e. either high (H%) or low (L%).

Earlier work in the *British school* framework (Crystal 1969, O'Connor & Arnold 1961) makes a clear distinction between nuclear and pre-nuclear inventories. There the nuclear, generally the last, accent is considered to have a different status to those that precede it. This, in the view of Pierrehumbert & Hirschberg (1990), misses important generalisations between nuclear and pre-nuclear accents which have a distinctly different status within the British framework. Pierrehumbert's (1980) theory drops the distinction between nuclear and pre-nuclear accent.

Pierrehumbert & Hirschberg also claim that that the use of level tones to describe intonation, rather than tone rises and tone falls as in the British school system, allows identical constituents of differing tunes to be equated. The example of H* H-H% and H* L-L% is given, where both tunes contain the tone H*. This information is lost in approaches that use rises and falls. Of course Pierrehumbert & Hirschberg's (1990) argument here is based on the assumption that it is indeed the high and low peaks in the contour that are important. This is reinforced by requiring only two tones, as opposed to the four suggested by Pike (1945) and Liberman (1975). Pierrehumbert's (1980) catathesis rule, which allows for H* accents to be *downstepped* in staircase style sequences is also argued to bring out similarities that would be otherwise missed.

2.2 Catathesis

Catathesis or downstep is a compression and lowering of pitch range, and is effectively what makes the Pierrehumbert theory work with only two tones H and L. Pierrehumbert considers catathesis to be triggered by an H L H tonal sequence which includs a bi-tonal pitch accent such as in the sequence H*+L H*+L L-L%. The result of catathesis is that the second H* is lower than expected at this point in the utterance as the pitch range has been lowered and compressed. This view of catathesis is questioned by Ladd (1983) where it is suggested that downstep is controlled by an independent feature which can be set on given pitch accents. The notation '!' is prepended to accent descriptions to show they are downstepped: H* becomes !H* when downstepped, L*+H becomes !L*+H and so forth. Pierrehumbert & Hirschberg (1990) however reanalyse catathesis in an attempt to resolve issues both with the H L H trigger and the feature interpretation of downstep. They say that downstep is triggered by the presence of H+L combination bi-tonal pitch accents alone.

2.3 ToBI

ToBI (Silverman et al. 1992), short for Tones and Break Indices, is a proposed standard for transcribing English prosody. Originally developed for varieties of American English it has been since adapted to other dialects and other languages (Mayo, Aylett & Ladd 1997, Reyelt, Grice, Benzmüller, Mayer & Batliner 1996, Campbell & Venditti 1995).

The system comprises of a series of parallel labelling tiers. The first is a tonal tier which contains pitch events based on Pierrehumbert's theory. This tier includes labels for the pitch events and often additionally includes 'HiF0' labels to mark the highest f0 peak in each phrase. A break index tier is used to mark breaks on a 0 to 4 scale (0 to 6 in an extended version) which measure the strength of association between adjacent words. And a miscellaneous tier is used for marking hesitations, disfluencies, non-speech and the like. An example of the system in use is shown in figure 2.1.

The break indices with values of 3 and above relate to prosodic boundaries. The 3 is a '-' boundary and 4 to a '%' boundary. In the extended version of the system 5 is used for '%' boundaries stronger than those marked 4, such as those with particularly long pauses, and 6 is used to signify end of utterance.

Part of the philosophy of ToBI is that it provides a framework within which different labellers can be consistent in their labelling. This is an obvious benefit where such labelling is intended for uses in speech synthesis, for example to train statistical models.

This is not to say that there are not any problems with this transcription system. There is an underlying problem in achieving the levels of consistency required for training statistical models brought about by both ambiguity in the tonal tier and the break index tier. The number of different types of accents in Pierrehumbert's theory and the similarity between certain accents, or at least the similarity in accents when realised in speech, presents a problem. On paper, distinct clear accent shapes seem like a good idea. However, in the real world making these distinctions is not always easy: the f0 traces that you see do not always exhibit the clear accent shapes that are expected, so it is not always obvious what type an accent is or even sometimes if there is an accent at all. Syrdal & McGory (2000)

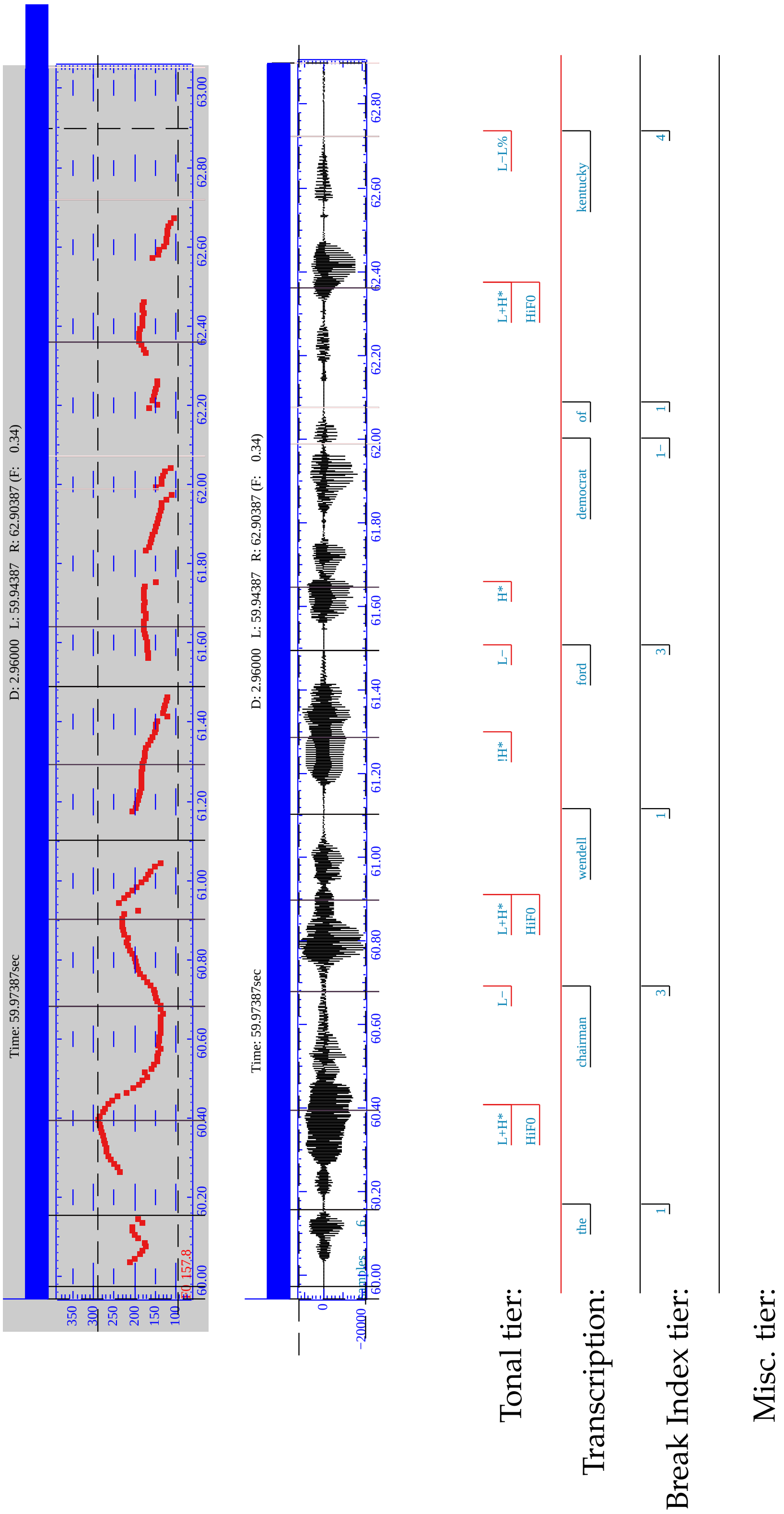


Figure 2.1: Example from ToBI corpus showing (from the top) Pitch contour, speech waveform, tonal tier, transcription, break index tier and misc. tier. In this example the misc. tier is unused and the original break index range of 0–4 is being used. ‘HiF0’ is additionally employed in the tonal tier to mark local pitch maxima.

show that although labellers tend to have a high agreement on the place of accents, the level of agreement on type is much lower, and possibly problematic for training TTS intonation models.

The shape of accents within categories has been shown to vary with the segmental material on which they occur. Grabe's (1998) thesis demonstrates how pitch accents can be compressed or truncated where there is less sonorant material for them to be realised on.

This raises the question of whether accent shapes in general form more of a continuum, and it is only the perception of them which is categorical. The Tilt model (Taylor 1998) (see section 2.7) treats accents in this way.

However, whatever ToBI's drawbacks are it has certainly furthered intonation research by the sheer fact that it allows researchers to present intonation patterns to each other in a standard way which can be easily understood.

2.4 Pitch Lowering Effects

We saw in section 2.2 that the progressive lowering of pitch in an utterance was accounted for in Pierrehumbert's (1980) approach by catathesis or downstep. However, this is not the only way to account for such pitch lowering. Some theories attribute pitch lowering to *declination* (Cohen & 't Hart 1967) rather than downstep. Declination describes a more general lowering of pitch across the utterance as a whole, rather than a stepped lowering related to specific pitch events. The models considered in the next section attribute pitch lowering to declination rather than catathesis.

Extra pitch lowering that occurs at the end of phrases or utterances, where the pitch reaches a lower level than elsewhere in the utterance, is often called *final lowering*. Final lowering is compatible with both declination and catathesis views of pitch lowering and can be found incorporated into models of both types. Whatever view of pitch lowering is taken at some point, either within an utterance or between utterances, pitch will reach a lowest point and then jump back to some higher level. This is usually called *reset* or *declination reset*.

2.5 Other Models of Interest

There are various other intonation models described in the literature which are aimed at speech synthesis which we consider for completeness:

2.5.1 The IPO approach

In the IPO approach, originally for Dutch intonation and later for English intonation (de Pijper 1983) the natural pitch contour is 'replaced' by a series of discrete stylised pitch movements, which have been specified in a standardised way to be perceptually equivalent to the original contour. A grammar specifies which pitch movements can be used at a given time.

The approach uses *pitch movements* rather than *pitch levels* as its atomic units, but these movements occur between three levels of pitch making eight distinct movements. These movements comprise of steep and shallow rises and falls between either two adjacent levels of the pitch range or across all three levels. Additionally each movement can be aligned with a syllable in three ways; denoted early, middle or late; leading to 24 movements in total.

The IPO approach takes the view that declination (see section 2.4) causes the downward trend in f0 as an utterance progresses. Both the top and bottom lines which control the position of the high and low targets decline through the utterance.

A typical example (taken from Ladd (1996)) is the 'hat pattern' which is a 'type 1 rise' (a low to high rise early in the accented syllable) followed by a 'type A fall' (a high to low fall early in the accented syllable). If these two movements occur as part of the same accent then the result is a 'pointed hat' otherwise they result in a 'flat hat' with a stretch of flat contour between rise and the fall. Phonologically the rise and the fall are considered obligatory, whereas the flat stretch between the rise and fall, along with flat stretched preceding the rise and following the fall are considered optional.

The approach makes direct use of resynthesis techniques to produce the stimuli for the perceptual evaluations of what parts of the intonation contour are salient to the listener for the purpose of developing the set of pitch movements. Willems, Collier & 't Hart (1988) extends de Pijper's (1983) description for English and provides quite a complex rule based approach for the synthesis of British English intonation.

2.5.2 *Superposition models*

The main other type of model is the overlay or superposition model. Here the complex pitch contour is regarded as being made up simpler signals superimposed on top of each other. These tend to separate out accent movement from declination as independent components.

The best known example of such a model is the Fujisaki model (Fujisaki 1983) which is a generative model. Here the f0 contour is comprised of a phrase component and an accent component. The components are referred to as 'commands' and are represented in the frequency domain by a sequence of impulse responses and a series of step functions respectively. The functions are then added in such a way as to produce a smooth f0 contour in the time domain. Figure 2.2 shows the structure of the model along with an example of what it produces.

The superpositional approach can also be seen in some more theoretical linguistic models of intonation, a good example being that of Grønnum (1992). Her hierarchical intonation model of Danish overlays components representing different temporal scopes, the longest being the length being the paragraph level and the shortest being that of a *stress group*.

This idea of dual components is often dismissed in intonation theory, possibly because this element of the intonation is often less interesting to the researcher. It is studied to some extent in the form of declination, and in English at least a general lowering of pitch is expected as a phrase progresses.

Realistically, any intonation model used for speech synthesis is likely to have an underlying phrase component, even if it is considered to be flat, to control the pitch range of a given speaker. It is likely in speech synthesis that downstep can accomplish the same effect as a phrase component could if used in the right way.

One of the goals of this thesis is to examine to what extent a phrase component can be teased apart from the f0 contours, and how accounting for phrasing effects on pitch range can give us more control over how we describe and position pitch events using theories which do not in their own right recognise a phrase component.

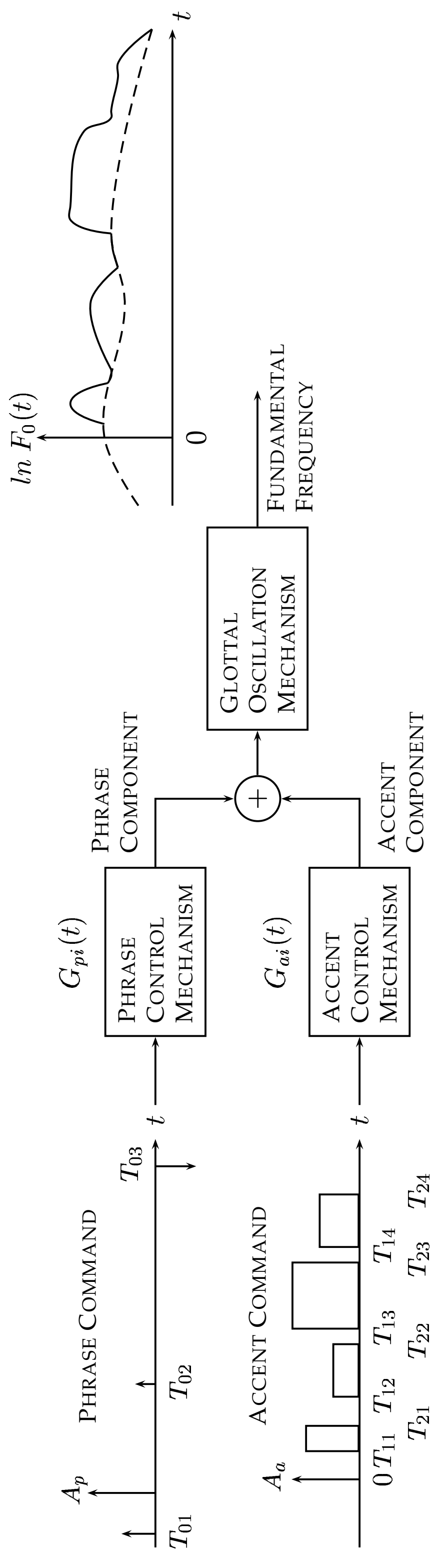


Figure 2.2: The Fujisaki model of intonation. In the frequency domain, a phrase component consisting of a series of impulses is added to an accent component consisting of a series of steps. The resulting contour in the time domain consists of a series of decays generated by the phrase component overlaid with a series of rise and falls generated by the accent steps.

2.6 Accent Alignment

So far we have only explicitly considered the placement of accents in terms of their pitch. Accents also align with the segmental material of an utterance in a particular way. To avoid confusion we adopt the term *positioning* to describe pitch placement and reserve *alignment* to describe the placement with respect to the segmental material.

Santen, Shih & Möbius (1998) discuss alignment based on the *sonorant rhyme* (or *s-rhyme*) which they define as being the part of the syllable from first non-initial sonorant through to the end of the last sonorant (see section 2.12.2 for more detail and a description of their proposed linear model of peak position dependent on various duration measures including this s-rhyme duration). The unusual definition of the s-rhyme is interesting as it gives an idea of the level of detail at which the syllabic structure needs to be examined, suggesting that the relationship between pitch event and syllabic structure is not necessarily a simple one.

Assuming then that pitch events align with syllables in some fashion, to model them we need to be aware of just how much of the speech we need to include to ensure that we have the full syllabic context which governs alignment. In other words, is considering the syllable that the pitch event aligns with in isolation enough? If not, what other syllables and syllabic information do we need to consider? Do we need the preceding and following syllables or do we need to look at the foot or the word? Arvaniti, Ladd & Mennen (1998) for example, show that for Greek, accent alignment can certainly occur with the next syllable under certain conditions.

With accent alignment in mind we now consider a model which explicitly accounts for an accent's alignment as well as its position.

2.7 The Tilt Intonation Model

The tilt intonation model (Taylor 1994, Taylor 1998) is a model orientated towards speech technology. It is a descriptive model which provides a parameterised representation of the change in pitch related to intonation events. It therefore makes very few assumptions about the underlying intonation theory. It assumes only that pitch events occur in a linear fashion at given times and have distinct starts and ends, and is appropriate to describe an intonation theory based around peaks and troughs (pitch movements) or rises and falls (pitch targets). The standard use of the theory assumes that there are two types of pitch events: accents and boundary tones, but, as they are treated in exactly the same manner, one could use only one type if a theory dictated.

Each event is characterised by a set of five independent parameters which completely describe the pitch movement:

Amplitude The amplitude of the event.
Duration The duration of the event.
Tilt A measure of the shape on the interval [-1:1]. -1 is a pure rise, 0 is a rise-fall and +1 is a pure fall. (See equations 2.1–2.4 below.)
Position A measure of the f0 position relative to a baseline (usually 0 Hz)
Time A measure of the time position of the event.

The amplitude and duration parameters can be extracted directly from a labelled intonation contour. Figure 2.3 shows the way the tilt parameter affects the accent shape. The time and position parameters are flexible in what they are measured relative to. Time is usually measured relative to the start of the vowel that the pitch event is assigned to.

The tilt model is usually used in conjunction with a labelling scheme which uses a simple set of labels to identify accents and boundaries. 'a' is used to signify all types of pitch accent and 'b' or sometimes 'rb' and 'fb', to distinguish between rising and falling, are used for boundaries. Where a single pitch movement can be attributed to an accent and a boundary a single label is used. For example 'afb' would be used to label an accent and falling boundary. From the point of view of a strict phonetic description, this may be better than a ToBI description as it makes a distinction between where there are separate pitch movements for each event and where a single pitch movement occurs for multiple events.

Along with the theory, Taylor provides an algorithm (an implementation of which is included with Festival) for automatic analysis of pitch events in terms of tilt. Given an f0 contour and the above labels marking pitch events, the necessary parameters can be extracted, or derived in the case of the tilt parameter, automatically from the data using the mathematical definition of tilt specified in equations 2.1–2.4. The pitch event is first split into a rise portion and a fall portion and the amplitudes A_{rise} and A_{fall}, and durations D_{rise} and D_{fall}, of each part are calculated from the pitch contour.

The amplitude and duration parameters are calculated by summing their respective components and the following intermediate parameters are then calculated (not to be confused with the actual amplitude and duration parameters themselves):

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{2.1}$$

$$tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{2.2}$$

(2.3)

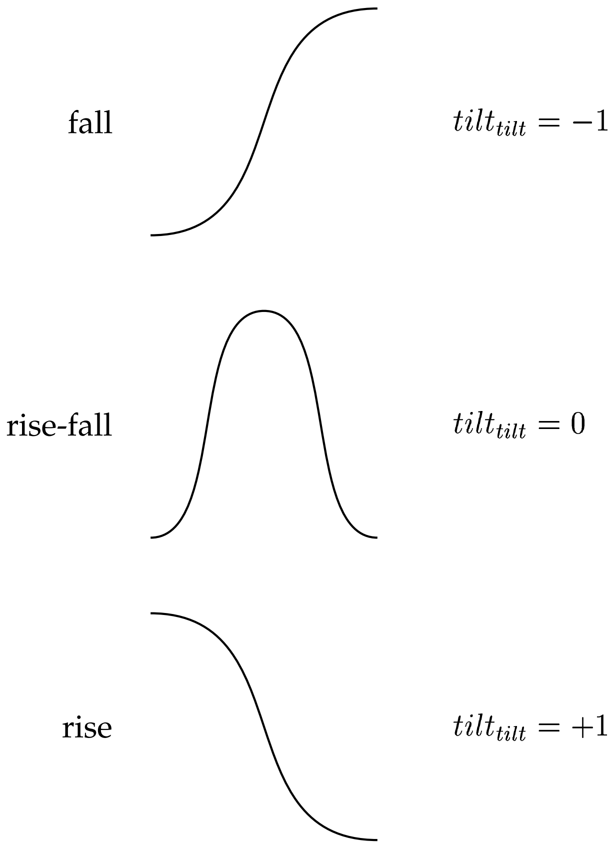


Figure 2.3: Pitch event shapes as they relate to Tilt parameter value.

and from them the tilt parameter is derived:

$$tilt_{tilt} = \frac{tilt_{amp} + tilt_{dur}}{2} \tag{2.4}$$

It is easier to produce more consistent hand labelled speech data with this model than with ToBI as the labeller, human or automatic, needs to make fewer decisions when assigning a label as most of the parameters are calculated automatically from the data. There are no specific accent or boundary type categorisations inherent in the model like in the Pierrehumbert theory, but Taylor (2000) illustrates how the parameter space could be partitioned into categories if so desired (see figure 2.4).

There is a particularly large overlap between the H*, L+H* and H*+L categories in figure 2.4. This suggests that there is not an clear invertible mapping between these phonological categories and their phonetic realisations. This is essentially saying that pitch contours of some H*, L+H* and H*+L accents look the same. This may have implications for the generation of intonation as it suggests that the data that models are trained on may contain ambiguities.

Tilt works well as a descriptive system in that it is relatively easy to assign tilt parameters to given stretches of pitch contour, and tilt can describe the variation

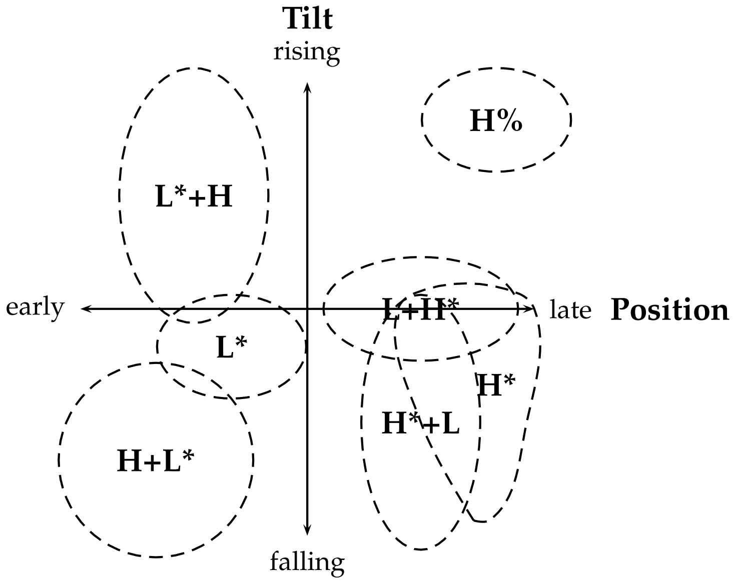


Figure 2.4: Taylor's schematic suggesting a comparison between Pierrehumbert style accent categorisations and tilt parameters.

in pitch event realisation to a great extent, but it does not provide a mechanism for accounting for that variation in terms of a linguistic interpretation, which effectively makes it unsuitable for modelling intonation for speech synthesis on its own. Rather, tilt is a good candidate to model the particular shapes of accents of predetermined phonological categories. The lack of a large dataset containing a variety of different pitch accent types that is consistently labelled has probably contributed to the lack of interest shown in tilt in speech synthesis systems to date.

Tilt is used in the analysis phase of the work described in this thesis as tilt parameters are available for pitch events in the dataset used and they provide a useful way of finding the peak position of pitch accents.

2.8 Prosodic Structure

When studying intonation alone, finding and describing an accent of a given type in a given place is often the final goal, but in the context of speech synthesis knowing that an accent is found in a certain place under certain conditions is not enough. We need to be able to recreate that accent, that is we need to know how to place such an accent with respect to its prosodic context in a given speaker's pitch range. To do this we need some understanding of the underlying prosodic structure and some understanding of how this relates to pitch range. Prosodic structure is used to account for high level patterns in intonation, such as the difference in pitch range between two phrases of the same utterance. Different theories take different views to how prosodic structure should be represented.

2.8.1 Hierarchical structure

We next need to concern ourselves with the questions: To what size chunks of speech do tune and pitch range apply, and how do these chunks relate to each other? The term IP (*intonational phrase*) is often used to refer to such chunks of speech, but there are different ideas concerning how big such a chunk is and what it should be called. As Ladd (1986) points out, the definitions of the phonemic clause (Trager & Smith 1951), the macro-segment (Hockett 1958), the tone group (Halliday 1967), the breath group (Lieberman 1967) and the intonational phrase (Pierrehumbert 1980), all describe potential IPs in one way or another.

Ladd (1986) then summarises these IP definitions as the largest phonological chunks into which utterances are divided. They extend from one *phonetically definable boundary* to the next, have a specifiable intonational structure, and are phonological units which are assumed to relate to syntactic or discourse level structure.

However, Ladd (1986) points out there is a problem with the definition of boundaries: the general assumption is that the domains over which phonological structure is specified are defined by the audible phonetic boundaries which occur in the speech stream. If something is structurally an IP then it is assumed to have boundaries, and if something has boundaries it is assumed to be an IP, leading to a somewhat circular definition.

Two such IP levels of phrasing are generally assumed. These are proposed in various ways, but basically they consist of big intonation phrases (IPs) containing littler intonation phrases (ips): e.g. single and double bar boundaries (O'Connor & Arnold 1961), major and minor tone groups (Trim 1959) and intermediate phrases and intonational phrases (Beckman & Pierrehumbert 1986). The general definition is that the little ip has the nuclear structure, that is contains one primary stressed unit, and the big IP consists of little ips and has an audible break associated with it. This is not to say that all of these definitions of big versus little are the same, just that they all accept two levels of phrasing.

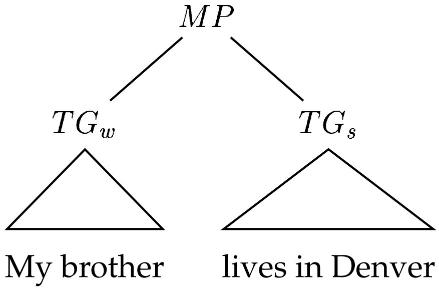
Ladd (1986) also takes this view with his major phrase (MP) and tone group (TG), but in a hierarchical fashion, defined thus:

> **Major Phrase** An MP is set off by audible prosodic breaks—rhythmically organised[1] pauses marked by actual silence and/or the prolongation of the pre-pause syllable, accompanied, in many cases, by the additional pitch movements (boundary tones, in the current terminology) such as a rise following an accentual fall. [Ladd's (1986) footnote]
>
> **Tone Group** A TG, on the other hand, is merely a structural unit of intonational phonology—the domain within which a nucleus is

[1] See Scott (1982) for instrumental evidence that the duration of boundary pauses depends on the place of the pause in the foot of the structure of the utterance; for a theoretical treatment see Selkirk (1984, ch. 6).

> defined—and the boundary between TGs need not be accompanied by any rhythmic break or additional pitch movement at all. That is an MP is broken down in to TGs such that each TG contains only one nuclear accent. The boundaries between TGs may not be marked by pauses as the MP boundaries are.

For example:



(4)

Where the ‘s’ and ‘w’ represent a ‘strong’ versus ‘weak’ metrical relationship between the two TGs.

The hierarchical approach is in distinct contrast to the main theoretical position of Pierrehumbert (1980), which is that intonation contours are considered linear strings of tones, although it is not completely incompatible with it, as there is no reason why TGs cannot be considered to consist of strings of tones.

The strict layer hypothesis (SLH)

A general assumption often made with prosodic structure is that it is non-recursive, that is any given level of structure has to be made up of only of units of the level below it. Ladd questions this in two respects; consider:

The book on the table, it seems to me, was a gift from my mother. (5)

Cooper & Sorensen (1981) found that declination is interrupted by the parenthetical and continues after it as if it wasn’t there. This suggests the structure:

[The book on the table [it seems to me]$_{MP}$
was a gift from my mother]$_{MP}$ (6)

rather than:

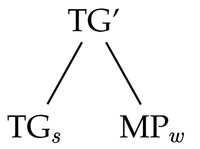
[The book on the table]$_{MP}$ [it seems to me]$_{MP}$
[was a gift from my mother]$_{MP}$ (7)

as would be imposed by the SLH.

Secondly consider:

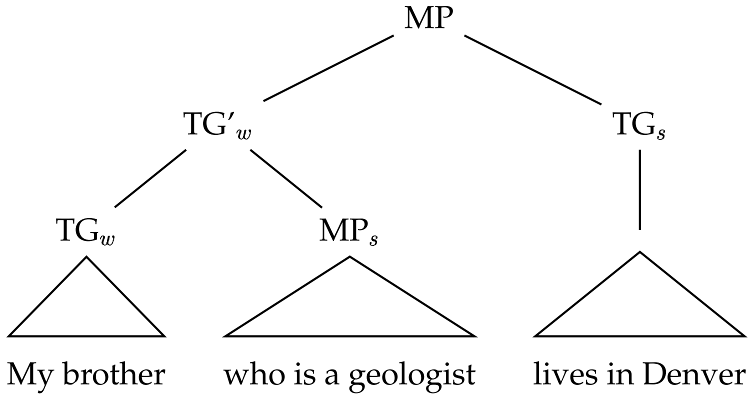
Would you like some more tea, Ian? (8)

The tag 'Ian' is obviously preceded by an intonational boundary, but does not exhibit the properties of a full IP. Pitch movement on the tag is usually a continuation of the pitch movement in the tail of the preceding phrase. Suppose the tag is treated as a MP, (it is after all set off by audible boundaries), leading to a recursive structure:



(9)

Here the 's' and 'w' again represent metrical weight, which we need not be over concerned with here, and the ′ superscript marks the TG as being a parent of a node with a similar status. This kind of structure would also allow the analysis:



(10)

However, this idea is problematic. It suggests that "who is a geologist" has the necessary intonational tune to stand on its own as an utterance, which is probably not the case, but would provide grounds for an interesting experiment.

2.9 The Effect of Prosodic Structure on Intonation

Some of the issues concerning prosodic structure and the effect it has on intonation are also discussed by Ladd (1990). Here, Ladd considers downstep as a metrical relationship between intonational constituents. This takes an alternate view to Pierrehumbert, and Ladd's (1983) earlier work where downstep was considered an independent feature, rather than a property of particular accents. This view attempts to answer Beckman & Pierrehumbert's (1986) criticisms that having down-step triggered by a feature forces a relation between the down-stepped accent and the preceding one, as this not only goes against the grain of Pierrehumbert's (1980) theory in the pitch events are supposed to be independent entities but also allows for nonsensical accent sequences. Downstepping the first accent in a sequence is meaningless for example, as there is nothing to down step from.

Ladd discusses a *register phenomenon* which is a factor placed orthogonal to other factors which contribute to the variation of pitch. Ladd asserts that a speaker has a predefined pitch range, which is idealised as constant and that register is defined as a frequently-changing sub-set of this range, and specific target types have fixed positions within the register. Nuclear accents are set to the top of the register, and the only scaling that is allowed is down scaling, i.e. the lowering of the pitch of pre-nuclear accents within the register.

This provides a two-way distinction as shown in Figure 2.5. Note here how the choice of tone is independent from the underlying metrical structure.

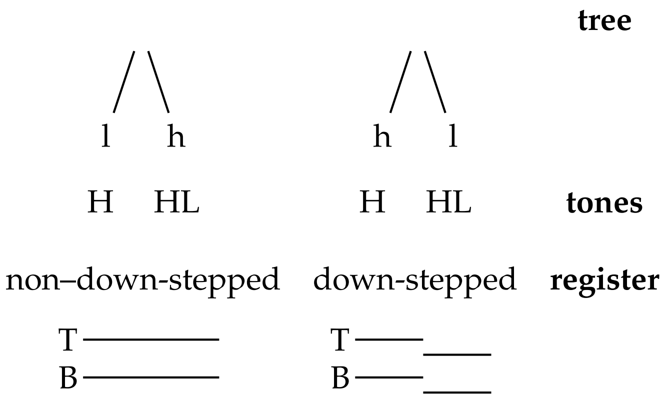


Figure 2.5: Ladd's two-way down-step distinction. The 'l' and 'h' represent the phonological high low relationship in metrical terms. The 'h l' ordering triggering downstep. The 'H' and 'L's represent the intonational tones, and the 'T' and 'B' represent the top and bottom of the pitch range.

Ladd shows the need for a metrical approach based on the results of an experiment (Ladd 1988) where sentences with different *and/but* constructions were produced by a variety of speakers. For example:

> Allen is a stronger campaigner, **and** Ryan has more popular policies, **but** Warren has a lot more money. (11)

Allen is a stronger campaigner, **but** Ryan has more popular policies, **and** Warren has a lot more money. (12)

Here the structure of the sentences is the same, but there are clear differences in pitch range between the two types, suggesting the need for a metrical distinction to cause the down scaling of pitch range outlined above.

Sentences like these are generally interpreted as the *but* opposing the conjoined propositions of the *and*, see Figure 2.6.

[A and B] but [C] or [A] but [B and C]

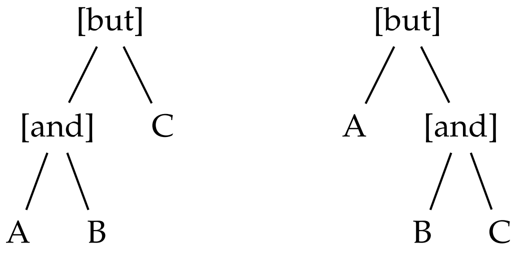


Figure 2.6: And/but distinction using trees

The *but* attachment can be thought of as being higher in the tree than the *and* attachment. Experimental evidence showed that there is a significant difference in the heights of the initial high tones of the sub-phrase dependent upon whether it is preceded by an *and* or a *but*. Ladd suggests that this shows the existence of an underlying hierarchical prosodic structure, but leaves unresolved the exact nature of a mapping from a tree structure to a set of tone heights.

While the idea of hierarchical structure controlling pitch range at the phrase boundary level seems appropriate, it is hard to judge whether it is a suitable explanation for the control of downstep within a phrase from the experiments described here. This point aside, from a practical point of view for speech synthesis, the need for large consistently labelled trees would currently make this theory unmanageable.

2.10 Intonation and Meaning

The general accord on pitch accents is that they render salient the material with which they are associated, irrespective of type. Pierrehumbert & Hirschberg also suggest that the lack of a pitch accent, where one would normally be expected, reduces the salience normally associated with a particular item, unless of course the item is being made salient by some other means.

In terms of Pierrehumbert's (1980) theory, Pierrehumbert & Hirschberg (1990) suggest that items marked salient with an H* are regarded as *new* to the discourse, and an intonational phrase of only H* accents signals that the proposition realised by the phrase should be added to the mutual belief space of the listener. When accompanied by an L phrase accent (and either boundary tone) the result is *neutral declarative intonation*. For example:

H* L L%
It is raining. (13)

If, on the other hand, the phrase accent is an H and the boundary tone an H%, then an element of questioning is introduced. Information is still being proposed by the use of H*s, but agreement is being sought after. For example:

H* H H%
You got my letter? (14)

The L* accent on the other hand is said to mark salience that is not intended to be added to the listener's mutual belief space. The accents comprising of complex tones are said to evoke salience of some *scale* and express different relationships between the accented item and others in the discourse.

	AGREED	
	+	−
θ	L+H*	L*+H
ρ	H*, (H*+L)	L*, (H+L*)

Table 2.1: Steedman's meanings of pitch accents.

Commitment	
$[S]$	L, LL%, HL%
$[H]$	H, HH%, LH%

Table 2.2: Steedman's meanings of Boundaries

Phrase and boundary H and L tones are used to signify continuation or separation respectively between phrases. The phrase tones express such a relation between intermediate phrases and boundary tones between intonational phrases.

An alternative interpretation of the meaning and intonation is proposed by Steedman (2002) reflecting ideas from Prevost & Steedman (1994). It differs from Pierrehumbert & Hirschberg's (1990) approach in that it looks at the meaning of pitch events within an *information-structural* framework. Steedman suggests that pitch accents mark words as *not-given* or *kontrast* (after Vallduví & Vilkuna (1998)) rather than strictly *new* to the discourse, and that there are only two further binary-valued dimensions along which the meaning of pitch accents need be distinguished. The first is *theme* and *rheme* (Halliday 1967, Bolinger 1958, Bolinger 1961). The second dimension is whether or not the particular theme or rheme is mutually agreed upon. In this theory boundary tones divide into two classes, marking if the commitment to the content of the phrase or utterance lies with the speaker or the hearer.

Steedman summarises with the information shown in tables 2.1 and 2.2. In table 2.1, θ and ρ are theme and rheme respectively and the ±AGREED feature signifies mutual agreement. So for example, L+H* is used to mark mutually agreed upon themes and L* to mark rhemes which are not mutually agreed upon. In table 2.2 $[S]$ and $[H]$ denote speaker and hearer commitment respectively, so LH% may be used as a boundary where the speaker is implying that it is the hearer who is committed to the content of the utterance. For example:

H: Congratulations. You're a millionaire!

          L*   LH% (15)
S: I'm a MILLIONAIRE?

Here the speaker implies that the hearer is committed to a non-agreed rheme.

The obvious advantage from a speech technology viewpoint of Steedman's (2002) approach when compared to Pierrehumbert & Hirschberg's (1990) is that rather than just interpreting the meaning behind various combinations of pitch accents and boundaries, it provides an algorithmic means of applying pitch events to text to convey a given meaning. Steedman goes on to formalise this in terms of a *Combinatory Categorial Grammar* (CCG). Steedman's theory is used to provide the accent descriptions used in the examples discussed in section 3.4.2 and for the pitch contours used for the evaluation in section 9.3.

2.11 Statistical Modelling Techniques

Historically, in intonation modelling as with many other aspects of speech technology, rule based approaches (by rule we tend to mean hand crafted heuristics based upon observation, as opposed to the equations which govern statistical systems) have been first superseded by brute-force statistical approaches which have later been refined to use linguistic knowledge. This results in a continuum of approaches, with strictly rule based systems at one end and strictly statistical approaches at the other.

A statistical basis for a system can achieve reliability and consistency with a certain level of accuracy which is considered a safe compromise compared to rules which may perform particularly well in many situations, but fail miserably in a small number of unpredictable circumstances.

Building a statistical model usually involves training on a corpus of data, where the model 'learns' an association between an input which is generally a parameter vector representing an entity, and the output which is the entity itself. Once trained, the model can generate a suitable entity from a given input vector.

There are many techniques for learning these kinds of association, neural network (NN), classification and regression trees (CART) and linear regression (LR) models are three particular examples used in the speech synthesis domain. We take a closer look at CART and LR models as they are of particular interest. Neural Networks are not discussed in detail as they are a less popular approach for intonation generation. They have however been used to do waveform synthesis (Karaali, Corrigan & Gerson 1996), to model intonation (Sun 2001, Holm & Bailly 2002) and to detect accents for speech recognition (A.Taylor 1995). Recurrent neural networks (RNNs) prove popular for these tasks as they have a limited ability to model time dependencies.

Linear regression models

Linear regression (LR) models assume that a predicted variable (p) can be modelled as the sum of a set of weighted real-valued factors.

$$p = w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \tag{2.5}$$

The factors (f_i) represent parameterised properties of the data, and the weights (w_i) are trained, usually using a stepwise least squares linear regression technique.

Often subsets of mutually exclusive binary valued factors are used. This partitions the model into parts where only a subset of weights contribute towards the predicted variable in a given context. For example w_1 and w_2 may contribute only to the predicted f0 of accented syllables, while w_3 and w_4 may only contribute to unaccented syllables.

A simple example of a linear regression model is shown in figure 2.7. Here f0 is predicted by the presence of an accent or a boundary and the number of syllables from the start of the utterance. An f_1 component causes a peak on accents, and an f_2 component causes a dip at a boundary. If a syllable has no pitch event associated with it then both f_1 and f_2 are zero valued so their associated weights make no contribution to the predicted f0. f_3 is always positive valued and causes the f0 to decline through the utterance.

Classification and regression trees

A classification and regression tree (CART) (Breiman, Friedman, Olshen & Stone 1984, Breiman, Friedman, Olshen & Stone 1993) model is a binary decision tree. As a classifier the model assigns a candidate class, from a predetermined set of classes to a target based on the values of parameters describing that target. As a regression tree the model estimates a parameter, where the possible values of that parameter fall into classes with different means and standard deviations. In

$$pitch_{f0} = 50 + 15f_1 + -5f_2 + -0.5f_3$$

f_1 Syllable is accented $\{0, 1\}$
f_2 Syllable is a boundary $\{0, 1\}$
f_3 Number of syllables from start of Utterance. $[1, n]$

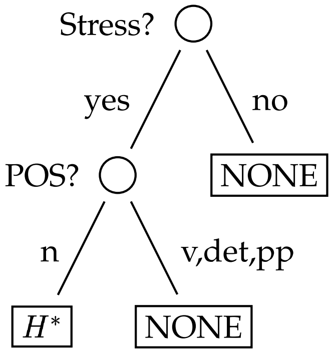
Syl	Event	$pitch_{f0}$
2	none	49
6	accent	62
8	none	46
12	boundary	39

Figure 2.7: An example simple linear regression model. The linear model on the left is used to calculate pitch targets for a selection of syllables from a hypothetical utterance of twelve syllables, where there is an accent on the sixth syllable and a boundary after the last syllable.

speech synthesis CARTs are widely used to model segment durations (for example Riley (1992)), but as we shall see they can also be used for accent prediction and pitch contour generation.

The leaves of the decision tree specify the class a target case falls into, and each node of the tree contains a binary decision based on one of the recorded parameters for the case. A very simple CART intonation prediction model is shown in figure 2.8. The example shows the sentence "The cat sat on the mat." Each syllable is parameterised by the features: POS (part of speech), Stress and Punc (following punctuation). The tree classifies accent class for each syllable. Parameter values for the words 'sat' and 'mat' are shown. 'Mat' receives an accent but 'sat' does not. An interesting point to note is that the tree does not necessarily use all of the parameters in its decision making process, and it may use different parameters for different circumstances depending on which branches are taken. In this example the Punc feature is not used at all and the POS feature is only used on the left branch of the tree.

The example tree here was constructed by hand, but trees in general are constructed by a data driven training process. For classification this works by repeatedly partitioning the data into a subsets of classes, where each split adds a node in the tree until all classes are accounted for individually at the leaf nodes. In reality the leaves of the tree tend to list probabilities of a case being in each class, and the class with the highest probability is chosen to classify an unknown case. In the above example the restricted set of classes representing accent type is 'H*' and 'NONE'.



Example: "The cat sat on the mat."

Word:	sat	mat
POS:	v	n
Stress:	yes	yes
Punc:	none	.
Accent Class:	NONE	*H**

Figure 2.8: Simple accent predicting CART.

For classification, the partitioning is performed to find the division which best partitions the data so that cases of the same class are placed in the same subset. For regression the partitioning is done to minimise the error in the predicted variable. The value of the predicted variable is taken to be the mean value of the class a case is classified as. If figure 2.8 were being used for regression, each class would be accompanied by a mean and standard deviation which are calculated from the individual members of that class in training.

One of the main advantages of CART models is that they are non-linear and can handle non-linear data potentially better than linear models like the LR model above. This doesn't however mean that they are problem free. They do not guarantee the best possible solution, just a local maximum, and a classifier may well classify a proportion of its training data incorrectly. Their efficiency however, usually makes up for these drawbacks.

2.12 Intonation in Speech Synthesis

We will now briefly look at a few speech synthesis systems and the approaches they take. However, as speech synthesis has become more commercial fewer details about exactly how intonation is generated are published, and when methods are published the details concerning actual implementations are often lacking.

As most speech synthesis systems are modular in design, they often allow for more than one approach to intonation generation to be used. This can be particularly useful for multilingual systems where intonation research has gone in different directions for the different languages.

2.12.1 Festival intonation generation

Festival is the speech synthesiser which we use in the work carried out here, so we start by examining its approach to intonation.

The Festival speech system (Taylor et al. 1998) allows for a variety of intonation modelling techniques to be used. Its default approach however is a statistical one where intonation generation happens in a number of steps. Each step in the intonation generation process builds upon the information generated previously, adding new relations or supplementing the information in the items of existing relations. The steps taken can be summarised as:

1. Assign phrase breaks.
2. Assign symbolic pitch events for syllables (ToBI accents and boundaries).
3. Assign durations to segments.
4. Assign contour target points to syllables.
5. Produce f0 contour from target points.
6. Produce speech using durations and f0 contour.

Steps 2 and 4 are of particular interest as these are the steps that are changed in the work described in this thesis.

Pitch event assignment

The default accent assignment module provided by Festival, works in the following way.

Pitch events are assigned using two CART classifiers trained on the f2b dataset described later in section 3.5. A portion of this data is saved for testing and not used in the training phase. One model predicts the presence of accents, the other predicts the presence of boundaries. The accent CART classifies syllables into the classes H*, !H*, L+H*, L*, L*+H and NONE based on punctuation, minimal part of speech information and the position of a syllable within the utterance. The outcome is that most nouns get accents assigned to them. Prediction of H* accents is reasonably good, around 70% correct on the data reserved for testing. Prediction of accents which occur less frequently in the training data, L* and L*+H in particular, is much worse. For example, none of the 176 L* accents in the saved test data were predicted correctly.

A similar CART model predicts boundaries from the set H-, L-, L-L%. L-H%, H-L% and NONE. This performs very well in predicting L-L% (90% correct), but much worse in predicting other boundaries. These results are due to the nature of the dataset used for training (see section 3.5).

The CART models are used to generate ToBI accent and boundary labels which are assigned to syllables. These CART models predict reasonable labels relating to some "general idea of a default intonation pattern", which usually means a sequence of H* accents followed by an L-L%. If the required intonation pattern for a particular utterance is different from this—in particular, if it is one containing pitch events which the models do not predict very well—the output from Festival is going to be wrong. This result can be particularly disappointing for concept to speech systems trying to use Festival, where the required accent sequence is known to the language system, but Festival predicts pitch events based only on the text. This is one of the issues we will address with the models we build in later chapters.

Contour generation

Contour generation in Festival (Black & Hunt 1996) is carried out using three LR models. Each model predicts the f0 at a different point of a syllable (start, middle and end). The factors incorporate information like the type of accent present, position of phrase breaks, syllable stress and syllable position. They consider this information for a five syllable window centred on the current syllable. This allows f0 on syllables around an accent to be affected by the presence of the accent, which means f0 movement is not restricted to occur on the syllable that is marked with the pitch event. For example the peak of an L+H* could occur in the syllable following the one the accent is assigned to.

The LR models are used to produce a series of pitch target points. Linear interpolation is used to fill in f0 values between the specified targets.

These models generate reasonable contours when the input comes from the above CART prediction models, but work less well when pitch events are otherwise specified. This is again thought to be related to the dataset which both models (for the prediction and contour generation stages) are trained upon. It is our intention to produce an alternative model which produces good intonation irrespective of the input source pitch events.

What is of interest here is the fact that these models rely predominantly on lexical information, particularly part of speech, and a predicted simple phrase structure. The result of this is that intonation produced is relatively neutral, and can produce neutral declarative pitch contours and simple question contours, but does not extend to being adaptable to express more than one meaning for a particular phrase.

2.12.2 The Bell Labs text to speech system

The most prominent recent detailed description of how intonation generation is performed outside of Festival is provided by Santen et al. (1998), where they discuss two approaches to intonation production. The first is referred to as ‘the tone sequence approach'. Here a number of abstract H and L targets are assigned to an utterance based on analysis of the text. The actual pitch values represented by these target points are determined in relation to parameters which specify an appropriate pitch range, typically represented as step functions. The second approach is a superpositional approach. This model attempts to address the variation in accents due to the segmental structure on which they occur. The position of accent peaks are predicted by a linear equation relating to the syllabic structure, this peak is then imposed upon an estimated phrase curve.

The tone sequence approach

The tone sequence approach relies on a series of predicted ToBI accent and boundary labels just like Festival does, although they include the additional use of initial boundary tones H% and L%.

A number of pitch target points are then assigned for each label. These target points are temporally aligned with respect to the syllable with which they are associated by values stored in a lookup table based upon five factors derived from experimental data.

Pitch range is modelled by three lines: a top-line, a reference line and a base line. Pitch target points are then scaled in relation to the lines controlling pitch range. Each pitch event is assigned a prominence factor. Boundary targets are positioned within the pitch range using the prominence factor. For example an initial H% with prominence factor of 0.5 will be positioned halfway between the reference line and the top line.

H* accents are modelled as a step up in pitch using three target points. A low and a high point are positioned close together to produce a sharp rise, followed by a second high point to produce a plateau. The position of the highs is governed by the following equations:

$$f0_h = Ref + p.(Top - Ref). \tag{2.6}$$

$$f0_l = f0_h - 0.6(f0_h - Ref). \tag{2.7}$$

where $f0_h$ and $f0_l$ are the high and low target positions, p is the prominence factor and Ref and Top are the reference and top-line frequencies respectively.

Phrase accents consist of a single target point positioned in a similar way to which the target points of accents are, and final boundary tones are placed the same way as phrase initial boundaries.

Downstep is modelled by lowering the top-line using the following equation:

$$Top_i = Ref + d.(Top_{i-1} - Ref). \tag{2.8}$$

Where Top_i and Top_{i-1} are the current and previous top-line frequencies and d is a downstep factor which varies from zero to one.

The superpositional approach

This approach uses a linear regression model to predict accent alignment within a superpositional framework, and is proposed as an improvement to the above tone sequence model particularly for English. The notion of *s-rhyme* (short for sonorant rhyme, previously mentioned in section 2.6) is introduced. The s-rhyme is defined as consisting of the first non-initial sonorant of the accented syllable through to the last sonorant, for monosyllabic groupings, or the last segment of the accented syllable for polysyllabic groupings. Material preceding the s-rhyme is considered to be the *onset* and material following the *remainder*.

Alignment *anchor points* for an accent's peak position and pre and post peak heights (measured as percentiles of total peak height) are modelled using a linear regression model. The model is trained on contours which have had an estimated phrase component removed, as this is modelled as a separate component.

Each alignment point (T_i) is then modelled with the following linear regression model:

$$T_i = \alpha_{o,c} D_{onset} + \beta_{o,c} D_{s-rhyme} + \gamma_{o,c} D_{remainder} + \mu_{o,c}. \tag{2.9}$$

where α, β and γ are trained coefficients parameterised over onset and coda type (c and o). μ is a intercept again parameterised on onset and coda type.

It should be noted that the linear regression model is being used here differently from the way it is used in the Festival. Festival uses linear regression to model pitch targets at particular points in every syllable, whereas here it was being used to model the alignments in time of specific pitch heights relating to pitch events.

This model is used within a superpositional framework where multiple component contours are used to compose a final pitch contour. Three components *minor phrase curves*, *accent curves* and *segmental perturbation curves* are discussed. The anchor points modelled above are used in computing the accent curve. The perturbation curves are used to control effects related to the transition from an obstruent to a sonorant and the minor phrase curves are simple declining phrase components.

Both of these approaches are to some extent rule-based approaches, but the rules are complex and derived from statistical analysis. The way in which statistical results are combined with linguistic knowledge is similar to the way in which the work described in this thesis is carried out by using linguistic knowledge to improve statistical models.

Black & Hunt (1996) demonstrate that Festival's linear regression approach is clearly better than the Bell Labs' tone sequence approach both in resulting output and ease of implementation. It is more difficult to judge how the Bell Labs' superpositional approach compares to other models as there is not a sufficiently detailed description available to re-implement it. Festival's linear regression approach has the advantage of modelling the pitch on every syllable rather than just modelling pitch movements related to pitch events—this means that pitch effects relating to lexical stress are inherently modelled. Additionally, not needing to make the assumption that each of the components of the superpositional model are sufficiently independent to be suitably extracted from the data is an advantage. And finally, as we shall see in chapter 9, using this approach will enable us to expand the model to produce effects not found in the data it was originally trained upon.

2.13 Summary

As we have seen there are numerous theoretical approaches to studying intonation, each focusing particularly on a specific aspect which the researcher finds interesting. Unfortunately generating intonation for speech synthesis is generally "uninteresting" as it is not the finer details that interest the theoretical researcher we are concerned with. A speech synthesis system needs to be able to produce a reasonable intonation pattern for any given input, interesting or otherwise.

Some intonation theories are highly descriptive in a phonological sense, for example, the ToBI system and the work of Pierrehumbert. This system can precisely describe pitch events at a phonological level, but tells us very little beyond an abstract description of what a particular piece of f0 contour would look like for a given accent configuration. Pierrehumbert (1981) addresses this problem to an extent by proposing a rule based model which generates f0 contours using a series of pitch targets and transition rules. However, this model is very simplistic and only generates neutral declarative intonation patterns.

Other models like tilt, give us a reasonable description of what the f0 contour looks like, but very little information about the phonological status of a pitch event. There is a clear divide here between the description of accents at the phonological level and at the acoustic realisation level. We work towards closing this gap and in chapter 9 we demonstrate a model for intonation generation which can generate particular shapes for specific phonological descriptions.

Other recent intonation research in the speech technology field which will not be discussed here includes MOMEL Hirst, Di Christo & Espesser (2000), an algorithm for the automatic modelling of f0 curves. This model uses what the authors call *modal regression*. This model shows potential as it does not require labels to specify the position of pitch events and is intended to model a wider range of curve shapes than the tilt model does. Prosodic models based around production and perception (Dogil & Möbius 2001*b*, Dogil & Möbius 2001*a*) and a phonetically motivated model of prosody (Möbius & van Santen 2000) have also been proposed.

CHAPTER 3

Tools and Resources

3.1 Further Defining our Goals

In chapter 1 we expressed one of our goals as being to understand how linguistic theory can be appropriately exploited to improve the generation of synthetic intonation. We now become more specific and express that goal as an attempt to answer the following questions:

1. What type of linguistic information, specifically concerning phrasing and accenting, is useful for automatically generating intonation for speech synthesis?
2. Can using the correct information give us intonation which is not only acceptably natural, but can it be of a recognised style?
3. How does this type of linguistic information relate to the type of information a language system is likely to provide, and can we use the information provided by a language system in a useful way?

We will use the answers to provide a better linguistic *framework* for intonation models for speech synthesis along with example models implemented in this framework. By framework we mean a set of constraints which restrict the variability of intonation by providing a *context* for individual components of the model. These ideas are described in detail in Chapter 7.

The approach we take to develop a better model is to try to use the the right linguistic knowledge in the right way to constrain a model. Instead of blindly throwing data at a modelling technique, we actively analyse a data set before starting to construct a model. We choose to use linguistic knowledge to build a framework within which a variety of models can be built. The idea of the framework is to provide a foundation on which either a full statistical model or a rule based model can be built. We also hope to gain a better understanding of how intonation works on the scale of a large corpus and learn how to cope with and use effectively the large amount of variation that occurs in a data set of this size.

3.2 Festival and its Methods.

Festival is used to produce the majority of the synthetic speech for the work carried out here. Festival was chosen in preference to other possible 'better' systems for a number of reasons. The better systems are generally commercial unit selection systems, which do not currently allow their prosody modules to be modified in a ways suitable for carrying out this kind of research. Prosodic improvements in such systems are achieved by choosing appropriate units which require little or no alteration to the underlying prosodic content. As unit selection system architectures become more open, it will be possible to carry out more research into how to choose these units appropriately with respect to prosody. The results of the work carried out here are applicable for use with unit selection and could be used either as part of the target cost or as part of a post selection prosodic modification mechanism.

Our main aim is to improve intonation by the use of richer linguistic information and in doing so to discover what linguistic input is needed to generate better intonation. This is important not only to improve intonation models for diphone synthesis, but also for applications where voice transformation is used, where the characteristics of one voice are being mapped onto another.

The commercial nature of unit selection synthesisers also dictates that they are generally closed systems which allow little or no manipulation of the processes they use to do speech synthesis. Festival on the other hand is a completely open and modular system which allows us to manipulate the synthesis process as we wish, and hence, provides an ideal platform for this research. In some ways the resulting speech synthesis output may not sound as good as that produced from a commercial unit selection system, but we will have more control over all stages of the synthesis process which is particularly useful when wanting to generate multiple intonation patterns, say as specified by a language system, rather than being restricted to a single pattern predicted by the system. This open access also allows us to see what is happening at intermediate stages in the synthesis and make it easier to understand and solve problems.

We have already seen how Festival usually generates intonation in section 2.12.1. This section introduces the data structures that Festival uses, concentrating on those which are related to intonation generation.

3.2.1 Festival internals

It is useful to have a basic understanding of the internal structures Festival uses to represent data as these often direct the formulation of models and the way data can be manipulated.

Festival stores data as Heterogeneous Relation Graphs (Taylor, Caley & Black 2001). The way in which Festival uses these is described from the bottom up. At the most basic level are structures called *items*. An item is a single data entity such as a pitch accent or a segment. Items consist of *features*, which are key-value parameters describing the entity. A `name` feature usually exists to identify the item and an `end` feature will exist if the item is time-aligned. Items are used to represent all the data components required for synthesis: tokens, words, syllables, segments, pitch events, diphones, etc.

Items then make up *relation* objects. A relation is a structured set of items. Relations are generally lists or trees with an item at each node. The SEGMENT relation, for example, is a list of the items describing the segments being used for synthesis. The SYLLABLESTRUCTURE relation is a list of trees: the top nodes are items from the WORD relation, the daughters of these are items from the SYLLABLE relation and their daughters are the items in the SEGMENT relation. This illustrates how items may be part of more than one relation.

Intonation event prediction involves creating a list relation called INTEVENT which contains symbolic accents and boundaries. An INTONATION tree relation associates the events with syllables. Contour generation involves creating an F0 relation which contains pitch target points. These structures are normally created by the models discussed in section 2.12.1. The models which we develop in later chapters to replace the current intonation models will take up the job of creating these relations.

3.3 Text to Speech vs Language System Input

The type of intonation model which is used by a particular speech synthesis system may be governed by how the system is intended to be used.

To generate useful intonation we need to be aware of what we are generating intonation for and what we are generating it from. The context a system is being used in determines what resources are available to the system and what is expected from the system.

As we saw in Chapter 1, the input to a speech synthesiser has in the past generally been plain text, just a list of words. Here the synthesiser has no point of reference or context to tell it what this text is for or where it is from. There will be different intonation contours which could be applied to the text to give different meanings. Picking one without knowing the context of the utterance may well be acceptable in some circumstances, but it others it may be catastrophic.

For example, intonation can be associated with the utterance: "Please proceed via the blue corridors and stairs to the emergency exit," to convey subtly different meaning and avoid ambiguity as to whether any stairs can be used or whether only blue stairs should be used.

There is increasing demand for speech output from *language systems*. Language systems are systems that generate textual descriptions to satisfy requests which are effectually database queries, although they tend to include background information so the answer to the current query can be influenced by the information that has previously been presented.

This type of system, and concept to speech systems in general, often need to convey a particular meaning for a given utterance, for example, to contrast the information currently being presented with that contained within the last response. This desired effect is precisely one where the general type of intonation produced by the statistical approaches used by TTS systems is unsatisfactory.

As the output text from a language system has been generated from a particular semantic concept, where the intended meaning is known, the language system is in a very good position to provide a lot more than just text output. The chances are that in generating the output text, all of the language processing information that the synthesiser needs, for instance, part of speech and focus information, were used in constructing the output. The language system may be able to provide phrasing and phrasal stress, syntax, semantics, accent position and possibly even duration information to the synthesiser.

The potential of this information is enormous. All of the information the synthesiser needs to predict can be provided by the language system with one hundred percent consistency with that used to generate the text.

However, as most speech synthesis systems are designed to be text-to-speech systems they may have difficulty using more complex input. If a speech synthesis system is not designed to use this information it may ignore it. It may even discard it and regenerate parts of it itself and come up with something which is entirely different from which the language system specified, something which is less appropriate and contains errors. Most speech synthesis systems are designed to be text-to-speech systems and have difficulty with more complex input.

3.4 Text Input From Language and Dialogue Systems

We turn our attention to some example marked-up input that we could be required to generate speech from. We look specifically at the output of language systems which have influenced the work carried out here by their need for intonation that is better than Festival is able to provide using the models it uses for text only input. We consider what each system is trying to achieve in terms of intonation and compare what can be made available to the synthesiser by each system to attain the desired result.

3.4.1 *Multilingual personalised information objects (M-PIRO)*

M-PIRO (M-PIRO 2000) is a system designed to deliver highly personalised descriptions of museum exhibits. The system generates texts that describe museum objects taking into consideration the user's interests, preferences and previous exchanges with the system.

The system is designed to work in a number of virtual environments. The primary environment is a web based virtual museum. Speech synthesis output is available here, but not of primary importance as the text can be read from the screen. The environment where speech synthesis plays a major role is the CAVE

(Cruz-Neira, Sandin, DeFanti, Kenyon & Hart 1992) environment. This is a virtual reality (VR) environment where a user sees projected three dimensional images of museum objects, together with speech commentary. Bad speech output here results in a confused user.

The language generation component of M-PIRO generates texts by natural language generation from internal descriptions of objects and the semantic relations between different objects and by incorporating canned text where phrases would be difficult to generate from scratch.

M-PIRO can output either text or syntactic structure expressed in SOLEML (based upon Hitzeman, Black, Taylor, Mellish & Oberlander (1999)), an XML mark-up language. The additional information provided by the mark-up is of interest here.

Table 3.1 shows an extract of SOLEML illustrating the kind of information which is provided to the synthesiser. The text actual text of the example is highlighted in bold to make it easy to pick out. Along with basic syntactic structure and accurate part of speech information, noun phrases (NPs) contain additional information marking them as new or important. In the example the NP at line 2 has the newness feature value `old` marking the text 'this complex' as not new to the discourse. The importance feature is used less frequently, and does not appear at all in the example. The actual use of these features in M-PIRO is still under review, but the intention is to make them useful for aiding the generation of intonation.

This mark-up falls short of directly specifying intonation. Suitable intonational phrasing does not necessarily correspond directly to any of the units in the syntactic structure. The work carried out here in adapting Festival to use this structure is discussed in section 10.2.

3.4.2 *Embodied believable agents (MagiCster)*

The MagiCster project (MagiCster 2002) is concerned with the development of believable conversational interface agents. This involves information delivery in a dialogue context using animated characters and synchronised speech synthesis. Although the project is still in its infancy and the language generation component of the dialogue system is still at the development stage, an agreed mark-

```
    <relation name="Syntax" structure-type="tree">
      <elem phrase-type="S">
        <elem phrase-type="NP" newness="old">
          <elem lex-cat="DT" href="words.xml#id(w1)">this</elem>
          <elem lex-cat="N" href="words.xml#id(w2)">complex</elem>
        </elem>
        <elem lex-cat="V" href="words.xml#id(w3)">was</elem>
        <elem lex-cat="V" href="words.xml#id(w4)">created</elem>
        <elem phrase-type="PP">
          <elem lex-cat="IN" href="words.xml#id(w5)">during</elem>
          <elem phrase-type="NP" newness="old">
            <elem lex-cat="N" href="words.xml#id(w6)..#id(w8)">
              the hellenistic period</elem>
          </elem>
        </elem>
        <elem lex-cat="CC" href="words.xml#id(w9)">and</elem>
      </elem>
      <elem phrase-type="S">
        <elem phrase-type="NP" newness="old">
          <elem lex-cat="PRP" href="words.xml#id(w10)">it</elem>
        </elem>
        <elem lex-cat="V" href="words.xml#id(w11)">dates</elem>
        <elem phrase-type="PP">
          <elem lex-cat="IN" href="words.xml#id(w12)">from</elem>
          <elem phrase-type="NP" newness="new">
            <elem lex-cat="N" href="words.xml#id(w13)..#id(w19)">
              between circa 230 and 220 B C</elem>
          </elem>
        </elem>
      </elem>
.
.
.
  </relation>
```

Table 3.1: An extract from a SOLEML example utterance.

up format and example dialogues in different domains are available to perform speech synthesis on.

The mark-up language that has been developed by this project is APML. APML was designed to directly incorporate elements which describe intonation based on Prevost & Steedman (1994) and Steedman (2002) discussed in section 2.10. The result is input which not only directly supplies appropriate intonational phrasing for the text but also provides consistently assigned pitch accent and boundary labels which relate directly to the meaning that is meant to be conveyed by the utterance.

An extract of APML is shown in table 3.2. All words which are to be be accented are embedded within emphasis elements which specify the appropriate accent type. Separate boundary elements specify boundary tones and intonational phrasing. The MagiCster system differs from the M-PIRO system in that MagiCster's design incorporates the requirements of speech output directly within the system, where in some respects speech is an afterthought in the design of the M-PIRO system. For this reason we concentrate on using MagiCster examples when evaluating the generation of pitch contours from ToBI accents (see section 9.3) as the APML mark-up gives us a appropriate accent specification with a known meaning to express. However, both the M-PIRO and MagiCster systems use the contour generation models developed by the work described in this thesis.

3.5 F2b – and The Boston Radio News Corpus

We now shift attention to the dataset we build models from and test against. The main body of data we have analysed is the *f2b* section of the Boston Radio News Corpus (Ostendorf, Price & Shattuck-Hufnagel 1995). This data set has been chosen because it is a reasonably sized data set of adequately complex sentences of read speech of a specific style from an individual speaker, and has been widely studied by the linguistics community. This data set is also currently used for the training of the default English intonation models for the Festival speech synthesis system, which can provide a good comparison for any final model.

We choose this data set over more variable speech styles such as found in the switchboard corpus (Godfrey, Holliman & McDaniel 1992) and the various map task corpora (Anderson, Bader, Bard, Boyle, Doherty, Garrod, Isard, Kowtko,

```
<turnallocation type="take">
  <performative type="greet">
    <rheme>
      Good
      <emphasis x-pitchaccent="Hstar"> morning </emphasis>
       Mr Smith <boundary type="LL"/>
    </rheme>
  </performative>
</turnallocation>

<performative type="inform">
  <theme belief-relation="gen-spec" affect="sorry-for">
       I’m sorry to
       <emphasis x-pitchaccent="LplusHstar"> tell </emphasis>
       you <boundary type="LH"/>
  </theme>
  <rheme>
    that you have been
    <emphasis x-pitchaccent="Hstar"> diagnosed </emphasis>
     as
     <emphasis x-pitchaccent="Hstar">suffering</emphasis>
     from a
     <emphasis x-pitchaccent="Hstar" adjecti-
val="small">mild</emphasis>
     <emphasis x-pitchaccent="Hstar">form</emphasis>
     of what we call
     <emphasis x-pitchaccent="Hstar">angina</emphasis>
     <emphasis x-pitchaccent="Hstar">pectoris</emphasis>.
     <boundary type="LL"/>
  </rheme>
</performative>
```

Table 3.2: An extract from an APML example utterance.

McAllister, Miller, Sotillo, Thompson & Weinert 1991) where we feel there is too much variation in speech style and insufficient data from individual speakers. We require at least an hour of clear speech from an individual speaker, annotated both with segment and intonation labels. The above corpora are both large in overall size but do not contain a suitable amount of data from individual speakers. We also choose f2b over less variable styles such as TIMIT (Garofolo 1988) where the short utterance style lacks the more complex prosodic structure which we wish to study.

The broadcast news style of speech provides clear and consistent intonation. It is also an appropriate style for a speech synthesiser, particularly when forming part of an information providing system, which is currently one of the main uses for speech synthesis.

The intonation labelling carried out on f2b is also useful to us, as there exists good hand coded ToBI labelling along with CSTR's hand coded and automatically coded accent/boundary labelling used for automatically deriving tilt parameters, both of which are available to us.

We intend to use the CSTR labelling predominantly for identifying pitch events. The more general labels of 'a' and 'b' are more useful to us here than a wide range of ToBI labels. The tilt parameters can subdivide the CSTR labels into smaller categories where appropriate. This classification is not as detailed as using the full ToBI inventory, but provides a simple and consistent set of pitch event types. A simpler model and the ability to generate intonation of an accent of unspecified type, for example where a language system just specifies a word to be emphasised, is deemed more important than being able to handle a full ToBI inventory. Furthermore, ToBI labels can always be mapped on to an appropriate more general classification. The problems with this mapping discussed in section 2.7 are not significant here since there are very few non H* accents from the categories that overlap.

Figure 3.1 shows an example utterance from the data set along with ToBI labels and CSTR a/b intonation labelling (IL). 'Fb' and 'rb' mark falling and rising boundaries respectively, whilst 'sil' marks silence, and 'c' marks connecting pieces of contours which contain no pitch events. The utterance consists of 5 phrases which are clearly marked at the ends with '%' boundary tones in the ToBI accent tier and fb/rb labels in the IL tier. In this example there is a one-to-one correspondence between the boundary labels in different labelling schemes.

There is also very good agreement between accent positions with the two labelling schemes, the main difference being that the penultimate accent, an 'L*', is not recognised in the IL tier. The form of IL labelling used here does not include a description of accents which are not of the general rise-fall type[1].

Figure 3.2 shows a longer more complex utterance, which spans two sentences, and shows two obvious levels of phrasing: that represented by '%' ToBI boundaries *within* the sentences and that represented by '%' ToBI boundaries *between* the sentences.

Towards the end of this utterance we see the use of the ToBI labels '!H-' and 'L-'. It is interesting to see how these correspond with the labels in the IL tier. The 'L-' (along with the preceding 'L+!H*') matches an 'afb' accent in the IL tier, as described in section 2.7 whereas the '!H-' has no corresponding boundary in the IL tier. This highlights some of the discrepancies that occur between label sets.

Another property of this particular style of intonation is hinted at by the ToBI 'HiF0' markers. When in phrases containing more than one accent, they are usually found associated with the first accent in the phrase, which is usually not the nuclear accent of the phrase. This is a property we shall be particularly interested in later on.

3.5.1 Problems with f2b

F2b is one of the best speech databases available for training intonation models, as it consists of a large amount of speech from a single speaker and is intonationally labelled. It does however have a major drawback in that the broadcast news style of speech does not provide much variation in accent specification. Table 3.3 shows the distribution of pitch events and boundaries. 83% of accents are H* and 58% of boundaries are either L- or L-L%. As we shall see this causes problems when trying to build models to generate other types of accents and boundaries. This also accounts for why Festival's TTS intonation models generally only predict H* accents and L boundaries.

[1] Labellers can signify such an accent but it will have been removed from the datafile shown as this example is intended for tilt model training purposes.

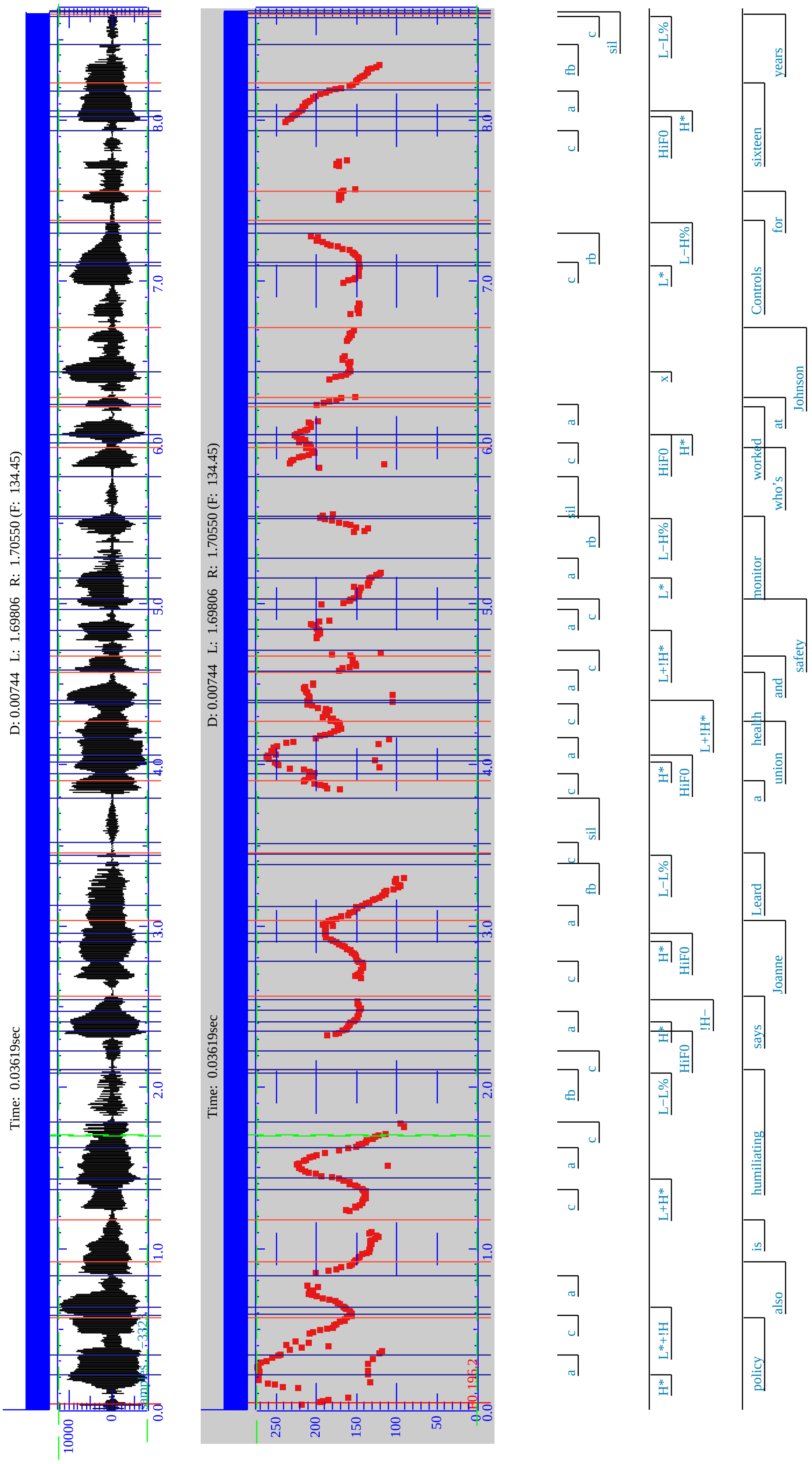


Figure 3.1: Example f2b utterance. First label tier shows CSTR's IL labels, second tier shows ToBI labels, third tier show words.

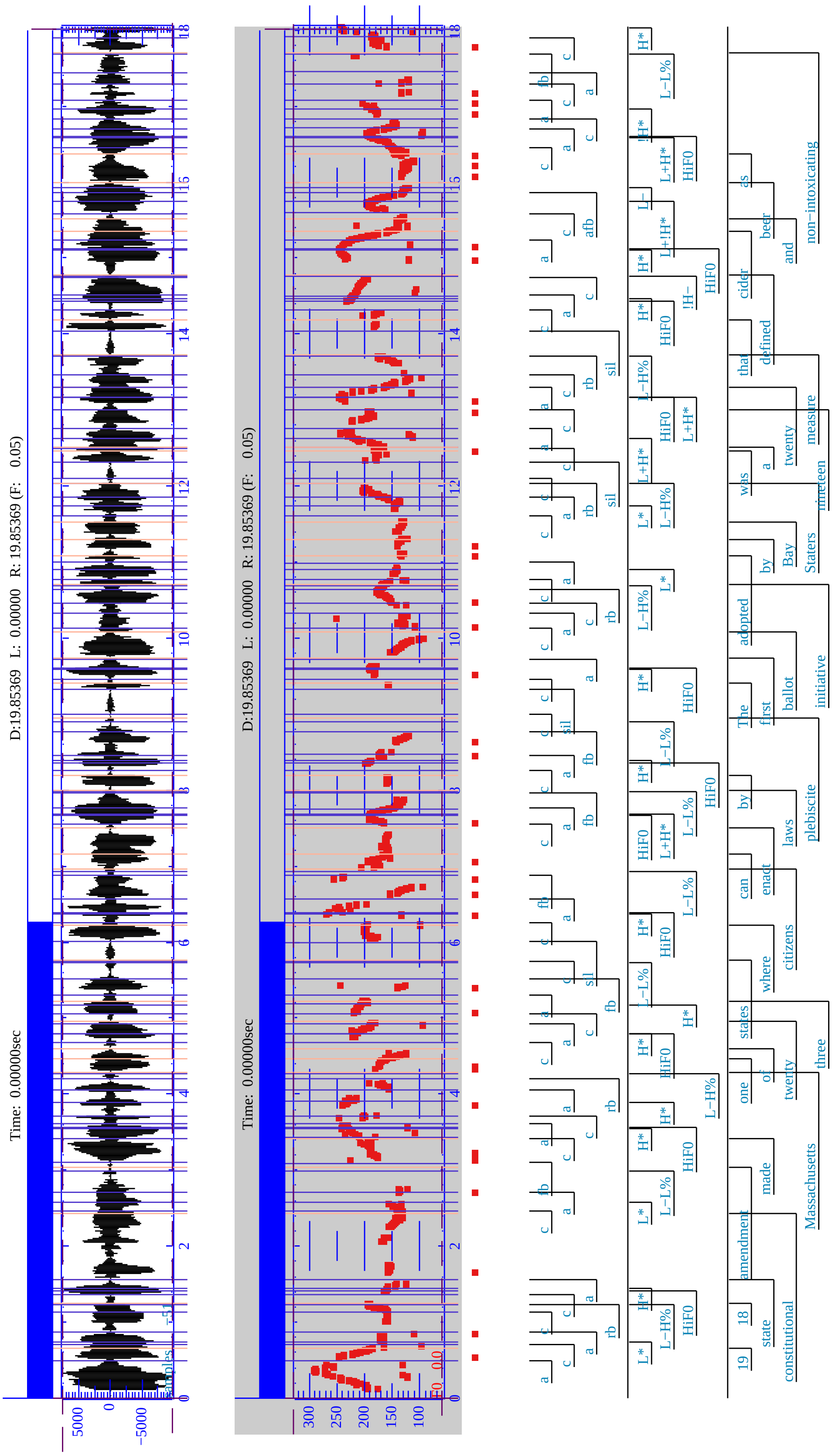


Figure 3.2: Example f2b utterance. First label tier shows CSTR's IL labels, second tier shows ToBI labels, third tier show words.

3.6 Processing F2b's Pitch Contours

For the work carried out here, the raw f0 files were processed in an attempt to provide an f0 from which more accurate measurements could be automatically taken than from the raw f0 file itself. The raw f0 is often suitable for manual analysis where the experimenter can correct for octave errors and other pitch tracking errors. The automatic tools are not able to discern such problems, so knowledge of what the f0 contour should look like is used in an attempt to minimise the errors found in it.

In an attempt to correct octave errors made by the pitch tracker any f0 values below 100Hz were considered halving errors and doubled accordingly. No maximum thresholding was done because visual inspection showed that this kind of error occurs less often, and correcting it would interfere with the some of the higher H* pitch accents which reached 300Hz and above. The corrected waveform was then median filtered by an order seven filter to remove outlier points. The order was chosen by visual inspection of the pre- and post-filtered pitch tracks, in an attempt to remove glitches, whilst not destroying some of the finer structure of the f0 contour.

3.7 Preparing F2b for Analysis and Model Building

We use tools provided by Festival to access the linguistic and acoustic data we are interested in in a meaningful way. Although Festival is primarily a speech syn-

Accent	Count
L*	190
H*	3846
L*+H	12
L+H*	553

Boundary	Count
H-	472
L-	339
H-H%	4
H-L%	40
L-H%	670
L-L%	1300

Table 3.3: Numbers of pitch accent and boundaries types in f2b

thesis engine, the way in which it stores information needed to produce speech is useful for analysis purposes. Tools developed to aid training of statistical models allow easy automated extraction of information related to individual components of a quite complex data structure.

We compile the f2b corpus of speech and subsequent additional labelling as a database in the format in which Festival stores and manipulates speech. Recall from section 3.2 that this database consists of utterances made up from a series of relations, items and features.

We wish to collate lists of features related to all the pitch events in the corpus which we can carry out statistical analysis on. So we need to extract features which provide information concerning where each pitch event is placed in the prosodic structure of its utterance along with features which describe the pitch event in terms of shape, alignment and positioning. The advantage of the Festival utterance structure over simple label files is that once built correctly the tree relations allow us to automatically access information that is indirectly related to a given entity. In the case of pitch events it allows us to access information regarding the syllable that they are associated with, and from there we can get information regarding the word or the individual segments.

We need to ensure that all of the corpus we are interested in is compiled into such a database. Fortunately this has been partially constructed from the f2b data whilst building previous intonation models for Festival, and already contains the following information:

- Text mark up: Phrase/Word/Syllable/Segment alignment.
- Prosody mark up: Phrasing/Tilt pitch event placing/Target f0 points

There are a number of issues concerning this structure with regard to the nature of this analysis. Most of the issues concern the fact that there is a lot of data we would like to use which is not available in the default utterance structure. This is primarily because we are using this tool out of its intended context.

Firstly, there is no representation of the actual f0 contour in the utterance. The utterances consist of primarily linguistic information. The f0 contour is a secondary property of the original f2b waveforms. To solve this, the f0 contour is extracted from the waveform and turned into a relation type object and incorporated into the utterance structure.

The second issue concerns voicing of segments. There are features which will tell us if a segment is phonologically voiced or not (i.e. whether the phoneme it is meant to represent is one of the vowels or voiced consonants). They do not however, tell us is that voicing has been phonetically realised. The value returned is based on the type of label of the segment, and does not necessarily mean that the segment is actually phonetically voiced. For example, the 't' segment in the word "stop" will be marked as unvoiced, as /t/ is an unvoiced segment. Phonetically however, the 't' may be voiced. To obtain an account of where voicing actually starts in the syllable we produce a relation which marks beginning and end of voiced sections within the utterance. The information within this relation is derived from the f0 track.

A third issue arises due to conflicting information regarding pitch events. Pitch event alignment information comes from two sources: the hand-labelled pitch event files and the files automatically generated by the tilt alignment program. The *start* and *end* times of a pitch event are regarded as the start and end times of the event itself as perceived by the labeller in the case of the hand labelled files. However, in the case of the tilt files the start and end points of a pitch event are the start and end points of a piece of pitch contour which the automatic aligner tool which accompanies the tilt tools thinks best represents the pitch event. These points do not necessarily coincide exactly.

It seems more appropriate for this analysis to use the start and end points as originally perceived by the original labeller, so this information is incorporated into the utterance.

The third point we need is the peak position of the pitch event. The only way we can acquire peak position information for a pitch event acurately is from the tilt information (as peak position is one of the parameters). The peak position lies within the start and end of the accent as defined by tilt, but this does not necessarily mean it lies within the accent start and end times as defined by the original labels as these may not coincide with the tilt start and end times as discussed above.

For example: consider a pitch event which is a falling boundary. If this is a straight fall in pitch then the peak position would be expected to be right at the start of the accent. If in the derivation of the best fit tilt parameters the start position was moved back a few milliseconds in time, then the peak position would be moved back accordingly. The peak position now lies before the start time of the hand labelled accent position, which is problematic. Interpretation of peak positions are therefore treated with care. Fortunately in the case of rise-fall accents where peak position is most meaningful, falling outside of the accent is less of a problem, as the peak position tends to be away from both the start and and end of the accent.

A fourth issue is that the phrasing derived for this study (see Chapter 4) is not the same as the phrasing already present in the utterance, so an additional relation with the new phrasing has to also be added. With the above information appropriately added to the utterance structure, all of the information we need for this analysis is now available.

Chapter 4

Phrasing: Analysis and Modeling

We begin our analysis of f2b by looking at the prosodic structure. Based on the literature we have discussed in chapter 2 we test two hypotheses.

4.1 Initial Hypotheses

Hypothesis 1: Our initial hypothesis is that the f2b utterances exhibit at least two levels of phrasing structure.

We test this hypothesis by investigating the effect of assuming that the data conforms to various different configurations of phrasing structure, looking for statistical effects relating to each of the structures in question.

Hypothesis 2: Our second hypothesis is that the speech of f2b can be *sufficiently* modeled by phrases containing a maximum of 3 distinct types of sub-phrases, namely an initial, a medial and a final sub-phrase type.

By sufficiently we mean where we gain more by having a simplified model than we lose by forcing the constraints that simplify the model. In simplifying the phrasing structure we will lose the ability to generate more complex phrasing patterns if they exist, but the simpler model will make the choice of assigning strucutre more robust as there will be only three categories to choose from. We will also be able to model those categories more accurately as we will have more data available for each category.

Our first step in analysing the data of f2b is to decide upon a phrasing structure which we will impose on the data to divide it up into sections which we will then analyse. As we have seen, the literature suggests that there are two or three levels which we should concern ourselves with, but how these levels actually manifest themselves is not so clear-cut and various structures are proposed usually accounting for two levels each. These two levels don't necessarily correspond to each other across different theories, suggesting that a possible third level may exist. The style of speech in question may also have an effect on the number of levels, with styles that employ shorter or simpler utterance structure not exhibiting the full structural range.

We choose our levels of prosodic structure based on our interpretation of the existing label sets, specifically the ToBI break indices, discussed in section 3.5. We attempt to show that such a classification is backed up by statistical analysis. This approach differs from what is usually considered the standard linguistic approach where particular examples of data are derived to exhibit a particular distinction to show a hypothesis is true. Here we apply statistical techniques to determine whether the data as a whole exhibits particular properties. We do this because rather than being interested in specific local effects in the data, we need this kind of judgment concerning the data set as a whole to produce a model for speech synthesis.

We take the ToBI break indices assigned to the data as our starting point. However, we do not use these directly because they pose a number of problems for us. Firstly there are five levels of break specified that are bigger than the break between words, where we are only interested in at most three: specifically, utterance level breaks and two levels that are related to the breaks in different levels of sub-utterance phrasing.

Here also there is not always the level of consistency within the labeling that we would like for this type of analysis. This type of data is often difficult to label, the distinction between levels '3' and '4' being a particularly hard distinction to judge at times. There are a few very notably long 'phrases' if we take a too literal interpretation of the break index data.

We actually partition the data in two separate ways: a simple method which takes the break indices as is and a second or complex way which attempts to simplify the structure of the data to help with our analysis and enable us to produce a well defined model of prosodic structure.

We will call our two levels of phrasing *IP* and *TG*. These terms are loosely based on those used by Ladd (Ladd 1996, ch. 6) but are not necessarily meant to relate to phrase units of the the exact same size and type as used there. An IP can be thought of as an *intonation phrase* but we do not wish to call it that explicitly because it may or may not be what others, particularly (Pierrehumbert & Hirschberg 1990) call an intonational phrase. Similarly TG can be thought of as a *Tone Group* which we consider to be a sequence of tones ending in some kind of boundary, and nothing more. An *utterance* (U) then consists of one or more IPs each of which in turn consists of one or more TGs.

Our basic approach then is to propose that the end of an IP requires a break index of at least 5, which is usually equivalent to a boundary tone accompanied by a pause. We then define the end of a TG as requiring a break index of 3 or 4. The justification for combining the levels 3 and 4 is that the TG would then end with either a ToBI tone of type 'X-' or a 'X-Y%' which means that our TG should contain one one accent carrying phrasal stress, zero or more other accents, and end with either a phrase tone or a boundary tone.

There is a reasonable argument that we should have grouped entities ending in indices 4 and 5 together as they are both marked by a boundary tone, but this would not give us a TG unit necessarily consisting of only one accent carrying phrasal stress. The existence of internal phrase boundaries and multiple phrasal stresses within the smallest phrasal unit suggests that a categorisation using a smaller smallest phrasal unit may be more appropriate.

The 4–5 break index distinction is particularly difficult one to make and to further investigate this issue, and to look at the options for producing a simpler model, the second more complex partitioning of the data concentrates on attempting to restrict the allowed prosodic groupings. This is done by insisting that an IP consists of at most three TGs. The partitioning is carried out based on the break indices as before, except where there would be an IP of more than three TGs, an extra IP break is inserted at the strongest (highest break index) TG boundary. If all boundaries have equal break index values, an IP break after three TGs is made. Some of the cases where IPs of more than three TGs occur can be accounted for by break indices of 4 being used instead of 5 due to the lack of a pause between phrases, where if it were not for this pause omission a index of 5 would have been used. There are other cases where this 4–5 distinction is not apparent and it is harder to justify the inclusion of an IP break, but nevertheless the idea is useful for examining the possibility of producing a simpler intonation model. If it is a bad choice, the statistical results should reflect it.

One possible problem with this analysis is the way TGs from IPs of different lengths are analysed together. What is to say that the second TG in an IP consisting of four TGs is expected to have the same characteristics as the second TG in an IP consisting of three TGs? Results from IPs of a fixed number of TGs are then compared to the overall results to investigate this. A further complication involves the TGs being of arbitrary length. Grabe (1998) shows that the heights of accents in a series are affected by the amount of intervening material. If this has consequences for the pitch range of a TG sized unit it may be complicating the results.

To address this issue we will also consider an analysis which compares TGs from IPs of different lengths to see if the pitch range characteristics of the nth TG in an IP are affected by the total number of TGs in the IP. This analysis concerns the behaviour of the overall pitch range structure as the number of TGs in the IP is varied. For example, if we compare an IP consisting of 2 TGs to one consisting of 3 TGs, are the pitch range characteristics of the second TG the same in each case? If we expect that pitch always lowers to the same level at the end of an IP (there may be multiple IPs to a sentence so we may or may not expect this phenomenon to occur) then we may expect the second TG which is IP final to be lower, or at least finish lower, than the second TG that is non final.

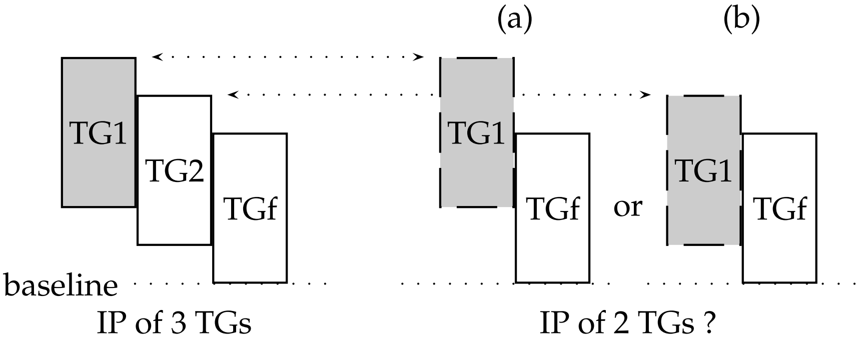


Figure 4.1: Two possible alignments for the first TG in an IP of 2 TGs compared to the first TG in an IP of 3 TGs.

Alternatively, do initial TGs in an IP align with each other, or is this alignment dependent on the number of TGs in the IP? Figure 4.1 shows a pictorial representation of the relationship: if a reset in pitch range, to a default starting pitch range, is expected at the beginning of the IP, we may expect the alignment as shown in figure 4.1a. However, if the IP is not the domain over which a full reset occurs we may get a reset as represented in figure 4.1b where the reset is such that there is space to step down once to the final TG in the IP. We may of course find neither of these alternatives, and find something completely different. There is also the question of IPs containing a single TG, does this TG look like a TG1 or a TGf or does it have distinct properties of its own.

This leads us to the following hypotheses which we test in section 4.3:

Hypothesis 3: If the TG is IP final it will exhibit different properties to TGs of the same position in longer IPs.

Hypothesis 4: Full pitch range reset occurs at the end of an IP.

4.2 Initial Analysis

An investigation considering our two levels of sub-utterance categorisation is now addressed. A selection of pitch range characteristics are measured for each phrase.

The analysis was carried out with three sets of TG categorisation:

type 0 which is based on the basic scheme outlined above and the TGs are consecutively numbered through the IP.
type 1 as type 0, but the final TG in an IP is always labelled f (for final).
type 2 based on the restrictive TG categorisation where only three TGs are allowed per IP, causing an IP break at the strongest break index.

The IPs are categorised in two different ways:

type 0 where IPs are just consecutively numbered.
type 1 where the final IP in an utterance is always labeled f.

These different combinations are then combined to produce the following data sets for analysis:

t00 : type 0 IP labels and type 0 TG labels
t01 : type 0 IP labels and type 1 TG labels
t12 : type 1 IP labels and type 2 TG labels

The motivation behind the groupings is as following:

The t00 grouping is considered a baseline control. It is the simplest application of a two level structure required to test Hypothesis 1. The t00 grouping compares the structure of two utterances by aligning them with each other from left to right. We tend to left align things by default probably because we read from left to right, but we should not necessarily assume that prosodic structure behaves this way. Aligning to the left groups all the TGs at the start of IPs together, but does not group the IP final TGs together. If we expect to see IP final effects, such as final lowering we need to consider right alignment as well. In other words, if the boundary at the end of an IP is a significant entity then it would make sense that the TGs that are IP final would share properties and that aligning them as a group in the analysis is the correct thing to do. t01 is introduced to do this, comparing the effect of the final categorisation of TGs. Finally t12 is used to test the feasibility of a simplified model and test Hypothesis 2.

4.2.1 Methodology

The variables that were measured for each TG are:

start f0 Measured as the first f0 point in the TG.
end f0 Measured as the last f0 point in the TG.
min f0 The lowest f0 value reached in the TG.
max f0 The highest f0 value reached in the TG.
Δf0 Calculated as max f0 - min f0.
mean f0 Calculated as the mean f0 value over the interval.
sd f0 The standard deviation of the f0 values over the interval.

Multivariate analysis of variance was then carried out on the data to find the statistically significant variations in the above dependent variables attributable to the given factors. The results presented in this section should be considered motivation for the further analysis carried out in section 4.3 because of the problems outlined below. They results of this initial analysis are still discussed to show us some general trends that cannot be seen from the more focused results of section 4.3.

The results presented here are possibly not accurate levels of significance because of technical complications due to the nature of the experimental design[1]. The re-analysis of the data in the next section addresses this problem and removes this uncertainty. It is also currently not possible to determine where the significant differences within a factor lie — be it between all levels of the factor or limited to between just two levels. This is also resolved with the reanalysis. Mean and standard deviation for all the f0 values in a TG are not analysed here but are considered in the re-analysis.

4.2.2 Results

Figures 4.2–4.4 show graphically the gross pitch range structure that is present in the database. Means and standard deviations for each variable for each factor grouping for each data set are shown in appendix A, tables A.1–A.18

The overall picture for the t00 data is shown in Figure 4.2. The regular structure of the pitch range starts to degrade towards the end of the third IP due to the sparseness of the data from that point onwards.

A particularly distinct feature concerns the first TG in each IP. The first TG in an IP appears to have a greater pitch range and a higher mean than the other groupings. The f0 mean of the non-initial TG is around 165-170Hz whereas the means of the f0 in the IP-initial TGs are around 200Hz. This shows that the first TG may have some special status. Another interesting feature is that the minimum f0 seems to be pretty constant across all of the TGs. There are no other clear effects shown by the graphical representation. The data is unable to show any effect caused by IP final TG, as final TGs are scattered throughout the groups, their position being dependent upon the number of TGs in a particular IP.

Statistical analysis (see Table 4.1) shows the interaction between IP type and TG type shows start_f0 to be significant at 1%, i.e the probability that the variability found within the variable start_f0 is due to chance and not due to the interaction caused by the grouping of tg type and IP type is less than 0.01.

This is the only time start_f0 shows up as significant, which is intuitively expected as it is the only grouping which includes all the IP initial TGs and only the IP initial TGs in the category TG1—the other groupings put some IP initial TGs

[1] The design of the model contains empty cells which the usual analysis of variance method cannot correctly account for.

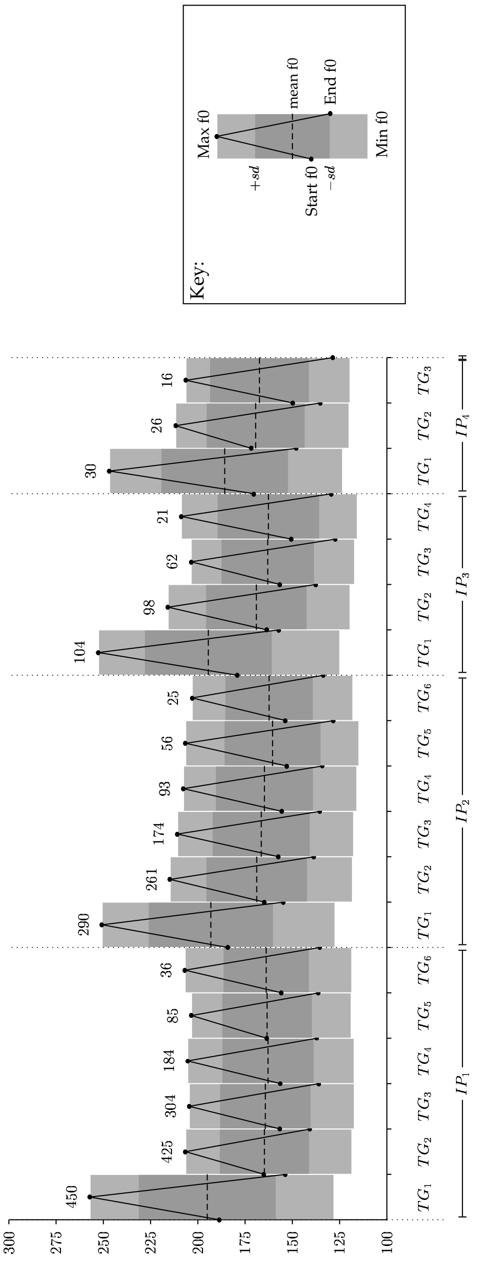


Figure 4.2: Graphical representation of pitch range for t00 data. The pitch range of each TG is represented by a grey bar. The mean pitch range is marked on each bar with a dashed line and +/- 1 standard deviation is indicated by the darker portion of the bar. Start, max and end f0 are also indicated by connected points drawn overlaying each bar.

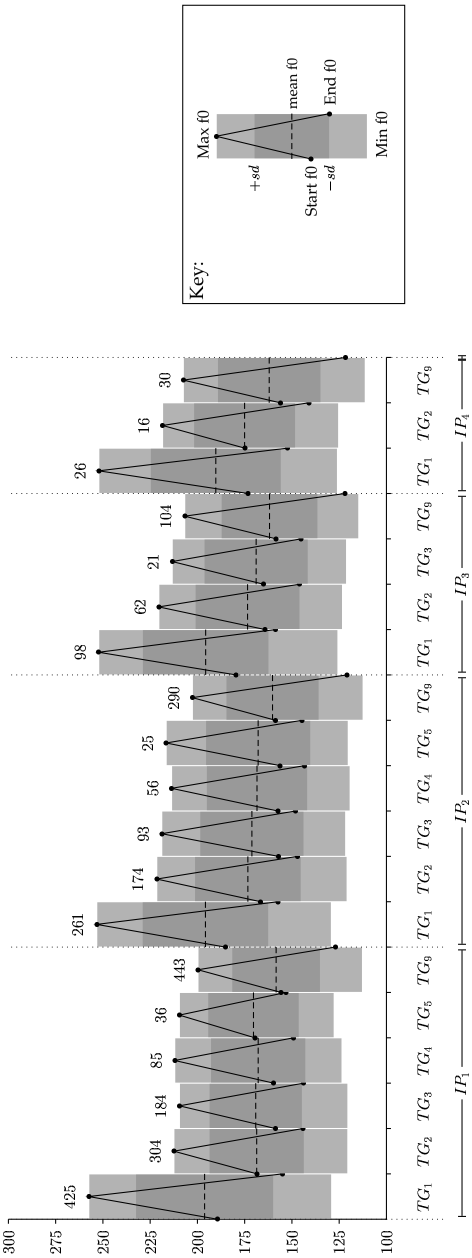


Figure 4.3: Graphical representation of pitch range for t01 data

IP_type by TG_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.05$

Univariate tests

Variable	F(22,2773)	significance
start_f0	1.49	$p < 0.01$
end_f0	1.00	
max_f0	1.46	
delta_f0	1.77	$p < 0.05$

TG_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.05$

Univariate tests

Variable	F(6,2773)	significance
start_f0	1.70	
end_f0	1.78	
max_f0	3.63	$p < 0.01$
delta_f0	2.19	$p < 0.05$

IP_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks)

Univariate tests

Variable	F(5,2773)	significance
start_f0	1.99	
end_f0	1.39	
max_f0	0.39	
delta_f0	0.85	

Table 4.1: MANOVA results for $f2b\_t00$ data.

IP_type by TG_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.05$

Univariate tests

Variable	F(22,2772)	significance
start_f0	1.10	
end_f0	1.48	
max_f0	1.09	
delta_f0	1.51	

TG_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.01$

Univariate tests

Variable	F(7,2772)	significance
start_f0	1.81	
end_f0	3.59	$p < 0.01$
max_f0	4.07	$p < 0.01$
delta_f0	1.75	

IP_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.05$

Univariate tests

Variable	F(5,2772)	significance
start_f0	1.37	
end_f0	1.97	
max_f0	0.36	
delta_f0	1.37	

Table 4.2: MANOVA results for $f2b\_t01$ data.

in the TGf category. The graphical representation suggests that the significant difference is between the first TG and the others—with no significant difference between the later categories.

There is a less significant interaction (at the 5% level) and a main effect for the variable delta_f0 (max_f0-min_f0, i.e. the overall pitch range), suggesting the position of the TG in the IP affects the overall pitch range used. Again it is thought that this is most likely to be a IP initial TG verses other TGs distinction. This result along with max_f0 showing up as significant at 1% for the main effect TG_type, reflect the differences we see in Figure 4.2.

The graphical representation of pitch range for the t01 in figure 4.3 shows similar characteristics to the t00 data, in that the IP initial TGs again seem to have a larger and higher pitch range. The IP final TG category shown here for the first time appears to be slightly lower than the other categories, but not to the same extent

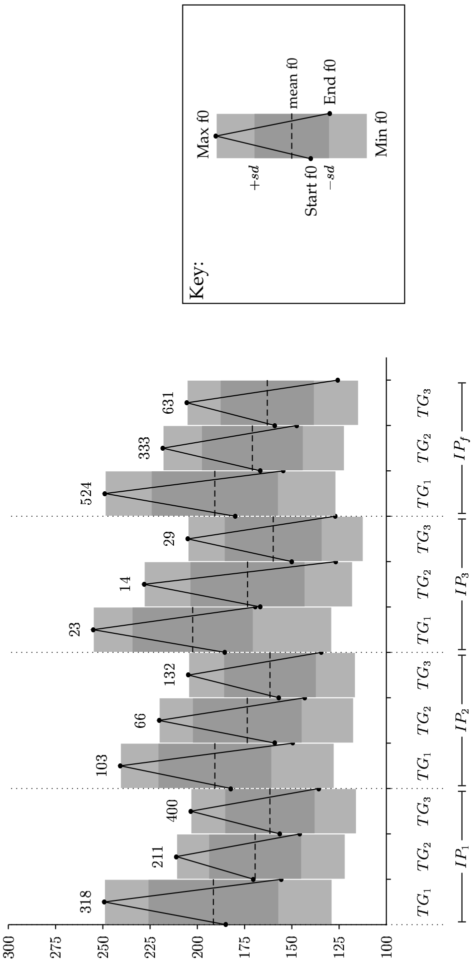


Figure 4.4: Graphical representation of pitch range for t12 data

IP_type by TG_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.01$

Univariate tests

Variable	F(10,2789)	significance
start_f0	0.77	
end_f0	2.65	$p < 0.01$
max_f0	1.69	
delta_f0	2.19	$p < 0.05$

TG_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks) $p < 0.01$

Univariate tests

Variable	F(2,2789)	significance
start_f0	1.67	
end_f0	6.33	$p < 0.01$
max_f0	17.91	$p < 0.01$
delta_f0	6.39	$p < 0.01$

IP_type

Multivariate Tests of significance

(Pillais,Hotellings,Wilks)

Univariate tests

Variable	F(5,2789)	significance
start_f0	0.25	
end_f0	1.41	
max_f0	1.68	
delta_f0	1.61	

Table 4.3: MANOVA results for $f2b\_t12$ data.

to which the IP initial TGs are higher. The only significant effects (see table 4.2) found here are for end_f0 and max_f0 found significant at the 1% level for the main effect TG_type. The max_f0 result reflects the IP initial TG high maximum, and the end_f0 result reflects the lower end to the IP final TG.

The t12 data in figure 4.4 shows an apparent downwards linear trend across TGs in all IPs where there is enough data to get a reasonable results. Start, end, max, min and mean f0 values all get lower as the TG position progresses through the IP, although min_f0 is affected less than the other variables. Delta_f0 and end_f0 are significant (see table 4.3) at the 1% level as an interaction between the two factors and as a main effect for TG_type, and max_f0 is significant as a main effect for TG_type, again reflecting what we see graphically.

4.2.3 Interpretation

We can accept Hypothesis 1 because the results clearly shows that the IP is a distinct level of intonational phrasing, with a reset occurring at the end of it, and the general trend of the results across all three methods of grouping is that effects are predominantly TG_type related, IP_type being significant as an interaction with TG_type. The t12 grouping additionally allows us to accept Hypothesis 2.

The possibly of there being some IP level effects is left open to debate, as there are some weak interactions attributable to IP categorisation, but no clear overall picture. If is difficult to tell exactly what role the TG plays from the data analysed so far as the trends found could just be the outcome of the TGs position within the IP, although the significant results for the effect TG_type suggest there is a distinct TG unit of phrasing. Further analysis of the internal structure of the TGs is needed to say more about this. This is an obvious next step as the results given here show the pitch range effects due to of the IP context of the TG, which would need to be normalised in some way when investigating the internal structure of the TG.

4.3 Second Analysis

The initial analysis of the pitch range characteristics suggested that there were only three types of distinct TG. A complete statistical analysis was hindered by the complexity of the model being analysed, mainly due to the classification of TG that was used.

We now consider adjustments and simplifications to the model which not only allow a full statistical analysis of the data but also provide a better classification for comparing TGs in different positions within the IP, and within IPs consisting of different numbers of TGs. The number of TGs which comprise an IP will be referred to as *IP length*. The position of a TG within an IP will be described numerically by the variable *TG type* and the position of an IP within an utterance will be described numerically by the variable *IP type*. The variables *TG type* and *IP type* take integer values from 1 to n where n is the *IP length* and *Utterance length* respectively. Under certain circumstances the nth TG takes the alphanumeric value f to denote the final TG in an IP. We then go on to consider the issues involved with taking this study further and analysing the alignment of the pitch events themselves which occur within the TGs.

4.3.1 *Analysing IPs of specific length*

A new analysis was carried out by splitting the data into groups of IPs of the same length. Two of the major problems concerning the analysis can be overcome by treating the data in this manner.

First, each TG can be considered the nth TG in an IP of length m, where m is fixed for each analysis. By doing this the data set is reduced into a series of subsets, each subset containing IPs consisting of a fixed number of TGs. This removes variation due to IP length in any given subset of data. Additionally this allows the comparison of TGs in particular positions across subsets with different IP lengths — this was not possible before as IP-final TGs were classified in such a way that their position in relation to the beginning of the IP was not known. This sub-setting also means that IP-initial TGs and IP-final TGs are both individually grouped together within a subset. This too could not be achieved with the basic analysis grouping methods—the variable IP length limited us to either IP-initial or IP-final TGs being grouped together but not both.

Secondly this sub-setting eliminates major problems with the statistical analysis involving empty cells in the design, as there no longer are any. For example, previously there was no data for the fifth TG in an IP of length 4, as it doesn't exist — resulting in the cell in the model being left empty.

The new design is still unbalanced so the analysis of variance needs to be carried out using a sequential sums-of-squares calculation[2], where each term in the model is only adjusted for those preceding it—resulting in an orthogonal decomposition. The more usual unique sums-of-squares calculation assumes that all the cell means are derived from the same number of data items. When this is not the case the sums-of-squares for various components of the model do not add up to the total sum of squares as required.

Repeated contrasts are also calculated on both factors ($IPtype$ and $TGtype$) which compare adjacent category values (e.g. initial TG and second TG, second TG and third TG, etc.). Cells where $n < 10$ are removed from the calculations as graphical representation of the data in these cases suggest that they should be treated as outliers. The following dependent variables were analysed as part of the design:

[2] This is usually an option which can be easily set in most statistical packages which carry out analysis of variance.

start_f0 The f0 value at the onset of voicing at the beginning of the TG.

end_f0 The f0 value at the cessation of voicing at the end of the TG.

max_f0 The maximum f0 value within the TG.

Δ_f0 The difference between the maximum and minimum f0 values within the TG.

mean_f0 The mean f0 value within the TG, calculated as $\frac{\sum f0}{n}$ for each pitch tracked f0 point within the TG.

std_f0 The standard deviation from the TG mean.

(min_f0 had to be omitted as it is a linear combination of max_f0 and Δ_f0.)

Main effects and interactions

The ANOVA results (see Table 4.4, additionally see tables A.19–A.24 for data summaries) are quite striking in that $TG\ type$ is found to be a main effect, significant at 1% for all of the variables in each of the subsets. This appears to be part of a weak interaction with $IP\ type$ for the subset containing IPs of length 2. The only other significant results are for $IP\ type$ as a main effect for $start\_f0$ in the subset of $IP\ length$ 3, this is only at 5% and as it is not consistently found in the other subsets will be ignored.

These results only show that for each dependent variable the distributions for each level of the $TG\ type$ factor are not all the same.

The results for each individual group are now examined more closely by looking at the repeated contrasts which involve the TG type. Repeated contrasts have been chosen over other contrast types as it is difficult to justify a suitable reference category from the $TG\ type$ factor. Initial TG or final TG could be argued as viable candidates, but it is felt that repeated contrasts will show any trends in the data more clearly as neither initial nor final can really be considered as a reference or default category at this stage. There is also little justification for comparing the pitch range parameters of an individual TG to those of the population as a whole as this would probably mask any downtrend effects.

Repeated contrasts show up significant differences between adjacent TGs. The initial analysis suggested that the initial and final TGs differ from each other and from medially positioned TGs (see figure 4.3), but that all medially positioned TGs are effectively the same. For this hypothesis to hold we would expect to see a contrast between first and second TGs and between penultimate and final TGs.

TGs per		TG type		IP type		TG type by IP type	
IP	Variable	*F*	*p*	*F*	*p*	*F*	*p*
		F(1,482)		F(2,482)		F(2,502)	
2	Start f0	62.54	$p < 0.01$	0.66		3.92	$p < 0.05$
	End f0	66.42	$p < 0.01$	1.21		3.32	$p < 0.05$
	Max f0	378.34	$p < 0.01$	1.37		3.35	$p < 0.05$
	Δ f0	200.57	$p < 0.01$	0.67		4.10	$p < 0.05$
	Mean f0	372.32	$p < 0.01$	0.47		2.53	
	Std f0	191.88	$p < 0.01$	0.70		5.24	$p < 0.01$
		F(2,750)		F(3,750)		F(6,750)	
3	Start f0	58.24	$p < 0.01$	2.47	$p < 0.05$	0.73	
	End f0	97.68	$p < 0.01$	3.15		1.37	
	Max f0	207.83	$p < 0.01$	0.94		0.73	
	Δ f0	76.19	$p < 0.01$	0.75		0.78	
	Mean f0	264.15	$p < 0.01$	1.14		0.29	
	Std f0	63.52	$p < 0.01$	0.31		1.52	
		F(3,580)		F(2,580)		F(6,580)	
4	Start f0	22.23	$p < 0.01$	0.47		0.23	
	End f0	28.02	$p < 0.01$	0.35		1.37	
	Max f0	86.70	$p < 0.01$	1.24		2.61	
	Δ f0	29.83	$p < 0.01$	0.18		1.43	
	Mean f0	124.77	$p < 0.01$	2.01		0.81	
	Std f0	31.16	$p < 0.01$	0.80		3.16	
		F(4,395)		F(1,395)		F(4,395)	
5	Start f0	10.28	$p < 0.01$	2.85		0.50	
	End f0	18.07	$p < 0.01$	0.20		1.55	
	Max f0	40.54	$p < 0.01$	0.18		0.43	
	Δ f0	19.71	$p < 0.01$	2.52		0.47	
	Mean f0	56.64	$p < 0.01$	0.06		0.54	
	Std f0	16.89	$p < 0.01$	0.07		1.24	
		F(5,240)		F(1,240)		F(15,240)	
6	Start f0	11.83	$p < 0.01$	0.68		0.92	
	End f0	6.44	$p < 0.01$	1.87		0.32	
	Max f0	20.18	$p < 0.01$	0.68		1.21	
	Δ f0	6.76	$p < 0.01$	1.23		1.20	
	Mean f0	34.12	$p < 0.01$	1.77		0.81	
	Std f0	6.56	$p < 0.01$	1.76		1.34	

Table 4.4: ANOVA Results for Analysis of IPs of Equal number of TGs

TGs per IP	TG type contrast	Dependent Variable t-test results											
		Start f0		End f0		Max f0		Δf0		Mean f0		Std. f0	
		t	p	t	p	t	p	t	p	t	p	t	p
2	1–2	5.90	$p < 0.01$	8.12	$p < 0.01$	16.15	$p < 0.01$	11.41	$p < 0.01$	16.23	$p < 0.01$	11.03	$p < 0.01$
3	1–2	3.47	$p < 0.01$	2.56	$p < 0.05$	9.67	$p < 0.01$	7.13	$p < 0.01$	10.35	$p < 0.01$	5.42	$p < 0.01$
	2–3	3.25	$p < 0.01$	6.15	$p < 0.01$	3.93	$p < 0.01$	1.84		4.05	$p < 0.01$	1.71	
4	1–2	3.29	$p < 0.01$	1.06		6.16	$p < 0.01$	4.37	$p < 0.01$	7.97	$p < 0.01$	3.58	$p < 0.01$
	2–3	1.05		1.13		1.09		0.85		1.92		-0.09	
	3–4	1.46		4.04	$p < 0.01$	2.13	$p < 0.05$	0.41		2.09	$p < 0.01$	0.98	
5	1–2	5.58	$p < 0.01$	5.76	$p < 0.01$	11.82	$p < 0.01$	8.44	$p < 0.01$	13.67	$p < 0.01$	7.46	$p < 0.01$
	2–3	1.28		1.16		-1.03		-1.42		0.16		-1.35	
	3–4	-1.87		0.50		-1.29		-1.01		-0.69		-0.67	
	4–5	-2.26	$p < 0.05$	0.14		-3.36	$p < 0.01$	-2.98	$p < 0.01$	-4.67	$p < 0.01$	-2.42	$p < 0.05$
6	1–2	6.79	$p < 0.01$	4.61	$p < 0.01$	8.42	$p < 0.01$	4.52	$p < 0.01$	11.82	$p < 0.01$	3.95	$p < 0.01$
	2–3	0.47		-1.38		-1.56		-0.79		-1.56		-1.14	
	3–4	-1.61		-0.25		-1.36		-0.62		-0.52		-0.70	
	4–5	-1.28		0.46		-0.25		-0.32		-0.81		0.42	
	5–6	-1.99	$p < 0.05$	0.04		-1.37		-0.71		-3.63	$p < 0.01$	-0.06	

Table 4.5: Anova Repeated Contrasts for Analysis of IPs of Equal number of TGs

The contrast results show clearly that the first and final TGs of an IP differ from those between them (see Table 4.5). For all groups the only repeated contrasts found to be significant are between the first and second TGs in an IP, or between the penultimate and the final TGs.

For IPs of length 2, all of the variables tested showed a contrast significant at 1%. For IPs of length 3, the final-penultimate position contrast is lost for Δ_f0 and $std.\_f0$, and the initial–second contrast for end_f0 has dropped to 5%, but the other 9 contrasts remain at 1%. For IPs of length 3, the end_f0 initial–second contrast is no longer significant, and in addition to the loss of the final–penultimate contrast for $std.\_f0$, it has also been lost for $start\_f0$. IPs of length 5 and 6 have all the initial–second contrasts at 1%, while IPs of length 5 have all but the end_f0 final–penultimate contrast, and IPs of length 6 have only this contrast for $start\_f0$ and $mean\_f0$.

The general trend seen here is that all initial–second contrasts for all variables are significant at 1%, as opposed to the final–penultimate contrasts which are only evident in 60% of the cases. The exception is for the variable $mean\_f0$ for which a contrast is always present — suggesting that there is a definite average pitch range distinction, even when this is not clearly shown by the other variables.

So we can safely conclude that there are clear pitch range differences between individual TGs in an IP, and those differences specifically manifest themselves as a very clear distinction between the TG which is IP-initial and those that follow it. There is also a distinction between the pitch range characteristics of the final TG in the IP and those that preceded it, although this may not always be as clear as the IP-initial distinction.

Across-subset effects and contrasts

We now turn our attention to Hypotheses 3 and 4. We look at results for the first three TG positions for IPs containing up to 6 TGs. We will also look at the TG category IP final, but this is considered in isolation, as the data is drawn from the t01 data set (where IP final TGs are classified together) and is hence not consistent with the t00 data which is used elsewhere in this analysis.

The analysis of variance results, shown in Table 4.6, reveal in general no interactions and show $IP\ length$ to be a clear main effect for most variables. There is also a significant main effect for IP type for some variables but which particular variables differs depending on the TG position. There are no significant IP type contrasts between individual values of IP type making interpreting this effect difficult. We therefore concentrate on the IP length main effect.

The repeated measures contrasts show that significant differences only occur between the categories which involve a TG in IP final position. For example, there is only a contrast for the third TG in an IP, between IPs of length 3 and 4; here the TG is in final position in the IP of length 3, and in non-final position in the IP of length 4. This seems to be generally true for all of the dependent variables measured.

This confirms hypothesis 3 that an IP final TG has special status, and the TG's finalness overrides properties defined by its relative position from the beginning of the IP. These results also show us that all IP initial TGs (excluding IP final ones of course) have the same properties, suggesting that the correct relationship is as shown in in figure 4.1a, confirming hypothesis 4.

4.3.2 Resynthesis using TG and IP structure

To verify that the above results would be a reasonable and useful addition for a prosodic model of speech synthesis, the following hypotheses were informally tested by way of resynthesis techniques:

Hypothesis 5: An utterance from the data set resynthesised in such a way that the pitch range for each TG, characterised by a mean and standard deviation, is equal to the average pitch range for that type of TG found in the data set, should be comparable to the original utterance. If this is not the case, then the TG categorisation is probably not a good categorisation.

Hypothesis 6: If an utterance generated by a TTS system does not incorporate such a detailed model of prosodic structure, then the prosody of this utterance should be able to be improved upon by imposing this finer level of structure upon it. This essentially imposes the structure we assume for the analysis onto an unstructured utterance. If no improvement occurs by using this structure then it suggests that the assumed structure does not capture the structure of the database.

TG		IP length		IP type		IP length by IP type	
Position	Variable	*F*	*p*	*F*	*p*	*F*	*p*
		F(5,709)		F(1,709)		F(5,709)	
1	Start f0	1.92		2.12		0.52	
	End f0	7.79	$p < 0.01$	0.60	$p < 0.05$	0.08	
	Max f0	2.04		4.31		0.65	
	Δ f0	0.38		2.79		0.87	
	Mean f0	15.38	$p < 0.01$	0.59	$p < 0.01$	0.38	
	Std f0	0.16		12.84		0.73	
		F(4,662)		F(1,662)		F(4,662)	
2	Start f0	2.74	$p < 0.05$	0.00		0.57	
	End f0	14.64	$p < 0.01$	0.64		2.67	$p < 0.05$
	Max f0	16.14	$p < 0.01$	10.81	$p < 0.01$	0.33	
	Δ f0	5.76	$p < 0.01$	8.52	$p < 0.01$	0.06	
	Mean f0	22.49	$p < 0.01$	8.69	$p < 0.01$	1.04	
	Std f0	7.27	$p < 0.01$	13.18	$p < 0.01$	0.40	
		F(3,456)		F(1,456)		F(3,456)	
3	Start f0	0.59		0.00		2.26	
	End f0	25.33	$p < 0.01$	0.31		2.29	
	Max f0	11.56	$p < 0.01$	4.97	$p < 0.05$	1.54	
	Δ f0	2.70	$p < 0.05$	4.82	$p < 0.05$	0.67	
	Mean f0	21.47	$p < 0.01$	1.45		1.16	
	Std f0	3.06	$p < 0.05$	5.28	$p < 0.05$	1.39	
		F(5,709)		F(1,709)		F(5,709)	
f	Start f0	5.43	$p < 0.01$	2.64		0.84	
	End f0	2.06		12.07	$p < 0.01$	0.99	
	Max f0	21.22	$p < 0.01$	2.67		2.16	
	Δ f0	20.95	$p < 0.01$	2.82		2.12	
	Mean f0	10.59	$p < 0.01$	3.94	$p < 0.05$	2.15	
	Std f0	14.99	$p < 0.01$	4.22	$p < 0.05$	2.18	

Table 4.6: Anova Results for Cross Comparison of TGs in the same Positions in IPs of Different Lengths. (Note: groups 1-3 are from t00 data sets and group f is from t01 data set and is a separate test.)

TG	IP length	Dependent Variable t-test results Start f0		End f0		Max f0		Δf0		Mean f0		Std. f0	
pos.	contrast	t	p	t	p	t	p	t	p	t	p	t	p
1	1–2	-1.25		-3.12	$p < 0.01$	-1.81		0.21		-4.48	$p < 0.01$	-0.60	
	2–3	-0.39		-2.59	$p < 0.01$	-0.92		0.99		-2.49	$p < 0.05$	0.64	
	3–4	0.39		-0.74		-0.01		0.31		-1.50		-0.04	
	4–5	0.36		-0.38		-0.39		-0.98		0.04		-0.46	
	5–6	-2.28	$p < 0.05$	-0.53		-0.42		0.61		-2.41	$p < 0.05$	0.54	
2	2–3	-2.49	$p < 0.05$	-7.36	$p < 0.01$	-7.01	$p < 0.01$	-4.28	$p < 0.01$	-7.17	$p < 0.01$	-4.83	$p < 0.01$
	3–4	0.26		0.89		0.11		0.30		-0.82		0.34	
	4–5	-0.46		-0.42		0.59		0.77		-0.67		0.81	
	5–6	0.05		1.70		0.32		-0.20		0.52		0.40	
3	3–4	0.37		-7.64	$p < 0.01$	-5.55	$p < 0.01$	-2.76	$p < 0.01$	-5.15	$p < 0.01$	-3.30	$p < 0.01$
	4–5	0.91		0.52		0.95		0.59		-0.95		0.81	
	5–6	0.85		0.51		0.31		-0.12		-0.88		0.06	
f	1–2	3.91	$p < 0.01$	1.06		9.91	$p < 0.01$	9.94	$p < 0.01$	6.80	$p < 0.01$	8.49	$p < 0.01$
	2–3	1.20		1.61		-0.93		-1.11		-0.88		-1.61	
	3–4	0.74		-0.36		-0.16		-0.48		-0.03		-0.49	
	4–5	-0.51		1.19		0.31		0.27		1.27		0.36	
	5–6	0.39		-1.54		-0.65		-0.02		-0.41		0.25	

Table 4.7: ANOVA Repeated Contrasts for Cross Comparison of TGs in the same Positions in IPs of Different Lengths.

These hypotheses were tested using only the TG mean and standard deviation values with respect to the total number of TGs to the IP and the TG position in the TG (IP_length and tg_type respectively).

Hypothesis 5. was tested using a crude fixed frame LPC resynthesis method. An informal comparison was made of utterances resynthesised with their original f0s to those with f0s normalised across each TGs and rescaled with respect to the average mean and standard deviation parameters for that TG position. The utterances with the altered f0 were found to be similar to those with the original f0. Although no formal evaluation was carried out, these results suggest that the average parameters are a reasonable representation of individual TGs.

Hypothesis 6. was tested by generating example sentences with Festival. The resulting f0 was normalised with respect to the f0 mean and standard deviation of the utterance as a whole, and then scaled appropriately with respect to the the mean and standard deviation found for that TG position. The normalisation was carried out in the usual way by calculating:

$$f0_{norm} = \frac{f0 - \mu}{\sigma}$$

where μ is the f0 mean and σ the f0 standard deviation. The rescaling is the reverse process using different mean and standard deviation values.

$$f0_{scaled} = f0_{norm}\sigma + \mu$$

Additional constant scaling and offset parameters were also included to compensate for the fact that the original f2b utterances were spoken by a female (only a male Festival voice was available). Festival was then used to resynthesise the utterance with this new f0.

The initial result of this process sounded rather unnatural, particularly at the TG1–TG2 boundary where a noticeably unnatural sounding drop in pitch often occurred. A better result than this was found by not using the values for the IP initial TGs, but using TG2 values for TG1s. Phrase initial highs were of course lost — phrase initial is used here and not sentence-initial as the data suggests phrase initial highs can occur at all IP level boundaries, whether a sentence boundary or not. This result suggests that mean and standard deviation calculated across the TG are generally sufficient to capture some sort of pitch range effect which can improve the naturalness of synthesised speech. However, what they do not seem able to capture is the more subtle effect of phrase initial highs. As this is the case it also seems unreasonable to expect to account accurately for f0 behaviour related to sentence final lowering.

The results also suggest that f0 mean and standard deviation alone would not be enough to position pitch elements within a TG if all TGs were treated equally and only distinguished by the f0 mean and standard deviation. There are two options for a model: either IP initial, IP medial and IP final TGs would have to be treated separately, or different and/or additional parameters would be needed to describe the f0 behaviour across the TG.

Therefore, the way in which we treat TG types differently will depend on how we consider the structure of TGs in differing positions to vary from each other. If we consider all TGs to be equal in structure and differ only in overall level or range then ideally we would want a general model which captures pitch range behaviour in this way. On the other hand if we consider the structure of an IP initial tone group to differ from an IP medial TG, to differ from an IP final TG, then perhaps a model which treats these TG types as different is more appropriate.

So, we need to decide if the phenomenon of phrase initial high is specifically associated with phrase initial TGs themselves or is just a pitch event which naturally occurs at the beginning of phrase initial TGs. That is to say, if we assume that the f0 contour is made up of pitch events plus an overlying phrase contour, we need to decide with which the phrase initial high is associated. Teasing apart pitch range of a phrase from the pitch of pitch events within it is quite difficult as to a large extent the pitch range is partially if not completely defined by the realisation of pitch events.

A similar argument exists for final lowering, which we have so far seen no evidence for in this data. That is not to say that there is any strong evidence against it either, as we have not yet examined descending sequences of accents and associated boundaries. We return to the issue of final lowering when we consider boundaries in chapter 6.

CHAPTER 5

Pitch Event Analysis

Our attention now shifts from looking at how prosodic structure affects pitch range to seeing how pitch events are realised within the prosodic units found to be significant in the previous chapter, namely tone groups (TGs).

Our intention is to learn how to represent pitch events within TGs in such a way as to be useful when modelling intonation for speech synthesis. We intend to use the knowledge about prosodic structure acquired in the previous chapter to group pitch events with like prosodic structure together and hence minimise variation within a given group. This approach helps us to reach our goal of being able to capture the intonation style of the database being analysed.

In this chapter we focus in on the tone group (TG) and continue the analysis started in the previous chapter by now looking at how pitch events are affected by both the phrasing structure which surrounds them and the segmental material on which they are placed.

5.1 Pitch Event Alignment

We can consider a pitch contour to be a function of time—at a given time there is a related pitch value. The nature of this function however is not simple and is really dependent on considerably more than just time. We can take a step back and consider the modelling of intonation to be the defining of this function which we can then plug values (e.g. time, TG number, etc...) into to generate a pitch contour.

We cannot currently expect to be able to fully define this function in a mathematical sense as we do not fully understand the underlying processes which result in intonation. We can however look at various aspects of the resulting function and consider pitch events in terms of these.

There are three particular aspects of pitch contours relating to pitch events which we can analyse. These are event shape, event time alignment and event pitch positioning. Shape considers the overall shape of the pitch event and is generally related to the type of pitch event in question, as specific accents and boundaries have specific shapes. Time alignment considers how this shape is positioned with respect to time (or the segmental structure of the utterance), and pitch positioning considers how the shape is positioned with respect to pitch.

Using the above aspects of pitch, we can consider a pitch contour representing a pitch event as being as being a particular shape which manifests itself at a particular pitch at a particular time. A graphical representation of shape, pitch positioning and time alignment is shown in figure 5.1. Here the pitch event is stylised as three points, a start, a peak and an end, with a smooth curve connecting these points. The shape is that of a rise-fall accent. The time alignment of the event is shown relative to the segmental material with which it is associated. This is usually a syllable or the voiced part of a syllable. The three points representing start, peak and are shown as these are the points we will use to characterise pitch events throughout the analysis. The small f0 tails at each end of the accent show that the start and end of the accent do not necessarily occur at the start and the end of the voiced portion of the syllable.

5.1.1 Pitch shape

Pitch shape governs how the f0 changes through the duration of the pitch event. The nature of this change accounts for an accent or boundary's type. This type can be described by associating a given linguistic theory that describes pitch events with the accent shape. Here we will currently only distinguish between general shapes which we will call falls, rises, and rise-falls, as this classification is easily available from the Tilt (Taylor 1994) parameters.

5.1.2 Pitch positioning

We need to associate this local f0 shape with the overall pitch range of the utterance, and more specifically we need to associate this locally to the parameters which control the pitch level for the TG in which the accent occurs, to give the accent position in the local pitch range.

5.1.3 Time alignment

Time alignment governs how the pitch event aligns itself with the syllabic structure with which it is associated. It is possible to consider the alignment of the pitch event with the syllable in a number of ways. First there is the issue of absolute alignment versus relative alignment. The points on the pitch event can be related absolutely in time to the syllabic structure (e.g. the pitch event start is 60 milliseconds from the start of the syllable). Tilt (see section 2.7) measures time alignment in this way. Or they can be related relatively (for example, the pitch event start is 30% of the way through the syllable).

There is also the question of which part of the syllabic structure the pitch event is best related to: the syllable as a whole, or specific parts of the syllable. If it is specific parts of the syllable then a decision needs to be made as to which parts — onset/rhyme, onset/nucleus/coda or some alternative division based on voiced/unvoiced distinctions, like the Bell Labs model (see section 2.12.2)

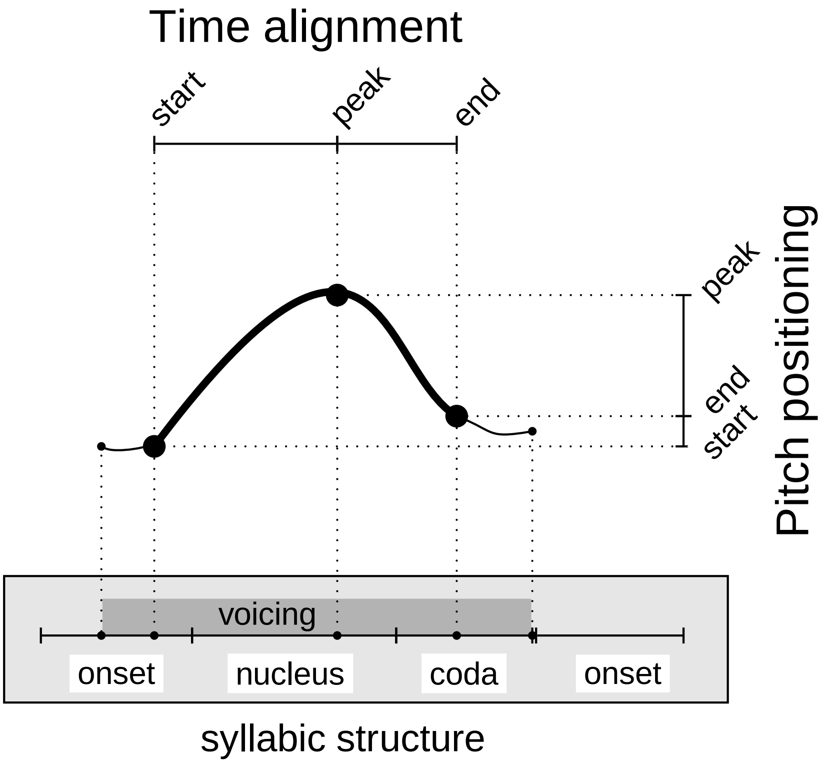


Figure 5.1: Stylised representation pitch event alignment with a syllable

which uses the s-rhyme to describe alignment. Unless pitch events align consistently at the same point in all syllables, however they are measured, there is also the need for a classification of syllables.

Grabe's (1998) finding that the peak positions of locally down-stepped accents are affected by the amount of segmental material between successive accents implies that the analysis here needs to consider not just details concerning the segmental material which the accent is associated with, but also at least some knowledge of the segments surrounding accents and the segmental distance between accents.

There is insufficient data available to address the general problem of pitch event alignment with different segmental material. We currently have just over 4100 accent type pitch events. If we assume this data is balanced in the sense that each accented syllable type is represented an equal number of times, then using say 25 onset types and 19 vowel types, then ignoring codas completely we would get 475 different syllable types. This would be fewer than 10 examples of each. Phonetic features could be used to group different dimensions of onsets and vowels, reducing the number of categories, but this still assumes that time alignment is only affected by the onset and nucleus of a one syllable context. We do not however abandon time alignment completely, we use it to distinguish sub classes of accent within the rise-fall category of accent as determined by tilt parameter value.

5.2 Accent Types and Alignment

Figure 5.2 is a scatter plot of *start time* vs *tilt_value* for accents in the f2b database. It shows that these accents fall into three groups, highlighted by lines in the figure. The main group of points (just looking at the distribution along the y-axis) in the centre of the plot with tilt values just greater than zero are "rise-fall" accents and two secondary groups representing "rise" and "fall" located at the top and bottom of the plot respectively with tilt values close to +/- 1.

Additionally there is a scattering of points with late start times. These points are few in number and are considered outliers and removed from the analysis.

Figure 5.3 is a further scatterplot. It shows *peak time* vs *tilt_value* for accents labelled 'a'. Here we see a somewhat linear relation between tilt value and peak time, the ‘falling’ accents having early peaks and the ‘rising’ accents having late peaks as we would expect.

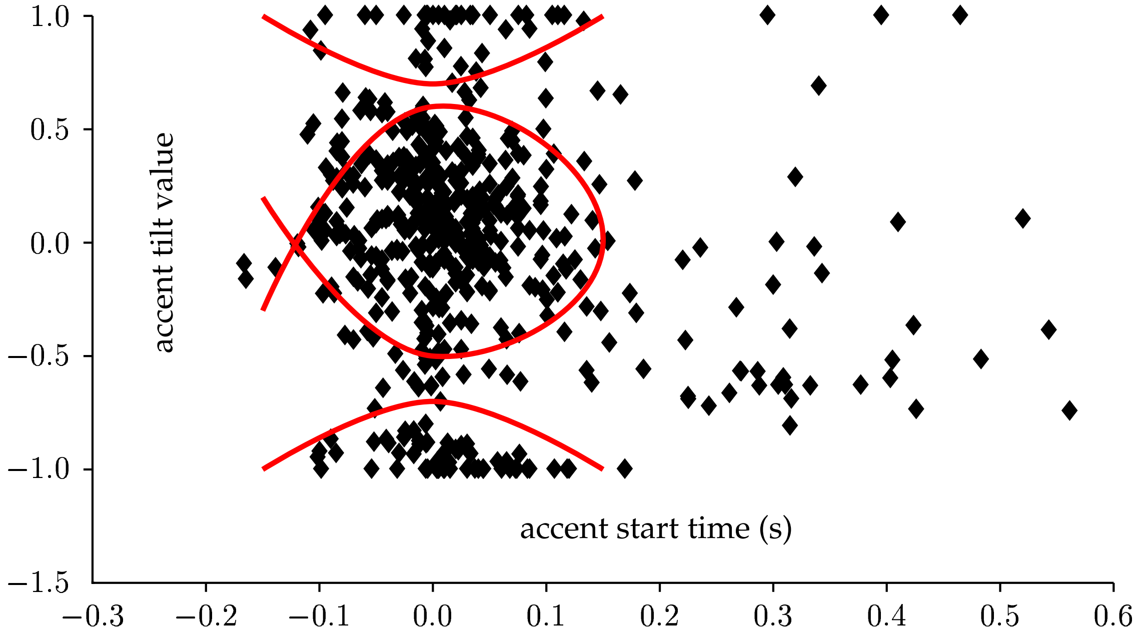


Figure 5.2: Scatterplot for accents, showing start time plotted against tilt value. This plot shows a sample of 12.5% of pitch events labelled ‘a’. Lines are added to highlight the groupings.

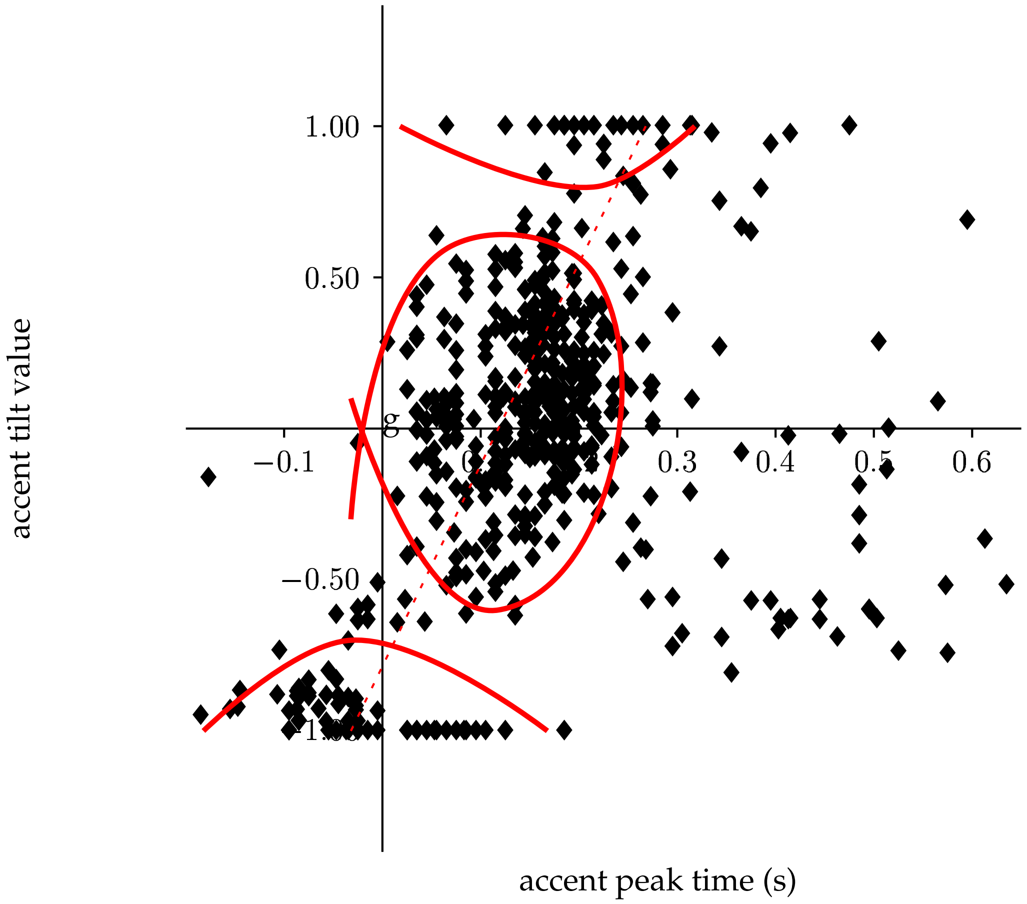


Figure 5.3: Scatterplot for accents. Showing peak time plotted against tilt value. This plot shows a sample of 12.5% of pitch events labelled ‘a’.

In the analysis we carry out we concentrate solely on the *rise-fall* accents. We don’t consider there to be sufficient data to produce a rigorous analysis of the other categories. However, the means and standard deviations for the accents which fit into these groupings are summarised in Appendix B so that the data is available to anyone who wishes to include them in a model.

5.3 Hypotheses

We are interested in how sequences of accents in a TG position themselves, and we are specifically interested in how this positioning relates to the pitch range we find in different TGs, with a particular interest in how the accents in IP initial TGs manifest themselves.

Listening to the data we are working with suggests that the first accent in an IP initial TG is higher than the first accent in non-ip initial TGs. That is, there is something about the start of a phrase which causes the first accent to be particularly high in pitch. This is one of the properties of the speech we are looking at that we try to demonstrate as significant.

The hypotheses we are testing here are:

Hypothesis 1: Accents in sequences have distinct pitch positioning and descend in pitch towards the end of the TG.

Hypothesis 2a: The pitch of the first accent in an IP initial TG is higher than that of first accents in non IP initial TGs.

Hypothesis 2b: The accents in IP initial TGs are positioned at a higher pitch than would be expected by a general downward trend in pitch from the start to the end of the TG.

The pitch range effects we saw in the previous chapter could be caused by either Hypothesis 2a or 2b, or it may be the case that both are true or of course neither.

5.4 Methodology

We use the Festival tools to extract the following information relating to each pitch event in the f2b corpus.

- Prosodic structure information: the position in the phrase of the tone group containing the pitch event, and the total number of tone groups in the phrase.
- Pitch event description: pitch event type, and its relative position in the tone group, and the total number of pitch events in the tone group. The tilt value of the accent to classify its shape.
- Voicing information: the start/end time of voicing segment which overlaps with the syllable which the pitch event is associated with.
- Pitch event timing information: start/peak/end time.
- Pitch event f0 information start/peak/end f0 values (Hz). Including f0 values normalised with respect to the tone group mean and standard deviation f0.

Time values are specified relative to the start of the syllable that a pitch event is associated with. This is considered to be more useful than absolute time values (which would be relative to the start of the utterance).

The utterance building process (see section 3.2) often highlights cases where different data sources do not mesh together well. This happens when the label files representing the original data do not align sufficiently with each other or data is just missing. Utterances which could not be built in a satisfactory manner were discarded and not used in the analysis.

5.5 Results

We first examine the relationship between the relative position of accents within the utterance structure and the pitch at which the accent is realised. With statistical analysis in mind, accent position is specified by the independent variables *accent_number* and *accent_count*. For example, if an accent has accent number 2 and accent count 4, then it is the second accent in a group of 4.

Accent f0 is measured in three places: at the start, peak and end of the accent. The peak should be distinct from the start and end of each accent as all the accents involved are 'rise-fall' accents so three points should be sufficient to capture a crude representation of shape.

Figure 5.4 shows the average f0 accent positions for accents in sequences of up to 4 accents in length. The accents are grouped by accent number; that is, the first accent from sequences, irrespective of sequence length, are grouped together. A consequence of this grouping method is that the final accents in sequences of different lengths are not grouped together. The closeness of the accents within the groups as shown and a noticeable difference between the groups (specifically the first three groups) suggests that grouping the accents like this is preferable to, say, grouping the accents as final, penultimate, anti-penultimate, etc.

The overall pattern of the accents is a distinct downward trend which suggests Hypothesis 1 holds. We see in figure 5.4 that there is a reasonable distinction between the pitches of the first three accents in a sequence, but the distinction between the pitch of third and fourth accents in a sequence of four accents is negligible. This may be due to the pitch range baseline being reached or just that there is less data available to make up this category, resulting in a higher variance.

Figures 5.5– 5.7 show the same data but separated into three subsets. This allows us to consider Hypothesis 2 in more detail, where we need to consider the accents in IP initial TGs and particularly the first accent in an IP initial TG.

Each subset contains the accent sequences from TGs with specific positions within an IP. Figure 5.5 shows data for IP initial TGs, figure 5.6 shows data from IP medial TGs and figure 5.7 shows data from IP final TGs. The analysis of the overall pitch range of specific TGs in the previous chapter showed significant differences between the pitch range of IP-initial/medial/final TGs. It could not however show whether this relationship was attributable to the TG as a whole or to specific pitch events within the TG. To some extent pitch range is an abstract concept in that it is predominantly realised by pitch events. The bits of pitch contour between pitch events are regarded as uninteresting because pitch is assumed to be resting at a default level in the absence of explicit pitch targets. With this in mind, the pitch range for a given TG can be seriously affected by the position of one pitch event within it. In our case a higher pitch range could be the result of just one accent which is raised in pitch within a TG or it could be the result of all of the accents being raised in pitch. The former we consider the effect of the pitch event, the latter we consider to be a more general pitch range property of the TG. Analysis of the pitch positions of these subsets for accent sequences is used to attempt to clarify this point.

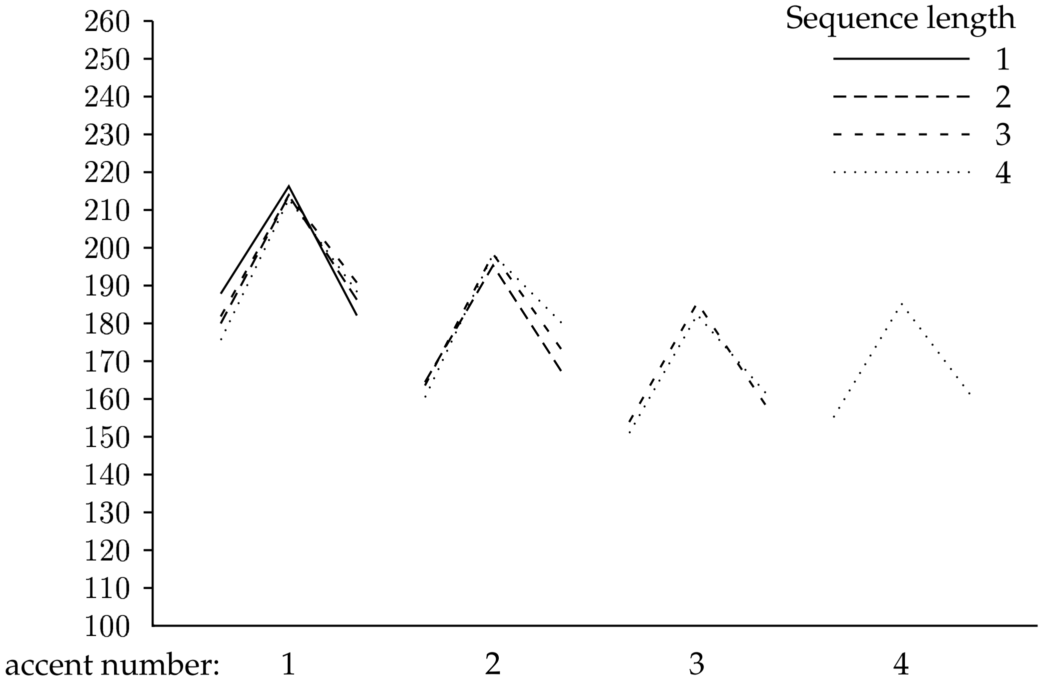


Figure 5.4: Pitch of rise-fall accents in sequences of up to 4 accents.

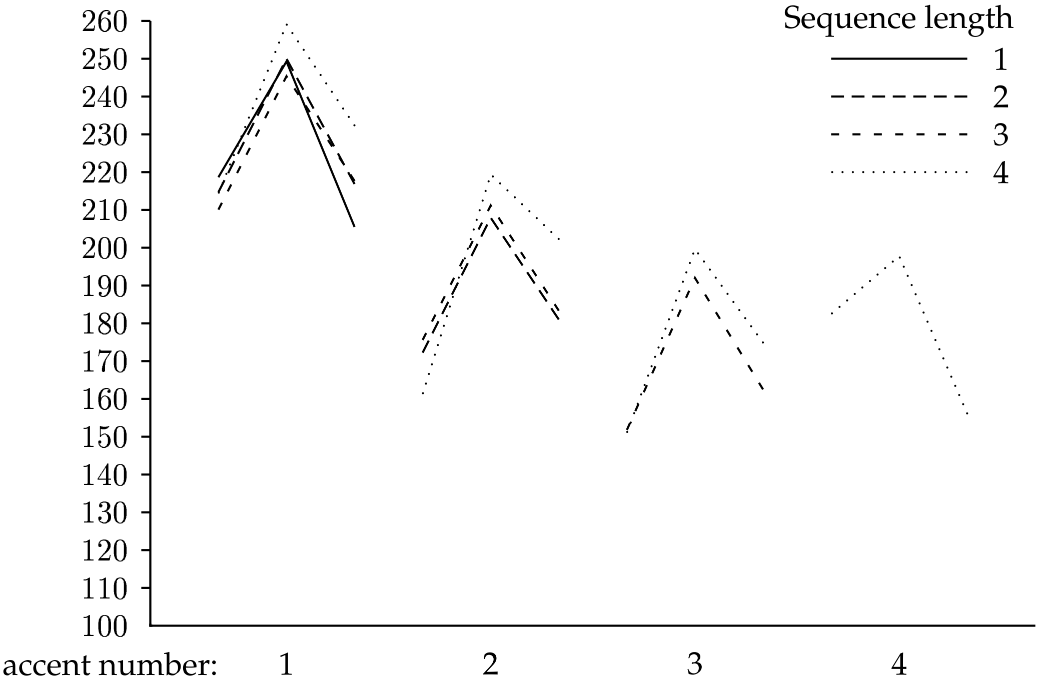


Figure 5.5: Pitch of rise-fall accents in IP initial TGs.

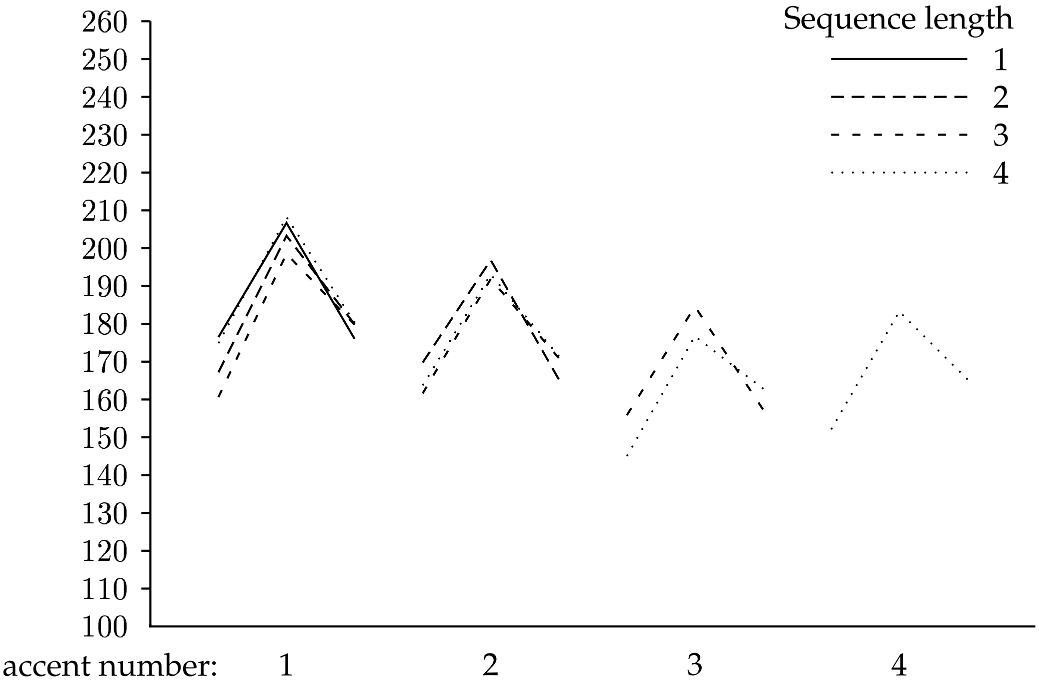


Figure 5.6: Pitch of rise-fall accents in IP medial TGs.

Firstly if we compare the IP medial TGs (figure 5.6) with the results for all TGs combined (figure 5.4) we see that there is very little difference in the position and height of the accents. This is what we would expect as most of the TGs are IP medial. However, if we compare the IP initial TGs (figure 5.5) with the IP medial TGs (figure 5.6) we find that the accents in the IP initial TGs are positioned higher than the accents in the IP medial TGs, although not uniformly so. Accent 1 is about 45Hz higher in the IP initial TG, but by accent 2 this difference has decreased to about 20 Hz, and by accent 3 it is negligible, suggesting that the pitch range effect is not simply an overall TG effect, nor an isolated effect affecting only the very beginning of the TG or only the first pitch event in the TG.

We now carry out an analysis of variance with the above data (see Tables 5.1 & 5.2). From the above observations we take *TG type* and *accent number* as independent variables, and use start, peak and end f0 values for dependent variables. Table 5.1 shows a clear effect for each independent variable with each dependent one ($p < 0.001$).

Moving to the contrasts, in the first row of table 5.2 we see a clear contrast between IP initial and IP medial TGs. The second row shows contrasts between IP medial and IP final TGs but only for peak and end position of the accent. The

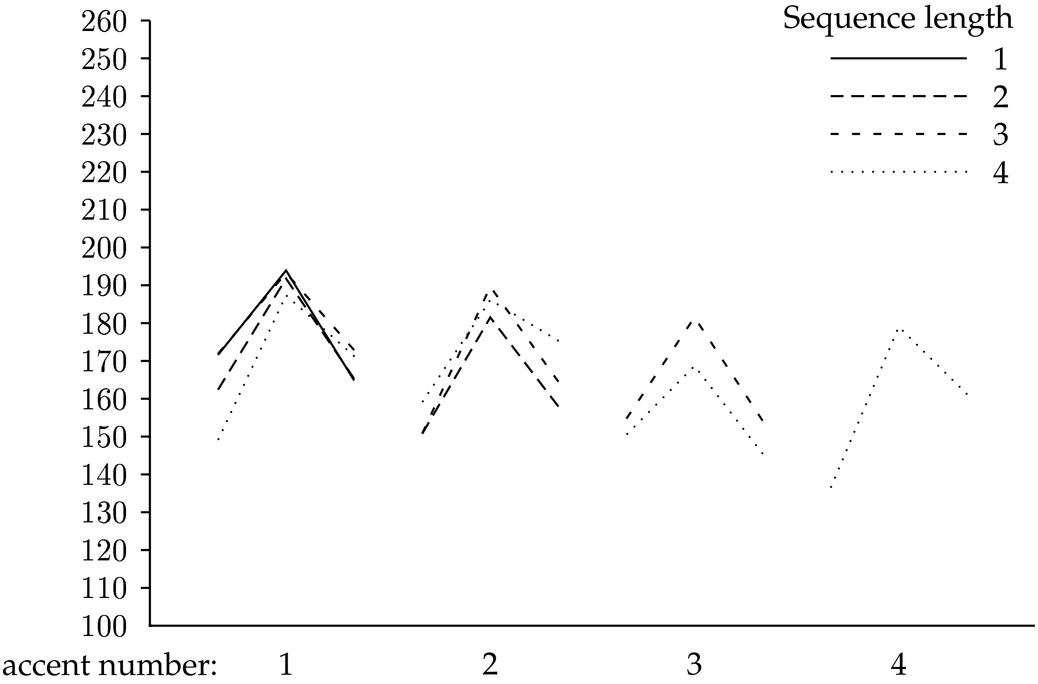


Figure 5.7: Pitch of rise-fall accents in IP final TGs.

third and fourth rows of the table show clear contrasts between the first and second accents in a sequence and the second and third accents in a sequence, but no distinction between third and fourth.

The accent position contrasts reflect what we saw graphically, and allow us to accept Hypothesis 1, for at least the first three accents in a sequence.

The interaction contrasts are a little harder to interpret but show a particular contrast between the interaction of TG position and accent position for the TG initial/TG medial vs. first accent/second accent interaction. This suggests that the pitch position of the first accent in an IP does have some special status, independent of being either in the initial TG or being the initial accent in a TG. This is true for the start, the peak and the end position of the accent and shows hypothesis 2a to be true. Hypothesis 2b is also shown to be true as the first row of the contrasts table shows significant differences in pitch for start, peak and end positions between IP initial TGs and IP medial ones.

Figure 5.8 shows a representation which could be the basis of a model for f0 at each accent position based on the group means. The start f0 points are grouped together on the left, a central group contains the peak positions and a final group contains the end positions. The different lines represent points from different TG positions and the points on each line represent the accents in a sequence in that TG. An example f0 sequence is shown by the dashed line joining equivalent points from the three groups. These points would represent the second accent in an IP initial TG. Some of the start and end points for the fourth accent in a sequence are higher than might be expected. This is probably due to there being less data for these accents than for the first three accents.

	TG position		accent position		TG pos. * accent pos.	
	F(2,3049)	p	F(3,3049)	p	F(6,3049)	p
start f0	19.23	< .001	54.50	< .001	49.85	< .001
peak f0	39.76	< .001	107.33	< .001	128.81	< .001
end f0	20.19	< .001	96.38	< .001	104.01	< .001

Table 5.1: ANOVA results

	start f0		peak f0		end f0	
TG pos. contrasts	t	p	t	p	t	p
initial/medial	4.57	< .001	6.21	< .001	3.62	< .001
medial/final	0.17	.169	2.65	.008	3.16	< .001
accent pos. contrasts	t	p	t	p	t	p
first/second	9.92	< .001	13.54	< .001	12.54	< .001
second/third	3.13	< .001	4.86	< .001	5.27	< .001
third/fourth	-0.10	.924	0.04	.969	-0.77	.439
interaction contrasts	t	p	t	p	t	p
i/m–1/2	8.13	< .001	9.32	< .001	6.24	< .001
i/m–2/3	0.99	.320	0.61	.539	2.14	.032
i/m–3/4	-2.07	.038	0.14	.887	1.66	.097
m/f–1/2	-2.19	.028	0.17	.868	1.79	.074
m/f–2/3	1.75	.080	1.35	.178	-0.28	.778
m/f–3/4	-0.22	.822	-0.17	.861	-0.26	.798

Table 5.2: Repeated contrasts

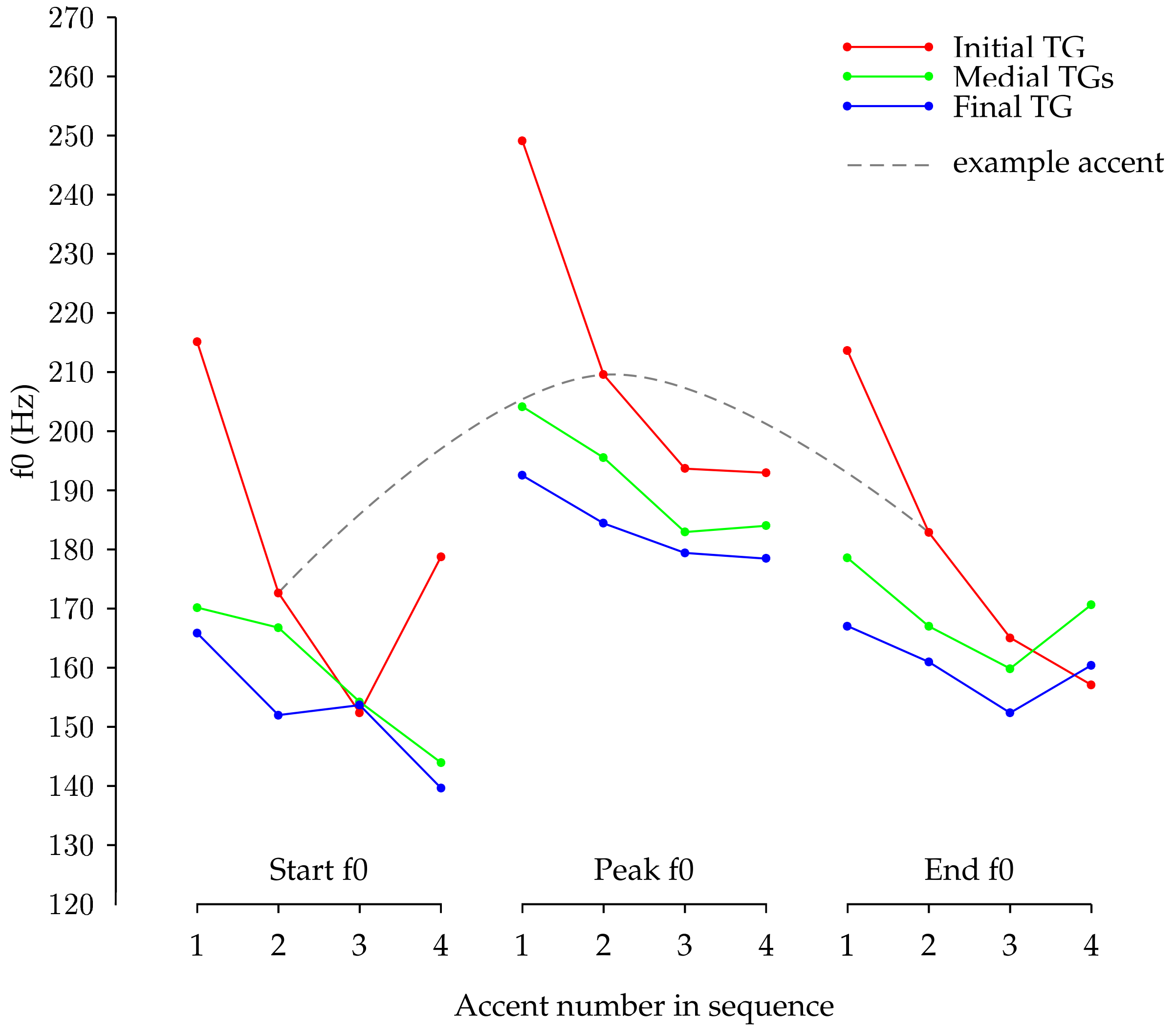


Figure 5.8: Summary of rise-fall accent group means, showing potential basis for a model.

CHAPTER 6

Boundaries

We continue our analysis by now considering the boundary pitch events at the end of TGs. We categorise boundary tones as two types: falling boundaries and rising boundaries.

6.1 Background

To describe accents we considered three points: a start point, a peak point and an end point. We can do this too with boundaries, but if we suspect the boundary to be a strict fall or a rise the peak position will be at the start of a falling accent and at the end of a rising accent and redundant. The peak position is also redundant as we intend to eventually model boundaries in the same way as simple falls or rises.

So before we start the formal analysis of boundaries, we take an informal look the peak position (as assigned by the tilt program) to confirm that ignoring it is acceptable. We first look at falling boundaries and we test to see if the peak of a falling boundary is indeed at the start of it.

Hypothesis 1: The peak of a falling boundary is at the start of it.

To test this hypothesis we simply compare the start time of the boundary with the peak time. Figure 6.1 shows a plot of the start time of falling boundaries plotted against the difference between the peak and the start time of the accent. If hypothesis 1 is true we would expect the difference between the peak and start times to be zero and give us something like a straight horizontal line. Start time

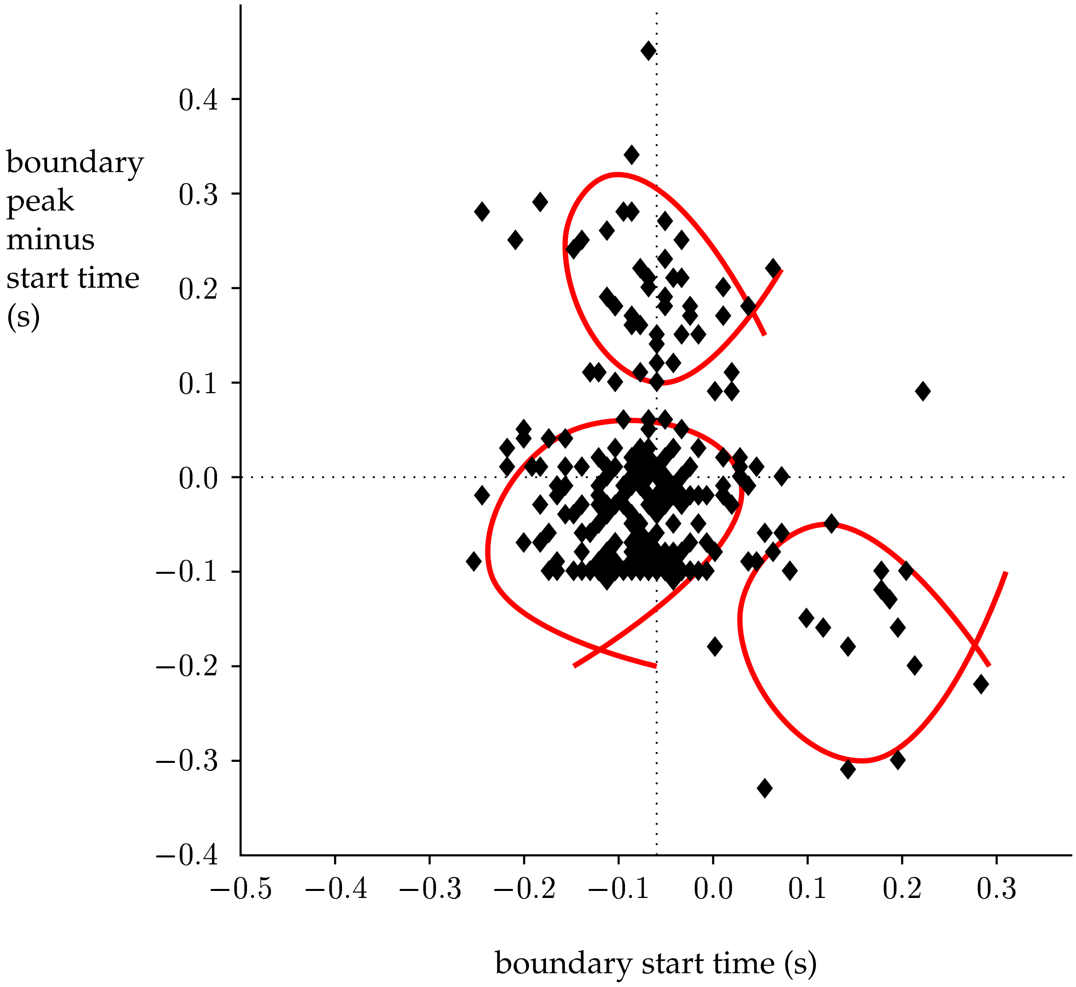


Figure 6.1: Scatter plot showing falling boundary start time plotted against time from boundary start to boundary peak.

plotted on the x-axis is really only used to space the points out and make the plots readable.

However, what we actually see are three groupings of points: a large grouping around the origin, a small grouping above the origin and a small group to the right of and below the origin.

The first oddity of this plot is that there are quite a few points with negative y values. A negative y value represents a boundary with a peak before its start. Peaks before starts come about because of the way in which the peak position is calculated. Recall from section 3.7 that the peak position is extracted from the tilt parameter representation that was calculated by the automatic tilt aligner. To interpret the graph we should consider all peak-time differences between -0.1 and 0 to be equivalent to zero (as the tilt aligner can move the start point by up to 0.1s). With this in mind we then see that the main group in the centre of the plot gives us our straight horizontal line.

The smaller grouping above it represents a small number of points where the peak is some way into the boundary. Within this group the points that have high peak-start times also have negative start times, suggesting that the boundary has been labelled as starting outside the voiced part of the syllable. The peak is placed by the tilt aligner after this point in the main part of the syllable.

The grouping to the bottom right is somewhat harder to account for. Here the boundary starts quite a long way into the syllable, but the peak is effectively where you would expect it to be near the start. This would appear to have been brought about by either a labelling inconsistency, or the tilt model choosing to represent a much larger section of f0 than the hand labels specify. Either way, the peak is generally at or before the start of the boundary, so these points are not treated as outliers as points in this quadrant of the graph were for accents. The general picture is that peaks do occur close enough to the start to accept hypothesis 1 and drop the peak position for falling boundaries.

We now consider rising boundaries. Rising boundaries are a more complicated class of boundary than falling boundaries, as they may manifest themselves as either a rise or as the lack of a fall, the latter resulting in a reasonably flat contour.

Our next hypothesis tests this assumption:

Hypothesis 2: Rising boundaries fit into two categories: a strictly rising category and a not-falling category.

Figure 6.2, a plot of f0 start position plotted against peak minus start position, shows us three distinct clusters of points like those we saw for rising boundaries. As with falling boundaries the start time can be adjusted by the automatic tilt labeller. Here the largest grouping is with accent start around 0 and accent peak minus start around 0.2. This represents the strict rising boundary, where the peak is clearly after the start of the accent. The two smaller clusters represent boundaries with the peak closer to the start, but from this plot there is no way of telling if they represent non-falling boundaries or otherwise.

An additional plot of f0 peak minus start time plotted against end minus peak time helps make the situation clearer. This is shown in figure 6.3. Here we see two clusters. The first is a cluster along the x-axis with end minus peak times

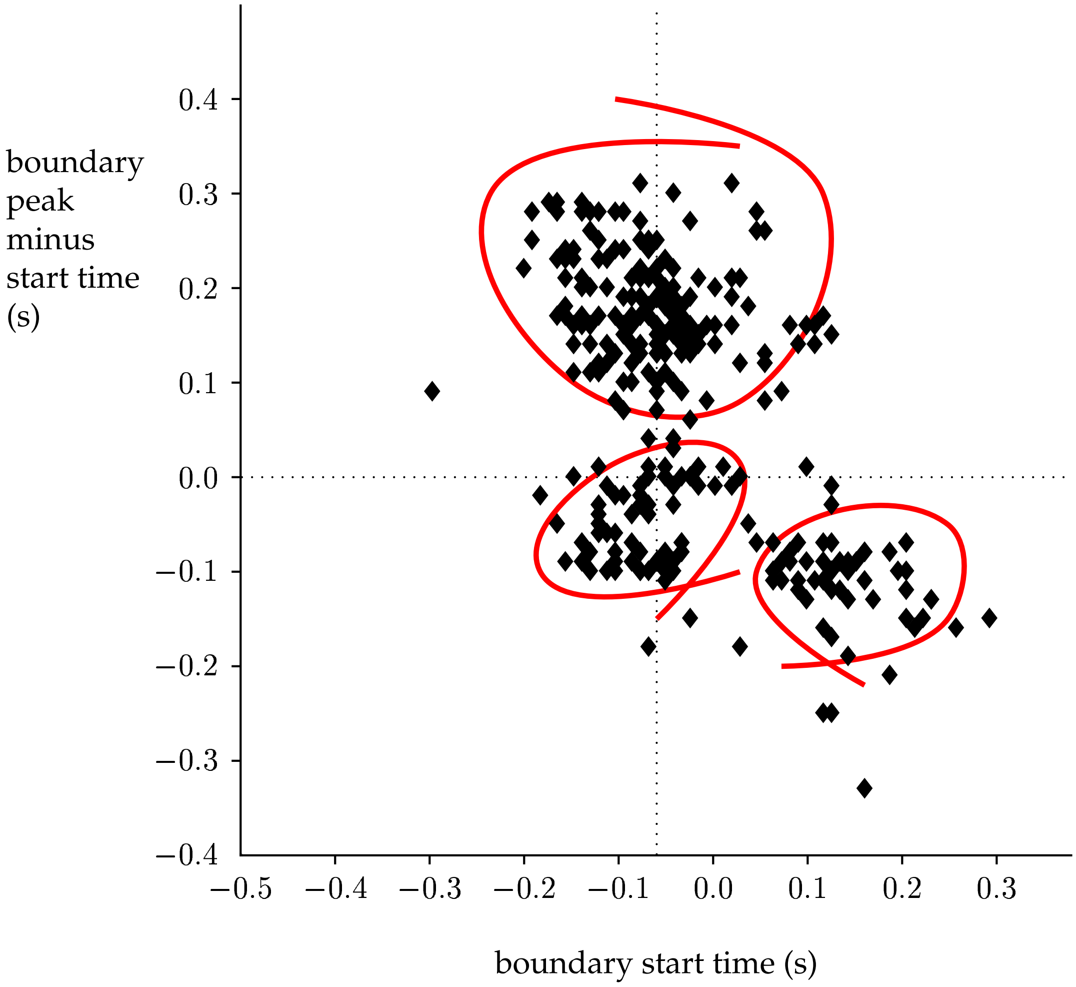


Figure 6.2: Scatter plot showing rising boundary start time plotted against time from boundary start to boundary peak.

close to. These are the group of strictly rising boundaries and match the top cluster in figure 6.2. A second cluster forms to the negative side of the y-axis just above the origin. This represents the other two clusters in the previous figure. The negative peak minus start time suggests the peak is in the beginning of the accent in a region added by the tilt aligner. This group is characteristic of the non-falling rising boundaries. However, as we shall see boundaries may occur in this cluster for other reasons.

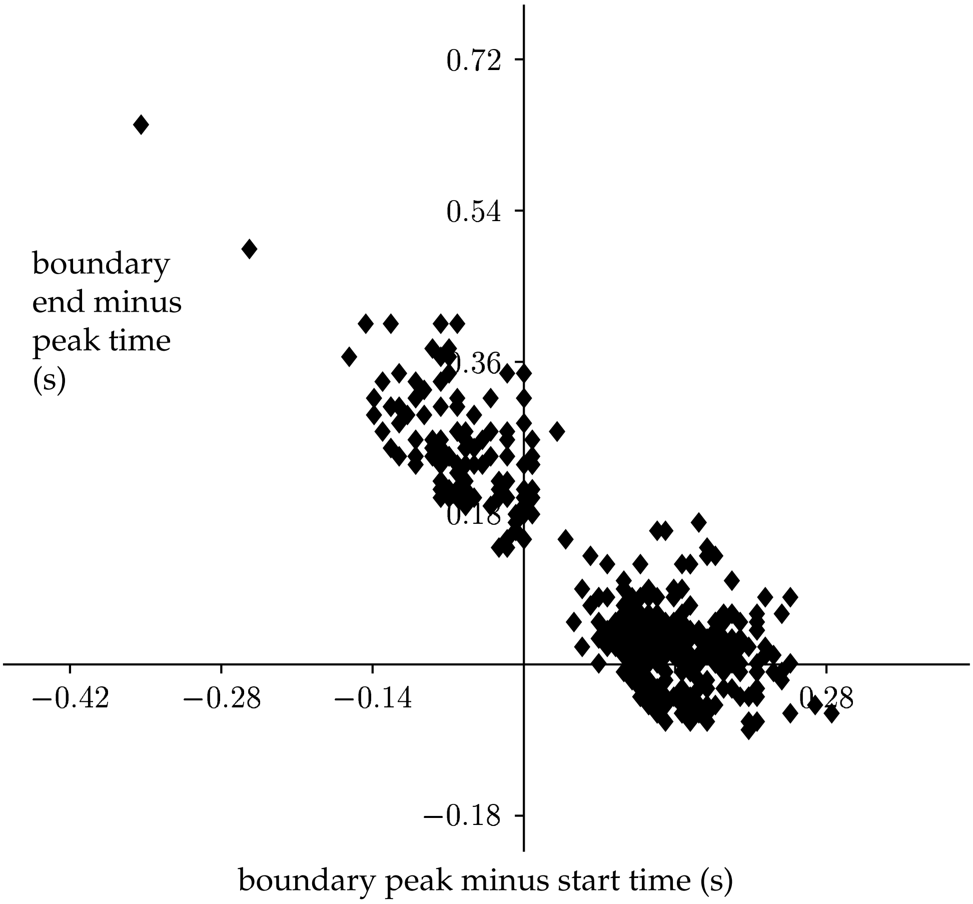


Figure 6.3: Scatter plot showing rising boundary start to peak time plotted against peak to end time.

Figures 6.4 and 6.5 show examples of rising boundaries. Figure 6.4 is an example from the lower cluster of figure 6.2, whereas figure 6.5 shows an example from the upper cluster of figure 6.2. Figure 6.4 shows nothing too surprising, a rising boundary with the peak somewhere near the end. The peak is not necessarily right at the end, which could be due to micro intonational effects or due to pitch tracking errors associated with the cessation of voicing at the end of the syllable. What we do see is that modelling this boundary with a simple rise would be reasonably appropriate.

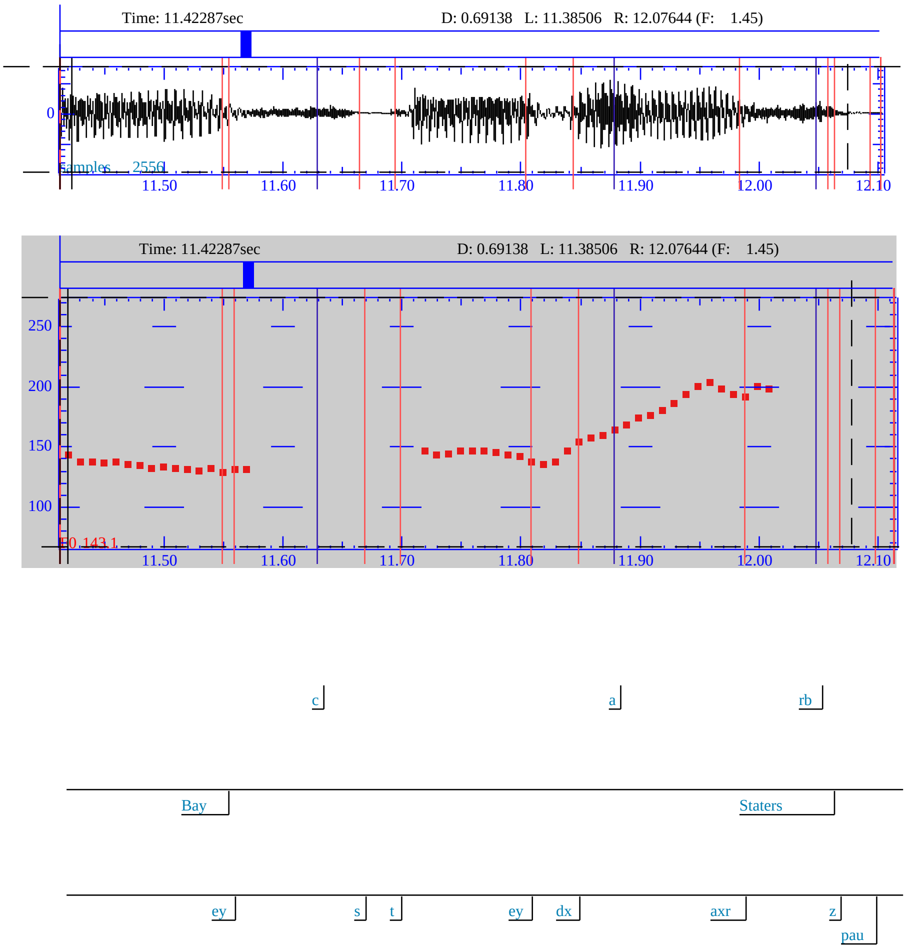


Figure 6.4: Figure showing a rising boundary with peak position located towards the end of the accent. Label tiers show (from the top) intonation labelling, words and segments.

Figure 6.5 is not as straightforward. Again, this shows what looks like a reasonably straightforward rising boundary, except for the fact that it occurs in the upper cluster of figure 6.2, which would indicate that the peak is close to the start of the boundary rather than the end. This has come about as a side effect of the tilt aligner moving the start of the boundary back into the previous accent, resulting in an overall fall, with the peak at the beginning. In this situation using the hand labelled start and end points to analyse the boundary produces much better results than using the tilt orientated data. It is also worth noting that over 50 of the rising boundaries in the database have tilt values between 0.2 and 1 which suggest an overall fall in pitch. This suggests that either the tilt aligner is not very good at modelling rising boundaries or that more than half of the rising boundaries actually fall.

To summarise, we see that the peak position in boundaries may be problematic and we are better to just use start and peak position alone.

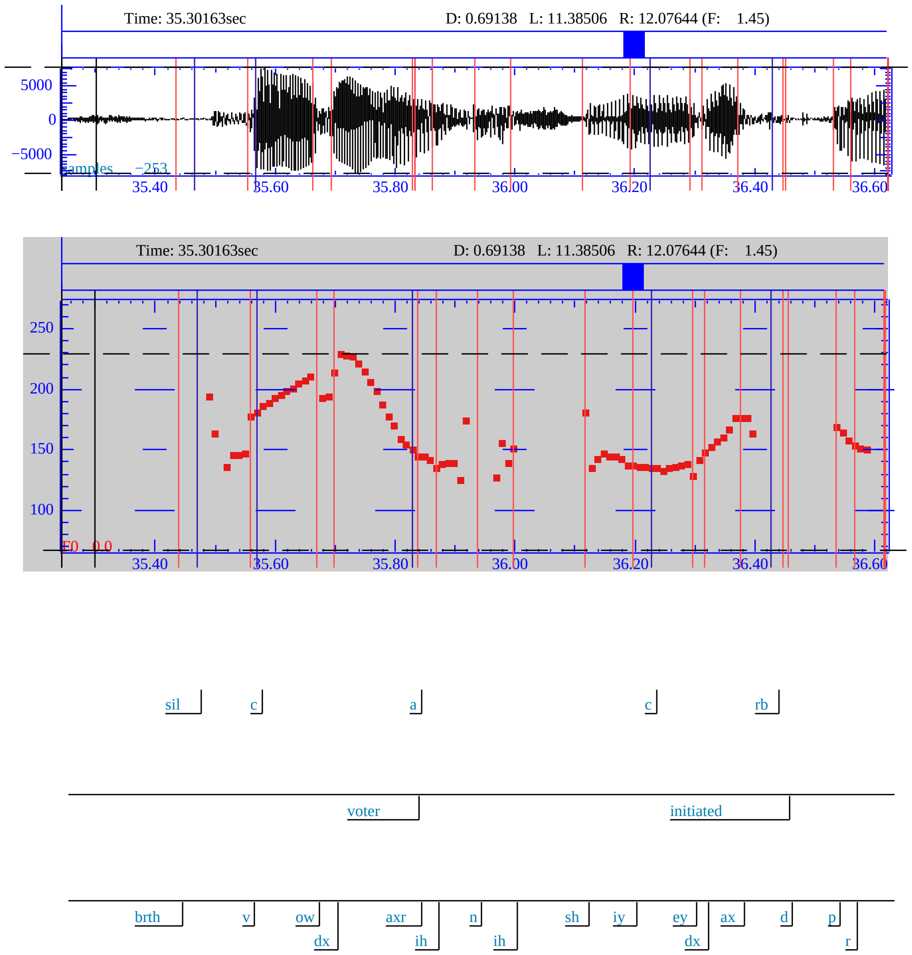


Figure 6.5: Figure showing a rising boundary with peak position located towards the begining of the accent. Label tiers show (from the top) intonation labelling, words and segments.

6.2 Falling Boundaries

We now begin to look at the pitch positioning of falling boundaries. We consider only the start and end points of boundaries as discussed in the previous section.

The overall pitch range results in chapter 5 were not too clear concerning what is happening at the end of each type of TG, but suggested that a falling boundary which is IP final should be lower than an IP internal one. So we test this hypothesis.

Hypothesis 3: The TG position affects the f0 positioning of falling boundaries. Specifically we expect to see the falling boundaries in final TG position being lower than those in medial or initial position.

The ANOVA results for the falling boundaries data are shown in tables 6.1 and 6.2. There is no significant effect for the starting-pitch position, and the significant effect for the ending-pitch position is a contrast between IP medial and final TGs.

So, although IP final falling boundaries end lower than non IP final ones they start at the same position. This allows us to accept a modified form of hypothesis 3, which states specifically that (the ending position of) falling boundaries in TGs in IP final position end lower. This is not particularly problematic as we would expect the position from which the boundary falls to depend more on the surrounding material than its TG position. For example an accent on the preceding syllable may have a drastic effect upon the boundary start position.

Finding no contrast for end position between initial and medial IPs provides evidence for the pitch range baseline remaining pretty constant except for final lowering.

	TG position	
	F(2,963)	p
start f0	1.89	.152
end f0	14.68	< 0.001

Table 6.1: Anova results for pitch position of falling boundaries

6.3 Rising Boundaries

There are two ways in which rising boundaries could behave. The point to which they rise could either be affected by the position of the TG in which they resides or it could be independent of the position in the TG. We take the former as our hypothesis and leave the latter as the null hypothesis.

Our hypothesis for how rising boundaries behave is then as follows:

Hypothesis 4: Rising boundary pitch position is dependent upon the position of its TG within an IP.

Anova results for these data are shown in tables 6.3 and 6.4. Here we see a clear effect for the end position, but also quite a strong effect for start position. The contrast results show that the end position is significantly different between all three categories IP initial, IP medial and IP final. This suggests that the pitch position of rising boundaries is affected by TG position and that the f0 points analysed in rising boundaries exhibit behaviour more like accents than of falling boundaries. This allows us to accept hypothesis 4.

The group means for both falling and rising boundaries are shown in figures 6.6 and 6.7. Along with the accents means shown in figure 5.8, these points would be sufficient to produce a simple rule based model for speech synthesis. We demonstrate this in chapter 7.3.

	start f0		end f0	
TG pos. contrasts	t	p	t	p
Initial/Medial	-0.46	.644	-0.22	.827
Medial/Final	-1.46	.144	4.90	.000

Table 6.2: Repeated contrasts for pitch position of falling boundaries

	TG position	
	F(2,963)	p
start f0	4.11	.017
end f0	16.04	.000

Table 6.3: ANOVA results for pitch position of rising boundaries

	start f0		end f0	
TG pos. contrasts	t	p	t	p
Initial/Medial	1.96	.050	3.38	.001
Medial/Final	1.30	.195	3.13	.002

Table 6.4: Repeated contrasts for pitch position of rising boundaries

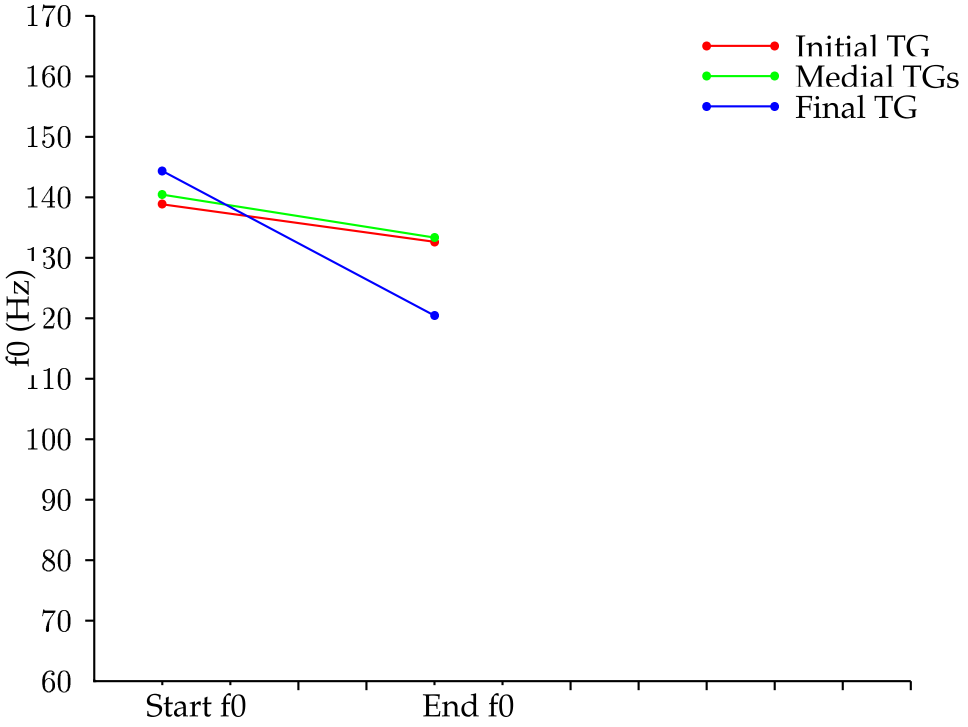


Figure 6.6: Summary of falling boundary group means, showing potential basis for a model.

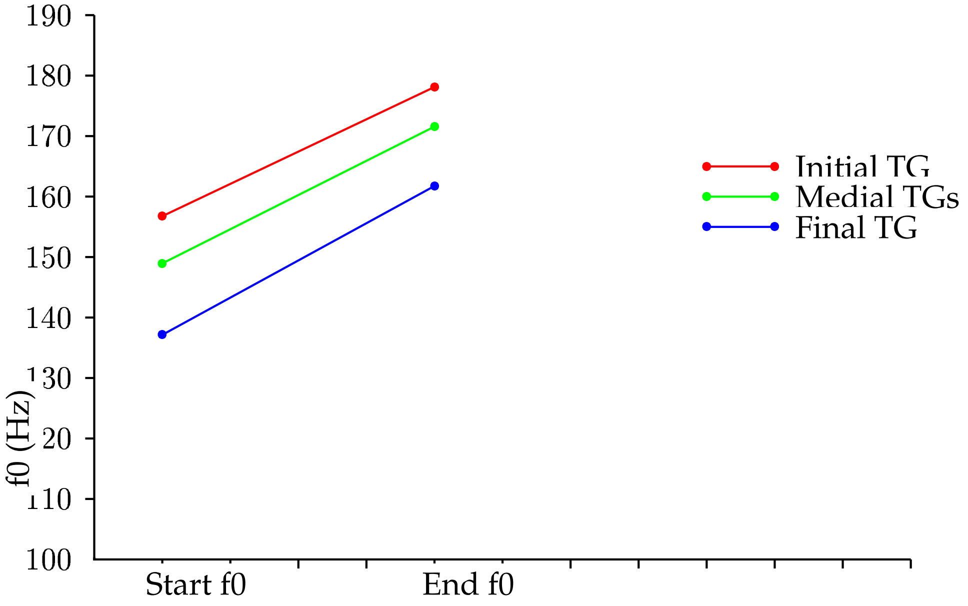


Figure 6.7: Summary of rising boundary group means, showing potential basis for a model.

6.4 Combined Accents and Boundaries

As well as basic rising and falling boundaries we need to consider pitch events where an accent and a boundary fall on the same syllable. These are labelled as a single entity in the data and are analysed as such.

We revert to using three points for analysis here as we expect these pitch events to incorporate an accent and thus to have a meaningful peak contained within them. As these pitch events are a combination of an accent followed closely by a boundary, we assume that there is some level of coarticulation and formulate our hypotheses accordingly.

Hypothesis 5: We expect the start and peak positions of a combined accent and boundary to follow that of an accent.

Hypothesis 6: We expect the end position of a combined accent and boundary to behave as if it were a boundary.

What we are suggesting then is that the front part, i.e. the start and peak of the accent and boundary combination behaves like an accent and the back part, namely the end, behaves like a boundary. In section 5.5 we saw that TG position was a main effect for start, peak and end position and contrasts were significant for all but start position in TG medial/final position. So for hypothesis 5 to be true we would expect to see these effects for the accent and boundary combinations.

In the previous sections in this chapter we saw that end position between the TG medial/final contrast was the only effect for falling boundaries and that end position was an effect for both TG initial/medial and medial/final contrasts for rising boundaries. This means that for our hypotheses to hold we would need to see effects for start and peak position to show accent-like behaviour and an end effect only between medial and final position to show boundary-like behaviour.

6.4.1 *Accent and falling boundary combinations*

The ANOVA results for combined accent and falling boundary combinations are shown in tables 6.5 and 6.6. Here we see that there is an effect for start and peak position but not for end position. Turning to the contrasts we see that it is only the TG IP medial versus TG IP final contrast which shows up as significant, at 5% for start position and at 1% for peak position.

This is not really sufficient evidence to accept either hypothesis 5 or 6 for accent and falling boundary combinations. The contrasts to accept hypothesis 5 are only partial and the end position effects that would allow us to accept hypothesis 6 are missing.

	TG position	
	F(2,242)	p
start f0	10.98	<0.001
peak f0	9.86	<0.001
end f0	2.75	0.065

Table 6.5: ANOVA results for pitch position of combined accent and falling boundaries.

	start f0		peak f0		end f0	
TG pos. contrasts	t	p	t	p	t	p
Initial/Medial	1.64	.101	0.24	.811	-0.62	.535
Medial/Final	2.53	.012	3.41	<.001	2.19	.030

Table 6.6: Repeated contrasts for pitch position of combined accent and falling boundaries.

6.4.2 *Accent and rising boundary combinations*

The anova results for combined accent and rising boundary combinations are shown in tables 6.7 and 6.8, and are somewhat uninspiring. Start position is the only effect, and only presents itself as a weak contrast between TG initial/medial position, not allowing us to accept either hypothesis 5 or 6 for accent and rising boundary combinations.

	TG position	
	F(2,46)	p
start f0	5.78	0.006
peak f0	1.01	0.373
end f0	1.34	0.271

Table 6.7: ANOVA results for pitch position of combined accent and rising boundaries.

6.4.3 Conclusions

There could be a number of reasons for poor results with combined accent and boundaries. Firstly there is limited data for these pitch events, particularly in the case of combined accent and rising boundary where there are only a total of 49 available for use in the analysis. This unfortunately prevents trying to split the data into further caregories.

Secondly, as the number of data points for these combined accent and boundary events is low we have not included the factor *accent number* which represents the number of preceding accents. This played a part in accent position, although *TG position* was a clear main effect in its own right, so an effect would still be expected here for TG position without accent number being used. There is insufficent data to carry out a full analysis including the number of accents in the sequence which ends in the combined accent boundary. But a partial analysis with the falling boundaries showed results no different to those without this factor included.

The outcome is that these results do not really show that these combined pitch events are a straightforward combination of an accent followed by a boundary. However, for modelling purposes there is no reason why they cannot be modelled as a class of their own. With this in mind, the mean values for these pitch events are shown in figures 6.8 and 6.9.

	start f0		peak f0		end f0	
TG pos. contrasts	t	p	t	p	t	p
Initial/Medial	2.15	.037	0.16	.877	0.36	.724
Medial/Final	1.49	.143	1.21	.230	1.30	.199

Table 6.8: Repeated contrasts for pitch position of combined accent and rising boundaries.

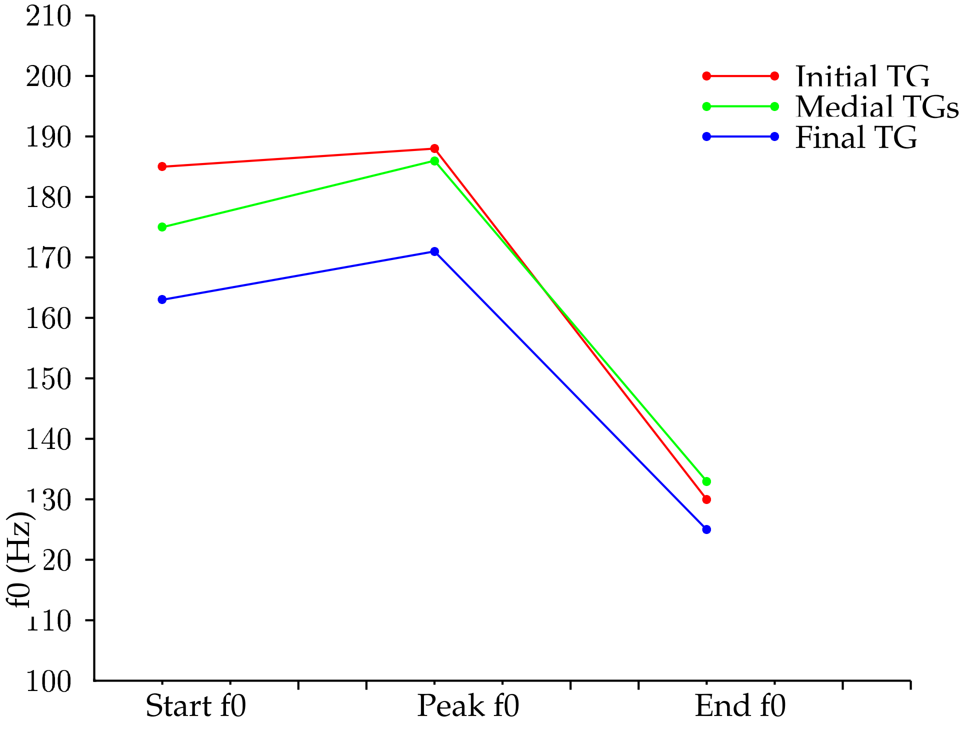


Figure 6.8: Summary of combined accent and falling boundary group means, showing potential basis for a model.

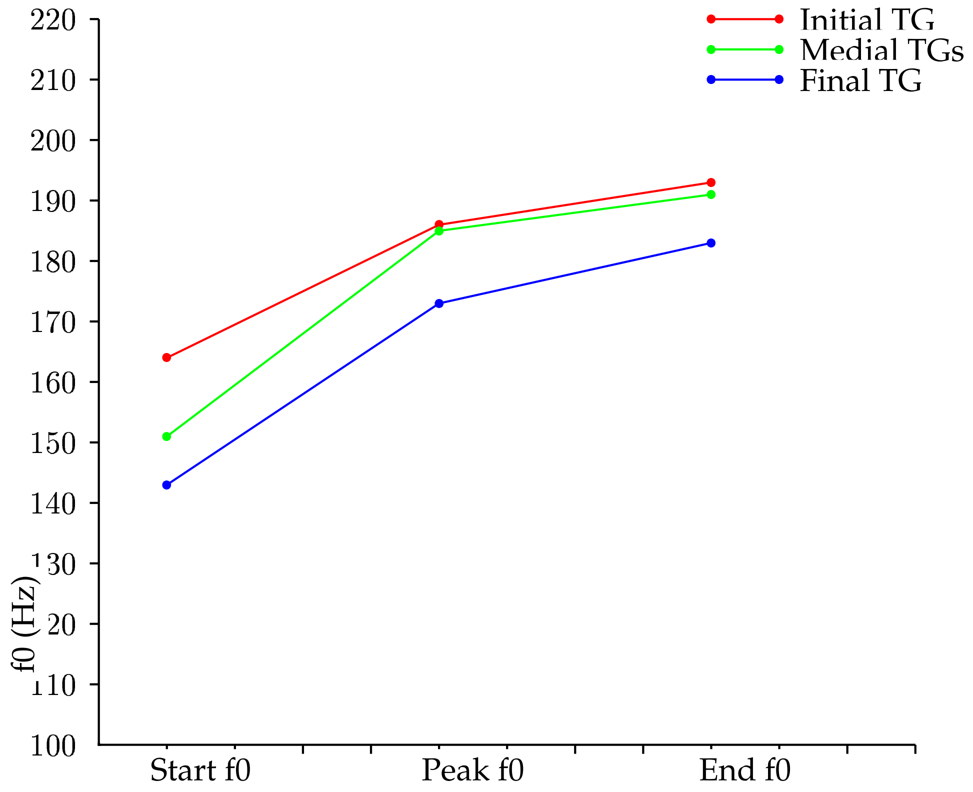


Figure 6.9: Summary of combined accent rising boundary group means, showing potential basis for a model.

CHAPTER 7

From Data to Model

This chapter deals with the issue of moving from data analysis to an intonation generating model. So far the analysis carried out in chapters 5 and 6 has collated a large amount of statistical information about f0 position for specific prosodic contexts within the data. We now turn the results of this analysis into a means of modelling intonation for speech synthesis. We first develop a framework using prosodic structure and we then demonstrate two models built within this framework.

7.1 Introduction

Our first aim here is to produce a *framework* within which intonation models can be built. The point of the framework is to allow different types of model, for example the rule based models or statistically trained models discussed in section 2.11, to be constructed with the same structure. The framework can be thought of as a set of theoretic constraints within which a model is built. These constraints may be represented by parameters but this is not crucial.

To clarify the distinction between model and framework, we can consider ToBI as an alternative framework. ToBI is not really an intonation model for speech synthesis, it is an intonation description system. It does not tell us what a given f0 contour looks like, it allows us to classify given intonation contours. However, ToBI can be used as a framework in which to build an intonation model for speech synthesis. The ToBI framework would specify that pitch events are to be modelled in sequences to correspond to those allowed by the description system and would constrain the assignment of phrase breaks to correspond to break indices in an appropriate manner.

The specifics of how a model is then constructed within this framework are flexible and left to anyone choosing to use it. For example Festival currently contains two intonation models built within the ToBI framework. The first is the statistical CART tree model, discussed in section 2.12.1, which predicts ToBI accents and boundaries, and then statistically generates an f0 contour from these. The second is the rule based model of Jilka, Möhler & Dogil (1997), which constructs f0 target points by rule from assigned ToBI labels[1]. These models are very different in nature but both adopt the ToBI ideas as their basis.

7.2 The Framework

The framework being proposed here is designed to constrain pitch event sequences in a similar way to which a framework built around ToBI would. However, the details we wish to concentrate on with this framework are not those that ToBI is concerned with. Where ToBI focuses on sequences of pitch event types and the relationship between neighbouring pitch events, the framework we are developing here focuses on the relationship between sequences of pitch events rather than the events themselves. This reflects the different goals of a description system, which needs to show fine differences in detail, and a generation system which needs to produce broadly acceptable intonation. The details within accent sequences are not ignored but are just considered less important—and are therefore not explicitly part of the framework. The primary difference then between the framework proposed here and a ToBI based framework is that this framework specifies prosodic structure at the phrase level, whereas ToBI primarily concentrates on specifying structure within the phrase.

To an extent the internal details of accent sequences in this framework are left to be general and flexible enough to incorporate existing frameworks such as ToBI or tilt within our framework. That is, the constraints laid down by the framework being developed here are designed to not contradict any of those which define the ToBI or similar frameworks. As they operate at different levels of the prosodic structure, constraints placed by the use of ToBI would not conflict with our constraints and hence a ToBI or similar framework could be incorporated within the

[1] This model is not discussed here because it is not in general use and is not capable of generating contours in all situations. It has particular problems when accents and boundaries occur close together.

framework proposed here and we could use this combined framework to build an actual model for generation. This approach is taken in chapter 9, where a linear regression model is built within our framework using ToBI.

The framework proposed here concentrates on providing prosodic structure to place sequences of accents within.

Using the utterance structure developed in previous chapters, we deem an utterance to consist of one or more IP phrase type units. Each IP consists of one or more TG phrase type units. The results of chapter 4 indicate that the relative position of the IP within the utterance is not regarded as important, whereas the relative position of the TG within an IP is. Following this, we use a three way initial-medial-final distinction rather than an n-way distinction. This structure is illustrated by figure 7.1. Each TG then consists of a sequence of accents followed by a boundary. This is illustrated in figure 7.2.

The framework can be regarded as a series of finite state networks. These are shown in figure 7.3. Here the top diagram shows that an utterance is just a sequence of one or more IPs. The middle diagram shows that an IP is a sequence of TGs. The status of the first and last TG are considered distinct from any medial TGs. If there is only one TG in an IP then it is treated as a final TG. This is so that any final lowering properties modelled by this TG are manifested. A separate class of TG representing IPs consisting of only one TG was not used to simplify

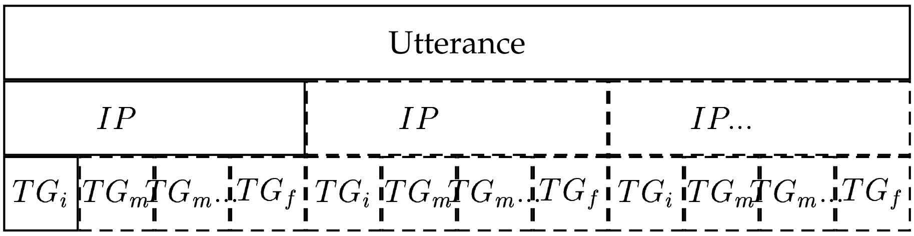


Figure 7.1: Utterance structure adopted by the framework.

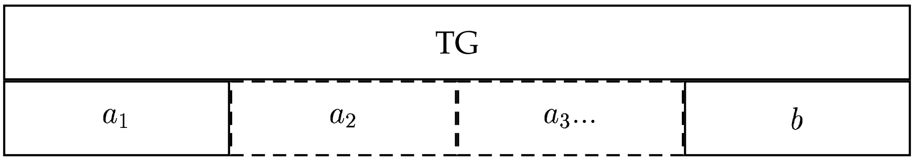


Figure 7.2: TG structure adopted by the framework

the model because the analysis that our distinctions are based upon was not sufficient to provide us with this distinction. Initial TG properties will be lost for IPs of one TG and in circumstances where it is considered crucial to accommodate this, for example in a system which generates many very short utterances, a separate TG category could be added.

The lower diagram shows the internal structure of a TG. A TG consists of a sequence of at least one accent followed by a boundary. Where there are more than four accents in a sequence, the third one is duplicated as our analysis showed that after the third accent in a sequence the variation between subsequent non-final accents was small and mostly insignificant.

In the case where there is only one accent in a TG, data representing the first accent in a sequence is used. This is chosen because the analysis showed the first accent in TGs was raised significantly in pitch compared to the other accents in the TG. This effect may be related to the broadcast news domain that the data is taken from and may not be universal. In a situation where this raising is not expected to occur or is just not wanted, it would be preferable to adapt the framework to use the last accent from the sequence so that the initial high position properties are lost.

If the assumption that the final accent in a sequence is the nuclear accent is being made, it may be more appropriate for the final accent to be the one which is used in a single accent sequence. We do not make this assumption here. Lone accents could alternatively be treated as a special case.

The three networks shown in figure 7.3 form a recursive transition network representing the utterance as a whole. However, the network is really only compositional as there are only calls to the TG network from the IP network and only calls to this from the U network. So, the whole system can be thought of as an abreviated description of a single large finite state network, where for example, there is a seperate TG network in the place of each of the TG_i, TG_m and TG_f nodes in the IP network.

Using the framework as described so far, each pitch event for which we need to generate an f0 contour can be parameterised by three variables: *TG type*, *Event type* and *Accent number* as follows:

TG type: TG_i, TG_m or TG_f

Pitch event type: $a_1, a_2, a_3, fb, rb, afb$ or arb
Accent number $1..n$ (accents only)

We refer to the parameters associated with a given accent which describe its prosodic position within the utterance as its *context*, opposed to those which describe the accent itself, which we refer to as its *description*. Description would include f0 position and time alignment with segmental structure, context would include TG type.

The principle behind the modelling techniques being developed here is that a pitch event should be modelled in its context and that context is provided by the framework. The framework can now be thought of as providing the set of context parameters specified above. A model is built within this framework by modelling pitch events which have the same parameter values together.

There are a couple of extra inclusions which could be made to the framework, which are not used here as the data analysed was considered insufficient to carry out the analysis appropriately to provide such a framework.

Pitch event context could be made more specific by including an extra parameter *Accent count* which holds a count of the total number of accents in the TG. This would then cause accents in different length sequences to be modelled independently. As this increases the number of contexts significantly, which would in turn increase the complexity of any model based on the framework, it is not used here.

A further reason for not using such a finer distinction between pitch events is that the accent context may need to include *Accent type*. By accent type we mean a categorisation which determines accent shape. Although accent type is generally considered to describe the accent rather than its context, if tune is considered to limit accent type choice this also makes it part of the context.

It is not always necessary to model accent type at all, as a model built within the framework may implicitly account for accent type, like Tilt does with the tilt parameter (see section 2.7). Most accent description systems however do make a distinction between accent types, for example H* versus L+H* in ToBI.

As the need for an accent type specification is model dependent, it is not included within the framework, but is suggested as an extension to the framework to be used where appropriate. If the ToBI pitch events, for example, were being used by a model, the ToBI framework would be incorporated within this framework using the accent type parameter.

One of the ramifications of providing a detailed context within which to model pitch events is that the framework also implicitly incorporates the modelling of declination and final lowering. This is because their effects on pitch events at different positions in the utterance are captured by the use of context. Thus any declination or final lowering present in data that a model is based upon would be retained with the framework. For example, if third accents in sequences are lower than their preceding neighbour because of declination, this will be captured by the use of context and seen in resulting synthesis.

We now consider the implementation of two models using the framework described so far in this chapter. The first is a simple rule based model and the second is a more complex statistical model.

7.3 Simple Hat Model

This section describes a Festival implementation of a very simple intonation model built within the framework described in section 7.2 This model extends the idea of a model which places simple 'hats' on accented syllables. Each accent is represented by three pitch targets, at the beginning, middle and end of the accented syllable. An f0 contour is then generated in a "join-the-dots" fashion. This generates an intonation pattern like that shown in figure 7.4

Festival incorporates a simple hat model as it is easy to implement and easy to use when no other model is available, for example when developing a synthesiser for a new language.

The intention here is to produce a better model from it by using the framework described in section 7.2. This serves well to demonstrate our framework and if successful will provide the groundwork and a baseline for building other models.

A simple hat model was implemented using the described framework. The pitch target points for each pitch event instead of being single fixed predetermined values are predetermined values based on the context provided by the framework. The specific values the model assigns are based on the mean values found by the analysis in earlier chapters. Our intention is to show that the variability of the target values introduced by the context turns the simple hat model into a viable model for speech synthesis.

7.3.1 F0 normalisation

At this point an issue of normalisation arises. We have mean start, peak and end position f0 values for all accent contexts prescribed by the framework. These were produced as a byproduct of the ANOVA tests carried out as part of the analysis. However, these values are specific to f2b. As there is not currently a Festival voice of this speaker, these values are not appropriate. Furthermore, as all of the standard Festival voices are male speakers and f2b is female the values cannot be used at all without a suitable transform to rescale to an appropriate male pitch range.

Therefore, to build a model from these values we have two options available to rectify the pitch range problem: we can either apply an arbitrary scaling to all the f2b values to move them into our speaker pitch range or we can define a contextual normalisation process which provides us with a set of normalised parameters which we can scale to any speaker they are being applied to. The normalised parameters will then need suitable scaling factors to map to a new speaker, which could be calculated by either a little analysis of that speakers pitch range, or by estimation. We choose the normalisation process as this allows the data to be easily adapted to any given voice, and provides a process for adapting other data. We realise that the actual differences between a male and female voice are much more complex than the simple pitch range scaling suggested here, but the deficiencies in the resulting quaility of diphone synthesis make it not worthwhile to do anything more complicated.

For our model then, we normalise each target point of a pitch event with respect to the TG component of the target. That is, we normalise target points relating to pitch events in initial TGs with respect to the measured mean and standard deviation of the f0 measured throughout all initial TGs in the database and likewise for other contexts.

The normalisation process is a simple calculation of a Z-score for each TG type. The resulting Z-scores are then used by reversing the process using a mean and standard deviation related to the pitch range of the voice being used rather than that found by the analysis. The equations used for this process are 7.1 and 7.2.

Equation 7.1 gives the transformation from pitch values within the database to the normalised values which are part of the model The mean value of particular target points in a TG context ($\mu_{t,tg}$) is converted into a z-score (z) using the database mean ($\mu_{D,tg}$) for the overall pitch range of the specific TG type and its related standard deviation ($\sigma_{D,tg}$). This normalises a given target point with respect to the pitch range relating to the type of TG that it is in, opposed to normalising with respect to the pitch range of the speaker as a whole which would be more variable.

The normalised target values for f2b are shown in table 7.1.

	Accent number	Start target	Peak target	End target
Initial TG				
	1	0.47	1.40	0.42
	2	-0.18	0.52	-0.24
	3	-0.48	0.23	-0.53
	4	0.02	0.43	-0.66
	arb	-0.59	0.04	0.20
	afb	-0.15	-0.06	-1.74
	rb	-0.86		-0.20
	fb	-1.19		-1.34
Medial TG				
	1	0.23	1.26	0.32
	2	0.15	1.02	-0.02
	3	-0.16	0.62	-0.26
	4	-0.15	0.83	0.20
	arb	-0.68	1.13	1.56
	afb	0.20	0.84	-1.24
	rb	-0.61		0.32
	fb	-0.80		-1.15
Final TG				
	1	0.37	1.24	0.24
	2	0.22	1.03	0.10
	3	0.05	0.74	-0.32
	4	0.04	0.92	0.03
	arb	-0.40	1.05	1.76
	afb	0.36	0.80	-1.54
	rb	-0.68		0.56
	fb	-0.53		-1.60

Table 7.1: Normalised target point frequencies for pitch events in each prosodic context.

Equation 7.2 then takes the normalised target value and scales it to the new speaker's pitch range for a given TG context using the appropriate mean and standard deviation ($\mu_{V,tg}$ and $\sigma_{V,tg}$), this produces a target point (t') which is used for synthesis.

$$z = \frac{\mu_{t,tg} - \mu_{D,tg}}{\sigma_{D,tg}}, \tag{7.1}$$

$$t' = z\sigma_{V,tg} + \mu_{V,tg} \tag{7.2}$$

To use normalised target scores we need to know means and standard deviations calculated over the set of TG contexts for the target voice ($\mu_{V,tg}$ and $\sigma_{V,tg}$ in equation 7.2. Unless there is sufficient data recorded from the speaker of the diphones to calculate this information from, these parameters will have to be estimated, and fine-tuned to produce an acceptable pitch range. This fine-tuning was done by producing a series of examples with different means and standard deviations and choosing the most natural sounding.

There are other, possibly better, ways to model and 'normalise' pitch range. Patterson's (2000) approach seems particularly promising. This approach is not used here as the tools to use such an approach automatically within a speech synthesis system are currently not available.

7.3.2 The model implementation

This simple model has been implemented for the *ked_diphone* voice for Festival. The model works on an utterance which has been hand labelled for accents and boundaries at the syllable level. This is done to reflect the input expected by a language system, as this is our target application. The voice specific mean and standard deviations required to produce actual target values from the normalised target values were estimated by starting with values suitable for a male speaker and then repeatedly adjusting them based on the results of synthesis. The selected values are shown in table 7.2. The mean for initial TGs was selected to be slightly higher than that of subsequent TGs to reflect that found by the original analysis of f2b.

Resulting synthesis shows that the effects of declination and final lowering can be modelled implicitly as expected and are clearly visible in the example shown in figure 7.5. The differences in the height of accents between the different TG contexts and within sequences in the same context are reflected by the resulting target points.

7.3.3 Conclusions

This model, even though simple, produces reasonable results especially when considering that there is no accent timing being taken into account and all accents are placed over complete syllables. Even a naïve listener can tell that this model is an improvement on the original simple hat model which has no variation between accents. As there is little extra cost involved to produce this model in place of the simple hat model, this suggests that the use of prosodic structure is worthwhile.

This model is still an order of magnitude worse than the standard statistical techniques used for text-to-speech (this can be seen from the evaluation results presented in section 8.3). This is not surprising considering that the standard statistical techniques account for many aspects of speech that this simple model does not. This model however, makes a good intonation model for testing a new language or dialect before a better model can be built.

The challenge is to show that the addition of prosodic structure can also be used to improve upon the standard statistical models to provide the flexibility required for intonation models where the intonation is dictated externally rather than by the synthesiser's analysis of the input text.

The simple hat model will be evaluated in the next chapter, but we first build an accompanying statistical model.

TG	$\mu_{V,TG}$	$\sigma_{V,TG}$
i	110	15
m	100	15
f	100	15

Table 7.2: Voice parameters for festival's ked_diphone voice.

7.4 A Statistical Model

In this section we describe the building of a statistical LR model for Festival using the framework described in section 7.2. As the f2b speaker is American, we build a new intonation model for one of Festival's existing American English voices.

7.4.1 Background

Both CART and LR models (see section 2.11) can be used to predict the f0 target values for a given syllable, although in practice LR models are prefered as they tend to produce better results in this situation. Like the hat models described in section 7.3 they predict an f0 value at three points for each syllable. Unlike the hat models they predict f0 values for all syllables rather than just the accented ones. This means that the peak of an accent is not modelled explicitly like it was in the hat models, where it was forced to be in the middle of the syllable. If for example, a peak falls at the end of a syllable then the f0 at the end of the syllable could well be the highest f0 value with a decline into the next syllable. Figure 7.6 shows a stylised example of what type of f0 variation these kinds of model can produce. Here, *accent 1* has a central peak just like accents generated by the hat models. *Accent 2* however, has a slight dip preceded by a late peak. This configuration of f0 points is more complex than what could be generated by the hat models.

The framework developed earlier in this chapter can be incorporated directly into a LR model, by including features such as 'TG type' and 'accent count' both for the current syllable and the surrounding syllables, so that a component which represents the variation between the different contexts specified by the framework is incorporated into the resulting model.

7.4.2 Evaluation measures

Results for LR models are usually reported using the *root mean square error* (RMSE) and *Correlation* (R^2). The original pitch contour from real speech is compared to the pitch contour resulting from resynthesis using the same text and other linguistic information. RMSE scores a point by point average f0 error for a target contour and correlation shows how well the contour's variation follows a target contour. The use of these scores for intonation is somewhat ungrounded as they do not correlate well with perceptual ratings of the difference between contours, which means that two contours which are perceived as conveying the same meaning may have a large RMSE score. See Clark & Dusterhoff (1999) for further discussion on this. However, in the absence of a robust well proven alternative we resort to reporting RMSE and R^2 values here.

We compare two new models to the default linear regression model which Festival uses. Our aim is to produce a model which is more flexible than the default Festival model and hence more suitable for use with language systems. We have reduced the number of labels identifying accents, but increased the richness of the explicit prosodic structure, with the intention that the richer prosodic structure allows us (or the language system) to be more prescriptive about where accents occur. Lists of specific features used in both the default model and the model trained here are given in Appendix C.

Model 1 is a model which is trained on the set of features that the default Festival model is trained on except that the features relating to accent/boundary type have been changed to accommodate the simplified pitch event labelling scheme being used here, and in addition features are added to represent the TG context defined by our framework. This model is later refered to as the Context LR model when compared with other models.

Model 2 is an an attempt to use the framework much more explicitly. Here each position in the framework is represented by a feature, so there is a feature for "second accent in a IP-medial TG". Features for previous and next syllable are included here but as this already results in 200 features, features for two syllables away from an accented syllable are not included. Also this model is not trained using a stepwise procedure as most of the features are mutually exclusive from each other. See section 9.2 for discussion of the consequences of mutually exclusive features.

Model	syllable start		syllable mid		syllable end	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Default Festival	27.88	0.25	28.29	0.33	27.14	0.29
Model 1	32.73	0.22	31.39	0.36	30.19	0.33
Model 2	37.48	0.21	31.42	0.38	-	-

Table 7.3: RMSE and R^2 comparisons of various linear regression intonation models

Results for these models are shown in table 7.3. Model 1 is shown to be comparable to the default Festival model. Its RMSE values are slightly worse than the default Festival model, but its correlation values are slightly better. This is the result of trading accent specifications for prosodic structure. If good synthesis can be achieved with this model then we have achieved the flexibility we require.

Results for model 2 proved much worse, suggesting that there was insufficient data to train the large number of parameters in that model. This model is abandoned at this point.

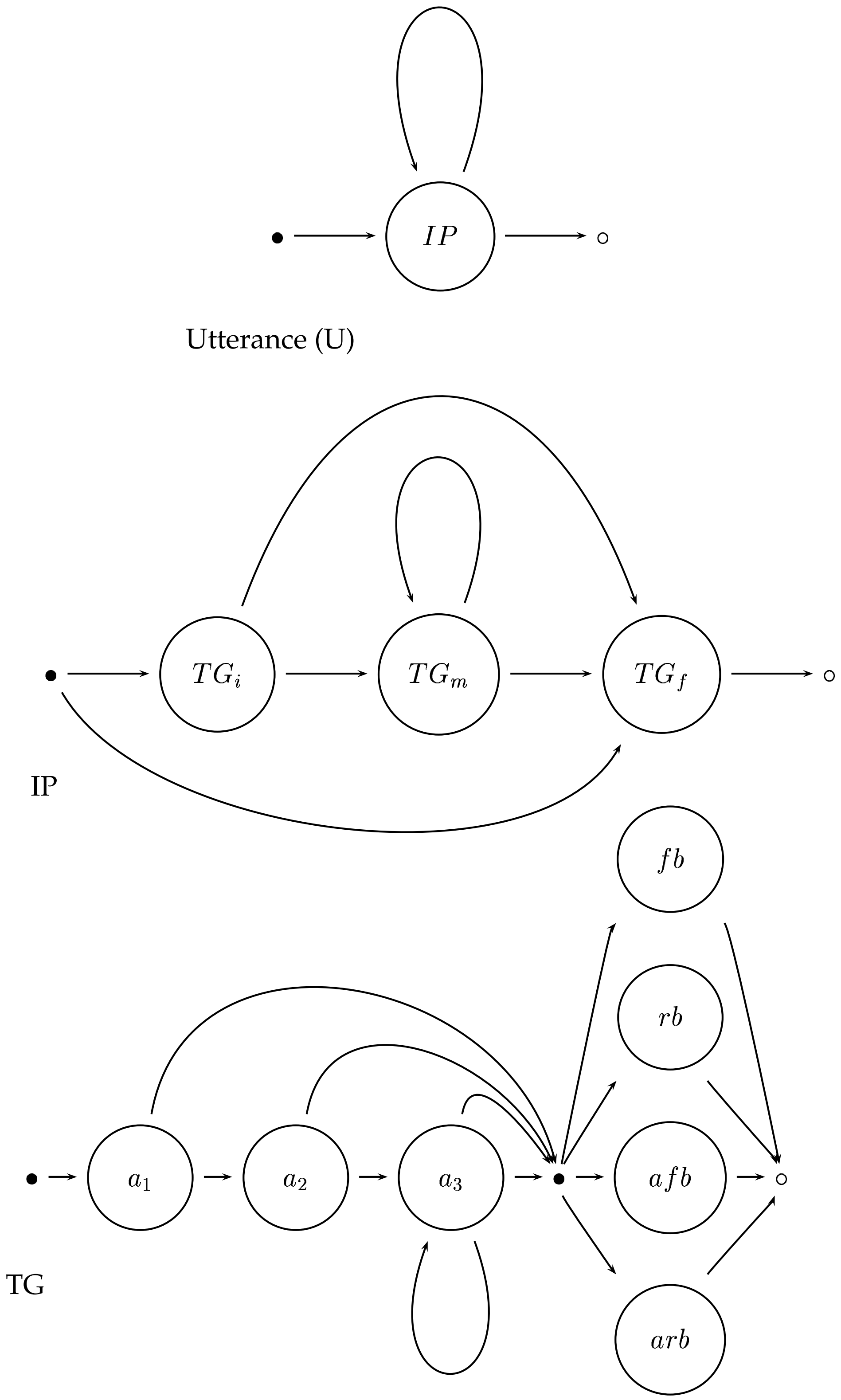


Figure 7.3: Finite state networks representing the prosodic framework. The top model represents the overall utterance structure, the middle model the IP structure and the bottom model the TG structure.

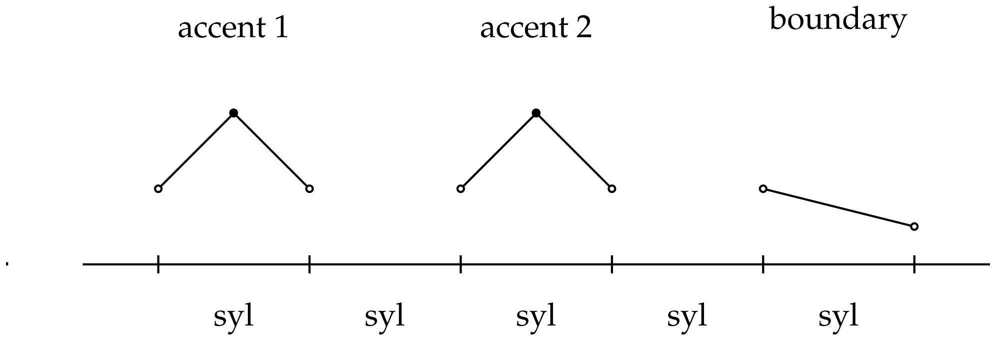


Figure 7.4: Example of the intonation pattern generated by a simple hat model.

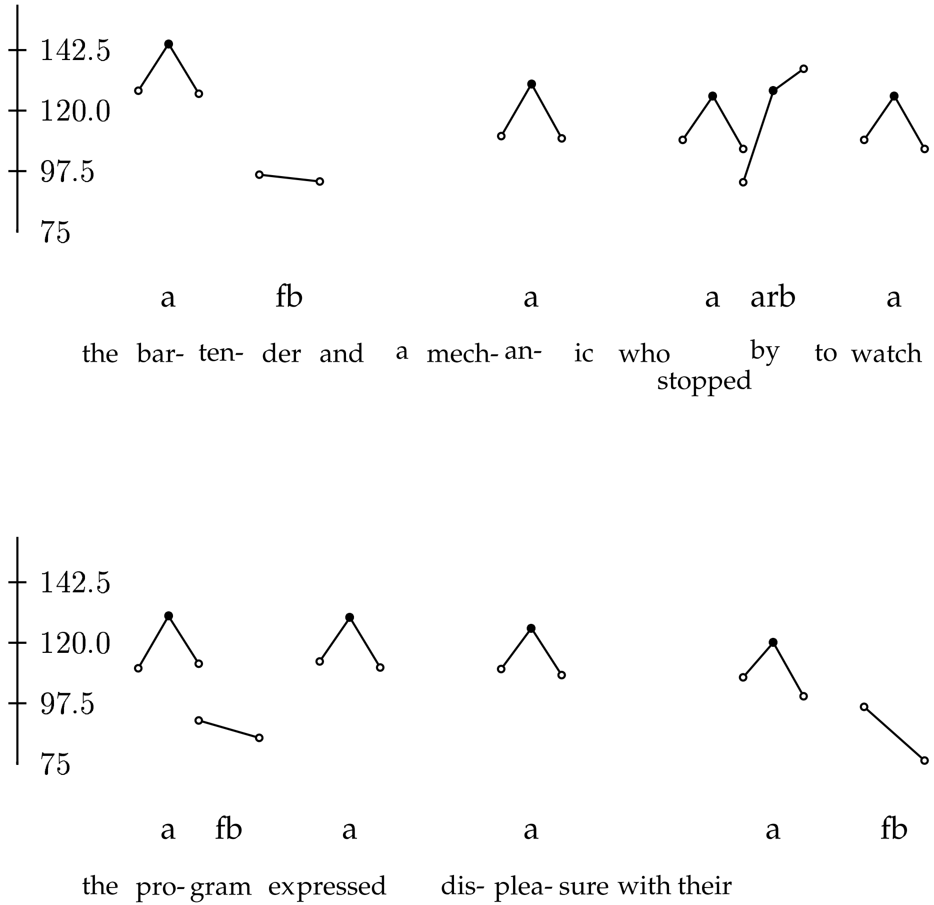


Figure 7.5: Example intonation contour for the utterance "The bartender and a mechanic who stopped by to watch the programme expressed their displeasure with their congressman."

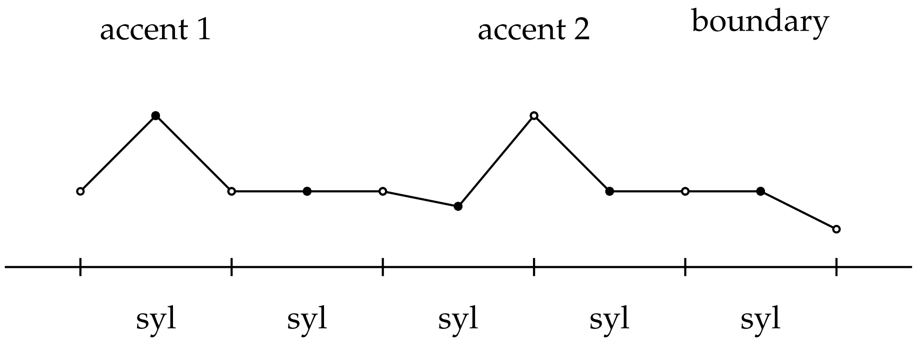


Figure 7.6: Example of a stylised intonation pattern generated by a statistical model.

CHAPTER 8

Evaluation

The aim of this chapter is to further evaluate the models described in the previous chapter. As well as taking a closer look at the output generated by these models we will verify that the synthetic speech produced by the models is acceptable to human listeners and an improvement on currently available intonation models.

One of our main goals was to try to capture the quality of the broadcast news style of speech, and we shall evaluate perceptually if we have been able to achieve this. We shall also consider if our models are suitable for more general use by looking at speech from other domains.

With other domains in mind, our hypothesis is that a large amount of what makes broadcast news sound like broadcast news is related to the placement of accents rather than the shape of the accents themselves. Broadcast news contains a lot of accents and they often occur in unusual places. As we are de-coupling the contour generation from the accent prediction, we should be able to generate speech which sounds less like read news by placing accents more appropriately.

8.1 The Broadcast News Domain

We first look at an example utterance from the data set that was used to train the models. This particular utterance is one that was removed from the training data (omission usually results from a missing label file). Figure 8.1 shows the natural speech and pitch contour. The most notable feature is the wide variation in f0 from a resting level of around 150Hz to peaks of around 280Hz, with very little evidence of declination across the utterance as a whole. There are however quite large dips to around 100Hz at the L-L% boundaries.

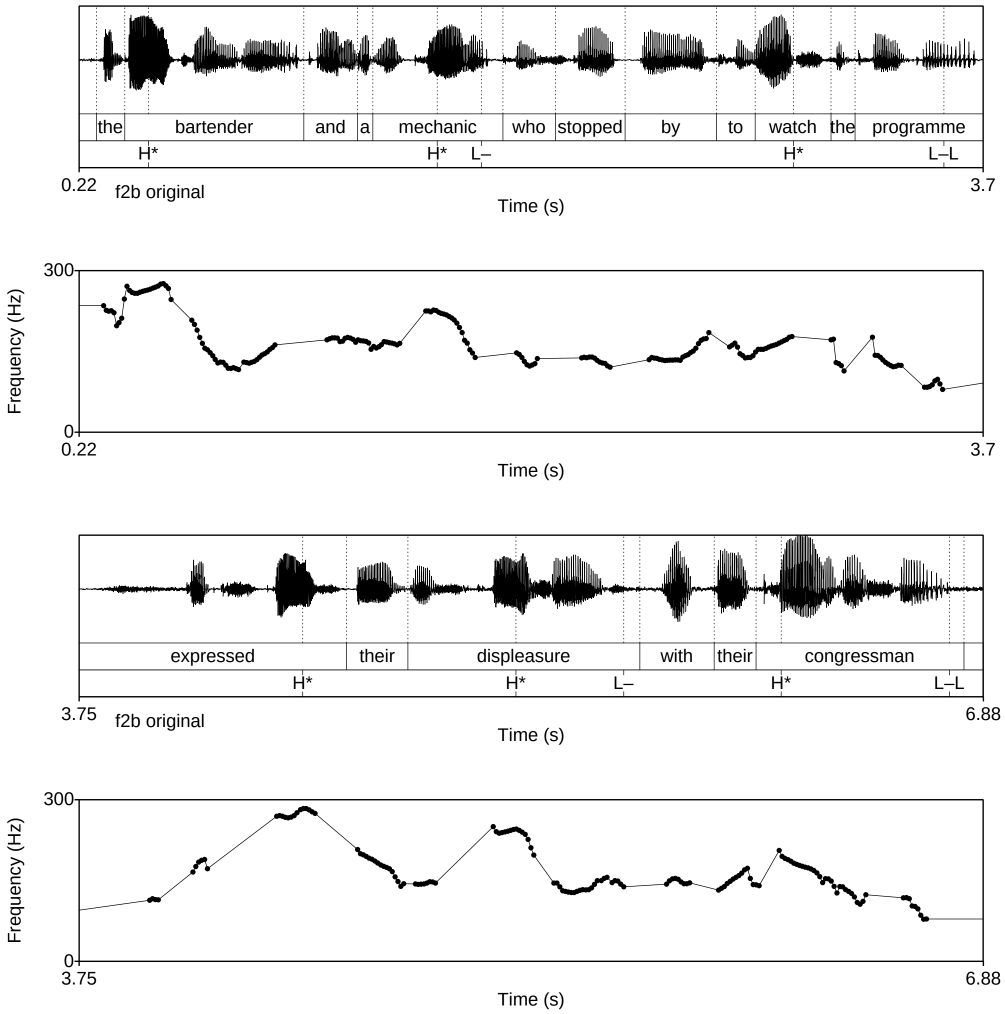


Figure 8.1: Example Broadcast News utterance from f2b. Natural Speech, female speaker.

We compare this contour with two synthesised contours. We use Festival as the synthesis platform, as it allows us to control the intonation in a way commercial synthesisers do not currently allow. This is more appropriate as we are particularly interested in using the input from language systems which would want to provide intonation patterns explicitly or at least provide strong hints to what an intonation pattern should be.

We compare Festival's default LR model, which is the best choice of model currently available for use by a language system using Festival, with the Context LR model developed in the previous chapter.

Synthetic speech was produced using Festival in two ways. One method was to use the hand marked ToBI labels in conjunction with the default LR target prediction model. The second method was to map the set of ToBI labels onto to the simpler 'a/b' labels used by the Contextual LR model, using the following simple mapping:

$$\begin{aligned}
(!)H^* \quad L-L\% &\longrightarrow afb \\
(!)H^* \quad L-H\% &\longrightarrow arb \\
(!)H^* &\longrightarrow a \\
L-(L\%) &\longrightarrow fb \\
(L\text{–})H\% &\longrightarrow rb
\end{aligned}$$

and including the relevant TG context information along with these symbols.

Figure 8.2 shows the synthetic pitch contours that are produced.

Comparison of the two synthetic contours reveals that they are reasonably similar. This is expected as they are essentially models of the same type trained on the same data. The differences however are interesting. Looking at what equates to the first TG of the utterance, the section relating to "The bartender and a mechanic", both contours produce a pronounced first accent, but only the Context model produces the dip towards the end of the word 'bartender' found in the natural speech. The default LR model then goes on to produce an accent of almost equal height on 'mechanic' where as the Context LR model produces a smaller accent, which again is a better match to the natural speech. In general the boundaries are reproduced more accurately and relative accent height is improved. These features are what we would expect to see based on the improvement in correlation that was found over the default LR model in the previous chapter. The context provides extra information about how individual pitch events differ from others that are close by.

Overall pitch placement and pitch range are harder to comment on as the synthetic speech is based on a male voice rather than the female voice of the original data. The pitch range of the Festival model declines a little over the utterance, where it seems to not do so, or reset more frequently, in the natural speech, but this effect is very marginal.

8.2 The Museum Domain

In recent years museums have moved from just having information cards next to exhibits to having pre-recorded speech which is accessed by various means when one looks at an exhibit. This has led to research into the automatic generation of this type of information with the aim of making the content more dynamic, with particular interest in the ability to make comparisons between recently visited exhibits and to personalise the information with a particular target group of visitors in mind. This research interest in the museum domain has led to a potential application for speech synthesis and has generated data to test synthesis with.

The domain is similar to broadcast news in that preprepared information is spoken aloud, but different in that the style is a generally calmer style than broadcast news, with less accenting, and less dramatic pitch variation.

The text shown in figure 8.3 is a sample from the M-PIRO project (see section 3.4.1). We consider how well we can generate intonation for it. This utterance has been generated by a language system, which can also supply information that should strongly influence where pitch events should be placed. As the system does not currently make this information available, the utterance was read by a male speaker and ToBI labels were assigned by a human labeller.

This utterance was used as a test to show how well the contextual LR model performs in this domain. We consider two issues: Does the model produce intonation acceptable for use in a domain other than that for which it was intended, and is it better than Festival's default LR model for input not generated by the CART prediction method that the default LR model is designed to work in conjunction with. More specifically, does the model produce reasonable intonation when the accent marking is externally provided and not predicted by Festival's own analysis of the text which would not necessarily provide the correct accents and would result in worse sounding intonation.

We again stick to using Festival as other 'better' synthesisers don't necessarily allow us to control the intonation in an appropriate way. Synthetic speech was produced using Festival in the same way as in the previous section.

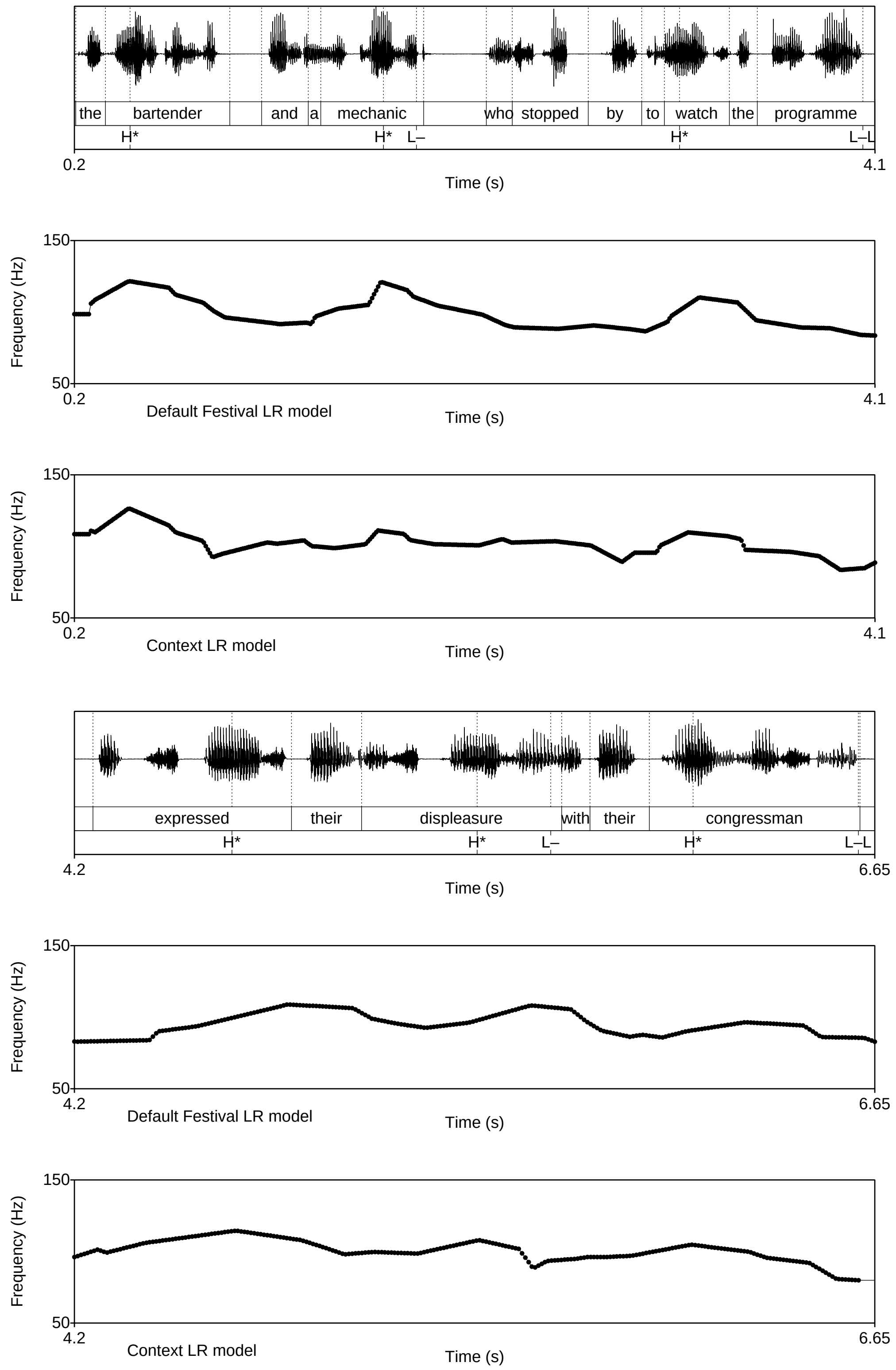


Figure 8.2: Example Broadcast News utterance from f2b. Synthesised speech.

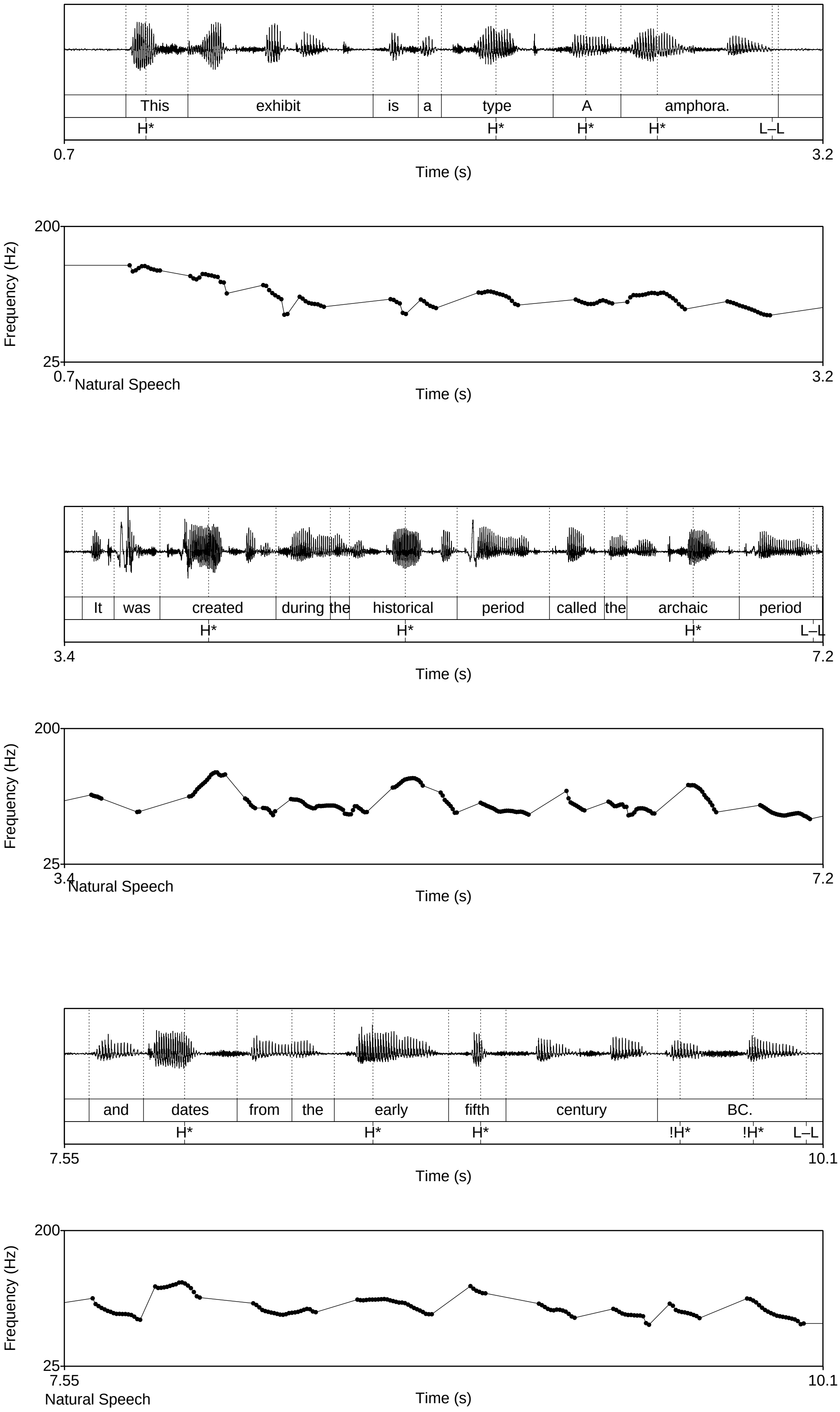


Figure 8.3: Example utterance from the Museum domain. Natural Speech, male speaker.

Figures 8.4–8.6 show the intonation contours of the synthetic speech. These are not shown with the natural speech as differences in segmental durations mean the utterances do not align well.

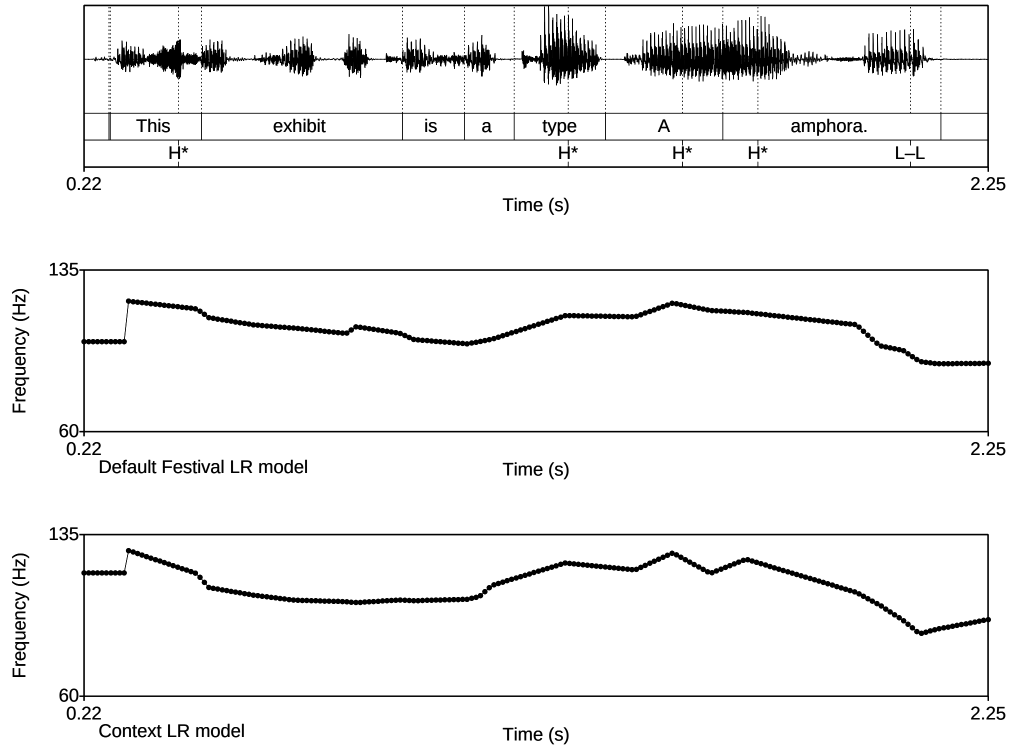


Figure 8.4: Museum domain synthetic speech (part 1)

Comparison of the synthetic contours reveals that they are similar as the previous examples were. On closer inspection the contour generated by the Contextual LR model is slightly 'bumpier'. In figure 8.4 the three accents towards the end of the phrase are more distinct and compare more favourably to the pattern found in the natural speech. In figure 8.5 the Contextual LR model shows clear small perturbations representing the lexical stress of unaccented words, which is not so obvious in the default Festival model. The default Festival model probably fails to account for lexical stress because it is expecting 'over accented' input, where an accent is likely to be present on much of the lexically stressed material. The Contextual LR model's ability to model unaccented material in this way, whilst still retaining clear accented material suggests that we may have achieved our goals in that the Contextual LR model produces good intonation where the accent marking is provided by an external source, and it is suitable for a domain other than that which it was trained on. We will see whether this is in fact the case in section 8.3.

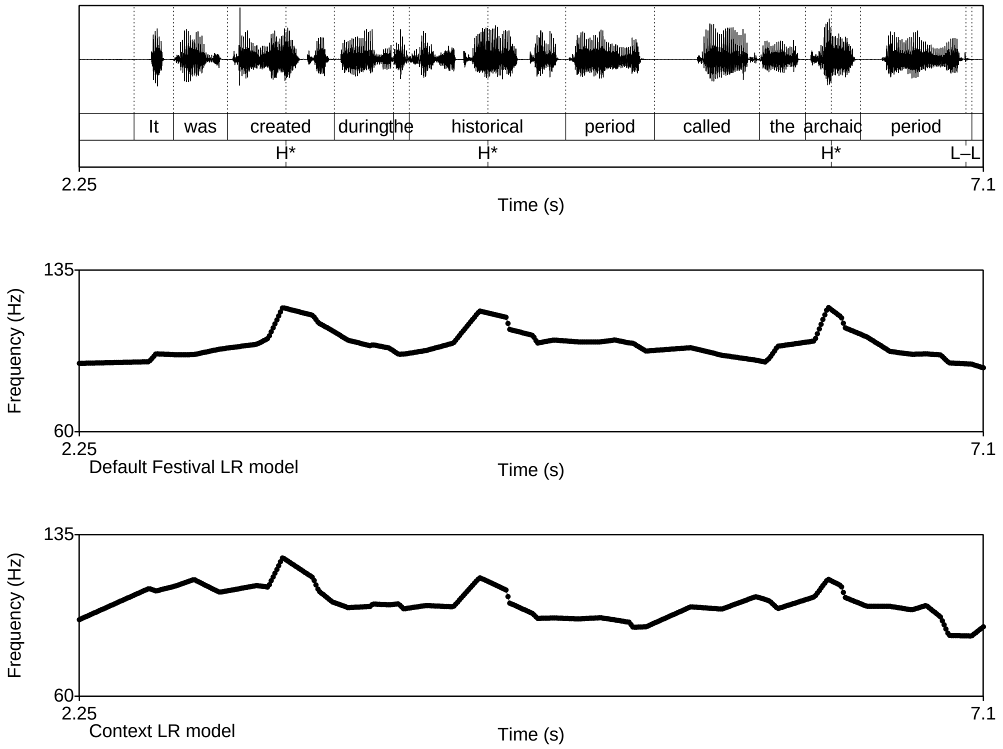


Figure 8.5: Museum domain synthetic speech (part 2)

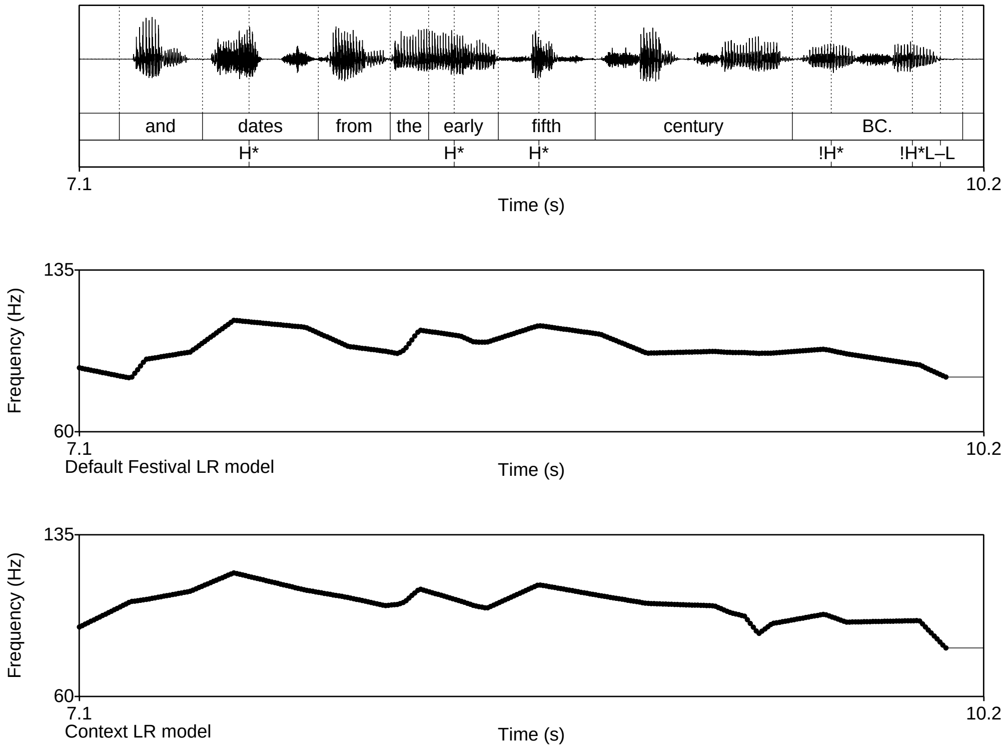


Figure 8.6: Museum domain synthetic speech (part 3)

Visually, the Contextual LR model also seems generates more distinct boundaries, particularly noticeable at the end of figures 8.4 and 8.6 which are IP final boundaries. This suggests that the use of prosodic structure that contributes to the 'context' part of the model is appropriate. So as long as the Contextual LR model proves to be perceptually acceptable, the example here suggests that we can generate better intonation than the Festival default model with a slightly reduced set of symbols.

8.3 Perceptual Evaluation

A simple perceptual evaluation experiment was performed, where listeners were asked to judge which of a pair of utterances was 'most appropriate'. The intention was to obtain a result to show that the intonation produced by the models is reasonable, based on a qualitative measure rather than the quantitative measures which we have previously expressed scepticism about, namely RMSE and correlation. We restricted ourselves to this form of evaluation, as more complex paradigms for evaluating synthetic speech are somewhat undeveloped, and it is not clear which type of questions really need to be asked to obtain results that can be interpreted in a meaningful way.

8.3.1 Methodology

We had two basic models we wished to evaluate: the Contextual Hat model and the Contextual LR model. There were also a number of natural and synthetic utterances we could compare to, ranging from completely natural speech through natural speech re-synthesised to completely synthetic speech produced by various other methods.

The experiment was designed to test two hypotheses. The first hypothesis is that subjects find discriminating between different, but quite similar, synthetic intonation patterns difficult. This hypothesis is based on the finding of Clark & Dusterhoff (1999). The second hypothesis is that subjects prefer the natural intonation over all the synthetic varieties, but prefer the new Context LR model over the other synthetic varieties.

To try to control for segmental quality all the speech was created by diphone synthesis using Festival. The overall speech rate was also controlled as much as possible to provide comparable utterances. Attempts to include the intonation from other speech synthesis systems in the experiment by overlaying pitch and duration information from these systems onto Festival's diphones produced very unnatural sounding speech. As the resulting speech was clearly distinct from the other samples, particularly with respect to overall pitch range and speech rate, it was not used in the evaluation.

This highlights a problem with manipulating the intonation of speech in general. Duration and f0 interact with each other in such a way that it makes it very difficult to manipulate one without the other, and realistically we do not know how to manipulate either well enough and consistently enough to change the duration of an utterance and keep it natural.

The result is that many of the manipulations that initially seem the right thing to do in producing synthetic speech for evaluation are not possible. For example, the only realistic way of altering the overall speaking rate of a natural utterance is to get the original speaker to say it again. Manipulating the duration by any means other than adjusting individual segment lengths relative to their context results in unnatural speech. Even if it were possible to adjust segments in this way, adjusting the f0 contour to match these adjustments is then itself a problem. This makes it very difficult to obtain natural speech that is similar enough to synthetic speech that listeners can make comparison judgements on and base those judgement on the particular aspect you are interested in.

Next, having obtained natural speech, we want to re-synthesise it to degrade its quality to match that of the synthetic speech. This can be done by taking the segmental timings and pitch contour of the natural speech and using them to produce synthetic speech. Unless the default phone durations and pitch values are very close to those of the natural speech, the quality of the resynthesis turns out worse than that of synthesis because more signal processing is required to produce the required result, This is particularly true of diphones. We could always degrade the synthetic speech as well, but we are then moving away from making comparisons to actual synthetic speech as it is output from the synthesiser.

Comparing two speech synthesis systems is generally harder as you tend to have no control over either durations or pitch to start with. If you do have control and you manipulate them, you are probably moving the quality of the synthesis away from its optimum and are no longer making a fair comparison.

With this in mind the following utterance types were chosen for use in the evaluation.

Natural (Nat) Diphone synthesis is performed using segment durations and a pitch contour extracted from a recording of the sentence spoken by a real speaker, so that the prosodic component of the utterance is completely natural. This type of synthesis emulates natural speech as closely as possible.

Festival Default (FD) Here the intonation is produced using Festival's default intonation model. This consists of a CART symbolic prediction stage followed by a LR target generation stage. Duration is predicted by a Z-score CART model. All models are trained on f2b. This type of speech provides a comparison with Festival as a text-to-speech system.

Festival Assigned (FA) Here pitch event labels are supplied to the same LR generation stage as above. This type of speech can be considered as the best Festival can do when intonation is assigned externally.

Contextual Hat (CHM) Intonation is generated by the Contextual Hat model discussed in section 7.3. Durations are based on the the same Z-score model as above.

Contextual LR (CLR) Intonation is generated by the Contextual LR model described in section 7.4. Durations are again based on the Z-score model above

The stimuli were made as similar as possible on non-prosodic dimensions. In particular an overall speaking rate was used to match that of the natural utterance, as any further manipulation of this utterance would make it less natural.

To reduce the load on the listener only a subset of all the possible pairwise comparisons was used. The comparisons that were used are shown in figure 8.7. The choices consider the Contextual LR model and Festival Assigned model as the likely candidates for a real synthesis application. The comparisons were designed to see how listeners judge them against each other and the other synthesis types.

Each sentence[1] was tested by 10 comparisons, the 5 comparisons above and the same 5 again but with the presentation order reversed. Where the same 'appropriateness' judgement was made irrespective of order a positive judgement was

[1] The M-PIRO sentence is actually two sentences, but we shall call it a sentence to minimise confusion.

considered to have been made; disagreement with respect to order was considered to signify uncertainty.

Three sentences were tested: a short Timit sentence, a longer f2b sentence and an M-PIRO sentence—a sentence generated by a real language system, an intended application for this research.

To judge appropriateness the listeners were asked to judge which of each pair they thought most appropriate for a given style of speech: the broadcast news style in the case of the f2b sentence (from this domain) and the Timit sentence (not from this domain) and a style suitable for a museum guide for the M-PIRO sentence.

A sample of natural broadcast news was provided for comparison for the broadcast news style, and it was suggested to listeners that they were helping to select a candidate to become a news broadcaster. For the museum style, the subjects were just asked to say which they would prefer as a museum guide.

Each block of 10 utterance pairs for the sentences were combined and randomised within the block for each listener. Listeners were presented with pairs of utterances on a web page. Individual stimuli were played by clicking on icons. Subjects could listen to each stimulus in a pair as many times as they liked and in any order before making a decision.

8.3.2 Results

Twenty seven native English speaking subjects took part in the experiment. Initially a second non-native group was also included, but this was dropped as dif-

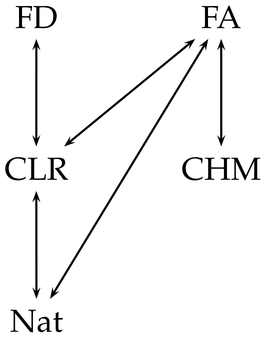


Figure 8.7: Comparisons used for perceptual experiment

ferent levels of competence in English and different first languages made it difficult to justify treating these subjects as a valid uniform group. The subjects were a mix of British English and American English speakers all listening to American English speech. No trends were found to suggest that their background affected their performance.

The first hypothesis, that speakers found discriminating between stimuli difficult was evaluated by analysing the consistency of the subjects judgements for each repeated pair of stimuli. If a subject made the same judgement the second time a stimuli pair was presented as they did the first time it was presented, they were judged consistent for that pair.

To be consistent a subject was required to judge a significant number of stimuli consistently. We calculate level of consistency as follows at the 5% level: there were 15 pairs of stimuli overall and the probability of making the same judgement twice for two pairs of stimuli is 0.5. The number of stimuli pairs needed to be judged consistent so that the probability that this would happen by chance is less than 0.05 can be calculated by examining the c.d.f. of a binomial distribution with $n = 15$ and $p = 0.5$. We find that $cdf(X = 10) = .9408$ and $cdf(X = 11) = .9824$. As this is a discrete distribution, we choose the $X = 10$ as being close enough to 0.95 to be suitable.

We now categorise subjects as consistent or inconsistent based on their ability to make 10 or more consistent judgements. 13 out of the 27 subjects, approximately half, were able to make a consistent judgement, validating our first hypothesis claim that this task is difficult. A closer look at the distribution of the level of consistency of individuals compared to how they would be expected to perform by chance alone strongly suggests that the subjects fall into two groups. Figure 8.8 shows a peak in the distribution around 8 pairs judged consistently. This reflects the behaviour we would expect by chance, and approximately half of the subjects fall into this category. A second much flatter component of the distribution then stretches out to the right relating to where subjects can perform the task. Some overlap between the components is to be expected as the subjects that are consistent may not show a preference for each stimuli pair.

We now consider the nature of the actual judgements made by the subjects. To do this we concentrate on those judgements made by the group of consistent subjects. The inconsistent subjects are not considered further at this point.

Figure 8.9 illustrates the judgements made by the subjects. The figure is split up into the three sentence types. For each sentence type, the different model types are shown as black boxes containing white text. The judgements between two tested model types are shown as a line, running between the two relevant boxes, with a number at each end indicating the percentage of times this stimulus was preferred. Preference values over 65%, which we will see are significant, are circled with a circle with a radius proportional to the preference value. This is designed to allow the significant trends shown by the diagram to be seen at a glance.

We see the following trends in the results. For both the Timit sentences and the f2b sentences, there is no obvious trend towards preferring the natural intonation contour over the synthetic ones that it was compared to. This is not necessarily surprising for the Timit sentences as the natural intonation here is not specifically broadcast news style that subjects were asked to rate as their preferred choice. The synthetic intonation patterns are a somewhat closer approximation to broadcast news style, although somewhat less natural than the natural Timit intonation. The natural f2b intonation obviously is broadcast news style, so it

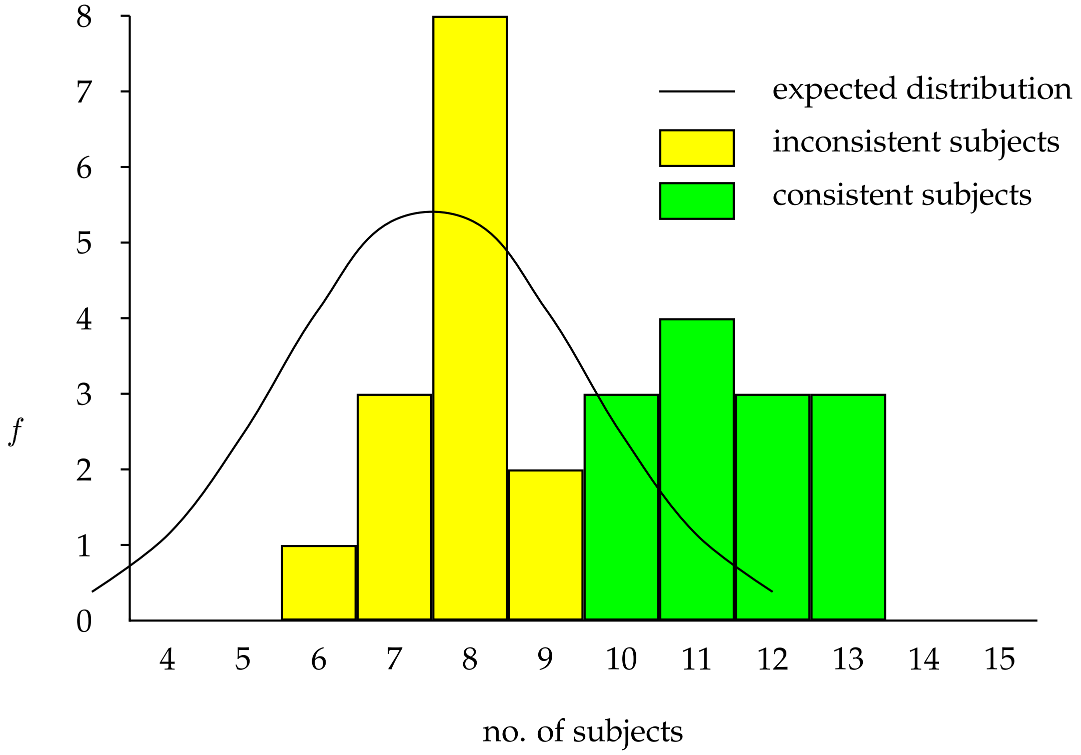


Figure 8.8: Frequencies of number of stimuli pairs judged as consistent by individual subjects.

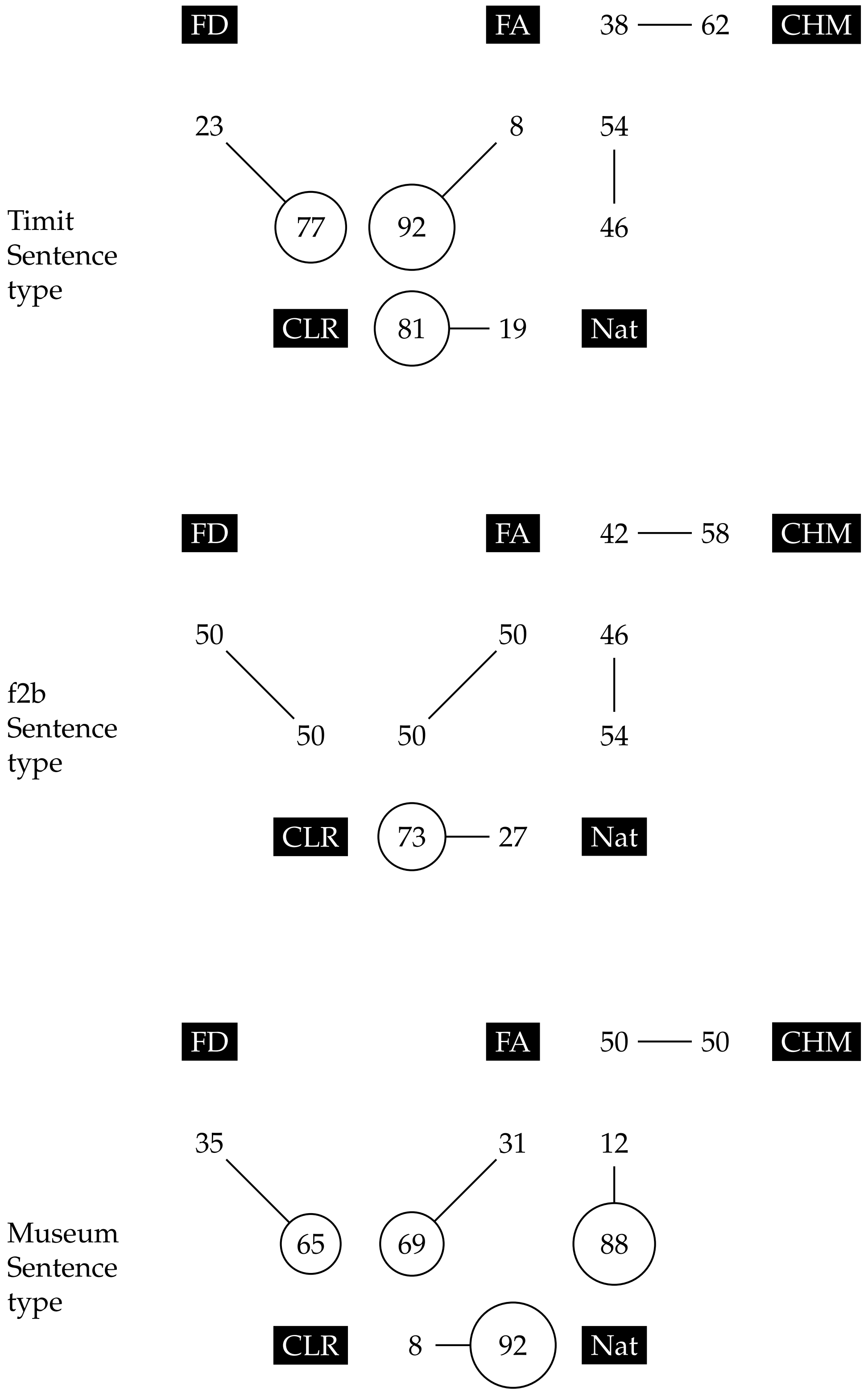


Figure 8.9: Preferred utterance pairs as judged by the consistent native-English speakers. See text for explanation.

is unclear as to why subjects don't find it the best example of the style. This is possibly because of the way in which the pitch range has been mapped from the original female speaker to the male synthesised voice, although this is done in the same manner as it was for the LR models.

For the Timit sentences there is a clear preference for the CLR model, in all three comparisons made against it, suggesting that the use of the TG parameters in the model do make a difference and help to capture the style. Statistically, as 26 judgements are being made for each stimuli pair, 2 from each of 13 consistent subjects, 17 of these judgements (65%) must be made in favour of one stimulus to produce a significant response with $p < 0.05$. For the Timit sentences only the comparisons involving the CLR model prove significant, and for the f2b sentences only the CLR comparison with natural discussed above is significant.

The preference for the CLR model against the other synthetic models is not maintained for the f2b sentences, in fact there are no clear trends at all for these sentences other than the CLR-natural comparison discussed above. This is most likely due to the length of the utterances. The f2b utterances are nearly twice as long as the Timit ones, with the intonation utilising the full range of TG structure. We can conjecture that with longer sentences it becomes harder to make a comparison judgement on the sentence as a whole. Instead, subjects tend to make a judgement based on a small part of the utterances which they find particularly salient, possibly because it is particularly good or particularly bad. Different subjects focusing on different thing and individual subjects focusing on different parts of the utterances at subsequent presentations result in no clear preferences.

There is also no preference between the rule based 'hat' model CHM and the FA model, suggesting that adding the TG information to such a simple model is not sufficient to make it as good as statistical models. The Festival assigned model is also never preferred to any of the other models, suggesting that accent assignment is less critical than the generation of a good contour once an accent is assigned.

For the museum sentences, the preference for natural intonation was significant. This is the expected result, and contrasts the above results for the broadcast news style sentences, adding more weight to the argument that the broadcast news style is not considered 100% natural by listeners. There is a also a significant trend towards liking the CLR model over the other synthetic models, which suggests that the CLR model is suitable for other uses than broadcast news. As far as hypothesis 2 is concerned it is accepted for the Timit sentences, rejected for the f2b sentences, and also accepted for the museum sentences.

8.4 Conclusion

We have seen that only half of subjects are able to make a simple preference judgement consistently. The results found here reinforce the ideas presented in Clark & Dusterhoff (1999), that untrained subjects find it very difficult to make judgements on synthetic intonation patterns, particularly when the differences between the synthetic patterns are small when compared to the differences between synthetic and natural examples. This highlights the need for careful screening when performing perceptual experiments involving synthetic intonation and particularly the need for consistency checking.

This consistency problem raises the questions: if so few subjects can consistently make useful judgements, then what is the value of the experiments and are the results from those that are consistent meaningful if they only reflect such a small portion of the general population?

One of the other problems encountered here concerns the length of the stimuli, particularly in the case of the f2b sentences, and it is not obvious how it can be resolved. Presenting shorter stimuli is one solution, but this is not an available option when wanting to specifically evaluate the overall pitch contour of a long phrase.

We believe that the problems are to a large extent related to the methodology being employed rather than in the data or with the subjects, and that because of methodological problems some of the results may be more useful than others. Problems with the methodology lie in the fact that very little work has been done specifically developing techniques for evaluating the perception of synthetic intonation, and the techniques adapted from other types of study, particularly where the stimuli are much shorter, do not work as well as they might. Part of the reason for carrying out these experiments then is to discover what the problems are and to refine the methodology to overcome them, as we have seen that careful processing of the obtained responses can still yield useful and meaningful results.

The results of the experiment show that the CLR model is a clear improvement on its predecessors and that incorporating TG context information does produce a better model. The fact that the CLR model was the preferred broadcast news style model, and also preferred for the less specific museum style, shows that the model achieves both of its primary goals of capturing the style of the data it was trained on, but also being useful for situations other than reading the news. It is likely that the model captures enough of the broadcast news style to be appropriate for that style, but captures it in a way that it can be used for other styles too.

As far as a building a better model for more general use than broadcast news, we recommend that the modelling techniques used here are appropriate, but the f2b dataset is not appropriate training data. Unfortunately other large, good quality, single speaker datasets that are intonationally labelled are not freely available. We attempt to find a resolution for this problem in the next chapter.

CHAPTER 9

LR Models with ToBI Input

9.1 Refining the Model

The new models discussed and evaluated so far have taken our simple accent and boundary classification to specify pitch event type. We recall that the decision to use this form of input was based on both wanting to have a model with a simple input which could be used without needing to subscribe to the complexities of ToBI and the fact that f2b is somewhat biased towards H* accents, with 83% of accents being H* or !H*. The outcome of this bias is that we can generate broadcast news style speech from an intonation description which is simpler than ToBI and this model is general enough for wider use, but it does only generate one type of accent, in ToBI terms, which may restrict its usefulness.

There are situations where the specific nature of the ToBI accents, particularly those other than H*, may be important. For example, Pierrehumbert & Hirschberg (1990) suggest that the L*+H accent is particularly good at expressing uncertainty, and the L* accent for accenting content already introduced into the discourse. Broadcast news rarely requires such constructs but they become much more important in a more general dialogue orientated systems. For example: L* can be used to express contrasts and L*+H can be used when asking for confirmation.

As just noted, the problem with the f2b data is that it contains relatively few low tone accents, and hence it is very difficult to train a model on f2b which can produce such accents. The LR model trained on f2b which comes with Festival fails to generate these accents accurately. A model trained with the TG structure framework developed here and ToBI labels for accents would also be expected to fail to capture these accents as none of the differences between the original Festival model and the models developed here focus on strengthening the importance of these types of accents, and hence there is no reason to expect our model to perform any better. This hypothesis was tested by building such a model. The model was built by replacing the simplified accent labelling we have employed up to now by the original ToBI labels and retraining the model. The resulting model is one which contains the prosodic structure components of the framework developed in chapter 7 but then relies on ToBI mark-up for pitch event descriptions.

Example output from this model is shown in figure 9.1. Here the utterance "the cat sat on the mat," is shown with the three different pitch contours: [H* H* L-L%] a normal statement contour, [L*+H L*+H L-H%] a contour expressing uncertainty and [L* L* L-H%] a contour to suggest that the hearer should already know that the cat was sitting on the mat.

The figure shows only very slight differences in f0 for these three quite different contours, specifically all the accents look like H* accents and all the boundaries look like L-L%. This clearly validates our hypothesis that there is insufficient data for f2b based models to reproduce these types of pitch events.

The format of the LR model allows us to gain more insight into how the model treats these pitch events. Recall that the LR model consists of a sum of weighted factors (see Appendix C for a full list). The factors representing different pitch events are mutually independent and hence the contributions of the factors for different pitch event types can be seen by examining their respective weights. The weights for the L* and L*+H pitch events are shown in table 9.1 along with those of H* which are shown for reference. For each pitch event the values shown

Accent	syllable has pitch event			previous syllable has pitch event		
	s	m	e	s	m	e
H*	21.0	15.0	7.3	0.0	0.0	0.0
L*	0.0	-27.2	-25.7	-14.2	0.0	0.0
L*+H	0.0	0.0	0.0	0.0	0.0	0.0

Table 9.1: Weights for features representing the f0 points for the syllable containing and the syllable following a pitch event.

are the contributions made to an f0 target point in Hertz by the presence of a particular pitch event. Target points are shown for the start, middle and end of the accented syllable and the following syllable. It should be noted that these values are each only the contribution of a singe factor in a sum of thirty or more factors, some which will increase pitch and others which will decrease pitch: for example the factor 'syllable is stressed' may contribute to a rise in pitch of a stressed syllable independently from a factor representing the presence of an accent. Zeros mean that a factor was not considered to make a significant contribution to f0 to be used as a factor and was therefore excluded by the stepwise training procedure. This is the case for all of the weights which contribute to the L*+H accents.

The stepwise procedure works by adding the factors to the model one by one in order of perceived importance. This order is judged by calculating which of

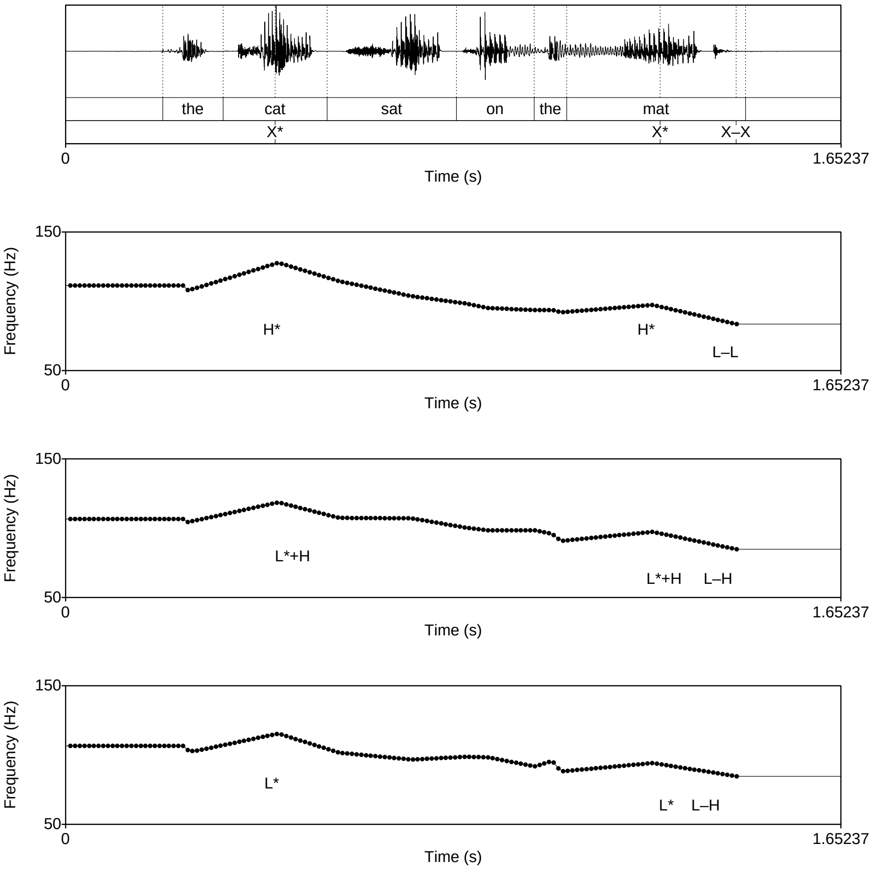


Figure 9.1: "The cat sat on the mat" with three different pitch contours generated by a ToBI based model.

the remaining factors can most improve the correlation of the model to a subset of test data. If no factor can significantly improve the correlation the procedure halts. The idea is that factors which do not uniquely contribute to representing the data, because they are noise or completely dependent on other factors, are excluded from the model.

The H* contributions raise the pitch in the accented syllable. The L* contributions lower the pitch at the end of the accented syllable and the beginning of the following syllable, but using the model shows that this is not sufficient to counter rises caused by other factors. The L*+H contributions as noted above are nil. This is not unexpected as these accents are effectively a small group of outliers in the data and a model which does not model them correctly is still going to model the 83% of accents and over 98% of syllables well. In the case of L*+H, there are only 12 accents in the training data of 12600 syllables. The problem is that in evaluating the L*+H factors the resulting change to all syllables is taken into consideration, and as 98% of syllables are unaffected, the factors are judged not significant.

Closer inspection of the L* and L*+H data shows that they are very noisy. Figure 9.2 shows a 20% sample of the L* accents from f2b. There is a clear mixture of accent shapes that one might expect to be categorised as an L* and shapes that one would not. Between 120 and 140 Hz there are a large number of roughly flat accents, some which dip slightly but more which peak slightly. Between 160–200 Hz are a selection of more varied shapes, but again most do not look like stereotypical L* accents. There is also one largely erroneous shape peaking above 280 Hz.

Figure 9.3 shows all of the L*+H accents in the training data. The general shape that would be expected for an L*+H accent would be a dip in pitch within the accented syllable followed by a late peak at the end or into the next syllable. It is clear here that there is pitch movement on both the accented and following syllables, but not necessarily as would be expected for an L*+H accent. Some accents peak rather early, and some, particularly those starting at around 200Hz seem to just fall. This could well be the result of bad labelling.

Better 'sample' accents can be seen by making a comparison with L* and L*+H accents taken from the ToBI examples database which accompanies te ToBI training materials. These are shown in figures 9.4 and 9.5. The ToBI data clearly shows a much cleaner distribution of points making up these accent shapes even though the data is from different speakers in a range of recording conditions.

We can however go further than just comparing these values. We can attempt to use the data from the ToBI examples to supplement the model in such a way that it is better able to generate L* and L*+H pitch events. The ToBI data cannot be added to the f2b data as the speaker pitch range characteristics are too varied and there is insufficient ToBI data to make a significant difference anyway. The structure of the LR model however can be exploited in a more subtle way.

Means of the measured f0 are calculated for syllable start, mid and end points for the accented syllable. For the L*+H, means of the start, mid and end points of the following syllable are also calculated. Example target accents are then synthesised using the model previously trained on f2b. Three sentences were used, with a rich enough prosodic structure to provide a total of eight accents, ensuring there is at least one accent sample in each TG context. From these, mean f0 target values were also computed to correspond to those obtained from the ToBI dataset above. The difference between the f2b target points and the equivalent ToBI mean is calculated and regarded as the error in the model. The ToBI means were compressed by 0.6 in pitch range and then lowered by 40Hz in an attempt to match the pitch range of f2b. This scaling is rather arbitrary but proved more effective than the standard normalisation procedure that has been used elsewhere. The problem is that the ToBI data comes from different speakers with different pitch ranges and there are insufficient examples to calculate the individual speaker pitch ranges in an accurate enough way to use the normalisation procedure.

The linear scaling that was used is demonstrated in figure 9.6. The initial compression was used to remove some of the excessively high values in the L*+H accents. The subsequent shift was then used to move the average pitch closer to f2b's average pitch.

One of the ToBI L*+H examples was however excluded from the calculations as an outlier as it appeared to be extreme and sound a little unnatural.

These means are summarised in table 9.2. The error value is then incorporated directly into the model. This is done by exploiting the fact that there are factors in the linear model which independently represent the presence of these accents.

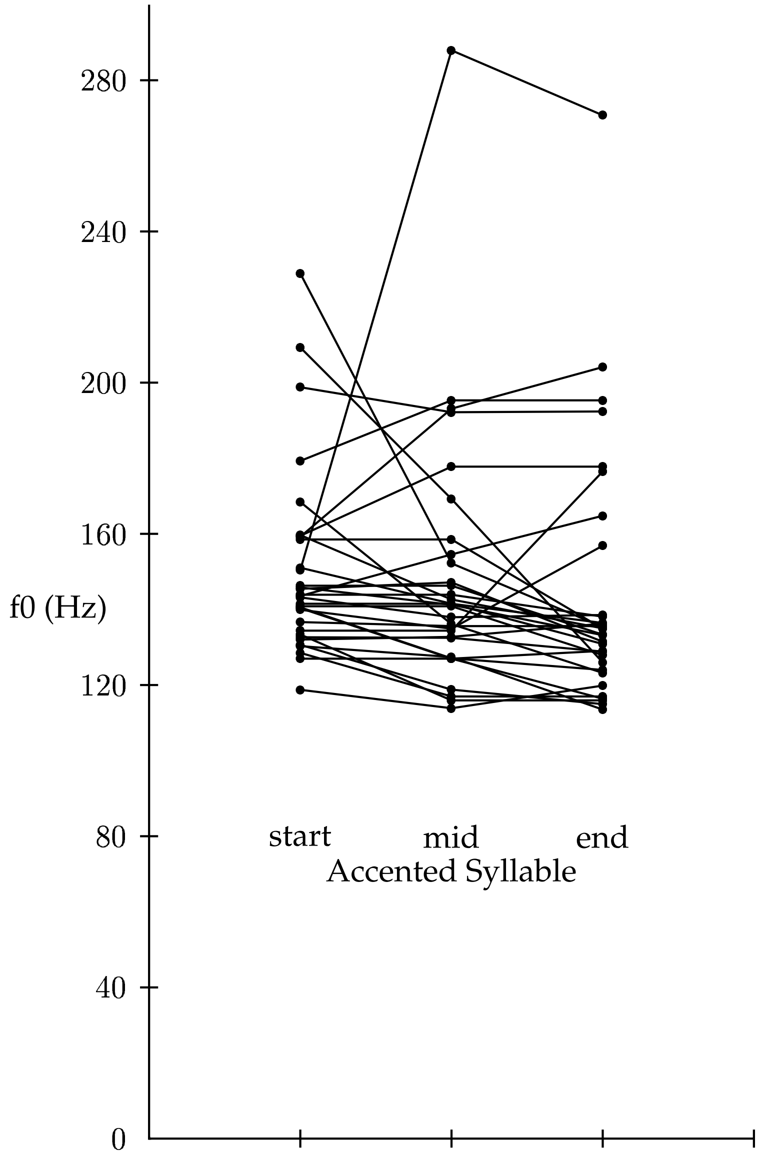


Figure 9.2: L* accents in f2b. This plot shows a random sample (20%) of the L* accents in the f2b data. Each line is a single L* accent. Each line consists of three points: the f0 at the start, middle and end of the syllable.

	accent syllable			next syllable		
	start	mid	end	start	mid	end
ToBI L*	177	145	167			
adjusted ToBI L*	147	124	140			
f2b L*	173	163	157			
error	-26	-39	-17			
ToBI L*+H	182	189	290	308	304	238
Adjusted ToBI L*+H	151	156	229	242	239	191
f2b L*+H	180	196	186	178	173	163
error	-29	-40	43	64	66	128

Table 9.2: Mean f0 target point values for f2b model and ToBI data.

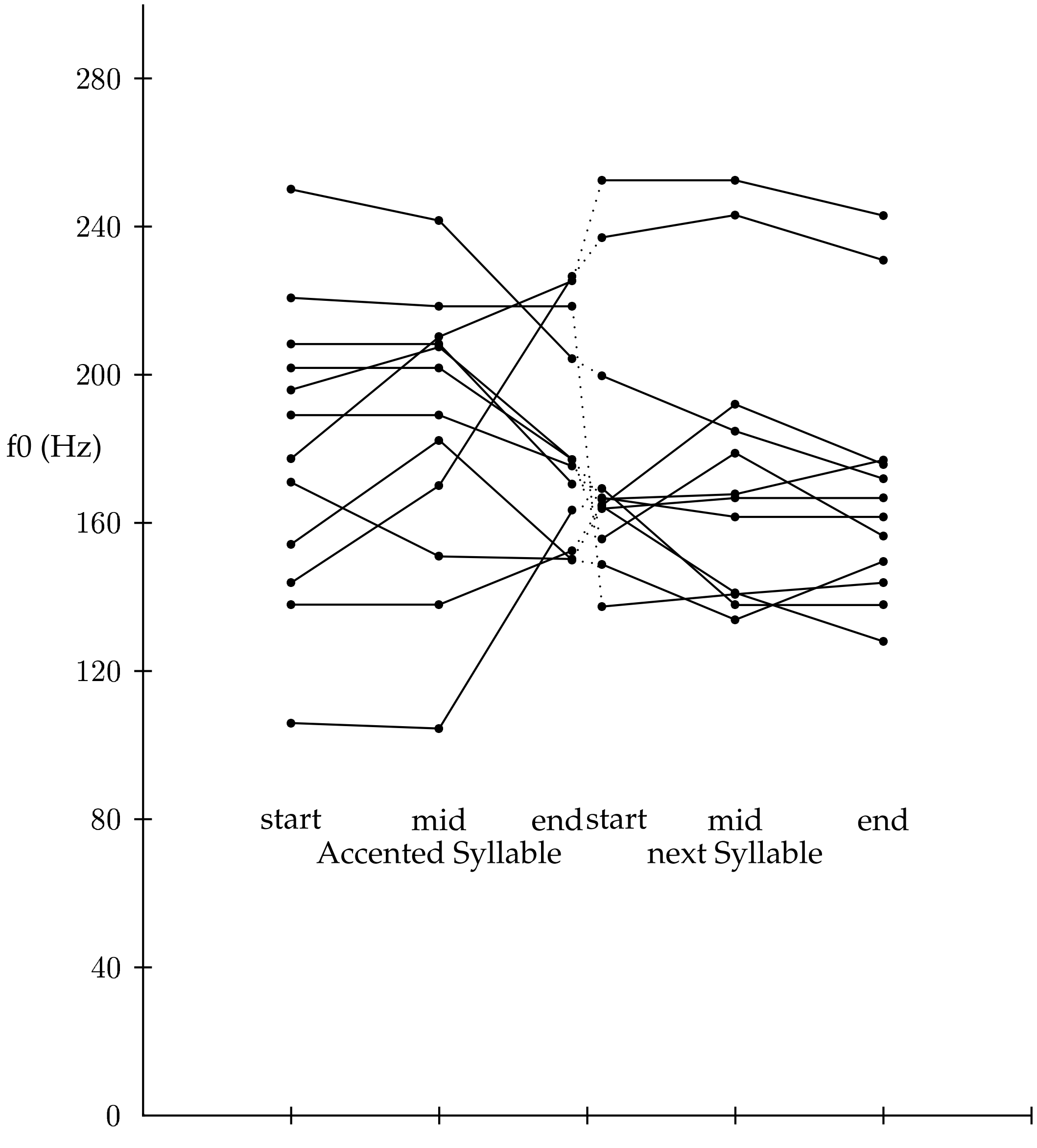


Figure 9.3: The 12 L*+H accents in the f2b data set. Each accent is represented by 6 points making up a line. The first three points make up the accent syllable and the later points make up the following syllable, the dotted lines show the connection between syllables. Large discontinuities at this join generally relate to an unvoiced section between syllables.

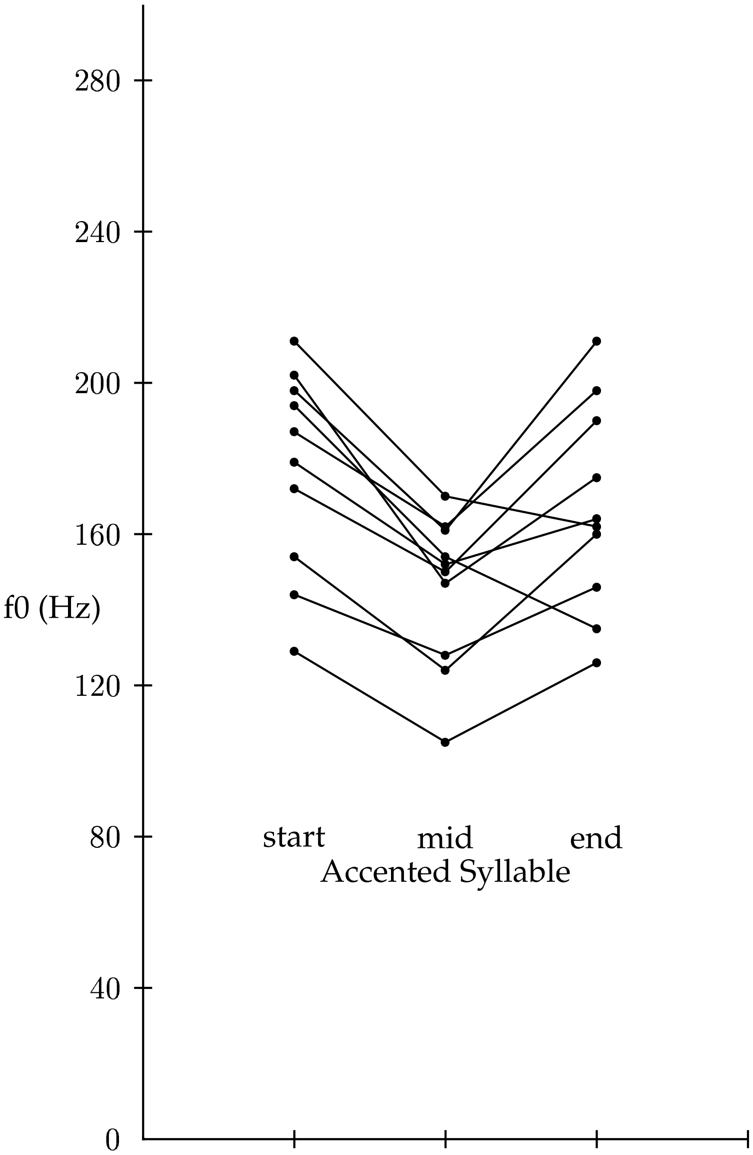


Figure 9.4: L* accents from ToBI examples

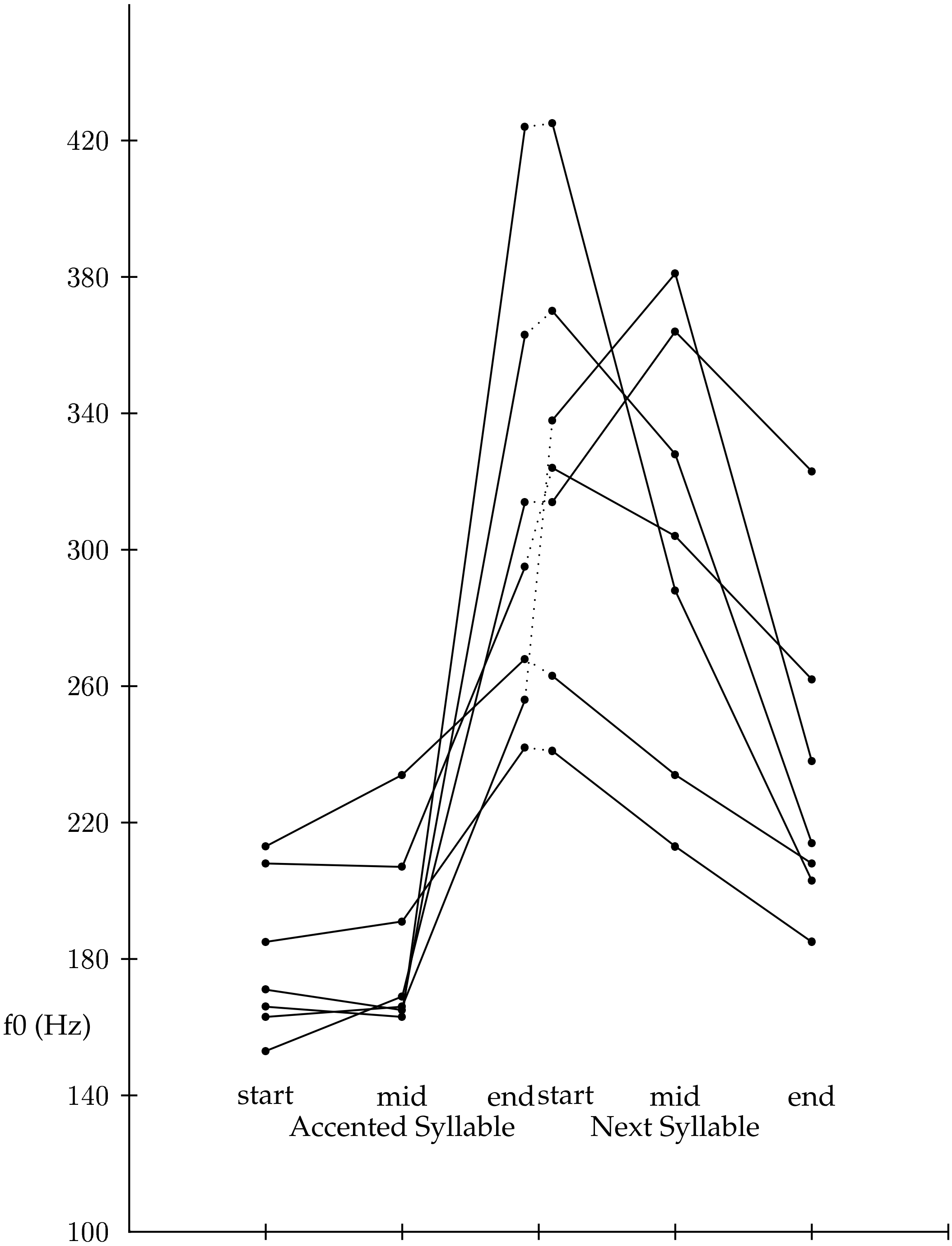


Figure 9.5: L*+H accents in ToBI.

The calculated error value for a target point was added to the value of the previously trained weight for that factor.

The new factor weights are shown in table 9.3. These are obtained by adding the error values from table 9.2 to the original model values from table 9.1.

Some of these values may seem extreme, but this is because they are intended not just to model the particular contour in question but to counteract the other factors in the default contour which may be opposing the required contour.

Synthesis from the resulting model gives pitch contours which are better than the unaltered model. The shapes of both the L* and L*+H accents are more like those that would be expected, but the pitch, especially with the H in the L*+H accents, sounds artificially exaggerated. This is seems to result from a combination of the exaggeratedness of some of the ToBI examples and the difficulty in mapping the

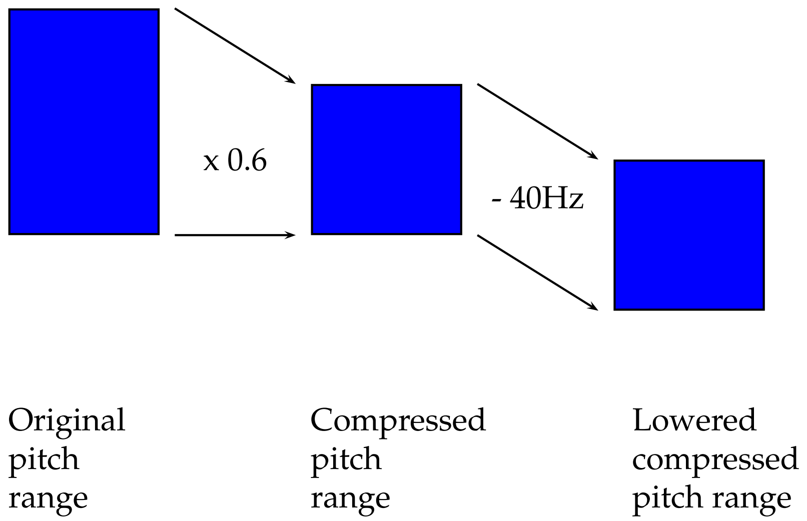


Figure 9.6: Linear scaling as applied to the ToBI example data. Pitch range is first compressed by 60% from the top and then lowered by 40 Hz.

Accent	syllable has pitch event			previous syllable has pitch event		
	s	m	e	s	m	e
L*	-5	-66	-43	-14	0	0
L*+H	-29	-40	43	64	66	128

Table 9.3: Manipulated weights for features representing the f0 points for the syllable containing and the syllable following a pitch event.

average ToBI pitch range to that of f2b. Example accents are shown in figure 9.7. Artificially large high stretches are apparent following both L*+H accents.

This leads us on to consider the notion of an idealised f2b accent. We have shown that the LR models can be manipulated by incorporating an error correction based on example ToBI data, but the correction relating to a 'sample' ToBI accent based on the average of a selection of ToBI accents is not really similar enough to f2b's pitch movements to provide a appropriate correction. This raises the question: is it possible to use our experience of the f2b data and our expectations of what we would expect to see with regard to L* and L*+H accents to produce 'idealised' f2b L* and L*+H accents and apply an error correction with respect to these?

The advantage of this approach is that an individual accent for each category can be tailored to fit directly into a single controlled context. This context can be created by the synthesis of a single sentence. The contour of this synthetic sentence can then be manually adjusted to represent the idealised accent. The error from the model can be calculated and the original model adjusted. As we will have only adjusted parameters which control the accent in question independently of its context, these adjustments should be applicable to other contexts as well and

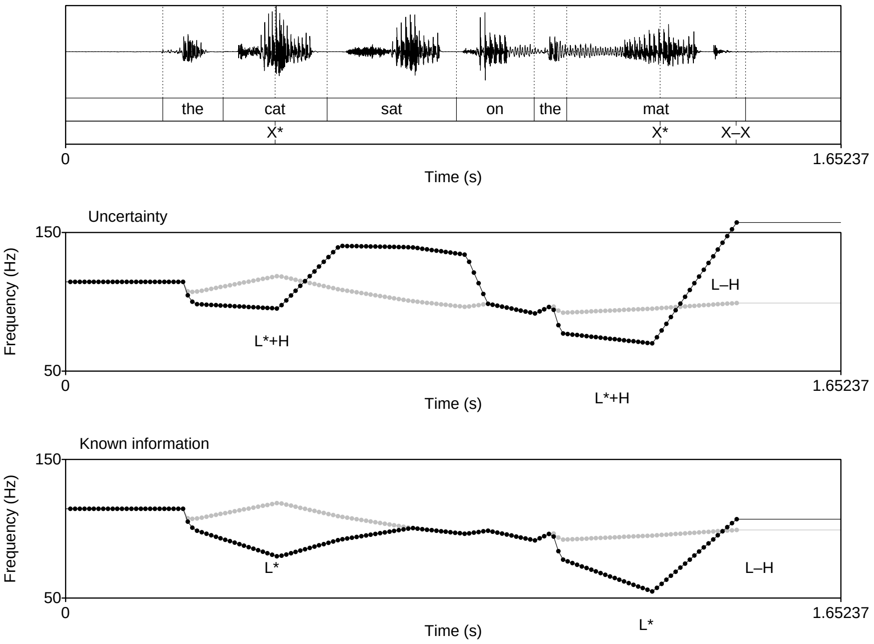


Figure 9.7: "The cat sat on the mat" with adjusted pitch contours generated by a adjusted ToBI model.

produce reasonable synthesis results for accents in these other contexts. The ability to do this is directly helped by the way in which the model parameterises the framework, as discussed in chapter 7 independently from the accent specification.

Table 9.4 shows the figures for such an adjustment for the L* and L*+H accents. Comparison with the previously adjusted weights of table 9.3 shows that these adjustments are similar to the ToBI adjusted weights but less pronounced.

Figure 9.8 shows synthesis using these adjustments. The first accent in each figure is the manually adjusted accent that was used to perform the calculation, and later accents in the example show the model producing an appropriate pitch pattern in other contexts. After an initial error correction to match a visual representation of 'idealised' accents further finer adjustments were made to improve upon the perceived quality of the accents (this is why the adjusted parameters are rounded to the nearest 5Hz). The resulting accents sound slightly exaggerated, but are less dramatic than the accents made with ToBI adjustments. This provides a suitable compromise which produces clearly distinct accent types which do not seem too artificial.

The discussion here has concentrated on L* and L*+H but the arguments are equally valid for other accents and boundaries. L-H%, H-H% and L+H* were also manipulated in the same way as L* and L*+H. Resulting weights for the ToBI adjustments and the idealised f2b adjustments are summarised in appendix D. The use of L-H% adjustments is also demonstrated in figures 9.7 and 9.8 as L-H% is used as an appropriate boundary tone for the example contours. There were however some problems with finding a suitable idealised L+H* accent. This is discussed later insection 10.3.

Accent		syllable has pitch event			previous syllable has pitch event		
		s	m	e	s	m	e
L*	adjusted weights	-5	-45	-20	-10	-5	0
L*+H	adjusted weights	-5	-30	20	60	60	5

Table 9.4: Calculations for manually adjusted model weights.

9.2 Changing the Model

We have shown that parts of the LR models can be manually altered based on new data or a combination of intuition and examining the resulting synthesis. This section takes a more detailed look at the extent to which we can do this and the implications and consequences of such actions.

We need to know whether altering a given parameter will have an undesirable effect on the resulting model. This is effectively asking how changing a given parameter will affect the resulting model.

To answer this question we need to understand both the structure of the variables being used to train the parameters in the model, including their relationship to

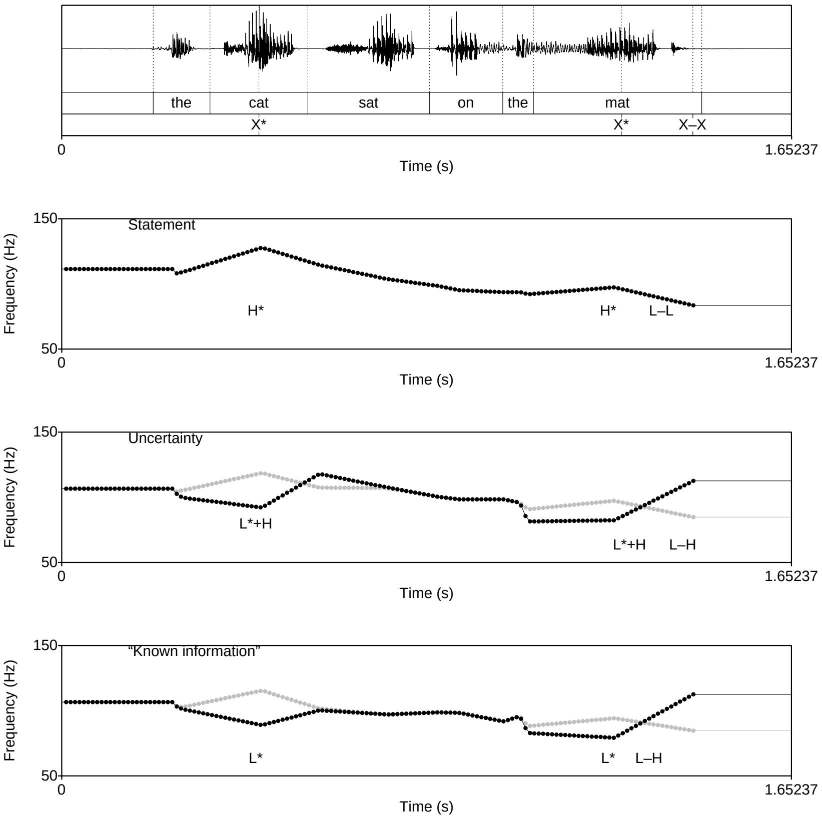


Figure 9.8: "The cat sat on the mat" with three different pitch contours generated by a manipulated ToBI based model. Grey contours show the equivalent unmanipulated contour.

each other, and the way in which they contribute to defining the dependent variable f0.

As we will be discussing how independent two variables are of each other in terms of the correlation between them, to avoid confusion we will refrain from calling the variables from which the parameters are trained 'independent variables' and call them 'factor variables' instead. We will also not call f0 the 'dependent variable' but rather the 'predicted variable'.

9.2.1 Correlation between factors

We start our analysis of the LR f0 model we have built here by looking at the correlation between the input factor variables by calculating the Pearson correlation coefficients between all pairs of input variables. As there are 88 input variables there are 3828 unique pairs of correlations. These are presented graphically in figure 9.9. A table of such values would be very difficult to interpret. The key to variable names and descriptions is given in table C.1. Recall that most variables come in groups where the main variable is supplemented by four others which provide a 4 syllable window. That is, if variable A registers some property of the current syllable, the variable p.A registers this property for the syllable previous to this syllable and pp.A registers this property for the previous-previous syllable. n.A registers it for the next syllable and nn.A registers it for the next-next syllable.

The general picture that figure 9.9 shows is that there is a large amount of low level correlation between pairs of factor variables. Correlations tend to show up as short diagonal strips in the figure. This is a direct result of the grouping of parameters representing a window of syllables. For example: if variable A correlates with variable B for a given syllable then for the next syllable, variable p.A will correlate with variable p.B by the same amount since these two correlations compare exactly the same properties of syllables. Also, where there is a strong correlation between such a series of variables there is often a weaker correlation surrounding the strip suggesting that when variable A correlates with variable B strongly, variable A will also correlate less strongly with variable p.B and/or n.B.

A example of this is visible between the variables 5–9 on the x-axis and 20–24 on the y-axis, representing a correlation between a syllable being TG medial and having an H* accent. Such a correlation is a little surprising, but can be interpreted as saying that there are a lot of H* accents in IP medial TGs. In such a situation a correlation between the next or previous syllable belonging to a medial TG and the presence of an H* accent, or a correlation between a medial TG syllable and the next or previous syllable carrying an H* accent is understandable due to the locality properties of the TG variable. If the current syllable is a medial TG syllable then the syllables either side are likely to be too.

Some groups of variables are somewhat more independent than others, the 30-34 group (L*+H accent), the 50-54 group (H-L% boundary) and 55-59 group (H-H% boundary) in particular. More generally the groups which are the most independent are the groups representing the accent and boundary classifications – with the exception of H* and L-L%.

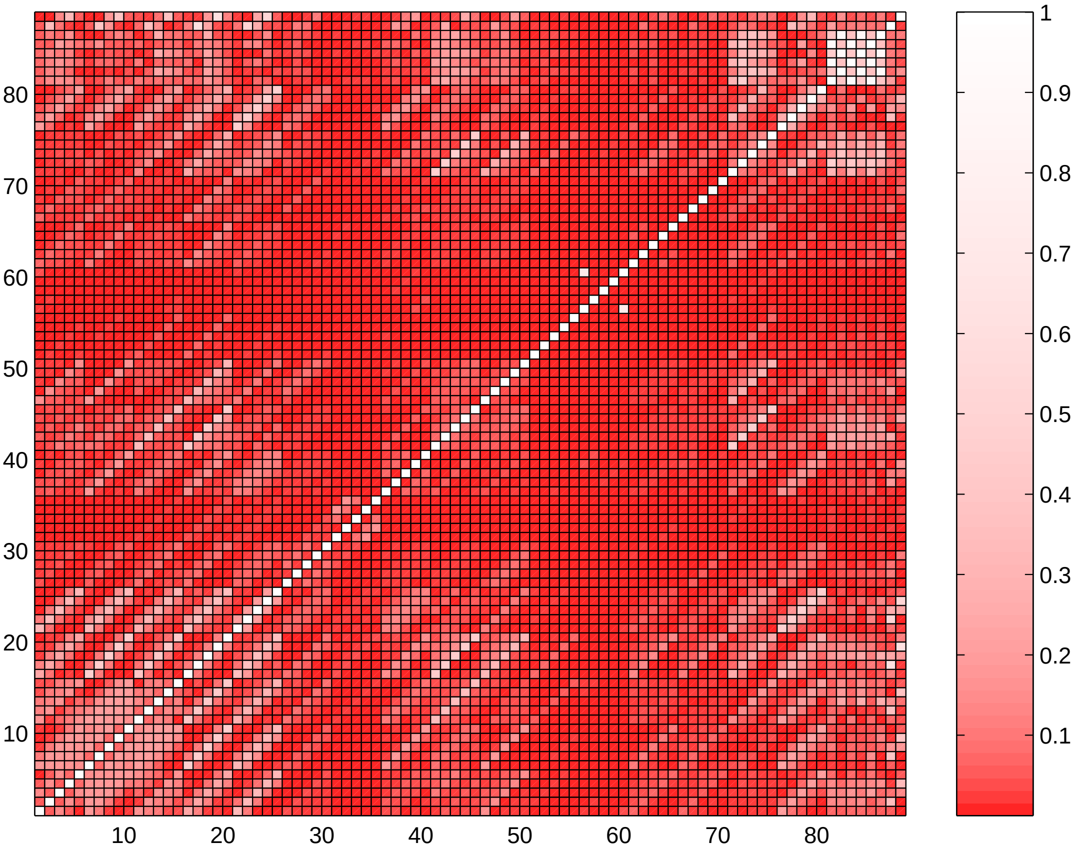


Figure 9.9: Correlations between factor variables used in LR model. Colour represents the r-value of the correlation. See table C.1 for key to variables names on each axis.

The variables representing the pitch events other than H* and L-L% could be independent, i.e. not correlated with each other for two reasons: they could be independent because the data is structured in a way that makes them truly independent or they could be independent because there is insufficient data for correlations between these variables to be seen. The latter case is most plausible as the H* variables do correlate with other variables and they have equal status with the other accent variables within the structure of the data. This is useful in its own right as we can use this lack of correlation where it would be expected to indicate that the data is insufficient to train the related parameter.

The variables representing the non H* accents and non L-L% boundaries are the ones that we wish to change, as these are the pitch events which occur less frequently in the data. The indepenence of these variables helps to ensure that any variance they contribute to the resulting model is unique, i.e. is additive and contains no redundancy, and that we will not be adversely affecting the contributions of other parameters.

9.2.2 Factor contributions to the predicted variable

In building the original model we used a greedy stepwise procedure to compute and add parameters, one by one, to the model whilst holding existing parameters constant to cope with the dependent contributions between the factor variables. When considering changes to individual parameters after training of the model has finished we need a better understanding of the relationship between the resulting parameters to judge how the changes will effect the overall model. We need to consider how an individual parameter contributes to the model with respect to other parameters in the model. To do this we need to know how individual factor variables depend on each other and more specifically how the interaction between factor variables accounts for the differences in the variance of the predicted variable f0.

To demonstrate this interaction we consider two simple examples where we have two factor variables (A and B) from which we wish to model a third predicted variable (X).

Firstly, if A and B are **independent**, i.e the correlation between them is zero, then each variable will make its own individual contribution to the variance of X depending on how well it correlates with X. Alternatively, if A and B are not independent and correlate with each other, then they still each contribute to the variance of X but any contribution which can be attributed to the region of dependence between A and B is redundant for one of the variables. So, if factor variables are not completely independent then the contribution they make to a predicted variable may not be unique.

In the model building process the consequences are as follows: variable A is added to the model and any variance in X that it can account for contributes towards the value of its parameter. Next variable B is added to the model, and the variance that it can account for is added to the model. This excludes any variance accountable by B which has already been accounted for by A as this is already present in the model. The result is that variable B contributes to less of the variance of X than it would do independently of A.

The order in which A and B are added to the model is chosen by the stepwise procedure, so that the one which correlates most with the predicted variable is added first. There is no general reason why the parts of A and B that contribute to the variance in X are the same parts of A and B that correlate with each other. There may or may not be some or complete overlap, and the size of this overlap determines the amount of independent contribution. The correlation between A and B alone is not sufficient to determine how much each can account for the predicted variable, although it should be obvious that the more the correlation between two factor variables the less likely they are to independently contribute towards the variation in the predicted variable.

From a statistical point of view, changing a parameter value can be thought of as the result of changing the distribution of the underlying variable that that parameter is trained on. If this distribution is changed, then the amount by which it correlates with another variable will be changed too. This in turn affects the amount of independent contribution that the other variables make to the resulting f0. Therefore if we change a parameter, we can no longer assume that we have the most appropriate parameter values for any other factor variables that it correlates with.

We can now make our first observation concerning how we should go about making changes to the parameters of the model. The observation is that if we change a parameter in the model, we should consider retraining all parameters that we subsequently added to the model after this parameter was added in the original training. This may or may not be practical, so we consider ways to relax this condition.

9.2.3 Structure underlying the variables

The factor variables that we used to build the model were specifically chosen to represent the underlying linguistic properties that we expected to be useful in predicting f0. The variables carry some of the underlying linguistic structure into the model.

The accent parameters, for example, that we changed in section 9.1 only contribute to groups of data points in the resulting model which are specific to where accents occur. The effect on the synthesis produced from the altered model is localised to the syllables immediately surrounding the type of accents for which parameters were changed. This localised effect is a direct result of the inherent structure of these specific factor variables. These binary valued variables are specifically designed to only affect localised areas in the resulting model, playing a specific role in the larger prosodic structure developed in Chapter 7.

Hierarchical structure

Recall that the prosodic structure is a hierarchical design. In terms of the resulting model which predicts f0, the higher a variable is in the structure, the more f0 points it directly affects. Also the nature of the design means that each variable affects localised stretches of f0 contour, which at the bottom of the structure relate to individual syllables, and at the top of the structure whole phrases. The nature of this structure, and its relationship to other parts of the data can be demonstrated with a subset of the parameters:

tgs TG is initial
tgm TG is medial
tge TG is final
stress Syllable is stressed
lstar Syllable has an L* accent
hstar Syllable has an H* accent

There is a clear hierarchical relationship between stress and the lstar and hstar parameters as shown in figure 9.10. The parent stress variable affects all stressed syllables, whereas the child accent variables only affect subsets of stressed syllables.

The consequence of a child–parent relationship in the hierarchy is that there is a clear overlap in the portions of the two variables which account for the variance of the predicted f0. Changes to one of the parameters representing such a variable, will interfere indirectly with the resulting contribution made by the other. As stated above, this interference is indirect in that the other parameter is not wrong per se, it just no longer reflects the value it would have been trained to given the change elsewhere in the model.

For example, changes to the stress parameter would affect target points for all stressed syllables, including all of those with H* and L* accents. The structure of the data is such that a change to the stress parameter may override the effects of the lstar and hstar parameters. The converse is also true, where changes to the accent parameters would affect at least a portion of stressed syllables. However, as the lstar and hstar parameters are leaves in the structure, if the stress parameter is already fixed, changes to the lstar or hstar parameters can be made safely, without interfering with the effect of the stress parameter. If we change an accent parameter this is akin to training this parameter with respect to other parameters as it was initially trained in the model building process, where the stepwise procedure attempted to work out the required order in which to include parameters to avoid interference. Furthermore, as the parameters representing the accents L* and H* are 'mutually exclusive' and each only has an effect on a localised part of the resulting f0 contour, alterations to either the L* or H* parameter will not have an effect one the other. For example altering the value of the L* parameter will have no knock-on effect on the H* parameter even if the H* parameter was added to the model after the L* parameter. The reason for this is that the contributions make by each of the parameters are independent, due to their mutual exclusivity. Although there is a correlation between the H* and L* variables, due to the fact that both variables have value zero for all syllables carrying neither H* or L*, the contributions they make to the predicted variable comes only from the uncorrelated parts.

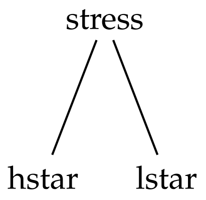


Figure 9.10: Hierarchical relationship between stress, hstar and lstar parameters

In general, where there are hierarchical dependencies in the data, adjusting leaf nodes of the structure is safe, but adjusting other parameters is not. This is not to say other parameters could not be adjusted, it just means that adjusting may have a knock-on effect on other parameters, namely their children in the hierarchical structure. For example in a working model, like the adjusted one we created in the previous chapter, the stress parameter would raise pitch on stressed syllables, and the L* parameter would compensate for this on L* accented syllables. This should come about naturally by the stepwise training procedure – in a perfect dataset stress would correlate more than L* with f0 and would be included in the model earlier. If the stress parameter were subsequently increased, the result would be to exaggerate stress on stressed syllables, but it would also produce an effect on L* syllables as the L* parameter was trained against the old stress value.

This brings us to our second observation: a variable that is mutually exclusive to the variable representing a changed parameter makes no contribution to the predicted variable, and so does not require a parameter adjustment, even if added to the model after the changed parameter.

Hierarchy and linear regression

It is worth stepping aside for a moment to relate the idea of hierarchy to the greedy algorithm used in the stepwise linear regression procedure. It is obvious that where hierarchy exists the parent parameter ideally needs to be added to the model first. The greedy algorithm knows nothing about the structure of the underlying data and only orders parameters by how much more variance of the predicted variable they can account for. In practise the parameters do get added to the model in the correct order, as the parent can account for more variance. This is due to the fact that compared to the number of stressed syllables there are relatively few syllables of each accent type, therefore the stressed variable naturally accounts for a larger portion of the f0 in the data as a whole.

For a situation where this is not the case, there are hierarchical linear regression techniques, where the parameter value representing a variable in a hierarchical relationship is a sub-equation in its own right. This is illustrated in equation 9.1. Here variable f_1 is the parent in the hierarchical relationship and the variables g_1 to g_m are its children. The parameter relating to f_1 is no longer a single value but a sub-model consisting of an independent component β_0 and other components relating to each of its children in the structure. The dependencies between parameters are much clearer here as they contribute directly to the model, and particularly changing a parent parameter involves taking into account the parameters of any children.

$$
\begin{aligned}
f0 &= \alpha_0 + \alpha_1 f_1 + \alpha_2 f_2 + ... + \alpha_n f_n \\
\alpha_1 &= \beta_0 + \beta_1 g_1 + \beta_2 g_2 + ... + \beta_m g_m
\end{aligned}
\tag{9.1}
$$

Cross hierarchy parameters

There are other parameters that by definition cross this stress–accent hierarchical structure. The tgs, tgm and tge parameters specified above are such an example. This arises because the TG context of a syllable is unrelated to that syllable's accent or stress properties. The mutual exclusivity of the tgs, tgm and tge parameters would allow the parameter representing one to be changed without the rest requiring adjustments.

As they represent a property of speech which is independent of stress and accenting, we make the assumption that their contribution to the predicted variable will be independent from the stress and accent parameters. The expected result of altering one of these parameters would be to shift the pitch range of the TG represented by the parameter up or down. This kind of adjustment is unlikely to be necessary as there is sufficient original data for these parameters to be reasonably trained. They could however be altered to emulate the speaker changing pitch range or in the process of using the model for a different speaker.

From this point onwards the relation between parameters becomes less clear as we consider the following additional parameters:

accent_count This is the nth accent in this TG.
syl_in Number of syllables since last TG break
syl_out Number of syllables to next TG break
ssyl_in Number of stressed syllables since last TG break
ssyl_out Number of stressed syllables to next TG break
asyl_in Number of accented syllables since last TG break
asyl_out Number of accented syllables to next TG break

The first parameter, accent_count sub-categorises accented syllables but in a way which is independent of the accent type categorisation parameters hstar and lstar.

The syl_in and syl_out parameters and the 's' and 'a' versions of them all count the relative position of the current syllable with respect to the TG it is in. It is unlikely that these parameters make a completely independent contribution to f0, particularly as there is a high correlation between these variables reducing the amount of possible independent contribution each could make. Altering any of these parameters would probably have repercussions elsewhere in the model.

Our specific model is further complicated by the fact that it does not just have parameters which consider the current syllable, but it has parameters which consider the two syllables to either side of the current syllable. As syllables are not randomly distributed within utterances, properties of adjacent syllables are not going to be independent of properties of the current syllable. If a syllable is stressed (or accented) then the chances are that the previous and next syllables are going to be unstressed (or unaccented). Or alternatively, if an accented syllable belongs to an initial TG then the chances are that the syllable next to this accented one does as well, and it is even more likely to belong to the same TG if the accent is the first accent in a sequence or the syllable is the first syllable in the TG. These relations lead to many interactions between the parameters, in a way that is difficult to express. However, as the correlations between a variable presenting a given syllable and one representing an adjacent syllable are generally lower that the correlations between parameters representing the same syllable, the likelihood is that the interference caused by these parameters is very limited and they can be treated in the same way as the relevant parameter for the given syllable.

In summary, we are led to the conclusion that the parameters we can safely change are those which are completely independent of other parameters, those which are clearly the children in a hierarchical dependency with other parameters and those which are mutually exclusive to each other. Independent parameters are easy to find through simple correlation calculations, but working out exactly which parameters meet the hierarchical dependency criterion is more difficult. To our advantage we find that where there is insufficient data to train a parameter properly we find the underlying variable correlates less with other variables that we might believe to really be the case, which both allows us to see a potential problem with the performance of the model and allows us to safely adjust the parameter in an appropriate manner. It is to our advantage that these parameters that we can change are the ones we generally will want to change.

It may seem surprising that the parameters that we want to change are the ones we can change, but this is more likely to be the result of our expectations and understanding rather than coincidence. Expectation of how the model should behave in a particular situation is brought about by our understanding of the model and the underlying processes. This understanding is also what allows us to develop ways in which to safely manipulate and improve upon the model. Conversely, we expect less from the parts of the model which we do not understand so well, and do not feel so compelled to control the behaviour related to those parts of the model.

9.3 Evaluating the Altered Model

This section tests the hypothesis that the altered ToBI CLR model is preferred by listeners to the unaltered ToBI CLR model. A perceptual experiment was carried out using the same basic methodology as used for the experiment carried out in section 8.3, although the number of stimuli and subjects were adapted to be more appropriate for making a comparison between two models.

9.3.1 Methodology

The sentences used in this experiment were taken from the MagiCster project (see section 3.4.2). The text is an example paragraph of a doctor giving a patient a diagnosis of what is wrong with them. The primary reason for using these texts is that they are marked up with ToBI pitch events which are generated from an underlying concept along with the text to convey a particular meaning – this removes the need to predict pitch events from the text itself – and removes one step where possible error or inconsistencies could be introduced.

Five relatively short sentences were taken from the paragraph of diagnosis and synthesised using both the ToBI trained CLR model and the altered ToBI trained CLR model. The texts are marked up using APML, the XML markup language described in section 3.4.2 which specifies simple semantic structure including pitch accents and boundary tones. Festival was adapted to read this markup and to use the pitch event and phrasing information from the AMPL directly instead of assigning its own accents, boundaries and phrase breaks.

The text of the whole paragraph was presented to subjects to provide context to the individual sentences. The sentences make use of the accents H* and L+H* and the boundaries L-L% and L-H% of which L+H* and L-H% have both been modified as described earlier in this chapter in the altered model. The full texts and ToBI mark-up are shown in table 9.5. Two sentences from the original paragraph were not used as they did not employ pitch events which had been altered in the altered CLR model.

Each sentence pair was presented 4 times to each subject, twice with the altered variant presented first and twice with the unaltered variant presented first. The 4 variants for each sentence were then combined making 20 stimulus pairs in total. These were then were presented to the subjects as a single block in a different random order for each subject.

Six of the subjects who were found to be consistent in the previous experiment were chosen to take part in this experiment giving a total of 120 responses, 24 for each sentence. The subjects were asked to decide which of each pair they thought sounded the most natural.

9.3.2 Results

The total responses for each sentence are shown in table 9.6. Overall 79 out to the 120 stimuli pairs presented showed a preference for the altered model over the unaltered model. This is significant at $p < 0.01$ in a binomial test. Looking at the results sentence by sentence it is clear that the altered model is preferred for sentences 1, 2, 4 and 5, but the unaltered model is preferred for sentence 3. Each of these individual results are significant at $p < 0.05$.

We also consider individual subjects results for each sentence, because although the overall results are very uniform with a 75% preference for the altered model (for all sentences except sentence 3), this 75% is made up of different scores from each subject in each case. Figure 9.11 shows the results for each subject. Here we see that the preferences of each subject are somewhat different. By themselves only subjects 2 and 4 make a significant preference (at $p < 0.05$) for the altered model over all sentences; the others only show a trend in this direction. There is also a reasonable amount of variation between how consistent subjects are in making judgements, subject 2 being completely consistent and subjects 1 and 3 being the least consistent.

H* L-L%
Good morning Mr. Smith.

(1) L+H* L-H% H* H*
I'm sorry to tell you that you have been diagnosed as suffering
H* H* H* H* L-L%
from a mild form of what we call angina pectoris.

H* H* L-L%
This is a spasm of the chest resulting from overexertion when the
H* L-L%
heart is diseased.

(2) L+H* L-H% H* L-L%
To solve this problem there are two drugs I would like you to take.

(3) L+H* L-H% H* L-L% H* L-L%
The first one is Aspirin which is an analgesic.

H* H* L-L%
That is it relieves the pain.

(4) L+H* L-H% H* L-L%
I have prescribed it to cure your angina.

(5) L+H* L+H* L-H%
The only problem is that this drug can be associated with some
H* L-L%
side -effects.

Table 9.5: Pargraph of text used in the perception experiment to evaluate altered ToBI CLR model. The numbered sentences were used in the experiment; the lighter unnumbered lines were not as there would have been no difference between the models.

	ToBI	Altered
Sentence 1	6	18
Sentence 2	6	18
Sentence 3	17	7
Sentence 4	6	18
Sentence 5	6	18
Total	41	79

Table 9.6: Distribution of responses for experiment evaluating the naturalness of the altered ToBI CLR model.

9.3.3 Conclusions

The level of consistency in this experiment was lower than that found in the previous experiment. This was to be expected as the difference between stimulus pairs here was much less than in the previous experiment, as the models being used only produce localised difference in pitch around particular pitch events. We interpret the low level of consistency as meaning that the subjects found this task particularly difficult. This interpretation was reinforced by subject's comments after the experiment saying that it was harder than the previous experiment.

The second sentence produced a significant result which was against the overall trend. The prosodic structure of the second sentence for which subjects preferred the unaltered model is very similar to that of the other sentences, so there is no obvious difference between this sentence and the others to suggest why listeners prefer the unaltered version of the model for this sentence and the altered model for the other sentences.

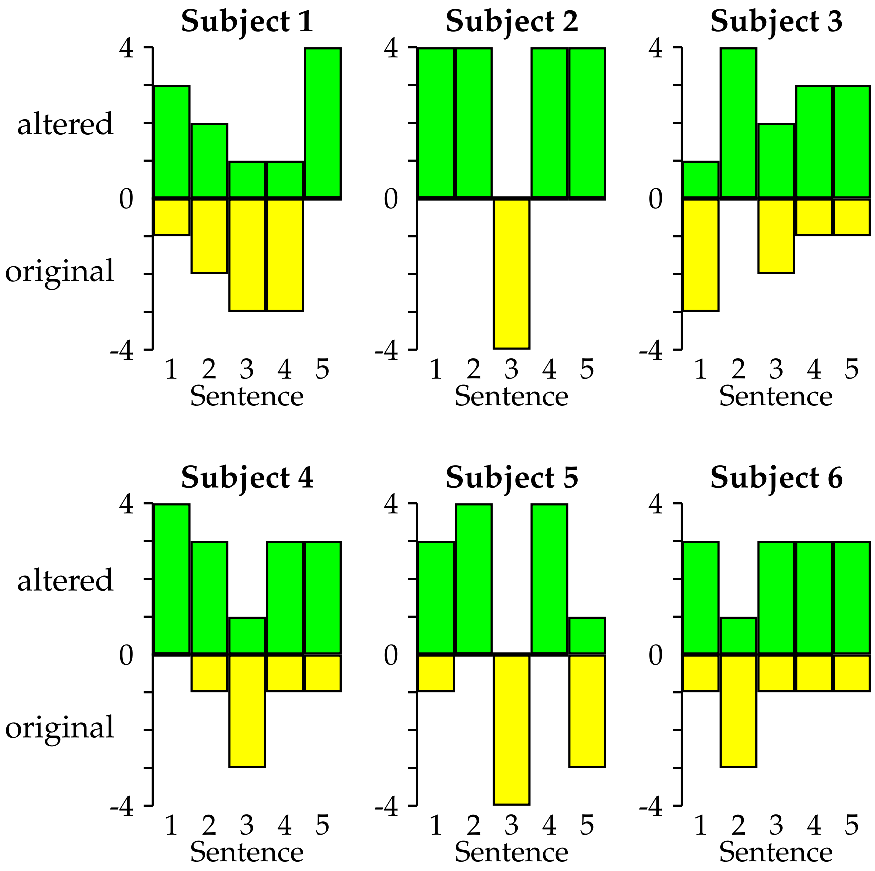


Figure 9.11: Results by subjects for experiment comparing the ToBI CLR model with the Altered ToBI CLR model.

The major difference between the resulting pitch contours for sentence 3 is that a clear rise in pitch is perceivable on the L-H% for the altered model, where it is clearly missing from the unaltered model. This leads us to believe that subjects either found the way in which this rise was generated to be unnatural or they thought that the allocation of a rise in that particular location sounded less natural than the absence of it. As the generated rise is very similar to rises generated elsewhere in the other sentences, we assume that it must be the allocation of it which is guiding subjects' judgements. The most likely reason for this is that the L-H% boundary sounds unnatural when placed at the end of such a short phrase in the middle of a sentence. The overall conclusion is that the modifications to the ToBI CLR model are worthwhile, and that listeners find the output of the modified model more natural.

CHAPTER 10

Discussion

We have been able to build models which can improve the quality of the pitch contours we can generate, in that our models can generate a wider range of pitch events than H* and L-L%. However there are a number of issues that have been raised during this building process which we need to address.

10.1 Accent Specification

A consequence of the models developed here is that we are now in a position to generate contours for utterances which are annotated with ToBI pitch accents and boundaries. The ability to do this raises some interesting issues with regard to accent specification. Subtle differences in accent specification now lead to clear differences in the final pitch contour.

The output of previous models was mostly determined by the location of accents, and accent type made little difference, as we saw in section 9.1. Specifying the correct accent type now becomes more critical as the choice is reflected in the resulting contour. Although in general this improves the quality of the synthesised intonation, there is the potential to produce bad, or at least confusing, pitch contours with this new model. With the other models, specifying a bad sequence of accent types would result in a non-specific neutral declarative type contour being generated, where now it will result in a contour which reflects the specified accent sequence, be it meaningful or otherwise.

ToBI is designed to be used for annotating real examples of speech and not to provide a specification of what makes a valid accent sequence for production purposes. Its only constraints are to ensure that tunes end with phrase and boundary tones. This lack of constraint is not really a problem with short utterances with only a single pitch accent, or two or three accents of the same type, as all such combinations will be meaningful. However, whether all combinations of different types of accents are acceptable as valid sequences in longer sentences is an area of research which has not really ever been addressed from a production point of view.

Steedman's theories (Prevost & Steedman 1994, Steedman 2000, Steedman 2002) provide us with a safety net in a model which can provide suitable accent specification for any reasonably formed utterance, guaranteeing a meaningful resulting contour. Whether this guarantees us a natural sounding contour (rather than a theoretically valid one) is unclear particularly where longer utterances are concerned as the production of such sentences has never been formally evaluated.

Long complex sentences pose particular problems because, as we have noted before, the literature tends to concentrate on short examples to illustrate specific points. For example, the average number of words in Pierrehumbert & Hirschberg's (1990) examples is five, and the longest example provided with fully specified intonation is only nine words long. There are longer and more complex sentences in the examples but their intonation is not specified as they are only used as contexts to elicit some of the more obscure intonation patterns that the authors wish to demonstrate. Other literature tends to follow these examples or variations of these examples.

In contrast, in a randomly chosen description of a museum object from the M-PIRO project (see section 3.4.1) comprising of 12 sentences, only one sentence contains fewer than 9 words; the average number of words is seventeen and the maximum is fifty-four! The compositional nature of Steedman's models could certainly provide mark-up for sentences of this length. The resulting intonation contours may or may not sound natural for a number of reasons:

Firstly, the assignment of accent and boundary type as prescribed by the theory is very rigid, and may produce an accent specification which sounds synthetic due to its regularity.

The differences in accent type which result from the given/new and plus/minus agreed status of the material being accented may not be sufficient to provide natural sounding intonation. For example M-PIRO descriptions of museum objects often contain a series of sentences describing various aspects of an object where the object is the theme of each sentence. After a few repetitions the default theme accenting strategy starts to sound stagnant and a human speaker would probably adjust their accenting strategy to compensate for this. The theory does not account for this.

There are other reasons why accent type decisions may not be so straightforward. For example, rhythmic considerations (Giegerich 1980) would require the noun phrase "The New York Metropolitan Museum of Art" to have multiple accents on it irrespective of whether it should be accented otherwise. The position of those accents may further be governed by the context the noun phrase is used in.

There are also unresolved problems with the actual contour generation stage of synthesis. The pitch range control needed to provide suitable pitch placement of individual sub phrases in relation to each other within long utterances is currently beyond our abilities. Speech data with a suitable amount of variation in prosodic structure along with a theory to account is currently unavailable. We return to this issue in section 10.4.

10.2 Appropriate Mark-up

Results using data from the M-PIRO project raise another issue. Synthesis results from SOLEML (see section 3.4.1) have so far proved disappointing. The reason for this is that unlike the APML mark-up (see section 3.4.2) used elsewhere the SOLEML mark-up does not explicitly specify prosodic information. Reasonable prosodic phrasing was obtained from the mark-up using the following simple rules:

- An Utterance break occurs following the last word in an S constituent as long as the next word is not a conjunction or a preposition (as this generally signifies a sentence being used as a constituent to another). An L-L% boundary is associated with this phrase break.
- An IP break occurs after the last word in an NP when the following word is not a verb. This type break has an L-H% boundary associated with it.

In general these rules provide reasonable intonation phrasing, but they are far from foolproof. For example, a sentence starting with a preposition would cause the sentence preceding it to not be ended with a big IP break and L-L% boundary.

The lack of pitch accent specification in the mark-up meant that pitch accents needed to be predicted as they would from plain text input. In an attempt to achieve better results than Festival's default prediction model, a simple set of rules using the newness information was proposed. The following accent types are assigned to the syllables in nouns carrying primary lexical stress in the following circumstances:

- H* is assigned to words which are marked as new.
- L+H* is assigned to words which are marked as old.

This is a little simplistic but without other information, such as theme and rheme, it was difficult to find a better alternative. There are also H* versus L+H* issues which are discussed in the next section which need to be considered. These rules also over-generate and place accents on all nouns, and do not allow for accents to be placed on other parts of speech. The importance feature is not used enough or at least not used in an appropriate way by the system to be useful in deciding what should be accented and what should not. It is hoped that by the end of the M-PIRO project additions to the XML output can be made to aid speech synthesis further.

The point to be made is that mark-up in itself is not sufficient to specify good intonation. The mark-up has to be appropriate. It is the case that for the museum domain within which the M-PIRO system works, L* and L*+H accents may not be needed as the system is only providing descriptive information which does not necessarily require L* and L*+H accents. Because of this, reasonable, but certainly not great, synthesis can be obtained from SOLEML mark-up, but it would be much harder to obtain reasonable results from SOLEML in a domain where a wider variety of accents are used.

It has been suggested that the resulting synthesis may be better than Festival's default synthesis primarily because phrase prediction from SOLEML does not incorrectly insert breaks in the middle of longer sentences which Festival otherwise tends to do. For this reason the Altered ToBI model developed here is being used by the M-PIRO project.

10.3 The L+H* Debate

We have so far neglected to discuss the L+H* pitch accent, both in the development and testing of the ToBI orientated model in chapter 9 and in the discussion in this chapter. The reason for this is that what the L+H* accent actually looks like, and specifically what makes it different from an H* accent is difficult to answer. Whereas Steedman posits this accent as the major accent associated with agreed themes, Ladd is entirely sceptical about its status as a phonologically distinct entity. Opinions also differ as to how the pitch contour associated with such an accent differs to that of an H*. Pierrehumbert & Hirschberg's (1990) stylised contours show the difference as an initial rise preceding the aligned H*. How the pitch moves into these accents is not shown on the diagrams making it difficult to know how L+H* and H* are really supposed to differ. Assuming the pitch is not high before these accents, they are *both* going to involve a rise into the H*. Could the nature of this rise be the difference?

Changing the shape of this rise by altering the parameters for L+H* in the model discussed in chapter 9 proved unfruitful. A delayed and therefore sharper rise was perceived as no different from the standard H*. Including more of a low before the H*, to emphasise the rise, resulted in the accent sounding like L*+H rather than H* or L+H*. In fact generating a perceivable three way contrast between the accents L*+H, L+H* and H* has so far proved impossible. It is possible that the approach to modelling a syllable's pitch at three points allows insufficiently fine modifications to produce the desired contrast, but unlikely as the suggested contour shapes have all been obtainable using this approach.

Ongoing research by Calhoun, Steedman and Ladd suggests the main difference between L+H* and H* is an alignment one, with L+H* having a delayed and possibly extended peak. So far, attempts to synthesise this sound no different from a standard H*. It is entirely possible that any effect of a delayed peak is lost if there is insufficient voiced material after where the peak would normally be as Calhoun's data is based upon voiced material designed to carry such an effect.

Our position then on the L+H*/H* distinction is that our model could in principle support such a distinction and does at the moment produce slightly different contour shapes for L+H* and H*. However we are not aware of definitions of H* and L+H* that yield two perceptually distinguishable natural sounding contours.

It is difficult to confirm or deny the existence of L+H* based on the data analysed here. The accents marked as L+H* in f2b are marked to identify shapes the labellers considered to be L+H*. What led them to make an L+H* decision is not known, but they were unlikely to have marked L+H* to identify themes. So we can only conclude L+H* does not seem to be distinct from H* in the way it is labelled in f2b. We cannot tell if there are distinct distributions in the shape of the combined sets of H* and L+H* as they are used in themes and rhemes as no information structure is available to us.

10.4 Suitable Prosodic Structure

We return to the issue of prosodic structure. We have demonstrated that improvements can be made in the synthesis of intonation by the use of prosodic structure. Specifically we have shown that the modelling of utterance initial high accents, the overall declination of pitch range and utterance final falls can be improved by employing an appropriate structure.

We based the structure purely on the results of analysing the f0 patterns in the f2b corpus rather than deriving prosodic structure from other other linguistic structures such as syntactic or semantic structure. We took this approach as there was no way to obtain a more detailed consistently annotated prosodic structure.

This provided us with a framework for prosodic structure which is independent from the specific semantics or syntactic structure of an utterance. This, however, does not mean such a relationship could not be used to decide upon how to use the prosodic structure during synthesis. Moreover, the lack of a predefined relationship allows complete flexibility in defining such a relationship.

The use of the prosodic structure developed here does result in improved synthetic intonation but there are areas in which further improvements could be made if more appropriate data were available. Employing the structure developed here results in better synthetic intonation than is obtained without the structure, but the currently used distinctions between structural units are not completely adequate to make all of the contrasts that we may be required to generate.

A contrast that cannot be made was demonstrated by Ladd (1988) (see section 2.9), where similar utterances, which in our framework would each consist of 3 TGs, have differences in prosodic structure. The utterances are of the form "A and B but C" or "A but B and C", A,B and C being three separate main clauses with the same number of accented words in each. Recall from section 2.9 that the differences found between the two forms are attributed to a different hierarchical prosodic structure for each form due to the fact that 'but' makes stronger boundary than 'and'.

The problem we are faced with is that our framework does not allow for such a distinction to be made. This is not surprising as such a distinction is not overtly marked in the data from which we derived our model. TGs were only marked as being IP initial, medial or final. Patterns like this contrast may or may not exist in the f2b data but they are not marked in such a way that they could be distinguished.

This type of distinction between TGs could be easily accommodated within the framework we developed by using a metrical classification if it were available. TG position specified by position within the metrical structure of an utterance rather than by linear position as is currently done. More data may be needed as there would be more categories and a suitably consistent structure would be needed to be specified.

It would be difficult to derive such a structure from the f2b data. The major difficulty would be getting consistent labelling, particularly where more than three TGs are involved, as the number of possible structure combinations becomes large. The structure of broadcast news is also somewhat regular, and may not be able to provide sufficient contrasts between TGs to be useful. Additionally a larger number of distinct TG types would probably reduce the number of pitch events available for training in each TG to a level where it was not possible to train a reasonable model anyway.

The preferred way to implement such a scheme would be to create a new corpus of speech where texts have been generated by a language system which can supply suitable metrical structure along with the text. This would simplify matters considerably in that the structure would be consistently marked-up across the corpus and in a meaningful way which related to some underlying concept being used by the language system. The recorded speech would of course need to be checked for its consistency with the generated structure, but this would be a much easier task than arbitrarily trying to decide what the structure of a given complex utterance was from scratch. The corpus would probably need to be somewhat larger than the hour or so of the f2b data to provide a balanced set of metrical structures.

10.5 The Need for Better Data

A particular problem that we found with the f2b data relates to the distribution of accent types. Recall that during the analysis phase of the development of our framework and models, we considered only peaked accents along with falling and rising boundaries. This reduced set of pitch accents was then used to build models which produced acceptable intonation for broadcast news. Attempts to progress further and build models using a wider variety of accent shapes using ToBI pitch event descriptions highlighted the fact that f2b is very unbalanced in its use of different pitch events (see section 3.5.1). This problem was overcome by taking advantage of the structure of the LR models being used and adapting parameters controlling the generation of pitch events to produce more suitable output.

The distribution of different pitch events that f2b uses is not necessarily unnatural, it is just that H* and L-L% are generally more frequent than other pitch events. The lack of intonationally balanced (with respect to pitch events, in the same way phonetic data would be phonetically balanced) is another pointer suggesting that current datasets that are available are not ideal for building structured intonation models for speech synthesis.

Based on the problems we have encountered with the data that is currently available we can summarise what we would like to see in a dataset ideally suited for this type of research.

- A large amount of data from a single speaker. This is probably the primary need for building intonation models. We ideally need multiple hours of data from an individual speaker.
- A large amount of variation in what is being spoken. Read news alone will not produce a model suitable for general use. A collection of data from different domains would be more appropriate.
- A balanced set of pitch events. The dataset needs to be constructed to include more of the less frequent pitch events than would occur naturally.

- Complex prosodic structure. The dataset should include longer utterances covering a variety of different prosodic constructions. Ideally these utterances would be provided with a predetermined structure eliminating the need for labellers to guess what it is.
- Clear speech which can be easily pitch tracked, or the inclusion of a laryngograph signal would also be helpful.

The above summary describes a dataset which would be needed to take the next step forward with the research carried out here, and it is hoped that it can provide a useful specification for anyone considering recording speech data to achieve similar goals.

10.6 Further Uses of the Altered CLR Model

The parameters of the CLR Model were altered in chapter 9 to improve the realisation of pitch events which were otherwise badly generated due to the lack of data. The same method could however be used to alter parameters for other purposes. The method could be used to modify the realisation of pitch events be more appropriate for a different dialect where data in that dialect is insufficient to train a complete model on. This could even be extended to a different language as many such language or dialectal differences concern only differences in peak alignment and shape which could be modelled appropriately within the three syllable window around the syllable the accent is assigned to. Adding new accent types is also possible.

The criteria for using a model on a different language or dialect from which it was trained would then be twofold: firstly the pitch movements on unaccented material would need to be considered similar enough or unimportant enough to be compatible with that produced by the model in its original form. Secondly the pitch events in the inventory of the target language or dialect need to be able to be specified as 'relative' pitch movements through the three syllable window centred around the syllable to which the event is assigned. Relative in this sense means relative to the pitch level the syllables would receive if they were unaccented, from this the parameter errors that need to be added into the model could be calculated.

10.7 Conclusions

Recall from chapter 1 that we set out with our primary goal was to improve synthetic intonation for speech synthesis in a way which helps us to understand more of the linguistic issues which need to be taken into accout to do this task well.

The Altered ToBI model developed in chapter 9 has achieved our primary goal. We have not only produced a model which we have demonstrated to be better than previous models but we have produced a model which is much more flexible than previous models in that it can generate intonation relating to a wide range of intended meaning for a given utterance. As an added bonus we have done this in such a way that a model can be adjusted to compensate for deficient data where accent types are not as frequent as we would like.

Our linguistic approach to developing this model has forced us to face a number of issues in linguistic theory. Although these issues have been quite diverse in nature and relate to different aspects of prosody and intonation from prosodic structure and pitch range to accent assignment, there is a common underlying theme which ties them together. Most of the problems and issues we have had to face have come about because linguistic theories have been developed around simple examples, which are constrained in ways which we cannot guarantee when being required to synthesise speech. Rather than being critical of the lack of research into more complex constructions it is hoped that this thesis has shown the need for research into this area and will lay down the challenge for further work.

APPENDIX A

Phrasing Analysis Summaries

First Analysis Summaries

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7
IP_1	mean	174.59	161.48	154.11	153.52	154.85	151.42	190.75
	sd	54.09	36.51	31.97	28.74	33.83	35.24	45.22
	n	450	425	304	184	85	36	7
IP_2	mean	175.00	166.13	153.63	156.61	146.59	148.30	148.46
	sd	43.79	34.22	31.03	32.52	29.16	28.06	25.19
	n	290	261	174	93	56	25	12
IP_3	mean	165.82	163.26	150.74	164.89	164.72	175.00	173.15
	sd	38.56	35.37	30.42	52.40	49.42	42.02	27.28
	n	104	98	62	21	9	5	4
IP_4	mean	172.40	171.74	147.47	114.12	173.11	116.85	131.51
	sd	47.03	31.14	31.26	13.14	43.06	6.37	.
	n	30	26	16	4	3	2	1
IP_5	mean	136.92	161.83	114.03	185.69			
	sd	29.50	32.87	14.59	.			
	n	7	7	3	1			
IP_6	mean	105.78	134.42					
	sd	.	.					
	n	1	1					

Table A.1: Start f0 by IP and TG for analysis t00

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7
	mean	157.16	145.46	142.22	140.42	141.07	133.32	151.72
IP_1	sd	37.90	30.72	29.13	34.37	32.76	30.66	22.42
	n	450	425	304	184	85	36	7
	mean	162.76	142.72	139.34	140.66	136.84	133.33	118.90
IP_2	sd	37.68	32.01	40.95	30.98	28.90	24.15	9.95
	n	290	261	174	93	56	25	12
	mean	164.24	140.18	137.80	136.96	146.42	161.72	118.52
IP_3	sd	36.77	33.84	32.42	25.03	43.97	40.35	6.30
	n	104	98	62	21	9	5	4
	mean	155.11	139.85	139.42	118.61	130.42	122.08	114.51
IP_4	sd	37.92	34.53	25.89	15.86	21.83	13.33	.
	n	30	26	16	4	3	2	1
	mean	149.21	124.73	117.85	105.78			
IP_5	sd	39.64	24.15	19.39	.			
	n	7	7	3	1			
	mean	191.67	132.92					
IP_6	sd	.	.					
	n	1	1					

Table A.2: End f0 by IP and TG for analysis t00

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7
	mean	128.72	119.04	116.85	114.95	116.81	118.26	132.57
IP_1	sd	27.66	16.48	14.75	13.46	17.32	16.35	12.04
	n	450	425	304	184	85	36	7
	mean	131.79	119.22	116.05	117.52	115.30	114.93	112.41
IP_2	sd	28.20	17.73	15.96	16.41	12.39	11.36	7.49
	n	290	261	174	93	56	25	12
	mean	130.39	120.16	114.77	114.64	128.14	153.63	112.02
IP_3	sd	26.37	20.26	14.94	12.79	30.67	40.82	5.46
	n	104	98	62	21	9	5	4
	mean	122.63	121.82	120.36	108.76	110.47	107.83	102.81
IP_4	sd	24.24	21.32	18.57	8.12	8.82	5.54	.
	n	30	26	16	4	3	2	1
	mean	128.13	114.32	104.77	110.15			
IP_5	sd	21.28	10.49	3.48	.			
	n	7	7	3	1			
	mean	105.78	111.21					
IP_6	sd	.	.					
	n	1	1					

Table A.3: Min f0 by IP and TG for analysis t00

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7
IP_1	mean	258.88	207.94	205.55	207.41	210.31	207.79	219.89
	sd	42.58	33.64	31.15	34.14	39.86	32.37	31.55
	n	450	425	304	184	85	36	7
IP_2	mean	253.12	218.09	213.14	208.90	208.93	204.68	195.56
	sd	33.21	39.79	35.82	27.81	30.60	27.78	26.75
	n	290	261	174	93	56	25	12
IP_3	mean	255.37	216.83	205.08	217.12	205.73	211.36	205.06
	sd	27.69	30.71	29.03	53.39	46.97	49.60	17.93
	n	104	98	62	21	9	5	4
IP_4	mean	248.35	212.42	207.64	213.16	240.70	232.53	210.51
	sd	35.51	35.03	21.50	40.41	47.10	39.88	.
	n	30	26	16	4	3	2	1
IP_5	mean	244.00	215.91	195.30	185.69			
	sd	22.72	27.45	14.02	.			
	n	7	7	3	1			
IP_6	mean	254.91	213.73					
	sd	.	.					
	n	1	1					

Table A.4: Max f0 by IP and TG for analysis t00

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7
IP_1	mean	130.16	88.90	88.71	92.46	93.49	89.53	87.31
	sd	46.62	33.97	32.04	34.18	40.24	30.23	32.69
	n	450	425	304	184	85	36	7
IP_2	mean	121.33	98.87	97.09	91.37	93.63	89.75	83.15
	sd	43.16	41.51	37.50	27.85	31.40	30.65	27.55
	n	290	261	174	93	56	25	12
IP_3	mean	124.98	96.67	90.31	102.49	77.60	57.73	93.03
	sd	39.84	34.59	27.51	53.93	38.64	33.30	23.20
	n	104	98	62	21	9	5	4
IP_4	mean	125.72	90.60	87.28	104.40	130.23	124.71	107.69
	sd	38.88	32.90	26.84	43.46	44.07	34.34	.
	n	30	26	16	4	3	2	1
IP_5	mean	115.87	101.59	90.53	75.54			
	sd	31.14	20.37	12.49	.			
	n	7	7	3	1			
IP_6	mean	149.12	102.52					
	sd	.	.					
	n	1	1					

Table A.5: Delta f0 (Max-Min) by IP and TG for analysis t00

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
	mean	189.41	168.53	158.79	159.68	169.64	178.76	189.42		155.69
IP_1	sd	45.62	36.53	31.14	33.73	34.81	38.24	34.67		31.16
	n	425	304	184	85	36	7	7		443
	mean	185.25	166.55	157.14	157.47	156.28	151.43			158.68
IP_2	sd	40.35	33.93	28.97	36.64	39.07	31.87			28.64
	n	261	174	93	56	25	12			290
	mean	179.51	164.15	165.05	150.89	144.79	172.93			158.43
IP_3	sd	33.62	32.16	29.52	39.21	34.81	50.87			28.16
	n	98	62	21	9	5	4			104
	mean	173.23	174.95	179.58	153.87	149.96	141.62			156.20
IP_4	sd	36.96	28.47	17.33	18.46	25.23	.			35.13
	n	26	16	4	3	2	1			30
	mean	167.51	184.51	130.01						156.38
IP_5	sd	20.47	41.44	.						21.61
	n	7	3	1						7
	mean	170.29								161.62
IP_6	sd	.								.
	n	1								1

Table A.6: Start f0 by IP and TG for analysis t01

Table A.7: End f0 by IP and TG for analysis t01

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
	mean	155.00	144.17	143.90	149.13	153.16	170.46	148.97		127.08
IP_1	sd	38.91	30.56	30.87	31.15	36.04	43.72	31.08		25.89
	n	425	304	184	85	36	7	7		443
	mean	157.40	147.19	148.16	143.34	144.69	141.98			121.02
IP_2	sd	36.85	33.70	32.59	36.41	34.05	27.04			18.87
	n	261	174	93	56	25	12			290
	mean	158.70	146.02	145.23	139.81	167.63	172.51			121.93
IP_3	sd	34.76	33.25	31.42	24.76	36.08	24.47			20.71
	n	98	62	21	9	5	4			104
	mean	152.31	141.11	162.98	117.06	143.57	119.80			121.55
IP_4	sd	38.64	31.81	22.02	13.83	16.05	.			21.53
	n	26	16	4	3	2	1			30
	mean	152.78	135.98	103.30						111.66
IP_5	sd	40.63	29.91	.						11.31
	n	7	3	1						7
	mean	190.27								111.57
IP_6	sd	.								.
	n	1								1

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
IP_1	mean	128.96	120.41	120.38	123.48	127.63	132.22	124.37		112.66
	sd	25.75	17.67	18.04	19.51	25.62	25.70	20.41		11.45
	n	425	304	184	85	36	7	7		443
IP_2	mean	129.10	120.79	121.55	119.18	120.18	121.50			112.32
	sd	26.74	17.72	18.21	18.51	19.13	14.17			9.54
	n	261	174	93	56	25	12			290
IP_3	mean	125.65	123.22	121.08	113.49	127.72	160.99			114.61
	sd	22.89	21.32	17.66	17.29	21.23	33.70			11.57
	n	98	62	21	9	5	4			104
IP_4	mean	125.89	125.27	142.21	115.77	115.18	100.52			111.19
	sd	25.88	16.99	11.48	15.60	.45	.			9.50
	n	26	16	4	3	2	1			30
IP_5	mean	138.90	118.37	103.30						107.79
	sd	23.54	15.66	.						6.71
	n	7	3	1						7
IP_6	mean	148.48								111.57
	sd	.								.
	n	1								1

Table A.8: Min f0 by IP and TG for analysis t01

Table A.9: Max f0 by IP and TG for analysis t01

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
IP_1	mean	257.47	212.44	209.51	211.99	209.53	210.65	219.58		199.71
	sd	41.58	32.34	31.31	33.32	31.34	46.06	31.84		32.52
	n	425	304	184	85	36	7	7		443
IP_2	mean	253.25	221.51	218.86	213.67	216.52	208.33			202.65
	sd	29.66	34.97	30.81	30.70	31.03	26.73			27.65
	n	261	174	93	56	25	12			290
IP_3	mean	252.21	220.41	213.32	200.54	211.11	223.15			206.71
	sd	26.83	29.39	36.03	47.24	50.56	46.78			30.83
	n	98	62	21	9	5	4			104
IP_4	mean	252.15	218.43	213.79	225.34	220.45	198.52			207.44
	sd	32.84	34.97	18.17	39.32	46.29	.			31.29
	n	26	16	4	3	2	1			30
IP_5	mean	244.37	229.92	177.54						201.20
	sd	22.29	18.40	.						21.65
	n	7	3	1						7
IP_6	mean	254.33								212.96
	sd	.								.
	n	1								1

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
IP_1	mean	128.52	92.03	89.13	88.51	81.90	78.43	95.21		87.05
	sd	48.08	36.13	34.58	35.33	32.00	43.78	39.78		33.68
	n	425	304	184	85	36	7	7		443
IP_2	mean	124.14	100.72	97.31	94.49	96.34	86.84			90.33
	sd	40.17	40.69	36.32	34.50	35.74	32.09			29.09
	n	261	174	93	56	25	12			290
IP_3	mean	126.56	97.18	92.24	87.05	83.38	62.16			92.09
	sd	36.98	35.74	35.07	43.22	42.67	36.55			31.25
	n	98	62	21	9	5	4			104
IP_4	mean	126.26	93.16	71.57	109.58	105.27	98.00			96.25
	sd	43.40	34.99	29.64	26.76	45.84	.			31.16
	n	26	16	4	3	2	1			30
IP_5	mean	105.47	111.55	74.25						93.41
	sd	25.11	2.74	.						24.74
	n	7	3	1						7
IP_6	mean	105.85								101.39
	sd	.								.
	n	1								1

Table A.10: Delta f0 (Max f0 - Min f0) by IP and TG for analysis t01

Table A.11: Mean f0 by IP and TG for analysis t01

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
	mean	196.19	168.63	169.14	167.84	170.33	167.78	176.27		158.35
IP_1	sd	26.36	17.61	19.65	19.06	25.17	26.50	16.12		15.53
	n	425	304	184	85	36	7	7		443
	mean	195.75	173.30	171.17	168.47	167.86	167.18			160.25
IP_2	sd	21.48	18.98	18.18	20.83	19.46	17.42			13.95
	n	261	174	93	56	25	12			290
	mean	195.63	173.46	168.92	159.43	178.30	188.23			161.84
IP_3	sd	20.06	18.57	21.47	31.40	41.55	35.27			13.36
	n	98	62	21	9	5	4			104
	mean	190.26	175.03	180.88	161.59	169.24	161.44			162.01
IP_4	sd	20.97	21.34	14.12	20.33	.96	.			16.42
	n	26	16	4	3	2	1			30
	mean	185.06	166.15	144.61						154.59
IP_5	sd	17.86	4.52	.						5.44
	n	7	3	1						7
	mean	198.19								171.77
IP_6	sd	.								.
	n	1								1

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
IP_1	mean	36.55	25.21	24.70	25.22	24.22	19.23	26.84		23.48
	sd	14.21	10.46	10.31	11.30	10.00	9.29	12.88		9.67
	n	425	304	184	85	36	7	7		443
IP_2	mean	33.48	28.23	27.59	26.81	27.84	23.81			24.66
	sd	11.57	12.37	9.97	11.29	11.08	10.99			8.20
	n	261	174	93	56	25	12			290
IP_3	mean	33.50	27.76	27.58	24.02	21.95	20.28			25.61
	sd	10.39	10.45	11.09	12.60	11.32	15.61			8.76
	n	98	62	21	9	5	4			104
IP_4	mean	34.65	26.99	22.50	33.45	26.23	28.59			27.40
	sd	12.69	10.86	9.32	5.18	8.56	.			9.17
	n	26	16	4	3	2	1			30
IP_5	mean	31.76	32.60	21.40						25.33
	sd	7.61	1.78	.						5.33
	n	7	3	1						7
IP_6	mean	28.88								23.16
	sd	.								.
	n	1								1

Table A.12: Standard Deviation f0 by IP and TG for analysis t01

		TG_1	TG_2	TG_3	TG_4	TG_5	TG_6	TG_7	TG_8	TG_9
	mean	1.26	1.07	1.06	1.03	1.01	.92	.97		1.27
IP_1	sd	.42	.46	.44	.41	.37	.43	.42		.48
	n	425	304	184	85	36	7	7		443
	mean	1.29	1.07	1.04	1.07	.94	1.02			1.24
IP_2	sd	.46	.46	.38	.40	.34	.42			.45
	n	261	174	93	56	25	12			290
	mean	1.31	.94	.90	1.04	.86	.50			1.18
IP_3	sd	.44	.36	.36	.52	.13	.26			.41
	n	98	62	21	9	5	4			104
	mean	1.23	.89	.72	.82	1.05	1.77			1.20
IP_4	sd	.33	.31	.16	.16	.19	.			.41
	n	26	16	4	3	2	1			30
	mean	1.05	1.23	1.36						1.15
IP_5	sd	.26	.36	.						.51
	n	7	3	1						7
	mean	1.68								1.79
IP_6	sd	.								.
	n	1								1

Table A.13: TG duration by IP and TG for analysis t01

		TG_1	TG_2	TG_f
	mean	173.07	165.32	154.45
IP_1	sd	53.59	35.71	34.22
	n	318	211	400
	mean	172.97	166.66	153.58
IP_2	sd	42.58	41.77	31.58
	n	103	66	132
	mean	164.87	176.60	143.19
IP_3	sd	46.45	35.70	28.02
	n	23	14	29
	mean	154.18	182.35	145.26
IP_4	sd	36.61	20.33	35.28
	n	7	4	8
	mean	135.85	151.44	158.10
IP_5	sd	.	.	39.02
	n	1	1	2
	mean	169.57	164.91	155.46
IP_f	sd	44.67	35.27	32.75
	n	524	333	631

Table A.14: Start f0 by IP and TG for analysis t12

		TG_1	TG_2	TG_f
	mean	158.59	151.71	139.75
IP_1	sd	38.53	29.76	29.52
	n	318	211	400
	mean	158.51	151.20	136.78
IP_2	sd	38.72	33.48	28.94
	n	103	66	132
	mean	175.73	126.10	135.36
IP_3	sd	33.16	26.28	28.66
	n	23	14	29
	mean	144.04	120.83	138.81
IP_4	sd	32.05	10.52	26.77
	n	7	4	8
	mean	151.44	143.74	104.26
IP_5	sd	.	.	2.15
	n	1	1	2
	mean	160.44	150.43	131.92
IP_f	sd	35.17	32.26	32.92
	n	524	333	631

Table A.15: End f0 by IP and TG for analysis t12

		TG_1	TG_2	TG_f
	mean	128.60	123.19	114.54
$IP_?$	sd	27.18	19.37	13.79
	n	318	211	400
	mean	133.20	120.90	114.60
$IP_?$	sd	29.62	19.52	14.39
	n	103	66	132
	mean	137.87	118.75	108.10
$IP_?$	sd	32.92	22.18	7.06
	n	23	14	29
	mean	119.27	119.12	109.41
$IP_?$	sd	20.03	7.38	11.13
	n	7	4	8
	mean	129.88	112.96	106.45
$IP_?$	sd	.	.	5.24
	n	1	1	2
	mean	129.04	123.19	114.23
$IP_?$	sd	25.52	19.37	13.21
	n	524	333	631

Table A.16: Min f0 by IP and TG for analysis t12

		TG_1	TG_2	TG_f
	mean	250.77	211.99	205.10
IP_1	sd	46.86	31.29	34.37
	n	318	211	400
	mean	242.33	225.58	206.00
IP_2	sd	37.39	50.69	30.62
	n	103	66	132
	mean	257.69	229.24	206.94
IP_3	sd	25.31	35.88	28.24
	n	23	14	29
	mean	251.97	198.08	220.64
IP_4	sd	47.90	29.99	32.78
	n	7	4	8
	mean	274.97	226.59	182.41
IP_5	sd	.	.	4.64
	n	1	1	2
	mean	250.74	221.31	208.04
IP_f	sd	35.79	36.37	34.69
	n	524	333	631

Table A.17: Max f0 by IP and TG for analysis t12

		TG_1	TG_2	TG_f
IP_1	mean	122.17	88.80	90.56
	sd	50.00	33.41	35.28
	n	318	211	400
IP_2	mean	109.13	104.68	91.39
	sd	45.98	53.12	33.48
	n	103	66	132
IP_3	mean	119.81	110.49	98.84
	sd	40.87	44.36	25.54
	n	23	14	29
IP_4	mean	132.70	78.96	111.23
	sd	50.31	37.03	32.77
	n	7	4	8
IP_5	mean	145.10	113.62	75.96
	sd	.	.	.60
	n	1	1	2
IP_f	mean	121.70	98.12	93.81
	sd	40.53	39.07	34.83
	n	524	333	631

Table A.18: Delta (Max-Min) f0 by IP and TG for analysis t12

Second Analysis Summaries

IP	TG	start f0	end f0	min f0	max f0	delta f0	mean f0	sd f0
0	0							
	μ	176.65	128.70	110.54	249.81	139.26	174.95	36.43
	N	24	24	24	24	24	24	24
	σ	45.77	28.29	9.76	47.70	48.24	17.75	13.27
1	0							
	μ	178.36	131.21	111.94	233.25	121.30	171.53	31.72
	N	25	25	25	25	25	25	25
	σ	36.44	25.97	9.14	32.74	34.92	15.91	9.95
2	0							
	μ	172.39	132.04	111.98	258.03	146.05	175.14	39.49
	N	6	6	6	6	6	6	6
	σ	24.82	37.19	10.05	31.90	32.34	9.63	9.30
3	0							
	μ	150.16	118.34	107.29	211.30	104.00	156.66	28.14
	N	4	4	4	4	4	4	4
	σ	6.21	10.93	5.96	31.17	26.18	10.32	7.94

Table A.19: Summaries for IPs containing 1 TG.

IP	TG	start f0	end f0	min f0	max f0	delta f0	mean f0	sd f0
0	0							
	μ	190.65	147.64	122.30	253.45	131.15	190.99	36.08
	N	121	121	121	121	121	121	121
	σ	44.84	38.93	20.80	40.24	45.42	23.23	13.40
	1							
	μ	155.84	131.47	113.55	191.27	77.73	155.12	20.54
	N	121	121	121	121	121	121	121
	σ	33.27	30.88	13.77	27.40	26.56	16.18	7.64
1	0							
	μ	181.69	149.52	124.83	251.22	126.39	189.11	34.59
	N	87	87	87	87	87	87	87
	σ	38.64	34.45	22.22	26.16	33.21	19.41	9.86
	1							
	μ	162.15	120.53	113.04	201.31	88.27	160.13	24.22
	N	87	87	87	87	87	87	87
	σ	27.72	19.80	9.71	28.10	30.05	14.19	7.94
2	0							
	μ	173.39	157.32	125.47	250.13	124.66	188.38	33.53
	N	36	36	36	36	36	36	36
	σ	29.86	31.39	21.48	27.98	36.85	16.04	10.49
	1							
	μ	161.89	122.93	112.78	206.47	93.70	161.14	25.16
	N	36	36	36	36	36	36	36
	σ	30.40	23.45	11.58	30.60	30.32	15.40	8.98
3	0							
	μ	165.17	154.64	118.04	249.90	131.86	184.35	39.87
	N	8	8	8	8	8	8	8
	σ	27.86	42.71	16.78	27.13	24.48	15.53	9.65
	1							
	μ	167.94	127.07	109.03	203.20	94.17	159.49	25.91
	N	8	8	8	8	8	8	8
	σ	29.87	27.93	6.98	33.85	34.90	18.51	9.32
4	0							
	μ	159.75	140.72	133.53	245.99	112.46	178.38	32.98
	N	4	4	4	4	4	4	4
	σ	10.03	10.03	6.54	20.65	20.76	12.81	5.03
	1							
	μ	150.81	114.72	107.95	204.77	96.82	156.06	24.84
	N	4	4	4	4	4	4	4
	σ	18.97	15.04	9.46	28.38	32.94	2.03	5.11
5	0							
	μ	170.29	190.27	148.48	254.33	105.85	198.19	28.88
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	1							
	μ	161.62	111.57	111.57	212.96	101.39	171.77	23.16
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.

Table A.20: Summaries for IPs containing 2 TGs.

IP	TG	start f0	end f0	min f0	max f0	delta f0	mean f0	sd f0
0	0							
	μ	190.85	155.07	132.17	256.31	124.14	195.48	35.88
	N	120	120	120	120	120	120	120
	σ	46.22	34.95	27.09	45.22	51.52	27.64	15.13
	1							
	μ	169.22	145.26	119.13	213.29	94.16	168.58	25.64
	N	120	120	120	120	120	120	120
	σ	35.44	28.84	15.18	32.83	35.88	17.43	10.77
	2							
	μ	153.02	124.17	112.28	196.91	84.63	156.87	23.46
	N	120	120	120	120	120	120	120
	σ	28.00	23.10	10.43	30.61	33.03	14.95	9.88
1	0							
	μ	184.93	161.51	130.80	255.28	124.49	196.38	33.11
	N	84	84	84	84	84	84	84
	σ	33.39	37.46	27.44	27.29	41.22	17.04	11.31
	1							
	μ	165.93	151.51	121.62	223.43	101.81	171.25	28.98
	N	84	84	84	84	84	84	84
	σ	33.40	30.63	19.06	31.71	36.81	18.17	11.52
	2							
	μ	157.82	120.33	112.94	200.99	88.04	160.93	24.09
	N	84	84	84	84	84	84	84
	σ	26.49	16.03	10.02	22.31	24.39	12.59	7.32
2	0							
	μ	178.94	158.58	124.61	253.47	128.86	197.21	33.43
	N	39	39	39	39	39	39	39
	σ	30.86	36.41	23.47	27.71	38.58	21.54	11.08
	1							
	μ	160.63	137.61	119.35	218.30	98.95	170.92	27.64
	N	39	39	39	39	39	39	39
	σ	33.08	28.14	19.25	28.45	35.73	16.92	10.70
	2							
	μ	150.17	118.72	114.81	199.81	85.00	160.32	23.53
	N	39	39	39	39	39	39	39
	σ	23.48	14.30	10.89	24.10	23.59	11.83	6.76
3	0							
	μ	168.74	135.86	116.95	262.65	145.70	191.37	37.17
	N	11	11	11	11	11	11	11
	σ	34.31	32.23	17.91	31.94	41.81	17.71	10.63
	1							
	μ	165.96	137.77	122.18	210.36	88.18	169.00	25.88
	N	11	11	11	11	11	11	11
	σ	26.49	29.76	16.13	24.11	27.90	15.31	9.48
	2							
	μ	142.98	115.53	112.48	201.69	89.21	162.27	27.17
	N	11	11	11	11	11	11	11
	σ	32.28	11.99	10.95	22.48	25.14	12.63	9.02

Table A.21: Summaries for IPs containing 3 TGs.

IP	TG	start f0	end f0	min f0	max f0	delta f0	mean f0	sd f0
4	0							
	μ	165.40	165.99	136.09	225.69	89.60	184.60	26.01
	N	2	2	2	2	2	2	2
	σ	26.77	90.54	48.26	6.11	42.15	21.90	12.49
	1							
	μ	189.48	146.45	120.04	231.92	111.87	163.65	33.63
	N	2	2	2	2	2	2	2
	σ	57.32	33.63	21.77	25.56	3.79	1.75	.21
	2							
	μ	154.30	107.19	107.19	203.23	96.05	150.70	28.10
	N	2	2	2	2	2	2	2
	σ	29.81	.28	.28	.55	.28	11.08	7.93

Table A.22: Summaries for IPs containing 3 TGs (cont).

IP	TG	start f0	end f0	min f0	max f0	delta f0	mean f0	sd f0
0	0							
	μ	182.83	159.32	130.84	261.46	130.62	200.11	37.67
	N	99	99	99	99	99	99	99
	σ	46.31	44.03	27.35	37.87	45.34	27.60	13.38
	1							
	μ	166.88	142.18	120.11	212.55	92.44	168.46	25.43
	N	99	99	99	99	99	99	99
	σ	37.09	32.29	19.70	31.12	36.46	17.28	10.59
	2							
	μ	156.36	143.76	117.99	209.77	91.78	166.26	24.94
	N	99	99	99	99	99	99	99
	σ	31.35	30.95	16.86	32.08	35.97	17.63	10.57
	3							
	μ	153.39	126.64	111.91	199.63	87.72	158.59	23.70
	N	99	99	99	99	99	99	99
	σ	27.61	25.80	9.60	27.79	29.24	13.65	8.88
1	0							
	μ	188.93	163.87	135.55	250.21	114.65	200.20	31.45
	N	38	38	38	38	38	38	38
	σ	49.00	38.24	31.77	38.10	52.94	26.97	14.33
	1							
	μ	166.14	148.14	122.35	223.36	101.01	174.71	28.37
	N	38	38	38	38	38	38	38
	σ	37.80	37.66	17.06	38.40	45.90	17.23	12.84
	2							
	μ	160.74	150.41	124.71	227.23	102.53	172.18	30.23
	N	38	38	38	38	38	38	38
	σ	28.01	29.31	15.38	32.43	37.84	14.60	10.66
	3							
	μ	152.24	119.85	111.01	199.35	88.34	159.32	24.84
	N	38	38	38	38	38	38	38
	σ	30.90	17.95	8.97	22.19	22.72	11.61	6.70

Table A.23: Summaries for IPs containing 4 TGs.

IP	TG	start f0	end f0	min f0	max f0	delta f0	mean f0	sd f0
2	0							
	μ	191.45	150.07	123.44	253.49	130.04	199.69	33.61
	N	11	11	11	11	11	11	11
	σ	42.22	33.96	21.47	23.96	34.57	23.27	11.52
	1							
	μ	169.85	165.05	130.84	228.47	97.64	178.97	28.90
	N	11	11	11	11	11	11	11
	σ	32.46	38.67	25.24	27.90	36.86	22.35	9.75
	2							
	μ	166.46	142.14	116.39	209.61	93.22	164.94	28.02
	N	11	11	11	11	11	11	11
	σ	26.96	35.73	16.65	30.11	36.84	15.10	12.13
	3							
	μ	151.17	121.74	117.27	212.89	95.62	165.08	29.15
	N	11	11	11	11	11	11	11
	σ	26.53	21.35	16.59	29.81	33.25	10.93	8.74
3	0							
	μ	150.77	181.33	150.77	227.39	76.62	189.07	21.19
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	1							
	μ	181.33	200.72	162.32	234.70	72.38	198.23	23.02
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	2							
	μ	200.72	177.78	150.81	200.72	49.90	168.24	14.87
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	3							
	μ	163.88	107.03	107.03	175.75	68.73	153.58	17.73
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
4	0							
	μ	202.80	174.59	166.03	275.27	109.24	212.69	38.34
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	1							
	μ	174.59	115.04	115.04	225.93	110.89	171.17	30.55
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	2							
	μ	130.01	103.30	103.30	177.54	74.25	144.61	21.40
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.
	3							
	μ	182.85	108.34	108.34	182.85	74.52	156.50	21.77
	N	1	1	1	1	1	1	1
	σ	.	.	.	.	.	.	.

Table A.24: Summaries for IPs containing 4 TGs (cont).

Appendix B

Accent Analysis Summary Tables

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
a	1.00	1.00						
		Mean	187.84	216.27	182.07	.24	1.29	.14
		StdDev	48.42	41.25	37.46	1.49	.63	1.04
		N	560	560	560	560	560	560
	2.00	1.00						
		Mean	179.96	213.81	186.25	.00	1.27	.35
		StdDev	54.35	45.18	40.81	2.13	.85	.90
		N	820	820	820	820	820	820
		2.00						
		Mean	164.43	195.32	167.39	-.39	.86	-.15
		StdDev	52.31	32.13	29.75	2.06	.86	.88
		N	654	654	654	654	654	654
	3.00	1.00						
		Mean	181.82	213.97	190.78	.20	1.34	.56
		StdDev	49.47	43.85	38.29	1.79	.77	.84
		N	307	307	307	307	307	307
		2.00						
		Mean	163.55	198.37	173.22	-.35	.94	.08
		StdDev	58.28	34.02	29.05	2.03	.76	.85
		N	286	286	286	286	286	286
		3.00						
		Mean	153.89	185.44	158.48	-.64	.59	-.38
		StdDev	48.55	30.99	24.10	2.05	1.00	.81
		N	239	239	239	239	239	239
	4.00	1.00						
		Mean	175.54	212.89	188.40	.06	1.42	.55
		StdDev	46.53	46.90	46.02	1.71	.83	1.05
		N	60	60	60	60	60	60
		2.00						
		Mean	160.37	198.28	180.24	-.31	1.02	.38
		StdDev	51.92	32.75	30.11	1.48	.76	.86
		N	65	65	65	65	65	65
		3.00						
		Mean	150.93	182.27	161.69	-.68	.45	-.33
		StdDev	51.35	28.60	24.85	1.81	.80	.79
		N	68	68	68	68	68	68
		4.00						
		Mean	155.08	185.21	161.39	-.52	.81	-.09
		StdDev	49.42	20.94	21.29	2.34	.81	.98
		N	48	48	48	48	48	48

Table B.1: Accent summaries for sequentially numbered TGs.

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
	5.00	1.00						
		Mean	195.08	216.79	191.89	.60	1.41	.55
		StdDev	46.79	39.62	34.80	1.27	.90	.98
		N	9	9	9	9	9	9
		2.00						
		Mean	163.29	203.03	163.66	-.11	1.26	-.19
		StdDev	62.42	33.51	25.98	1.83	.96	1.06
		N	11	11	11	11	11	11
		3.00						
		Mean	168.85	194.11	177.71	-.09	.64	.05
		StdDev	22.99	32.16	35.65	.68	.53	.80
		N	13	13	13	13	13	13
		4.00						
		Mean	138.28	180.33	172.49	-1.11	.35	.09
		StdDev	65.98	29.42	24.49	2.65	.86	.70
		N	12	12	12	12	12	12
		5.00						
		Mean	168.65	190.13	167.08	-.36	.42	-.15
		StdDev	21.33	24.10	38.15	.61	.96	1.32
		N	4	4	4	4	4	4
	6.00	1.00						
		Mean	.00	254.62	218.61	-4.62	2.29	1.31
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		2.00						
		Mean	229.53	229.53	185.19	1.61	1.61	.40
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		3.00						
		Mean	165.32	220.55	132.71	-.13	1.36	-1.02
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		4.00						
		Mean	148.21	176.33	176.33	-.60	.16	.16
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		5.00						
		Mean	140.04	140.04	127.76	-.82	-.82	-1.15
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		6.00						
		Mean	173.70	173.70	161.62	.09	.09	-.24
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1

Table B.2: Accent summaries for sequentially numbered TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
afb	1.00	1.00						
		Mean	162.35	179.64	127.04	.36	1.28	-1.48
		StdDev	34.32	23.48	21.87	1.29	.54	.90
		N	50	50	50	50	50	50
	2.00	2.00						
		Mean	157.60	178.54	125.68	-.08	.88	-1.51
		StdDev	46.73	25.07	22.58	1.95	.79	1.04
		N	95	95	95	95	95	95
	3.00	3.00						
		Mean	161.53	179.61	129.15	-.17	.45	-1.39
		StdDev	38.19	29.49	15.16	1.16	.93	.60
		N	45	45	45	45	45	45
	4.00	2.00						
		Mean	163.10	166.32	170.21	.78	.97	1.21
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		4.00						
		Mean	166.74	183.75	132.34	.02	.70	-1.39
		StdDev	21.28	17.96	19.10	.91	.47	.59
		N	13	13	13	13	13	13
	5.00	5.00						
		Mean	112.68	180.79	131.58	-1.26	.69	-1.22
		StdDev	159.35	63.02	9.94	5.32	2.57	.35
		N	2	2	2	2	2	2
arb	1.00	1.00						
		Mean	158.75	177.97	193.13	-.47	.44	1.46
		StdDev	23.83	32.72	25.09	.49	.78	.79
		N	8	8	8	8	8	8
	2.00	2.00						
		Mean	158.68	182.69	188.88	-.55	.34	.69
		StdDev	15.60	26.61	17.62	.59	.96	1.32
		N	14	14	14	14	14	14
	3.00	3.00						
		Mean	167.67	176.85	196.43	-.45	.13	.60
		StdDev	24.37	46.37	20.21	.50	1.67	.82
		N	4	4	4	4	4	4

Table B.3: Accent summaries for sequentially numbered TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
fb	.00	.00						
		Mean	164.64	164.64	126.88	.98	.98	-1.67
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
	1.00	.00						
		Mean	77.16	123.67	118.04	-3.53	-1.60	-1.84
		StdDev	67.66	14.24	42.44	3.24	.95	1.75
		N	38	38	38	38	38	38
	2.00	.00						
		Mean	76.30	125.69	110.81	-3.81	-1.41	-2.62
		StdDev	70.68	15.50	46.81	4.02	.81	3.33
		N	68	68	68	68	68	68
	3.00	.00						
		Mean	71.26	129.02	114.20	-3.15	-1.61	-2.04
		StdDev	75.41	28.37	45.14	2.68	.92	1.41
		N	26	26	26	26	26	26
	4.00	.00						
		Mean	137.55	111.13	135.51	-.98	-2.03	-1.06
		StdDev	9.87	4.07	7.86	.53	.49	.51
		N	3	3	3	3	3	3
rb	.00	.00						
		Mean	138.14	178.63	178.52	-1.59	1.23	1.22
		StdDev	26.66	13.97	13.79	.84	.53	.52
		N	4	4	4	4	4	4
	1.00	.00						
		Mean	117.05	175.58	181.75	-2.46	.04	.22
		StdDev	75.81	33.66	22.95	3.44	1.41	.98
		N	110	110	110	110	110	110
	2.00	.00						
		Mean	104.03	165.71	175.26	-2.98	-.19	.23
		StdDev	75.36	34.23	22.39	3.91	1.61	1.01
		N	116	116	116	116	116	116
	3.00	.00						
		Mean	91.03	164.26	173.94	-2.81	-.40	.04
		StdDev	71.84	43.63	31.74	2.68	1.51	.88
		N	46	46	46	46	46	46
	4.00	.00						
		Mean	61.97	140.93	156.58	-3.70	-.92	-.44
		StdDev	75.51	30.12	18.44	3.07	1.68	1.17
		N	14	14	14	14	14	14
	5.00	.00						
		Mean	62.11	136.19	152.28	-4.44	-1.35	-.93
		StdDev	85.07	27.20	27.18	4.17	1.11	1.17
		N	5	5	5	5	5	5

Table B.4: Accent summaries for sequentially numbered TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
a	1.00	1.00						
		Mean	218.64	249.27	205.49	.39	1.30	.01
		StdDev	42.10	40.35	42.53	.88	.57	1.22
		N	166	166	166	166	166	166
	2.00	1.00						
		Mean	214.69	250.01	216.85	.46	1.38	.54
		StdDev	45.31	45.88	44.20	1.02	1.09	.90
		N	234	234	234	234	234	234
		2.00						
		Mean	172.23	207.92	181.00	-.73	.39	-.37
		StdDev	60.51	34.12	32.77	1.84	.85	.79
		N	185	185	185	185	185	185
	3.00	1.00						
		Mean	210.07	245.36	217.66	.57	1.56	.78
		StdDev	39.63	41.34	37.60	.99	.74	.87
		N	94	94	94	94	94	94
		2.00						
		Mean	175.59	211.17	183.34	-.28	.74	-.05
		StdDev	56.85	33.98	28.53	1.48	.70	.71
		N	91	91	91	91	91	91
		3.00						
		Mean	151.78	192.01	162.51	-1.05	.30	-.59
		StdDev	57.39	34.06	26.46	2.03	1.00	.76
		N	79	79	79	79	79	79
	4.00	1.00						
		Mean	214.31	259.11	232.33	.60	1.68	1.03
		StdDev	43.60	41.58	49.07	1.09	.61	.83
		N	12	12	12	12	12	12
		2.00						
		Mean	161.23	219.38	202.28	-.68	.91	.49
		StdDev	68.46	39.27	37.62	1.91	.84	1.03
		N	16	16	16	16	16	16
		3.00						
		Mean	150.99	199.62	174.89	-1.01	.43	-.27
		StdDev	61.77	22.98	21.42	1.90	.61	.65
		N	19	19	19	19	19	19
		4.00						
		Mean	182.44	197.77	156.09	.14	.57	-.70
		StdDev	36.12	29.75	24.09	1.24	1.17	.90
		N	12	12	12	12	12	12

Table B.5: Accent summaries for accents in initial TGs.

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
	5.00	1.00						
		Mean	223.91	270.63	202.66	1.64	2.98	1.03
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		3.00						
		Mean	177.43	199.80	171.04	-.03	.66	-.16
		StdDev	7.03	19.15	15.90	.56	.42	.18
		N	3	3	3	3	3	3
		4.00						
		Mean	156.92	164.31	163.17	-.67	-.40	-.45
		StdDev	9.61	16.04	14.42	.28	.02	.08
		N	2	2	2	2	2	2
afb	2.00	2.00						
		Mean	178.09	182.61	129.55	.25	.43	-1.51
		StdDev	26.75	38.09	14.04	.98	1.24	.37
		N	6	6	6	6	6	6
	3.00	3.00						
		Mean	157.56	190.56	131.92	-.96	-.20	-1.64
		StdDev	61.76	42.03	17.61	1.38	.84	.75
		N	12	12	12	12	12	12
	4.00	4.00						
		Mean	191.52	200.06	144.47	.44	.82	-1.21
		StdDev	5.92	9.27	35.45	.40	.75	1.21
		N	3	3	3	3	3	3
arb	2.00	2.00						
		Mean	165.46	190.18	192.66	-.60	.03	.09
		StdDev	9.24	24.63	19.48	.56	.69	.60
		N	9	9	9	9	9	9
	3.00	3.00						
		Mean	167.67	176.85	196.43	-.45	.13	.60
		StdDev	24.37	46.37	20.21	.50	1.67	.82
		N	4	4	4	4	4	4
fb	1.00	.00						
		Mean	62.87	129.50	121.64	-2.57	-1.26	-1.44
		StdDev	72.94	10.94	16.00	1.55	.24	.21
		N	4	4	4	4	4	4
	2.00	.00						
		Mean	69.57	121.42	140.73	-3.17	-1.76	-1.28
		StdDev	74.67	5.36	26.86	2.28	.46	.61
		N	8	8	8	8	8	8
	3.00	.00						
		Mean	94.77	114.23	130.72	-2.56	-2.17	-1.67
		StdDev	77.99	7.91	19.94	1.93	.59	.43
		N	6	6	6	6	6	6

Table B.6: Accent summaries for accents in initial TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
rb	1.00	.00						
		Mean	140.59	182.40	190.37	-1.78	-.31	-.09
		StdDev	72.13	32.73	17.74	2.53	1.31	.76
		N	35	35	35	35	35	35
	2.00	.00						
		Mean	111.97	169.05	178.28	-2.55	-.33	.04
		StdDev	75.53	34.67	22.88	3.46	1.50	.96
		N	51	51	51	51	51	51
	3.00	.00						
		Mean	89.49	172.55	185.40	-2.67	-.70	-.13
		StdDev	74.45	55.30	38.99	2.14	1.58	.89
		N	18	18	18	18	18	18
	4.00	.00						
		Mean	64.85	137.77	151.80	-4.15	-1.27	-.69
		StdDev	76.51	30.28	19.10	4.08	1.44	.86
		N	4	4	4	4	4	4
	5.00	.00						
		Mean	103.52	132.58	165.23	-2.05	-1.43	-.37
		StdDev	89.69	27.69	4.50	2.27	1.31	.24
		N	3	3	3	3	3	3

Table B.7: Accent summaries for accents in initial TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
a	1.00	1.00						
		Mean	176.47	206.67	176.02	.04	1.29	.20
		StdDev	48.43	34.36	31.41	1.90	.72	1.00
		N	262	262	262	262	262	262
	2.00	1.00						
		Mean	167.17	203.27	179.88	-.27	1.22	.35
		StdDev	55.17	35.49	32.82	2.34	.64	.88
		N	328	328	328	328	328	328
		2.00						
		Mean	169.75	196.85	165.31	-.06	1.03	-.11
		StdDev	45.44	32.97	27.63	1.53	.87	.93
		N	265	265	265	265	265	265
	3.00	1.00						
		Mean	160.61	198.51	179.77	-.32	1.17	.45
		StdDev	56.75	35.91	34.20	2.28	.83	.88
		N	97	97	97	97	97	97
		2.00						
		Mean	161.60	191.84	170.95	-.34	.99	.23
		StdDev	57.02	31.04	27.15	2.27	.69	.85
		N	94	94	94	94	94	94
		3.00						
		Mean	155.84	184.57	157.36	-.46	.72	-.31
		StdDev	43.06	32.03	25.77	2.18	1.06	.92
		N	81	81	81	81	81	81
	4.00	1.00						
		Mean	174.82	208.24	180.38	.25	1.50	.50
		StdDev	38.01	41.04	37.03	1.16	.90	1.12
		N	28	28	28	28	28	28
		2.00						
		Mean	163.68	193.03	171.49	-.14	1.10	.28
		StdDev	50.22	26.02	20.82	1.39	.54	.86
		N	26	26	26	26	26	26
		3.00						
		Mean	144.97	176.69	162.86	-.73	.42	-.14
		StdDev	47.83	27.52	22.76	1.65	.85	.80
		N	27	27	27	27	27	27
		4.00						
		Mean	152.05	183.23	165.39	-.88	.89	.14
		StdDev	43.81	16.88	21.14	2.97	.54	.91
		N	20	20	20	20	20	20

Table B.8: Accent summaries for accents in medial TGs.

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
	5.00	1.00						
		Mean	206.12	225.57	195.59	.72	1.51	.35
		StdDev	46.41	39.13	51.24	1.06	.65	1.47
		N	4	4	4	4	4	4
		2.00						
		Mean	130.64	206.33	161.62	-1.10	1.51	-.24
		StdDev	66.33	42.00	29.55	1.97	1.23	1.34
		N	6	6	6	6	6	6
		3.00						
		Mean	171.61	188.52	177.45	.09	.52	.04
		StdDev	24.42	40.65	44.91	.58	.62	.98
		N	7	7	7	7	7	7
		4.00						
		Mean	120.85	186.30	185.79	-1.70	.55	.60
		StdDev	83.38	37.07	22.72	3.40	.99	.37
		N	7	7	7	7	7	7
		5.00						
		Mean	173.92	198.83	168.24	-.23	.68	-.08
		StdDev	22.72	20.43	46.63	.68	.98	1.61
		N	3	3	3	3	3	3
afb	1.00	1.00						
		Mean	177.47	203.95	140.99	-.01	1.53	-.89
		StdDev	41.64	28.08	16.04	.93	.35	.75
		N	7	7	7	7	7	7
	2.00	2.00						
		Mean	155.11	195.08	132.47	-.71	1.04	-1.34
		StdDev	71.29	23.24	16.92	3.21	.60	.52
		N	21	21	21	21	21	21
	3.00	3.00						
		Mean	173.28	187.54	125.56	.20	.99	-1.73
		StdDev	32.88	20.54	11.58	1.17	.87	.52
		N	8	8	8	8	8	8
	4.00	4.00						
		Mean	142.13	179.10	122.06	-.73	.50	-1.60
		StdDev	5.41	41.09	13.51	.47	.88	.38
		N	2	2	2	2	2	2
arb	1.00	1.00						
		Mean	162.67	184.42	200.18	-.56	.45	1.43
		StdDev	22.78	29.34	16.46	.45	.84	.85
		N	7	7	7	7	7	7
	2.00	2.00						
		Mean	146.38	174.80	182.29	-.69	.99	1.36
		StdDev	20.99	27.77	14.63	.51	1.37	1.55
		N	4	4	4	4	4	4

Table B.9: Accent summaries for accents in medial TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
fb	.00	.00						
		Mean	164.64	164.64	126.88	.98	.98	-1.67
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
	1.00	.00						
		Mean	85.74	119.34	134.75	-3.42	-1.82	-1.19
		StdDev	71.40	10.44	35.49	3.52	1.12	.81
		N	13	13	13	13	13	13
	2.00	.00						
		Mean	58.87	130.66	131.31	-4.48	-1.40	-1.38
		StdDev	71.87	21.80	18.11	3.82	.95	.81
		N	19	19	19	19	19	19
	3.00	.00						
		Mean	45.90	156.14	129.04	-5.65	-.91	-1.95
		StdDev	79.50	60.84	21.45	3.71	1.69	.33
		N	3	3	3	3	3	3
	4.00	.00						
		Mean	136.76	111.99	134.72	-1.29	-2.04	-1.35
		StdDev	13.83	5.35	10.94	.05	.69	.05
		N	2	2	2	2	2	2
rb	.00	.00						
		Mean	138.14	178.63	178.52	-1.59	1.23	1.22
		StdDev	26.66	13.97	13.79	.84	.53	.52
		N	4	4	4	4	4	4
	1.00	.00						
		Mean	107.35	175.84	180.24	-2.57	.30	.40
		StdDev	74.27	32.39	22.79	3.52	1.38	1.04
		N	65	65	65	65	65	65
	2.00	.00						
		Mean	98.45	168.74	175.83	-3.50	.09	.42
		StdDev	77.16	33.63	21.58	4.43	1.64	1.04
		N	50	50	50	50	50	50
	3.00	.00						
		Mean	115.91	159.15	165.98	-1.94	-.32	-.01
		StdDev	61.95	32.99	23.10	2.56	1.43	.81
		N	20	20	20	20	20	20
	4.00	.00						
		Mean	68.05	143.43	163.14	-3.01	-.90	-.41
		StdDev	85.35	34.77	18.92	2.71	2.03	1.47
		N	7	7	7	7	7	7
	5.00	.00						
		Mean	.00	141.60	132.84	-8.02	-1.24	-1.77
		StdDev	.00	36.44	40.70	4.04	1.21	1.73
		N	2	2	2	2	2	2

Table B.10: Accent summaries for accents in medial TGs.

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
a	1.00	1.00						
		Mean	171.58	193.91	164.83	.45	1.30	.19
		StdDev	37.24	28.11	25.86	1.07	.51	.84
		N	131	131	131	131	131	131
	2.00	1.00						
		Mean	162.37	191.75	165.29	-.13	1.21	.17
		StdDev	45.56	33.01	28.71	2.57	.83	.88
		N	242	242	242	242	242	242
		2.00						
		Mean	150.73	181.50	157.94	-.52	1.08	.01
		StdDev	49.95	22.70	24.85	2.73	.69	.87
		N	193	193	193	193	193	193
	3.00	1.00						
		Mean	171.87	193.24	172.86	.28	1.22	.43
		StdDev	35.78	30.66	23.42	1.86	.65	.73
		N	98	98	98	98	98	98
		2.00						
		Mean	150.80	189.64	164.50	-.52	1.07	.08
		StdDev	62.35	32.20	29.39	2.36	.84	.99
		N	86	86	86	86	86	86
		3.00						
		Mean	154.74	181.53	154.19	-.38	.89	-.24
		StdDev	50.26	25.88	17.09	2.14	.81	.64
		N	60	60	60	60	60	60
	4.00	1.00						
		Mean	149.06	187.35	171.26	-.60	1.13	.35
		StdDev	46.06	38.58	42.82	2.54	.81	1.08
		N	18	18	18	18	18	18
		2.00						
		Mean	159.07	186.12	175.34	-.07	1.03	.53
		StdDev	23.58	31.15	28.25	.81	1.01	.72
		N	19	19	19	19	19	19
		3.00						
		Mean	150.49	168.68	145.48	-.52	.44	-.67
		StdDev	45.72	26.71	21.53	2.00	.92	.81
		N	19	19	19	19	19	19
		4.00						
		Mean	136.44	179.02	161.26	-.60	.96	.13
		StdDev	59.45	13.59	19.93	2.14	.78	1.02
		N	15	15	15	15	15	15

Table B.11: Accent summaries for accents in final TGs .

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
	5.00	1.00						
		Mean	159.58	181.42	175.50	-.32	.55	.37
		StdDev	49.16	18.81	11.75	1.48	.19	.11
		N	3	3	3	3	3	3
		2.00						
		Mean	197.47	198.31	157.61	1.07	1.10	-.25
		StdDev	34.60	33.47	28.53	.65	.60	1.01
		N	3	3	3	3	3	3
		3.00						
		Mean	146.77	197.99	173.27	-.59	.87	.01
		StdDev	37.59	34.84	39.14	1.42	.64	1.00
		N	2	2	2	2	2	2
		4.00						
		Mean	159.49	175.34	146.32	-.01	.62	-.54
		StdDev	24.03	14.43	4.55	1.14	.75	.00
		N	2	2	2	2	2	2
	6.00	1.00						
		Mean	.00	254.62	218.61	-4.62	2.29	1.31
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		2.00						
		Mean	229.53	229.53	185.19	1.61	1.61	.40
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		3.00						
		Mean	165.32	220.55	132.71	-.13	1.36	-1.02
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		4.00						
		Mean	148.21	176.33	176.33	-.60	.16	.16
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		5.00						
		Mean	140.04	140.04	127.76	-.82	-.82	-1.15
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		6.00						
		Mean	173.70	173.70	161.62	.09	.09	-.24
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1

Table B.12: Accent summaries for accents in final TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
afb	1.00	1.00						
		Mean	159.89	175.68	124.77	.42	1.24	-1.58
		StdDev	32.90	20.39	21.99	1.34	.56	.90
		N	43	43	43	43	43	43
	2.00	2.00						
		Mean	156.81	173.83	123.34	.11	.92	-1.56
		StdDev	38.55	21.99	25.29	1.42	.74	1.23
		N	63	63	63	63	63	63
	3.00	3.00						
		Mean	162.22	170.70	127.76	.25	.70	-1.19
		StdDev	23.45	22.12	8.93	.84	.84	.28
		N	21	21	21	21	21	21
	4.00	2.00						
		Mean	163.10	166.32	170.21	.78	.97	1.21
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
		4.00						
		Mean	164.79	177.01	128.18	.22	.78	-1.37
		StdDev	18.82	11.50	11.03	.99	.24	.39
		N	7	7	7	7	7	7
	5.00	5.00						
		Mean	112.68	180.79	131.58	-1.26	.69	-1.22
		StdDev	159.35	63.02	9.94	5.32	2.57	.35
		N	2	2	2	2	2	2
arb	1.00	1.00						
		Mean	131.29	132.83	143.81	.17	.35	1.62
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
	2.00	2.00						
		Mean	146.87	146.87	181.25	.50	.50	3.44
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1

Table B.13: Accent summaries for accents in final TGs (cont).

name	count	num	st_f0	pk_f0	en_f0	st_f0n	pk_f0n	en_f0n
fb	1.00	.00						
		Mean	74.58	125.24	107.01	-3.79	-1.52	-2.31
		StdDev	67.20	16.45	47.19	3.38	.92	2.17
		N	21	21	21	21	21	21
	2.00	.00						
		Mean	87.24	124.32	94.09	-3.54	-1.31	-3.55
		StdDev	69.76	13.10	53.68	4.41	.80	4.13
		N	39	39	39	39	39	39
	3.00	.00						
		Mean	74.29	124.66	94.75	-2.75	-1.72	-2.56
		StdDev	80.94	20.29	58.42	2.96	.77	1.91
		N	12	12	12	12	12	12
	4.00	.00						
		Mean	139.13	109.42	137.09	-.37	-2.00	-.48
		StdDev	.	.	.	.	.	.
		N	1	1	1	1	1	1
rb	1.00	.00						
		Mean	97.74	150.00	161.37	-4.09	-.44	.17
		StdDev	86.52	36.16	27.04	5.09	1.68	1.11
		N	10	10	10	10	10	10
	2.00	.00						
		Mean	92.74	142.34	162.00	-2.88	-.68	.19
		StdDev	72.58	28.48	20.96	3.61	1.90	1.04
		N	14	14	14	14	14	14
	3.00	.00						
		Mean	36.87	156.61	168.28	-5.26	.04	.55
		StdDev	63.03	42.44	30.16	3.11	1.71	1.02
		N	7	7	7	7	7	7
	4.00	.00						
		Mean	43.94	139.29	147.63	-4.71	-.51	-.20
		StdDev	76.11	29.16	16.52	3.27	1.51	1.03
		N	3	3	3	3	3	3

Table B.14: Accent summaries for accents in final TGs (cont).

APPENDIX C

Features Used in the Linear Regression Models

Feature descriptions:

tgs Initial TG
tgm Medial TG
tge Final TG
accent_count The number of an accent in the TG.
hstar H* accent on syllable
!hstar !H* accent on syllable
lstar L* accent on syllable
lstarplush L*+H accent on syllable
lstarplush+ L*+H or L*+!H accent on syllable
accent_other Other ToBI accents not matching any of those specified above.
a Tilt style accent.
afb Tilt style accent and falling boundary
arb Tilt style accent and rising boundary
fb Tilt style falling boundary
rb Tilt style rising boundary
lminuslpc L-L% boundary tone
lminus L- boundary tone
lminushpc L-H% boundary tone
hminus H- boundary tone
hminushpc H-H% boundary tone

hminushpc+ H-H% or !H- boundary tone

endtone_other Other ToBI boundaries not matching those specified above.

syl_break The strength of the break after this syllable.

old_syl_break The strength of the break after this syllable.

stress Syllable is stressed

syl_in Number of syllables since last phrase break.

syl_out Number of syllables before next phrase break.

ssyl_in Number of stressed syllables since last phrase break.

ssyl_out Number of stressed syllables before next phrase break.

asyl_in Number of accented syllables since last phrase break.

asyl_out Number of accented syllables before next phrase break.

last_accent Number of syllables since last accented syllable.

next_accent Number of syllables before next accented syllable.

sub_phrases Number of minor phrase breaks since the last major phrase break.

Features used in each model

Description	Default Festival	Context LR	Context ToBI
TG context:		tgs	tgs
		tgm	tgm
		tge	tge
		accent_count	accent_count
Accents:	hstar	a	hstar
	!hstar		
	lstar		lstar
	lstarplush+	afb	lstarplush
	accent_other	arb	
Boundaries:	lminus		lminus
	hminus		hminus
	lminuslpc	fb	lminuslpc
	lminushpc	rb	lminushpc
	hminushpc+		hminushpc
	endtone_other		hminuslpc
Misc:	old_syl_break	syl_break	syl_break
	stress	stress	stress
	syl_in	syl_in	syl_in
	syl_out	syl_out	syl_out
	ssyl_in	ssyl_in	ssyl_in
	ssyl_out	ssyl_out	ssyl_out
	asyl_in	asyl_in	asyl_in
	asyl_out	asyl_out	asyl_out
	last_accent	last_accent	last_accent
	next_accent	next_accent	next_accent
	sub_phrases	sub_phrases	sub_phrases

Table C.1: Table showing which features are used in which models. TG context, accent and boundary features, syl_break and stress features are duplicated for 'previous previous', 'previous', 'next' and 'next next' syllables. The remaining features are not.

Appendix D

Adjusted Parameters in the Altered ToBI Model

	accented syllable			next syllable		
	s	m	e	s	m	e
Original Model	0	-27	-26	-14	0	0
Adjusted Parameter	-5	-66	-43	-14	0	0

Table D.1: Parameter adjustments for L*

	accented syllable			next syllable		
	s	m	e	s	m	e
Original Model	0	0	0	0	0	0
Adjusted Parameter	-29	-40	43	64	66	128

Table D.2: Parameter adjustments for L*+H

	accented syllable			next syllable		
	s	m	e	s	m	e
Original Model	11	13	10	0	0	0
Adjusted Parameter	12	21	7	11	6	0

Table D.3: Parameter adjustments for L+H*

	accented syllable			next syllable		
	s	m	e	s	m	e
Original Model	0	-34	-33	0	0	0
Adjusted Parameter	-13	-35	-36	-1	0	0

Table D.4: Parameter adjustments for L-L%

	accented syllable			next syllable		
	s	m	e	s	m	e
Original Model	0	-20	-5	0	0	0
Adjusted Parameter	-20	0	40	0	0	0

Table D.5: Parameter adjustments for L-H%

References

Allen, J. (1987), *From Text to Speech: The MITalk System*, Cambridge University Press, Cambridge, UK.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. & Weinert, R. (1991), 'The HCRC map task corpus', *Language and Speech* **34**, 351–366.

Arvaniti, A., Ladd, D. R. & Mennen, I. (1998), 'Stability of tonal alignment: The case of Greek prenuclear accents', *Journal of Phonetics* **26**(1), 3–25.

A.Taylor, P. (1995), Using neural networks to locate pitch accents, *in* 'Eurospeech '95', Madrid.

Beckman, M. E. & Pierrehumbert, J. B. (1986), Intonational structure in Japanese and English, *in* C. Ewen & J. Anderson, eds, 'Phonology Yearbook', Vol. 3, Cambridge University Press, Cambridge, UK, pp. 255–309.

Black, A. & Hunt, A. (1996), Generating f0 contours from ToBI labels using linear regression, *in* 'ICSLP 96', Philadelphia, Penn.

Bolinger, D. (1958), 'A theory of pitch accent in English', *Word* **14**, 109–149.

Bolinger, D. (1961), 'Contrastive accent and contrastive stress', *Language* **37**, 83–96.

Breiman, L., Friedman, J. H., Olshen, J. A. & Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Breiman, L., Friedman, J. H., Olshen, J. A. & Stone, C. J. (1993), *Classification and Regression Trees*, London : Chapman & Hall.

Bruce, G. (1977), *Swedish Word Accents in Sentence Perspective*, Lund: Gleerup.

Bruce, G. & Gårding, E. (1978), A prosodic typology for Swedish dialects, *in* 'Nordic Prosody', Gleerup, pp. 219–228.

Campbell, N. & Venditti, J. J. (1995), J-ToBI: An intonation labelling system for Japanese, *in* 'Proceedings of the Autumn meeting of the Acoustical Society of Japan', Vol. 1, pp. 317–318.

Carroll, L. (2000), *The Annotated Alice: The Definitive Edition*, Norton, New York. annotated by: Gardener, M.

Clark, R. A. J. & Dusterhoff, K. E. (1999), Objective methods for evaluating synthetic intonation, *in* 'Eurospeech 1999', Vol. 4, pp. 1623–1626.

Cohen, A. & 't Hart, J. (1967), 'On the anatomy of intonation', *Lingua* **19**, 177–192.

Coleman, J. S. (1990), YorkTalk: "synthesis-by-rule" without segments or rules, *in* 'Proceedings of the ESCA Workshop on Speech Synthesis'.

Cooper, W. & Sorensen, J. (1981), *Fundemental Frequency in Sentence Production*, Heidelberg : Springer.

Cruttenden, A. (1997), *Intonation*, Cambridge University Press, Cambridge, UK.

Cruz-Neira, C., Sandin, D., DeFanti, T., Kenyon, R. & Hart, J. (1992), 'The CAVE: Audio visual experience automatic virtual environment', *Communications of the ACM* **35**(6), 65–72.

Crystal, D. (1969), *Prosodic Systems and Intonation in English*, Cambridge University Press, Cambridge, UK.

de Pijper, J. (1983), *Modelling British English Intonation*, Foris, Dordrecht.

Dogil, G. & Möbius, B. (2001*a*), 'Toward a perception based model of the production of prosody', *J. Acoust. Soc. Am.* **110**(5), 2737.

Dogil, G. & Möbius, B. (2001*b*), Towards a model of target oriented production of prosody, *in* 'Proceedings of the European Conference on Speech Communication and Technology', Vol. 1, Aalborg, Denmark, pp. 665–668.

Fujisaki, H. (1983), Dynamic characteristics of voice fundamental frequency in speech and singing., *in* 'The production of speech', Heidelberg: Springer-Verlag, pp. 39–55.

Garofolo, J. S. (1988), *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Giegerich, H. J. (1980), 'On stress-timing in English phonology', *Lingua* **51**, 187–221.

Godfrey, J., Holliman, E. & McDaniel, J. (1992), Telephone speech corpus for research and development, *in* 'The International Conference on Acoustics, Speech, and Signal Processing', Vol. 1, pp. 517–520.

Grabe, E. (1998), Intonational Phonology: English and German, PhD thesis, Max-Planck-Institut for Psycholinguistics and University of Nijmegen.

Grønnum, N. (1992), The Groundworks of Danish Intonation: an introduction, PhD thesis, University of Copenhagen/Museum Tusculanum Press.

Halliday, M. A. K. (1967), *Intonation and Grammar in British English*, Mouton.

Hirst, D., Di Christo, A. & Espesser, R. (2000), Levels of representation and levels of analysis for the description of intonation systems, *in* M. Horne, ed., 'Prosody: Theory and Experiment', KAP, pp. 37–88.

Hitzeman, J., Black, A. W., Taylor, P., Mellish, C. & Oberlander, J. (1999), An annotation scheme for concept-to-speech synthesis, *in* 'European Workshop on Natural Language Generation', pp. 59–66.

Hockett, C. (1958), *A Course in Modern Linguistics*, Macmillan.

Holm, B. & Bailly, G. (2002), Learning the hidden structure of intonation: Implementing various functions of prosody, *in* 'Speech Prosody 2002', Aix-en-Provence, France.

Holmes, J. N., Mattingly, I. G. & Shearme, J. N. (1964), 'Speech synthesis by rule', *Language and Speech* **7**, 127–143.

Jilka, M., Möhler, G. & Dogil, G. (1997), 'Rules for the generation of ToBI-based American English intonation', *Speech Communication* **28**, 83–108.

Karaali, O., Corrigan, G. & Gerson, I. (1996), Speech synthesis with neural networks, *in* 'World Congress on Neural Networks', San Diego, pp. 45–50.

Ladd, D. (1983), 'Phonological features of intonational peaks', *Language* **59**, 721–759.

Ladd, D. R. (1986), 'Intonational phrasing: The case for recursive prosodic structure', *Phonology* **3**, 311–340.

Ladd, D. R. (1988), 'Declination "reset" and the hierarchical organization of utterances', *J. Acoust. Soc. Am.* **84**(2), 530–544.

Ladd, D. R. (1990), Metrical representation of pitch register, *in* J. Kingston & M. Beckman, eds, 'Between the Grammar and the Physics of Speech', number 1 *in* 'Papers in Laboratory Phonology', Cambridge University Press, Cambridge, UK, pp. 35–57.

Ladd, D. R. (1996), *Intonational Phonology*, Cambridge University Press, Cambridge, UK.

Lawrence, W. (1953), The synthesis of speech from signals which have a low information rate, *in* W. Jackson, ed., 'Communication Theory', London: Butterworth Scientific Publications, pp. 460–471.

Levelt, W. (1989), *Speaking: From Intention to Articulation*, MIT Press, Cambidge, MA.

Liberman, M. (1975), The Intonational System of English, PhD thesis, MIT. Distributed 1978 by IULC.

Liberman, M. & Pierrehumbert, J. (1984), Intonational invariance under changes in pitch range and length, *in* M. Aronoff & R. Oehrle, eds, 'Language Sound Structure', MIT Press, Cambidge, MA, pp. 157–233.

Lieberman, P. (1967), *Intonation, Perception and Language.*, Cambridge, MA : MIT Press.

M-PIRO (2000), 'Multilingual personalised information objects', IST-1999-10982. http://www.ltg.ed.ac.uk/mpiro.

MagiCster (2002), 'Embodied believable agents', IST-1999-29078. http://www.ltg.ed.ac.uk/magicster.

Mayo, C., Aylett, M. & Ladd, D. (1997), Prosodic transcription of Glasgow English: An evaluation study of GlaToBI, *in* 'Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications', pp. 231–234.

Möbius, B. & van Santen, J. (2000), Phonetically motivated modeling of prosody, *in* 'Prosody 2000: Speech Recognition and Synthesis', Kraków, Poland, pp. 161–166.

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

Ostendorf, M., Price, P. J. & Shattuck-Hufnagel, S. (1995), The Boston University Radio News Corpus, Technical Report ECS-95-001, Boston University, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA.

Page, J. H. & Breen, A. P. (1996), 'The Laureate text-to-speech system — architecture and applications', *BT Technology Journal* **14**(1), 57–67.

Patterson, D. (2000), A Linguistic Approach to Pitch Range Modelling, PhD thesis, University of Edinburgh.

Pierrehumbert, J. (1981), 'Synthesizing intonation', *J. Acoust. Soc. Am.* **70**(2), 985–995.

Pierrehumbert, J. B. (1980), The Phonology and Phonetics of English Intonation, PhD thesis, MIT.

Pierrehumbert, J. & Hirschberg, J. (1990), The meaning of intonational contours in the interpretation of discourse, *in* P. R. Cohen, J. Morgan & M. E. Pollack, eds, 'Intentions in Communication', MIT Press, Cambidge, MA, chapter 14, pp. 271–311.

Pike, K. (1945), *The Intonation of American English*, Ann Arbor : University of Michigan press.

Prevost, S. & Steedman, M. (1994), 'Specifying intonation from context for speech synthesis', *Speech Communication* **15**, 139–153.

Reyelt, M., Grice, M., Benzmüller, R., Mayer, J. & Batliner, A. (1996), Prosodische etikettierung des deutschen mit ToBI, *in* 'Natural Language and Speech Technology, Results of the third KONVENS conference', pp. 144–155.

Rhetorical Systems Ltd. (2000), 'The rVoice Text-to-Speech engine'. http://www.rhetorical.com/rvoice.html.

Riley, M. (1992), Tree-based modelling of segmental durations, *in* G. Bailly, C. Benoit & T. R. Sawallis, eds, 'Talking Machines', New York: Elsevier Science Publishers, pp. 265–274.

Santen, J. V., Shih, C. & Möbius, B. (1998), Intonation, *in* R. Sproat, ed., 'Multilingual Text-to-Speech Synthesis: The Bell Labs Approach', Kluwer Academic Press, chapter 6, pp. 141–171.

Scott, D. R. (1982), 'Duration as a cue to the perception of phrase boundary', *J. Acoust. Soc. Am.* **71**, 996–1007.

Selkirk, E. (1984), *Phonology and Syntax : The Relation between Sound and Structure*, Current studies in linguistics series, MIT Press.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), ToBI: A standard for labeling English prosody, *in* 'Proceedings of the 1992 International Conference on Spoken Language Processing', pp. 867–870.

Steedman, M. (2000), 'Information structure and the syntax-phonology interface', *Linguistic Enquiry* **31**(4), 649–689.

Steedman, M. (2002), 'Information-structural semantics for English intonation', LSA Summer Institute Workshop on Topic and Focus, Santa Barbara July 2001, Draft 2, May.

Sun, X. (2001), Predicting underlying pitch targets for intonation modeling, *in* '4th ISCA workshop on Speech Synthesis', pp. 143–147.

Syrdal, A. K. & McGory, J. (2000), Inter-transcriber reliability of ToBI prodosic labeling, *in* 'Proceedings of ICSLP 2000, Beijing, China'.

Syrdal, A. K., Wightman, C., Conkie, A., Stylianous, Y., Beutnagel, M., Schroeter, J., Strom, V., Lee, K. S. & Makashay, M. J. (2000), Corpus-based techniques in the AT&T NextGen synthesis system, *in* 'ICSLP 2000', Vol. 3, pp. 410–415.

't Hart, J., Collier, R. & Cohen, A. (1990), *A perceptual study of intonation: An experimental phonetic approach to speech melody*, Cambridge University Press.

Taylor, P. (1994), 'The rise/fall/connection model of intonation', *Speech Communication* **15**, 169–186.

Taylor, P. (1998), The tilt intonation model, *in* 'ICSLP 1998'.

Taylor, P. (2000), 'Analysis and synthesis of intonation using the tilt model', *J. Acoust. Soc. Am.* **107**(3), 1697–1714.

Taylor, P., Black, A. & Caley, R. (1998), The architecture of the Festival speech synthesis system, *in* 'The Third ESCA Workshop in Speech Synthesis', pp. 147–151.

Taylor, P., Caley, R. & Black, A. (2001), 'Heterogeneous relation graphs as a mechanism for representing linguistic information', *Speech Communication* **33**, 153–174.

Trager, G. L. & Smith, H. L. (1951), *An Outline of English Structure*, Norman, OK: Battenburg Press. Reprinted 1957 by American Council of Learned Societies.

Trim, J. L. M. (1959), 'Major and minor tone groups in English', *Le Maître Phonétique* **112**, 26–29.

Vallduví, E. & Vilkuna, M. (1998), On rheme and kontrast, *in* P. Culicover & L. McNally, eds, 'Syntax and Semantics', Vol. 29, San Diego, CA: Academic Press, pp. 79–108.

van Santen, J. & Hirschberg, J. (1994), Segmental effects on timing and height of pitch contours, *in* 'ICSLP', Vol. 2, Yokohama, pp. 719–722.

Willems, N., Collier, R. & 't Hart, J. (1988), 'A synthesis scheme for British English intonation', *JASA* **84**(4), 1250–1261.

Xydas, G. & Kouroupetroglou, G. (2001), The DEMOSTHeNES speech composer, *in* '4th ISCA Workshop on Speech Synthesis', pp. 167–172.