A Phonetic Model of English Intonation

Paul Alexander Taylor

A thesis submitted for the degree of
Doctor of Philosophy
University of Edinburgh

1992

Abstract

This thesis proposes a *phonetic model* of English intonation which is a system for linking the phonological and F_0 descriptions of an utterance.

It is argued that such a model should take the form of a rigorously defined formal system which does not require any human intuition or expertise to operate. It is also argued that this model should be capable of both analysis (F_0 to phonology) and synthesis (phonology to F_0). Existing phonetic models are reviewed and it is shown that none meet the specification for the type of formal model required.

A new phonetic model is presented that has three levels of description: the F_0 level, the intermediate level and the phonological level. The intermediate level uses the three basic elements of *rise*, *fall* and *connection* to model F_0 contours. A mathematical equation is specified for each of these elements so that a continuous F_0 contour can be created from a sequence of elements. The phonological system uses **H** and **L** to describe high and low pitch accents, **C** to describe connection elements and **B** to describe the rises that occur at phrase boundaries. A fully specified grammar is described which links the intermediate and F_0 levels. A grammar is specified for linking the phonological and intermediate levels, but this is only partly complete due to problems with the phonological level of description.

A computer implementation of the model is described. Most of the implementation work concentrated on the relationship between the intermediate level and the F_0 level. Results are given showing that the computer analysis system labels F_0 contours quite accurately, but is significantly worse than a human labeller. It is shown that the synthesis system produces artificial F_0 contours that are very similar to naturally occurring F_0 contours.

The thesis concludes with some indications of further work and ideas on how the computer implementation of the model could be of practical benefit in speech synthesis and recognition.

Declaration

All the work contained in this thesis is my own unless otherwise stated and has not been submitted for another degree at any university.

Acknowledgements

Steve Isard was my main supervisor for the duration of my PhD study. He spent endless hours discussing various aspects of my work with me and I am very grateful for his advice, encouragement and friendship. Jim Hieronymous was my second supervisor. I learnt a considerable amount about speech technology from my discussions with him. Bob Ladd gave me advice and constructive criticism, and I thank him for introducing me to intonational phonology. I am also grateful to Martin West, of Salford University, who first introduced me to the speech field, and encouraged me to take up postgraduate study.

During my course of study, I was employed at the Centre for Speech Technology Research. I am grateful to Mervyn Jack for giving me the opportunity to work there and the time to conduct the research presented in this thesis. Many others at CSTR deserve thanks: Fergus McInnnes for advice on subjects ranging from shell programming to spelling; Mark Schmidt for the use of his data; Paul Bagshaw for the use of his F_0 tracking software; Mike Steele for his graphics programs; John Elliot for advice on LaTeX and Richard Caley for giving me help on just about everything to do with computers. Nick Campbell (unknowingly) gave me the idea of trying to formalise prosodic processes, which eventually evolved into the study of intonation presented here. I would also like to thank my colleagues Iain Ballantyne and Alan Wrench who were good friends throughout and faithfully put up with the more bad-tempered apsects of my personality which emerged as the thesis neared conclusion.

I would like to thank all my friends in Edinburgh, especially the residents of Graham Brown House, the Mountaineering Club, and all the others who made my time in Edinburgh so enjoyable.

I owe a great deal to my family for all they have done for me over the years. It is to them that this thesis is dedicated.

Contents

List of figures

List of tables

Chapter 1

Introduction

Phonetic modelling concerns the relationship between two different representations of intonation: fundamental frequency and phonology. Fundamental frequency (F_0) is the acoustic correlate of pitch and is defined as the frequency of vibration of the vocal folds. Fundamental frequency descriptions are normally represented as F_0 *contours* which are plots of F_0 against time. Phonological representations are used by linguists to describe intonational sounds of similar meaning and phonetic form. The two representations are quite different: the F_0 contour is numerical and continuous whereas the phonological description is qualitative and (usually) discrete.

The reason there are different representations (in many areas, not just intonation) is that it is often easier to work with information in different forms: it is useful to have acoustic evidence on which precise measurements can be made; it is also useful to be able to talk about things which belong to a certain class although their exact acoustic representations may differ.

1.1 A Formal Approach

The aim of the work described in this thesis was to provide a rigorously defined method for deriving one intonation representation given the other.

What we want to find is a system that can derive the phonology of any utterance given the F_0 contour, and also the F_0 contour given the phonology. A system can only be tested on a finite number of utterances, so it will be impossible to conclusively prove the system correct. However, finding an existing system which can model even a small number of contours is very difficult, so the number of utterances need not be huge to prove that a new system is better than

those already in existence.

My approach is a *formal* approach. It aims to link the two representations explicitly so that a machine can perform the mappings between the two descriptions.

In many ways, this approach is similar to that of generative phonologists such as Chomsky and Halle (1968). An explicit grammar will be defined that can be used either to map between the F_0 contour and the phonological representation, or to map from the phonological representation to the F_0 contour. How these mappings are to be performed must be rigorously defined as a fully formal system will operate without the need for human language intuition, expertise or intelligence. It could be argued that the provision of a system that can derive the phonology of an utterance from its F_0 contour is impossible as the phonology can only be determined by using higher level linguistic information in conjunction with evidence from F_0 contours. However it is surely the case that *some* phonological information can be extracted from the F_0 contour; and if the proposed grammar can only produce a partial phonological description, that will still be of interest.

The approach taken here departs from that of Chomsky and Halle in two important aspects. Firstly, the aim is to provide a grammar that links the phonology and the acoustic level, rather than the phonology and a surface phonetic level. Thus instead of providing a grammar which relates two discrete systems of description, the grammar presented here relates the discrete phonological level to the continuous acoustic (F_0) level. The consequence of such an approach is that the standard symbolic rule-based techniques of traditional phonology are somewhat inappropriate as we are dealing with a continuous description.

The second major divergence from the traditional approach is that the aim is not to propose a grammar that describes the performance of a fictional ideal speaker; the grammar is intended to be able to model any contour produced by a healthy[1] native English speaker.

It could be claimed that as the proposed grammar has to cover a greater "distance" in the speech process (i.e. between the phonology and acoustics rather than phonology and phonetics), and as it is not dealing with an ideal speaker, the design of such a grammar will be considerably more difficult than the design of a traditional phonological grammar. This may be so, but I think this increase in difficulty is balanced by the fact that the relationship between the phonology of intonation and the F_0 contour is simpler than the equivalent process

[1] Not suffering from a pathological speech disorder.

in segmental phonology. Intonation production is governed by a single articulator, the glottis, whereas segmental production is controlled by a number of articulators working in parallel. The phonological intonation systems described in this thesis seem to be simpler than most segmental phonological systems.

There are many reasons to use a formal approach. The main advantage is the reason why other generative phonologists use their frameworks: it provides a rigorous, coherent, concrete system which can easily be tested. Chomsky and Halle explain:

> One of the best reasons for presenting a theory of a particular language in the precise form of a generative grammar, or for presenting a hypothesis concerning general linguistic theory in very explicit terms, is that only such precise and explicit formulation can lead to the discovery of serious inadequacies and to an understanding of how they can be remedied. In contrast, a system of transcription or terminology, a list of examples, or a rearrangement of the data in a corpus is not "refutable" by evidence. It is for just this reason that such exercises are of very limited interest for linguistics as a field of rational inquiry.

In addition to the theoretical advantages of using a formal approach, there is the very useful practical advantage that such systems can be implemented on computers. My interest in intonation arose from work in speech synthesis and I felt that the model presented here could in principle be incorporated into either a synthesis or recognition system.

The work presented here is concerned with *English* intonation; many of the principles and ideas discussed here may be directly applicable to other languages, but such considerations are outside the scope of this thesis.

1.2 Overview

Chapter 2 is a review of previous work in phonetic modelling. Preceding the review itself are two preliminary sections. The first presents a formal framework for phonetic modelling, the second introduces the intonational issues that are relevant to the phonology/F_0 relationship.

The main part of this chapter explains the theories belonging to the British School (Palmer, 1922), (Crystal, 1969), (Halliday, 1967), (O'Connor and Arnold, 1973), the Dutch school (t'Hart and Cohen, 1973), the Pierrehumbert school (Pierrehumbert, 1980), (Ladd, 1983b) and the Fujisaki school (Fujisaki and Kawai, 1982). These systems either provide a non-formal account of the F_0/phonology relationship, or provide an account that is only partly specified in

formal terms. The workings of the models are explained and it is shown how well each would serve as the basis for a fully formal system. The chapter concludes by arguing that fundamental problems prevent any of these models being used as they stand. The chapter not only questions the (usually under-specified) F_0/phonology relationships of these models, but in many cases argues that the difficulties with these models derive from more general problems which cannot simply be solved by specifying the behaviour of the model more precisely.

Chapter 3 presents a new phonetic model. The first part of this chapter presents a new intermediate (a kind of phonetic) level and description. It is shown how this description system can accurately synthesize and analyse many types of pitch accent. A mapping process is defined which produces a synthetic F_0 contour given an input in the form of the intermediate description. An equivalent analysis mapping is defined, which aims to produce a description on the intermediate level when given an F_0 contour. The analysis mapping is not fully formal as it relies on human labeller expertise to locate pitch accents. The consequences of this lack of full formality are examined in chapter 4.

The second part of chapter 3 presents a new phonological description. This section of the thesis starts with a discussion of what issues are relevant in the design of an intonation phonology, and in the design of a grammar that will link this new phonology to the intermediate level. The tune aspects of the phonological system are described in full. Other aspects of the phonology are not dealt with as thoroughly. It is argued that problems with how to represent phenomena such as phrasing, prominence and pitch range are still too controversial and intractable to be dealt with here. A tune grammar is defined which shows how to convert phonological tune descriptions into intermediate descriptions and vice-versa.

The chapter concludes with a discussion of the new model. In this discussion the feature based nature of the new phonological level is examined, and it is seen that many uncertainties still exist about how to describe pitch accents.

Chapter 4 explains work in designing analysis and synthesis computer algorithms that can map from one level to another. These algorithms were designed to prove that the model could work in a formal way without having to rely on human labelling expertise. Also, it was possible to use the computer implementation to measure the accuracy with which the computer could analyse and synthesize F_0 contours.

The most difficult implementation task was the extraction of the intermediate level descrip-

tion from the F_0 contour. An assessment method is explained which is able to compare two transcriptions of the same utterance and give a "similarity" score. The hand labelled version of the utterance is taken as correct, and the computer transcription is compared to this to see how well the computer has performed the mapping. This assessment method is used to train the thresholds of the automatic system, and to evaluate the system's final performance. Results are given showing how well the analysis performs on open and closed tests. The automatic system performs very well, but is still significantly worse than the human labeller.

A very important conclusion drawn from the work on implementing the analysis system concerns why it performs less well than a human. If the system is to be called fully formal, both the computer and the human should be capable of following the same analysis procedure and their performance should not be different. What appears to be the problem is that the phonetic model makes slightly incorrect assumptions about the nature of F_0 contours, specifically about the effect of segmental influence on the contour. It is seen from this that more work needs to be carried out either in pre-processing the F_0 contours so that they fit the expectations of the phonetic model, or else to explicitly build in a segmental element into the model.

The method of synthesising F_0 contours from the intermediate level is also described. This synthesis system is shown to be very successful, with the synthesized contours often being very accurate reconstructions of naturally occurring ones.

Work in the automatic derivation of a phonological tune description from an intermediate level description is also shown, but other aspects of the phonology are not dealt with as the theory is incomplete in these areas.

Chapter 5 summarises the main conclusions from the work presented in the thesis and suggests topics for further work. The suitability of the computer implementations for speech synthesis and recognition are also discussed in this chapter.

Chapter 2

A Review of Phonetic Modelling

This chapter presents a review of previous work in phonetic modelling. Preceding the main review are two introductory sections. The first explains a framework for a formal phonetic model and the second discusses some of the issues involved in phonetic modelling.

An important aspect of this thesis is the attempt to define the phonology/F_0 relationship formally. Section 2.1 presents a framework that clearly states what structure phonetic models should take, and what criteria should be used to assess these models. This formal framework is later used to examine existing models and is also used as the basis of the new phonetic model presented in chapter 3.

Section 2.2 introduces some of the issues that are relevant to phonetic modelling. A brief overview is given of how linguists have traditionally described intonational phenomena and what issues are still the subject of current debate. The issues in this section are those of intonation in general, but are described with the intention of showing their relevance for phonetic modelling.

The rest of the chapter takes an in depth study of previous work in phonetic modelling. Several models are assessed in terms of the framework of the formal model outlined in section 2.1.

The chapter concludes with a comparison of these models and shows that considerable effort would have to be expended before these models could be formally complete. It is also argued that some of the fundamental assumptions of these models are problematic and so it would be difficult to implement any of these models in formal manner no matter how much effort was expended.

2.1 Outline of a Formal System

2.1.1 Levels, Grammars and Mappings

In formal terms, we can explain a phonetic model as comprising of levels, grammars and mappings. The two basic levels are the phonological level and the F_0 level. Here, a grammar is defined as a device which relates one level to another. Often it is useful to specify in which direction the grammar is being used, and a process which uses a grammar in a particular direction is called a *mapping*. Two mappings then exist that convert information on one level to information on the other level. These mappings are the phonology-F_0 mapping, and the F_0-phonology mapping. The phonology-F_0 mapping is a synthesis mapping whereas the F_0-phonology mapping is an analysis mapping. Many of the models reviewed below have at least one intermediate level between the phonology and the F_0. Often this intermediate level is referred to as the phonetic level. Here, the term "intermediate" will be used, primarily because these intermediate levels do not fall in the same position across theories, and so it is difficult to compare them directly. Avoiding the term "phonetic" also saves us from having to claim that there is any articulatory reality in these descriptions: they may exist simply to break the complicated mapping into two easier mappings. If an intermediate level exists there will be four mappings, the phonology-intermediate and intermediate-F_0 synthesis mappings, and the F_0-intermediate and intermediate-phonology analysis mappings. The use of an intermediate level will necessitate two grammars, the phonology-intermediate grammar and the intermediate-F_0 grammar[1]. The phonology-F_0 grammar is the overall grammar which links the two main levels.

It is also useful to define a level above the phonological level, which we will call the *linguistic* level. For purposes of the work presented here, this level can be thought of as being everything that is above the phonological level, i.e. syntax, semantics, pragmatics and also para-linguistic and non-linguistic phenomena such as emotion. Because we are not concerned with the relationship between these effects, they can be grouped together to form a single level. Any study which wishes to model behaviour above the phonological level will have to separate the linguistic level into its individual components.

Below the F_0 level, one can define another level, the *waveform* level. F_0 contours can be extracted from waveforms, and speech synthesizers can be used to produce waveforms given

[1] Here the order of the words has arbitrarily been chosen in the synthesis direction.

F_0 (and other) information. The waveform level is not entirely necessary as F_0 can be measured more or less directly by using a laryngograph (see section 3.2.1).

The main requirement of a formal system is that everything must be defined explicitly and exactly. A formal system does not require any native language intuition or draw on linguistic expertise or knowledge. If a system is formally defined, it is possible to implement the system on a computer and achieve exactly the same performance. It is often the case that systems are implemented on computers so as to test their formality, and the model presented here was implemented on a computer for that reason.

2.1.2 Definitions of Levels of Description

The F_0 Level

F_0 is defined as the frequency of vibration of the vocal folds; a F_0 contour is a plot of this frequency against time. By measuring the time taken between equivalent parts in the vocal fold vibration cycle, say the time taken between the instant of glottal closure of one vibration and the next, one can measure the *period* of the vibration. By knowing the time taken to complete a glottal cycle, it is possible to calculate the *frequency* of vibration, which is the number of cycles per second. The most rudimentary F_0 contour therefore consists of a discrete plot of vocal fold vibration frequency at the point of each closure of the glottis.

As changes in vibration frequency between one pitch period and the next are usually small, it is possible to join the points on the rudimentary F_0 contour and create the illusion of a continuous function. This continuous contour can then be sampled at any rate so that a F_0 contour can be described as a list of frequency values at regular intervals. Commonly, F_0 contours are described every 10ms or every 5ms.

Borrowing from traditional generative phonology and syntax (Chomsky, 1965), we can say that the function of the phonology-F_0 grammar is to produce all the legal F_0 contours of the English but none of the illegal ones. Also, using the F_0-phonology mapping, the grammar should be able to generate the correct phonological description for an utterance given the F_0 contour. Chomsky (1971) argues that generative grammars should be evenly balanced between generating surface structures from deep structures and generating deep structures from surface structures. This view is taken here, in that an ideal phonology-F_0 grammar should be capable of being used for both analysis and synthesis.

What constitutes the "possible set" of F_0 contours and the "correct" phonological description will now be discussed.

For a moment assume that for every F_0 contour, an expert labeller can give a single, unambiguous phonological transcription of an utterance's intonation. If we design a phonology-F_0 grammar that when given a F_0 contour produces the same phonological transcription as the labeller, we can say that the phonology-F_0 grammar has correctly analysed that utterance.

It is an easy matter to assess if the two transcriptions agree; if the labeller using Pierrehumbert's terminology[2] marks an accent **H* + L** and the phonology-F_0 grammar produces **H* + L**, the grammar has performed correctly; if however the mapping process produces **H + L*** we can confidently say that this is the wrong transcription. Moreover, if the mapping produces a transcription of **H* +L* +H** for the accent, we can say that this is not only wrong, but *illegal*, as such a sequence disobeys the well-formedness conditions for Pierrehumbert's system.

F_0 contours, on the other hand, are much more troublesome. For purposes of this thesis, the term "F_0 contour" is used to refer to a particular instance of an F_0 contour, and not of a general class or type. The *universal* set of F_0 contours is the set that contains every possible sequence of F_0 values, the vast majority of which will be unpronouncable by humans. The set of legal native speaker F_0 contours contains those which native English speakers can produce. There are no established well-formedness conditions that can be used to tell if a F_0 contour is a member of the native speaker set or not. This is usually not a problem, as we can say that any F_0 contour produced by a native speaker is a member of the set. However, this is an *empirical* criterion, and is of no use when we want to tell whether an artificial F_0 contour (produced by the grammar) is a member of the set. The only way of resolving this problem is to have access to an extensive number of tokens from the native speaker set, and judge artificial contours as being legal if they are the same as a humanly produced member of the set.

This empirical judgement criterion is still troublesome as it relies on us being able to say whether two contours are the same. As has been mentioned before, F_0 contours are effectively continuous which makes it is very difficult to say if two similar F_0 contours are in fact *exactly* the same. Thus we can only say to what degree two contours are similar to one another.

In attempting to assess the similarity of F_0 contours two methods are commonly employed. The first method involves using a distance metric and comparing the F_0 values at equivalent

[2] Pierrehumbert's system is explained in section 2.5.

points on the contour. The result of such a process is a similarity score, which if within a defined limit would imply that the contours were acceptably similar to be called the same. The problem with this approach is that the acceptance limit is arbitrary. If set too loosely, human subjects may be able to perceptually distinguish contours which are judged the same. If the limit is too stringent the system will be expending effort modelling variation in F_0 which is irrelevant.

Alternatively, a perceptual measure might be designed, whereby two contours would be deemed the same if human subjects could not distinguish the difference. The problem with this approach is that it is difficult to generalise the findings from a particular perceptual experiment and it is impractical to have subjects judge every contour produced by the grammar.

No matter which approach is taken, firm yes/no decisions are difficult to come by, and we will see that this problem in assessing the similarity of contours has important repercussions.

There are also difficulties associated with the measurement of F_0. In most work, F_0 is measured with the use of a computer F_0 detection algorithm which produces an F_0 given a digitised waveform. (Cheng and O'Shaughnessy (1989), and Medan et al. (1991), describe F_0 tracking algorithms.) These algorithms often produce errors due to the inherently difficult task of extracting F_0 from speech waveforms. Thus two different F_0 tracking algorithms may produce slightly different F_0 contours from the same waveform. In the work presented here, a laryngograph was used to measure F_0 (see section 3.2.1). This provided a much more reliable method for extracting F_0 but there was still a limit to the precision of the device.

A common solution to the lack of well-formedness conditions is to say the F_0 contours produced by a phonetic model constitute the legal set. This is the approach taken by most phonetic models that are reviewed in this chapter. There is nothing wrong with this approach if the set of contours produced by the phonetic model is similar to the native speaker legal set. However if the model's legal set is very different to the native speaker set, the model's grammar may find it very difficult to analyse a contour it cannot produce as such contours are "alien" to the phonetic model.

If a system cannot model a group of contours from the native speaker legal set we can say with certainty that it is an insufficient model. If it can model these contours we can claim that there is a good chance that the model is sufficient, but it will be impossible to say with certainty that the model is totally correct unless we can test it on every contour in the native speaker legal

set. One might think that the amount of data needed to test a phonetic model would therefore have to be very large. In fact, this is not the case. We can prove that all the existing models are insufficient by testing them on only a small amount of data. Thus we are not at the stage at the development of phonetic modelling where the amount of data needed to test models is a problem. The difficulty is that any new system which models a set of data can only be claimed to be an adequate model with a certain *confidence*. The confidence that this new model is the correct one will increase as the set of test data that it correctly models increases.

The Phonological Level

We have a strict definition of what an F_0 contour is, but it is not possible to give a correspondingly strict definition of the phonology of intonation.

The temptation might arise to propose a phonological system that makes the the phonology-F_0 grammar very simple. If the phonology was free to be designed in an arbitrary way, the easiest phonology-F_0 grammar would arise from having a phonology that was very close to the F_0 level. The reason for not having a very "phonetic" phonology is that the phonology, as well as being linked to the F_0 level, must also be linked to the linguistic level. The phonological system should express differences in meaning, and just as the optimal system will have a simple and accurate phonology-F_0 grammar, the grammar that links the phonology to the linguistic level should be as simple and accurate as possible. The phonological level is not free to be manipulated so as to make the design of the phonology-F_0 grammar simple; the phonological level must be relevant and accessible to higher levels. Therefore, there are constraints on where the position of the phonological level can be with regard to the linguistic and F_0 level. The exact position of the phonological level is still somewhat arbitrary, but as we shall see later in this chapter, there is considerable agreement between existing theories as to where the phonological level should be positioned.

However, just as the phonological level does not have a strict definition, neither does the linguistic level. The definition of the linguistic level does not concern us directly, but as shown later, the imprecise definitions of higher level linguistic functions do present some problems for phonological description systems.

The problems of deciding whether two F_0 contours are the same was discussed above. It is a much easier matter to decide if two phonological descriptions are the same due to their

discrete nature. However, a problem arises in that we have no way of determining what the correct phonological description for an utterance is. Unlike the F_0 contour, which has a strict "physical" description, the phonology is a linguistic invention and is not directly measurable. The only practical solution is to compare human transcriptions with transcriptions derived from the formal system. If the two compare well we can say that the formal system is mimicking the ability of the human. If the transcriptions differ, it may be because the formal system is at fault, or it may be due to the an error with the humanly produced transcription. Thus it is difficult to say whether a formally produced transcription is correct as we have no independent specification of what correct is.

In summary we can say that both the phonological level and the F_0 level have significant problems associated with them that make the design of a formal model difficult. The F_0 level has a strict definition, but it is difficult to say whether a particular F_0 contour belongs to the native speaker legal set. While it is straightforward to say if a phonological description is legal, or if two phonological descriptions are the same, the problem lies with discovering what the correct phonological transcription for the utterance is.

2.1.3 Testing Phonetic Models

In the previous section the "all and only" traditional generative phonology/syntax criterion was suggested. We can divide this criterion into a number of tasks.

For the time being, we will ignore the problems associated with phonological descriptions and state that for every humanly produced utterance, there is a correct phonological description and a correct F_0 contour description. The mappings can then be tested using the following criteria:

phonology-F_0 Mapping. Given the phonological description for each utterance in a set of data, is it possible to derive an F_0 contour that is indistinguishable from the measured F_0 contour for that utterance?

F_0-phonology Mapping. Given the F_0 contours for a set of data, is it possible to derive the correct phonological descriptions?

Another useful testing procedure is the *analysis/resynthesis test.* In its general form, this test can be described as follows.

Analysis/Resynthesis Test. A description on a particular level is to be converted to a different level. The relevant mapping is performed. If the complementary (inverse) mapping is then performed on the newly derived description, the resultant description should be indistinguishable to the original. If result is different, then we can say that there is a fault in the model.

As we have discussed in the previous section, it is difficult to decide what the correct phonological transcription is, and it is difficult to tell if two F_0 contours are the same. These problems have to be kept in mind when assessing the practical performance of a model.

2.2 Important Intonational Issues in Phonetic Modelling

The purpose of this section is to introduce some of the terminology of intonation which is used in the review that starts in section 2.3. The review sections discuss in more detail the problematic aspects of different theories as regards phonetic modelling.

This section gives a brief explanation of five intonational issues that are relevant to phonetic modelling. These are: *tune, phrasing, scaling, timing* and *segmental influence.* Some of these issues, such as segmental influence, are not relevant at all above the phonological level, but are vital in the automatic analysis of F_0 contours. Hence this section is heavily weighted to the needs of phonetic modelling and takes a somewhat different angle from a normal introduction on intonational phonology.

2.2.1 Tune

The intonation tune can be broadly described as the pitch pattern of an utterance. Tunes differ from one another in *type* and in *association.* By using different *types* of tunes, the speaker can convey a wide variety of effects such as surprise, disbelief, excitement and sarcasm. By varying the *associating* of the tune one can emphasize certain words. By shifting emphasis from "john" to "match" in examples 1a and 1b, one can convey different effects. By varying the type of tune, one can also express different effects as in example 1c.

Example 1a *John* went to the match (*as opposed to Harry*)

Example 1b John went to the *match* (*not the theatre*)

Example 1c *John* went to the match (*disbelief: but he hates football !*)

Describing tune association is often simple: words or syllables are described as being accented or as being the focus or nucleus of the phrase. The word with which an accent is associated is partly dependent on the syntactic, semantic and pragmatic structure of the utterance; why pitch accents occur where they do is not under discussion here. The location and type of pitch accent used is determined by the effect the speaker wishes to produce.

Describing tune type is much more difficult and a wide variety of description schemes have been proposed. These schemes can be broadly divided into those which classify tunes using dynamic features (rises and falls) and those which use static features (tones).

Theories also vary in the size of the units they use. *Global* descriptions make use of a few basic patterns that cover the entire phrase, *atomistic* theories make use of smaller units that combine together to from larger patterns. Jones (1957) is at the global end of the scale, the British school (O'Connor and Arnold, 1973), (Halliday, 1967) uses sub-phrase units, while the American school (Pike, 1945), Pierrehumbert (Pierrehumbert, 1980) and the Dutch school (t'Hart and Cohen, 1973) use units which are smaller still.

Much of the discussion on the subject of tune centres around how to describe *pitch accents*. A pitch accent is commonly manifested in the F_0 contour as a (relatively) sudden excursion from the previous contour values. This excursion attracts attention to the syllable with which it is associated. Pitch accents can only occur in association with stressed syllables (see section 2.2.6 on stress), but need not occur on all stressed syllables.

Most work agrees that the intonation phrase is the basic structural unit of intonation (see section 2.2.2). In each intonation phrase there is a focus word which is perceptually the most important pitch accent. This accent is often referred to as the *nucleus*. Traditionally, the nucleus is also the last accent in the intonation phrase.

The British School (O'Connor and Arnold, 1973) uses a separate system of classification for nuclear accents and non-nuclear accents. All the pre-nuclear accents in an intonation phrase are described with a single unit, which is different from the approach taken in the American School (Pike, 1945) and the Pierrehumbert School (1980) where each pre-nuclear accent receives a classification.

The other main area of interest in tune description concerns what happens at the ends of intonation phrases. Often F_0 is low at a phrase boundary, but in many circumstances F_0 is high. For instance, if another phrase directly follows the current one, a *continuation rise* may

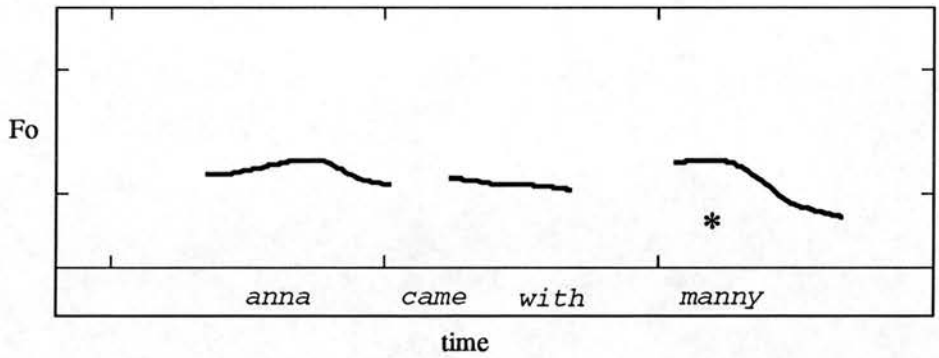


Figure 2.1: *Typical Fall Accent. "Anna came with Manny." The nuclear fall is on the stressed syllable of "manny" denoted by an *. The fall is a commonly found intonation accent and is often used in neutral declarative situations.*

be present. If the tune is that of a yes/no question, the final pitch may also be high. The British school deals with these effects by using different nuclear accent and tail configurations. Pierrehumbert's bitonal system makes use of high and low *boundary tones* which distinguish the different types of contour.

Tune Effects

O'Connor and Arnold (1973) give a thorough account of many of the uses of intonation in colloquial British English. They show how different combinations of pre-heads, heads nuclear accents and tails produce different effects in the listener's perception of the utterance. It would be impossible to show all the possible types of F_0 contour for English, but six common nuclear accent types are shows in diagrams 2.1 to 2.6. These examples are not comprehensive and other theories may classify these contours differently. The examples merely demonstrate some of the intonational effects that can be produced.

2.2.2 Phrasing

The study of prosodic phrasing concerns what types of prosodic constituents exist, how they relate to one another, how prosodic structure is manifested, and what factors determine this prosodic structure. In the previous section, the term *intonation phrase* was used, which was described as the domain in which tunes are realised. This is the primary unit of intonation structure.

Nearly all theories use this unit of phrasing which is called a *word group* (O'Connor and

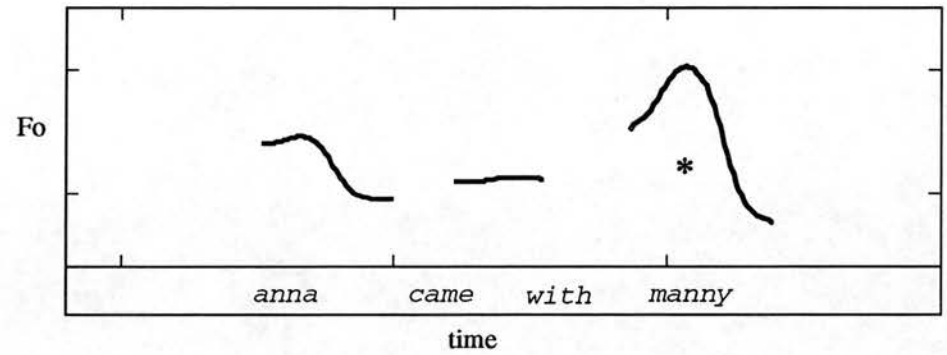


Figure 2.2: *High Fall, "Anna came with Manny !". This shape corresponds to a British "high fall", +raised or pitch level 4. In this particular utterance there is still a single intonation phrase, and the word "anna" also has an accent, but this accent is pre-nuclear. Some may argue that there is no phonological distinction between fall and high fall, and that the high fall is really just a extra prominent fall (see section 2.2.3).*

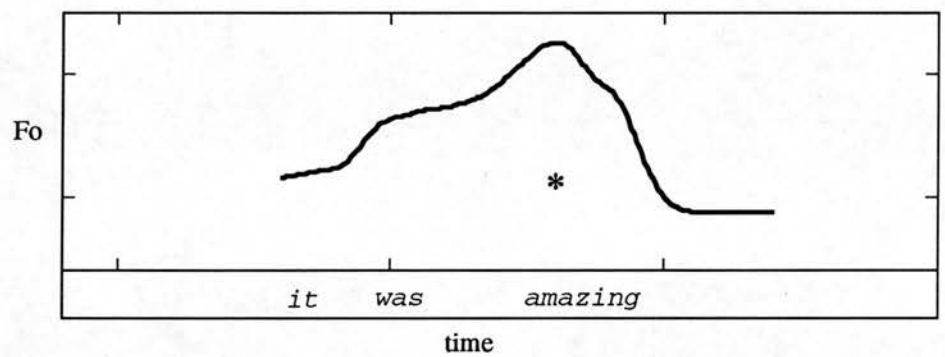


Figure 2.3: *Rise fall accent, "It was Amazing !". Spoken with a sense of wonderment, this accent is similar to a fall, but with a much larger preceding rise. The peak value of the F_0 contour is also later than with a simple fall accent.*

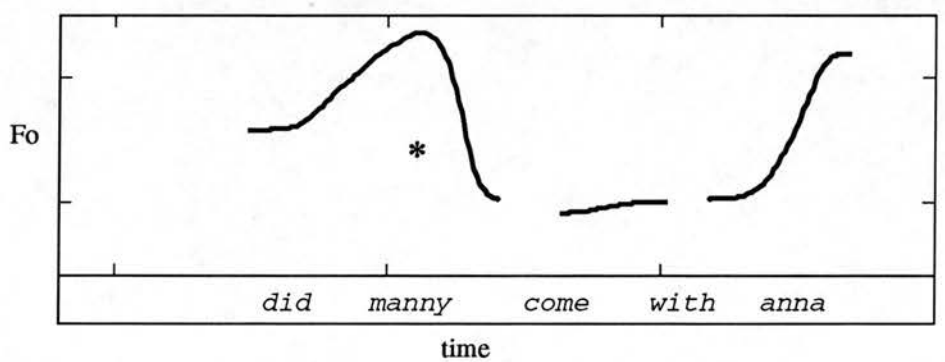


Figure 2.4: *Fall rise accent, "Did Manny come with Anna ?" A peak in the F_0 contour occurs in the stressed syllable of "manny" (*). After coming down from the peak, the contour rises slowly and finishes with a sharp rise at the end of the phrase. This type of accent is often used for simple yes/no questions.*

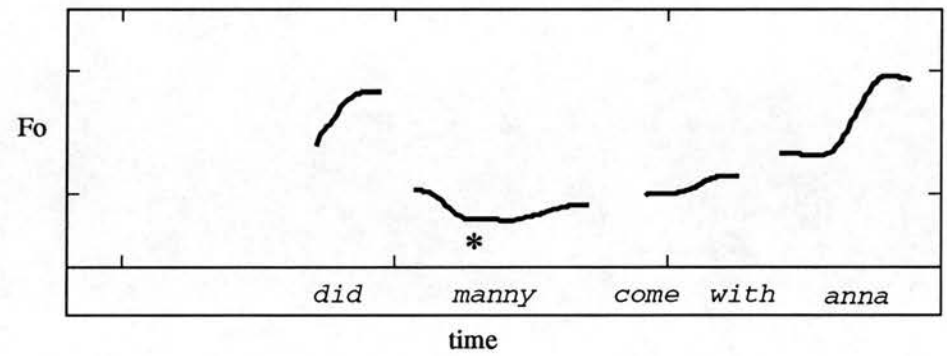


Figure 2.5: *Low Rise, "Did Manny come with Anna ?!". This accent shape may at first glance look similar to the fall-rise, but differs in that the stressed syllable (*) of the word which carries the nuclear accent is not a peak but a valley. Thus the F_0 contour rises from the nuclear accent. Quite often this accent is preceded by a falling F_0. This accent can be used to convey incredulity or disbelief.*

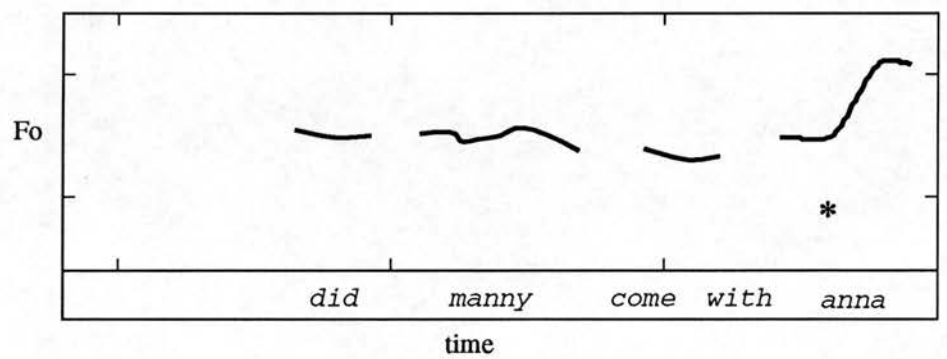


Figure 2.6: *High Rise, "Did Manny come with Anna ?". Here the accent falls on the first syllable of Anna. There is no valley as with the low rise, and the F_0 on the nuclear syllable is much higher. High rise accents are often used for yes/no questions where the speaker is looking for confirmation in a statement, as in "ok?" or " right?". It is similar in many ways to the low rise, with the F_0 contour rising from the nuclear accent, the main difference being that the nuclear accent occurs considerably higher in the speaker's pitch range, and is often not preceded by a falling section of contour.*

Arnold, 1973), *tone unit* (Crystal, 1969), *tone group* (Halliday, 1967) or *intonation phrase* (Pierrehumbert, 1980). Although the strict definitions may vary, they all describe the same basic unit. The term *intonation phrase* or simply *phrase* will be used here.

As traditionally defined, an intonation phrase is often delimited by non-hesitation pauses and contains at least one (nuclear) accent. Simple sentences such as example 2 have one intonation phrase. More complex sentences such as example 3 have two (the first ending after "come").

Example 2 My sister lives in Edinburgh.

Example 3 Even if he does come he won't be able to stay very long.

Intonational phrasing can help with syntactic disambiguation. Example 4 has three phrases (delimited with a "|"), while example 5 has two.

Example 4 My sister, | who lives in Edinburgh, | has just had twins.

Where the sister lives is just additional information, not essential to the correct interpretation of the utterance. The relative clause is non-defining

Example 5 My sister who lives in Edinburgh | has just had twins.

This distinguishes the sister who lives in Edinburgh from the sister who lives in Glasgow, i.e. the relative clause is defining.

Older theories have tended to use the intonation phrase nearly exclusively as the main unit of prosodic structure (Crystal, 1972), (O'Connor and Arnold, 1973); Halliday (1967), uses four levels, the intonation phrase, the foot, the syllable and the phoneme; but this still implies one level of intonation phrasing. Pierrehumbert (1980) made use of the *intonation phrase*, and more recently introduced the *intermediate phrase* which is dominated by the intonation phrase (Beckman and Pierrehumbert, 1986). Selkirk (1984) uses intonational phrase, phonological phrase, prosodic word, foot and syllable as units of description.

Ladd (1986), (1992a) describes these theories as holding to the "strict layer hypothesis" (originally proposed by Selkirk (1984)) and claims that even with these extra levels, the theory of prosodic structure is too strict. He proposes that prosodic structure, like syntactic structure, can be recursive. He states that in principle a prosodic tree is not limited in depth, and nodes

of a given type may dominate nodes of any other type. A major motivation of Ladd's is to do away with the somewhat arbitrary names and definitions that have been proposed for units below the intonation phrase. More recently, he proposes an amended version using limited recursion, where a hierarchy of constituent types exists, and a node may only dominate a node of similar type but not of higher class. Although there is disagreement as to how many levels of prosodic phrase exist, there is a consensus that prosodic structure is "flatter" than syntactic structure, and if recursion does exist, it is still manifested in a "flatter" tree (Bachenko and Fitzpatrick, 1990).

In looking for evidence of different levels of phrasing, the well documented phenomena of phrase-final syllable lengthening is often used (Klatt, 1975), (Campbell and Isard, 1991), (Crystal and House, 1988). These studies have shown that syllables at the end of phrases tend to be longer than normal. Using this idea, researchers have tried to determine how many levels of phrasing can be distinguished. Wightman et al. (1992) studied acoustic lengthening at different types of possible prosodic boundaries and found it possible to distinguish three distinct categories *below* the level of the intonation phrase. Another experiment on different data also showed acoustic evidence for at least four levels of phrasing (Ladd and Campbell, 1991).

As for what determines prosodic structure, the debate is just as heated. A simple proposition would be that prosodic structure is determined by syntactic structure. Few people think the relationship is simple; the debate hinges around whether syntax determines prosody via a complicated mapping, or if syntax is only one of a number of factors affecting prosodic structure (Bickmore, 1990), (Chen, 1990), (Selkirk, 1984). Consider examples 6 and 7 (Chomsky and Halle, 1968).

Example 6 This is $[_{np}$ the cat that caught $[_{np}$ the rat that stole $[_{np}$ the cheese]]].

The prosodic structure, as shown in 7, is clearly aligned differently.

Example 7 This is the cat | that caught the rat | that stole the cheese.

Often (as in examples 6 and 7) prosodic structure is not *directly* related to syntax. Metrical factors seem to have some effect, with the result of splitting the sentence into approximately equal chunks, which can position prosodic boundaries in the middle of syntactic constituents.

The debate about what determines prosodic structure is of course made all the more difficult by the uncertainty of how many levels of prosodic structure exist and how they relate to one another.

2.2.3 F_0 Scaling: Downdrift, Pitch Range and Prominence

Downdrift

It has been observed by many people that there is often a gradual downdrift in the value of F_0 across a phrase (t'Hart and Cohen, 1973), (t'Hart and Collier, 1975), (Pierrehumbert, 1980), (Cooper and Sorensen, 1981), (Liberman and Pierrehumbert, 1984), (Fujisaki and Kawai, 1988). How downdrift (often referred to as declination) is dealt with by different theories varies widely. Ladd (1984) gives a review of some of the different theories.

Many treat downdrift as an automatic physiological effect arising from changes in sub-glottal pressure during the course of an utterance (Lieberman, 1967), (Cooper and Sorensen, 1981). This account gives the speaker little conscious control over declination.

The approach of the Dutch School (t'Hart and Cohen, 1973), has been to use three parallel declination lines, which refer to a baseline, a mid-line and a line corresponding to the top of the speaker's normal range. The contour must follow one of these lines or be rising or falling between them. Fujisaki's model is more flexible in that the rate of declination and initial starting value can be varied, but the overall effect is still automatic (Fujisaki and Kawai, 1988). Liberman and Pierrehumbert (1984) show that the final F_0 value for utterances is invariant under a wide range of utterance lengths and pitch ranges which is inconsistent with the view that declination slope is constant. They propose an exponential decay downdrift effect, with the additional feature of "final lowering" at the end of the phrase. Figure 2.7 shows three different views of declination.

A major claim of Pierrehumbert's thesis (1980) was that more than one factor was responsible for the downdrift of F_0 contours. As with many other theories, she proposed that the phonetic declination effect exists, but also argued that the major contribution to the downdrift of utterances was *downstep* which was a phonological effect and therefore controllable by the speaker. Figure 2.8 shows a downstepping and non-downstepping version of the same sentence. These two sentences are not only different in F_0 shape, they also have subtly different meanings. (The first sounds more excited, the second sounds more relaxed and confident.)

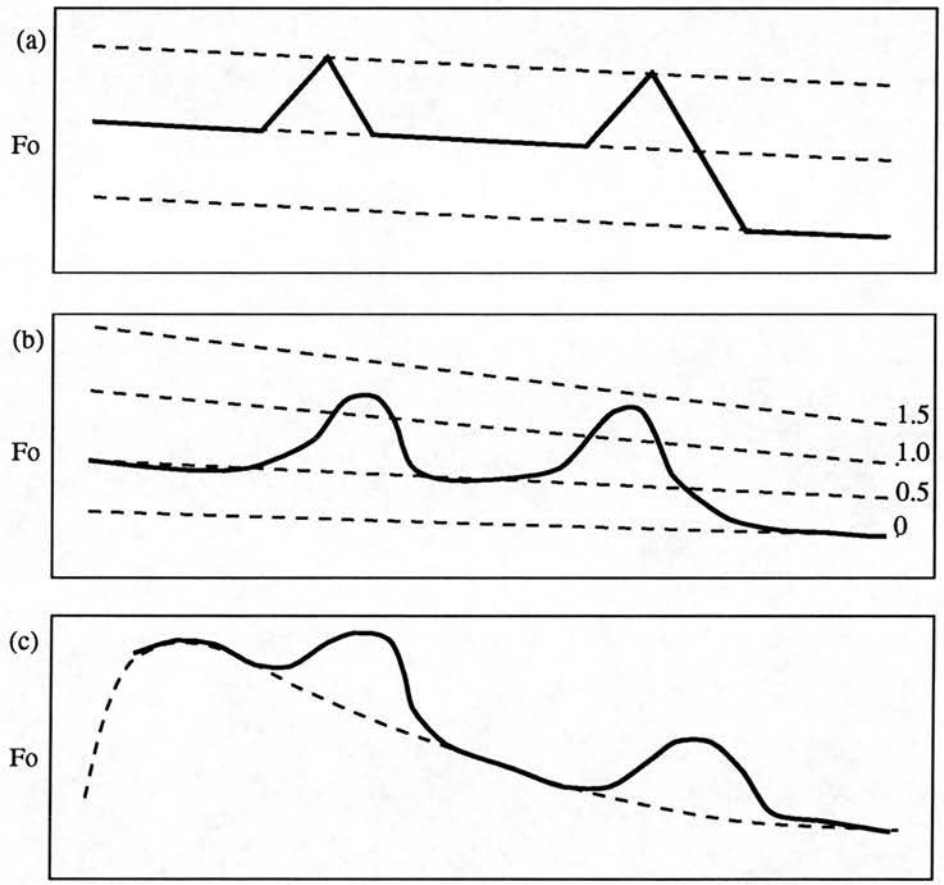


Figure 2.7: *The Dutch model, figure(a), has three declination lines, which refer to a baseline, a mid-line and a line corresponding to the top of the speaker's normal range. The contour must follow one of these lines or be rising or falling between them. Pierrehumbert's system (figure (b)) scales the pitch range 0.0 to 1.0 for normal speech but allows higher levels. The contour is not required to follow any of these declination lines - they are merely the "graph paper" on which the F_0 contour is produced. Note how the lines converge with respect to time. The Fujisaki model (figure (c)) only specifies a baseline, which decays exponentially.*

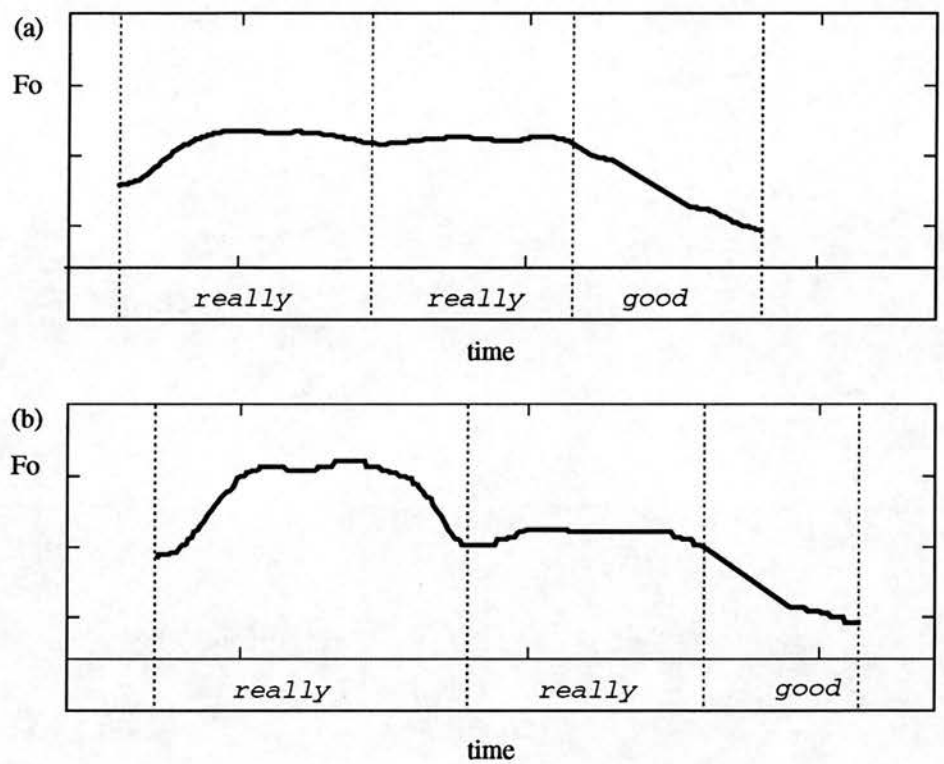


Figure 2.8: *Two utterances of the phrase "really, really good". Figure (a) has the two "really"s at the same pitch level, with a fall on "good". In Figure (b) each word is downstepped relative to the previous word.*

Prominence

Prominence is a measure of a pitch accent's perceived strength; by increasing an accent's prominence more emphasis is perceived on the word on which the accent occurs.

The prominence of an accent does not have a straightforward relationship with the F_0 contour. Because of the declination effect, the later an accent occurs in a phrase, the lower its peak F_0 value will be. The effect of declination is compensated for by the listener, thus two accents occurring at different times in a phrase can have equal prominence, but widely differing F_0 values (Pierrehumbert, 1980). The bitonal phonology of Pierrehumbert makes use of low accents, the F_0 of which *decrease* with increasing emphasis (Liberman and Pierrehumbert, 1984).

It has been shown in a number of experiments (Ladd, 1992b), (Gussenhoven and Rietveld, 1988) that the perception of prominence is far from straightforward. These experiments showed that speakers do not decide the strength of each pitch accent independently and that the relationships between accent heights and prominence can only be judged in context. Thus the strengths of accents within phrases follow a pattern.

Sometimes a measure of prominence is included within the specification of an accent's tune. In many phonologies, one comes across classifications such as "High Fall" (O'Connor and Arnold, 1973), "+ raised peak" (Ladd, 1983b), and "pitch level 4" (Pike, 1945). These terms refer to what the researchers feel are distinct phonological categories, that is, "High Fall" is not just a big fall, it is a separate accent. (See figures 2.1 and 2.2). Pierrehumbert (1980) argues against this view by stating that prominence is paralinguistic and therefore outside the realm of phonology. The counter argument is that these higher accents appear in predictable circumstances and therefore deserve a phonological category of their own.

Pitch Range

In music, if a sequence of notes is repeated an octave higher than the original, the tune remains the same, even though the frequency values of the notes are different with respect to the original. Rather it is the constancy of the *pattern* that gives the perceptions of the tunes being the same.

The same effect is observable in intonation: by increasing the overall pitch of an utterance while keeping the basic tune pattern constant, the perceived tune remains the same. The relationship between intonational tunes in different pitch ranges is not as simple as the musical

equivalent. It has been shown that the increase or decrease in pitch range need not be constant throughout the phrase, as utterances always tend towards a fairly constant final F_0 value.

Pitch range varies for a number of reasons. In single isolated utterances it can be used for increasing the overall emphasis of the utterance. When one "raises one's voice" in anger one is using an increased pitch range. Pitch range factors also have a role to play in longer utterances. If a speaker started at the same F_0 level with every intonation phrase, the speech would sound very bland. Speakers use a variety of pitch ranges to be more expressive. Some think that pitch range is paralinguistic (Pierrehumbert, 1980), (Beckman and Pierrehumbert, 1986) and that it is a "free" choice the speaker must make at the start of every phrase. Ladd (1983a), (1992a), (1988) argues that speakers do seem to exhibit systematic behaviour in the choice of pitch range, which can be linked to discourse structure.

The boundaries between pitch range and prominence effects are not clearly defined. For example, in many systems it is difficult to say if an unusually high accent is due to extra (local) prominence or a temporary shift in pitch range (see section 2.5.4). Sometimes the effect of raising the pitch range for a phrase is to increase the emphasis for the *entire* phrase. In example 8 the raised phrase is in italics.

Example 8 First of all he turned up late, then he said he had forgotten the tickets and then he had *the nerve to claim it was all my fault.*

2.2.4 Timing

The position of an intonation feature (say a peak) can vary as a result of the influence of three aspects of intonational timing. The tune association discussed in section 2.2.1 is a large scale effect and the position of the peak can be described by stating which syllable it is associated with. On a smaller scale, the position of the peak within a syllable can have phonological significance. The differences in peak alignment between fall accents and rise-fall accents helps distinguish these two accents types. On a smaller scale still, there are slight non-phonological differences in peak position. The segmental structure of the syllable may affect the location of the peak, and peaks occurring at the ends of phrases may occur sooner than normal (Silverman and Pierrehumbert, 1990).

The reasons for why tunes are associated as they are is not directly relevant to phonetic modelling. How differences in timing contribute to different phonological accent types is of

importance, as are the details of how these broad phonological categories are realised in F_0 contours.

2.2.5 Segmental Influence

Intonation is not the only factor influencing F_0. It has been noted that segmental environment has a noticeable effect on the F_0 contour. Silverman (1987) gives an thorough review of the more important studies on segmental effects.

These effects are not within the realm of intonation but they must be taken into account when analysing F_0 contours. Their effect on the contour is to make direct measurements of F_0 difficult to interpret unless either segmental context is kept constant (a method used by many researchers) or the contour is processed in some way to normalise for segmental context.

For purposes of this thesis we can identify three main aspects of segmental influence. It should be noted that while much of the literature concentrates on the influence of segments on F_0, here we are concerned with the segmental influence on F_0 *contours*, and thus again segmental influence is viewed from a slightly different angle than is usual in the literature.

The most noticeable segmental effect is that in unvoiced segments, by definition, there is no F_0. Evidence exists that listeners do not distinguish the intonation pattern in F_0 contours with different voiced/unvoiced patterns. The listeners achieve this by interpolation through the voiced regions and thus create the perception of a continuous contour (Kohler, 1991a). Although unvoiced segments may not change the perception of a utterance, they certainly change the appearance of the F_0 contour, and unvoiced segments must be taken into account when analysing contours.

The second type of influence, *segmental perturbation* arises from consonants, especially obstruents. The general tendency is for obstruents to produce sharp spikes in the contour. These segmental perturbations are normally short in duration (typically less than 30ms), but do cause F_0 excursions that are comparable in size to some smaller pitch accents (Silverman, 1987).

The third type, *segmental scaling*, includes the often documented phenomena of *intrinsic* F_0. Investigation into this phenomena has shown that high vowels consistently produce F_0 values higher than low vowels (Lehiste and Peterson, 1959), (Zee, 1980). Thus two accents that are similar in every way except for the vowel of the accented syllable may have different F_0 values. Similarly, the type of vowel, and the overall voicing pattern within the syllable can

affect the durational patterns of the F_0 contour (Gartenberg and Panzlaff-Reuter, 1991). Thus the type of vowel and the structure of the syllable must be taken into account when analysing F_0 contours.

In his thesis, Silverman demonstrates that segmental F_0 has an influence on listeners ability to correctly identify segments and argues that these effects must be taken into account when building synthetic F_0 generators for synthetic speech systems.

2.2.6 Stress

The term *stress* is used in the literature with a bewildering variety of definitions. Jones (1957) describes stress as the "degree of force with which a sound or syllable is uttered", Abercrombie (1967) talks in terms of "force of breath impulse" while Crystal (1969) uses loudness as the main indicator of stress. Others regard stress as being controlled by the relative durations of syllables. Lehiste and Peterson state that the perception of stress is produced by variations in intensity, F_0 phonetic quality and duration. Often the terms "stress" and "accent" are used analogously, which adds to the confusion.

Explanations such as the ones given above are primarily concerned with the provision of a definition of stress that is in some sense phonetic. However, a viewpoint taken by many is to make a firm distinction between stress that appears in utterances and a more abstract lexical definition of stress. In all bi-syllabic words spoken in isolation, one syllable is perceived as having more emphasis than the other. Thus in the word "table" the stress falls on the first syllable, in "machine" the stress falls on the second. In words with many syllables, such as "information" the main stress falls on the "at" syllable, but "in" can also be regarded as having stress. Traditionally, such terms as "primary" and "secondary" stress were used to described syllables in longer words, but more recently work such as metrical phonology (Liberman, 1975) has argued that there is a more complicated underlying structure to word's stress patterns. The theories of metrical phonology are also useful in explaining why, when these words are spoken in a sentence or phrase, the surface stress structure is often quite different from the concatenated lexical stress patterns of the individual words.

It is not necessary to give a strict definition of stress here as its lexical form, phonetic substance and acoustic properties are not under investigation. The relationship between stress and accent is occasionaly mentioned, however, and it is helpful to give some idea of what is

meant by stress. Bolinger provides a useful definition which is sufficient for our purposes:-

> I reserve *accented* for the syllable which actually *is* highlighted in a sentence - to show the importance of the word - and apply *stressed* to the particular syllable in the word that gets the accent *if* the word is important enough to get one.

Review

No-one has proposed a complete, formal phonetic model of intonation. Most previous work concentrates on a describing a part of the phonology-F_0 grammar or a particular level. Some systems such as the Fujisaki model do propose a formal intermediate-F_0 mapping and phonetic level of description but this constitutes only part of a phonology-F_0 grammar. Other models cover the range of the phonology-F_0 grammar but fail to explain their workings in the formal way that is required here.

Several "schools" are reviewed using the formal testing criteria explained in section 2.1.3. The use of such testing criteria is somewhat harsh as they are much stricter than the criteria that the developers of the models used. However, this is necessary as it is central to this thesis that a formal approach must be adopted.

2.3 The British School

2.3.1 The Phonology of the British School

The British School of intonation includes contributions made as far back as Palmer (1922). Other major contributions in this school have come from O'Connor and Arnold (1973), Crystal (1969), and Halliday (1967). All these variants on Palmer's original theme use dynamic features such as "rise" and "fall" to describe intonation.

In the account given by Crystal, the most important part of the contour is the nucleus which is the only mandatory part of an intonation phrase. The nuclear accent can take one of several configurations, e.g. fall, fall-rise, low rise. Other parts of the contour are termed the *tail* (**T**) which follows the nucleus, the *head* (**H**) which starts at the first accented syllable of the intonation phrase and continues to the nucleus, and the *pre-head* (**P**) which precedes the head. The intonation phrase has a grammar of (**P**) (**H**) **N** (**T**), where the brackets denote optional elements.

The phonology-F_0 grammar of this system is the most loosely defined of all the models reviewed here. This is hardly surprising as none of the originators of this system had the technology to analyse F_0 contours in detail. The phonological descriptive terms are related to actual contour shapes that are found, but the descriptions should not be interpreted too literally. Both "fall" and "rise-fall" accents have rises followed by falls, the difference being that the rise

in the fall accent is much smaller and earlier in the syllable than the rise in the rise-fall accent. Halliday describes his tones using rise and fall terminology, but does away with the standard naming system preferring simply to name his tones 1, 2, 3, 4 & 5.

All of the British School phonologies are broad in their coverage of the intonational effects of English. Very little direct reference is made to F_0 contours and this has "freed" the British School from having to give strict accounts of various phenomena. This has allowed the authors of this system to cover a large amount of data more quickly. The terminology of the system is rather impressionistic: it matters little in this system that the "rise" and "fall" terms are not always used to indicate the same phenomenon, since these descriptions are seen as being purely mnemonic. This impressionistic feel in the British school does present problems when a formal grammar has to be developed.

2.3.2 Phonetic Modelling in the British School

Some more formal descriptions have been proposed for use with the British School phonology. In particular, two models which have been designed for speech synthesis purposes are those of Isard and Pearson (1988) who use Crystal's phonology and Vonwiller *et al* (1990), who use Halliday's. Both these synthesis models use the full range of the British school tune descriptions and Isard and Pearsons's scheme is capable of variations in prominence and pitch range.

2.3.3 Problems with the British School Phonetic Models

There are a number of fundamental problems which prevent these systems from be expanded to becoming formal grammars for the British School phonology.

Straight Line Approximation

The Isard and Pearson system does not attempt to accurately model F_0 contours as straight lines are used to construct the contour. It is clear from looking at any contour that F_0 often follows a curved path. However, Isard and Pearson, like many others, have justified their use of straight lines because experiments have shown that straight line approximations to real F_0 contours are perceptually equivalent to the original so long as the approximation does not deviate from the

real contour too much[3]. Straight line approximation is commonly used in speech synthesis phonetic models. This is mainly due to the belief of the researchers that straight lines are somehow more simple than curves. Often the straight line contours produced by synthesis models are passed through smoothing filters which result in more natural looking curved F_0 contours.

This approach is difficult to reverse for analysis purposes. One could "inverse smooth" the F_0 contour and extract the straight line approximation, but this would be technically very difficult and also inadvisable as it can be argued that real F_0 contours do not have an underlying straight line form.

Trying to match straight lines directly to real F_0 contours is also problematic as there will obviously be a large difference between the real and the straight line F_0 contour, so optimising a fit between the two will be difficult. Many different possible fits of equal distance may be found for a single curve, implying that there would be many intermediate level descriptions for a single contour.

Flexibility

As is common in intonation speech synthesis systems, Isard and Pearson only give one set of parameters for their model. They use synthesis rules such as:-

> "There is a standard rate of -15 intervals/sec at which falls, rise-falls, fall-rise and rise-fall-rises descend from their target heights to the baseline."

The "-15" is an arbitrary constant which is unsubstantiated. It is not claimed if this is constant across speakers and if not, how it might vary, or how the slope-rate for a new speaker might be measured. It is also claimed that the peak of a fall "occurs 60msec into the voiced portion of the syllable" which is another unsubsantiated arbitrary constant. There is no indication of where these constants come from; whether they are approximations, or just guesses at real timing behaviour.

If we were to take the Isard and Pearson model as it stands and adopt it for analysis purposes, we would run into trouble the first time a peak did not occur 60ms into the syllable. A more flexible timing approach could be derived, but there are no indications as to how this might be

[3] see section 2.4 on the Dutch School for more information on straight line perceptual equivalence.

achieved. Thus although it may be possible to bypass the straight line approximation problem, the lack of flexibility as regards timing and rates of fall make the system difficult to adapt for analysis purposes.

The Isard and Pearson model was developed as a phonology-F_0 mapping scheme. As such, it may perform quite well in being able to model the F_0 contours of the speaker it was based on. However there are no principles for adapting the system to cope with the variations between speakers, and as such the system only models a small portion of the native speaker set of legal F_0 contours.

The use of these constants is quite common in speech synthesis literature. This is not a problem if systems are designed specifically for speech synthesis purposes, as it is only necessary to generate a contour that is an acceptable approximation to a real one. The problem lies in that the use of these constants restricts the system to being able to produce only a very small number of the human set of legal F_0 contours, and therefore the system cannot be used as the basis for an analysis system.

2.3.4 British School: Summary

The difficulty in using the British School as the basis for a formal system is not restricted merely to problems with the synthesis models mentioned above. The need for a grammatical specification such as (**P**) (**H**) **N** (**T**), has been brought into question by work such as Pierrehumbert's which has argued that the head is more easily described using a separate classification for each pitch accent rather than a single unit covering all the pre-nuclear accents. For example, the use of a single classification unit for the head makes the description of phonological downstep difficult. Similarly, Pierrehumbert adopts a more flexible approach to describing post-nuclear phenomena than the system of tails employed by the British School.

Pierrehumbert's system is strong from a production point of view in that the course of the intonation contour does not have to be planned out more than a few syllables ahead. She and others (Ladd, 1983b), (Silverman, 1987) have criticized theories such as the British one (termed a *contour interaction* theory by Ladd) as requiring too much "lookahead" or "pre-planning" to be practical. In a long intonation phrase, there may be many pre-nuclear accents. In Pierrehumbert's system the speaker is free to choose the type of each pitch accent just before it is uttered; in the British school the speaker must make a single choice at the start of the first

accent which then determines how all the pre-nuclear accents are to be spoken.

It is clear that much more work will have to be carried out before a F_0-phonology mapping system could be developed for the British school. However, considering the above mentioned problems in the phonology, the usefulness of building such a model must be questioned. Designing a phonology-F_0 grammar that is inherently linked to a problematic phonology will hamper the grammar in its ability to analyse and synthesize F_0 contours.

2.4 The Dutch School

2.4.1 The Dutch Phonetic Model

The Dutch School (t'Hart and Cohen, 1973), (t'Hart and Collier, 1975) proposed a clearly defined phonetic level between the F_0 level and the phonological level.

Unlike the Isard and Pearson phonetic model, the Dutch system is designed for both analysis and synthesis of F_0 contours. The system is based on the principle of *stylization* of F_0 contours. Stylization in the Dutch system involves taking a F_0 contour and attempting to fit a series of straight lines as closely as possible to the original contour. This stage is useful as it reduces the amount of data needed for further analysis: a small number of straight lines are easier to deal with than a continually varying F_0 contour. From these stylizations, a series of basic patterns can be found - this process is called *standardization*.

The version of the theory presented in t'Hart and Cohen (1973) describes contours in terms of three declination lines - high, middle and low. Pitch accents are realised by rising and falling between these declination lines. An example of a stylized and standardized contour is shown in figure 2.9 (from Willems (1983)).

Because of the stylization process, the continuously varying nature of the F_0 contour is eliminated, and because of the standardization process, the contour description is further reduced into a small number of units (rises, falls etc). This stylization and standardization process constitutes a F_0-intermediate mapping procedure. The standardized phonetic description can then be easily mapped onto a phonological description. Thus the full range of mappings are available and the Dutch School exhibits a fair degree of formality.

Willems (1983) shows a set of standardized patterns (based on Halliday's phonology) for use in a British English speech synthesizer. By providing robust methods for performing the

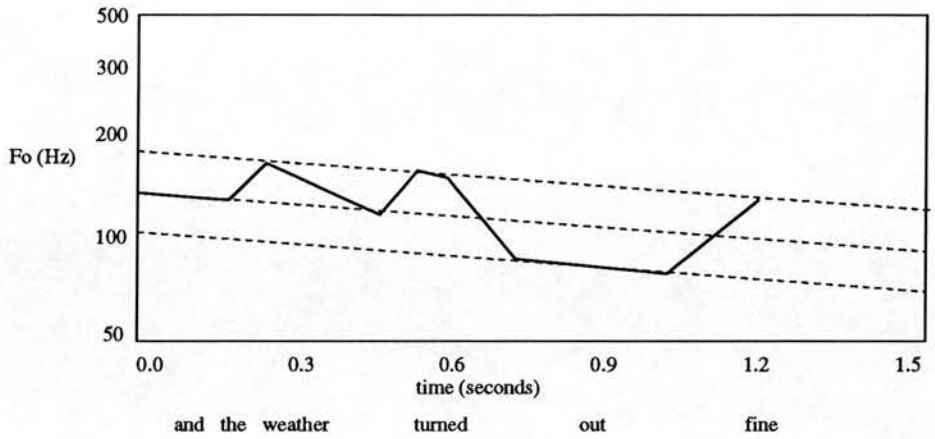


Figure 2.9: *Example of a standardized contour in the Dutch system. The dotted lines denote the three declination lines and the thicker solid line shows the path of the F_0 contour. The first excursion to the top declination line is a head accent (British school). The second accent which rises to the top line and then falls to the baseline is a fall accent. The rise at the end is a continuation rise.*

stylization and standardization processes on an F_0 contour, we would have a formal version of the Dutch system.

2.4.2 Problems with the Dutch Model

A significant aspect of the Dutch model is that it is intended to be used for analysis as well as synthesis. The Dutch school literature often reports experiments where F_0 contours are analysed using the stylization and standardization procedure. Thus the Dutch school is in line with idea that a phonetic model should be capable of both synthesis and analysis. A balance between analysis and synthesis is an important step towards the fulfillment of the requirements of a formal phonetic model as outlined in section 2.1. However, significant problems prevent the Dutch model from fulfilling all the criteria for a formal phonetic model.

Liberman and Pierrehumbert's (1984) experiment on downstep shows that at least 5 distinct F_0 levels can be consistently articulated by a variety of speakers. They use this fact to argue very strongly that *any* system which proposes a strict division of levels is incapable of describing English intonation properly. (They use this argument to criticise Pike's (1945) 4 level system). The Dutch system uses three rigidly defined levels, and therefore has problems dealing with any sort of downstep. This strict three level distinction also poses problems with changing the pitch range or describing accent prominence (see section 2.5.4 for Ladd's solution to controlling

pitch range).

The phonetic, intermediate level is incapable of expressing all the necessary distinctions between downstepping and non-downstepping contours. To see why this is problematic, consider the following situation. The standardization procedure is used on a F_0 contour that exhibits more than two levels of downstepping, and the phonetic description extracted. An F_0 contour is then reconstructed from the phonetic description. This reconstructed contour will not exhibit the correct downstepping pattern and will be substantially different from the original. Thus the F_0-intermediate and intermediate-F_0 mappings are not the analysis and synthesis equivalents of each other.

This is an example of a system failing the *analysis/resynthesis test* described in section 2.1.3.

The fault in this case lies with "forcing" the F_0 contour to be analysed in terms of a the three line declination system. If there is a large discrepancy between the behaviour of real F_0 contours and what the model proposes, then the model will run into severe difficulties if used as an analysis system. This is another example of the system failing because its definition of the legal set of F_0 contours (i.e. those which it is capable of synthesizing) is substantially different from the native speaker set ofF_0 contours. The model will have difficulty analysing any contour that is not within its own legal set.

The account of the system adapted for English given in Willems (1983) also contains arbitrary constants as in the Isard and Pearson model. The sizes, slopes and positions of the rise and fall elements are given for each accent type with no indications as to where these constants came from.

Problems also exist in that there is no strictly defined analysis procedure for labelling F_0 contours in terms of the intermediate level. Willems admits that the system cannot be properly formal by explaining his labelling criteria. He claims -

> The Selection of a Pattern. The user (labeller) has to make a choice from the set of six basic patterns. Since the grounds on which selections are made in normal speech situations are not clear, this choice has to be a matter of *good taste*" [4]

Using "good taste" is not a formal approach as this requires human labelling expertise and intuition. Often with aspects of the models reviewed in this chapter, the formal specification

[4] my italics.

of a particular mapping is merely a question of "tidying-up" the existing theory. However, in the case of the Dutch model, it would be very difficult to formalise the analysis procedure as there is no principled method of analysing contours which are not within the model's legal set. When the F_0 contour under examination is within the model's legal set, the analysis procedure is probably quite straightforward; it is when the contour is not within the legal set that "good taste" is required. Thus the lack of formality stems from fundamental problems with the model and is not simply a question of "tidying-up".

2.4.3 Dutch School: Summary

The Dutch model suffers from its phonetic level being too inflexible with regard to pitch accent size and scale. The downstep experiments of Liberman and Pierrehumbert show most clearly why three levels ar not enough, but examination of F_0 contours[5] shows that F_0 contours can not easily be described within this framework.

A positive aspect of the Dutch School is that it has made an attempt at providing the type of mappings that are required for a complete formal system. In a way this is the cause of the model's downfall. If the phonetic level were not so clearly defined it would be more difficult to criticise it. However, since all the features of the system are explicitly described, it is easy to find fault where fault exists.

2.5 The Pierrehumbert School

2.5.1 Pierrehumbert's Intonational Phonology

The version of Pierrehumbert's phonological system as explained in her thesis (Pierrehumbert, 1980) describes an intonation contour as a series of high and low tones. This system is in some ways the extension of Pike's (1945) theory which used a system of four tones numbered 1 to 4. By using a system of diacritics which distinguish tones located on accented syllables from those occurring at boundaries and between accents, Pierrehumbert showed that the four level description could be reduced to two tones, which she called **H** (high) and **L** (low).

Pitch accents can be represented as either a single or double tone. Every pitch accent has a *starred* tone (*) which signals that it is that tone which is directly associated with the accented

[5]Note the wide variation in accent size in the contours show in appendix B.

syllable. Double tone accents have an additional tone, referred as a a floating tone, which is marked with a (-). Floating tones are not directly aligned with a stressed syllable, but are associated in a more indirect way. The possible pitch accents are **H***, **L***, **H* + L-**, **H- + L***, **L- + H*** and **L* + H-**.

At phrase boundaries, *boundary* tones can be found, which are marked with a (%). *Phrase tones* are used to show path of the contour from the last (nuclear) accent to the phrase boundary. These are also marked with a (-).

Unlike the British school analysis, there is no strict division of the contour into regions such as head and nucleus. Both nuclear and pre-nuclear accents can be any one of the six types described above. The nucleus accent is distinguished because the phrase and boundary tones that follow it allow a much larger inventory of intonational effects.

Each tone forms a target from which F_0 contours can be realised by using interpolation rules. The target value for each tone can be scaled independently, and the pitch range for each phrase is a "free" choice as well. As with many other theories, Pierrehumbert retains the idea of a declination baseline, but argues that the downdrift commonly observed in F_0 contours is mainly due to the phonological effect of *downstep* which again is controllable by the speaker. Pierrehumbert proposes that the downstep effect is "triggered" by the speaker's use of a sequence of **H L H** tones, using evidence from African tone languages as justification (see figure 2.8 for examples of downstepping and non-downstepping contours).

The version of her theory outlined in her thesis (Pierrehumbert, 1980) uses only one level of phrasing, the *intonation phrase*, but later work proposes the extra levels of the *intermediate phrase* and the *accentual phrase* (Beckman and Pierrehumbert, 1986).

2.5.2 Problems with Pierrehumbert's Phonology

The key points of Pierrehumbert's phonology are :-

- **H** and **L** are the fundamental intonational units.
- Downstep is a phonological effect triggered by sequences of **H L H** tones.
- Nuclear and pre-nuclear accents use the same description system.
- Very little lookahead is needed.

- F_0 contours are constructed by using each tone as a target and using rules to interpolate between the targets.
- Pitch range and prominence are paralinguistic effects over which the speaker has a free choice.

The first question that must be asked is "are **H** and **L** really the fundamental units of English intonation ?" Since the publication of her thesis, the Pierrehumbert phonology has been widely accepted. The "levels versus configuration" debate[6] is now thought by many to have been resolved on the side of levels, mainly due to the wide acceptance of the Pierrehumbert system. A possible viewpoint is that the Pierrehumbert system has become widely accepted not because of her use of tones, but rather because her system adopts a very useful approach to lookahead, downdrift, nuclear/pre-nuclear accents, pitch range and prominence. Her system deals with these effects convincingly, and as her explanation of these effects is intrinsically tied in to the use of tones, the argument for tones being the basis of intonation is strengthened.

However, the argument for using tones has not been worked through thoroughly and is in no way an obvious solution to the problem of how to describe tune.

Pierrehumbert's main argument for the use of tones is that by using two tones instead of the four (used in the Pike (1945) and Liberman (1975) systems) she is able to counter many of the arguments previously put forward for not using tones. Much of the discussion on tone usage is taken from the point of view of correcting previous tone theories rather than starting from first principles and proposing solid reasons for tone use.

Empirical evidence for tones is given in Liberman and Pierrehumbert (1984). This experiment shows that peak F_0 of downstepped accents is predictable (it follows an exponential decay line). Their experiment shows that accents in downstepping contours fit a target model well. They compare the predication made by their target model with the observed behaviour of their F_0 contours and show that the target model predicts F_0 behaviour better than a dynamic rise/fall model would. They admit that the test is not totally fair in that they spent longer developing their target model than their dynamic model, but do claim "We argue that our observations strike a blow in favour of static features"[7].

[6] This debate centres around whether the Pike/Liberman/Pierrehumbert system of tones underlies intonation, or if the Crystal/O'Connor & Arnold/Halliday/ Dutch School system is correct. See Ladd (1983a) and Bolinger (1951).

[7] Page 165

Even taking into account Liberman and Pierrehumbert's natural bias towards tone based phonology, it does seem that their results point clearly to some sort of target based model. However this experiment only proves that downstepping accents when spoken from a list follow targets. It is not possible to extend this fact to confidently conclude that tones and levels underlie the intonation of English.

Interpolation Rules

It is necessary at this stage to talk a little about the system of interpolation rules which is the basis of Pierrehumbert's phonology-F_0 mapping system. Pierrehumbert makes sure that we see her tones as being *phonological* tones, which depend on a number of factors to give their exact F_0 target value. Potentially, there is nothing to stop a particular **L** being higher than a particular **H**, just as in music we cannot say that a "C" note is lower than a "D". (If the C occurs in the next octave up, it will be higher than the D[8].) In Pierrehumbert's system an **L** at the start of a phrase may be higher than a **H** later on. The problem with Pierrehumbert's intonational phonology is that we know that **H** and **L** are abstract, and cannot be mapped to a single F_0 value, but we don't know to what *extent* they are abstract. It seems that she uses different degrees of abstraction in different places: sometimes the tones are taken fairly directly as target points with straight line interpolation between them, sometimes very sophisticated rules are needed to perform the interpolation. There seems to be no principled method of knowing when to interpret tones literally as targets, or when to start using some sophisticated interpolation rules.

The classic example of this confusion concerns the interpolation rules used between two **H*** accents. The default interpolation rule is simple: give each tone a target and use straight lines to join the targets. The actual interpolation rules may be slightly more sophisticated so as to produce curved F_0 contours, but straight line interpolation is representative of the true F_0 shape. However, a different interpolation strategy is needed between two **H*** accents. If the two accents occur within one syllable of each other, straight line interpolation is used. If the accents are further apart the contour "dips" between the two peaks as shown in diagram 2.10. Pierrehumbert states that "this complication in the interpolation rules is unattractive" but argues against the obvious possibility of placing a **L** between the two **H** accents, as this would trigger

[8] see section 2.2.3 for more on analogies between intonation and music.

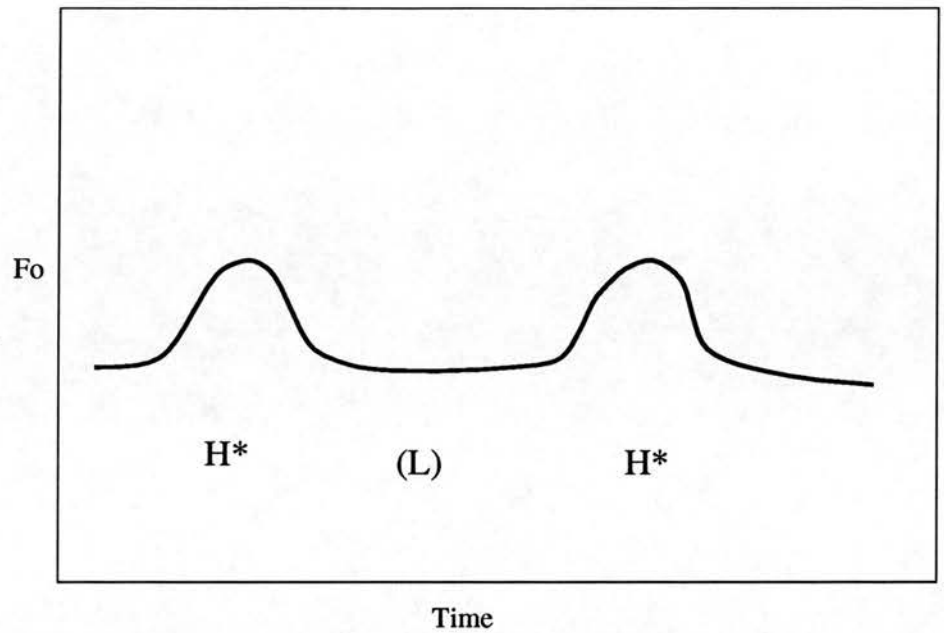


Figure 2.10: *This diagram shows the dipped interpolation between two **H*** accents. The **L** in brackets is where the low tone would be positioned if this were not forbidden by the downstep rule.*

downstepping[9]. This problem of a different system of interpolation being used for one special case is worrying, and casts doubt on the whole concept of using targets and interpolation in this way. If a new set of rules may be presented to get out of one tricky situation, why not use complicated rules any time a tricky situation arises? Thus if some contour seems to violate the tonal system, we can always invent complicated rules to explain them. This also implies that one must consider context before deciding how literally to interpret a particular set of tones[10].

Tonal Configurations

Setting aside the question of interpolation rules, many other problems exist with the use of tones. Ladd (1983b) questions the use of Pierrehumbert's classification of what the British school describes as the rise-fall accent. Pierrehumbert uses **L* + H-** for this tone. The problem is that this accent is similar in F_0 shape and meaning to the more common fall accent, which Pierrehumbert uses **H* + L-** to describe. Ladd argues that a relatively minor change in accent

[9] see sections 2.5.4 and 2.5.3 for a solution to the downstepping problem.

[10] Non-linear interpolation exists for many other tonal contexts, in particular between the nucleus and the boundary tone. Nuclear tone sequences which need additional non-linear interpolation rules include **H* L- H%, H* H- L%, H* H- H%, L* H- L%, L* H- H%, L* L- H%, H-+L* H- H%, H*+L- H- L%, H*+L- H- H%**. These examples have been taken from the appendix of Pierrehumbert's thesis.

type should not be reflected in using a totally different tonal specification. He argues that these accents are of a similar type, and proposes a feature, [± delayed peak] to distinguish them. The use of **L*** for a rise-fall accent is inappropriate as the behaviour of this **L** is totally different from that of the **L** in **H- + L*** or the single tone **L***, both phonetically and in meaning. Liberman and Pierrehumbert (1984) report that as the prominence of an **L*** increases, its F_0 decreases. This is true of the **L*** and **H- + L*** accents but not of the **L* + H**, which shows an increase in the F_0 of the **H** tone with increased prominence, similar to the behaviour of the **H* + L** accent. To use **L*** for two such dissimilar accents is misleading as it does not group together accents which are similar in meaning and in phonetic behaviour.

Other Tones

Tones can be marked to indicate that they occur on a stressed syllable, that they are floating or phrase tones, or that they are boundary tones. What is not clear is in what sense is a low boundary tone similar to a low starred tone. It is clear from studying contours that some end high and some end low, but again there is no evidence to suggest this is due to some tonal phenomenon. By examining Pierrehumbert's analysis of nuclear accents and tails, one can quickly see that the fitting of tones to contours is again non-trivial, with complicated interpolation rules being needed to make the tonal configurations fit the observed data[11].

2.5.3 Amendments to the Original System

Pierrehumbert's system has been widely accepted by many intonation researchers. Here we will discuss some amendments to the original system that have been suggested.

Pierrehumbert proposed that much of the downdrift observed in intonation contours is phonological in nature. This has been widely accepted, although many have disputed that downstep is triggered by sequences of **H L H**. Ladd (1983b) argues that downstep can be treated as a *feature* which is independent of tonal sequence. This makes the entire system easier as it allows the placement of **L** tones to be conducted in a more straightforward way. **L**s can be placed where they seem to belong without the neeed to worry about placing an **L** between two **H** tones that are not in a downstep relation to one another. Ladd also suggests using the features [± delayed peak] and [± raised peak]. The delayed peak feature is used to

[11] see Pierrehumbert (1980), Appendix to Figures.

distinguish between **H* + L-** and **L* + H-**. Ladd uses **H L** for both these tones, and the delayed peak feature to distinguish them. (see section 2.5.2).

Ladd also argues for a raised peak feature which is similar to Pike's pitch level 1[12]. Ladd states that this tone is not simply a **H*** with increased prominence and argues that these tones are phonologically different, giving empirical evidence in Ladd (1992b). Ladd's system therefore uses much simpler tonal configurations than Pierrehumbert's original system and uses an extra layer of features to describe the more subtle differences between accents. Beckman and Pierrehumbert (1986) consider these arguments but decide to stay more or less with the original system.

Ladd also contests the view that intonation structure is strictly flat and proposes recursive or semi-recursive descriptions of phrasing. As evidence, he uses data from an experiment which shows that the pitch ranges of intonation phrases occurring in sequence follow a pattern (Ladd, 1988). Ladd likens his hierarchical phrase structure in which each successive phrase starts lower than the last to a "downstepping" of phrases. The main thrust of Ladd's argument in these issues is that pitch range and prominence are not purely paralinguistic phenomena, as Pierrehumbert claims.

Silverman (1987) notes problems with pre-planning in the Pierrehumbert system, particularly when dealing with post-nuclear phrase tones. He also concludes that downstep exists, but argues that it is dealt with in a cumbersome way because of the **H L H** tonal sequence rule.

2.5.4 Phonetic modelling in the Pierrehumbert School

Pierrehumbert's phonological system is more directly linked to the F_0 level than the phonological description of the British School. To generate an F_0 contour from her phonological specification all one need do is position the tones with respect to time, apply their scaling values, and then use the interpolation rules to generate the F_0 values between the tones.

One would therefore think that phonology-F_0 mapping procedure might be made substantially easier by using the Pierrehumbert phonology. Pierrehumbert argues that her system does not need any intermediate level; that the mapping from phonology to F_0 is a single process. Her work therefore does not use an explicit intermediate level but relates F_0 values directly to the

[12] Pike's original system of having "1" represent the highest tone and "4" represent the lowest was reversed by later linguists.

phonological specification. If we ignore the inelegance of the interpolation rules for the time being, and forget any worries as to whether or not tones are a useful intonational phonology, it is possible to claim that the phonology-F_0 mapping procedure may be capable of generating the set of legal F_0 contours. The numerical downstep model proposed by Liberman and Pierrehumbert shows what factors affect downstep, new/old relations and the F_0 lowering effect that is often observed at the ends of phrases. This provides some of the numerical specification needed for a formal phonology-F_0 mapping, but we cannot say for certain that the mapping process is adequate as not all the necessary numerical detail is available for a fully defined phonology-F_0 mapping. However, when we come to the problem of F_0-phonology mapping in Pierrehumbert's system, we find that this is very difficult to define in a formal manner.

Pierrehumbert argues that there is no intermediate level in her system. The strongest candidate for an intermediate level is the set of targets which are derived from the phonology. This is discounted by Pierrehumbert as she sees that there is no one-to-one mapping between the target level and the phonological level. Thus the phonology-intermediate mapping is well defined by the interpolation rules, but the complementary mapping is impossible as the interpolation rules cannot be trusted to work correctly in reverse.

Pierrehumbert clearly recognises this and states

> The tonal sequence underlying the contour is entirely unaccessible; specifically, the types, locations, and phonetic values of the tones (can not be) accessed.[13]

Thus the originator of the model claims that the intermediate-phonology mapping, and therefore the F_0-phonology mapping is impossible. Presumably Pierrehumbert thinks that some sort of top-down analysis is what helps listeners decide what accent type it is that they are hearing. This idea is not justified as, in my opinion, it is clearly the interpolation rules that are at fault. As the form of the interpolation rules is inherently linked to the use of the two tones, one must question the basis of the entire system. One cannot say that this difficulty in providing a F_0-phonology mapping *proves* that Pierrehumbert's system is at fault, but it does make it clear that the system would be very difficult to use as the basis of any formal phonetic model.

[13] Page 28 of Indiana Linguistics Club Edition.

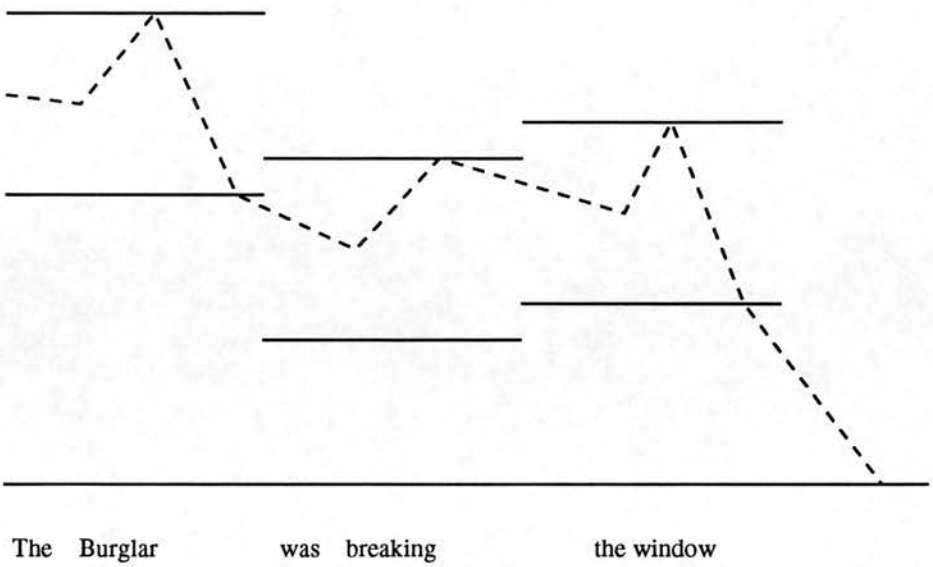


Figure 2.11: *Ladd's register based model. The register settings are used to control pitch range and downstep. Usually the registers decrease in height throughout a phrase, except for the nuclear accent, which uses a slightly higher register.*

Ladd's Phonetic Model

Ladd (1987) proposed an implementation of the Pierrehumbert phonology that does away with the need for a separate interpolation rule for F_0 between **H*** accents. The intermediate level of his system is somewhat similar to the Dutch system in that it uses three lines as references. Where Ladd's model differs is that these lines are set in a *register* which can be vertically shifted so as to allow downstep, different pitch ranges, and prominence variation. Figure 2.11 shows a typical contour in his system.

Ladd only discusses the model for speech synthesis purposes, but it is clear from his account that this type of model could be adapted for analysis purposes. He only discusses the implementation of **H*** and **H* + L-** accents, but it should be possible to model **L*** and other accents. The F_0 contours produced by the model's intermediate-F_0 mapping do not match real F_0 contours particularly accurately as straight lines are used, but as the model makes allowance for downstep and pitch range, a closer fit should be possible than with the Dutch model. The intermediate level is phonologically relevant too, in that the use of registers does not seem unnatural - one can imagine "raising one's voice" in the way Ladd proposes. The registers give a sort of global pattern to the intonation contour, whereas the F_0 contour itself shows the individual course of F_0 through each accent.

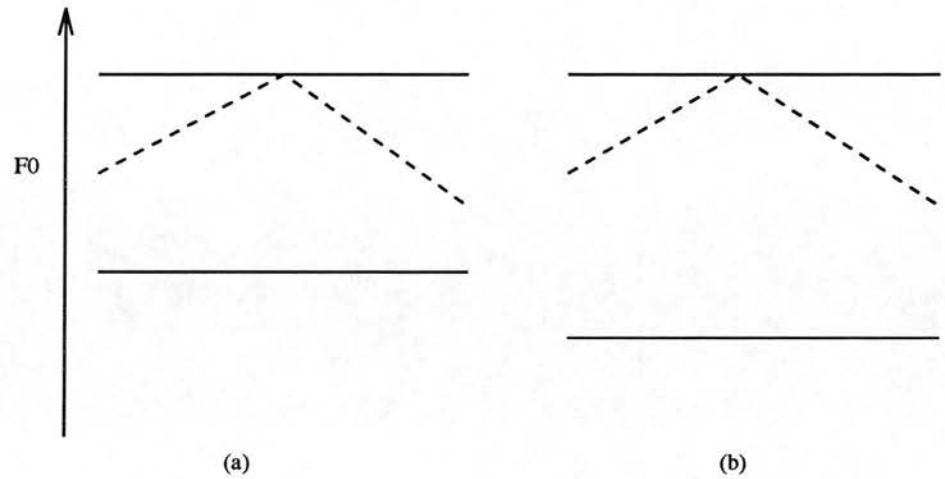


Figure 2.12: *The same F_0 contour, shown with the dotted line can be placed in a high narrow register (a) or a lower, wider register (b). This uncertainty makes analysis difficult.*

Problems arise with Ladd's model in the way the widths and heights of the registers are determined. The *width* (the F_0 difference between the upper line and the lower) of a register is variable, with wide registers being used to accommodate prominent accents. Pitch range is determined by register *height* (the distance from the top line to the baseline).

This presents no problem if this model is used in its original synthesis mode. The difficulty arises when the heights and widths of the registers are to be determined from analysis of an F_0 contour. The synthesis model states that the F_0 contour behaves in a similar way to the Dutch model with the F_0 contour following one of the declination lines, or else rising or falling between them. If real F_0 contours followed this behaviour exactly there would be no problem, but this three line system is only an approximation and in reality F_0 contours will not follow this pattern. Thus it is difficult, from examination of an F_0 contour, to say precisely where the register settings should change, and what their heights and widths will be.

This problem arises from there being too many free variables in the model. The system needs more constraint in its specification of register behaviour if it is to used as the basis of a robust analysis system. This is the opposite problem from that of the Dutch model. There the system was too constrained, forcing different effects to be grouped together. In Ladd's model two effects which are virtually identical may be marked with different register positions and settings. Figure 2.12 demonstrates this problem.

2.5.5 Pierrehumbert: Summary

Pierrehumbert has made a very significant contribution to intonation in recent years. Her treatment of downstep, declination, nuclear/pre-nuclear accentuation, pitch range and prominence has resolved many previous problems. However it has been shown above that there are fundamental problems, not least with the central tenet of the theory, that English intonation is a tone-based phenomenon. The changes proposed by Ladd make the system easier to deal with, but the problems do not disappear completely.

We have shown that the use of the somewhat arbitrary interpolation rules is unattractive and makes any formal analysis difficult. The need for rules such as these is a direct consequence of the tonal phonology, and so the phonology must be questioned. If we take the view that all we are trying to do is to build a phonetic *model* of intonation, rather than attempting to discover the "reality" of intonation, it makes sense to choose the simplest description possible. Pierrehumbert's initial ideas may look simple, but using tones in the way she advocates makes the phonology-F_0 grammar over-complex and cumbersome.

Ladd's phonetic model seems more promising and it solves some of the problems associated with other models, such as the Dutch one. However specifying the register is still a major obstacle to using it for analysis. If the top and bottom lines of the register could be strictly defined from contour analysis, the model might become usable, but there has been no proposal for how this might be accomplished.

2.6 Fujisaki's Model

2.6.1 Fujisaki's Filter-based Phonetic Model

Fujisaki's phonetic intonation model (Fujisaki and Kawai, 1982) takes a quite different approach to the models previously discussed in that it aims for an accurate description of the F_0 contour which allegedly simulates the human production mechanism. Fujisaki's model was developed from the filter method first proposed by Öhman (1967).

Fujisaki states that intonation contours are comprised of two types of components, the *phrase* and the *accent*. The production process is represented by a *glottal oscillation mechanism* which takes phrase and accent information as input and produces a continuous F_0 contour as output. The input to the mechanism is in the form of impulses, used to produce phrase shapes,

and step functions which produce accent shapes.

This mechanism consists of two second order critically damped filters. One filter is used for the phrase component, the other for the accent component.

The F_0 contour can be represented by equations 2.1, 2.2 and 2.3.

$$\ln F_0(t) = \ln F_{min} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{0_i}) + \sum_{j=1}^{J} A_{a_j}(G_{a_j}(t - T_{1_j}) - G_{a_j}(t - T_{2_j})) \quad (2.1)$$

where

$$G_{p_i}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (2.2)$$

and

$$G_{a_j}(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \theta] & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (2.3)$$

F_{min}	baseline
I	number of phrase components
J	number of accent components
A_{p_i}	magnitude of the ith phrase command
A_{a_j}	magnitude of the jth accent command
T_{0_i}	timing of the ith phrase command
T_{1_j}	onset of the jth accent command
T_{2_j}	end of the jth accent command
α_i	natural angular frequency of the phrase control mechanism of the ith phrase command
β_j	natural angular frequency of the accent control mechanism of the jth accent command
θ	a parameter to indicate the ceiling level of the accent component.

Although the mathematics may look quite complicated, the model is in fact very simple. Each phrase is initiated with an impulse, which when passed through the filter, makes the F_0 contour rise to a local maximum value and then slowly decay. Successive phrases are added to the tails of the previous ones, thus creating the type of pattern seen in figure 2.13. The time constant, α, governs how quickly the phrase reaches its maximum value, and how quickly it

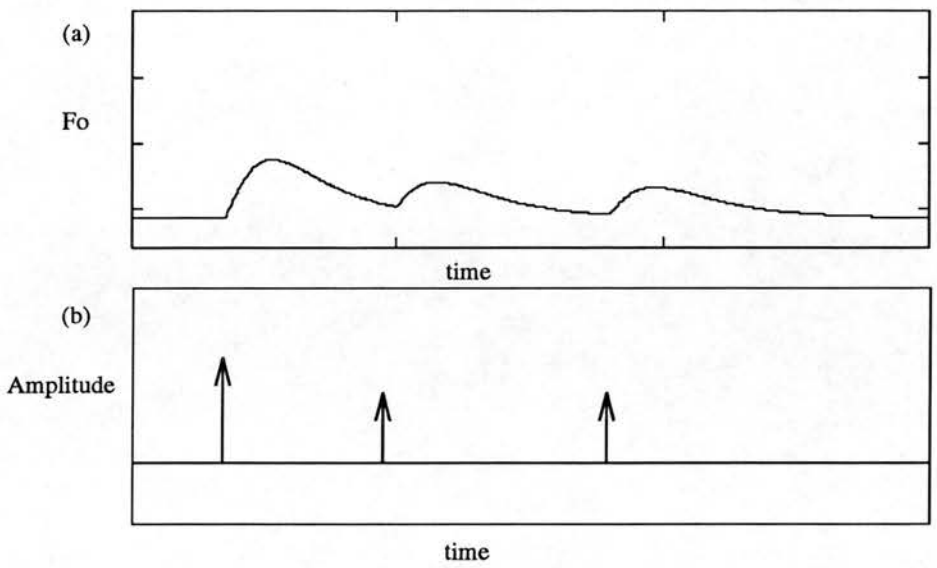


Figure 2.13: *Three phrase components of differing amplitude. Graph (a) shows the path of the F_0 contour, while graph (b) shows the input impulses.*

falls off after this.

Accents are initiated by using step functions. When these step functions are passed through the filter they produce the responses shown in figure 2.14. The accent time constant, β, is usually much higher than α, which gives the filter a quicker response time. This means the shape produced from the accent component reaches its maximum value and falls off much more quickly than the phrase component.

The impulse phrase input is essentially a reset of the declination line. Many other models describe phrase behaviour in terms of a rapid initial rise, followed by a slower fall back to the baseline (Ladd, 1988), (Pierrehumbert, 1980). Fujisaki provides an explanation of this, which matches the acoustic evidence well (i.e. exponential decay) without having to resort to using implausible production schemes (such as rigid declination lines). His phrase scheme shows how the path of the contour can be specified for 2 or 3 seconds from one initial input, with no need for the speaker to devote any conscious effort to producing the phrase shape after the impulse has occurred; the rest of the phrase shape is automatically determined. Figure 2.13 shows the F_0 contour produced by the phrase component.

Using step functions as input is also attractive. The step function has a definite start and end which are related to the boundaries of the stressed syllable with which the accent is associated. The accent duration can be increased by lengthening the step, the prominence of the accent can

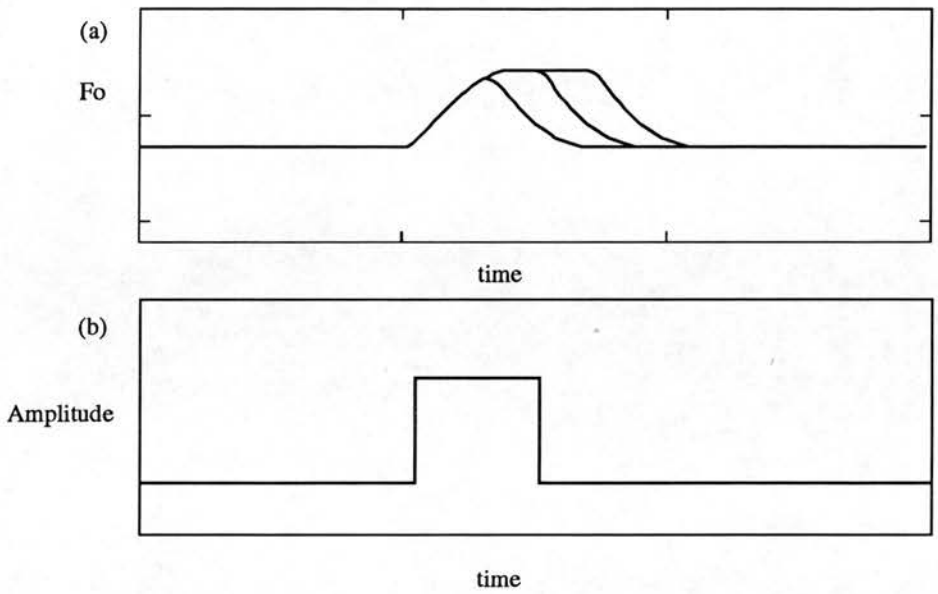


Figure 2.14: *Three accents of differing duration. Graph (a) shows the path of the F_0 contour, graph (b) shows the input step function for the second accent. As the accent duration increases, the accent becomes less like a peak and develops a flat top.*

be varied by using different amplitudes of step functions. Figure 2.14 shows three F_0 contours produced by using step functions of different durations.

2.6.2 General Points on the Fujisaki Model

Japanese, while still being an intonation language, has a very different intonation system from English. This makes it difficult for a non-native speaker to assess if the model is sufficiently general to cover all the intonation effects in Japanese. Fujisaki claims that as his model is based on human production mechanism, it should be applicable to all languages (Fujisaki, 1992). He shows his model's operation on English, Estonian and Chinese intonation[14]. Möbius has adpated the model for German intonation and has shown good results (Möbius et al., 1991a), (1991b).

The Fujisaki model is the only model reviewed here that attempts an accurate synthesis of the F_0 contour. Fujisaki has demonstrated in many papers that the model can fit real F_0 contours very accurately. It is not clear how much effort was needed in specifying the input parameters so that the model output was optimally matched, but satisfactory matches were made for all

[14]The English intonation contour shown in this paper is a declarative utterance.

contours (i.e. the analysis method is not formally defined).

The concentration on a viable production strategy makes it is possible to understand why the intonation contour is the shape it is, which must surely be preferable to simply stating that contours are made up from interpolated targets or rises and fall as other theories do. The Fujisaki model gives substance to the intonation contour.

The model is also attractive in that proper control can be exercised over accent prominence and pitch range. By increasing the amplitude of the phrase component, one can raise the pitch range; by varying the amplitude of a step input, one can change the F_0 of the accent peak, thereby controlling its prominence. These two effects can be controlled independently of one another.

Like the Dutch model, the Fujisaki model is able to perform both synthesis and (non-formal) analysis which makes it a very strong candidate for the type of model we are looking for. Above all, the Fujisaki model is simple in its operation presenting no problems for lookahead, production, prominence and pitch range. No complicated interpolation rules are needed, and the underlying form of the contour, consisting of impulses and step functions, is highly plausible.

2.6.3 The Fujisaki Model for English

I have adapted the Fujisaki model for English. This involved writing a computer program that could synthesize F_0 contours given the step and impulse input. The synthesized contours could then be overlayed on real F_0 contours for comparison. Impulses were marked with a location and an amplitude, and each accent was marked with a location, a duration and an amplitude. The time constants α and β could be varied independently for each accent, although this may not have been needed as Fujisaki claims that these are constant for a particular speaker. After working with the model for some time, it was possible to examine an F_0 contour by eye and guess suitable parameters to feed into the synthesis program. These parameters could then be adjusted by hand so as to make the synthesized contour closely resemble the original.

For the types of sentences shown in Fujisaki's literature, the model worked very well. Making the accent shapes fit the contour's accents was much easier than the equivalent process for the phrases. The accents do not overlap and it is easy to spot where an accent occurs in a phrase by examination of the contour. As the phrase shapes are often hidden beneath the

accent shapes it was more difficult to tell if they were correctly positioned. It was quite easy to place the phrase shapes so that the contour was reproduced accurately, but it was difficult to be sure that the phrase shapes occurred in meaningful positions (i.e. near some kind of prosodic boundary).

The model worked well for neutral declarative types of F_0 contour. However problems became apparent when the analysis was widened to others types of English intonation.

Downstepping Contours and Final Falls

An inherent feature of the Fujisaki system is that most accents finish at or near their starting F_0 level. Any decrease in F_0 from accent beginning to end is due to the declination effect of the phrase component which is usually small. This makes it difficult to account for downstepping accents.

It is often the case that after the last accent in the phrase, the F_0 contour falls to a baseline. This is difficult to represent in the Fujisaki model because it is the phrase component which controls the declination pattern, not the accent component. Fujisaki is aware of this and suggests using a negative impulse after the nuclear accent to make the contour fall to the baseline. The locations of the positive phrase components are assumed to be related to the prosodic boundaries in some way, but if negative impulses are used, their position is determined by a different mechanism, i.e. the same mechanism which governs accent placement. This seems to complicate the original model. Liberman and Pierrehumbert (1984) note this and go further to show that the utterance final lowering effect they observe cannot be modelled even when using this negative component.

Low and Rising Contours

The accent component in the Fujisaki system is good at modelling the **H*** accent in the Pierrehumbert system. Rise-fall accents (**L* + H-**), are difficult to model as their rise time is often longer than their fall time. The accent component uses the same time constant for its rising section and its falling section, so it is unclear how the model is to deal with rise-falls.

More troublesome are low accents (**L***). There is no mention of how the model is to reproduce these accents so we have to postulate some means ourselves. Using negative step functions would be a possibility, though this fails to mimic the observed shape of low accents.

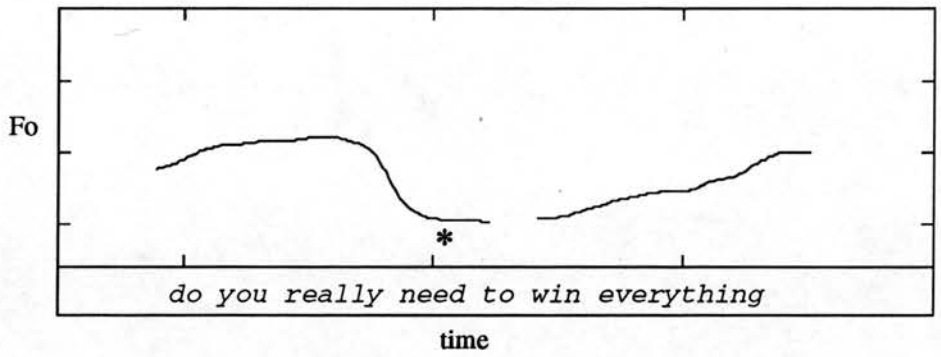


Figure 2.15: *This low-rise **H- + L*** contour is difficult to explain using the Fujisaki system. The sharp fall of the F_0 contour to the low accented syllable * is difficult to model using the accent component. This low accent is followed by a gradual rise, which is also difficult to model.*

While low accents are often preceded by rapid falls, it is rare they are followed by a rise as rapid as would be specified by using a negative accent shape. An alternative would be to use the falling part of a normal accent shape to represent the fall into the low and use a phrase shape to produce the subsequent rise. This too has problems. First, it means that the accent shape is not aligned with any stressed syllable, and second, even the slower time constant of the rising part of the phrase shape is often too rapid for the gradual rises which follow low accented syllables. Other schemes such as using backwards phrase components may achieve the desired effect, but this would require a total rethink as to how to make the production mechanism behave in a plausible manner. Figure 2.15 shows a contour that is difficult to analyse using the Fujisaki model.

Möbius's model of German seems to adequately describe many F_0 contours for that language, but crucially, it seems as if he restricts his work to non-interogative utterances, which avoid the problematic rising intonation patterns (Möbius and Pätzold, 1992). Hence the model may be problematic for languages other than English.

It doesn't seem possible to model low accents or gradually rising intonation effects in the Fujisaki system without radically changing the model. The problem with any of the suggested changes is that they strain the possibility of being able to map these inputs to a phonological description. Introducing accents that don't occur on stressed syllables and impulses which are determined by accent position greatly reduces the initial simplicity and phonological relevance of the model.

I have tested all of the above suggestions for producing low accents and gradual rises.

As well as being phonologically unintuitive, none seemed to fit the F_0 contours as well as the original model fitted the declarative contours.

2.6.4 Mapping within the Fujisaki System

If we consider the subset of native speaker F_0 contours that the Fujisaki model can synthesize and forget those which it can't, the model is the closest of the ones described in this chapter to being an accurate formal system. Although the actual implementation may be difficult, it should be possible to design an inverse filtering mechanism which can derive the inputs to the damped filters given their outputs (the F_0 contour). Thus it should be possible to determine the parameters of the Fujisaki model given an an F_0 contour. Alternatively a system could be set up where the parameters of the model were continuously varied until the best fit was achieved between the synthesized contour and the original F_0 contour. Such a system would constitute a F_0-intermediate mapping system. As the inputs to the filters represent easily interpretable quantities, i.e phrase and accents, it should not be difficult to design an phonology-F_0 grammar.

However, there is no point in attempting any of this considering the system's inability to model large classes of contours.

2.6.5 Fujisaki: Summary

It is disappointing that the Fujisaki model has proved to lack enough generality to be able to cover the intonation patterns of English. It is quite possible that the model does work very well for Japanese, but even so, its attractiveness for Japanese is diminished in light of these findings. However, the Fujisaki model is a useful system in that it points in the direction of models which aim for a good degree of accuracy and are not simply straight line approximations. It may be possible to introduce new components into the system to allow the additional contour shapes to be modelled, but I can see no way of doing this at present.

2.7 Comparison of Models

2.7.1 Redundancy

The previous sections have explained existing models and shown some of the problems associated with them. Figure 2.16 is an impressionistic diagram showing how each model relates

its phonology to its intermediate level and its intermediate level to the F_0 contour. This diagram is somewhat simplified, and we have shown Pierrehumbert's system as having a single intermediate target level (but see section 2.7.3 and figure 2.17). The arrows on the diagram represent the various mappings.

In all cases the synthesis mappings are straightforward, although the Fujisaki and the Dutch systems cannot model all native speaker F_0 contours and so in this sense their synthesis mappings can be said to be at fault.

The intermediate levels are shown in different positions relative to one another (e.g. the Fujisaki intermediate level is further to the left than the Pierrehumbert level). This is deliberate and is intended to indicate where these intermediate levels lie in the phonology-F_0 grammar of each system.

Let us take the intermediate level in Pierrehumbert's system to be a set of target points which need only to be joined by straight lines in order to create a F_0 contour. In principle these target points are unconstrained and can occur anywhere, it is the phonology-intermediate mapping which makes them occur in well behaved patterns. A set of these F_0 target values can be converted into a continuous F_0 contour with little effort. Likewise, it is easy to describe any F_0 contour as a series of unconstrained target positions. What would be difficult (impossible in Pierrehumbert's view) would be to link a set of F_0 derived target values to the phonological description. Likewise it is difficult to derive the target values from Pierrehumbert's phonology as complicated interpolation rules are needed. Putting this another way, we can state that the intermediate-F_0 grammar is simple whereas the phonology-intermediate grammar is complicated. Thus it is possible to describe the intermediate level in this model as being *close to the contour.*

The Dutch model is quite different. There is a simple relationship between the intermediate level and the phonology, but analysis of F_0 contours in intermediate level terms is difficult. As the intermediate-F_0 mapping can only synthesize a small number of the legal set of contours, we can say that this mapping is lacking, and as it is difficult to capture the features of a given F_0 contour accurately using the intermediate level, we can say that the F_0-intermediate mapping is troublesome too. The Dutch system therefore has a simple phonology-intermediate grammar and a complicated intermediate-F_0 grammar and so we describe this system as being *close to the phonology.*

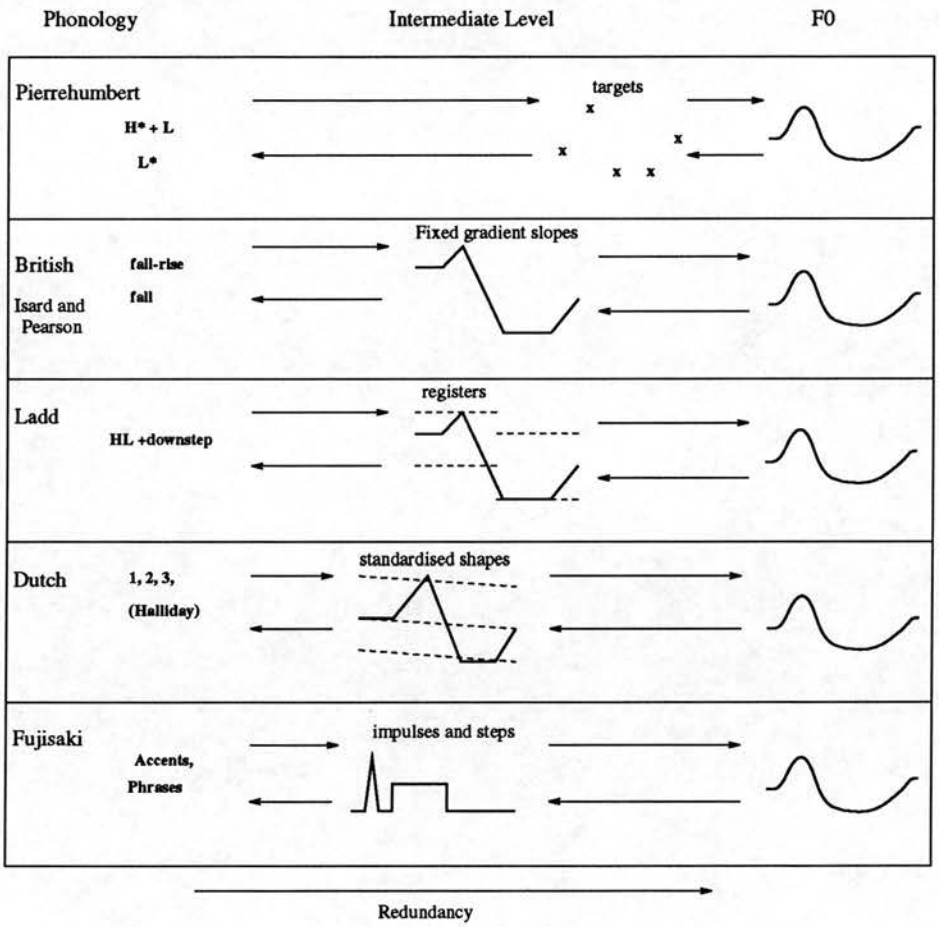


Figure 2.16: *A comparison of five phonetic models.*

The interesting fact is that when the various systems are compared in this way, we can state the general principle that the difficulty in performing a particular mapping between two levels is proportional to the "distance" between them.

Looking at Figure 2.16, we can say that the difficulty in performing a particular mapping is proportional to the length of the arrow between the two relevant levels.

The Dutch system is probably closest to the phonology, with the Isard and Pearson, Fujisaki and Ladd models being somewhere in the middle of the picture. In Isard and Pearson's model the phonology-intermediate mapping process still has to deal with accent size, while in the Fujisaki model the phonology-intermediate mapping has to specify accent size and duration. In Ladd's model, register settings and boundaries need to be determined by the phonology-intermediate mapping.

Another way to analyse this diagram is to think of the horizontal axis as representing *redundancy*. Redundancy is a very powerful concept that is frequently used in both linguistic research (Chomsky and Halle, 1968) and information engineering (Lathi, 1983). F_0 contours have high redundancy in that a single phonological description may describe many very similar F_0 contours. These contours can be described in the same way because of a consistent regularity in their shape. If a method of phonological description is a accurate one, and the F_0-phonology mapping is adequate, a large number of F_0 values specified at (say) 10ms second intervals can be completely described by a few phonological symbols. By using the synthesis phonology-F_0 mapping, an F_0 contour that is very similar to the original one can then be reconstructed. Expressing the continuous F_0 contour with a small number of discrete units means the F_0 contour has effectively been *coded*. The analysis mapping processes can be regarded as coding mechanisms which convert descriptions with high redundancy to descriptions with low redundancy.

Thus descriptions at the phonological end of the redundancy axis are tightly constrained in that there are only a small number of possibilities per unit time. At the F_0 end, the number of possible F_0 contours is vast. We can describe intermediate levels using this concept of redundancy. The Dutch intermediate level has a low redundancy as the number of possible combinations of the units is small. This tight constraint and low redundancy is what makes the phonology-intermediate grammar simple. A target system still has a large degree of redundancy in that there are many possible combinations of targets per unit time that will be described by

the same phonology.

In addition to redundancy changes between the F_0 level and the phonological level, there is also the question of numerical content. We have said before that the F_0 is a quantitative description whereas the phonology is a qualitative description. So somewhere in the mapping from F_0 to phonology the numerical content must be "lost". All the intermediate levels we have described still have a large numerical content. This is what makes the Dutch system "easy" to criticise: having an intermediate level with so little redundancy while still maintaining numerical content makes the description too inflexible. This might lead us to believe that a target system is therefore better. It is true that the numerical target system is flexible enough to describe most F_0 contours, but it is highly redundant and the intermediate-phonology mapping is then responsible for reducing this redundancy. This makes the intermediate-F_0 mapping more difficult.

Thus, there is no reason for saying that a close-to-the-contour level is better than a close-to-the-phonology one. The position of the intermediate level merely dictates where the "redundancy reducing" part of the work is to be carried out. The Dutch system can easily be shown to be inadequate as it stands, whereas a target system, due to its greater redundancy, still maintains the possibility of being powerful enough to model any F_0 contour. But any intermediate-phonology mapping in a target system will be difficult to design, and only when a system is formally proposed can we say with any certainty if it is better or worse than the Dutch system.

In the design of a phonetic model, the choice of "where" to put the intermediate level (or levels) is crucial as it decides the complexity of the phonology-intermediate and intermediate-F_0 grammars.

Many-to-One and One-to-One Mappings

The F_0-phonology mapping in the British school is a many-to-one mapping in the sense that many F_0 contours can be given the same phonological description. The notion of a many-to-one F_0-phonology mapping agrees with the traditional attitude to phonology which aims to describe a wide range of utterances with a small inventory of symbols. The many-to-one mapping principle also fits in with the idea of the F_0-phonology mapping being a redundancy reducing mechanism.

However, it would be wrong to assume that a F_0-phonology mapping is it necessarily a many-to-one mapping. In the Dutch model, there is a unique intermediate description for each type of accent, and this in turn represents a unique F_0 contour. Thus there is a one-to-one mapping between the phonological level and the F_0 level. The Dutch system does have a many-to-one mapping in that an F_0 contour from its legal set is intended to represent a much larger number of naturally occurring F_0 contours. Likewise, the Fujisaki model has a unique intermediate level description for each contour that is in the model's legal set.

If we think of a fall accent as representing a set of F_0 contours or parts of F_0 contours, then this is a many-to-one system. However, if we include pitch range and prominence in the phonological description, then a single phonological description will describe fewer F_0 contours. Segmental influence accounts for much of the remaining variation between F_0 contours of the same phonological class.

The F_0-phonology mapping may be thought of as a many-to-one mapping, but this normally arises due to one F_0 contour in the model's legal set being intended to represent many F_0 contours in the native speaker set. Often there is a one-to-one or a near one-to-one mapping between the F_0 level and the phonological level.

2.7.2 Well-Formedness Conditions for F_0 Contours

As no-one proposes independent F_0 well-formedness conditions, it becomes the case that every model's intermediate-F_0 mapping defines its legal set of F_0 contours. All the explicitly defined models (Isard and Pearson, Dutch school and Fujisaki) define the legal set as being those contours which are produced by the synthesis mapping. It is precisely this which makes these models bad at analysing F_0 contours: any contour which lies outside the model-specific set of legal F_0 contours is indescribable in the system, and therefore the model has no ability to understand or describe such a contour.

On the other hand, an unconstrained target based system has the ability to synthesize many contours that will not be in the native speaker set. This is problematic in that it allows the analysis procedure too much choice when determining the intermediate description for an utterance; effort is wasted as the analysis procedure may look for contours which are not in the native speaker legal set.

Thus the need for a synthesis mapping to be able to create *all* and *only* the contours of

the language is not merely a desirable property, it is a *necessity* as this is what determines the expectations the model has of naturally occurring F_0 contours.

2.7.3 Comparing Phonological Descriptions

Tune

Section 2.7.1 explained the concept of redundancy as applied to phonetic modelling and showed how this could be represented diagrammatically (figure 2.16). Figure 2.17 is a similar diagram designed to demonstrate the differences between the tune aspects of the phonologies of Pierrehumbert, Crystal and Halliday.

At first one might think that the methods of phonological tune description outlined above are radically different from one another. Halliday uses the numbers 1 to 5, Pierrehumbert uses **H** and **L** tones and Crystal uses dynamic features such as rise and fall.

The main differences between these theories lie in what terms are used and what each theory states as being the underlying substance of intonation (tones, rises and falls etc). However, if we consider only the high level phonological *descriptions* of each system, and set aside the issue of mappings and grammars, we actually find that these tune descriptions are all very similar to each other. The evidence for this equivalence of description lies in the fact that it is easy to translate a description of an accent in one phonology to a description in an other phonology. Thus final **H*+L** is equivalent to "fall" is equivalent to "1". There is no one-to-one mapping between the phonological descriptions, but there is a large degree of overlap. All theories have the idea of pitch accent and boundary effects, all state that the nucleus accents are somehow different from pre-nuclear accents and so on.

Figure 2.17 aims to show that the high level phonological tune descriptions are similar and that it is the lower levels of the theories which diverge. Pierrehumbert uses tones and Crystal uses rises and falls, but what is significant is that at the these terms are seldom used as single items - they are grouped together to form compound descriptions, hence, rise-fall, fall–rise, **H* + L**, **L- + H*** etc. Thus in these phonologies there is a level above that of the basic static or dynamic unit. On this higher level there is more agreement between theories than on the static/dynamic level which is why it is possible to convert a description in one theory to a description in another. This is a very important point, as it shows that there is a large amount of agreement across theories as to what the highest level of the phonology is describing. At the

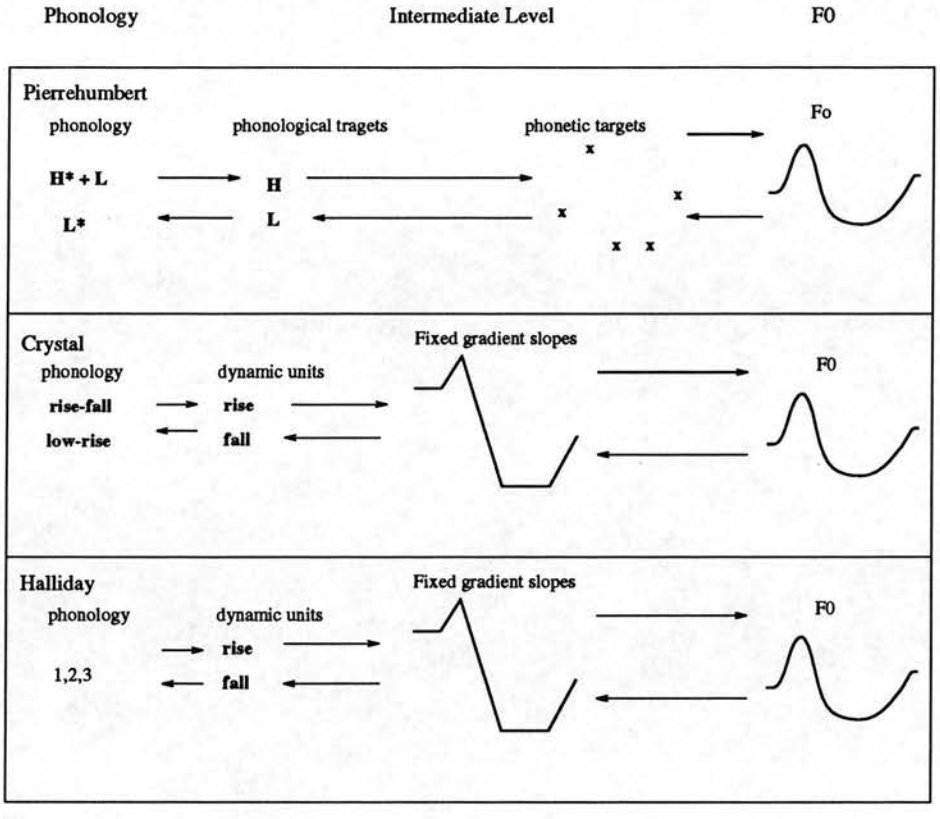


Figure 2.17: *A comparison of three phonologies.*

start of the chapter there were doubts expressed about the solid foundations for phonological descriptions. The F_0 level had a rigid definition, but the phonological level was less clearly defined.

This might have presented problems if we had shown that there was little agreement between theories as this would imply that the phonological level was as arbitrary as the intermediate level. However, by noting that there is a large amount of agreement between phonological descriptions, we can see that the phonological level is better established than the intermediate level.

Comparing other Aspects of Intonation Phonologies

There is less agreement between theories on the subject of phrasing. All agree that there is a basic intonation phrase unit but disagreement exists over whether other levels exist and what their relationship with one another might be. Section 2.2.2 discussed this in some detail. Much disagreement also exists regarding the issues of prominence and pitch range, as mentioned in section 2.2.3.

2.8 Conclusion

This chapter has given a review of the most important existing phonetic models. None of these have proved sufficiently powerful for our purpose of finding a formal method of linking intonational phonology and F_0. Some systems, such as the Dutch model are inadequate due to their inflexibility. Others, such as the Pierrehumbert system are potentially still capable of being used as the basis of a formal system, but much more work would need to be carried out in order to fully specify a phonology-intermediate grammar. Fujisaki's model proved particularly attractive, but if this system was to be used, it would have to be adapted so as to model low accents and slowly rising contours.

Chapter 3

A New Phonetic Model of Intonation

3.1 Introduction

This chapter explains the theory of a new phonetic model. The framework used is that of section 2.1 where the model is described in terms of levels, grammars and mappings. The model presented here is not complete, and therefore cannot be claimed to be fully formal. However, the division between the complete and incomplete parts of the model is quite distinct, and indications as to what is required to complete the model are given at the end of this chapter.

The first part of this chapter describes an intermediate level of description and the relationship between this intemediate level and the F_0 level. The intermediate-F_0 grammar is fully defined in that it completely specifies how to classify pitch accents once they have been located, and how to synthesize F_0 contours given an intermediate level description for an utterance. There are still unresolved areas, most importantly the lack of a formal method of detecting the presence of a pitch accent from an F_0 contour.

A new phonological system of description is presented in the second part of the chapter. The phonological system is complete with respect to tune; but phrasing, scaling and timing are left incomplete, largely due to the considerable uncertainty that prevails concerning how to describe these phenomena. Owing to the incompleteness of the phonological description, the phonology-intermediate grammar is only completely specified as regards tune. The chapter finishes with indications as to how to complete this grammar.

Chapter 4 explains the computer implementation of the new model. The main objective of the implementation was to test that the model really was formally defined. The implementation of the F_0-intermediate mapping was the most difficult due the lack of any formal specification

of how to locate pitch accents. The performance of this mapping was tested by using an objective assessment method that compares the transcriptions produced by the computer with hand labelled transcriptions. In many cases the computer transcriptions were very similar to the hand transcriptions but in a number of cases the computer was significantly worse. The difference in performance was nearly always due to the human labeller's ability to detect pitch accents reliably.

A intermediate-F_0 synthesis mapping was also developed. This was assessed by labelling F_0 contours, synthesizing new contours from this description, and comparing them to the original contour. An objective assessment method was developed which measured the similarity between the synthesized contours and the originals.

Appendix B shows many examples of F_0 contours which have been labelled by a human labeller (the author) and a computer. Also shown in the appendix are synthesized contours produced using the intermediate-F_0 mapping. The comparison of these contours with the originals is helpful in forming an impression of the synthesis mapping's accuracy.

Chapter 3 explains the theory of the new model, while chapter 4 explains its implementation and testing. This chapter is therefore necessarily theoretical in approach, while the following chapter gives more solid evidence for the claims proposed. Throughout, it may be helpful to refer to Appendix B so as to gain an impression of how the model labels and synthesizes F_0 contours.

3.2 Data

3.2.1 Material

It was stated in section 2.1.2 that the general aim of a phonology-F_0 grammar was to be able generate all the F_0 contours in the native speaker legal set and no others. Due to the lack of well-formedness conditions, the only sure way to assess a phonetic model is to test it on every contour in the native speaker set. This is clearly impossible as the set of legal F_0 contours is effectively infinite. A way around this problem is to choose a subset of the legal set and test the model on that subset. A humanly manageable amount of data might be typically between 10 and 1000 contours, which only constitutes a minute fraction of the legal set. To derive some significance from such a small fraction of data, careful selection procedures must be used

so that the contours used are *representative* of the larger set. This is a common procedure in statistics where the behaviour of a large population is derived from the behaviour of a small population. The crucial issue that arises is how best to pick the representative set.

A dilemma immediately confronts any experimenter when choosing a representative set of F_0 contours. On the one hand, the experimenter usually wishes to investigate a particular area of intonation, and therefore wishes to collect data which relates to that area. The temptation is therefore to choose F_0 contours which exhibit this particular intonational effect. This approach can distort the data in that as the experimenter has so much control over the form of the data, he or she can consciously or subconsciously influence the data in order to support or reject a particular hypothesis.

If the experimenter does not follow this approach, and collects more unconstrained speech data (say by choosing contours at random) the chances are that the occurrences of the phenomenon under investigation will be very sparse in the data and may be influenced by so many other effects that it may be difficult to deduce anything about the particular phenomenon under investigation.

To try to counter this problem, two sets of data were chosen - one that was defined using the first phenomenon-specific criteria, and the other using a more random selection procedure.

Set A consisted of utterances carefully designed and spoken by the author, each intended to exhibit one or more particular intonation effects. All the nuclear pitch accent types described by O'Connor and Arnold (1973) were included. Most of the variation was in tune association and tune type, but a variety of phrase effects, pitch ranges and prominences were also covered. The data was therefore recorded before the model described below was developed and many of these sentences were chosen to test the Dutch and Fujisaki models. It was thought that the Dutch system had difficulty in expressing variation in accent prominence and that the Fujisaki system would find it difficult to model low accents, so many examples of these effects were included. Although it may be claimed that the new model was developed to fit the data, it would be wrong in this case to think that the data was chosen to support the model.

Set B was recorded for another set of experiments unconnected with this thesis. The author had no control over the design and collection of these sentences. The text material for these sentences was taken off the UNIX news network and read by a speaker who had no specialist knowledge of intonation. The material was often in the form of "chatty" text which frequently

contained unusual grammatical constructions and was often more similar to spontaneous speech than normal text. This was advantageous in that as the speaker was instructed to speak in a chatty style, a wide coverage of intonation accents and effects was present. The speaker was told to read the messages "expressively" so as to capture the chatty nature of the text which was often sarcastic and humorous. Because of the free nature of the material and the speaker's lack of specialist intonation knowledge, the intonation content of this data could not be specified before the recordings took place.

Many sentences in set A were designed with a "minimal pair" philosophy, where two sentences of the same text would differ only in which word carried the intonation accent or on which type of intonation accent was used on a word. Thus it was hoped that the differences between accents types and accent alignment could be investigated. Set B was not as unconstrained as free, conversational speech would be, but was considerably more "random" than set A.

It would be impossible to test a model on every contour in the native speaker legal set. The subsets of the native speaker set that these utterances represent cannot be guaranteed to contain all the important intonational phenomena of the language, but data set A did contain most of the intonational phenomena commonly discussed in the literature. Because of the clean and controlled nature of the speech, set A was used to test and develop the intonation model. Data set B, on the other hand, was intended to be a tougher test of whether the intonation system developed on the controlled data was general and robust enough to cope with more natural intonation.

When hand labelled, data set A contained 164 pitch accents and 136 intonation phrases. Data set B contained 301 pitch accents and 156 intonation phrases. Appendix A has a full listing of the text of the sentences.

3.2.2 Collection

The data in both sets were recorded in a sound insulated room. The speech was recorded at 20KHz sampling rate using a close-talking microphone.

To make the measurement of F_0 easier, all the recordings were made with the use of a *laryngograph* which is a device which fits around the neck and measures impedance across the vocal folds (Abberton et al., 1989). This device outputs a waveform which has a much more

readily detectable period than a speech waveform. A simple peak picking algorithm was used to produce a F_0 contour from the laryngograph waveform. Occasionaly, the algorithm made an error in determining F_0: these errors were corrected manually by examining the laryngograph waveform and measuring the time interval between cycles, which could then be converted to a frequency value.

3.3 A New Intermediate Level and Intermediate-F_0 Grammar

This section describes a new intermediate level of intonational description and the intermediate-F_0 grammar that exists between it and the F_0 contour. Section 3.4 describes the phonology of the new system and the phonology-intermediate grammar.

3.3.1 Developing the Fujisaki System

My initial idea was to produce a completely formal analysis/synthesis system, and I examined the existing phonetic models to assess their suitability. Initially, the Fujisaki system looked the most promising of all the phonetic models. Its attractiveness lay in its ability to model some accents very accurately, particularly the fall accent. Indeed, if we were to constrain our task to fall accents, Fujisaki's system might suffice. The problems with the system arose when accents other than simple fall (**H*** or **H*+L**) accents were examined. Because of this ability to model the fall accent well, the system was chosen as a starting point, and efforts were made to see if the system could be altered to accommodate other types of pitch accent.

From examination of the contours that were difficult to synthesize with the Fujisaki system, a persistent pattern emerged which explained why the system had difficulties.

An inherent feature of the Fujisaki accent shape is that the size of the rise part is always approximately the same as the size of the fall part. From the function defining the accent component, it is clear that in isolation, the rise and fall parts of the accent must be equal. As the accent shape is superimposed on a phrase component, the phrase component may give a tilt to the accent shape which results in either an overall rise across the accent or an overall fall. The important point is that the overall rise or fall across the accent is governed by the phrase component, and as the phrase component usually does not have a steep gradient, the overall rise or fall will be slight.

What was noticeable in the data from set A was that a significant number of accents did not have equal rise and fall amplitudes and that in many cases the fall amplitude was considerably larger than the rise. Figure 3.1 shows three typical pitch accents from data set A. Only one of these contours is easily analysed within the existing Fujisaki framework, the other two contours show pitch accents with large falls.

It was obvious that the Fujisaki model was insufficient as it stood. Two amendments seemed possible:

1. The phrase component could be altered so as to provide sharply rising and falling phrase shapes, thereby giving more control over the amplitudes of the rise and fall parts of the accent component.

2. Proposing a new accent component that has different rise and fall characteristics.

Option 1 is unattractive as this means that the timing of the phrase component is influenced by the accent position, which implies that the phrase and accent shapes are not independent. The phonology-intermediate grammar for the original Fujisaki system is quite simple: if the phrase component was allowed to determine accent shape, this simplicity in the phonology-intermediate grammar would be lost. Option 2 seemed more attractive and was examined.

There seemed to be a consistency amongst the shapes of all the falls in the data; likewise there was consistency amongst all the rises. Thus if the rise and fall parts of the accents could be modelled independently, a common accent-describing formula might be possible.

From extensive examination of F_0 contours, the basic specification for a fall shape was deduced. This specification stated that the fall shape had to start with a short level section which gradually fell off, have a straightish mid-section which was falling rapidly, before tailing off to a level section again. This specification can be fulfilled by the shape of a cosine curve between 0 and π radians (see figure 3.2). However, when this shape was tested on the accents, it was found to be too "straight"; the transition between the initial level portion and the steep fall happened too quickly. An amendment to this basic curve shape proved much more suitable in accurately modelling fall shapes.

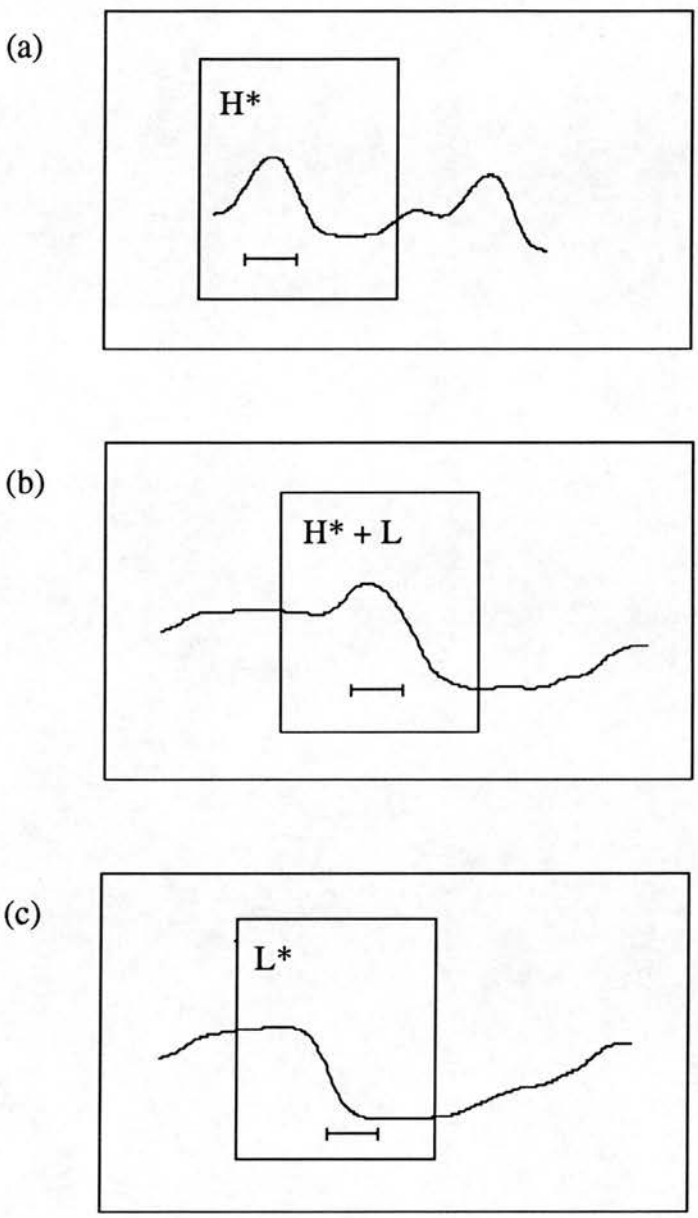


Figure 3.1: *Three F_0 contours from data set A. The relevant pitch accents are marked by rectangles and their Pierrehumbert accent type is shown in the corner of the rectangle. The horizontal bars show the position of the vowel of the accented syllable. Accent (a) can easily be modelled by the Fujisaki system as the rise and fall parts of this accent are approximately equal in amplitude. Accent (b) clearly has a larger fall than rise, whereas accent (c) has no rise near the accented syllable.*

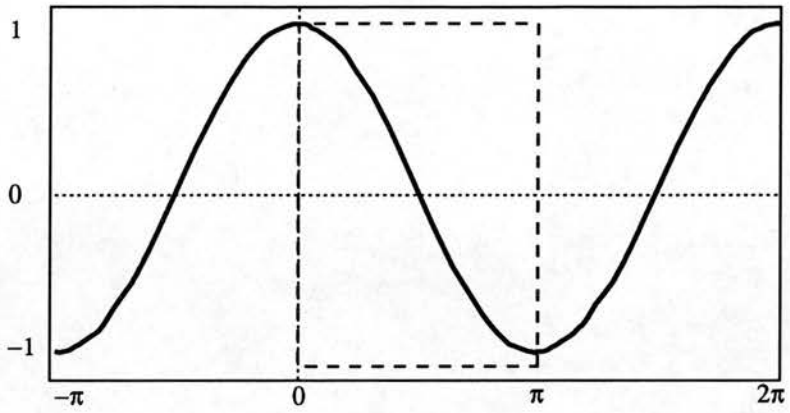


Figure 3.2: *A cosine curve. The area in the dashed rectangle is a similar shape to the shape of the falling part of a pitch accent.*

3.3.2 An Equation to Describe Pitch Accents

Two general functions were discovered that could produce the type of curve required. Both of these had a variable degree of curvature which allowed control over how sharp the level-to-falling "corner" was.

The general form of the first equation is given below for x and y values, where γ denotes the degree of curvature.

$$y = \begin{cases} sin(x) & \gamma = 1 \\ C_1.sin(sin(x)) & \gamma = 2 \\ C_2.sin(sin(sin(x))) & \gamma = 3 \end{cases} \tag{3.1}$$

Here, a sine function is recursively applied to the output of a sine function. The heavier the recursion, the longer the level sections become and the steeper the fall section becomes. A large order of recursion (> 100) will produce a shape with sharp right angle corners. The constants C_n are used to normalise the amplitudes so that each function's output lies between -1 and 1.

The second equation is derived from a monomial[1]. A monomial of the form $y = x^n$ will have a zero gradient at the origin and will pass through the point (1, 1). By copying and rotating the section of the curve between (0, 0) and (-1, 1) to the space between (-1,1) and (-2,2), we can

[1] A monomial is a function of the form $y = a.x^n$, i.e. a function with only one x term.

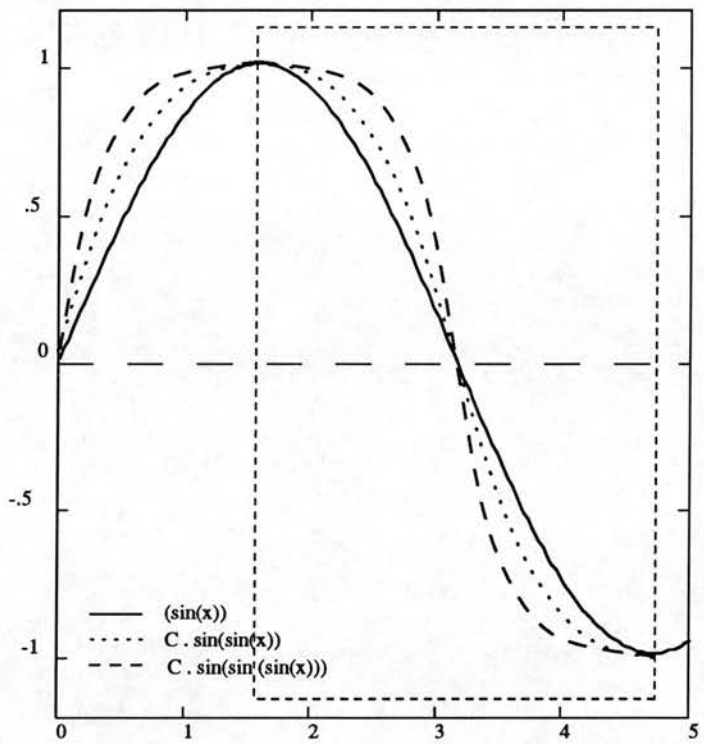


Figure 3.3: *The family of recursive sine functions. The rectangular area shows the part of the curve that can be used to model a pitch accent fall. The solid line shows a normal sine function. The dotted line is the sine of a sine function. This curve is clearly flatter near y = 1 and y = -1. The dashed line takes the level of recursion one more stage and the corner of the curve after y = 1 is delayed still further.*

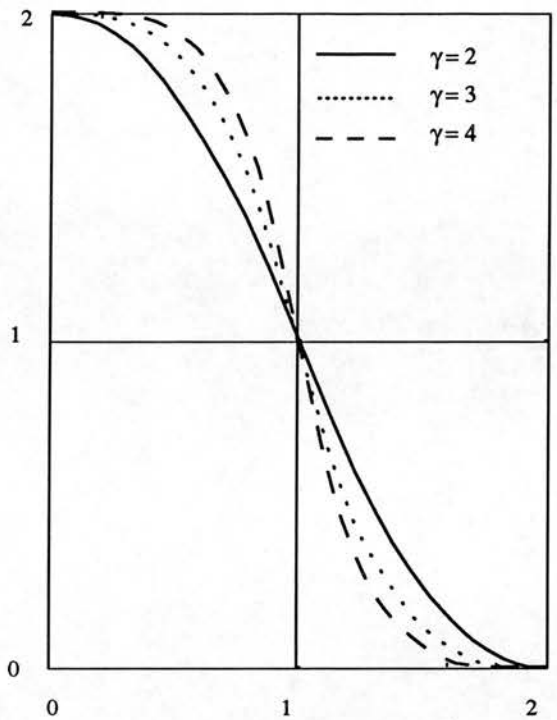


Figure 3.4: *The family of monomial functions. As the power of the function increases, the "corners" of the curves bulge further out. For very high values of γ the corners will tend to the points (1, 2) and (1, 0).*

produce a shape similar to the family of sine curves. The equation (3.2) shown below has be shifted and scaled so as to lie in the region (0,0) to (1,1). See appendix C for a full derivation of this function.

$$\begin{aligned} y &= 1 - 2^{\gamma-1}.x^{\gamma} \quad 0 < x < 0.5 \\ y &= 2^{\gamma-1}.(1-x)^{\gamma} \quad 0.5 < x < 1.0 \end{aligned} \tag{3.2}$$

The two families of functions produce similar curves, but there are some slight differences. The sinusoid equation can only produce functions for discrete values of γ whereas the monomial equation can produce curves for fractional values of γ. This means that in practice the monomial equation is more flexible. However, the sinusoid equation seems more "natural"; the continuous nature of the function ensures all of its derivatives will be continuous. The monomial equation was used because of the ability to vary γ continuously,

This equation is defined in the space (0, 0) to (1, 1). In order to be fitted to real F_0 contours, the function needs to be scaled on both the x and y axes. This is achieved by the use of two

constants, A and D, which represent the amplitude and duration of the fall. The equation in terms of timing and frequency is shown in equation 3.3.

$$\begin{aligned} f_0 &= A - AC.(t/D)^\gamma \quad 0 < t < D/2 \\ f_0 &= A.C.(1 - t/D)^\gamma \quad D/2 < t < D \end{aligned} \quad where \;\; C = 2^{\gamma - 1} \tag{3.3}$$

A suitable rise shape was found by reflecting the fall shape in the y-axis. Although it was pleasing to discover a generic accent function, the independence between the rise and fall parts of the accent still had to be preserved because there did not seem to be any simple mechanism relating the amplitudes and durations of the rise and fall parts of accents.

Figure 3.5 shows the contours in figure 3.1 with the rise and fall shapes superimposed. It can be seen quite clearly from this diagram that the rise and fall shapes fit very closely to the original F_0 contour.

A computer program was developed that could synthesize the monomial rise and fall shapes given their amplitudes, durations and γ (the "coefficient of curvature")[2]. All the pitch accents in data set A were then tested to see if they could be modelled with the rise and fall shapes. Nearly every accent in the data set was modelled accurately by using the rise and fall shapes, and in nearly every case a γ of about 2 was the most suitable. With these encouraging results, it was clear that the rise and fall shapes could overcome many of the problems associated with the modelling of pitch accents in Fujisaki model.

3.3.3 Modelling Non Pitch Accent Parts of the Contour

Between Pitch Accents

Fujisaki used the output of impulse excited critically damped second order filters to model phrase effects. These shapes can often fit F_0 contours well, and the physiological plausibility that Fujisaki claims for this approach is another desirable feature. However, this shape was also shown to be incapable of modelling sections of slowly rising contour such as that in figure 3.1 (c).

From examination of the F_0 contours in data set A it was seen that most of the movement in a F_0 contour occurred in the vicinity of its pitch accents. Except at the beginnings and ends

[2] See chapter 4 and specifically section 4.5 for more on computer synthesis of F_0 contours.

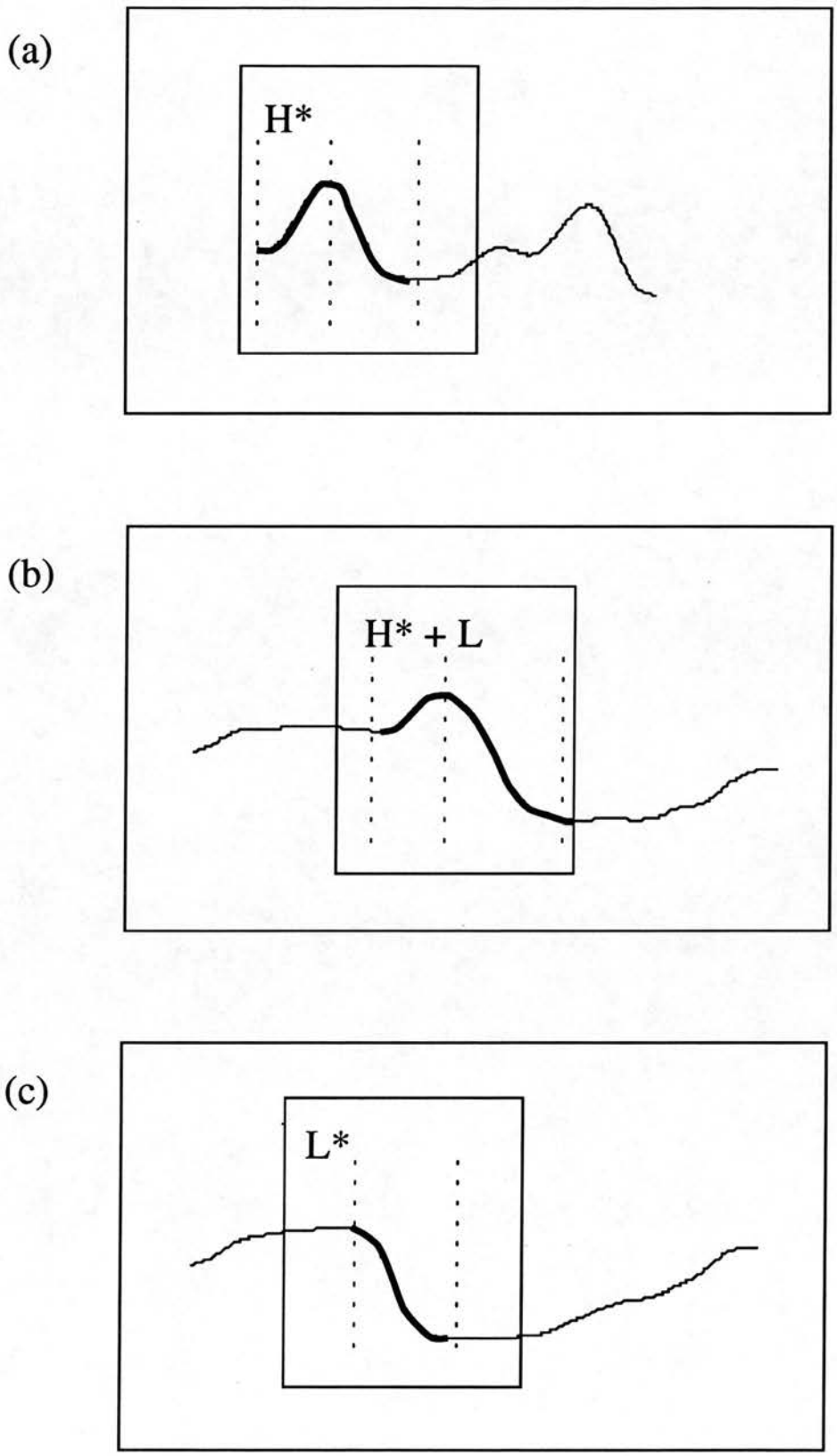


Figure 3.5: *Three pitch contours with the rise and fall shapes superimposed in bold. Contour (a) has a rise and fall of approximately equal amplitudes. Contour (b) has a fall that is much bigger than the rise, while contour (c) can be modelled by only using a fall.*

of phrases, the F_0 contour nearly always followed a straight line.

It therefore seemed appropriate to allow a third element to join the rise and fall elements. This element was described by a straight line which in principle could be of any length and gradient. This new element was termed the *connection* element, due to the fact that it connects the rises and falls of different pitch accents.

In most cases, the connection elements of a contour have no phonological function, the F_0 contour is merely "coasting in neutral" to the next pitch accent. However, in the low-rise accent (such as that shown in figure 3.6 (c)) the connection element has a phonological function (it is the "rise" in "low-rise").

The rise and fall shapes have zero gradient at their start and end which means that their join is continuous and smooth. This is not the case if a rising or falling connection element is joined to one of these shapes. In most cases the connection elements are not steep, and so although a singularity occurs at the boundary, it will be small.

The Starts and Ends of Phrases

Often, the start of a phrase rises fairly sharply, levels off and then gradually declines. Fujisaki's phrase component captures this behaviour accurately. Alternatively, this behaviour could be interpreted as a short section of rising contour followed by a connection element.

At the ends of phrases, F_0 contours often rise sharply (figure 3.6 (b)), particularly in some types of question. This behaviour can be interpreted as a connection element followed by a sharp rise.

The shapes of these phrase-initial and phrase-final rises were very similar to that of the rise shape that had been developed for pitch accents. It therefore seemed appropriate to allow the rise shape to be used for the modelling of these phrase boundary rises as well as the rises of pitch accents.

Sequences of Elements

The question arises, "are there well-formedness conditions for sequences of elements?" To put it another way, can the elements occur in any order? From examination of the labelled contours in data set A, it seemed that in principle any element could follow any other. The only restriction was that there was no occurrence of two contiguous connection elements. This seems quite

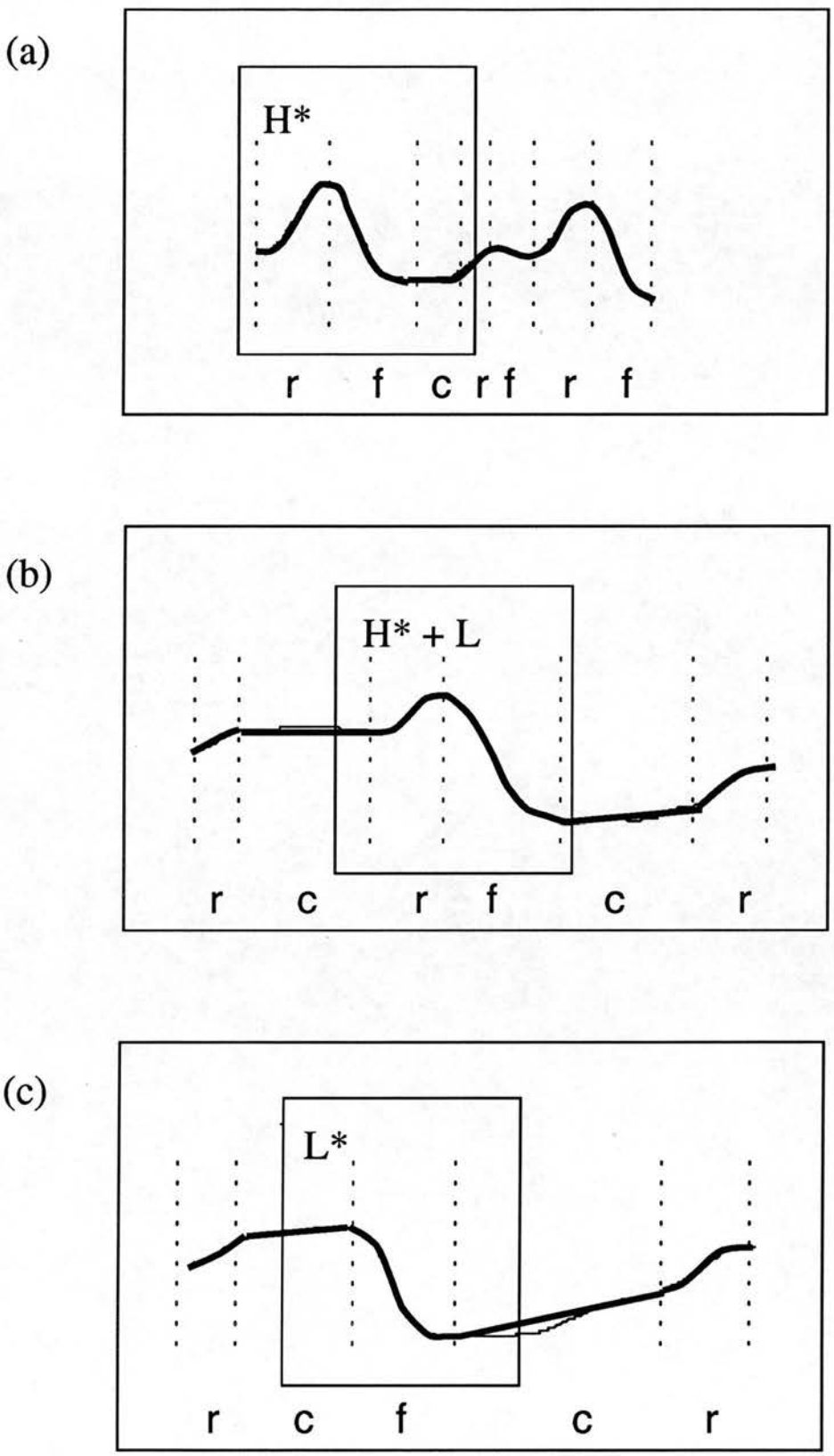


Figure 3.6: *The three contours have been marked with rise, fall and connection elements. The boundaries between the elements are shown with dotted lines, with the type of each element shown below. The thick solid lines show contours which have been synthesized from the RFC description for each utterance. These lines have been superimposed on the original F_0 contours for comparison.*

a reasonable restriction as the rise and fall elements can be regarded as conscious "actions", whereas the connection element carries little intonational meaning. It could be assumed that the only way to end a connection section is by a rise or fall, and that a new connection element cannot be started without a similar conscious action. This issue touches on the phonetic and psychological "reality" of the rise, fall and connection elements which will be discussed in more detail in section 3.5.1.

There is obviously a practical limit to the number of fall and rise elements that can occur in sequence: a series of rise elements will eventually hit the top of the speaker's pitch range and prevent the speaker from uttering any more. Likewise with falls and the bottom of the pitch range.

3.3.4 Downdrift

Most of the intermediate level descriptions discussed in chapter 2 deal with downdrift/declination. There did not seem to be any simple way of including a declination effect within this model. Each element has an amplitude which can be used to control all F_0 scaling effects. Downdrift arises in the system through the use of connection and fall elements. The majority of connection elements have a gradual fall off, and as most fall elements have larger amplitude than the rise elements, this also contributes to the downdrift effect.

3.3.5 Outline of the New Phonetic Model

- F_0 contours can be divided into a linear sequence of non-overlapping, contiguous *elements*.
- Each section is labelled with one of three fundamental elements: *rise*, *fall* or *connection*.
- The elements can occur in any order, with the exception that two connection elements cannot occur in sequence.
- Rise and fall elements are given by equation 3.3. They can be scaled to any extent on the frequency or time axis.
- Connection elements are straight lines of any gradient or duration.

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

Table 3.1: *Example of RFC description*

- Fall elements are only used to represent falling parts of F_0 contours which are associated with a pitch accent. All falling parts of F_0 contours associated with pitch accents are represented by a fall element.
- Rise elements are used to represent rising parts of F_0 contours which are associated with a pitch accent. All rising parts of F_0 contours associated with pitch accents are represented by a rise element.
- Rise elements may also be used at the beginnings and ends of phrases where there is a sharply rising section of contour.
- Connection elements are used everywhere else; specifically to model parts of contour which do not have a pitch accent or a phrase boundary rise.

The new intermediate description is a list of elements, each with a *type*, an *amplitude* and a *duration*. For example, utterance A.13 (the first phrase of which is shown in figures 3.1 and 3.6) was given the hand labelled RFC description shown in table 3.1.

Although this intermediate description system grew from the Fujisaki model, the two systems have diverged considerably. Among the main differences are that the new model uses

a linear sequence of elements whereas the Fujisaki system superimposes its accent component on its phrase component. Also, by stipulating that the amplitudes of the rise and fall elements of a pitch accent can vary independently, the number of possible accent shapes is greatly increased. In many ways the above system could be considered similar to the Dutch model as it uses linear sequences of rises, falls and straight lines to model contours. This may be true, but the above model is not constrained by the strict levels and declination lines of the Dutch model.

The above outline forms the basis of the formal specification of the model. The major remaining problem with the system is that there is no definition of how to determine what is and what is not a pitch accent from examination of a F_0 contour. When hand labelling, it was not usually difficult to determine pitch accents, although sometimes it was helpful to listen to the utterance before labelling. Phrase final boundary rises were easy to mark, but phrase initial rises were often very small and sometimes it was difficult to decide whether to mark one or not. Hence, the labelling of F_0 contours still requires some linguistic expertise. In chapter 4 the consequences of this will be examined further.

Also of importance is that the system makes no attempt to model segmental influence. Segmental perturbations caused by obstruents are simply ignored, and it is at these points that the synthesized contours differ most from the original contours. Ignoring this form of segmental influence is justified because it is not really an *intonational* phenomenon.

The importance of recreating these segmental effects in synthesized contours is debatable (see Silverman (1987), but the fact that the contours that the intermediate-F_0 mapping produces are free of segmental perturbations has the effect of making the model's legal set different from the native-speaker legal set. As shown in the review of models in chapter 2, this can have important reprecussions for a model's analysis capability. This issue is important, and is examined in more detail in chapter 4.

Segmental scaling effects are still present in the RFC description as the descriptions are influenced by intrinsic vowel height and duration.

3.3.6 Intermediate Level: Summary

It was shown that the rise/fall/connection (we will use the shorthand "RFC" hereafter) model was able to model all the F_0 contours in the two data sets. The synthesized contours were often very similar to the original F_0 contours, so the intermediate-F_0 synthesis mapping was found to

be very accurate. Chapter 4 gives empirical evidence for the closeness of the fits.

It was also relatively easy to hand label the contours in the database. There were few ambiguous labelling decisions and most of these concerned the question of where to place element boundaries rather than which elements to use.

By allowing the elements to be scaled freely on the time and frequency axes, we have made this intermediate description close to the F_0 level. Hence the redundancy is greater for this model than the Fujisaki model, which helps to explain why the intermediate-F_0 grammar is more successful in modelling intonation than the Fujisaki model. As was pointed out in section 2.7.1, the creation of an intermediate level with a good F_0-intermediate grammar does not prove anything on its own. As our overall goal is to design a phonology-F_0 grammar, this new intermediate level's worth can only be proven by showing that the increase in redundancy has not made the design of the phonology-intermediate grammar too difficult.

It is worth noting that although the RFC system is very free in its numerical variables, the elements themselves occur in strictly defined *phonological* situations. The rise and fall elements are only used to described pitch accents and boundaries. It is this fact that should enable the phonology-intermediate grammar to be simple, regardless of the fact that the redundancy in this intermediate description is relatively high.

3.4 A New Intonational Phonology

3.4.1 Issues in the Design of a New Phonological Description

I have designed a new phonological system of description and a phonology-intermediate grammar to link it to the RFC description. The reasons for designing a new phonological system, instead of using the Pierrehumbert or British school ones, are discussed below.

The RFC description does not sit comfortably with a tone-based phonology, as the RFC description is a "configuration" theory[3]. One could argue that it is the nature of the RFC description that is at fault: Pierrehumbert's system is fully valid and an intermediate level designed specifically for this phonology should have been designed. This issue can only be resolved when the new phonological system and phonology-F_0 grammar is compared to a completed phonology-F_0 grammar for the Pierrehumbert system. However it is surely worth

[3]However, see section 3.5.4.

building a phonological description and grammar to complement the RFC system - if only to perform the comparison with the Pierrehumbert model.

The British system was not used because it classifies pre-nuclear accents and nuclear accents differently. Pierrehumbert's method of describing pre-nuclear accents individually, and using post-nuclear sequences of units to produce a full inventory of nuclear accents seemed more attractive.

The design of a new phonology is governed by the need to link it with the intermediate level and also to the linguistic level. Problems arise as the linguistic level is not as fully defined as the intermediate level, which gives rise to a certain imbalance. We will take the attitude that where controversial linguistic issues govern an aspect of the phonology (such as how to describe prosodic phrase structure), that aspect will be left "open". This means that a mechanism for expressing the behaviour of that aspect will be provided, but the strict behaviour will be left for further work.

The design of a new phonology must address the issues of tune, phrasing, scaling and timing. The subject of tune is perhaps the most important, but also may be the easiest to resolve due to the large amount of agreement between theories about how to describe accents (at least as far as nuclear tones are concerned (see section 2.7.3)).

Intonational phrasing is a more difficult issue because of the large amount of disagreement about how many levels of intonational phrasing exist, how they relate to one another and what linguistic factors determine them. In principle, our new phonological description assumes a system similar to Ladd's *compound prosodic domains* (1992a) which is suitably flexible and open-ended. However, in practice, the means of expressing different levels of phrasing in RFC elements will not be examined. The RFC system shows phrase structure by means of boundary rises and pauses, so different levels of phrasing should be expressed by systematic behaviour in the sizes of the declination resets (boundary rises) and pauses.

The issue of scaling seems to be the most controversial in the literature. Most of the literature agrees that accents can have different prominences, but disagreements exist over how much choice speakers have in deciding these prominences. Regardless of whether one takes a phonological view of prominence (such as that of Ladd (1992b)) or a paralinguistic view (Pierrehumbert, 1980), a phonological system of description should have the ability to express the prominence of accents, even if the phonology itself is not responsible for determining the

sizes of pitch accents. Pitch range will be treated in a similar manner. Downdrift in the new phonology will be treated as primarily a phonological effect controllable by the speaker.

Section 2.2.4 identified three levels of timing. The first (what Halliday terms "tonicity") concerns the description of where pitch accents occur. This is easy to describe in all phonologies, as all we need say is that an accent is associated with a particular syllable. The second kind of timing, which is phonologically relevant to the description of accent type, is what helps to separate "fall" accents from "rise-fall" accents. This will be dealt with in some detail but must remain partially complete due to the influence of the third type of timing, segmental timing. Segmental timing determines how the segmental nature of a pitch accent's syllable governs the location and duration of the RFC units.

The phonology and phonology-intermediate grammar presented below is incomplete. It is probably better to leave a phenomenon open for the time being rather than be committed to a system which will easily prove inadequate. However, these problems must eventually be addressed as an incomplete formal system is not really a formal system at all. Many of the other systems that were reviewed here gave no indication of what factors control segmental timing, or of what the exact relationship between prosodic phrase structure and declination reset is. The system presented below may be lacking, but it is no worse than any other system as regards completeness.

3.4.2 Issues in the Design of a Phonological-Intermediate Grammar

A somewhat obvious principle learnt from the F_0 analysis work was that it is much easier to provide a way of linking two descriptions if they adequately describe the phenomena they were designed to describe. It is easier to label F_0 contours with the RFC system than the Fujisaki system, because it is difficult in the Fujisaki system to classify accents which don't fit into the system.

Thus the quality of the descriptions is of prime importance when trying to define a grammar that links them. The phonology-intermediate grammar and the phonological description system were designed together with this principle in mind. So long as the phonology is accessible from the linguistic level, and so long as it is powerful enough to describe intonational effects adequately, there is nothing problematic about designing the phonology so as to make the phonology-intermediate grammar as simple as possible.

The Redundancy of the RFC system

The primary purpose of the F_0-phonology mapping (with respect to tune) is to convert the continuous, numerical F_0 contour into a discrete set of qualitative units. The scaling of the RFC system has considerable flexibility in that for most pitch accents, which contain a rise and a fall, there are two scaling parameters to be produced by the phonology-intermediate mapping, rather than the single figure in most systems. Compared to the Fujisaki system, this system has a more complex phonology-intermediate mapping. This increase in complexity is acceptable as it is this flexibility in the amplitudes and durations of the rise and fall elements which makes the RFC system overcome the problems of the Fujisaki model. It may be possible to express the prominence of a particular accent with a single number or symbol; this would entail a study of the relationship between the amplitudes of rise and falls in a large data set. If no relationships were found, it could then be assumed that the two elements were phonologically independent of one another.

The phonological effect of downstep was very obvious in the data sets and is dealt with below. The RFC description says nothing directly about declination or downdrift as no simple model was seen to be appropriate. Although on average, F_0 contours started in higher positions than they finished, there were many contours which ended in rises. Declination was not directly modelled in the RFC description or the intermediate-F_0 grammar. Thus it is the task of the phonology-intermediate grammar to specify any downstep or declination effects.

The F_0-intermediate mapping filters out segmental perturbations as these are simply ignored. There is no mechanism for synthesizing these from an RFC description but this was not though necessary as they are not intonational effects. Silverman (1987) argues that while not being intonationally significant, these effects are important in the perception of the segments themselves, so if one was using the intermediate-F_0 mapping to construct F_0 contours for a speech synthesis system, it might prove wise to develop a method of adding these segmental perturbations to smooth F_0 contours. Intrinsic vowel height and duration is still present in the RFC description and must be normalised for in the phonology-intermediate grammar.

Segmental Normalisation

The segmental mapping problem is unique in that segmental effects exist in the F_0 contour but should not have any place in a phonological description. An RFC description derived

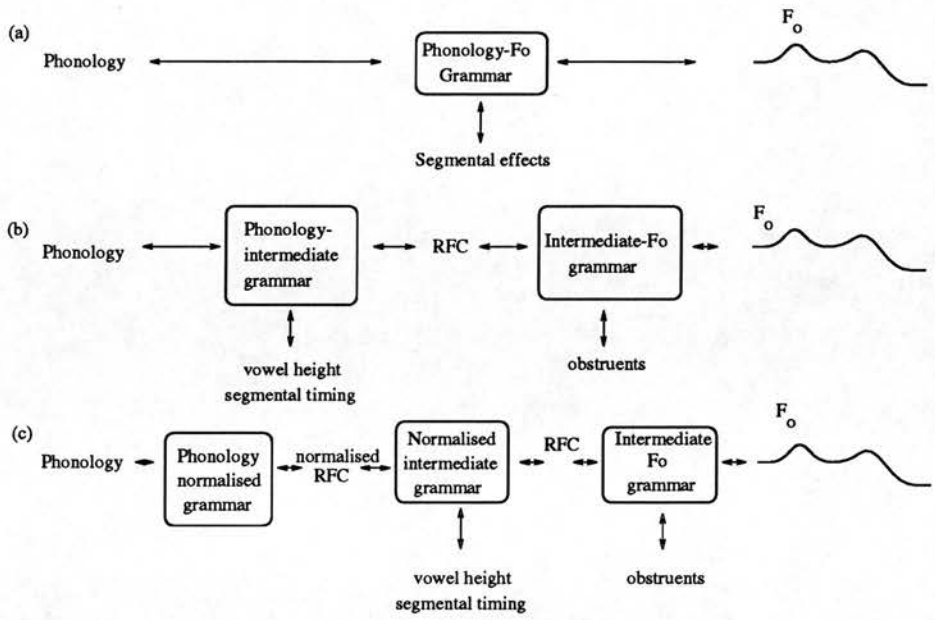


Figure 3.7: *Figure (a) shows the general form of the phonology-F_0 grammar. Figure (b) shows the division of labour in the current system and figure (c) shows the effect of introducing the normalised intermediate level.*

from a natural contour will be influenced by segmental scaling, which affects the amplitudes, durations and positions of the elements. It is therefore the task of the phonology-intermediate grammar to deal with these segmental effects. This is somewhat unattractive as the phonology-intermediate grammar should really be responsible for higher level aspects of the phonetic modelling process. A possible solution is to provide a *normalised intermediate level* between the intermediate level and the phonological level. This is shown in figure 3.7.

If we adopt this proposal, the phonology-intermediate grammar is then separated into the phonology-normalised and the normalised-intermediate grammars. This method has the advantage that the phonology-normalised grammar (largely corresponding to the old phonology-intermediate grammar) is only concerned with issues which are phonologically relevant. Any differences in the duration or amplitude of the normalised RFC elements will be due to phonological factors alone. This idea has not yet been adopted as the data sets were inappropriate for a study of segmental normalisation. The discussion at the end of this chapter and chapter 5 give indications of what further work is needed in this area.

3.4.3 Classification of Pitch Accents

The design of the phonological system took much the same approach as the design of the RFC system. Data set A was examined to see what patterns emerged and what classifications could be made.

The majority of the pitch accents in the database appeared as peaks in the F_0 contour. In RFC terms, these were realised as a rise element followed by a fall element, with the peak being the boundary between the two. The peak normally occurred somewhere near the middle of the vowel of the syllable. Most of the pitch accents which did not have F_0 peaks were simple fall accents occurring on syllables with low intonational emphasis, and were realised in the RFC system as a fall element occurring in the latter part of the accented syllable. The remaining pitch accents were of "low-rise" or "high rise" (**L***) type and had a rise section leading out of the accented syllable often accompanied with a fall element leading into the accented syllable,

All the accents except the **L*** group have a somewhat similar meaning (Ladd, 1983b) and therefore deserved to be included in the same class. The **L*** accents were grouped in a different class.

The names **H** (high) and **L** (low) were chosen to represent the two groups, although the terms "peak" and "valley" were also considered. This use of **H** and **L** is somewhat similar to Pierrehumbert's use and all the accents marked by a **H*** in Pierrehumbert's system are classed **H**, and all the accents marked **L*** are classed **L**, except for **L*+H** which is classed **H**. However, there the similarity ends. The new system uses two classes, but these are not meant to represent tones. The classes are only used to classify pitch accents (no boundary tones), and an accent can belong to only one class (no combination tones). In figure 3.1, (a) and (b) are marked **H** and (c) is marked **L**.

H Accent Features

The **H** class covered a wide variety of pitch accents and needed further subclassification. The most appropriate scheme seemed to be that of Ladd (1983b) who used features to classify the accents. Ladd's system had three features that distinguished **H** accents.

downstep Usually realised as a **H** accent with a fall and little or no rise. Often occurs in a sequence.

delayed peak used to differentiate a British fall from a rise-fall and a Pierrehumbert **H*** from a **L* + H**. In RFC terms, [+delayed] is realised by a rise section of greater duration than normal and a peak occuring later in the syllable.

raised peak This is equivalent to Pike's level 4. In RFC terms both the rise and fall sections have unusually large amplitudes.

A crucial difference between Ladd's description system and the one proposed here is that although Ladd's features make the tone phonology much easier, his system still distinguishes between **H** and **HL**. A **HL** accent has a larger fall than rise. Therefore Ladd has potentially twice as many accent types as the system presented here. Instead of merging the H and HL into one category - which would be unwise as these accents are different, it was decided to use the downstep feature to distinguish between them. Hence the downstep feature adopts a slightly lower level meaning, in that it marks any **H** accent whose fall is significantly large than its rise. The use of this feature in different contexts may be used to impart different meanings. Perhaps a feature name such as "fall" might have been more appropriate, but this name was discounted due to possible confusions with the RFC terminology [4].

The names of Ladd's features were changed, resulting in the following three features.

downstep Similar to Ladd's feature but also distinguishes H from HL.

late The same as "delayed peak".

elevated The same as "raised peak".

The features are *non-exclusive* so that an accent can be marked with more than one feature.

The initial **H/L** classification was unambiguous but the subclassification of accents using these features was more difficult. Sequences of downstepping accents were easy to mark but often it was difficult to decide if an isolated accent deserved the downstep feature. Clearly some accents had larger falls than rises, and some accents had roughly equal size rises and falls, but there seemed to be a large grey area in between. Therefore, only those with very large differences were marked. This labelling problem would still be present in Ladd's system.

[4] The main reason for choosing the names of the features was a very practical one. When labelling F_0 contours, the accent class (**H** or **L**) was subscripted with the first letter of the feature, thus it was important to have feature names that began with different letters.

Late accents were marked with little confusion. What proved most difficult was the use of the elevated feature. There seemed to be a continuum in accent size and although the largest accents received the elevated feature with confidence, others were labelled more arbitrarily.

L Accent Features

The number of **H** accents outnumbered the number of **L** accents in data set A by a factor of about 10:1. Although it is difficult to give a precise ratio, it seems that in natural speech **H** accents are also more common than **L** accents. Maybe because of the small numbers of **L** accents, there seemed less variety and the only subdivision that was made was whether or not a fall occurred leading into the accented syllable. This feature was termed *antecedent fall* (the default case being the fall leading *out of* **H** accents). This distinguished between a simple **L*** and a **H+L***, the latter having the antecedent fall.

A characteristic of all the **L** accents was that the contour after the accented syllable was always rising, either in the form of a rising connection element or a boundary rise element. A discussion of why the choice of phonological element is restricted after **L** is given in section 3.4.6.

3.4.4 Classification of Non Pitch Accent Phonological Phenomena

The two main remaining areas of phonological description involve the classification of boundary rises and connection elements.

Boundaries

A phonological element "**B**" was used to mark boundary rises. In the data, two main reasons were found for the phrase-final boundary rise. Firstly, as a *continuation rise* in which the contour rose at the end of the phrase to indicate that another, related phrase was about to follow. Secondly, the boundary rise was used as part of a compound accent construction such as a fall-rise (see figure 3.6).

Although the two types of phrase final **B** have quite different linguistic function, there seemed no obvious difference in the way they were realised. Hence, it seemed probable that the difference in perception between these two **B**s was indicated by higher level factors which were not expressed at the RFC level. Therefore no feature was used to distinguish between

these two types of boundary rise. (However, see section 3.4.6 on the different contexts of boundary units).

The other position in which the boundary rise occurred was at the start of phrases (figures 3.6 (b) and (c) both have this). This could be viewed as a "declination reset" (Ladd, 1988). Distinguishing this type of boundary rise from the phrase final one is simple owing to the clear-cut difference in position. Boundary rises at the starts of phrases were given the feature *initial.*

Often phrases are delimited by silence. In the cases where they are not, it was originally thought that there might be ambiguity as to whether the boundary rise was initial or non-initial. However, it soon became clear that all the occurrences of boundary rises that were used as part of a fall-rise accent were followed by silence, and as the continuation rise may be considered a form of declination reset, any boundary rise not bordering silence was labelled initial.

Classification of Connection Sections

In nearly every case, the RFC connection units were phonologically irrelevant in that they carried no meaning. They are still relevant in the RFC description as they are needed to reproduce F_0 contours accurately.

Occasionally, differences in the use of the connection elements themselves *were* found to change the perception of the meaning of the utterance. For example, take the case of a low-rise type of accent where contour is at a low F_0 value on the accented syllable, and slowly rises from that point until the end of the phrase. The use of a *rising* connection element here imparts a different meaning than when a standard level of falling connection element is used. This is one of the intonational effects that was most difficult to capture in the Fujisaki system. Figure 3.6 (c) shows this effect.

Although only a single phonological use had so far been found for the connection element, it seemed in keeping with the rest of the phonological system to give connection elements the phonological category **C** and use a feature *rising* to mark the connection elements that were rising.

Rising connection elements were occasionally found to be phonologically significant in pre-nuclear positions also. The "surprise-redundancy" contour (Liberman and Prince, 1977) has a rising connection element between the **L** accent and the **H** accent (as in utterance A.38 in

appendix B). If this sequence of accents is produced without a rising connection element, the surprise-redundancy effect is lost.

3.4.5 Summary of Phonological Elements and Features

H Made up of a fall element, optionally preceded by a rise.

downstep The fall is much larger than the rise. Often there is no rise element.

late The peak (the boundary between the rise and fall elements) occurs later than usual.

elevated The rise and fall elements have greater amplitude than usual.

L F_0 is at a minimum on the accented syllable.

antecedent A fall element leads into the accented syllable.

C The phonological class for the connection element.

rising The connection element is rising.

B Rises which occur at phrase boundaries.

initial The boundary element does not occur immediately before a major phrase break.

3.4.6 Well-Formedness Conditions for Phonological Elements

Initially all the contours in data set A were labelled without any restrictions on what sequences of elements could occur. After the contours were labelled, the well-formedness grammar was designed. Thus there is no real theoretical justification for the form of the grammar presented below, beyond adequately describing what sequences of units occurred. The grammar shown below is a finite state grammar, and it should be noted that not every sequence of elements that can be generated from this grammar actually occurred (there are infinitely many sequences). This grammar is simply the most compact grammar that could be written for the sequences of elements found in the data. The grammar is described below and shown in figure 3.8.

(B) {(C) {H | L}} + (C) (B)

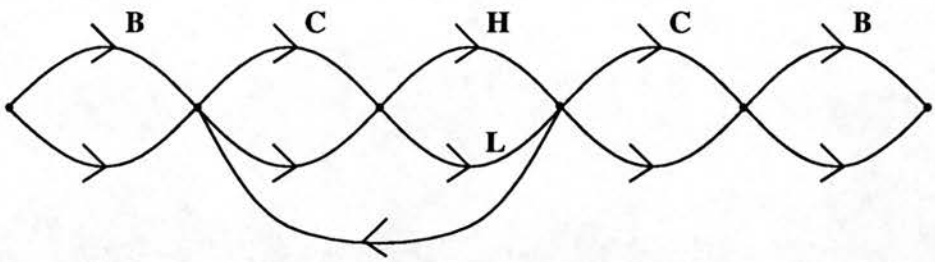


Figure 3.8: *A diagrammatic representation of the finite state grammar of well-formedness conditions. The arrows represent the directions in which one can trace a path through the network.*

Brackets denote optional elements, curly brackets group symbols together, "|" means "or" and "+" means "one or more occurrences of".

The grammar states that there must be at least one pitch accent in every phrase. A phrase can optionally start or end with a boundary element, and connection elements can occur between any other elements.

This well-formedness grammar, like Pierrehumbert's, implies that nuclear and pre-nuclear accents are described by the same accent classification system. However, it is clearly wrong to say that there is no difference between nuclear and non-nuclear accents. The distinctive feature of the nucleus is that it is the last accent, and this has the effect of allowing the last pitch accent to be associated with the "tail" of the phrase. Thus the "tail" (the last connection element and the boundary element) is grouped together with the nuclear accent to form a *nuclear configuration*, which has a different *overall* classification system from a pre-nuclear accent.

Nuclear Configurations

Table 3.2 shows most of the nuclear accent configurations allowed by the well formedness conditions. The British school name for each configuration is given. This chart makes clear that there is a many to one mapping between the phonology presented here and the classes of the British school. This is primarily because the new phonology is capable of making finer distinctions than the impressionistic phonology of the British School.

H

Any **H** accent which has no rising contour after it corresponds to a British school fall accent. Within this class, "rise-falls" and "high-fall" are included which are marked by the features **late**

Nuclear Configuration	British School Description
H	fall
H_d	fall
H_l	rise-fall
H_e	high fall
H C	fall
H B	fall, continuation rise
H C B	fall, continuation rise
H C_r B	fall, continuation rise
H C_r	fall, continuation rise
H_d B	fall-rise
H_d C B	fall-rise
H_d C_r B	fall-rise
H_d C_r	fall-rise
H_l B	rise-fall-rise
H_l C B	rise-fall-rise
H_l C_r B	rise-fall-rise
H_l C_r	rise-fall-rise
L_a C_r B	low-rise
L_a C_r	low-rise
L_a C B	low-rise
L_a B	low-rise
L C_r B	high-rise
L C_r	high-rise
L C B	high-rise
L B	high-rise

Table 3.2: *Nuclear accent configurations*

and **elevated** respectively. Combinations of these features can produce elevated-late accents. Connection elements occur phrase finally when the nucleus does not occur near the last syllable of the phrase.

Continuation Rises

The "default" use of either a rising connection element or a boundary element indicates a connection rise, i.e. that a related phrase is about to follow. The difference in amplitude between the end of the fall element and the end of the phrase may be important in signalling some relationship between the two phrases.

H_d + *rising configuration*

H_d + *rising configuration* (**H_d C B**, **H_d B**, **H_d C_r** or **H_d C_r B**) contours differ from normal **H** accents followed by continuation rises in that the fall element is usually of greater amplitude in the **H_d** case. There are four different versions of this configuration shown. The configuration with no connection element simply occurs because the nucleus is near the end of the phrase and so there is no room. The configuration with the rising connection element gives the impression of more emphasis to the nucleus than the configuration with the non-rising connection element. If no boundary element is present, the percept of this type of accent is still present, but the speaker sounds slightly uninterested or lethargic. The use of a **B** does not give rise to a big difference in meaning, but there is still a noticeable effect.

H_{dl} + *rising configuration*

These operate in a similar way to the previous group of accents. This type was quite uncommon in the data and not all possibilities shown here existed in the data sets. There is no reason to think that these do no occur: a **H_{dl}** + *rising configuration* accent is simply a **H_d** + *rising configuration* configuration with a late peak.

L_a and L

There are no examples of **L_a** accents in the data which are not followed by a rising contour. It is possible to say an **L_a** accent without a rise if one tricks oneself into saying a normal low accent and then not allowing the intonation to rise after the accented syllable. This sounds highly

unnatural, but it is possible to produce. The general tendency was always for the intonation to rise after the syllable, either with a rising connection element or a boundary element. Non-rising connections elements only ever occurred when the nucleus was near the boundary element, and so this element tended to be very short. **L** accents which did not have antecedent falls typically occurred much higher in the speaker's pitch range and can be considered equivalent to the British school high-rise.

All possible sequences of nuclear and post-nuclear elements occur (as generated from the well-formedness grammar) apart from the case of **L** followed by a non-rising contour. This one exception is still possible to produce, but does not seem to be part of the English intonational "lexicon".

3.4.7 The Phonology-Intermediate Grammar

The phonological system outlined above dealt mainly with the description of intonational tune. Other aspects of the phonological description will not be strictly defined here. Because of this somewhat incomplete phonology, the phonology-intermediate grammar is also incomplete. This section describes in as much detail as possible the relationship between the RFC level and the phonological level.

Both the RFC description and the phonological description use linear sequences of elements to describe intonation. It should therefore be possible to write a grammar that produces a sequence of elements in one description, given a sequence of elements in the other description. To a certain extent, such a grammar can be written, but problems arise because it is not just the type of RFC element that is important, but also its amplitude, duration and timing. Amplitude and duration are included within the RFC description, but timing relates to the relative position of the elements with respect to the syllable structure of the utterance. Therefore timing information is dependent on the position of the syllables and specifically the vowel of the syllable. Thus a complete phonology-intermediate grammar is impossible to specify unless timing information, which cannot be deduced from F_0 contour analysis alone, is allowed in the grammar.

Another problem arises with the definition of the phonology-intermediate synthesis mapping. The phonological description may use a single category (say **H**) for a group of RFC

elements of different amplitudes and durations. How then does one decide in the phonology-intermediate mapping what size the RFC elements should be? This also begs the question, "*why* do the RFC elements have different durations and amplitudes if they belong to the same phonological class?" This question will be addressed below in section 3.5.1, but it is useful to explain why some variations occur. Segmental effects account for some differences in element amplitude due to intrinsic vowel height, and also account for differences in element duration arising from intrinsic vowel duration (Kohler, 1991a). Other prosodic, non-intonational effects also influence the duration of the elements, for example in sentence B.5 (appendix A) the word "honest" is lengthened to achieve a particular prosodic effect though the pitch accent associated with this word is a simple **H**.

Proper phonological variation also occurs, most notably from the influence of prominence. The accents that are marked with **H** in the data have a wide variation in prominence, and until a phonological mechanism for describing prominence is formulated, this will remain unaccounted for.

Thus in the phonology-intermediate mapping synthesis mapping, we are in the position to say what RFC elements to use, but not what their amplitudes and durations should be. Of course, if one wanted to use the system as the basis for the intonation component in a text-to-speech synthesis system, one could construct simple amplitude and duration rules, i.e. that peaks in **H** accents occur 60ms into the vowel of the accented syllable[5]. Although the specification of a set of simple rules would allow a phonology-intermediate mapping system, it would not work for the intermediate-phonology mapping, and it is clearly against the whole approach of this thesis, which aims to provide an account of every intonational effect in English without resorting to gross over-generalisations. Such a phonology-intermediate mapping would be no more difficult to design for the RFC system than for any other system, but it is preferable not to commit the system to these sorts of rules. Therefore the phonology-intermediate mapping must remain incomplete until a thorough study is performed into the mechanism of element amplitude and duration.

The fact that the phonology-intermediate mapping can provide the type but not the size of the RFC elements should come as no surprise. In our previous discussions on redundancy it was shown that the workload of a grammar or mapping was proportional to the distance

[5] As in the Isard and Pearson model, section 2.3.2.

r f		**H**
f	(starts late in syllable)	$\mathbf{H}_d$
f	(starts early in syllable)	$\mathbf{L}_a$
r f	(late peak)	$\mathbf{H}_l$
r f	(high peak)	$\mathbf{H}_e$
r	(on accented syllable)	**L**
r	(not on accented syllable)	**B**
r	(non-phrase final)	$\mathbf{B}_i$
c		**C**
c	(rising)	$\mathbf{C}_r$

Table 3.3: *The phonology-intermediate tune grammar*

between the relevant levels. In the case of the *symbolic* RFC description, the elements are close to the phonology in that they are discrete and occur in strictly defined places. Thus the relationship between phonological element type and RFC element type is straightforward. However, in terms of *numerical* content, the RFC description is clearly quite some distance from the phonological level, in that the RFC description is numerical and the phonological description is not. The symbolic part of the RFC description has very low redundancy whereas the numerical part of the RFC description is still highly redundant. Thus a simple grammar can be constructed for the part of the RFC description which is "close to the phonology", a more complicated grammar is needed for the part which is "close to the contour".

Tune Grammar

Table 3.3 shows the phonology-intermediate tune grammar. Some of the terms used, such as "late peak", are obviously somewhat subjective, and will be discussed further below.

This grammar could be used to analyse sequences of RFC elements and deduce their phonological nature. The only essential information is the timing of the elements with respect to the vowel of the stressed syllable.

A real ambiguity appears to exist when a **L** with no antecedent fall occurs on the last (and sometimes next to last) syllable of the phrase. Here the rise element is indistinguishable from a normal boundary rise (which may, for example, have been caused by the need for a continuation rise). The only difference between the two is that a **L** accent occurs in relation to a stressed syllable, and is perceived as the nucleus, whereas a continuation **B** is not associated with a stessed syllable - another syllable is clearly perceived to be the nucleus. This apparent

ambiguity is easily explained if we think of a series of utterances, starting with one whose **L** accent nucleus is many syllables from the end of the phrase. The end of the phrase will end in a **B** and the connection element between the two will be rising. An utterance with the nucleus closer to the end of the phrase will simply have a shorter connection element. In the case of the nucleus being the last or the next to last syllable, the rising connection element will be so short as to have disappeared, leaving the **B** as the only indication of the **L** accent. This case is particularly interesting in that it is the only case that needs information other that the RFC description to show that an accent exists. In the other cases where non-RFC information is needed, this has been to subclassify an accent, not to show whether the accent is present or not.

Some Numbers

The exact timing of pitch accents is not addressed here, but an indication of how timing varies can be given. An isolated fall accent can either give rise to a $\mathbf{L}_a$ or a $\mathbf{H}_d$ accent. In data set A, the contours were hand labelled with the new phonological description system. A "vowel-onset-to-fall" (VF) distance was defined as being the distance from the start of the vowel to the beginning of the fall. All the relevant sentences in data set A had the VF distance marked by hand. Distinguishing between $\mathbf{H}_d$ and $\mathbf{L}_a$ was easy: H accents typically had VF distances of about 60ms with the shortest being 30ms and the longest being 120ms. Thus the start of the fall in a downstepping accent always occurred at least 30ms after the onset of the vowel. The average VF distance for the $\mathbf{L}_a$ was -140ms. This distance ranged from -100ms to -165ms. The start of the fall occurred on average 140ms *before* the vowel onset. Thus there is a very clear distinction in timing between the two types, with no overlap in the timings between the two groups. One might even go so far as to say that the more general rule is that falls occurring after the vowel onset are $\mathbf{H}_d$, falls occurring before the vowel onset are $\mathbf{L}_a$

Distinguishing **H** from $\mathbf{H}_l$ was less easy. The data showed that the average VF distance in the late accents was about 100ms, compared with an average of 60ms for the **H** accents. There did seem to be some overlap in the timings of the two types of accents, but this could be due to the syllables in question having different vowels, and hence different intrinsic durations. Gartenberg and Panzlaff-Reuter (1991) studied late peaks in German and found their timing heavily dependent on the syllable structure and type of vowel.

The amplitudes of accents varied more considerably, with rise elements ranging from 10Hz

to 96Hz and fall elements ranging from 11Hz to 140Hz. Although those accents which were marked with the elevated feature occurred in elements with large amplitudes, there was no distinct boundary between those marked elevated and those not marked. This is evidence that there is a more subtle distinction between elevated and non-elevated accents, in that some sort of high level context governs whether an accent is perceived as being elevated or not. It is also evidence for the "paralinguistic" view of prominence where an accent can have any amplitude.

Phrasing

Phrasing is expressed intonationally by the use of boundary elements. These varied less in amplitude (20 Hz to 70 Hz) than the rise elements which formed pitch accents. Phrase initial boundary elements tended to be smaller than phrase final elements. There did seem to be some relation between the size of the phrase initial boundary unit and the perceived strength of the phrase boundary preceding it. This issue was examined in Ladd (1988) who also found a relationship.

3.5 Discussion of the New Phonetic Model

The model presented above is an attempt at defining a formal phonetic model of intonation. The model has three levels of description, the F_0 contour, the RFC level, and the phonological level. The intermediate-F_0 grammar which links the RFC description to the F_0 contour is complete and provides a method of describing intonation on one level given a description on the other level. The phonology-intermediate grammar is incomplete owing to uncertainties about the linguistic level and the mechanisms of segmental influence. These unresolved parts of the phonology-intermediate grammar form well defined subject areas which will be left for further work. The phonology-intermediate grammar could be made more complete than it currently is by positing general rules as to how certain mechanisms work. This approach has been avoided as it was seen to be against the spirit of the formal approach taken elsewhere in the thesis.

The following sections discuss some interesting points about the new model.

3.5.1 Phonetic Reality?

An obvious question is "is the RFC description a phonetic description?". I think that for a description system to be "phonetic", it must have some articulatory basis. The following account attempts to provide a possible phonetic explanation for the RFC description.

Consider the situation where the glottis is vibrating at a certain constant frequency. This would correspond to a level connection element. A rise "command" is issued, which causes the glottis to start changing frequency. A short amount of time is taken before the glottis has reached a state where it is quickly changing the rate of vibration. As it approaches its intended new F_0 value, it relaxes and the rate of change drops again. This explanation would account for the shape of the rise and fall elements. This argument for the phonetic basis of the rise shape is further strengthed by the fact that the rise element is used in two quite different phonological situations; boundary elements and **H** accents. When the glottis is not involved in a rise or fall process, its frequency of vibration changes slowly, if at all, and this can be modelled be the use of straight lines.

It seems unlikely that the monomial function is anything other than an approximation to the "real" function. We demonstrated that at least two mathematical functions were capable of producing these shapes, and there may be still more. I can find no explanation for why the monomial function should fit the curve as well as it does. It is certainly not the case that the shape of the rise and fall is arbitrary, and that any function could produce the correct shape. Many mathematical functions were tried, and only the monomial and sinusoid curves seemed suitable.

Although slight variations in γ may have allowed better fits, the fact that a value of 2.0 was used throughout adds weight to the evidence that there is something special about this function. "Why 2.0?" is of course an even more difficult question to answer than "why the monomial", the only evidence is that it works well.

Saying "it works well" is often an excuse for using an unprincipled "hack". However I think that the monomial shape does model contours well, and the evidence can be seen by looking at how well RFC synthesized contours compare to natural contours (see appendix B). Finding a more theoretically justifiable shape is an appropriate topic for further work, but as I believe the monomial shape is one of the best aspects of the model, I don't see much point in looking for a new shape until the other remaining problems have been resolved.

Why do the Elements have Strict Boundaries?

A curious aspect of the RFC system is that the elements have strict boundaries. The model implies that the speaker instantaneously changes from producing one element to producing an other, with no transition period. This could be seen as being implausible, as other aspects of speech, such as segments, do not have strict boundaries and do have transition periods. I can offer no real explanation as to why this should be, other than proposing that the curved starts and ends of the fall and rise elements are in fact transitions and that the middle part of the monomial shape is the element proper.

3.5.2 Features

The use of features in phonology is widespread (Chomsky and Halle, 1968), (Jakobson et al., 1952) (Clark and Yallop (1990) give a review). In traditional segmental phonology, there is usually some articulatory basis to the feature, e.g. voiced/unvoiced. The voicing feature is binary in that a segment is either voiced or unvoiced. Other features such as "high" (when referring to the place of articulation of a vowel) are often not so obviously binary and represent more of a continuum. What then is the justification of the features presented here, and can they be truly considered binary?

H Accents

The features used here were first proposed by Ladd (1983b). These features were used in the new phonological description to avoid Pierrehumbert's cumbersome tone description system. Initially, Ladd's features were used in the new phonological system as they provided a simple method for describing intonation accents: as work progressed, the phonological system diverged from Ladd's system considerably until the only similarity was in the used of the same features to classify **H** accents.

Ladd (1983b) does not provide any articulatory (phonetic) justification for his feature system - it is simply a way of differentiating types of accent. Ladd's system was adopted permanently because it was seen that the RFC system could be used to provide a solid articulatory basis for these features, as the following explanation shows.

The elevated feature is an expression of F_0 amplitude. An accent marked with this feature has a rise and fall element of greater amplitude than usual. Hence the elevated feature is an

expression of the amplitude of an accent's rise and fall elements.

Accents which are late have their peak in a later position relative to the vowel onset than usual. This also normally produces a longer duration rise element than usual. Thus the late feature concerns the position of the elements and the relative durations of the elements.

Downstepped accents have noticeably larger falls than rises, and so it is the difference in amplitude between the rise and the fall which distinguishes the downstepped accent from the non-downstepped accent.

Thus "elevated" is controlled by the amplitudes of the elements, "late" is controlled by the timing of the elements, and "downstep" is controlled by the relative amplitudes of the elements. If we examine the default **H** accent, which is made up of a rise followed by a fall, it becomes clear that there are only so many ways of adjusting the rise and fall elements so as to produce new shapes: the rise and fall elements can only be scaled in two ways, either vertically (amplitude) or horizontally (duration). The only other way in which different accent shapes can be produced is by varying the positions of the elements with respect to the syllable. So there are three dimensions of control for producing different accent shapes: amplitudinal, durational and positional. If two elements make up the accent, as in the case of the **H** accent, the elements can be varied independently to produce a wider range of accent shapes.

Both the durational and positional controls are tightly constrained by the duration of the syllable. If an accent is produced too late in a syllable, it risks being perceived as being associated to the following syllable, likewise an early accent risks being perceived as a late accent associated with the preceding syllable. Within the duration of the syllable, there can only be so many perceptually distinct positions that the accent can have. It would be unlikely that two accents with their peaks occurring within 10ms of each other would be perceived as being different. For two accents to be perceived as separate types of accent, their positions must differ considerably. This therefore limits the number of possible accent positions that are perceptually distinct from one another. In Kohler's model of German intonation, there are three distinct positions for **H** accents; early, medial and late (Kohler, 1991c), (Kohler, 1991b). In English intonation, there are only only two positions, **H** and $\mathbf{H}_l$, which would correspond to the German medial and late. (The early type may not exist in English owing to a possible confusion with $\mathbf{L}_a$. See the end of this section for more on this.)

The duration of the elements could also be systematically varied so as to produce different

accent shapes, but again this is limited by the duration of the syllable. $\mathbf{H}_l$ accents do seem to have longer rise elements than **H** accents, but this variation of duration can only produce slight differences in accent shape.

The amplitudes of the rise and fall elements of an **H** accent are not constrained by the duration of the syllable. There will be some minimum pitch excursion which sets the lower limit on how small an accent can be and still be audible; the upper limit is set by the maximum F_0 of the speaker's pitch range. The amplitude variability is much greater than the variability in duration or position. The minimum duration of a fall element in data set A was 100ms and the maximum duration was 380ms, about 4 times bigger than the smallest duration. By comparison, the largest amplitude was 140Hz, about 13 times bigger than the smallest, 11Hz. The elevated feature is used to distinguish extra high pitch accents, but as stated before it was difficult to define any value which would serve as a boundary between elevated and non-elevated accents.

Variation in the *relative* amplitudes of the rise and fall elements is what gives rise to downstepped accents. In principle, upstepped accents may exist where the rise element has considerably larger amplitude than the fall element. Although some accents were found which did have larger rise amplitudes than fall amplitudes, none had differences as large as the $\mathbf{H}_d$ accents. In principle, there is the possibility that another type of pitch accent exists, but no occurrences of upstepped accents were found.

If the relative amplitudes of the elements can give rise to different types of accent, why do differences in the relative durations of the elements not operate in the same way? To some extent, the $\mathbf{H}_l$ accent does express differences in the relative durations of the elements, but the major difference between $\mathbf{H}_l$ and **H** is in peak position. I think relative durational differences are not so significant due to the constraints placed on the ability to vary element duration. It is quite easy to produce an accent with a fall five times greater than the rise, but difficult to produce an accent with elements which vary this much in duration. To ensure that the accents are perceived properly, the differences between the elements must be quite large and as it is easier to produce rise and fall elements with bigger variation in amplitude than duration, this explains why relative amplitude plays a major part in accent differentiation, while relative duration only plays a small part.

Other Phonological Elements

The features used to subclassify **L** and **C** elements are original to this thesis. The use of a fall element as part of an **L** accent seems to add extra emphasis to the accented syllable, presumably because it makes the difference in F_0 between the syllable and its environment more exagerated. **L** accents which do not have antecedent falls occur at a higher position in the speaker's pitch range and are similar to British school high-rise accents.

The connection elements are straight, and by default are level or slowly falling. Rising connections elements are used to contrast this behaviour and are used in situations such as $\mathbf{L}_a$ $\mathbf{C}_r$ and $\mathbf{H}_d$ $\mathbf{C}_r$ **B**.

We have stated above that connection elements have durations and amplitudes in the same way that rise and fall elements do. This Cartesian description has been used throughout, but an alternative is to describe connection elements using polar coordinates, whereby each element would be described by a duration and a *gradient.*

The polar representation is often easier to work with owing to the fact that the duration of the connection elements is not really a property of the elements themselves, but rather is governed by the positions of the pitch accents. Consider the case of two rising connection elements, one with twice the duration of the other. If they are given the same amplitude, the shorter one will be rising twice as fast as the other, which may lead to a difference in perception between the two connection elements. However, if they have the same gradient, they are more likely to be perceived as being similar. Thus connection elements with the same gradient are more similar than connection elements with the same amplitude.

If we accept that it is the gradient which is the controlling property of connection elements, it makes sense that a feature should be used to distinguish elements of differing gradient. It is also important to note that this is really the only feature that could be used for connection elements: the duration is governed by the accent position and is not a property of the element itself.

Are These Features Binary?

When hand labelling the data sets, it was quite easy to decide which RFC element to use, and there was little uncertainty as to where the boundaries between the elements were. It was also easy to decide if an accent was a **H** or a **L**. The features were more difficult to mark.

The late feature was often easy to mark as this type of pitch accent sounded different from a normal **H** accent and its peak was noticeably later. Many accents were obviously downstepped, particularly those **H** accents which had no rise element. Many **H** accents were definitely not downstepped as their rise and fall elements were of approximately equal size. There was a grey area in between (approximately 25%) of **H** accents for which it was difficult to decide.

This problem may arise from the fact that the downstep feature is not a binary feature but is more continuous and really just an indication of the relative amplitudes of the rise and fall elements (as explained in 3.5.2). Speakers wishing to produce downstepping contours would choose patterns at the more extreme end of the downstep scale to ensure that the effect was perceived properly.

There does not seem to be any clear evidence in the data presented here for the existence of an elevated feature. The amplitudes of **H** accents varied considerably, but the **H** accents did not seem to fall into two distinct groups. However, studies investigating extra high pitch accents (Ladd, 1992b) have shown consistent patterns in the heights of accents and so it might be premature to abandon the notion of an elevated feature.

Superficially, the antecedent feature does seem to be binary in that the fall element either exists or does not. However, it can be argued that this feature is continuous and represents the amplitude of the antecedent fall. A **L** with no fall is really just a **L** with a zero size fall.

The connection feature, "rising", is used to differentiate connection elements in terms of their gradient. Again, this feature could be argued to be continuous. However, in the post nuclear tail positions, where the difference between rising and non-rising connection elements is most important, the gradients of the elements did fall into two distinct groups. Elements were either clearly rising or else level or falling. There was little confusion about what feature to mark these elements with.

There has been much argument in the literature concerning the description of accent prominence (see section 2.2.3). Ladd and others think that a feature such as "elevated" exists and that accents can be classified as either "+ elevated" or "- elevated". Pierrehumbert and others argue that such an analysis is mistaken and that accent prominence is continuous, only bounded by the maximum F_0 of the speaker's pitch range. What the previous discussion should have made clear is that this problem is not unique to accent prominence; downstep and other features behave similarly.

The approach of the thesis has been to develop a formal system, and where it has proved impossible at present to complete a part of the system, this part has been left open. This means that the descriptions in the system have more redundancy than necessary, but this is preferable to having descriptions systems which will be easily be proved to be wrong. Thus the downstep and elevated features can not at this stage be said to be binary. It is simple to invent a continuous phonological measure to describe them: accent prominence can be described in terms of Hertz or on a logarithmic scale (see section 3.5.3). The reservation about such an approach is that it almost certainly means that more redundancy exists in the phonological description than is needed.

The confusion that exists in the literature probably indicates that the elevated feature is not continuous, but that its mechanism is complicated, certainly more complicated than a binary mechanism. Thus its seems wise to let the elevated feature and other features be continuous for the time being, hoping that their mechanisms will be discovered at a later date, and that these mechanisms will eventually be included in the theory.

Are H and L Features?

Is it possible that there is a general category "accent" and the feature distinction high/low should be used to differentiate **H** from **L**?

Such an idea seems attractive in that this would make our system more homogeneous in its uses of features. One could even go so far as to say the fundamental category is "element", with features "accent", "connection" and "boundary".

The main reason for arguing that **H** and **L** are not features is we would then have to account for why these two types of accents have different sets of features. **H** has three features, while **L** has one. We can hypothesize as to what the effect of the three **H** features would be on the **L** accent if we consider the discussion on the phonetic reality of the features outlined above.

"Elevated" was simply an indication of the amplitude of the rise and fall components. This feature could be used to control the F_0 value of the valley of the **L** accent.

"Late" was an indication of the position and relative positions of the elements. This feature could be used to indicate the position of the F_0 valley.

"Downstep" gave an indication of the relationship between the amplitudes of the elements. In a **L** accent, the relative size of the fall as compared to the following connection element

or boundary element could be systematically adjusted to produce different perceptions in the listener. This would therefore represent the antecedent feature.

There is no evidence for the use of these features in the data, but the possibility remains that **H** features can be used to model **L** accents. It may also be the case that speakers can produce these variations in the basic **L** accent, but do not because it is not conventional in English (i.e. it does not appear in the intonational "lexicon").

If **H** and **L** were features, our model would be somewhat similar in structure to Kohler's (1991a) feature based model of German intonation. Kohler uses "peak" and "valley" instead of **H** and **L** giving rise to the feature "± valley".

Whether **H** and **L** are features is an interesting topic for debate, but does not greatly affect the workings of the model in its current form. We will use **H** and **L** as though they are not features, if only for the reason that (I think) they are easier to work with as mnemonic descriptions units.

Is $\mathbf{L}_a$ really an early H?

The Kiel model of German intonation has early, medial and late peak accents (Gartenberg and Panzlaff-Reuter, 1991), (Kohler, 1991a). The system for English has the equivalent medial (**H**) and late ($\mathbf{H}_l$) accents. Falls do occur in earlier positions, but these have been classed as $\mathbf{L}_a$. Is it possible that the $\mathbf{L}_a$ accent is really an early **H** accent?

From examination of some F_0 contours this may seem a plausible explanation as often $\mathbf{L}_a$ looks very similar to $\mathbf{H}_d$. However, I believe that $\mathbf{L}_a$ is not an early **H** accent as the behaviour of the two types of accent differs in a number of ways. Firstly, increasing the prominence of an **L** accent has the effect of lowering the F_0 value of the accented syllable, rather than increasing it as in **H** accents (Liberman and Pierrehumbert, 1984). Secondly, the F_0 contour always rises after **L** accents either as a connection or boundary element. The F_0 contour sometimes rises after **H** accents, but does not have to. $\mathbf{L}_a$ is therefore not an early **H** accent.

3.5.3 Units of Scale

One may think that the use of Hertz in a phonological description, and even an intermediate description, is unattractive as this is such a "low-level" measure.

It is preferable for the RFC and phonological descriptions to have as low a redundancy

as possible. Although no experiments have been carried out to prove it, I would guess that a fall element of 161ms would not be perceived differently from a fall element of 162ms; thus the *precision* of the millisecond description is too high. The principled way to reduce the redundancy would be to carry out perceptual tests to discover the "just noticeable differences" in element amplitude and duration, and use these as the numerical units. This would ensure that the redundancy in the RFC description was minimized.

Such an approach may be theoretically attractive, but may not be practical. Using Hertz and milliseconds is more practical for the time being, simply because these are units that are familiar to us.

Linear Hertz frequency scales have been abandoned by some in favour of logarithmic (semitone) scales, the argument being that the perception of frequency is logarithmically based (Silverman, 1987), (t'Hart and Cohen, 1973). Perceptual scales of frequency have been proposed such as the Mel scale and Bark scale (Lindsay and Norman, 1972), (Buser and Imbert, 1987), and more recently scaling systems have been proposed specifically with intonation in mind. Hermes and van Gestel (1991) conducted experiments and found that human auditory perception of intonation was best represented by a system that was intermediate between linear and logarithmic scales. Regardless of whether the perception of frequency is logarithmic, linear or otherwise; transforming from a linear to another numerical scale is not a redundancy reducing exercise, and therefore does little to solve the problem of how best to describe scaling in intonation.

One might be tempted to use Pierrehumbert's notation whereby a pitch accent's prominence is described on a scale where 0.0 represents the baseline, and 1.0 represents an accent's "normal" height. Such a description system may seem attractive, but this is no less redundant than a Hertz frequency scale. For Pierrehumbert's scale to be less redundant, there would have to be a limit of the *precision* of the scale, i.e. a fixed number of decimal places. Such a limitation would be arbitrary. Thus unless we adopt a discrete system, it doesn't really matter if the scale uses Hertz, logarithmic Hertz, or some other scale; they are all continuous and have the same redundancy.

Using Hertz to represent prominence is unusual, but as the above explanation should have pointed out, this is really no different from using any other continuous scale. Ultimately, perceptual experiments probably will have to be carried out, and a proper prominence scale

devised.

3.5.4 Levels or Configurations?

No original work on intonation would be complete without comment on the "levels vs configurations" debate. This debate hinges around whether dynamic (rise and fall) or static (tone) processes are the fundamental substance of intonation. Ladd (1983b) points out that much of this debate has been based on argument over theories which are different in more ways than there choice of levels or configurations. I have previously argued that the accepted status that the Pierrehumbert phonology currently enjoys stems from the resolution of many problems such as declination/downstep rather than because levels are the substance of intonation.

At first it would seem that the model presented here is a configuration theory, as the RFC elements are described as being shapes (configurations). On the phonological level, **H** is described as being commonly realised as a peak, which again is a configurational term. However, it is possible to re-interpret this model as being a target based model.

The RFC system as a Target Model

I have previously criticised the cumbersome complexity of Pierrehumbert's interpolation rules, and have tried to show that the RFC system provides a simpler, more elegant way of describing intonation. A peculiar feature of the RFC system is that it has very low *symbolic* redundancy but high *numerical* redundancy (see section 3.4.7). Considering the flexibility of the durations and amplitudes within the RFC system, it could be argued that the system is really a *target* system, which uses two kinds of interpolation rules, straight line and monomial.

The main advantage of thinking that the RFC system is not a target system is that this avoids the need for any lookahead. Thus a long connection element is not an interpolation between the present point and some point in the future, but a F_0 pattern that has been uttered with a specific gradient. When started, it is not necessary for the speaker to know when the next element will begin. Utterance A.17 in appendix B has a rising connection element which is nearly 2 seconds in duration. In examples such as this the look-ahead for a target system would be considerable.

On the other hand, the apparent paradox about the connection element only being phonologically significant when it is rising could be explained by using the target interpretation of the model. In such a system, the rising connection element would simply be the interpo-

lation between a pitch accent and another element which was higher in the pitch range than normal. This argument is difficult to resolve one way or the other, and may indicate that the levels/configurations distinction is not as clear-cut as others have thought.

A useful future experiment would be to examine contours of utterances with rising post-nuclear intonation. By studying a wide variety of utterances with F_0 contours similar to that shown in utterance A.17, it might be possible to determine whether it is the position of the start of the final boundary element or the gradient of the connection element that exhibits the most consistent behaviour. From this study and others (Liberman and Pierrehumbert, 1984), it might eventually be possible to determine whether levels or configurations underly the intonation process. Such studies would carry more weight in resolving the debate than any amount of theoretical argument.

3.5.5 Points on Hand Labelling

It was stated in chapter 1 that for a system to be properly formal, it must be possible to map from one description to another without relying on human linguistic intuition. When hand labelling the F_0 contours, it was often easier to label after listening to the utterance. Labelling in this manner is relying on intuition and so this labelling approach is not a formal one. However, it was also possible to label the contours just by examining the F_0 traces. This was more difficult, but still possible. From examination of F_0 contours alone it was difficult to distinguish between **H**, $\mathbf{H}_t$ and $\mathbf{L}_a$, but if the position of the vowel was shown superimposed on the F_0 contour, this distinction could be made.

The "deaf" labelling was certainly slower than when listening to the speech, but did not seem to be any less accurate. Deaf labelling in itself does not prove that the labelling system is a formal one, but it does help to show that all the information can be extracted from the F_0 contour from analysis of F_0 contours and vowel timing alone, without the need to resort to language intuition.

Of course the only way to properly prove that the system is formal is to get a transcriber who has no language intuition to do the labelling. As it is practically impossible to get a human with no language intuition, the sensible alternative is to get a machine to act as the transcriber. This is the subject of the next chapter.

Chapter 4

Computer Implementation of the New Model

4.1 Objectives

This chapter explains the work carried out in trying to implement the phonology-F_0 grammar on a computer. By far the greatest effort was spent on implementing the F_0-intermediate mapping, which was the system that took an F_0 contour and produced an RFC description. A complete intermediate-F_0 mapping system was developed, and a phonology-intermediate mapping system was developed to the extent that the theory allowed.

The work presented in this chapter had two main objectives; to implement the model on computer, thereby proving its formality; and to objectively measure its results, thereby testing its performance.

Nearly all this chapter concentrates on the implementation of the F_0-intermediate mapping. By comparison, the intermediate-F_0 mapping was trivial. A intermediate-phonology mapping system was developed, but this was only able to perform part of the required mapping because of uncertainties in the theory of the phonology-intermediate grammar. A phonology-intermediate mapping system was not developed for reasons which are explained at the end of this chapter.

Section 3.3 explained the RFC description and intermediate-F_0 grammar in considerable detail. What was not defined explicitly was how to locate pitch accents from examination of F_0 contours. This was simple for a human labeller, but the formalisation of this task proved difficult. As with the phonetic models reviewed in chapter 2, it is the case with this model that the synthesis mapping defines the set of legal contours. It was shown in chapter 2 that F_0 contours which cannot be synthesized by a model are difficult to analyse as the model's legal

set of contours is different from the native-speaker set.

Considerable effort was spent on ensuring that the new model could accurately synthesize all the contours in the data sets, which were intended to be representative of the larger set. However, the model only aimed to synthesize the intonationally significant aspects of these contours, and made no attempt to model segmental perturbations. As the synthesis mapping had no ability to create these effects in the F_0 contour, it proved difficult to directly analyse F_0 contours influenced by significant segmental perturbations. To overcome this problem, a *contour preparation* module was developed which removed much of the segmental influence from an F_0 contour, thereby making it more like the contours which were produced by the synthesis mapping.

Section 4.2 describes the contour preparation process, the module which locates RFC elements in the contour, and the module which determines the precise boundaries of the RFC elements.

Section 4.3 explains in more detail the operation of the system's thresholds and parameters. A training method is described here whereby these thresholds can be statistically optimised. This training method uses an *objective assessment criterion* which measures the differences between transcriptions. This assessment method was used to give the training system a score for the system's performance when using a wide variety of thresholds. The set of thresholds which gave the lowest overall score were chosen as the optimal set.

This assessment method was also used to test the overall performance of the analysis system. Open and closed tests were performed, to assess how particular the thresholds were to the data they were trained on. Also, thresholds trained on set A were used to analyse the data in set B and vice versa, thus attempting to show how particular the thresholds were to an individual speaker. Results are given for all these tests.

The implementation of intermediate-F_0 synthesis mapping is explained in section 4.5. The ability of this mapping to accurately produce F_0 contours is discussed and a simple objective method of comparing F_0 contours is explained. Results are given showing how well this mapping models the contours of data sets A and B.

Section 4.6 discusses the computer implementation of the intermediate-phonology mapping. This section is restricted to the analysis of tune due to the incompleteness of the theory regarding other aspects of phonology. It is shown that the phonological description of an utterance can

reliably be extracted from the RFC description, so long as information on vowel timing is available to distinguish certain pitch accents.

4.2 Automatic RFC Analysis System

4.2.1 Overview of F_0-Phonology Mapping System

The *automatic RFC analysis system* was designed to be the computer implementation of the F_0-intermediate mapping. This system had three main modules which functioned relatively independently of each other.

Contour Preparation Involved extracting and processing the F_0 contour so as to be as free as possible of pitch and segmental perturbations.

Broad Classification Transcribed the utterance into sections labelled *rise, fall* or *connection*. The boundaries between these sections were only losely defined at this stage.

Optimal Matching Determined the precise boundaries between sections.

4.2.2 Contour Preparation

Considerable effort was spent ensuring that the intermediate-F_0 mapping could accurately produce contours representing all the intonational effects of English. However, this mapping was concerned with creating the *intonational* effects of the language and did not attempt to model segmental influence on the contour. Thus there was still a large discrepancy between the set of contours that the model produced and the native speaker set. As stated before, differences in these sets can lead to analysis difficulties. One solution might have been to incorporate a segmental model in the intermediate-F_0 mapping. This would have been difficult and would have required a thorough investigation into the precise mechanism of how segments influence F_0. An alternative solution was to modify the contours of the native-speaker set so as to make them more similar to the contours that the intermediate-F_0 mapping produces. The amount of modification had to be carefully controlled so as not to distort the intonation of the contour. Thus only slight modifications were allowed, and these were restricted to changing

segmental influence and not intonational content. This approach of modifying F_0 contours was implemented as the *contour preparation* module.

It soon became obvious that it would be impossible to fully eliminate segmental influence from the F_0 contour, since a full phonetic segmentation of the utterance would be required. No automatic method has yet been devised which can do this with total accuracy, and even if a method could be found, it would add considerably to the complexity of the system. Therefore, an approach was taken that tried to eliminate segmental effects from the contour without knowledge of what types of segment were present.

Segmental content influences the contour in many ways. Some of these influences can easily be compensated for, while others are more troublesome. The next sections examine how segments affect F_0 contours and explains how these effects were dealt with in the analysis system.

Pitch Perturbations

Pitch perturbations are not segmental effects but arise as a consequence of the glottal production mechanism. The pulses coming from the glottis do not arrive at an exactly equal intervals but vary slightly causing the F_0 contour to appear uneven as seen in figure 4.1.

Pitch perturbations were removed by performing 7-point median smoothing. The contour in figure 4.1 (b) shows a contour which has been processed by 7-point median smoothing, where each value is replaced by the median of itself and six neighbouring values, three on each side. Values on the edges of unvoiced regions are smoothed by extrapolating the curve from the next nearest fully smoothed values.

Obstruents

Obstruents have a significant effect on F_0 contours. (Silverman, 1987), (Kohler, 1991a). The specifics of which particular obstruent produces which effect is not of great importance here, but from study of the data in set A, it is apparent that they cause the F_0 contour to make large short-term deviations in the path of the contour. These deviations can be seen in figure 4.1.

These sharp deviations can sometimes be as large as actual pitch accents, so their elimination is important if the deviations are not to be incorrectly classified as accents. Although these segmental deviations may sometimes have similar amplitude to pitch accents, they are easily

distinguished by their much shorter duration, as they typically only last for 10 or 20 milliseconds. Median smoothing was used to eliminate these deviations .

The smoothing is a vital stage in the contour preparation process. It eliminates short-term deviations in the contour while preserving the long term features which represent the intonation itself. However, one must be careful not to smooth the contour too heavily. The order of the median smoother represents the length of contour "window", i.e. how many values are looked at to find the median. If the order is too high, the intonation features themselves will start to be eliminated; if the order is too low, the segmental features will still be observable. So a balance must be achieved. Unfortunately it was often the case that not all the obstruent influence could be eliminated before the intonation content started to be affected. Therefore smoothing can only remove most of the obstruent influence, not all. 15 point median smoothing was found to remove most of the segmental influence without greatly distorting the intonational content of the contour.

Unvoiced Consonants

The most obvious segmental influence on the F_0 contour is that voicing is absent altogether during unvoiced segments, resulting in breaks in the F_0 contour. Kohler (1991a), claims that although unvoiced obstruents affect the F_0 contour in the way described above, this influence is very short term, and the does not affect the overall pattern of the F_0 contour. He claims that the contour is *masked* in unvoiced regions. Kohler goes on to give explanations of how the listener perceives a contour containing an unvoiced region to be perceptually similar to one containing no unvoiced region. In Kohler's account, the listener simply interpolates through the unvoiced region.

By default, the intermediate-F_0 mapping produces fully voiced contours. Although it would be a simple matter to mask off a region of the contour, it would be difficult to know *where* to do the masking as this this would be dependent on knowledge of the segmental structure of the utterance. An alternative is to interpolate through the unvoiced regions of the contours in the native set so that all contours are fully voiced and continuous (except in pauses).

If Kohler's observations are correct, this interpolation should produce an F_0 contour that is free from the unvoicing problem, but still containing exactly the same intonation information as the original.

Unvoiced regions were eliminated by using straight line interpolation. It is possible that other types of interpolation could be used, but straight line interpolation is straightforward and reliable, whereas more sophisticated techniques, such as spline interpolation, sometimes produced unpredictable results. Figure 4.1 (c) is the result of the smoothing and interpolation. The straight line nature of the interpolation can leave sharp edges where the interpolated region joins the rest of the contour. To round off these edges, further 7 point median smoothing was performed. The resultant contour is figure 4.1 (d).

How the System Eliminates Segmental Influences in Practice

In practice, the pitch and segmental perturbations were removed together in a single 15 point median smoothing operation. After the unvoiced regions had been interpolated, 7 point median smoothing was used to remove singularities from the joins between the original contour and the interpolated region. Figure 4.1 (d) shows a F_0 contour that has undergone the contour preparation process.

Unlike data set A, data set B was not constrained so as to use mostly voiced phonemes. Therefore the segmental influence, particularly from obstruents, was much higher. Although it might have been possible to ignore the contour preparation for data similar to set A, the smoothing and interpolation process was essential for data such as set B.

4.2.3 RFC Labelling

It was straightforward to hand label the data in terms of the rise, fall and connection elements. Labelling in this way was not a strictly formal process as it made use of the human labeller's ability to locate pitch accents and ignore segmental perturbations. Much of the segmental influence had been removed from the contours which were to be analysed but the computer labelling system still had to deal with the problem of locating pitch accents.

Various analysis techniques were examined before the basis of the final system was settled upon. It is not necessary to give a history of the analysis system's development, but it may be of interest to explain some failures.

The first idea was to try a top-down phonological classification. This involved separating the **H** accents from the **L** accents first, and subdividing the groups later. As the **H** accents mostly represented peaks of some sort, a peak-picking algorithm was developed.

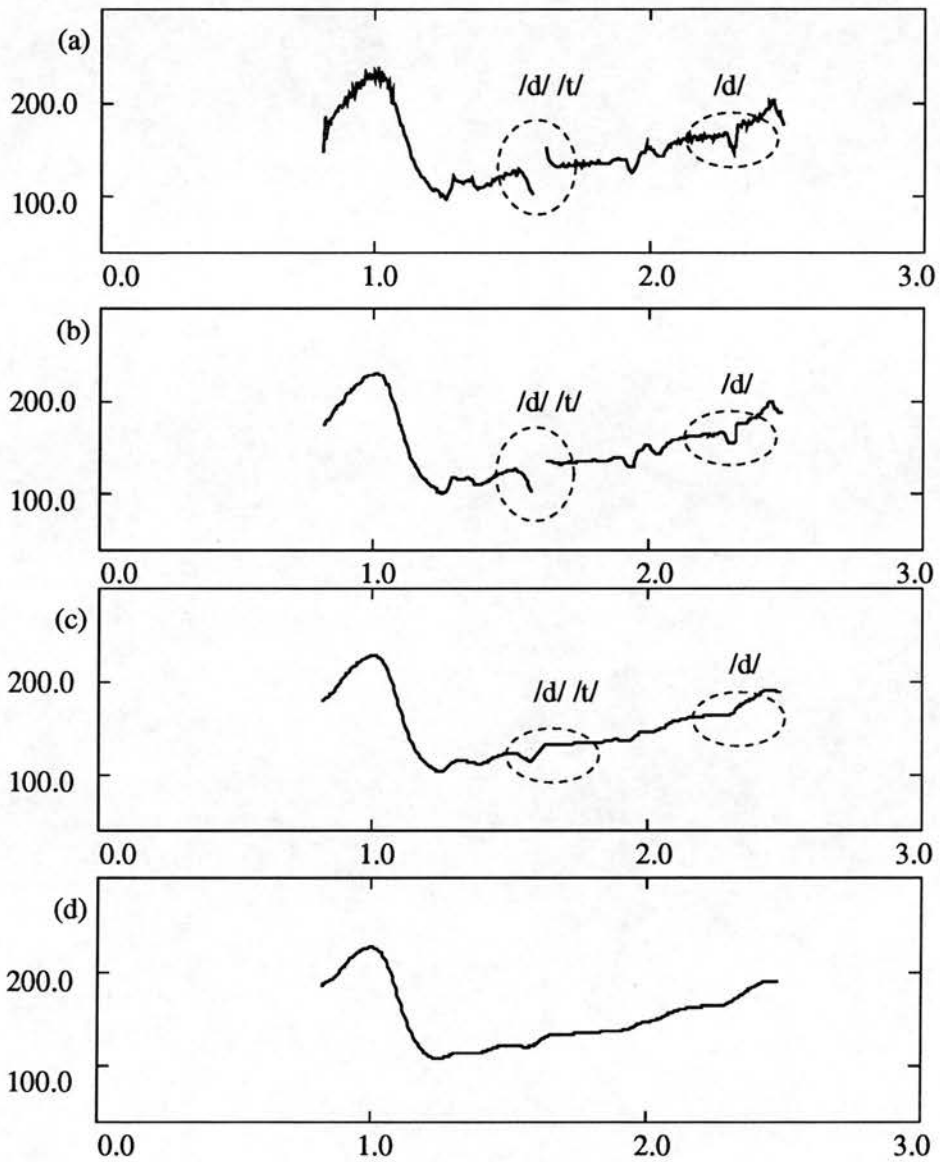


Figure 4.1: *The processing of a raw F_0 contour into a smooth fully voiced F_0 contour*

The design of such a peak-picking algorithm proved difficult due the very large variation in the types of peaks in the data (some of the downstepping accents had no peak at all). Little progress was being made in the design of a rule-based peak-picker, so it was thought that it might be worth trying a statistical approach.

All the peaks in the database were labelled. A neural network was trained using the peak/no peak decision as output, and the values of the F_0 contour surrounding the peak as input. This was not particularly successful either, with only about 80% of peaks being correctly identified - far too many false insertions of peaks were produced, mainly from residual segmental influence.

Although much more effort could have been spent on these two approaches, it was decided that a bottom-up approach might be more suitable. In this type of system, phonological classification would be left to last and the first step would be to analyse the F_0 contour in terms of the RFC elements. After some initially promising results, this approach was adopted and in the end proved quite successful.

The following sections described the system that was finally developed.

Frame Labelling

The F_0 contours produced by the laryngograph pitch-tracker gave a F_0 value every 5 milliseconds. As intonational information varies more slowly than this it is possible to use a much lower sampling rate. In order to keep the amount of information presented to the system as small as possible, the F_0 contour was re-sampled so that a value was available every 50 ms. Re-sampling at this much lower frequency is justified in that this module was solely concerned with detecting rises and falls; the original contour was reused later when more precise analysis of the contour was required.

A simple classification system was devised whereby each of the 50 ms F_0 values was compared to the previous F_0 value. If the F_0 between the frame and its predecessor increased by a certain amount, this frame was marked as a rise. If the F_0 decreased by a certain amount, the frame was labelled fall. The remaining frames were classed as connection. A rise threshold and a fall threshold were defined which controlled the classification process. If the F_0 difference between two frames was greater than the rise threshold, the frame was labelled rise; if the F_0 difference was lower than the fall threshold, the frame was labelled fall. These thresholds

were defined in terms of F_0 difference per second and in principle could be set to be any value. Initially these thresholds were set by hand, but the procedure explained in section 4.3 was later used to ensure that these thresholds were optimised.

All adjacent frames of the same class were grouped together to form a labelled *section*. From examination of the contours marked in this way, it was apparent that this method was quite successful and labelled most parts of the contours correctly.

If the rise threshold was too low, it was possible for rising connection sections to be mislabelled as rises. This type of error was referred to as an *insertion*. Conversely, if the threshold was too high, legitimate rises were mistakenly labelled as connection elements. This type of error was known as a *deletion* error. The optimal threshold would have the number of both these errors as low as possible. Insertion and deletion errors also occurred for fall sections. Theoretically another type of error, *substitution* could have been present, where a fall section was marked as a rise section. This type of error never occurred due to the simple fact that rise thresholds were always positive and fall threshold were always negative, thus excluding any direct confusion.

Assimilation

Often two fall sections that were obviously part of the same pitch accent would be separated by a small connection section. Examination of the contour showed this to be due to some unevenness in the F_0 contour, probably caused by residual segmental effects. Rise sections were often "interrupted" in this way by connection elements, and connection sections were also often interrupted by frames labelled as rise. This type of insertion error was largely eradicated by introducing the *assimilation module* into the system.

The hand labelled data contained no sections which were shorter than 100ms, therefore any automatically labelled section smaller than this was probably mislabelled. The assimilation module looked for cases of a small section interrupting two larger sections of the same type. If this section was smaller than a defined assimilation threshold, it was relabelled with the name of the neighbouring sections, and these three sections were grouped into one larger section. Figure 4.2 shows a contour undergoing the assimilation process.

The assimilation threshold is definable in the same way as the rise and fall thresholds. Using a very short assimilation threshold meant that no assimilation occurred, using a very

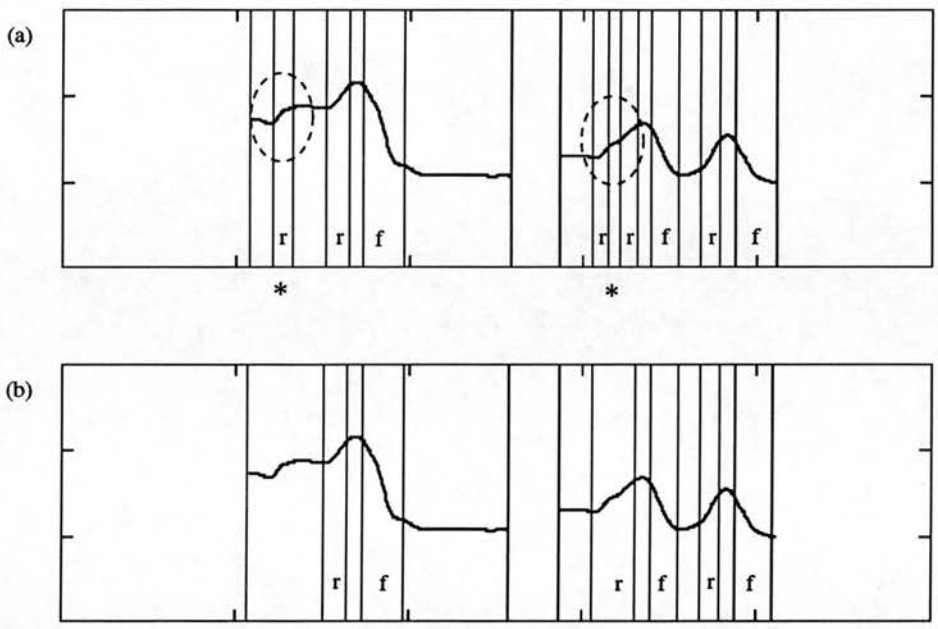


Figure 4.2: *Graph (a) shows the results of the classification module. Rises are marked "r", falls are marked "f" and connection elements are unmarked. The classification is correct except for two spurious sections which are indicated by the dotted circles. The rise is spurious because it is labelling a segmental perturbation which has no pitch accent associated with it. The spurious connection element is an insertion error because it is interrupting a rise section associated with the first pitch accent in the second phrase. Both these sections are small and are surrounded by sections of the same type. The assimilation module deletes the spurious sections resulting in the transcription shown in graph (b).*

long threshold meant that legitimate sections were deleted. As with the rise and fall thresholds, initially this was set by hand, but the training process described in section 4.3 was eventually used to ensure the best results.

The assimilation module greatly reduced the number of insertion errors.

Output of Classification Module

At this stage in the automatic analysis process a rough (to the nearest 50 ms) transcription had been performed. This broad classification was then used as input to the next module, which was designed to find the precise boundaries between the sections.

4.2.4 Optimal Matching of RFC Sections

The broad classification module worked best on the re-sampled 50ms contour. The optimal matching module required a finer specification of the F_0 contour, and so the original 5ms

sample-rate contour was used.

Finding the precise boundaries involved synthesizing candidate rise and fall shapes of different duration and amplitude, and comparing these candidate shapes to the original contour. The shape that fitted the contour best was kept, and its start and stop times were taken as the precise boundaries of the element.

The number of shapes that were tried varied in each case. Four thresholds were specified which defined "search areas". (Again, these thresholds were initially set by hand. The training process described in section 4.3 was eventually used.) A search area was created around the start and stop boundaries of each element. Typically the beginning of the start search area was chosen to be about 50 ms before the element's rough boundary, and the end of the area about a third of the way through the element. The stop region was defined in a similar way. Within each search area, there were typically about 20 frames of F_0 contour. Every possible candidate shape that started or stopped on one of these frames was synthesized and compared to the original contour. If there were 20 start and 20 stop positions, this would create a total of 400 candidate shapes. Each shape was synthesized and the euclidean distance between the candidate shape and the original F_0 contour was calculated. This distance was normalised with respect to time, as longer shapes would naturally have larger distances. The candidate with the lowest normalised euclidean distance was chosen and its positions defined the start and stop boundaries of that element.

This method was only used for the rise and fall sections as the connection section boundaries were defined by the boundaries of optimally matched rise and fall sections. The connection elements were therefore not directly matched to the F_0 contour, but so long as the theory regarding the connection elements is correct, and that they are indeed straight lines, the fact that they are not matched should not present a problem. If a rise section neighboured a fall section, as in a typical **H** accent, the optimal matching process could choose a different end boundary for the rise than the start boundary it chose for the fall. This happened quite often, but the discrepancy between the two values was often small. A common boundary was positioned half-way between the two originally chosen boundaries.

The amplitude of each section was determined from the duration: the F_0 value of the contour was measured at each boundary and the resultant difference in F_0 across each section was taken to be its amplitude. The output of the optimal matching module was a list of elements, each

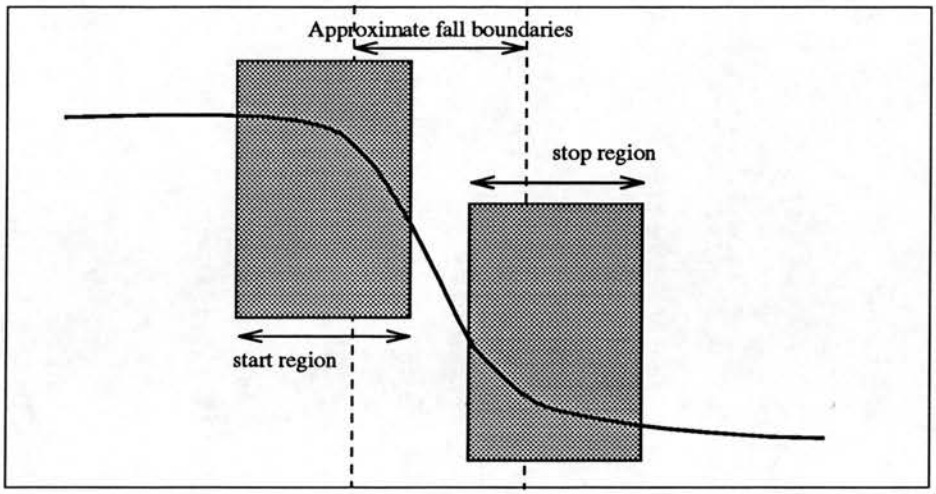


Figure 4.3: *A typical H_d accent. The dotted vertical lines show the boundaries that have been marked by the classification module. A region (covered by the shaded box) is defined around each of these lines. The durations of these regions vary, but typically they are about 100 ms long. As each frame is 5ms long, 20 possible positions are present in each region, giving 400 possible accent shapes.*

with a duration and an amplitude, similar to that shown in section 3.3.5.

Figures 4.3 and 4.4 show the optimal matching process.

4.3 Assessment and Training

During development, the RFC analysis system was improved and assessed primarily by impressionistic means; the analysis system was run over some of the data and the output examined visually by superimposing the resultant transcription on the F_0 contour. By studying the performance of the system in this way, the analysis system was improved until it evolved into the system described above. This type of assessment has limited worth as it is inherently subjective in nature. A more objective, quantitative assessment system needed to be devised.

The main reason for developing the objective assessment method was so that figures would be available that gave a measure of the system's performance. This would enable the techniques used in this system to be objectively compared against other techniques. Also it would help in the development of improvements in the system; by measuring the scores of an improved version with an existing version one could judge how much the improvements increase the system's performance. Finally, the assessment method was used in the training procedure outlined below. By calculating scores for the system's performance when using every set of

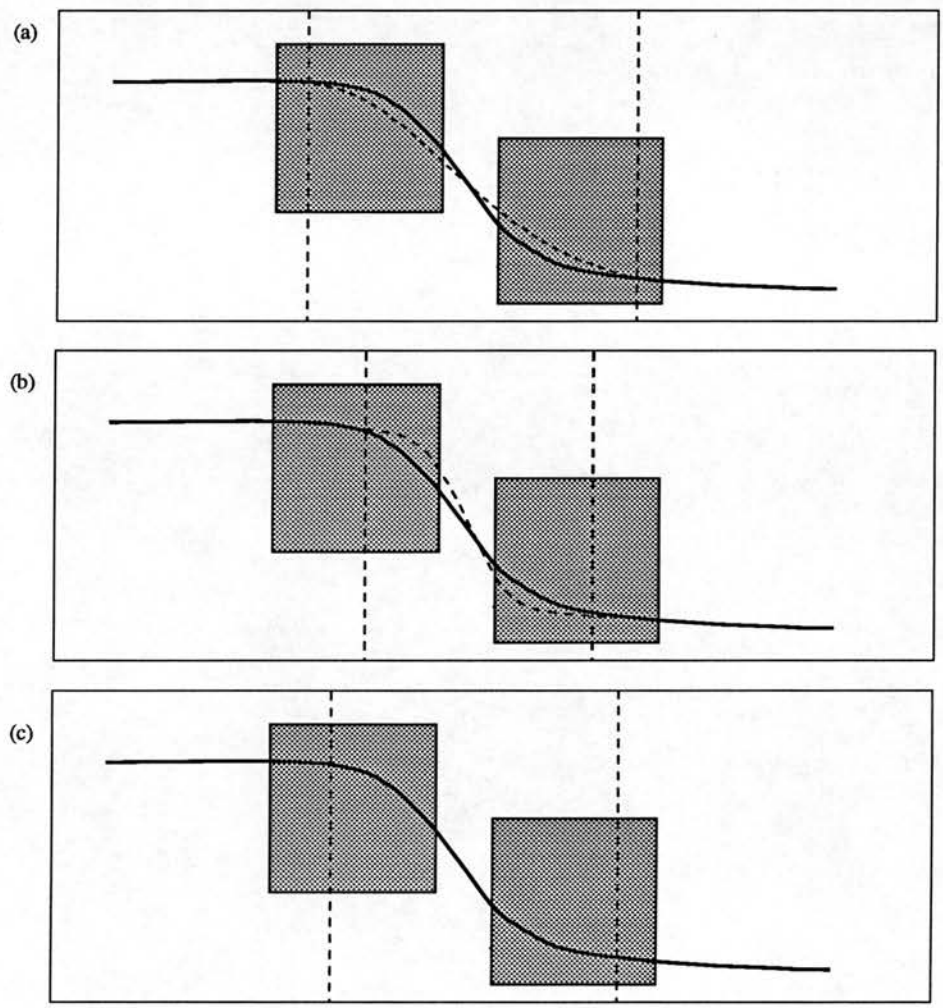


Figure 4.4: *These three figures show the matching process in operation. Graph (a) shows a fall element (the dotted curve) that is too long. Its start and end position are shown by the dotted vertical line in the shaded box. Graph (b) shows a curve which is too short, and graph (c) shows the curve with the optimal duration.*

possible thresholds, it was possible to find the set of thresholds which gave the best performance.

It is simple to design some metric which will give figures as to how well a particular system performs; what is more difficult is to design an assessment method that agrees with impressionistic views. The assessment method must be *weighted* properly so that important effects which are analysed incorrectly receive a harsh penalty, while relatively unimportant errors receive a more lenient penalty.

A very significant problem when assessing the performance of a labelling system such as this is we have no way of knowing what the correct segmentation is. When assessing F_0 contours produced by the synthesis system we can compare these to a naturally occurring contour which has been determined through measurement. There is no way to measure the RFC description of an utterance. The only solution is to hand label the contours using linguistic expertise, and compare them to the automatically produced transcriptions. If the two agree, we can say the formal, computer system can label with the same ability as the expert labeller. The problem with such an approach is that it relies on the human labeller being able to correctly transcribe the utterance. The RFC labelling criteria, as described in section 3.3.5, were intended to be as rigorously defined as possible so as to avoid arbitrary labelling decisions. The only area that required real expertise was in the labeller's ability to detect pitch accents and ignore segmental effects.

With the proviso that the human labelling process may be prone to some error, we can use hand transcriptions as a reference with which to test automatic transcriptions.

In the assessment method that was eventually developed, two types of error were measured. The first concerned the *identity* of a particular section and the second concerned the *precision* of the boundary positions.

The possible identity errors were:

Insertion occurs when the automatic system claims a section exists for which there is no corresponding hand labelled section.

Deletion occurs when the automatic system misses a section which exists in the hand transcription.

Substitution occurs when the automatic system gives a section a different class from the hand transcription.

The second type of error concerned the precision of the labelling: assuming there are no identity errors, how closely do the automatically chosen boundaries align with the hand chosen boundaries?

Every error incurred a quantitative penalty, which tried to reflect the impressionistic measure of how bad that type of error was.

A penalty of 0.1 was given for every 10ms of misaligned boundaries. Any misalignment below this would not incur a penalty. The most important type of identity error occurred when pitch accents were labelled wrongly. It was judged that a rise or fall element that underwent an identity error would receive a penalty of 3.0. This implies that an alignment error of 300ms is equivalent to the insertion or deletion of a section.

The system was best at labelling large elements; small elements were often ignored. Large intonation accents are often perceptually more important than smaller ones, so it would have been pleasing to incorporate this weighting into the assessment method. Unfortunately this was difficult to achieve, because no real measure of accent importance could be found, and it was thought that it would be better to keep to the simple criteria rather than construct a more elaborate assessment method that might be misleading.

By adding the penalties for an utterance, a combined score was calculated, which was normalised per unit time.

Having an error score per unit time is useful in being able to tell if one analysis system is better than another, but it does not tell *how* much better. For that to be known, some measure of the worst or maximum error is useful. A rough method was devised whereby the score for a "random" transcription would be used as a guide to the maximum possible error.

A random transcription could be simulated by comparing an utterance's transcription with the transcription for a different utterance. The score for this comparison should indicate what score would be produced by an analysis system that produced "rubbish" output. To find an average random transcription score, the procedure was carried out on all the utterances in data set A, which resulted in a maximum score of about 30. It is possible to produce scores worse than this; a transcription consisting of hundreds of 10ms sections would produce a score in excess of 1000. However, this is unrealistic as one would have to deliberately design the analysis system using knowledge of the assessment criteria in order to achieve such a bad score.

The scale of the scores are purely arbitrary, and the numbers "0.1" for misalignment and

"3.0" for identity errors were chosen so as to keep the scores in manageable single or double figures.

4.3.1 Training Method

With the use of the quantitative assessment method, it was possible to optimise the thresholds of the analysis system to ensure maximum performance. The thresholds fell into two groups; those which governed the initial broad classification of sections and those which governed the optimal matching of the precise boundaries.

Thresholds

rise-thresh defined the gradient that the contour had to exceed in order to be labelled **rise**.

fall-thresh defined the (negative) gradient that the contour had to fall below in order to be labelled **fall**.

rise-assim was the duration below which sections neighboured by rise sections were subjected to assimilation processing (see section 4.2.3).

fall-assim was the duration below which sections neighboured by fall sections were subjected to assimilation processing.

There were two sets of thresholds for the optimal matching module, one for rise sections and one for fall sections. These thresholds defined the search areas in which the candidate shapes were fitted.

start-begin Distance before the marked start of the section. Measured in seconds.

start-end Distance after the marked start of the section in percentage duration of the section.

stop-begin Distance before the marked end of the section in percentage duration of the section.

stop-end Distance after the marked end of the section. Measured in seconds.

There are thus four groups of thresholds: rise classification, fall classification, rise optimal matching and fall optimal matching. These four groups operate independently in that the rise thresholds don't affect the fall analysis, and the optimal matching thresholds don't affect the classification analysis. This independence allows the groups of thresholds to be trained separately.

Training Procedure

For the rise and fall threshold training, it is not the overall score, but rather the number of insertions and deletions that are important, as this module is only concerned with the detection of elements. The threshold training for the optimal matching procedure used the objective assessment scores.

The training procedure was a simple one. Each of the thresholds was varied systematically, scores for each set of thresholds were recorded, and the set of thresholds achieving the lowest score was kept.

Four training experiments were thus designed, one for each of the four groups of thresholds. The classification experiments involved systematically varying the gradient threshold against the assimilation threshold. The optimal matching experiments involved varying the four relevant thresholds against each other.

4.3.2 Rise and Fall Threshold Training

Two variables influenced the ability of the system to find rise elements: **rise-thresh** and **rise-assim**, of which **rise-thresh** was the more important. A matrix was constructed covering **rise-thresh** values from 20 Hz/s up to 800 Hz/s and **rise-assim** values from 0.025 seconds to 0.525 seconds. This matrix had 96 entries. A separate experiment was run for each entry in the matrix and the average number of rise insertion and deletion scores per utterance was recorded (there were no substitution errors found in any of the experiments). Figure 4.5 shows the average number of insertion and deletion errors as a function of **rise-thresh** for 32 of the utterances in data set A. As one might expect, the number of insertions decreases and the number of deletions increases as **rise-thresh** increases. If one accepts the principle that an insertion is as bad an error as a deletion (which is an assumption that was used when designing the assessment method described in 4.3), then the optimal point on the curve is where the

combined number of insertions and deletions is at a minimum. This value of **rise-thresh** occurred at about 120 Hz/s. The combined curve is also the shape one would expect, showing that extremes of threshold give bad results and the optimal point is a compromise. Figure 4.6 shows the numbers of combined insertions and deletions for a different value of **rise-assim**. The trends displayed in this graph are similar to 4.5.

The minimum point in the combined curve is the optimal value for **rise-thresh**. For each value of **rise-assim**, there is a separate combined curve. By finding the lowest values of all the curves, it was possible to find the optimal value for **rise-thresh** and **rise-assim**. From figure 4.7 it can be seen that apart from small values it makes little difference what the value of the assimilation threshold is. The lowest score was found when **rise-assim** was 0.125s and **rise-thresh** was 120Hz/s. Although it seems that even large values of **rise-assim** give good results, it is safer to use a middle range value, as the large values may join two different accents by mistake. This did not happen in the training data, but if the size of **rise-assim** was very large (> 1.0s) this would be common.

The fall threshold was trained in exactly the same way as the rise threshold. Figures 4.8 and 4.9 show how varying the fall threshold influenced the error rate. Figure 4.10 shows the significantly worse performance that can be expected if the assimilation threshold is set too low. This graph clearly shows the usefulness of the assimilation process. The lowest number of combined errors and deletions (18) occurred when **fall-assim** was 0.0125. Two values for **fall-thresh** gave this result: 75Hz/s and 120Hz/s.

4.3.3 Rise and Fall Optimal Matching Training

This training procedure was concerned with finding the best values for the search area in which the optimal fitting of the rise and fall shapes would take place. The method was as before, with a matrix of values being used. This time the matrix was four dimensional as there were four interdependent variables governing the definition of the search areas. Two search regions were defined, one for the start of the shape and the other for the end of the shape. The variables defining these areas are explained below.

start-begin The distance before the marked start boundary.

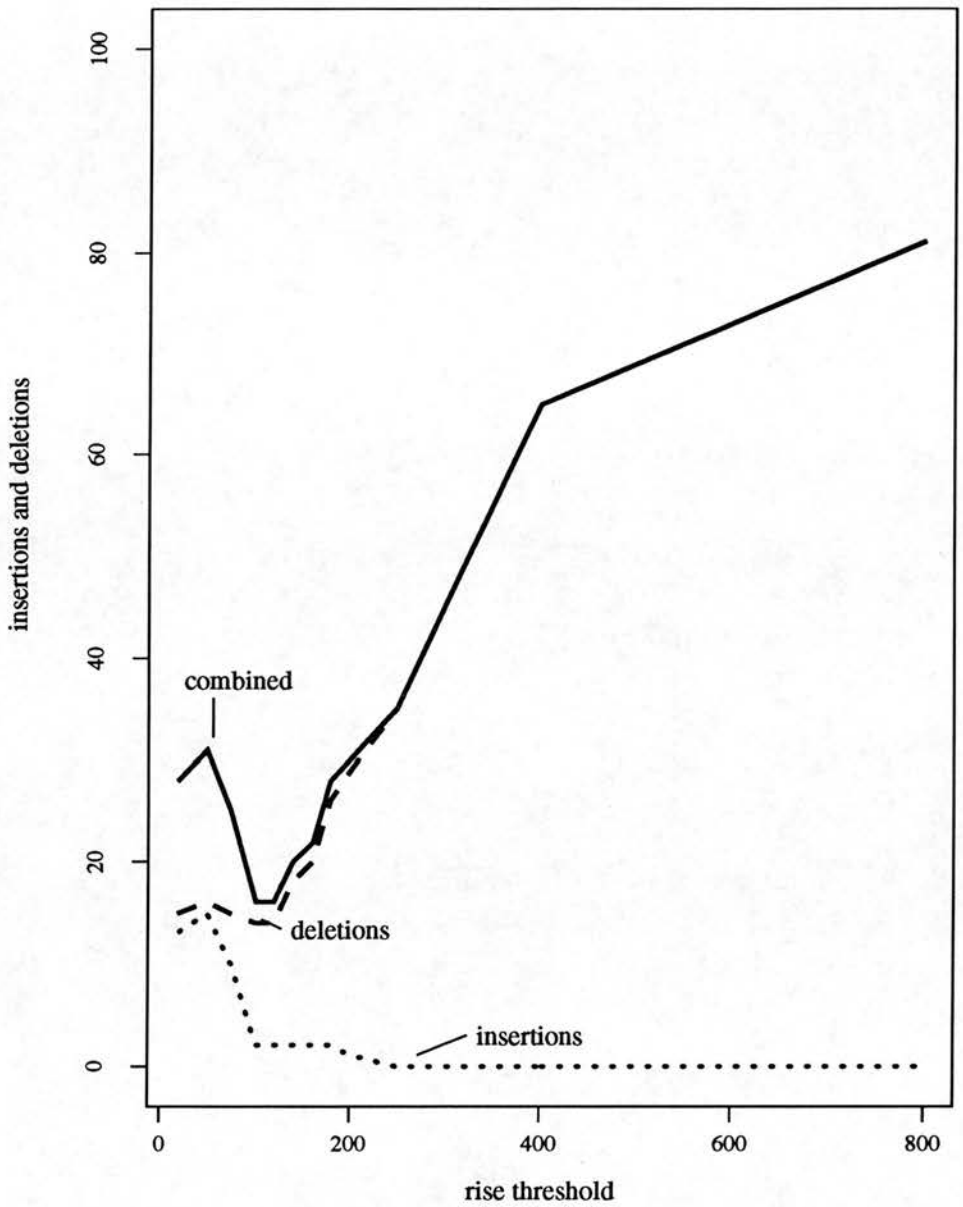


Figure 4.5: *Rise threshold performance. Rise insertion and deletion errors are plotted against values for the* ***rise-thresh****, with* ***rise-assim*** *at 0.175s. As the number of insertions decreases, the number of deletions increases. The minimum in the combined curve is the optimal* ***rise-thresh*** *for this value of* ***rise-assim****.*

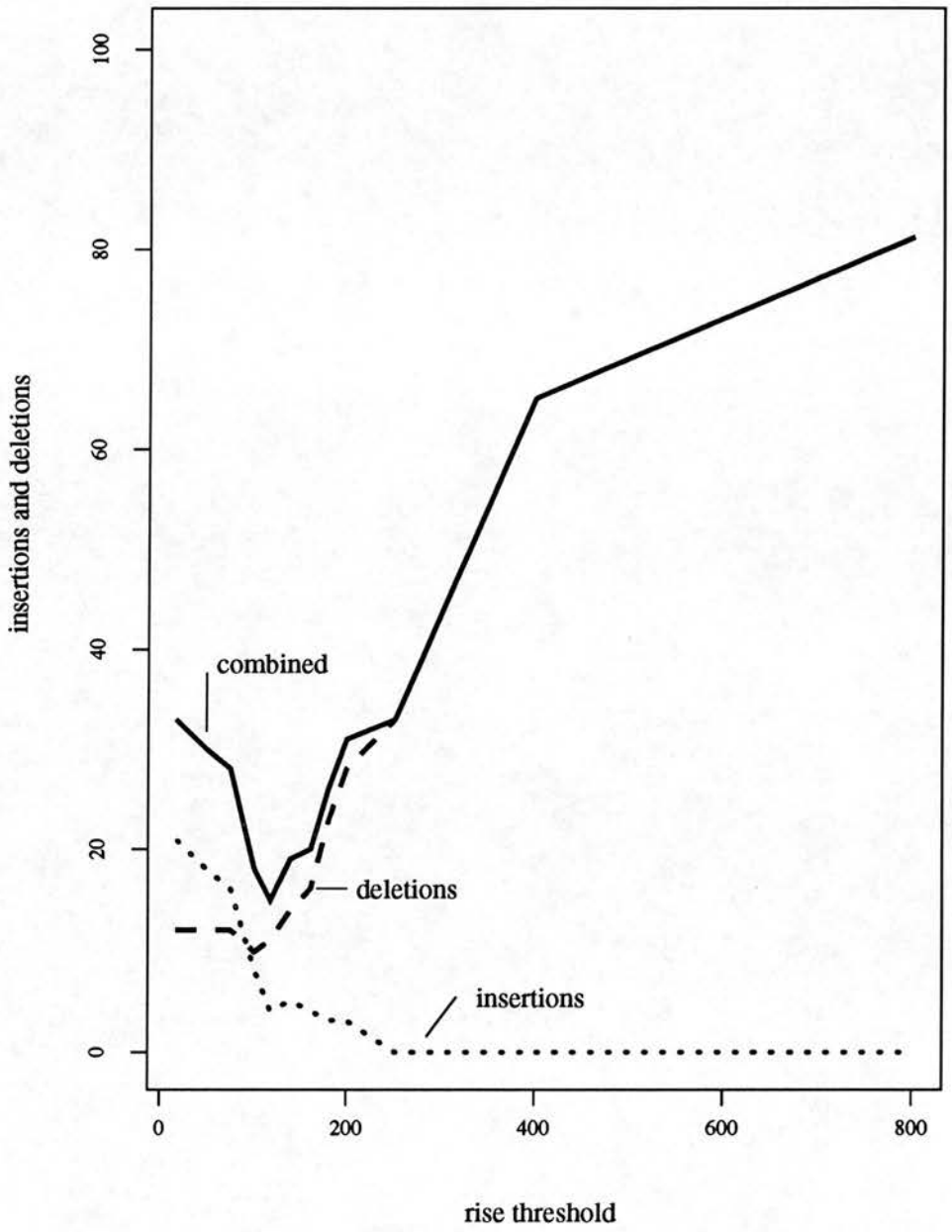


Figure 4.6: *Rise insertion and deletion errors are plotted against values for the* ***rise-thresh****, with* ***rise-assim*** *at 0.125s.*

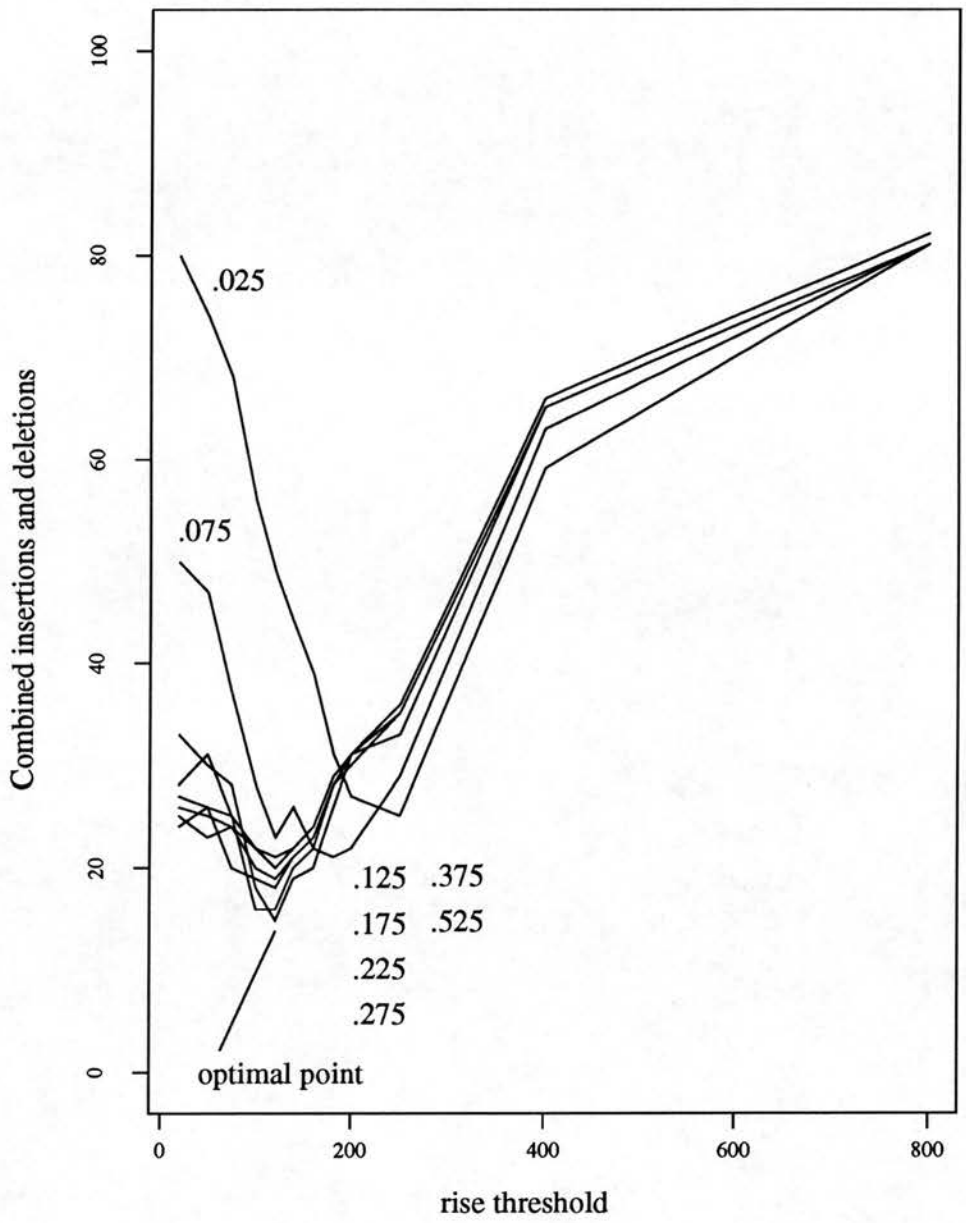


Figure 4.7: *Combined rise insertion and deletion errors plotted against* ***rise-thresh****. 8 curves representing different values of* ***rise-assim*** *are plotted here. The smallest values of* ***rise-assim*** *(0.025 and 0.075) have the worst scores, with all the values between 0.0125 and 0.525 having approximately the same performance. The lowest score on the graph is when* ***rise-assim*** *is 0.125 and* ***rise-thresh*** *is 120Hz/s.*

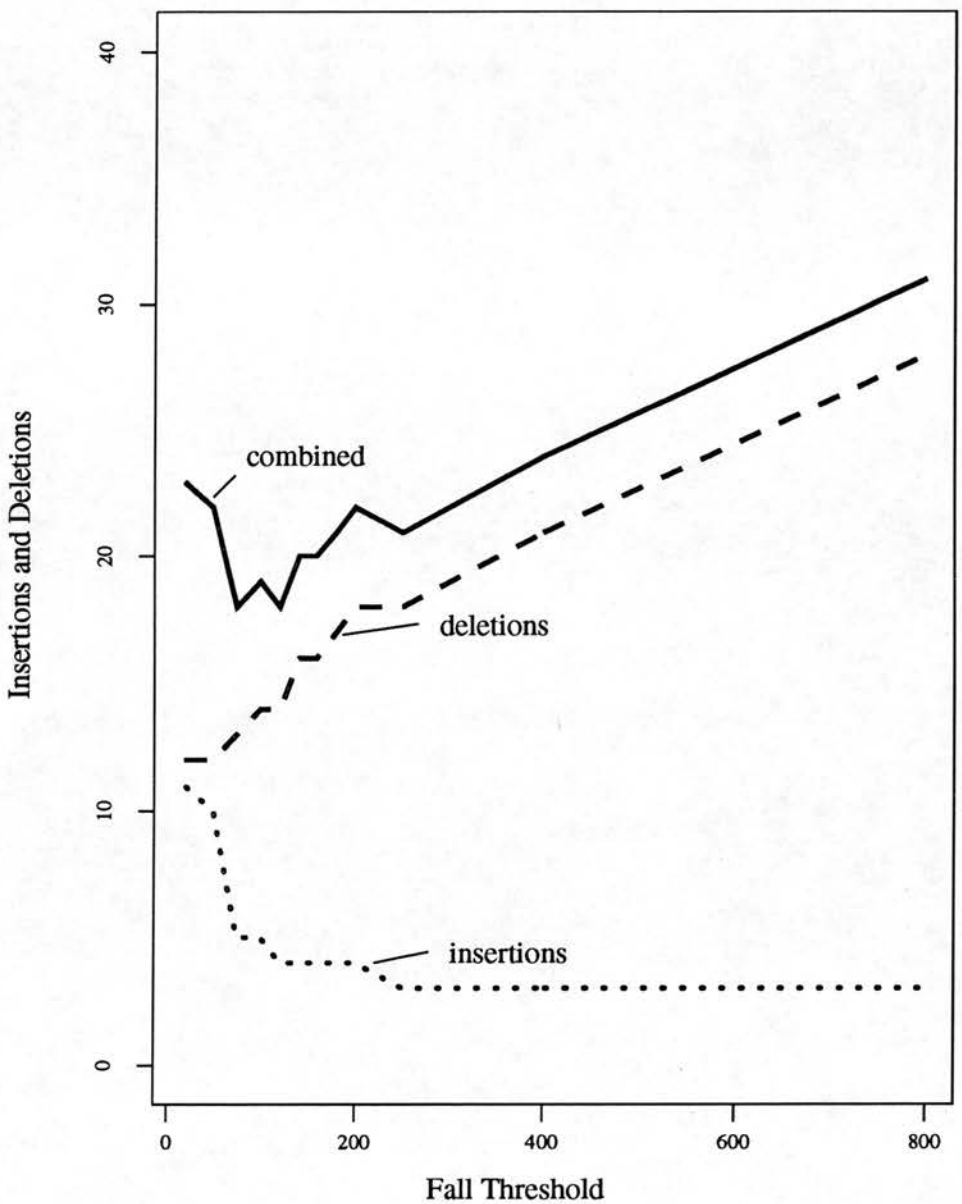


Figure 4.8: *Fall insertion, deletion and combined errors for an assimilation threshold of 0.125. The pattern in this graph is very similar to the graph showing rise performance.*

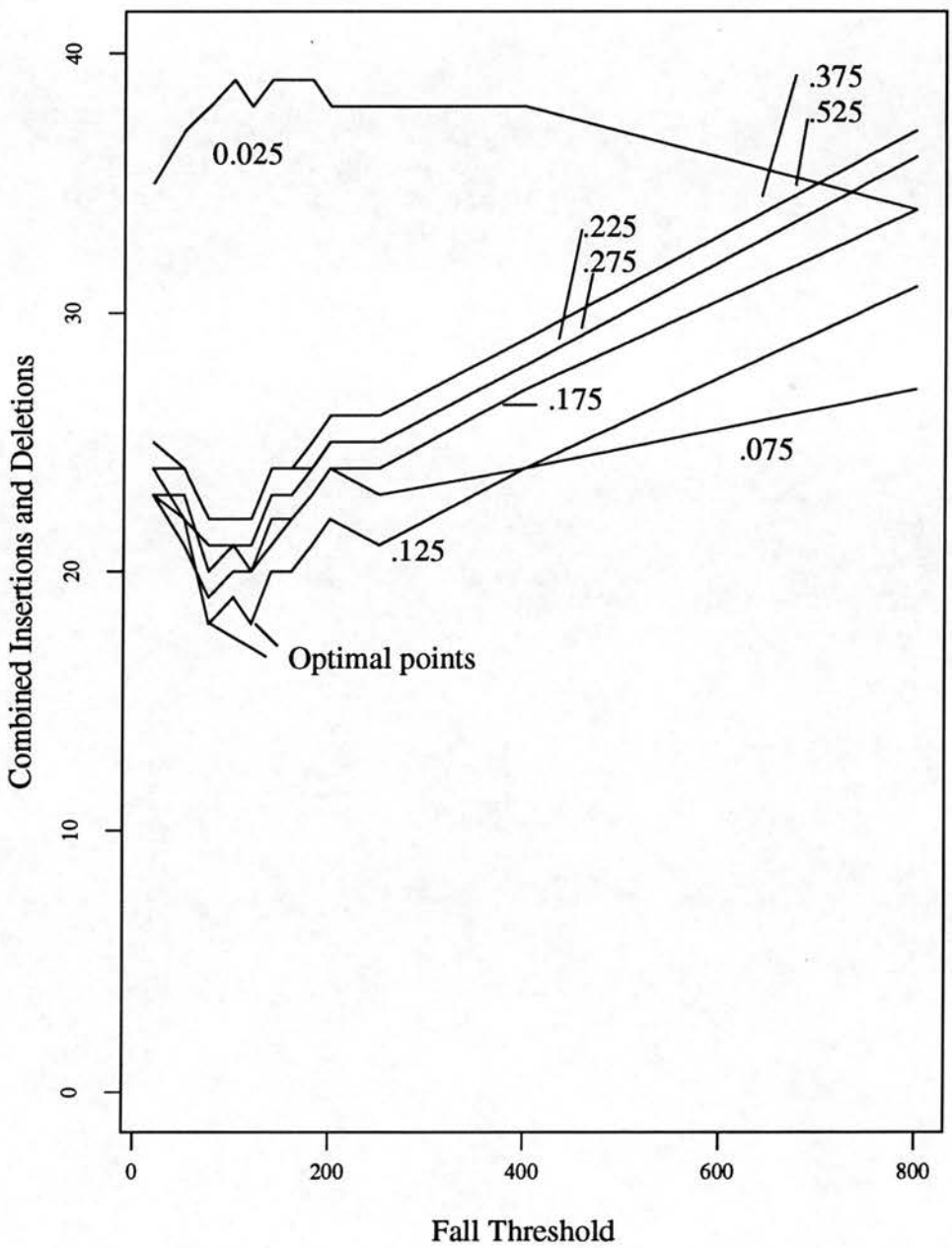


Figure 4.9: *Combined fall insertion and deletion errors for an all* ***fall-assim*** *values. Every curve apart from that representing* ***fall-assim*** *= 0.025 is similar, with the best results when* ***fall-assim*** *= 0.125.*

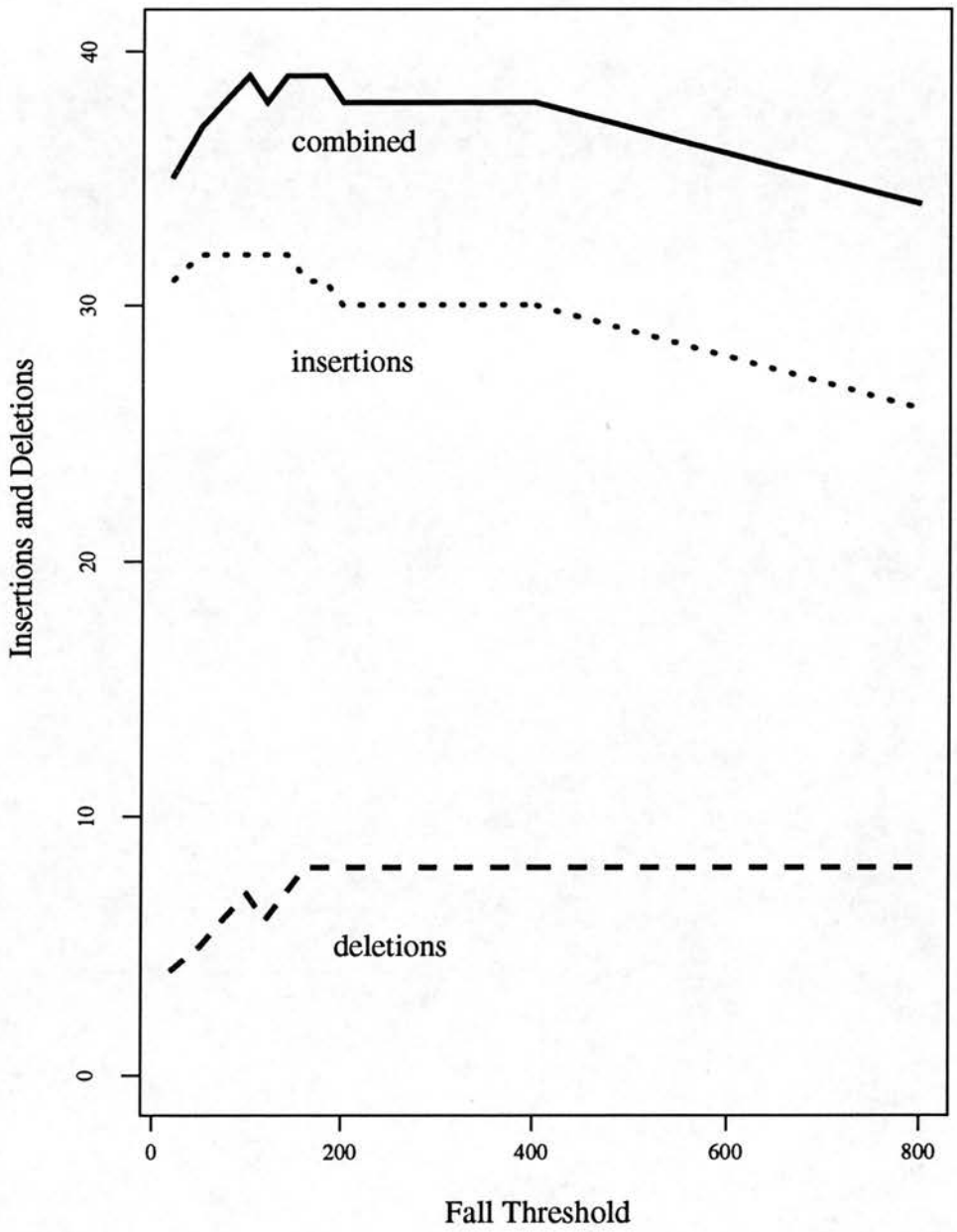


Figure 4.10: *This graph shows the errors when fall-assim was at its lowest value, 0.025. The unusually high error is due to a much larger number of insertions than usual. This results from pairs of fall sections separated by a connection section being classed individually instead of being assimilated.*

start-end The fraction of the marked section into which to look for the start of the shape.

stop-begin The distance after the marked end boundary.

stop-end The fraction of the marked section into which to look for the end of the shape.

The fractional variables, **start-end** and **stop-begin** were tested with values 0.0, 0.1 and 0.2. The other, absolute, variables were given values of 0.03s, 0.06s, 0.1s and 0.15s. All possible combinations of these four variables resulted in a matrix of 144 values. An experiment was run for each of these values and this time the assessment method as defined in section 4.3 was used to produce a score for each set of thresholds in the matrix.

Figures 4.11 and 4.12 show the distributions of the scores. These graphs show that the variation in overall score was very small. For both the rise and fall scores, the distribution is bimodal. In the case of the rise thresholds, all the high scores (> 52.0) occur when the start-begin variable is set to 0.0. In the case of the fall scores, the high scores occur when the stop-end variable is also set to its minimum value. Both these results imply that the rough transcription generally marks boundaries slightly later than they actually occur, especially when the section is neighboured by a connection element.

The worst score for both the rise and fall thresholds was when all the variables were at their lowest values, i.e. when the optimal fitting was at its most restricted. This is evidence for the need for the optimal fitting process. If little optimisation was allowed, the scores were significantly worse.

However, if a reasonable amount of optimal matching search area was allowed, all the resulting scores seemed to be similar, implying that so long as the thresholds are above a certain minimum, one can be guaranteed a good fit. Perhaps even more significantly, this shows that no matter what the optimisation thresholds were, the same precise boundaries were chosen. Thus the optimal matching process was *stable*, which is a very desirable property. By using this method, one can be reasonably sure that the fitting of the rise and fall shapes is non-arbitrary.

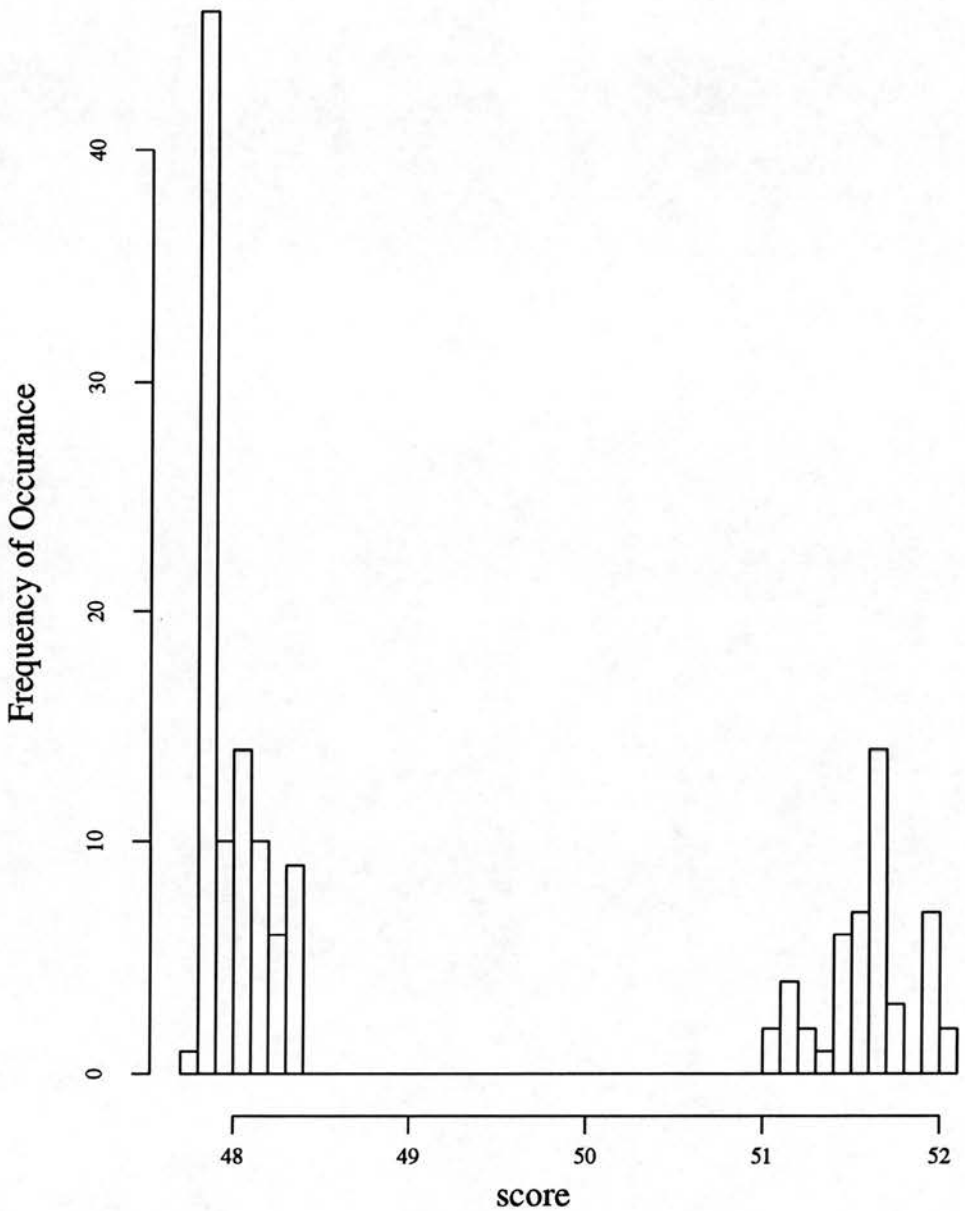


Figure 4.11: *Distribution of fall scores for different values of the optimisation thresholds. There is a clear bi-modal distribution shown here. All the scores relating to the right hand portion of the graph are distinct in that they have a stop-end value of 0.0.*

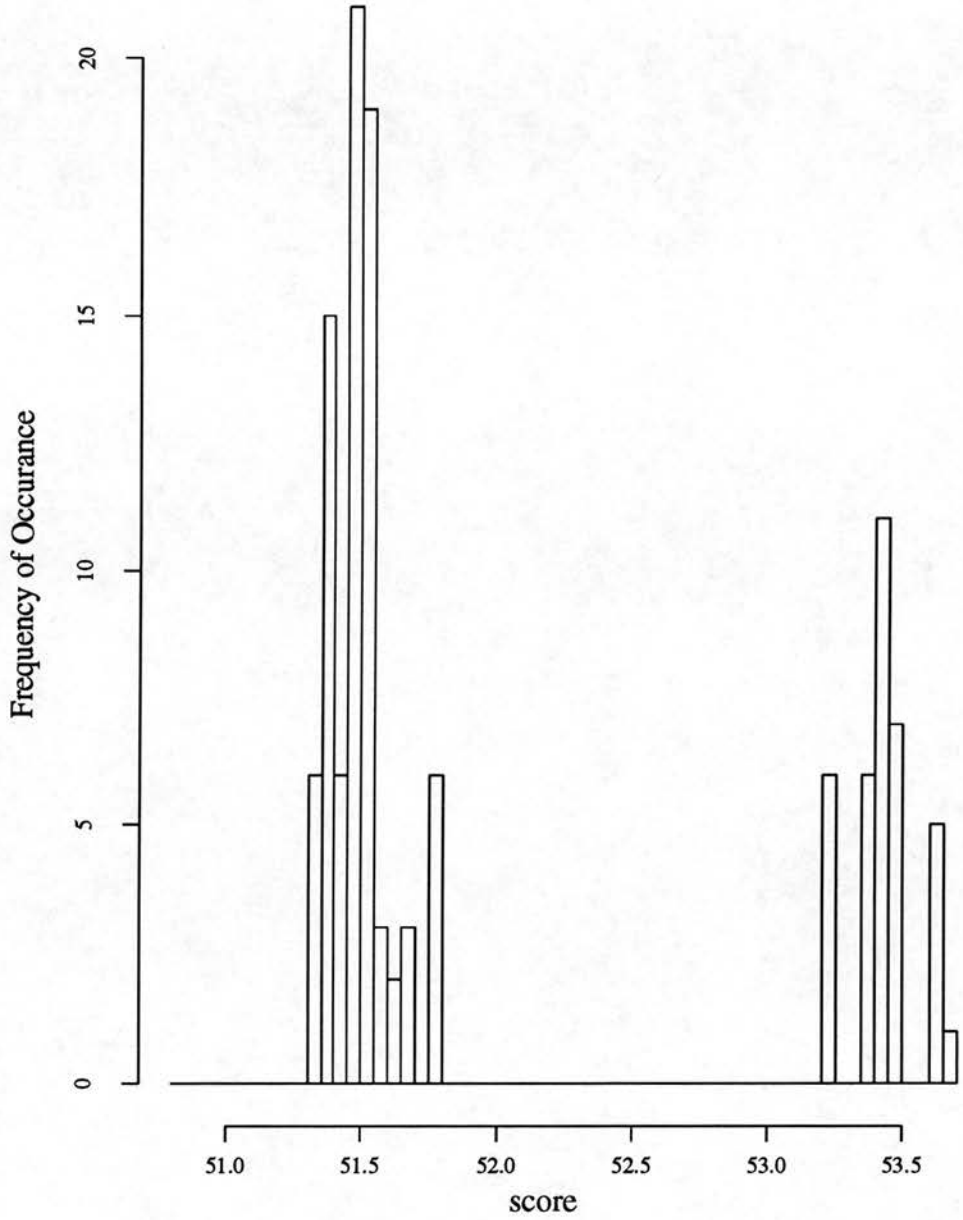


Figure 4.12: *Distribution of rise scores for different values of the optimisation thresholds. The same bi-modal distribution is apparent. All the scores relating to the right hand portion of the graph are distinct in that they have a start-begin value of 0.0.*

Threshold	Value
Rise Threshold	120Hz/s
Rise Assimilation Threshold	0.125s
Fall Threshold	120Hz/s
Fall Assimilation Threshold	0.125s
Rise start-begin	0.06s
Rise start-end	0.2
Rise stop-begin	0.1s
Rise stop-end	0.1
Fall start-begin	0.15s
Fall start-end	0.1
Fall stop-begin	0.1s
Fall stop-end	0.2

Table 4.1: *Optimised thresholds for data set A*

4.3.4 Final Thresholds

The thresholds shown in table 4.1 achieved the lowest number of insertions and deletions from training data set A. It should be remembered that the optimal matching thresholds are not particularly significant and any values above 0.1 will provide more or less the same results. Using these thresholds, the score for each utterance in data set A was calculated. The average score for this data set was 1.25.

4.4 Performance

4.4.1 Results

A series of tests was conducted using data sets A and B to assess the performance of the analysis system. Tests were carried out to judge the difference between open and closed test performance and to see how well the system analysed data from one speaker when trained on data from another.

The two data sets were divided into two equal sized groups. Thus four data sets were created: A1, A2, B1, and B2. Using the training procedures outlined above, a set of optimised thresholds was chosen for each of the four sets. A series of tests was carried out using these trained thresholds. The results are shown in table 4.2.

From these results we can average the scores for data set A and B so as to predict the overall open test performance. These results are shown in table 4.3.

Training data	Test Data	Nature of Test	Average Score
A1	A1	closed test	1.25
A2	A2	closed test	1.70
A1	A2	open test	1.85
A2	A1	open test	1.31
B1	B1	closed test	2.45
B2	B2	closed test	2.75
B1	B2	open test	2.79
B2	B1	open test	2.47
A	B	speaker independent open test	3.45
B	A	speaker independent open test	1.65

Table 4.2: *Results for closed, open and speaker-independent tests*

Data Set	Open Test Average Score
A	1.56
B	2.63

Table 4.3: *Open test average scores for data sets A and B.*

4.4.2 Discussion of Results

It is difficult to say with any certainty how well the recogniser has performed based on these results. There are no other equivalent systems with which to compare performance. All we can look at is how the system performs on different types of data, and how well it performs against the score of the random transcription. The average scores of 1.56 for set A and 2.63 for set B compare well with the random transcription score of 30.0. Even the worst score in the experiments of 3.45 is still much better than the random transcription score.

After arguing throughout the thesis about how important the formal approach is, and after spending considerable effort in designing and implementing the objective assessment criteria, ultimately the judgement of how well the system has performed must be somewhat subjective. Although the scores are very useful, I think the best indication of the systems performance is the perhaps most subjective one. Appendix B contains many F_0 contour plots from set A and set B marked with hand and automatic transcriptions. By examination of these plots, it is possible to obtain a good impression of how well the system performs, and where the remaining problems lie.

The system is not perfect, and four areas were judged to be the potential causes of error:-

1. Errors in F_0 measurement.

Data Set	Laryngograph	F_0 Tracker
A2	1.85	3.12
B2	2.79	5.17

Table 4.4: *Comparison of analysis system on laryngograph and F_0 tracked contours*

2. Automatic RFC labelling techniques not being sophisticated enough.

3. Legitimate discrepancies between computer and human labelled versions.

4. Problems with the RFC model.

F_0 Measurement

The laryngograph produced reliable F_0 contours for most utterances. Small errors may have occurred, but these would have had only a slight effect on the overall scores. The use of a laryngograph is a much more reliable means of extracting F_0 than the use of automatic F_0 extraction algorithms. To assess how badly inaccurate F_0 tracking would influence the recogniser, the speech waveforms of sets A and B were F_0 tracked using the algorithm described in Medan et al. (1991). These contours were analysed by the analysis system and the results compared with those obtained from the laryngograph contours.

Using the optimised thresholds for data sets A1 and B1, results were obtained for data sets A2 and B2 which are shown in table 4.4.

The analysis system performed approximately twice as badly on F_0 tracked contours than on laryngograph contours. Different tracking algorithms will no doubt produce different scores, but it is important to show that tracking errors do adversely affect the performance of the analysis system. The reason for the deterioration is that the analysis system has no expectation of F_0 contour errors. The performance for F_0 tracked contours could be improved by designing a more sophisticated contour preparation module.

Analysis Method

A major difference between the two sets of data was that the utterances of set A were designed so as to have mainly voiced speech. Data set B had no such restriction and had many more unvoiced segments than set A. One might think that this should not matter if the interpolation principle described in section 4.2.2 is correct, i.e. the unvoiced sections of the contour can be

compensated for by interpolation.

However, the problem doesn't seem to be with the unvoiced regions themselves, but with the behaviour of the F_0 contour before and after the unvoiced regions. It was explained in section 4.2.2 that F_0 contours often had sharp deviations before and after obstruents. If the part of contour on either side of this obstruent is of sufficient duration, the smoothing algorithm is powerful enough to smooth this short term deviation. However, in a short section of contour bounded by two obstruents, it is much more difficult to determine the underlying path of the contour and the smoothing algorithm is less likely to perform its smoothing properly. This was perceived as the major cause for the discrepancy of scores between the two data sets. This is a difficult problem to overcome: more or heavier smoothing alone is not the answer as it was explained in section 4.2.2 that too much smoothing can distort the intonation of the contour.

Because of the residual segmental effects, the thresholds in the broad classification module had to be higher than they otherwise would have been as the thresholds were used to distinguish pitch accents from segmental perturbations instead of their intended purpose, which was to distinguish pitch accents from connection elements. Owing to these high thresholds, it was occasionally the case that the rises of small **H** accents were labelled as connection elements. What enables the human labeller to avoid this confusion is partly that the rise section looks as if it belongs to a pitch accent (drawing on the labeller's previous labelling experience), or that the rise section is occurring in a logical position, e.g. on a stressed syllable, (drawing on the labeller's ability to listen and understand the utterance and also on the knowledge of stress patterns within the utterance).

The most natural way to attack this problem is to include knowledge of the segmental structure of the utterance as input to the analysis system. This has two potential benefits. First, if the locations of obstruents are known, it should be possible to concentrate smoothing in these regions thereby reducing the influence of the segmental perturbations. Using very heavy smoothing in these specific areas should not distort the intonation of the utterance greatly. Secondly, the location of vowels may be useful in giving indications as to where accents are most likely to occur.

One could go further and say that the whole principle of the automatic analysis system is at fault. Other techniques could be used for the detection of pitch accents or RFC elements. A neural network was tried previously, but with little success. It may be the case that such

an approach is worth persevering with, and a much more sophisticated neural network system might eventually perform better than the rise/fall threshold system described here.

Hand Labelling

Occasionally, when comparing the hand and automatically labelled versions, I came to the conclusion that I had made a mistake in my hand labelling and that the automatically labelled version was in fact better. Nearly always this involved small pre-nuclear **H** accents There were no cases of arbitrary labelling of nuclear accents.

If two experienced human labellers were to mark the same set of data it would be unusual for them to produce exactly the same labels. A major aim of this thesis was to try an reduce the amount of arbitrariness in labelling to a minimum. Arbitrary labelling decisions have largely been removed from this system, but still slight differences will linger, especially with the classification of small pitch accents.

Problems with the RFC system itself

It is important to discover if the discrepancies between the automatically labelled and hand labelled result from difficulties with the model itself. If so, this implies that the model is flawed and needs improvement.

It is very difficult to say for certain that none of the analysis errors arose from problems in the model, but it seems true that if any errors did occur they were very rare and none have been located explicitly. Nearly all the errors arose from residual segmental influence in the F_0 contour. The remainder arose from discrepancies in labelling, but none have been found that indicate that the model itself is at fault. This of course does not really prove anything, but it does show that on this data the model has no problems. Therefore as the model has been able to account for all the data in a satisfactory way, its worth is dependent on how representative the contours in set A and B are of the native speaker set.

Considerable thought was put into the choice of the data sets (see section 3.2.1), but it is impossible to claim that every intonation effect has been covered.

4.4.3 A Note on Assessment

The assessment criteria were designed to reflect human observations about the quality of transcriptions, i.e. utterances which looked like they had bad transcriptions would receive high penalty scores. However, the sophistication of the assessment criteria could only be taken so far - sometimes what looked like a good transcription turned out to receive a high score. This was mainly owing to the fact that all pitch accents were treated equally: a small accent in a pre-nuclear position would receive the same score for being labelled incorrectly as a large nuclear accent. Although the exact criterion depends on what the system is to be used for, in most cases it will be more important to label nuclear accents correctly than small head accents.

From studying the automatic transcriptions of the utterances, it is obvious that the system labels large accents more correctly than small accents. Nuclear accents are often the largest in the phrase, so that the system marks nuclear accents more successfully than pre-nuclear accents. By using these factors, a more sophisticated assessment method might give the system better scores.

4.4.4 Variation in Scores for Different Data Sets

In the speaker dependent open tests, the average score for data set A was 1.56 whereas the average score for data set B was 2.63. Thus the automatic analyser works significantly better on data set A than on data set B. There were many differences between the two data sets (the two data sets were chosen so as to be as different from one another as possible) but the difference which was seen as being responsible for the poorer results was that data set B had many more obstruents than data set A.

Although this may have accounted for the majority of the errors, other factors may be relevant. The analysis system was developed using data set A. Development involved trying a certain analysis technique, looking at the transcriptions that were produced and then adapting appropriate parts of the system. This was the way in which the assimilation module was introduced. Just as data set A benefited significantly from the introduction of a new module into the system, so new modules or approaches may be needed to improve the performance of the system on data set B. It is difficult to guess what improvements might be necessary until the more major problem of segmental influence is tackled.

Differences in Thresholds across Data Sets

The differences between open and closed tests in all cases were small. This was due to the simple fact that the training algorithm produced very similar thresholds for each data set. If the thresholds of two data sets are the same, there will be no difference between the results of an open and closed test. In all cases, the preferred assimilation threshold was 0.125 seconds. The size of this threshold was probably more a consequence of the spacing of points on the re-sampled 50ms contour than any inherent feature of the speaker's intonation. Both groups in data set A chose rise and fall thresholds of 120Hz/s. The thresholds in data set B were between 100Hz/s and 140Hz/s but there was not much difference in score between these thresholds. In every case the optimisation thresholds were different, but as these made only slight differences in the overall score, this variation was not important.

Only two speakers were used and although differences between the intonation of these speakers could be observed, more speakers would be needed for a proper study on speakers' intonational variance. We can say that the types of rise and fall element produced by these speakers are quite similar. If a method of eliminating segmental influence could be found, it might well be possible to construct a speaker independent labelling system. The differences in performance between the two speakers are more a result of the prescence of obstruents than the result of fundamental differences between speakers' intonation.

4.5 Synthesizing F_0 Contours from an RFC description

The computer implementation of the intermediate-F_0 mapping process was very straightforward. A computer program was written that took a list a of RFC elements as input and produced an F_0 contour as output. Each element in the input list had an entry for *type*, a *duration* and an *amplitude*, similar to the list shown in section 3.3.5.

This particular mapping is interesting in that it existed *only* as a computer implementation; this mapping was never performed manually because it would have been very tiresome to have to draw F_0 contours by hand.

In section 2.1.3 the following criterion was given for testing the phonology-F_0 mapping.

> Given the phonological description for each utterance in a set of data, is it possible to derive an F_0 contour that is indistinguishable from the measured F_0 contour for that utterance?

We can use the same basic criterion for the intermediate-F_0 mapping. By synthesizing a F_0 contour from its RFC description and comparing this to the utterance's original contour, the intermediate-F_0 mapping can be tested.

As should be clear from the discussion in section 2.1.2 on F_0 representation, it is not a simple matter to determine the similarity of F_0 contours. However, it is helpful to give *some* indication of how accurately the intermediate-F_0 mapping synthesized F_0 contours.

It is important to remember that the intermediate-F_0 mapping only attempted to synthesize the intonational part of the F_0 contour, and segmental and pitch perturbations were not modelled. Two comparison tests were therefore carried out; one on the original F_0 contours, complete with unvoiced regions, pitch and segmental perturbations; the other on the smoothed, continuous contours that were used in the analysis experiments.

A simple assessment method was developed based on the euclidean distance between equivalent points on the two contours. The number of points in a contour is dependent on the duration of the contour, and also the sampling rate at which the F_0 values are specified. To normalise for this, the final distance score was taken as the average root mean square distance between all pairs of points. The use of the average distance is helpful in that this can be expressed in Hertz. Also, one can think of this value as representing the average distance that the synthesized contour differs from the original at any given point.

In regions where one contour was voiced and the other was unvoiced, no score was given. This was seldom important in the smoothed contour, as these regions were very small and only occurred at the starts and ends of phrases. For the unsmoothed contours, there were many more unvoiced regions owing to unvoiced segments. The synthesized contours were given the same voiced/unvoiced patterns as the original unsmoothed contours, and so there was zero difference in these regions.

Such an assessment method is crude, and takes no account of a listener's ability to perceive F_0. However, the design of a perceptual test that could give clear answers as to whether F_0 contours are the same or not would be very difficult, and it would be even more difficult to generalise the results of such a test to conclude that *every* contour the intermediate-F_0 mapping produces is acceptable.

4.5.1 Synthesis Results

For data set A, the synthesized contours deviated from the original smoothed contours by an average of 5Hz. The deviation in data set B was on average 10Hz. For the unsmoothed contours, the deviation was approximately twice as much, with data set A having an average deviation of 11Hz and set B 18Hz.

This numerical evidence supports the subjective view that the synthesized contours match the originals well, as shown in appendix B. It would be useful to compare how well the RFC synthesis system models contours compared to other synthesis systems, but this is difficult as we have no well defined method for deriving the intermediate descriptions for these systems. It would probably be the case that the originators of these systems could achieve better fits with their models than I could. However, it is interesting to note that in a simple test, the Fujisaki model had an average distance of 7.0 Hz on selected **H** accents from data set A. Thus on the types of contour which the Fujisaki system can model, its accuracy is similar to that of the RFC system.

4.6 Implementation of the Phonology-Intermediate Grammar

The problems of producing a phonological description of an utterance's intonation from a RFC description were discussed in detail in the previous chapter. The incompleteness of the theory makes it is impossible to design a complete implementation of the intermediate-phonology mapping system. The system described below deals with the part of the mapping which was discussed is most detail in chapter 3, namely *tune*.

4.6.1 Intermediate-Phonology Tune Mapping

The tune analysis system worked by using a system of rules which examined each RFC section in the context of its immediate neighbours and gave it a phonological class.

The rules are given in table 4.5. An asterisk (*) indicates that any element is allowed in the indicated position. Any sequence of elements that is not given below does not receive a phonological classification. Sometimes a single phonological class is associated with two RFC elements, as in the case of $\mathbf{H}_l$. In the output of the computer program, a standard was defined whereby the phonological element was aligned with the first RFC element.

Previous Element	Current Element	Next Element	Phonological Category
*	rise	fall	**H**
fall	fall	*	$\mathbf{H}_d/\mathbf{L}_a$
connection	fall	*	$\mathbf{H}_d/\mathbf{L}_a$
*	connection	*	**C**
fall	rise	rise	$\mathbf{B}_i$
fall	rise	connection	$\mathbf{B}_i$
connection	rise	rise	$\mathbf{B}_i$
connection	rise	connection	$\mathbf{B}_i$
*	rise	silence	**B**

Table 4.5: *Rules for phonological classification of tune*

Decisions which could not be made at this stage were left open; for example it is impossible to chose between a $\mathbf{H}_d$ with no rise element and a $\mathbf{L}_a$ from F_0 contour analysis alone.

Timing information was later used to distinguish between $\mathbf{H}_d$ and $\mathbf{L}_a$. If the fall occurred early in the syllable, the accent was of type **L**; if the fall occurred later the type was **H**. Timing was also used to distinguish normal **H** from $\mathbf{H}_l$. If the peak was later than usual, the accent was marked as late.

To distinguish these types of accents, vowel timing information was necessary. Vowel boundaries were marked by hand on all the utterances where timing information was necessary. The "vowel-onset-to-fall" (VF) distance discussed in section 3.4.7 was used to distinguish accents which differed in their alignment to the vowel. The small number of $\mathbf{H}_l$, $\mathbf{H}_d$ and $\mathbf{L}_a$ accents made it impossible to formulate rigorous rules as to how these accents differ in timing, but the timing behaviour discussed in chapter 3 was used to make a provisional rule that could distinguish these accents.

Any fall element that started before the vowel onset was classed as $\mathbf{L}_a$, and any fall element occurring more than 80ms after the vowel onset was classed as $\mathbf{H}_l$. This rule correctly distinguished all the $\mathbf{L}_a$ accents and most of the **H** and $\mathbf{H}_l$ accents, but it would be prudent to assume that such a rule is very specific to this particular database.

H and $\mathbf{H}_d$ were distinguished by a similar rule which stated that if an accent's fall element had more than twice the amplitude of that accent's rise element, the accent was classed as $\mathbf{H}_d$. This did not always agree with the hand transcriptions, but as discussed in section 3.5.2 the marking of some features was somewhat arbitrary, and disagreement between the hand and automatic systems is more due to problems with the theory than with the automatic system.

The elevated feature was not implemented in the automatic system as any thresholds between **H** and $\mathbf{H}_e$ would have been purely arbitrary.

Other Aspects of the Intermediate-Phonology Mapping

It was not possible to give a detailed prosodic phrase structure description, but phrase boundaries were marked were pauses and boundary elements occurred.

Scaling information regarding prominence and pitch range was not analysed, again due to the incompleteness in the theory.

4.6.2 Phonology-Intermediate Mapping

The process of producing a sequence of RFC elements from a phonological description is quite straightforward. The other aspect of the phonology-intermediate mapping process, namely the provision of amplitude and duration information for the RFC description, is considerably more difficult and prevented a computer phonology-intermediate mapping from being implemented.

The completion of this part of the system is dependent on suitable description systems being found for prominence, pitch range and phrasing, and also the completion of the theory relating to the numerical specification of the RFC elements.

4.7 Discussion of the Computer Implementation

The automatic analysis system was quite successful at deriving the RFC description and the phonological tune description of utterances. However, its performance was significantly worse than that of a human labeller. Two main problems caused errors in the analysis system; the differences between the model's legal set of contours and the native-speaker set, and the inability of the system to locate pitch accents reliably.

The main differences between the two sets of F_0 contours lay in the presence of unvoiced regions, obstruents and pitch perturbations in the native speaker set. Processing was used to lessen these effects in the contours of the native-speaker set, resulting in contours which were free of unvoiced regions and less influenced by the effects of the pitch perturbations and obstruents. However, the smoothed contours were not entirely free of segmental perturbations and this often resulted in segmental bumps being incorrectly labelled as pitch accents.

The other major problem was that the theory provided no formal method for locating pitch accents. In practice, the analysis system used the gradient of the contour as a guide and labelled steeply rising sections as rises and steeply falling sections as falls. In most cases this was adequate, but in a significant number of cases, segmental bumps were labelled with rise or fall elements.

Accents always occurred in some relation to the vowel of syllables. If the vowel structure of the utterance was known, this might be useful in giving indications as to where pitch accents are likely to occur. Thus rises which occurred in improbable positions could be ignored as it would be likely that these were due to segmental influence.

The problems associated with segmental perturbations and pitch accent location are interrelated as it is the usually the presence of segmental perturbations which causes the system to incorrectly label pitch accents. An obvious future improvement in the system would be to use a segmental transcription of the utterance. This would enable the contour preparation module to concentrate smoothing in obstruent regions, help give the element location part of the system potential locations for pitch accents and give timing information that could distinguish between phonological accents classes.

The intermediate-F_0 mapping produced F_0 contours for an utterance given that utterance's RFC description. Often, these synthesized contours were very similar to the original contours. The main differences were due to the original, smoothed contours still being influenced by the utterance's obstruents.

The intermediate-phonology mapping was straightforward to implement, but it was impossible to distinguish between certain phonological classes without reference to vowel timing.

4.7.1 Conclusions

As there are no equivalent system with which to compare the one presented here, it is difficult to give an objective judgement as to the system's performance. It is clear that the RFC analysis system still makes errors, but the assessment results of 1.56 and 2.63 for data sets A and B are considerably better than that for the approximated worst case of 30.0. Similarly, from the graphs in appendix B, and the average distance scores of 5Hz and 10Hz for data sets A and B, we can conclude that the intermediate-F_0 mapping accurately synthesizes F_0 contours.

Perhaps the most important point for discussion is whether the performance of the automatic

analysis system tells us anything about the ability of the the theory to accurately model the intonation of English. It was the intention of the RFC labelling rules described in section 3.3.5 to define as strictly as possible the manner in which F_0 contours should be hand labelled. From the large amount of agreement between the human and automatic transcriptions, we can conclude that these labelling criteria are non-arbitrary. The fact that the RFC description can then easily be linked to the phonological description of the utterance (as least as far as tune is concerned) indicates that the RFC description level is a useful method of describing intonation within the phonology-F_0 grammar.

Chapter 5

Conclusions

5.1 Summary of Main Findings

It was proposed that a formal approach, commonplace in traditional generative phonology, could be used to describe the relationship between the phonological description of an utterance's intonation and its F_0 contour.

Chapter 2 used this formal approach to review some of the existing phonetic models. It was shown that it would be difficult to use Pierrehumbert's theory for the basis of a formal model because of the difficulty in providing a F_0-phonology mapping for this theory. The Fujisaki and the Dutch models used a more formal approach, but these systems were insufficiently powerful to model all the commonly occurring types of English F_0 contours. Because of the difficulties with existing phonetic models, work focused on designing a new model.

The use of the formal approach, which described systems in terms of levels, grammars and mappings, was useful in showing how different theories tackled the phonetic modelling task. The concept of redundancy was introduced which was useful in comparing the intermediate levels of the various systems. It was argued that the mapping between the F_0 contour and phonology could be seen as a redundancy reducing exercise and intermediate levels occurring in different positions in the phonology-F_0 grammar would have different redundancy content. Thus it was seen as unwise to directly compare intermediate levels as each was individual to its particular theory. It was also shown that the position of an intermediate level merely dictated the division of labour between the phonology-intermediate and intermediate-F_0 grammars, and did not in itself make the provision of a phonology-F_0 grammar easier.

Chapter 3 introduced a new phonetic model. The intentions underlying the design of this

model included:-

- The model should be able to synthesize F_0 contours which closely resemble naturally occurring contours.
- The system should be able to model all the intonational effects of English.
- The grammar for this system, which governed the relationship between the phonological and F_0 levels, should be amenable to both analysis and synthesis. Given any F_0 contour, the model should produce the correct phonological description; given any phonological description, the model should produce the correct F_0 contour.

The new model used an intermediate level of description, which was based on the principle that F_0 contours could be described as a sequence of rise, fall and connection elements. Each of these elements was described by an equation, which when given the appropriate parameters, could be used to synthesize an F_0 contour. Fall shapes were used to describe the falling parts of pitch accents; rise shapes were used to describe the rising parts of pitch accents and the rises often observed at the beginnings and ends of phrases; connection elements were used everywhere else.

A new phonological description system was developed which also used sequences of elements to describe intonation. Pitch accents were described as belonging to one of two basic classes, **H** and **L**. Within the **H** class, accents were subcategorised using the features "downstep", "elevated" and "late"; within the **L** class, a single feature "antecedent" was used. In addition to the pitch accent elements, two other phonological elements were used: **B** for boundary rise elements and **C** for connection elements.

Chapter 3 also described the grammars which link the F_0 RFC and phonological levels. The intermediate-F_0 grammar, which linked the F_0 level and the RFC level showed how to create F_0 contours given an RFC description, and also how to describe an F_0 contour using the terminology of the RFC system. The phonology-intermediate grammar, which linked the RFC and phonological levels, was less complete. The completion of this grammar depended on providing a mechanism for specifying the durations and amplitudes of the RFC elements from the phonological description.

Chapter 4 described the computer implementation of the grammars in the new phonetic model. Most of this chapter concentrated on describing the automatic RFC analysis system (the

implementation of the F_0-intermediate mapping). The system worked by analysing smoothed F_0 contours in two stages. The first stage involved the detection of rise and fall sections; the second stage involved determining the precise amplitudes and durations of these sections.

An objective transcription assessment method was developed that could compare two transcriptions. The assessment produced a score representing the similarities between the transcriptions, with similar transcriptions receiving low scores. Hand labelled transcriptions of the F_0 contours in data sets A and B could be compared with automatically produced transcriptions to assess the analysis system's performance. A training method was developed which used the assessment criteria to optimise the various thresholds in the analysis system.

Final results showed that the analysis system achieved an average score per utterance of 1.56 on data set A and 2.63 on data set B. These scores compared well with an estimated worst score of about 30. Errors in the automatic transcriptions stemmed mainly from the differences between the native-speaker set of F_0 contours and the model's legal set. The main difference was due to residual segmental influence in the F_0 contour which the smoothing part of the analysis system had failed to remove.

Chapter 4 also explained the computer implementation of the intermediate-F_0 synthesis mapping. It was claimed that this part of the phonetic model was successful due to the similarity of the F_0 curves produced by this mapping to naturally occurring contours. The implementation of the intermediate-phonology mapping was also explained, but this was left unfinished due to the incompleteness of the theory.

A major motivation behind the computer implementation of the mappings was to prove that the model could be formally defined, and also to objectively test the ability of the model in analysing and synthesizing F_0 contours. The automatic analysis system was worse at analysing contours than a human labeller, and the synthesis system did not produce F_0 contours that were indistinguishable from naturally occurring F_0 contours, but it is argued that the system performed considerably better than any of the models review in chapter 2.

5.2 Further Work

5.2.1 Numerical Mapping in the phonology-intermediate Grammar

The largest area of remaining work is the completion of the numerical mechanism in the phonology-intermediate grammar. This involves two main areas; segmental scaling and phonological timing and scaling.

It is recommended that the approach taken for the completion of this work should that which was advocated in section 3.4.2, where a segmentally normalised intermediate level is added to the phonology-intermediate grammar. The introduction of the segmental normalised RFC level would split the phonology-intermediate grammar into parts, the phonology-normalised grammar and the normalised-intermediate grammar.

The aim of the proposed intermediate-normalised mapping will be to convert a standard RFC description into one which is free of segmental influence. The amplitudes of the elements of a pitch accent are mainly governed by the accent's prominence, but intrinsic vowel F_0 also plays a significant role. From a study of pitch accents in different segmental contexts, it should be possible to derive relationships showing how vowel-type influences F_0. Once such relationships have been found, they can be used to normalise the RFC amplitudes.

Slightly more complicated is the process of normalisation with respect to duration. As with intrinsic vowel height, each vowel can be thought of as having an intrinsic duration (Allen et al., 1987). However, there is also the additional factor of *syllable structure* to be compensated for. Syllables can vary in the number of segments they contain, and syllables with long voiced regions are liable to have pitch accents with longer duration (Gartenberg and Panzlaff-Reuter, 1991). As with the proposed study on intrinsic vowel F_0 the recommended approach is to analyse data where pitch accents are realised in a wide range of syllables with different structures. Any relationships that can be deduced from this study can be used in normalisation process of the intermediate-normalised grammar.

Once the influence of segmental effects has been removed from the normalised RFC description, the remaining differences between element durations and amplitudes should be due to phonological factors. *These* differences in durations and amplitudes will have to be modelled by the phonology-normalised grammar.

From the discussion in chapter 3, we can identify the following factors in the phonology-

intermediate grammar which influence RFC duration and amplitude.

- The timing differences between **H** and $\mathbf{H}_l$.
- The timing differences between $\mathbf{H}_d$ and $\mathbf{L}_a$.
- Pitch accent prominence. The provision of a mechanism to relate RFC amplitude to phonological prominence depends crucially on the design of how to express this prominence at the phonological level. The amplitudes of the elements in a normalised RFC description should be related to the prominence of that accent. Studying the behaviour of segmentally normalised accents should resolve questions as to whether prominence can be expressed (a) using the "elevated" feature; (b) using a more complicated phonological mechanism; or (c) using a continuous scale, which would imply that prominence is truly paralinguistic as Pierrehumbert claims.
- Pitch range. As with prominence, a systematic method for describing pitch range needs to be developed.
- Declination resets at phrase boundaries. A major issue in intonational phonology involves how to describe and determine phrasing (see section 2.2.2). Phrasing is manifested in the RFC system by the use of boundary elements. The completion of the phonology-intermediate grammar depends partly on finding a suitable way to describe the strengths of boundary rises on the phonological level, and partly on the resolution of what prosodic phrase units exist and how they relate to one another.
- Declination. The scaling of all the elements will depend on the declination of the system - the amplitudes of elements cannot be compared directly unless declination effects are compensated for.

It should be straightforward to determine the timing differences between **H**, $\mathbf{L}_a$ and $\mathbf{H}_l$. The analysis of the data analysed so far suggests that there are clear cut differences between these accents.

The **H**, $\mathbf{L}_a$ and $\mathbf{H}_l$ accent timing problem is simply a matter of providing a mechanism that can link a phonological and RFC description. The other effects are somewhat different in that an appropriate phonological description system has to be designed as well as the phonology-intermediate grammar.

However, one could argue that a description system for prominence, pitch range and phrasing already exists, as the numerical specification of the RFC elements could be used for the phonological description. Prominence, pitch range and phrasing could all be described in terms of the Hertz value of the appropriate RFC elements. The disadvantage to such an approach is that the phonological description system would have an undesirable amount of redundancy and would not reflect categorial differences in phonological classes, i.e. everything would be represented as being continuous.

The literature shows quite clearly that there is much disagreement over these aspects of intonation, which may be an indication that we are still some way from a proper understanding of them. If this is the case, it is unlikely in the short-term that a phonology-intermediate grammar will be developed that uses an ideal low redundancy phonological description system. A more practical solution may be to use the high redundancy Hz description and then conduct investigations so as to reduce the redundancy in the phonological description as much as possible, with the end goal of having no redundancy in the phonological description whatsoever.

5.2.2 Algorithm Improvements

The completion of the implementations of the phonology-intermediate grammar is largely dependent on the completion of the theory underlying this part of the model. Once the mechanisms to control tune, phrasing, scaling, timing and segmental influence have been finalised, it should be straightforward to implement these mappings on computer.

The major problem with the implementation of the F_0-intermediate mapping is that segmental influence is still present in the contour even after the smoothing operations have been performed. In the conclusion to chapter 4, it was argued that if the position of the obstruents were known, heavier smoothing could be applied in these areas and this would help reduce the segmental perturbations in the contours. If a segmentation was available to the system, this might be useful in given the system indications as to where accents are likely to occur, e.g. in some relation to the vowels.

The implementation of the intermediate-F_0 mapping follows the theoretical specification of the mapping exactly, and the F_0 contours which are produced by the synthesis program are very similar to real F_0 contours. It could be argued that it is the function of the intermediate-F_0 mapping to re-introduce the segmental effects on the contour. If this is seen as being

needed, then segmental modelling could be incorporated into the intermediate-F_0 mapping. Alternatively, a post-processing module could be included in the phonetic model which adds the segmental effects to the synthesized contour. This extra module would be the synthesis equivalent of the contour preparation module.

5.3 Applications for the Model and Computer Implementation

5.3.1 Speech Synthesis

It was stated in chapter 1 that a motivation for starting this work was to design a phonetic model that would be useful for speech synthesis purposes. The model as it stands cannot be used a speech synthesis system owing to the incompleteness in the phonology-F_0 mapping.

It should be easy to provide simple rules to control the numerical content of the RFC elements. Many of the phonetic models used for speech synthesis that were discussed in chapter 2 used such rules. This type of approach would only be recommended as a short term solution because it does not make use of the power of the phonetic model. The longer term solution would be to complete the phonology-intermediate grammar and implement this in a computer phonology-intermediate mapping. If this was done properly, the benefits of using the phonetic model should become apparent, and the intonation produced should sound considerably more natural than that of conventional synthesizers.

The use of a comprehensive phonetic model that can produce natural sounding intonation should have many benefits over conventional intonation models in speech synthesis systems. At the simplest level it may help produce more natural sounding speech as it synthesizes F_0 contours that are very similar to natural occurring ones. More importantly, I think the system will be able to produce many intonational effects that other systems cannot. While the input to the phonetic model is not of a sufficiently high level to say "pronounce this sentence in a sarcastic manner", it should be possible, with a little knowledge of how sarcastic utterances are described in the phonological system, to specify the phonological specification for a sarcastic sentence and have the model produce an appropriate F_0 contour.

The parts of the phonetic model which have been completed to date have all been speaker independent; e.g. the monomial function is the same for both speakers in the data. Eventually some speaker dependent parameters will be found (e.g. females are certain to have different

scaling mechanisms from males). It should be possible to account for differences between speakers with the system, and allow for the production of F_0 contours which are characteristic of a particular speaker's voice. Just as the diphone synthesis method can capture segmental differences across speakers (Taylor and Isard, 1990), so should the intonation model.

5.3.2 Speech Recognition

The analysis system presented here could constitute a "front end" for a speech recognition system's language model. The analysis system would take the F_0 contour output from a F_0 tracking algorithm and produce a phonological description of that utterance's intonation.

Most speech recognition systems concentrate on analysing the segmental content of utterances and do not make use of prosody. More recently, some systems have made use of prosody, but this has been mainly limited to the analysis of stress (Waibel, 1986), (Hieronymus, 1992).

The intonation content of utterances has largely been ignored as a source of useful information. This is primarily because the type of speech that most speech recognition systems deal with is simple enough so that utterances can be understood from analysis of their segments alone. Many speech recognition systems deal only with spoken commands and declarative sentences. In such situations, intonation is largely unnecessary for the recognition task.

It is only in more complicated dialogue situations that intonation will start to play an important role. If a dialogue system is intended to recognise relatively unconstrained speech, there will be situations where questions and statements risk the possibility of being confused. It is in situations such as this that the automatic recognition of intonation information will be useful (Longuet-Higgins, 1985).

Each of the four main areas of intonational phonology (tune, phrasing, scaling, timing) is potentially useful in a speech recognition system. Some phonological information could probably be of use in the language models present in the more sophisticated contemporary recognition systems. Other aspects are representative of much higher level linguistic processes and may not be of use to practical systems for some time to come. The potential uses of the four main phonological aspects of intonation will now be examined.

In the past, speech recognisers have used very simple language models often employing finite state grammars to control syntax (Woods, 1985). As the vocabulary and complexity of

the language domain increases, such approaches become unsuitable and more sophisticated grammars are needed. Speech recognition systems do not tend to use traditional linguistic approaches to syntax[1] as these grammars often deal with ideal speech and don't take account of disfluencies and "ungrammatical" utterances which often occur in real-world situations. Much current work concentrates on providing language models which are probabilistic, and so can deal with "noisy" input, but also make use of aspects of more traditional syntactic grammars which provide general powerful rules to guide parsing (for examples see Seneff *et al* (1992) and Morimoto (1992)).

These systems try to recognise words from analysis of the segments of the utterance, using the language model to restrict the order in which words can occur. Prosodic phrase structure information would be useful in such a system as this would give indications of where phrase-breaks might occur, independently of the segmental side of the recognition system. Veilleux *et al.* (1992) describe a system which makes use of durational cues to locate potential phrase boundaries. This phrasing information is then used to resolve ambiguous parses. Analysis of the boundary elements produced from the RFC analysis system could be used in much the same way, as these show that an intonation phrase break has occurred.

The location of pitch accents is also a readily usable part of a phonological description. On the simplest level, the fact that a syllable is associated with an accent is a indication that the syllable is stressed. So long as potential stress shifting effects are taken into account, this could be used to determine the lexical content of an utterance. A more sophisticated language model could make use of the fact that words with nuclear accents are more important than those without, and so special attention should be given to nuclear accent bearing words during semantic analysis.

Knowledge of tune type will help distinguish questions from statements: utterances which end in either rising connection elements or boundary elements will most probably be yes/no questions. The type of nuclear accent gives an indication of the speaker's intentions; a question that has a **L** nuclear accent has a different meaning than a question with a **H** accent, and will require a different response from a dialogue system.

Pitch range and prominence are perhaps the highest level intonation effects. These effects

[1] For example Chomsky's Government-Binding Theory or Gazdar's Generalised Phrase Structure Theory (Sells, 1985).

could be used by a sophisticated system to judge speaker emotions or intentions.

Potentially, all the information that is contained within the phonological description of an utterance is of use to a speech recognition system. Some parts of the phonological description are governed by very high level linguistic factors and it is unlikely that any speech recognition system will be able to make use of such information in the near future. Other parts of the description will be of more practical, immediate benefit and could be incorporated into current language models without too much difficulty.

The Suitability of the RFC analysis system for Speech Recognition

The above discussion focused on how phonological descriptions of intonation could be of use in speech recognition. This section will examine the suitability of the automatic analysis system described here for this task.

It is important to realise that the phonological content of an utterance is determined by a variety of linguistic phenomena, hence different parts of the phonological description will be of use to different parts of a language model. It has already been said that most language models are relatively simple, and because of this, not all the information available in a phonological description will be of use to the recognition system.

The analysis system attempts to provide a full RFC and phonological description of an utterance, and gives equal weight to the provision of each aspect of intonation. In a practical recognition system, it may be the case that only one aspect of the phonological description is needed and most of automatically deduced information will be of little use. Hence the automatic analysis system will be exhibiting a poor spread of resources if considerable effort is spent on detecting intonation effects that are not needed by the language model. A more practical short-term solution would be to design a system for each part of the phonological description that is needed. This should result in more accurate analysis of the intonational phenomena that are of use to the language model.

Eventually recognition systems will be able to make use of all the aspects of intonation and the full power of the phonological description can be used. Until then it is probably wiser to design individual analysis algorithms for each aspect of intonation that is needed in the language model.

5.4 Concluding Remarks

This work proposed that existing phonetic models were insufficiently powerful to properly model all the effects of intonation. It was argued that those models which did aim to cover a wide range of intonational effects were not defined in a strict enough manner, and that those which were more strictly defined did not cover all the intonational effects of English. Thus a new phonetic model of intonation was designed with the intention of being able to cover all the intonational effects of English, and do so in a carefully constructed, fully-specified way.

The interplay between the theoretical side of the work and the formal, computer side was interesting. Knowing that the theory would have to be implemented on computer helped ensure that the theory was strictly defined and comprehensive; having a carefully designed theory made the computer implementation easier.

Although it is clear that more work needs to be done in completing the phonology-intermediate grammar, and certain parts of the intermediate-F_0 grammar, the work presented here is considered a positive step towards a fully formal, comprehensive phonetic model of English intonation.

Appendix A

Text of Speech Data

Two sets of sentences were used in the experiments.

Data Set A

Data set A consisted of 64 sentences designed and spoken by the author. These sentences were designed to cover all the pitch accents of English, with each pitch accent occurring in many different positions within the phrase. A variety of phrasing situations were used, ranging from simple one-phrase sentences to sentences containing lists and relative clauses.

The words in the sentences were made up of mainly voiced phonemes so as to make the F_0 contours as continuous as possible. This made the contours easier to analyse by eye and also had the advantage that F_0 tracking algorithms could perform more reliable analysis.

Many of the words were borrowed from the literature, so many sentences may look familiar. This was simply to avoid having to think up of lots of words containing only voiced phonemes.

The hand labelling of data set A produced a total of 164 pitch accents and 136 intonation phrases.

In data set A, British school terminology was used to ensure the utterance was spoken with the desired accent type.

In the following sentences, a "*" denotes the nuclear accent. If the nuclear accent is not a fall accent, the type of nuclear is given in brackets after the text.

1. Do you really *need to win everything? (*fall-rise*)

2. Do you really *need to win everything? (*low-rise*)
3. Do you *really need to win everything? (*fall-rise*)
4. Do you *really need to win everything? (*low-rise*)
5. Do you really *need to win everything you do? (*fall-rise*)
6. Do you really *need to win everything you do? (*low-rise*)
7. Do you really need to win *everything you do? (*fall-rise*)
8. Do you really need to win *everything you do? (*low-rise*)
9. Do you *really need to win everything you do? (*fall-rise*)
10. Do you *really enjoy playing the game? (*low-rise*)
11. Do you really *enjoy playing the game? (*low-rise*)
12. Do you *really enjoy playing all the games you have said? (*low-rise*)
13. The large window stays closed: the small window you can open.
14. The window stays closed, but you can open the door.
15. Should I open the window *wider?
16. Must you use so many large unusual *words when you argue? (*fall-rise*)
17. *Must you use so many large unusual words when you argue? (*low-rise*)
18. Was the weather dry in November, or was it rainy, as usual? (*fall-rise*)
19. Was the weather *really dry in November, or was it rainy, as usual?
20. Was the weather really dry in *November, or was it rainy, as usual? (*low-rise*)
21. Was the region's weather unusually dry in November, or was it rainy, as usual?
22. Was the region's weather unusually dry in November, or was it rainy, as usual? (*fall-rise*)
23. Blueberries, brambleberries and loganberries are my favorite pie filling.

24. Blueberries, brambleberries and loganberries are my favorite pie filling, but not all at once.

25. I'm *sure these are the ones.

26. I'm absolutely *sure these are the ones.

27. Are you *sure these are the ones, or is there some doubt? (*low-rise*)

28. Are you sure these are the *real ones, or is there some doubt? (*fall-rise*)

29. Are you sure these are the real ones, or is there some *doubt?

30. My *brother lives in Denver.

31. My brother lives in *Denver.

32. My brother, who lives in Denver, is an engineer.

33. My brother who lives in Denver, is an engineer.

34. I gave the money to *Andrew.

35. I *gave the money to Andrew.

36. I gave the money to *Andrew! (*surprise-redundancy*)

37. The formula may have *errors, but the idea is generally *true.

38. But you would have gone there anyway! (*surprise-redundancy*)

39. I really believe Ebenezer was a dealer in Magnesium.

40. Was Ebenezer a dealer in *Magnesium? (*fall-rise*)

41. Was Ebenezer a *dealer in Magnesium? (*fall-rise*)

42. Was *Ebenezer a dealer in Magnesium? (*fall-rise*)

43. There isn't *any money.(*surprise-redundancy*)

44. There isn't *any money? (*high-fall*)

45. There isn't any *money!?

46. Was the weather *dry in November?

47. Was the weather dry in *November?

48. *Was the weather dry in November? (*low-rise*)

49. Was the weather unusually dry in November?

50. Aluminium is higher up the table than *magnesium?

51. *Aluminium is higher up the table than *magnesium?

52. I really believe Ebenezer was a dealer in Magnesium.

53. Was Ebenezer a dealer in *Magnesium?

54. Was Ebenezer a *dealer in Magnesium?

55. Was *Ebenezer a dealer in Magnesium?

56. There isn't any *money.

57. There isn't any *money? (*high-rise*)

58. There isn't any *money!? (*fall-rise*)

59. Was the weather *dry in November?

60. Was the weather dry in *November?

61. *Was the weather dry in November?

62. Was the weather unusually dry in November?

63. Aluminium is higher up the table than *magnesium?

64. *Aluminium is higher up the table than *magnesium? (*low-rise*)

Data Set B

The text for data set B was taken from the UNIX news system, which is an electronic bulletin board allowing open communication between UNIX users. A number of new groups exist that focus on particular areas such as a sport or hobby. The messages are usually informal and often humorous.

Forty-five messages, or parts of messages were spoken by a speaker who was familiar with the terminology of the text, but was linguistically naive. The speaker was instructed to speak these messages expressively, with the intention to capture the mood in which the message was written. A laryngograph was used to measure the F_0 of the speaker.

This data was used as it had many interesting intonational effects. More importantly, as I had no control over the design and recording of the data, it was intended to be a more open test of the model.

When the utterances were labelled by hand, a total of 301 pitch accents and 156 intonation phrases were marked.

1. Looks to me as though your mind rot has already set in.
2. I'm being bothered by an inconsistency that I'm hoping someone with a longer memory than mine, or better references, can resolve.
3. The ones I once owned went to my niece a couple of years ago.
4. Besides, who's going to refuse to do anything to a guy who's holding a uzi.
5. Boy, doesn't that make you feel like this is an honest man.
6. I would vote that the lack of upsets is an upset.
7. But I still wish to fight it out with Arijt.
8. Agreed but I bet that the reason for that is we would not be able to beat either Patrick McEnroe or Shapers even if we were to choose the surface, the spectators (only my family members and other supporters etc).
9. Heck, we might not be able to beat either of them in Paraguay even with the crowd behind us!

10. That would be like Bush never having any summit meetings with Gorbachev (he won't have any more anyway, looks like it!)

11. I mean Agassi can always crib that he is the moral winner, if he wins (hopefully), cause he "owns" Becker, right?

12. I bet if he had gotten that one win he would have been back, and you, Arijt, would (might) have written the same piece substituting Borg for Connors.

13. I wanted to write a whole lot, but I am too busy.

14. Basically, my forecast is based on my whimsical conclusion that the maximum number of matches that Haarhuis can last, on his top class player beating spree, is two.

15. One might be tempted to ask Connor's opponents for a second opinion.

16. You can't be serious - I've eaten more rabbits than you've had hot dinners!

17. That's not a bug, it's a feature!

18. Nah mate - real catchers play darts.

19. Does just what I always wanted DOS to do.

20. As far as over-generalisation goes, perhaps English teachers are more often the target than the perpetrator.

21. Can you imagine what this group would become if everybody just posted surf reports daily for his spot worldwide?! Especially mine would be quite distressing - flat, flat, flat.

22. Pity - if it was you'd probably be less distressed.

23. Oops sorry! I forgot everyone in netland needs the electronic laugh-track to understand humour.

24. Well, lots of university students, obviously.

25. I found the book interesting from the point of view of a skinny guy who didn't get bigger.

26. Feel free to differ.

27. Also mentioned was Gundes Fann, probably not competing; and Marilio Desultz, probably retiring after these games - is 41 too old?

28. My parents' shrinks always warned me about the kids of school teachers - especially the public school variety.

29. And it's cheap too.

30. You have a pleco in your tank and don't understand why there's no algae?

31. Same thing is true of the international morse code, the flag hoists on sea-going shapes and the Dvorzak simplified keyboard (refers typewriters, not Beckstein pianos)

32. They never did catch us!

33. Has nothing to do with dietary laws!

34. Nice business, maybe I should start watching for the next seat on the NYSE or ASE that comes up for sale.

35. But all the engineers I've met seem to fit this stereotype.

36. If I were interviewing you, I'd be curious about why you waited five months to leave!

37. But developing a new system as exciting as the Macintosh (or even the PC!) is not the same thing as writing COBOL.

38. Further, the entire roof lifts off the greenhouse to simulate a convertible, but there's no place to store the removed roof so it's no more practical than a Mercedes.

39. I think that all Americans are weirdos that will (at the drop of a hat) change into Hawiian shirts and kilts.

40. Oh goodness, I just used that as a complementizer.

41. I've been asked by some friends (who are members of a parent-teachers association) to get some educational software suitable for primary school kids aged five up to twelve, and games (suitable for the teachers!)

42. When I bought my IBM (with departmental funds), the technicians who came to start it up, always referred to that thing as a SCSI.

43. But that topic is far removed from free or inexpensive access to the internet. Hey, I provide inexpensive internet access and I assure you, it isn't illegal!

44. Hey, its only money!!

45. For another 10K a year I'll let you sleep in my spare room on cheap furniture and let you play loud bad music all night!

Appendix B

Labelled F0 Contours

This appendix contains examples of F_0 contours from the two data sets. All the figures show three graphs relating to the original F0 contour, the hand labelled transcription and the automatically labelled transcription.

All three graphs show the original (smoothed) F_0 contour; graphs (b) and (c) also show synthesized contours (in bold).

(a). The original F0 contour in isolation.

(b). This graph shows the hand-labelling for the utterance.

The graph is divided into sections which are delimited by the dotted lines. In the box immediately below the graph is the RFC element for each section. The box below this gives the phonological transcription.

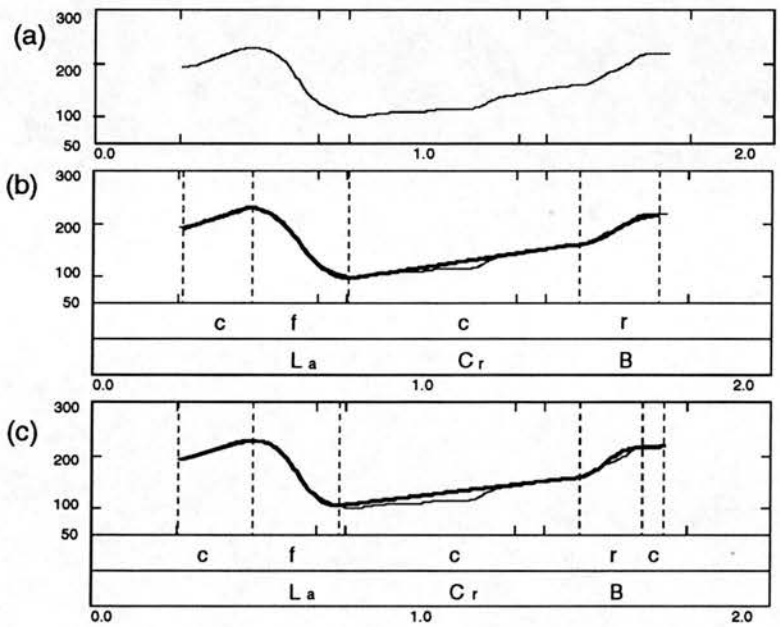
The RFC labels are either "r" "f" or "c". Occasionally, when there is insufficient room, the "c" labels are omitted. Any section without a label is a connection element.

The phonological labels are "H" "L" "C" and "B". The subscripts "d", "l", "e", "a" and "r" are used to denote the features.

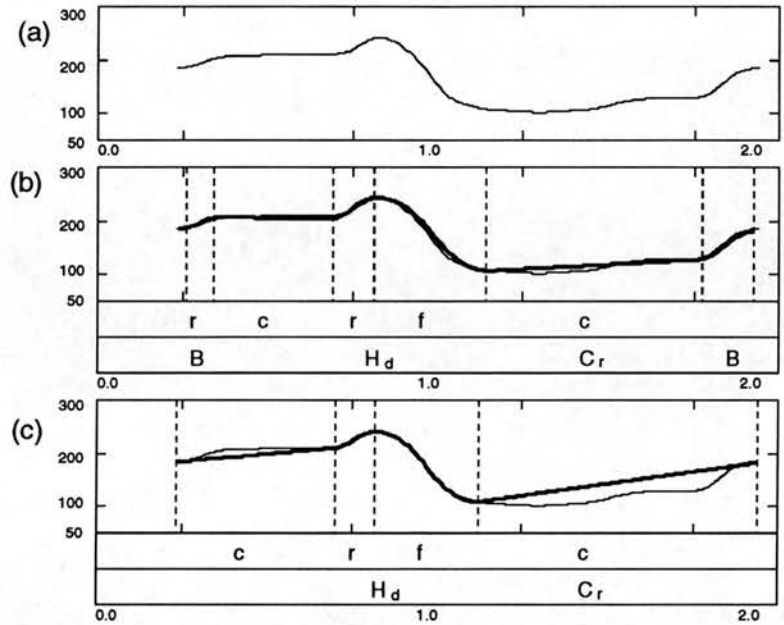
The thick black line represents the synthesized F_0 that is produced when the hand-labelled RFC parameters are used as inputs to the IF synthesis program. By comparing the synthesized F_0 and the original, one can judge the accuracy of the synthesis mapping.

(c). This is the automatically labelled version. As with graph (b), the dotted lines denote boundaries and the two boxes represent the RFC and phonological labels.

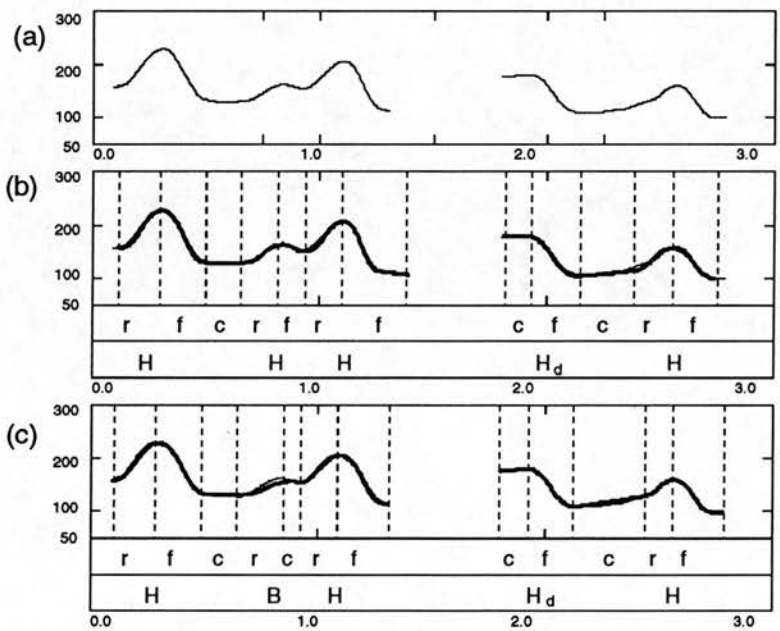
All graphs are scaled between 50 and 300Hz on the y-axis. Depending on the length of the utterance, the x-axis varies from 2.0 seconds to 5.0 seconds.



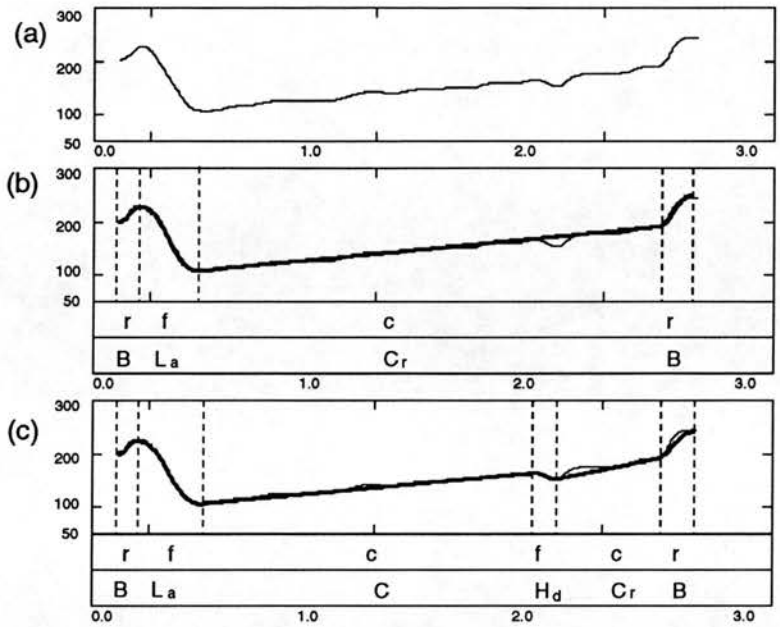
Utterance A.4 *Do you *really need to win everything.*
This phrase has a L_a *nuclear accent. The automatic system produces a similar RFC description to the hand-labelled version.*



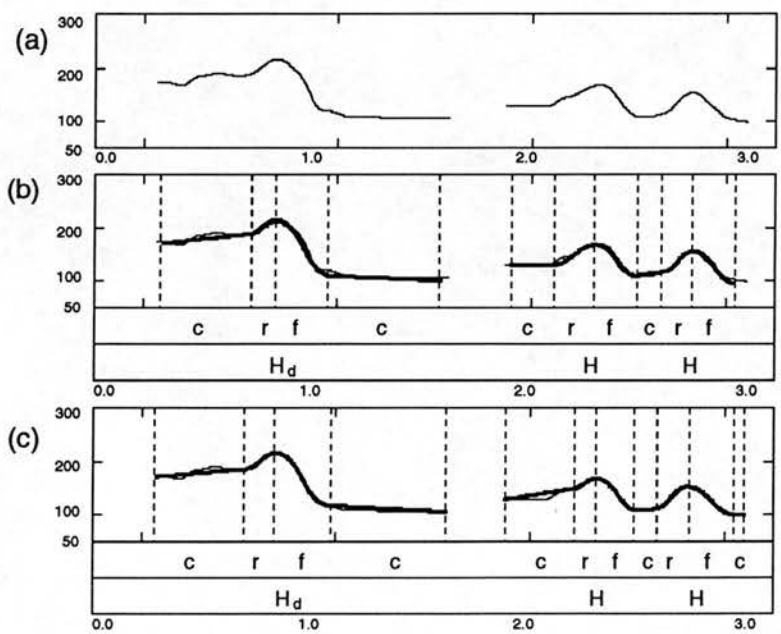
Utterance A.5 *"Do you really *need to win everything you do"*
*An example of a H_d C_r **B** accent which is described as a fall-rise in the British school. The automatic version has not placed a boundary element at the end of the phrase. Even so, the phonological classification is that of a fall-rise accent as the boundary element is not mandatory.*



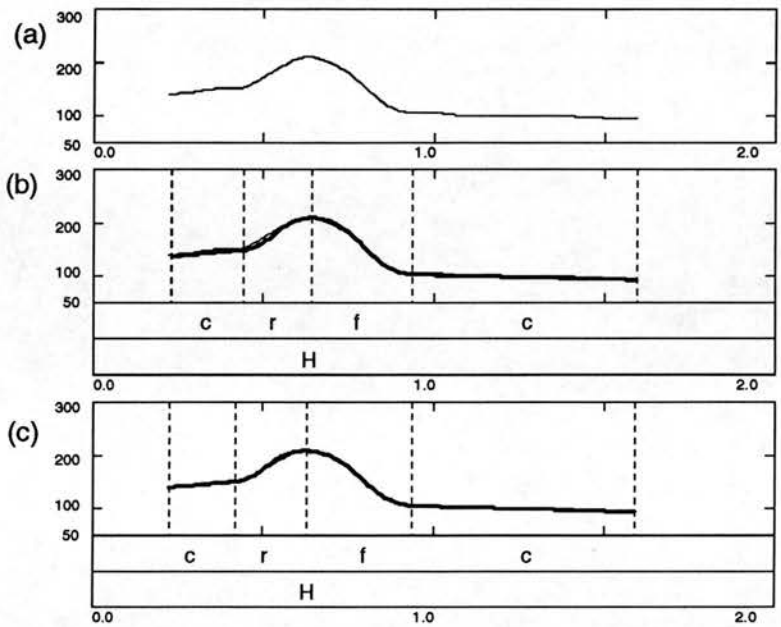
Utterance A.13 *"The large window stays closed: the small window you can open."*
*In this example we can see that the synthesized curve fits the original very well, and that the automatically labelled RFC description is very close to the hand labelled version. The only error is that a fall has been missed in the second accent of the first phrase. This results in that element receiving a phonological class of **B**, which is an error.*



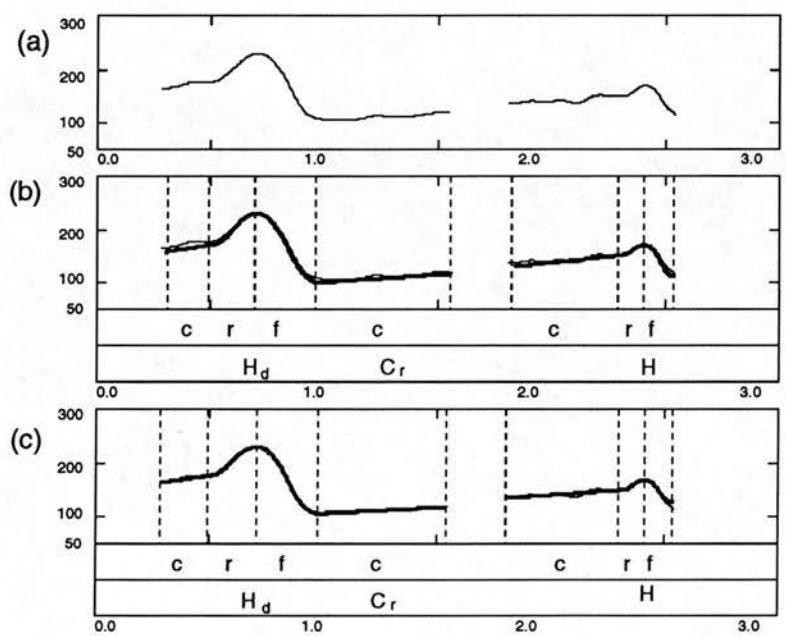
Utterance A.17 *"*Must you use so many large unusual words when you argue"*
This is an unusual example of a L_a accent in that the nucleus is very near the start of the phrase. This contour clearly demonstrates the suitability of using long straight connection elements in certain situations. The automatic system has mistakenly used a fall element to model a large segmental glitch.



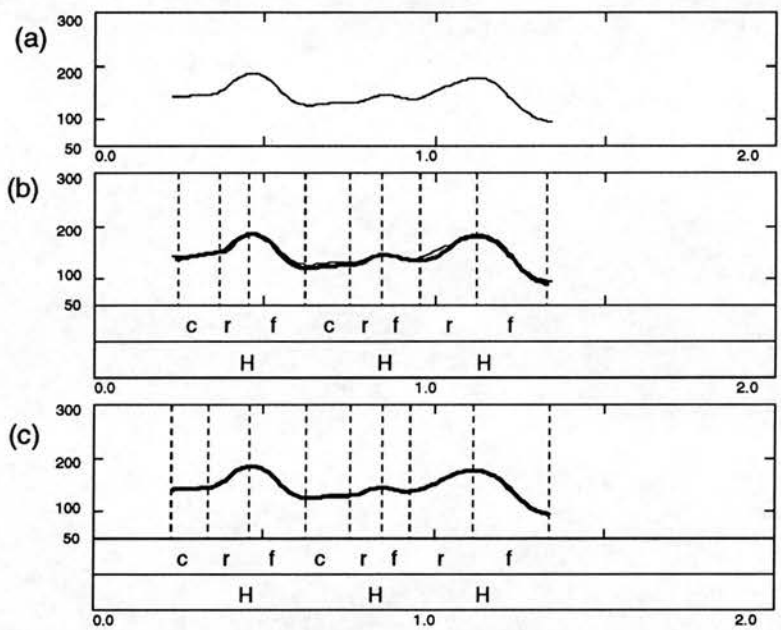
Utterance A.20 *"Was the weather *really dry in November, or was it rainy, as usual?"*
Here the hand and automatic versions correspond nearly exactly. The segmental bump early in the first accent of the second phrase confused the automatic system somewhat, causing it to start the rise element later than in the hand version.



Utterance A.24 *"I'm *sure these are the ones"*
Very close fit between original and synthesised contours, and very close agreement between automatic and hand labels.

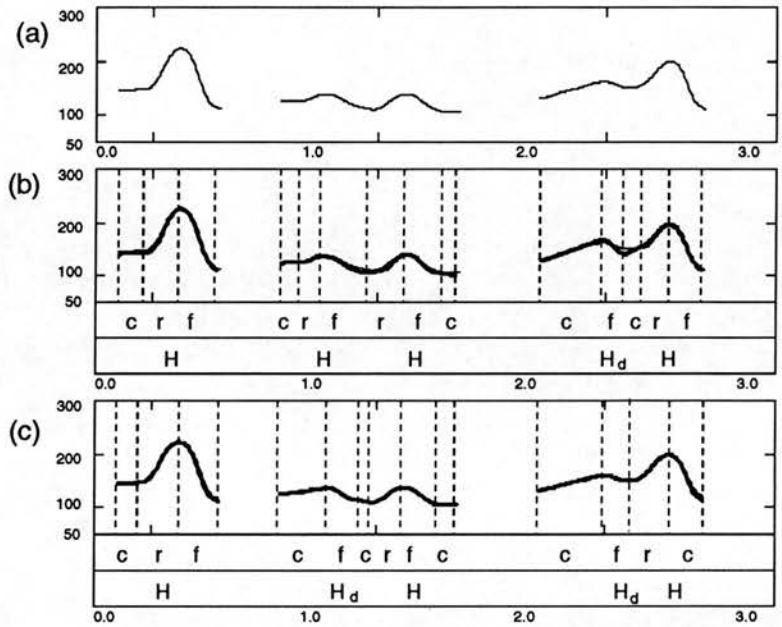


Utterance A.27 *"Are you *sure these are the real ones, or is there some doubt?"*
The two transcriptions agree very well and the synthesized contours also fit the original well. The nuclear accent in the first phrase is an example of the uncommitted sounding $H_d\ C_r$ *accent that was discussed in section 3.4.6.*

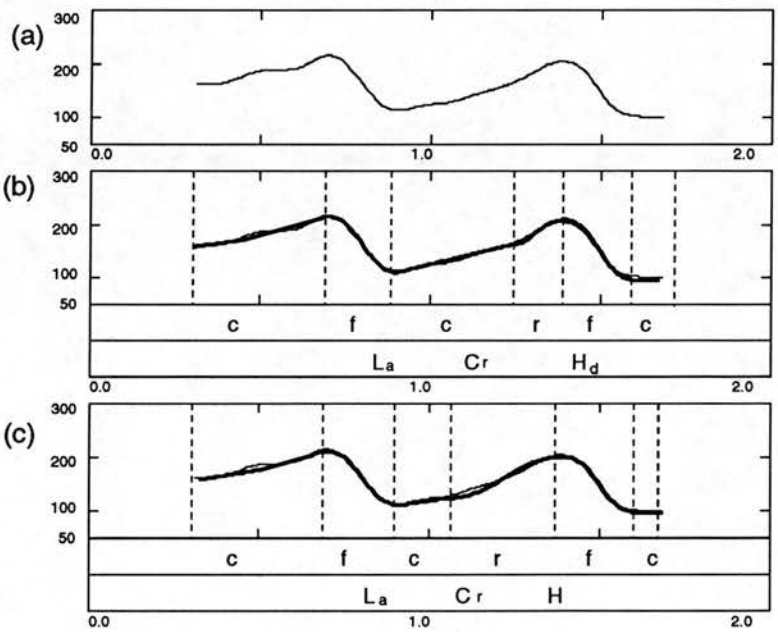


Utterance A.30 *"My brother lives in *Denver."*

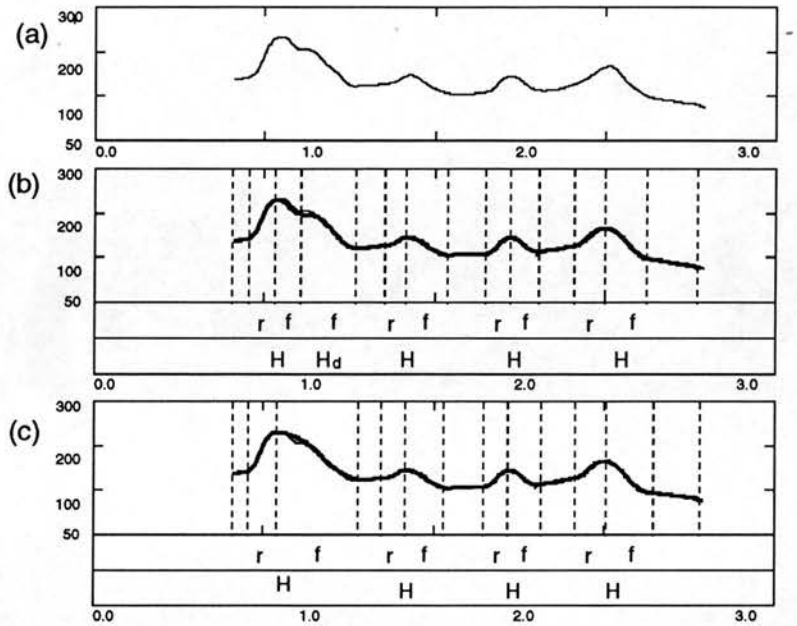
Again good agreement between the automatic and hand transcribed versions. It is interesting to note that the contour created from synthesizing the automatic transcription is closer to the original than the contour synthesized from the hand transcription. This is due to the automatic optimal matching system trying hundreds of slightly different shapes to find the closest fit. In the hand labelling, the number of shapes tried depends on the labeller. By trying more shapes a better fit could be obtained, but this is a lengthy process for a human labeller.



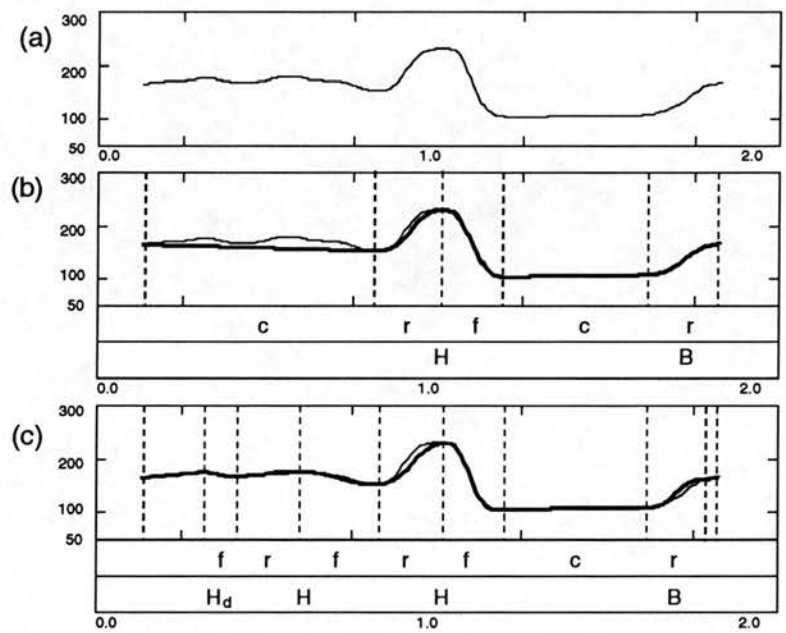
Utterance A.32 *"My brother who lives in Denver, is an engineer."*
The two RFC descriptions do not differ greatly, but the rise deletion error in the second phrase produces a downstepped BF H accent. This contour shows how the analysis assessment method is insensitive to the relative importance of accents. The method gives equal weight to the labelling of the first accent and second accents in the utterance, even though they are substantially different in size.



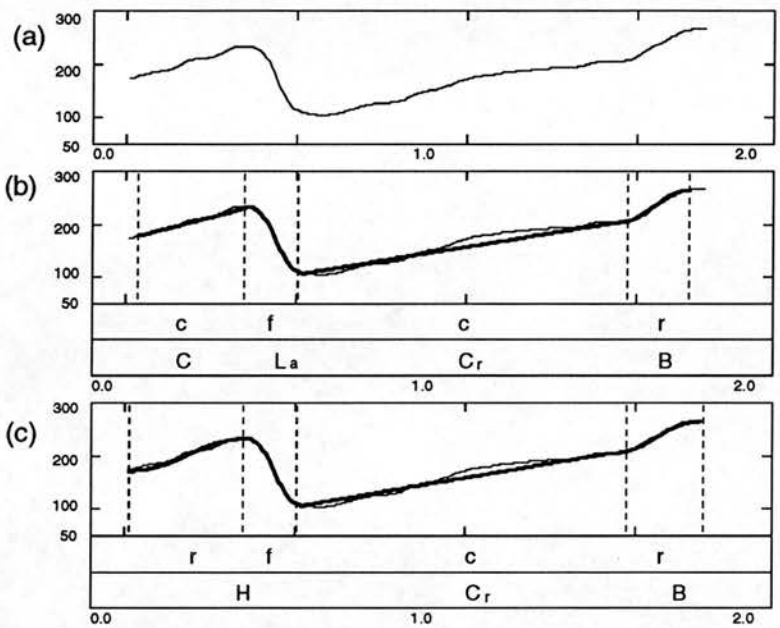
Utterance A.38 *"But you would have gone there anyway!"*
This is an example of the surprise-redundancy contour. The two transcriptions are similar in symbolic terms, but the automatic system clearly puts the start boundary for the last rise element considerably before the equivalent boundary in the hand labelled version.



Utterance A.39 *"I really believe Ebenezer was a dealer in Magnesium."*
The automatic system interprets the two downstepping sequence of the two **H** *accents as a single* **H** *accent. This is the consequence of the assimilation module grouping two nearby fall elements together.*

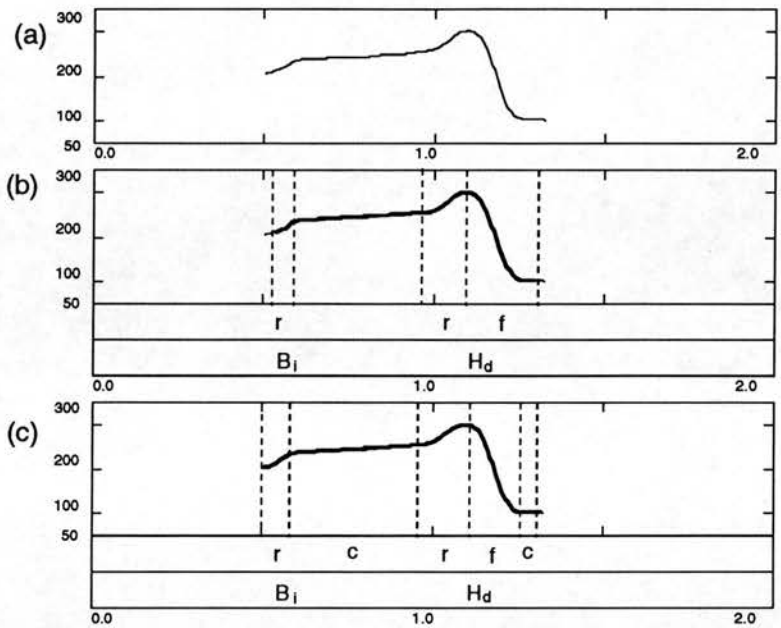


Utterance A.41 *"Was Ebenezer a *dealer in Magnesium?"*
The automatic system misinterprets the segmental bumps at the start of the phrase and labels with rise and fall elements. When hand labelling, it was difficult to tell if these bumps arose from segmental influence or indicated the prescence of a pitch accent, it was only from listening to the utterance that the decision could be made.

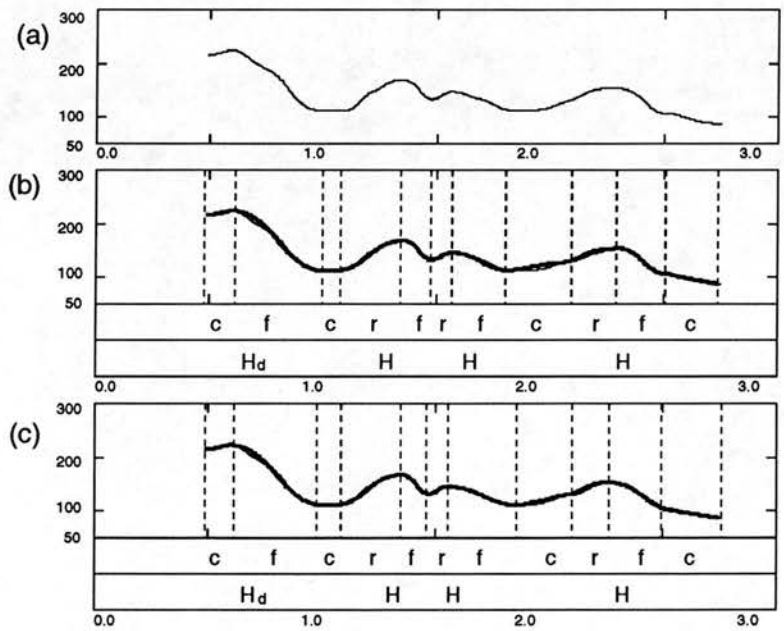


Utterance A.42 *"Was *Ebenezer a dealer in Magnesium?"*

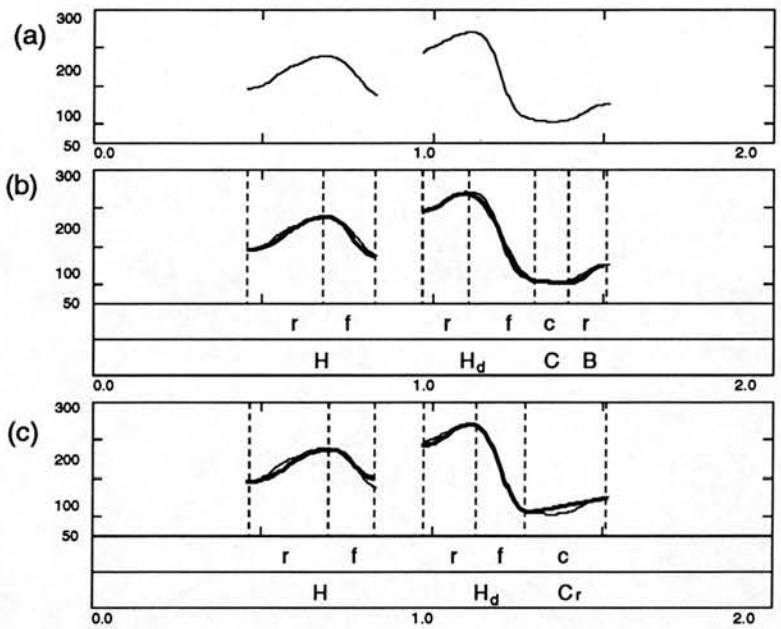
The sharply rising section of F_0 contour at the start of the utterance is mislabelled as a rise element. This in turn means that the phonological classification system labels the accent with **H** *instead of* **L**. *Timing information is only used when a fall element occurs in isolation. If it had been available, it would have been clear that the fall occurred in a very early position which would have indicated a* **L** *accent. It is also interesting to note that the duration of the rise section is significantly longer than usual. In a more sophisticated labelling system, this unusualness could have been used to show that incorrect labelling might have occurred.*



Utterance A.45 *"There isn't any money!?"*
This is a good example of a nuclear accent with a fall very much greater than the rise. Such an accent would be difficult to account for in Fujisaki's system. The hand and automatic systems agree nearly exactly in this case.

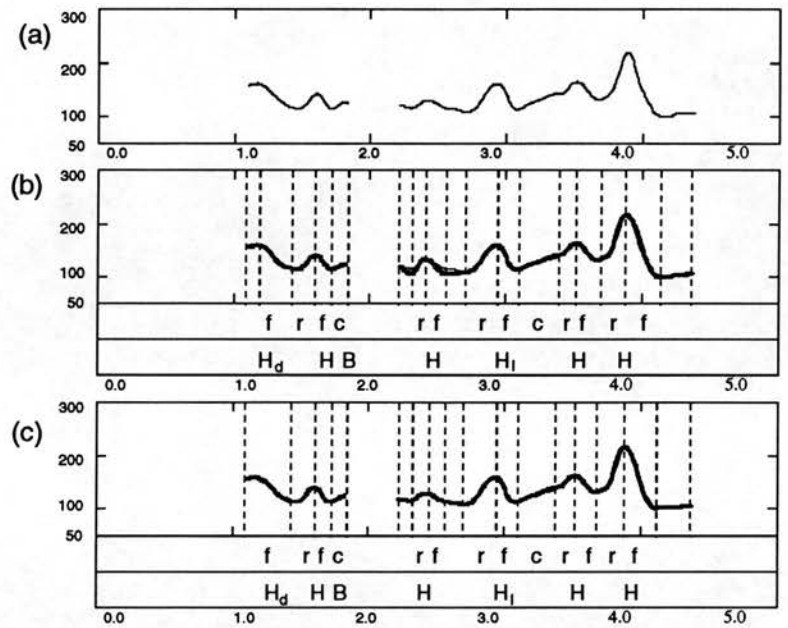


Utterance A.50 *"Aluminium is higher up the table than *magnesium?"*

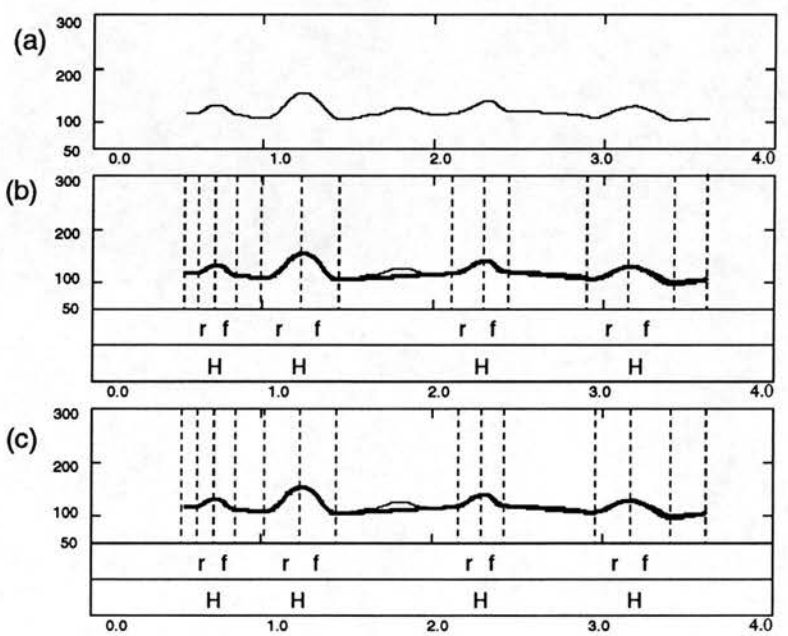


Utterance A.58 *"There isn't any money!?"*

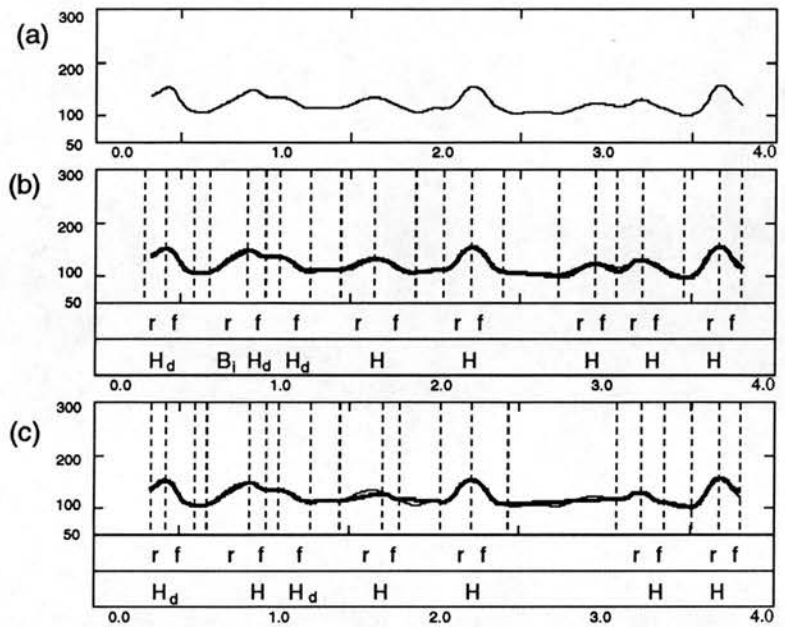
The automatic transcription differs from the hand transcription at the end of the utterance. Even though the final boundary element is missed, the two phonological classifications fit into the British school category of "fall-rise".



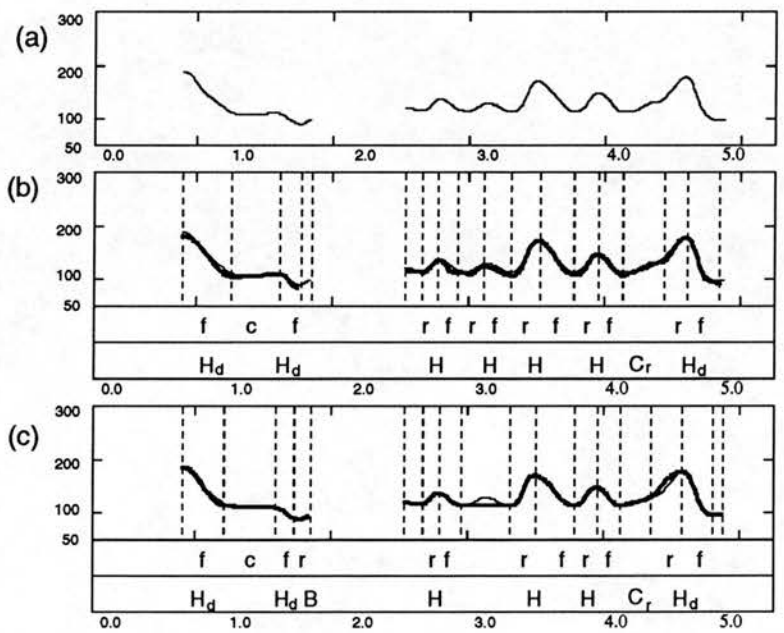
Utterance B.1 *"Looks to me as though your mind rot has already set in."*
Note that the rise section of the H_l *is clearly longer than the fall section.*



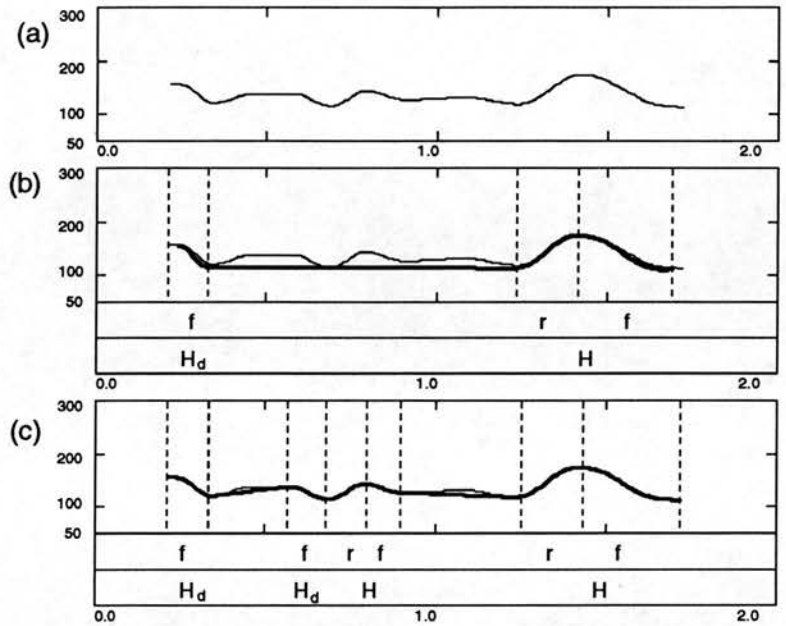
Utterance B.3 *"The ones I once owned went to my niece a couple of years ago."*
The segmental bump between the second and third **H** *accents could easily be mistaken for a pitch accent. Its rates of rise and fall were just below the trained threshold levels.*



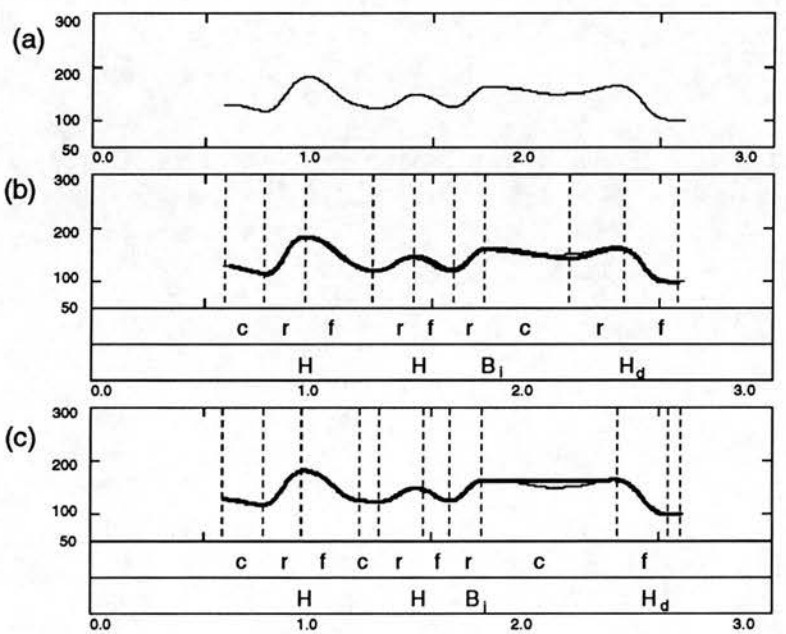
Utterance B.4 *"Besides, who's going to refuse to do anything to a guy who's holding a uzi."*
Normally rise fall sequences are marked ***H****, but when the rise occurs phrase-finally, and the fall is the first element in the next phrase, the correct phonological description is* $\boldsymbol{B_i\ H_d}$*. The automatic system did not use this rule, as there is no way to distinguish rise elements in accents from rise elements occurring at boundaries.*



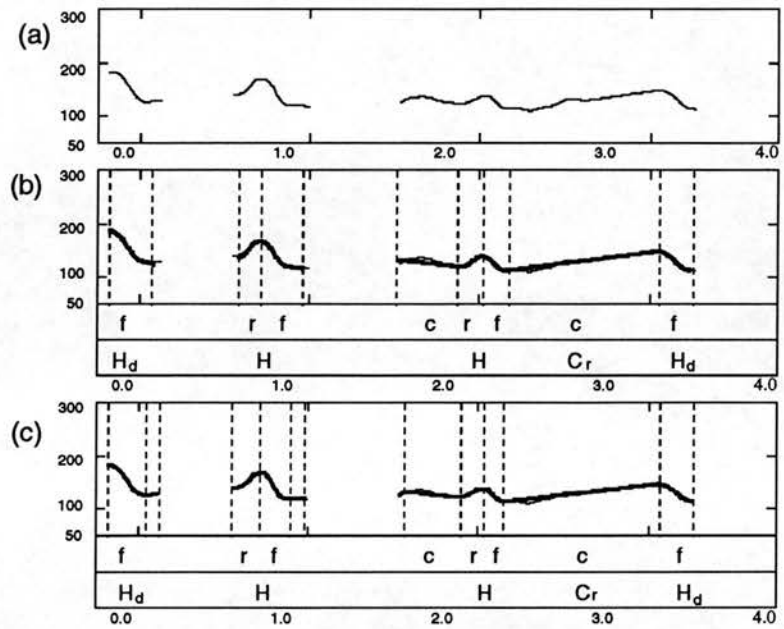
Utterance B.16 *"You can't be serious - I've eaten more rabbits than you've had hot dinners!"*
A minor pitch accent is missed in the automatic transcription of the second phrase. This utterance contains a rising connection element in a pre-nuclear position which is typical of some whimsical utterances.



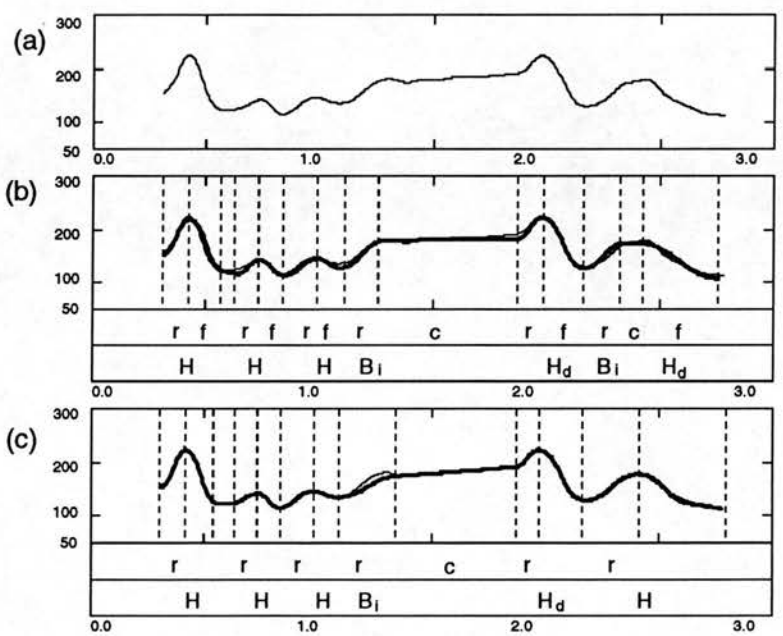
Utterance B.17 *"That's not a bug, it's a feature!"*
This utterance clearly demonstrates the problem of segmental influence. There are only two accents in the phrase, but the automatic system marks four.



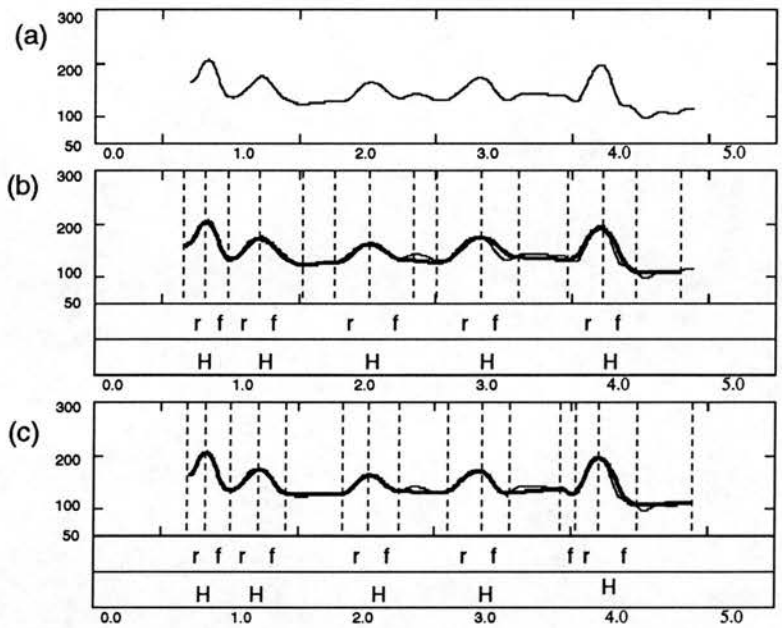
Utterance B.19 *"Does just what I always wanted DOS to do."*
This is a good example of where the two phonological transcriptions are similar, but the RFC transcriptions vary.



Utterance B.22 *"Pity - if it was you'd probably be less distressed."*
The two transcriptions are very similar.



Utterance B.24 *"Well, lots of university students, obviously."*
The word "obviously" is in a tagged phrase. The automatic system does not mark this as such because it finds it can fit the rise and fall elements very well without the need for an intervening connection element.



Utterance B.25 *"Oops sorry! I forgot everyone in netland needs the electronic laugh-track to understand humour."*

Another example of the synthesized contour from the automatic transcription being closest to the original contour, specifically in the fourth accent.

Appendix C

Mathematical Derivation of the Monomial Function

Section 3.3.2 gave the equation for the monomial function. Below is the full derivation of this function.

We are trying to derive an equation defined in the region between $(0, 0)$ and $(1, 1)$ that has a gradient of zero at $(1, 1)$ and $(0, 1)$, passes through the point $(1/2, 1/2)$ and has a variable degree of curvature.

$$y = (-x)^{\gamma}$$

gives a family of equations in the region $(x < 0)$ which all have gradient zero at the origin. Rotating this family through 180 degress and translating by $(-2, 2)$ gives:-

$$y = 2 - (2 + x)^{\gamma}$$

These two equations combined can give the required shape, defined in the area $(-2, 2)$ to $(0, 0)$.

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

Translated into the positive quadrant, the equations becomes

$$y = (2 - x)^{\gamma} \qquad 1 < x < 2$$
$$y = 2 - (2 + (x - 2))^{\gamma} \qquad 0 < x < 1$$

Scaled so as to lie in the region betwen (0, 0) and (1, 1), the equation becomes

$$y = \tfrac{1}{2}.2^{\gamma}(1 - x)^{\gamma} \qquad \tfrac{1}{2} < x < 1$$
$$y = 1 - \tfrac{1}{2}.2^{\gamma}(x)^{\gamma} \qquad 0 < x < \tfrac{1}{2}$$

Substituting the constant C for $2^{\gamma-1}$ gives

$$y = C(1 - x)^{\gamma} \qquad \tfrac{1}{2} < x < 1$$
$$y = 1 - C(x)^{\gamma} \qquad 0 < x < \tfrac{1}{2}$$

The variables are changed to F_0 and t and scaling is allowed in the x and y directions, denoted by A and D, giving the final equation.

$$F_0 = AC(1 - t/D)^{\gamma} \qquad D/2 < t < D$$
$$F_0 = 1 - C(t)^{\gamma} \qquad 0 < t < D/2$$

Appendix D

Computer Implementation Details

All the algorithms were implemented in the ANSI C programming language. The programs were developed on a UNIX workstation, but due to the portability of the C language, these programs should work on other hardware platforms.

The RFC analysis system consisted of approximately 5000 lines of C code, the synthesis system 1000 lines and the phonological analysis system added another 1000 lines.

As far as practical use is concerned, every part of the system worked fast, and there should be no difficulty in incorporating these programs into real-time systems. The only part of the system which took any appreciable time was the optimal matching module of the analysis system. This was time-consuming as approximately 400 monomial shapes had to be calculated for every element, and each of these shapes had to have the distance between itself and the F_0 contour measured.

Although this part of the analysis process was slightly time consuming, this should not present a major problem for a sufficiently powerful computer. If speed is critical, the number of possible shapes which are calculated for the matching process could be reduced. This may provide sub-optimal matches, but the overall performance should not deteriorate greatly.

Appendix E

Published Work

During the course of the work on this these, five conference papers were published by me. These were:-

(a). Taylor, P. A. and Isard, S. D., (1990). Automatic Diphone Segmentation using Hidden Markov Models. *In: Proc. 3rd International Australian Conference on Speech Science and Technology, SST-90.*

(b). Taylor, P. A. and Isard, S. D., (1991). Automatic Diphone Segmentation. *In: Proc. Eurospeech '91, Genova, Italy.*

(c). Taylor, P. A., Nairn, I. A., Watson, G. W., Sutherland, A. M. and Jack M. A., (1991). An Interactive Speech Generation System. IEE Colloquium on *Systems and Applications of Man-Machine Interaction using Speech I/O*. Digest no. 1991/066.

(d). Taylor, P. A., Nairn, I. A., Sutherland, A. M. and Jack, M. A., (1991). A Real Time Speech Synthesis System. *In: Proc. Eurospeech '91, Genova, Italy.*

(e). Taylor, P. A. and Isard, S. D., (1992) A New Model of Intonation for use with Speech Synthesis and Recognition. *In: International Conference of Speech and Language Processing, Banff, Canada.*

Only the last paper describes work from this thesis and is included here for reference.

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

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

Abstract

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

1 Introduction

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

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

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

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

1.1 An Analysis/Synthesis System for Intonation

In our system, five levels of representation were used.

1. **Acoustic Waveform**
2. **F_0 Contour**
   Fundamental frequency (F_0) is most often measured from digitally sampled waveforms using automatic analysis algorithms.
3. **Phonetic Description**
   The phonetic description is a description of the contour as a sequence of discrete units.
4. **Normalised Phonetic Description**
   The normalised phonetic description is a discrete description which has been normalised to filter out phonemic effects. Phonemic effect include differences in F_0 depending on vowel type, and differences in accent position depending on syllable structure [1].
5. **Phonological Description**
   The phonological description is a qualitative abstract description of the intonation of an utterance.

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

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

Four main systems were planned:-

1. **Signal Processing**
   Extracting F_0 from the acoustic or laryngograph waveform.
2. **Contour Analysis**
   Extracting the phonetic description. from the F_0 contour
3. **Normalisation**
   Filtering out the phonemic effects so as the contour is normalised with respect to phoneme information.
4. **Phonological Abstraction**
   Deriving the phonological description from the phonetic one.

1.2 General Aims

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

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

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

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

2 Signal Processing: F_0 Extraction

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

3 Phonetic Descriptions and F_0 Analysis

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

3.1 The Dutch Model

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

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

3.2 The Fujisaki Model

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

3.3 A New Phonetic Description

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

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

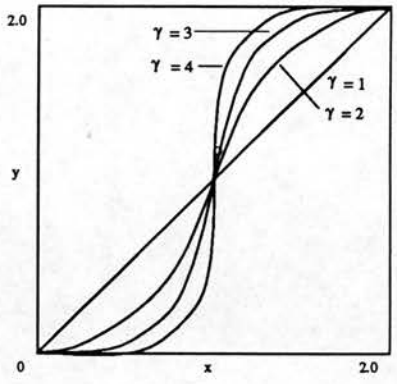


Figure 1: The polynomial functions used for rise and fall shapes. As γ increases the curve becomes less like a straight line and more curved.

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

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

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

Usage

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

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

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

Performance

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

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

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

4 Automatic F_0 Labelling

An automatic labelling system was built to label F_0 contours.

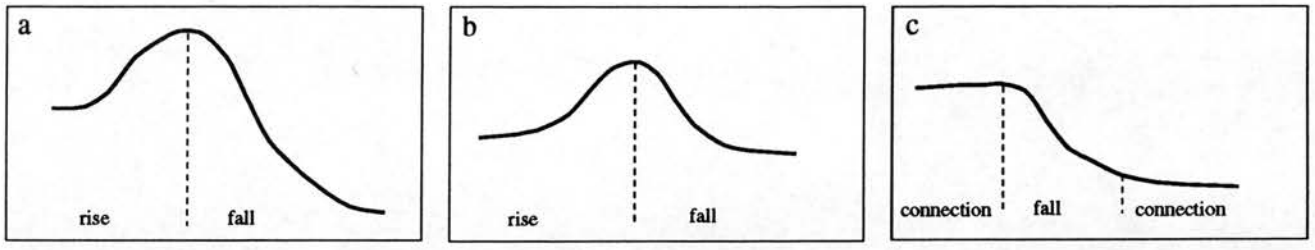


Figure 2: Three typical fall accents shown with different rise and fall combinations. (a) is a typical nuclear fall, (b) is likely to be found in a prenuclear position and (c) is often seen in sequences of downstepping accents.

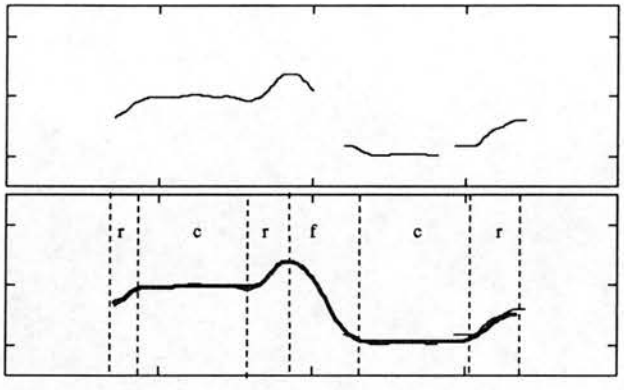


Figure 3: The top graph shows the F_0 contour, the bottom graph shows the same F_0 contour but with the reconstructed F_0 contour shown superimposed in bold. The utterance is "Do you *need to win everything?" (* denotes accented word).

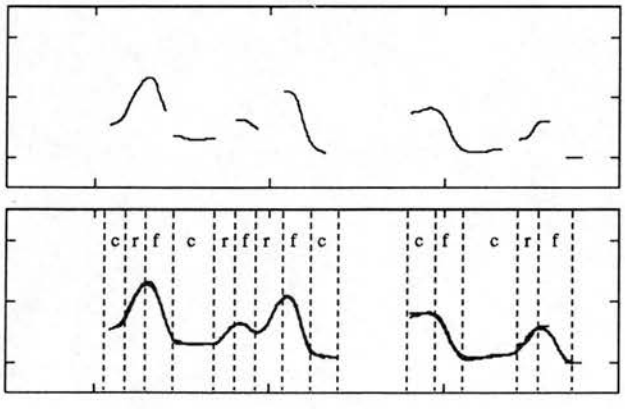


Figure 4: The top graph shows the F_0 contour, the bottom graph shows the same F_0 contour but with the reconstructed F_0 contour shown superimposed in bold. The utterance is "The large window stays closed, the small one you can open"

4.1 Contour Preparation

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

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

4.2 Broad Class Frame Labelling

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

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

4.3 Optimal Matching

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

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

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

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

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

4.4 F_0 Analysis: Assessment

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

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

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

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

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

5 Phonemic Normalisation and Phonological Abstraction

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

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

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

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

6 Conclusion

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

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

References

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

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

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

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

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

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

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

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

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

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

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

Bibliography

Abberton, E. R. M., Howard, D. M., and Fourcin, A. J. (1989). Laryngographic assessment of normal voice: A tutorial. *Clinical Linguistics and Phonetics*, 3(3):281–296.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

Allen, J., Hunnicut, S., and Klatt, D. (1987). *From Text to Speech: the MITalk System*. Cambridge University Press.

Bachenko, J. and Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.

Beckman, M. E. and Pierrehumbert, J. B. (1986). Intonational atructure in Japanese and English. *Phonology Yearbook 3*, pages 255–309.

Bickmore, L. (1990). Branching nodes and prosodic categories. In Inkelas, S. and Zec, D., editors, *The Phonology-Syntax Connection*. The University of Chicago Press.

Bolinger, D. (1951). Intonation: Levels versus configurations. *Word*, 14:109–149.

Buser, P. and Imbert, M. (1987). *Audition*. Hermann.

Campbell, W. N. and Isard, S. D. (1991). Segmental durations in a syllable frame. *Journal of Phonetics*, 19:37–47.

Chen, M. Y. (1990). What must phonology know about syntax. In Inkelas, S. and Zec, D., editors, *The Phonology-Syntax Connection*. The University of Chicago Press.

Cheng, Y. M. and O'Shaughnessy, D. (1989). Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Signal Processing*, 37:1805–1814.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT press.

Chomsky, N. (1971). *Chomsky: Selected Readings.* Oxford University Press.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* Harper and Row.

Clark, J. and Yallop, C. (1990). *An Introduction to Phonetics and Phonology.* Basil Blackwell.

Cooper, W. E. and Sorensen, J. M. (1981). *Fundamental Frequency in Sentence Production.* Springer-Verlag.

Crystal, D. (1969). *Prosodic Systems and Intonation in English.* Cambridge Studies in Linguistics. Cambridge Universtity Press.

Crystal, D. (1972). The intonation system of English. In Bolinger, D., editor, *Intonation.* Penguin.

Crystal, T. H. and House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America,* 83(4).

Fujisaki, H. (1992). The role of quantitative modeling in the study of intonation. In *Symposium on Research on Japanese and its Pedagogical Applications, November 4-7, Nava, Japan.*

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

Fujisaki, H. and Kawai, H. (1988). Realization of linguistic information in the voice fundamental frquency contour of the spoken Japanese. In *International Conference on Speech and Signal Processing.* IEEE.

Gartenberg, R. and Panzlaff-Reuter, C. (1991). Production and perception of F_0 peak patterns in german. In Kohler, K. J., editor, *Studies in German Intonation.* Universitat Kiel.

Gussenhoven, C. and Rietveld, T. (1988). Fundamental frequency declination in Dutch: Testing three hypotheses. *Journal of Phonetics.*

Halliday, M. A. K. (1967). *Intonation and Grammar in British English.* Mouton.

Hermes, D. J. and van Gestel, J. C. (1991). The frequency scaling of speech intonation. *Journal of the Acoustical Society of America.*

Hieronymus, J. L. (1992). Use of acoustic sentence level and lexical stress in hsmm speech recognition. In *International Conference on Speech and Signal Processing.* IEEE.

Isard, S. D. and Pearson, M. (1988). A repertoire of British English contours for speech synthesis. In *SPEECH '88, 7th FASE Symposium.* FASE.

Jakobson, R., Fant, G. M. C., and Halle, M. (1952). *Preliminaries to Speech Analysis: the Distinctive Features and their correlates.* MIT press.

Jones, D. (1957). *An Outline of English Phonetics.* Cambridge: Heffer & Sons, 8 edition.

Klatt, D. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3:129–140.

Kohler, K. J. (1991a). A model of german inotnation. In Kohler, K. J., editor, *Studies in German Intonation.* Universitat Kiel.

Kohler, K. J. (1991b). The perception of accents: Peak height versus peak position. In Kohler, K. J., editor, *Studies in German Intonation.* Universitat Kiel.

Kohler, K. J. (1991c). Terminal intonation patterns in aingle-accent utterances of German: Phonetics, phonology and semantics. In Kohler, K. J., editor, *Studies in German Intonation.* Universitat Kiel.

Ladd, D. R. (1983a). Levels vs. configurations, revisited. In Agard, F. B., Kelly, G., Makkai, A., and Makkai, V. B., editors, *Essays in Honor of Charles F. Hockett.* Leiden: E. J. Brill.

Ladd, D. R. (1983b). Phonological features of intonation peaks. *Language*, 59:721–759.

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonlogy Yearbook 1.*

Ladd, D. R. (1986). Intonational phrasing: the case for recursive prosodic structure. *Phonlogy Yearbook 3*, pages 311–340.

Ladd, D. R. (1987). A model of intonational phonology for use with speech synthesis by rule. In *European Conference on Speech Technology.* ESCA.

Ladd, D. R. (1988). Declination reset and the hierarchical organization of utterances. *Journal of the Acoustical Society of America*, 84(2):530–544.

Ladd, D. R. (1992a). Compound prosodic domains. *Edinburgh University Dept Linguistics Ocassional Paper*. Submitted for publication in *Language*.

Ladd, D. R. (1992b). Constraints on the gradient variability of pitch range. *Edinburgh University Dept Linguistics Ocassional Paper*. Submitted for publication in *Papers in Laboratory Phonology*.

Ladd, D. R. and Campbell, W. N. (1991). Theories of prosodic structure: Evidence from syllable duration. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, pages 290–293.

Lathi, B. P. (1983). *Modern Digital and Analog Communication Systems*. Holt-Saunders International Editions.

Lehiste, I. and Peterson, G. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, 31.

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

Liberman, M. Y. (1975). *The Intonational System of English*. PhD thesis, MIT. Published by Indiana University Linguistics Club.

Liberman, M. Y. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry 8*, pages 249–336.

Lieberman, P. (1967). *Intonation, Perception and Language*. M.I.T. Press.

Lindsay, P. H. and Norman, D. A. (1972). *An Introduction to Psychology*. Academic Press.

Longuet-Higgins, C. (1985). Tone of voice: The role of intonation in computer speech understanding. In Fallside, F. and Woods, W., editors, *Computer Speech Processing*. Prentice Hall International.

Medan, Y., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, 39:40–48.

Möbius, B., Demenko, G., and Pätzold, M. (1991a). Parametric description of German fundamnetal frequency contours. In *Proc. 12th International conference on Phonetic Sciences.*

Möbius, B., Demenko, G., and Pätzold, M. (1991b). Parametrische beschreibung von intonationskonturen. In Hess, W. and Sendlmeier, W. F., editors, *Beiträge zur angewandten und experimentellen Phonetik.* Stuttgart: Steiner.

Möbius, B. and Pätzold, M. (1992). F_0 synthesis based on a quantitative model of German intonation. In *International Conference on Speech and Language Processing '92.*

Morimoto, T. (1992). Continuous speech recognition using a combination of syntactic constraints and dependency relatioship. In *International Conference on Speech and Language Processing '92.*

O'Connor, J. D. and Arnold, G. F. (1973). *Intonation of Colloquial English.* Longman, 2 edition.

Öhman, S. (1967). Word and sentence intonation: A quantitative model. *STL-QPSR 2-3.*

Palmer, H. (1922). *English Intonation with Systematic Exercises.* Cambridge Universtity Press.

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

Pike, K. (1945). *The Intonation of American English.* University of Michigan Press.

Selkirk, E. O. (1984). *Phonology and Syntax.* MIT Press.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theories.* Centre for the Study of Language and Information.

Seneff, S., Meng, H., and Zue, V. (1992). Language modelling for recognition and understanding using layered bigrams. In *International Conference on Speech and Language Processing '92.*

Silverman, K. (1987). *The Sructure and Processing of Fundamental Frequency Contours*. PhD thesis, University of Cambridge.

Silverman, K. and Pierrehumbert, J. B. (1990). The timing of prenuclear high accents in English. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology 1*. Cambridge University Press.

Taylor, P. A. and Isard, S. D. (1990). Automatic diphone segmentation using hidden markov models. In *SST-90, Third International Australian Conference in Speech Science and Technology*.

t'Hart, J. and Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1:309–327.

t'Hart, J. and Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, 3:235–255.

Veilleux, N. M., Ostendorf, M., and Wightman, C. W. (1992). Parse scoring with prosodic information. In *International Conference on Speech and Language Processing '92*.

Vonwiller, J. P., King, R. W., Stevens, K., and Latimer, C. R. (1990). Comprehension of prosody in synthesized speech. In *SST-90, Third International Australian Conference in Speech Science and Technology*.

Waibel, A. (1986). *Prosody in Speech Recognition*. PhD thesis, C.M.U.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vacinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91:1707–1717.

Willems, N. J. (1983). A model of standard English intonation patterns. *IPO annual Progress Report*.

Woods, W. A. (1985). Language processing for speech understanding. In Fallside, F. and Woods, W., editors, *Computer Speech Processing*. Prentice Hall International.

Zee, E. (1980). Tone and vowel quality. *Journal of Phonetics*, 8:147–258.