Decomposition of Fundamental Frequency Contours in the General Superpositional Intonation Model

Taniya Mishra
B.A., Saint Mary's University of Minnesota, 2001

Presented to the Division of Biomedical Computer Science within
The Department of Science & Engineering
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Computer Science and Engineering

September 2008

The dissertation "Decomposition of Fundamental Frequency Contours in the General Superpositional Intonation Model" by Taniya Mishra has been examined and approved by the following Examination Committee:

Jan van Santen
Professor
Thesis Research Adviser

John-Paul Hosom
Assistant Professor

Deniz Erdogmus
Assistant Professor

Bernd Möbius
Associate Professor
University of Stuttgart

Dedication

This thesis is dedicated to my parents and my husband for their constant love and complete support. Thank you for always believing that this day would come. I will love you always.

Acknowledgements

It takes a village to raise a child — and also to write a dissertation. I certainly have many people to thank for helping me achieve the latter.

First and foremost, I thank my advisor, Jan van Santen, for introducing me to the field of speech synthesis and teaching me much of what I know about it; for pointing out the interesting topics in the field yet allowing me the freedom to pursue the topic that interested me the most; for posing many challenging questions that helped shape my thesis; for encouraging me to not give up when the answers eluded me; and last but not the least, for always being there to guide me in all my academic endeavors.

I also thank the other members of my thesis committee, John-Paul Hosom, Deniz Erdogmus, and Bernd Möbius for reading my dissertation and making many constructive suggestions and comments.

I owe much to Esther Klabbers for always being available to discuss ideas, to collaborate on projects, and to read over and comment on early drafts of my research papers and dissertation. I want to also thank Emily Tucker for her help in editing my dissertation.

I also want to acknowledge the many students, faculty, and staff at OGI who willingly participated in the series of listening experiments that I conducted. An extra thanks to John-Paul Hosom, Rachel Coulston, and Meg Mitchell for lending their voices for the speech recordings.

Last but not the least, I want to thank my officemates and colleagues for sharing many adventures and misadventures of graduate school with me: Xiaochuan Niu, Emily Tucker, Fan Yang, Kristy Hollingshead, Akiko Kusumoto, and Qi Miao.

Contents

List of Tables

List of Figures

Abstract

Decomposition of Fundamental Frequency Contours in the General Superpositional Intonation Model

Taniya Mishra, B.A.

Doctor of Philosophy,
Division of Biomedical Computer Science within
The Department of Science & Engineering
and the Oregon Health & Science University
School of Medicine

September 2008
Thesis Advisor: Jan van Santen

The general superpositional approach to intonation posits that the fundamental frequency contour can be quasi-additively decomposed into component curves such as phrase curves, accent curves, and segmental perturbation curves. Currently, these component curves can only be estimated if one assumes a specific superpositional intonation model, such as the Fujisaki model (Fujisaki 1983). The central objective of the presented research therefore is to develop an algorithm for decomposing any arbitrary fundamental frequency contour into its component curves using only the most general assumptions of the superpositional approach. The decomposition algorithm is applied in two ways in this work: (1) to examine the assumptions about the relationship between the shape of the F_0 contour and the perception of prominence in the superpositional context, and (2) to enable the

implementation of the recombinant synthesis method (van Santen et al. 2004:i, van Santen et al. 2005).

Recombinant synthesis, also called multi-level unit sequence synthesis, involves searching one corpus for acoustic units while searching another corpus for components of the F_0 contour, namely, phrase curves, accent curves, and segmental perturbation curves, that are then additively recombined according to the general superpositional approach, and imposed on the acoustic unit sequences using standard pitch modification methods. In this synthesis method, the components of the F_0 contour are obtained from natural F_0 contours, and extracting them requires a decomposition algorithm.

Besides the two specific applications that are demonstrated in this thesis, the stated objective also has a wider significance. If one can estimate component curves of the F_0 contours using only the most general assumptions of the superpositional approach, then this would enable accurate determination of (i) the effects of linguistic control factors on the component curves, (ii) their time courses, and (iii) interactions between curve classes. Barring the discovery of fatal flaws in the superpositional concept along the way, the ultimate result of this research would be the development of a - potentially complex - superpositional intonation model that contains a richer collection of curve classes than current models allow, specifies a clear and phonologically well-justified mapping between each curve class and the F_0 contour, and generalizes well to fit across a wide range of speech materials.

Chapter 1

Introduction

1.1 Background

Text-to-speech synthesis, the technology of artificially creating (*synthesizing*) speech from text, is being used with increasing success in common user applications as well as in assistive technology applications for the visually and vocally impaired (Jans 2000). Researchers predict growth of this technology and greater societal impact (Sproat et al. 1999). Despite its promise, speech synthesis is still not a solved problem (Schroeter et al. 2002, Sproat et al. 1999). Most users are unhappy with the quality of synthesized speech, while some outright reject it. Also, many commercial application developers deliberately avoid the use of speech synthesis in their products (Sproat et al. 1999). In all cases, the main complaint is that synthesized speech sounds "unnatural", i.e., it does not sound like human speech. Incorrect or inexpressive intonation, poor voice quality, and lack of variability in speech have been cited as the main causes for the unnaturalness of synthetic speech (Murray and Arnott 1996, Keller and Zellner Keller 2000). The presented research aims to address the problems stemming from incorrect or inexpressive *intonation*.

Although it has many definitions[1], in this work we define intonation as Jones (2002:237) did: "In popular language ... [intonation is] the way the voice goes up and down in speaking. In scientific language ... [intonation is] the variations in the *pitch* of the musical note produced by the vibrations of the vocal cords." Pitch is the primary perceptual correlate of intonation. It is "that auditory property [of a sound] that enables a listener to place it

[1]See Rossi 2004, p.32–34, for a comprehensive summarize of the many definitions of intonation.

on a scale going from low to high, without considering its acoustic properties" (Ladefoged 1993:186). This means that the term "pitch" only relates to the human perception of intonation. When discussing the acoustic properties of intonation, it is more appropriate to talk about *fundamental frequency* (also referred to as F_0). Fundamental frequency is the primary acoustic correlate of intonation. It is the frequency of the vibrations of the vocal cords. Because of this strong correlation between intonation on the one hand and pitch and fundamental frequency on the other, the latter two terms are sometimes used to refer to intonation. *Prosody* is another term that is often used interchangeably with intonation. However, prosody encompasses more than intonation; it also includes speech duration, loudness, and voice quality[2].

Intonation is crucial to human communication. Some of the important functions of intonation in communication, outlined in Crystal (1987), are as follows. (1) *Emotional*: Expressing speaker emotion (such as anger, fear, joy, or happiness) and attitude (such as excitement, boredom, surprise, friendliness, or reserve). (2) *Grammatical*: Segmenting utterances into clauses and sentences, and indicating utterance modality (such as declarative, exclamative, injunctive, or interrogative); for example, the use of rising intonation for "Mary called?" versus falling intonation for "Mary called.". (3) *Information structure*: Marking the distinction between given and new information. (4) *Textual*: Shaping the structure of large sections of the discourse; for example, in news-reader style speech, different paragraphs corresponding to different news items are marked by different melodic shapes. (5) *Psychological*: Organizing discourse into units that can be easily perceived and memorized, such as the distinctive intonational pattern used to mark lists in utterances. (6) *Indexical*: Indicating personal identity and group memberships of the speakers; for example, members of certain professions have distinctive speaking styles, such as newsreaders, sportscasters, preachers, etc.

[2]Voice quality is the perception of the laryngeal settings (e.g., creaky voice versus breathy voice, each dictated by the pattern of the vocal fold vibrations), supralaryngeal settings (e.g., palatalized voice in which the tongue is positioned higher and further forward than is commonly positioned, versus pharyngealized voice in which the pharynx is constricted more than usual) and overall muscular tension settings (e.g., tense voice versus lax voice) used by a speaker during articulation (Laver 1980, 1994).

Given its important functions in human communication, generating correct and expressive intonation is extremely important for speech synthesis systems. However, it is also considered to be one of the fundamental challenges of speech synthesis (van Santen et al. 2005). The onus of meeting this challenge falls to the *intonation model* that underlies a speech synthesis system. *A model* is a representation of an observable phenomenon in terms of abstract theoretical entities; it is based on well-known assumptions about the phenomenon that it is representing (van Santen et al. 2000). There are two main categories of intonation models, *symbolic intonation models* and *quantitative intonation models*. A symbolic intonation model (e.g., ToBI (Silverman et al. 1992)) represents intonation using an alphabet of symbols and reduces intonation to a set of discrete categories. However, a symbolic model is not adequate for generating intonation in speech synthesis systems because it does not provide the required quantitative information about intonation and its temporal relationship to the underlying speech stream (Mixdorff and Fujisaki 2000). A more suitable model is a quantitative intonation model. A quantitative intonation model represents intonation numerically. More particularly, it represents intonation as a sequence of (time, F_0) pairs; this sequence is called the *fundamental frequency contour* (also called the F_0 contour).

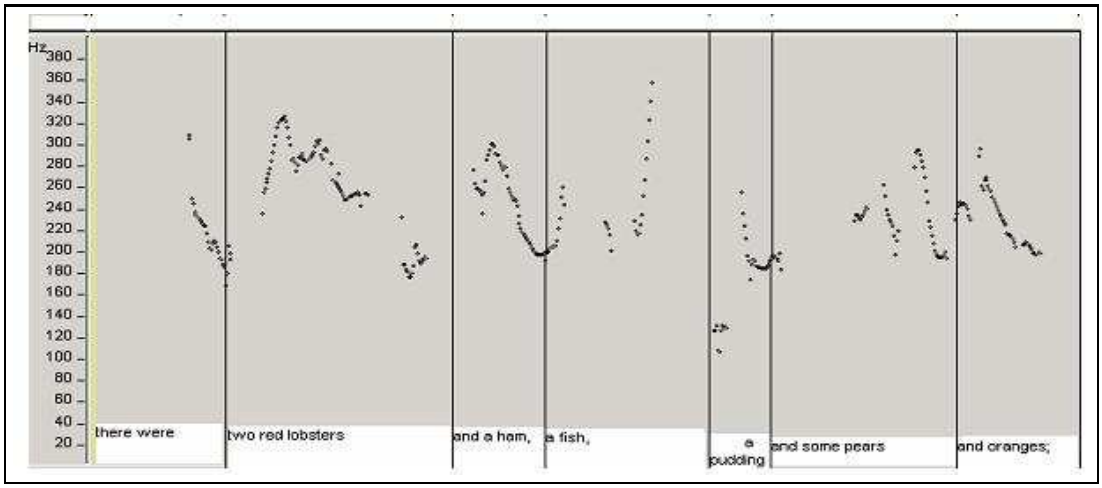

Figure 1.1: Example of a natural F_0 contour.

Quantitative intonation models can be further categorized into those based on the *sequential approach* and those based on the *superpositional approach*. These two approaches

to intonation modeling differ in terms of F_0 contour characterizations as shown in Figure 1.2. The sequential approach characterizes the F_0 contour as a sequence of distinct F_0 movements or targets that are generated left to right. Some examples of intonation models based on the sequential approach are the IPO model ('t Hart et al. 1990), the Pierrehumbert model (Pierrehumbert 1980), the Tilt model (Taylor 2000), and the RNN model (Traber 1992). The superpositional approach characterizes the F_0 contour as an overlay (or *superposition*) of several simpler component contours. Some examples of intonation models based on the superpositional approach are the Fujisaki model (Fujisaki 1983), the Linear Alignment model (van Santen 1997), and the SFC model (Bailly and Holm 2005).

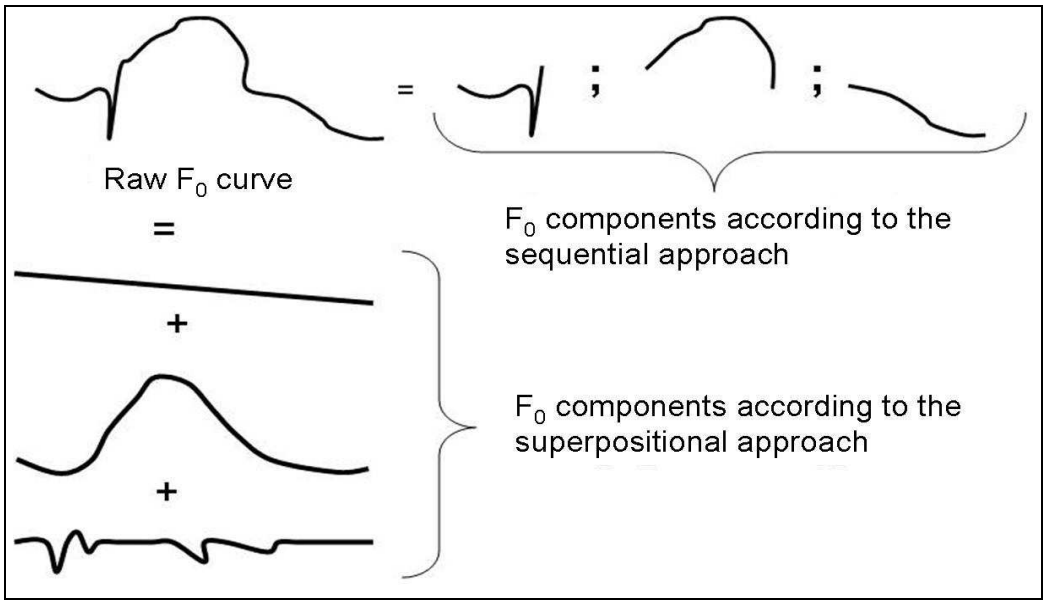

Figure 1.2: Sequential versus superpositional approach.

The two approaches also differ in terms of the assumptions they make in order to develop a mapping between linguistic (and non-linguistic) factors and the F_0 contour. In the superpositional approach, the F_0 contour is assumed to be determined by local factors (e.g., word emphasis, syllable stress) as well as global factors (e.g., sentence modality, overall phrasing structure), whereas in the sequential approach, it is assumed that the F_0 contour is completely determined by local factors while remaining generally unaffected by non-local factors (Pierrehumbert 1980). However, there is sizable empirical evidence that

shows that the F_0 contour is affected by both local and non-local factors, as is discussed in Chapter 3. Consequently, this research is based on the superpositional approach to intonation modeling.

1.2 Statement of the problem

In spite of the potential of the superpositional approach to quantitative intonation modeling, existing intonation models based on the superpositional approach (e.g., the Fujisaki model, the SFC model, and the Linear Alignment Model) fail to meet the challenge of generating a meaningful and expressive F_0 contour for any arbitrary utterance. This failure can be attributed to the specific assumptions about F_0 contour production and/or perception made by particular superpositional models. So far, however, no truly critical test of these assumptions has been performed (van Santen et al. 2004:ii) owing to the challenging problem of decomposing a natural F_0 contour into its component curves (van Santen et al. 2005, Mixdorff 2004).

Decomposing a natural F_0 contour into its component curves is challenging for the following reasons: (1) There is no unique solution to the decomposition of a given F_0 contour. Any number of component contour shapes can add up to form a single F_0 contour. (2) The F_0 contour is often not smooth; it is interrupted by gaps (as shown in Figure 1.1) that represent either unvoiced sounds — when the vocal cords are not vibrating — or pauses in the utterance. (3) Few constraints governing the shape of *the phrase curve* underlying the F_0 contour have been established. Phrase curves are a class of F_0 component curves that are associated with *intonational phrases* (defined in Chapter 3) and represents the global shape of the F_0 contours. (4) Successive *accent curves* in a F_0 contour may overlap in time. Accent curves are a class of F_0 component curves that represent the local peaks and valleys in the F_0 contour, and serve to intonationally highlight certain syllables and words. (5) *Segmental perturbations* cause local maxima and minima to occur in the F_0 contour that may obscure the shapes of the phrase curves and accent curves. Segmental perturbations are the sudden changes in the F_0 contour caused by post-obstruent vowels, nasality effects, intrinsic pitch, etc. The class of F_0 component

curves that represents segmental perturbations are called *segmental perturbation curves*.

Some attempts have been made to solve the decomposition problem, but they are incomplete solutions. In solutions proposed by Fujisaki and colleagues (Mixdorff 2000, Narusawa et al. 2002), decomposing an extended range of F_0 contours is only possible by imposing overly-strict assumptions on the shapes of the component curves. It has been observed, however, that because of these too strong shape constraints, this decomposition technique lacks the generality to fit certain observable F_0 contours (Taylor 1992, 2000); consequently, the linguistic meaningfulness of the F_0 components extracted from such F_0 contours by this decomposition technique is questionable (Ladd 1996, p. 30, Aguero et al. 2004). In a different solution (Bailly and Holm 2005, Bailly and Gorisch 2006), no shape-related assumptions are made; rather, the prototypical shapes of the component curves are learned from different instances of the same class in the corpus using neural networks. However, the results of this decomposition technique are unsatisfactory when sufficient training data is lacking (Raidt et al. 2004). A third solution (Sakai 2004) also makes no assumptions about the shapes of the component curves, and thus only returns the *average* shape of the component curves, which the decomposition algorithm has "learned" from the training corpus (van Santen et al. 2005).

A complete solution to the decomposition problem is needed, not only for its indirect effect on speech synthesis through the examination of assumptions made by particular superpositional models, but also to impact speech synthesis more directly via the implementation of the *recombinant synthesis* method. The recombinant synthesis method, also called the multi-level unit sequence synthesis method (van Santen et al. 2004:i, van Santen et al. 2005), is a superpositionally-motivated approach to speech synthesis. It involves searching one corpus for acoustic units, while searching another corpus for components of the F_0 contour, namely, phrase curves, accent curves, and segmental perturbation curves, that are then additively recombined according to the general superpositional approach and imposed on the acoustic unit sequences using standard pitch modification methods. A key concept of the recombinant synthesis method is that the entire corpus of F_0 component curves is obtained by the decomposition of natural F_0 contours, which means that the existence (and usage) of a robust decomposition algorithm is imperative for the

implementation of this synthesis method.

1.3 Goals of the thesis

The central goal of this research is to develop an algorithm for decomposing any arbitrary natural F_0 contour into its component curves using only the most general assumptions of the superpositional approach, in order to enable the testing of assumptions made by intonation models based on the superpositional approach and to enable the implementation of the recombinant synthesis method.

Using only the *general assumptions of the superpositional approach* (van Santen and Möbius 1999) is a key difference between the decomposition algorithm presented in this work and the standard F_0 analysis techniques based on the Fujisaki model (Mixdorff 2000, Narusawa et al. 2002). The general assumptions of the superpositional approach include few, if any, assumptions regarding the shapes of the F_0 component contours. Consequently, in contrast to the Fujisaki model-based standard F_0 analysis techniques, the presented decomposition algorithm prescribes very few constraints governing the shapes of the pitch component curves, thus enabling truly reliable tests of the model-specific assumptions.

Another important way in which our decomposition algorithm differs from the Fujisaki model-based standard F_0 analysis techniques, as well as other decomposition techniques (such as those presented in Sakai 2004, and Bailly and Holm 2005), is the explicit modeling of segmental perturbations. The reason why modeling segmental perturbations is important is because they are not random Gaussian noise; rather, they represent highly systematic effects of certain segmental classes on the F_0 contour. Here are some examples: segmental perturbations related to post-obstruent vowels cause the F_0 contour to be shifted upward in vowel regions during the first 50-100 msecs after the offset of an obstruent (van Santen and Hirschberg 1994); segmental perturbations due to the presence of nasals cause 5-10 Hz depressions in the F_0 contours that are perfectly aligned with the start and end of the nasal regions (van Santen et al. 1998); and segmental perturbations related to intrinsic pitch cause variations in F_0 contour that are positively correlated with

vowel height, i.e., other things being equal, a high vowel generally has a higher F_0 than a low vowel (Beckman 1986:129).

A key assumption of the general superpositional approach is that different phonological entities are tied to different curve classes. Expanding on this assumption, our decomposition algorithm explicitly models not just component curves tied to phrases and accent groups but also those tied to phone segments. In comparison, other decomposition techniques either only explicitly model curves tied to accent groups and phrases (as in Sakai 2004, Mixdorff 2000, and Narusawa et al. 2002), or use an admittedly simplistic model of segmental perturbations (as in Bailly and Holm 2005). Since the shape of the F_0 contour is influenced by segmental perturbations, modeling these curves adds to the robustness of our decomposition algorithm.

We have divided the central goal into the following three subgoals, each intended to address one key aspect of the central goal.

1.3.1 Subgoal 1: decomposition algorithm development

The first subgoal is to develop a decomposition algorithm for extracting the component curves of natural F_0 contours using only the general assumptions of the superpositional approach. Specifically, this algorithm has the following key features:

- It assumes that the natural F_0 contours it encounters are labeled and segmented, which is a valid assumption in the area of speech synthesis.
- It makes very general assumptions about the shapes of the component curves.
- It explicitly models segmental perturbations.
- It addresses the issue of successive accent curves overlapping in time.
- It appropriately deals with the gaps that occur in natural F_0 contours due to unvoiced phones or silences.
- It decomposes complicated utterances of three common intonation types, namely, declarative, continuation rise, and interrogative; it provides the theoretical framework and extensible implementation to address other intonation types in the future.

- It avoids extracting average shapes of component curves (unlike (Sakai 2004)); instead, it extracts individual natural shapes of component curves on a per utterance basis.

1.3.2 Subgoal 2: use of decomposition algorithm in speech synthesis

The second subgoal is to implement a proof-of-concept (POC) of the recombinant synthesis method in order to demonstrate the essential role of the decomposition algorithm in this synthesis method. The proof-of-concept implementation of the recombinant synthesis method involves

- extracting component contours of natural F_0 contours,
- creating a corpus of the extracted pitch components labeled with information about their relationship to *phonological entities* (i.e., phone segments, syllables, *left-headed feet*[3], and phrases) such as their size in terms of number of left-headed feet or number of syllables, their relative position in the phrase or utterance, and the broad phonemic classes that they contain (vowel or diphthong, sonorant consonant, and other),
- selecting F_0 component curves from the corpus that match the target contour in terms of the aforementioned labels,
- and, additively recombining them to generate a target F_0 contour.

This implementation does not include creating an acoustic unit corpus or searching it for appropriate acoustic units; rather, the target F_0 contours are imposed on pre-recorded natural speech so that the strengths and weaknesses of this intonation generation method are observed unconfounded by the effects of acoustic unit search and concatenation.

[3] A left-headed foot is defined as a sequence consisting of an accented syllable followed by all unaccented syllables that precede the next accented syllable or phrase boundary. The sequence of syllables preceding the first accented syllable in a phrase is called the anacrusis. (Klabbers et al. 2002, Klabbers and van Santen 2003)

1.3.3 Subgoal 3: use of decomposition algorithm in testing assumptions

The third subgoal is to test assumptions regarding the perception of *prominence* and the shapes of F_0 contour components in order to show the decomposition algorithm's utility in testing assumptions about intonation in the superpositional framework. Prominence is what we hear when a word (or group of syllables) "stands out" from those around it in an utterance. The perception of prominence is often assumed to be related to height, position, or even slopes of the fundamental frequency peaks (Gussenhoven et al. 1997); thus, in the context of the superpositional approach, it can be related to the height, position, and slopes of the underlying F_0 component curves. These assumptions are tested by a series of perceptual experiments. Specifically, we explore the following questions:

- What is the combined and individual influence of accent height, accent peak location, phrase onset, and phrase slope on perceived prominence?
- Is there a line of reference that listeners use to determine the prominence of peaks in the F_0 contour? If so, is the reference line abstract or observable?
- Does contrastive accent influence perceived prominence differently than normal accent?

1.4 Organization of the thesis

Chapter 2 presents a review of quantitative intonation models that have been used for text-to-speech synthesis and their relative advantages and disadvantages. The following three chapters represent the technical part of this thesis, each chapter providing a solution corresponding to one of the three subgoals outlined in Section 1.3. Chapter 3 presents (1) the motivation for a superpositional approach to the decomposition problem, (2) a description of the general superpositional model of intonation, which will be the basis of the decomposition algorithm, and (3) a description of the decomposition algorithm that decomposes natural F_0 contours into their component curves. Chapter 4 describes the proof-of-concept implementation of recombinant synthesis, and discusses its comparative performance against other synthesis methods. Chapter 5 describes a series of perceptual tests performed to test assumptions regarding perceived prominence and the shapes of F_0

contour component curves. Chapter 6, the last chapter of this thesis, describes a summary of the main findings of this work, its overall significance, and the future directions of this area of research.

Please note that significant portions of the material presented in many of these chapters have been previously published by the author as conference papers, book chapters, and journal articles.

Chapter 2

Intonation Models in Speech Synthesis: A Literature Review

Several quantitative intonation models — based on either the superpositional approach or the sequential approach — have been used in existing text-to-speech synthesis systems. Here we review ten such intonation models, five using the sequential approach, three using the superpositional approach, and two that are neither sequential nor superpositional but contain certain features that may be considered sequential or superpositional. For each model, the implicit assumptions underlying the model, the explicit representation of F_0 chosen by the model, and the way in which the model can be used to compute intonation in a text-to-speech system are outlined.

2.1 Intonation models in the sequential approach

Models based on the sequential approach to intonation modeling consider the F_0 contour to be a sequence of distinct F_0 movements or targets that are generated left to right. The five sequential models that are reviewed here are Pierrehumbert's theory of intonation, ToBI-based approaches, the recurrent neural network (RNN) model, the tilt model, and the IPO approach.

2.1.1 Pierrehumbert's Theory of Intonation

The Pierrehumbert theory of intonation (Pierrehumbert 1980), developed by Pierrehumbert as part of her doctoral dissertation, is a phonological model of intonation. It is based

on autosegmental-metrical (AM) phonology (Ladd 1996, Goldsmith 1990). In keeping with the AM theory, the Pierrehumbert model considers intonation to be a sequence of high(H) and low(L) tones. The H and L tones are in *phonological opposition*, i.e., the difference in sound between them serves to distinguish intonational meaning. The two types of tones never interact with each other; rather, they follow each other sequentially in an utterance.

The H and L tones are the building blocks of three larger tone units: *pitch accents*, *phrase accents*, and *boundary tones*. *Pitch accents* mark prominence. They are either single tones (H*, L*), or pairs of tones (L+H*, L*+H, H+L*, H*+L); the * denotes the association of the tone with a stressed syllable. One or more pitch accents comprise an *intermediate phrase*. One or more intermediate phrases comprise an *intonational phrase*, the largest prosodic unit posited by this theory. The edges of the intonational phrase are marked by *boundary tones*. The boundary tones are single tones (%H, %L, H%, L%); the % denotes the association of the boundary tone with the pitch onset or offset of the intonational phrase. Pitch movement between a pitch accent and a boundary tone is indicated by a *phrase accent* (H-, L-), denoted by the diacritic, -.

To ensure that the model renders well-formed intonational representations, Pierrehumbert defined a finite state grammar that specifies the combinations in which pitch accents, phrase accents, and boundary tones can occur. She also devised a set of *phonetic realization rules* (Pierrehumbert 1981) to produce F_0 contours from the phonological model of intonation described above.

Generating the F_0 contour of a target utterance in a TTS system using the Pierrehumbert theory of intonation involves three main steps. First, determine the tonal representation of the utterance using the finite state grammar. Second, specify the target F_0 values of the high and low tones depending on the metrical prominence of the associated syllables, and the F_0 values of the preceding tones using the phonetic realization rules. Third, using the rules again, connect the target F_0 values to generate a F_0 contour: If two neighboring targets are far apart, connect them with a sagging contour implemented via a quadratic function. Otherwise, connect them via monotonic curves.

2.1.2 ToBI-based approaches

ToBI (Silverman et al. 1992) stands for *tones and break indices*. Based on Pierrehumbert's theory of intonation, it was developed in four research meetings between 1991 and 1994 as a standard for describing American English intonation. It has since been extended to transcribe other languages and dialects (Mayo et al. 1997, Reyelt et al. 1996, Campbell and Venditti 1995).

ToBI consists of three parallel labeling tiers. The first tier is the *tone tier*. The tones specified by Pierrehumbert's theory are labeled in the tone tier. The second tier is the *break index tier*. In the break index tier, *break indices*, ranging from 0 to 4, are marked. Break indices mark the boundary strength between adjacent words; 0 indicates no boundary, 3 indicates an intermediate phrase boundary (- in Pierrehumbert's model), and 4 indicates a intonational phrase boundary (% in Pierrehumbert's model). The third tier is a miscellaneous tier, where hesitations, disfluencies, laughter, non-speech sounds, etc., are labeled.

It is important to note that ToBI is a labeling system. It does *not* specify the means to produce quantitative intonation from the labels. However, there are both rule-based and statistically trained approaches that can be applied to the ToBI labels to generate F_0 contours. An example of the rule-based approach is Jilka's hand-crafted rule system for specifying the F_0 contour of American English from ToBI labels (Jilka et al. 1999). Jilka's approach is similar to Pierrehumbert's *phonetic realization rules*. These rules specify the target F_0 values associated with ToBI labels, depending on pitch range and the voiced part of the syllable. The target F_0 values are calculated from left to right, taking into account only preceding ToBI labels, not subsequent ones. The F_0 contour is produced by linear interpolation between target points.

An example of the statistically trained approach is Black and Hunt's linear regression based approach for generating F_0 contours from ToBI labels (Black and Hunt 1996). This approach simply involves predicting three target F_0 values for every syllable — one at the start of the syllable, one at mid-vowel position, and one at the end of the syllable — by means of linear regression. The prediction formula is defined by Black and Hunt (1996)

as follows:

$$F_0\ value = I + w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \tag{2.1}$$

The f_i variables indicate the features that contribute to the F_0 value of a syllable, such as the ToBI label associated with the syllable, syllable position in the phrase, and syllable stress. I and w_i are parameters that are estimated by linear regression.

2.1.3 The RNN Intonation Model

The recurrent neural network (RNN) intonation model, developed by Traber (1992), uses *neural networks* to predict the F_0 contour. Simply put by Traber (1993), a *neural network* can be considered to be a non-linear statistical model with many parameters. These parameters are estimated in the training phase so that they can return an optimal set of outputs from the corresponding input. The neural-network-based approach was motivated by the goal of using minimal human effort to obtain high quality intonation, that is, while humans would specify *which* phonological units were relevant for the phonetic realization of intonation, clever machine learning techniques would figure out *how* the phonological units map to the F_0 contour.

In the implementation of this intonation model in the SVOX TTS system (Traber 1995), the F_0 contour was generated per syllable. Per syllable, the accent value of the syllable and segmental properties relating to its position in the phrase and the sentence, as well as the accent values of the neighboring syllables, are given as input to the neural network. The network outputs the F_0 region related to the syllable, represented by eight samples of the F_0 contour. The resultant F_0 regions are concatenated together to produce the complete F_0 contour.

2.1.4 The Tilt Intonation Model

The tilt intonation model, developed by Taylor and Black (Taylor 2000) at the Center for Speech Technology Research of the University of Edinburgh, considers intonation to be sequence of intonational events, which are parameterized by *tilt parameters*. The model posits four basic types of intonational events: pitch accents, boundary tones, connections

(regions in the F_0 contour between two pitch accents, two boundary tones, or a pitch accent and a boundary tone), and silence.

Pitch accents and boundary tones are each modeled by piecewise combinations of quadratic functions; these quadratic functions may be rising or falling functions. Connections are modeled by straight-line interpolations. The amplitude and duration of the rising and falling quadratic functions, the position of the associated intonational event in the time-F_0 plane, together with a tilt value associated with each event constitutes the set of tilt parameters associated with the intonational events. The tilt parameter represents the amount of rise and fall of each accent. The tilt value is the difference in the amplitudes of the rise and fall functions, divided by their sum (Dusterhoff and Black 1997), as shown in the equation below.

$$tilt = \frac{|Amp_{rise}| - |Amp_{fall}|}{|Amp_{rise}| + |Amp_{fall}|}$$

The tilt value ranges from -1 to 1, where -1 indicates a pure fall, 1 indicates pure rise and 0 indicates a rise followed by a fall of equal magnitude. Thus, the tilt model uses continuous parameters rather than imposing categorical classification on the intonational events.

Dusterhoff and Black (1997) have shown that the tilt model can be successfully used to predict F_0 contours in a text-to-speech system. The tilt-based F_0 generation process has two stages: a training stage and a testing stage. The training stage requires a training database. The database is labeled with tilt events, either automatically or by hand. For each syllable in the database marked with a tilt event, a set of linguistic-prosodic features are extracted. The features are grouped into separate training sets depending on event type. A CART (Classification and Regression Tree; Breiman et al. 1997) training algorithm is applied to each of the training sets to develop a decision tree for every tilt parameter. The decision trees thus describe the tilt parameters in terms of an optimal subset of the extracted features.

The training stage described above is performed offline. The tilt parameter descriptions obtained in the training stage are used in the testing stage for F_0 contour generation of given text. The given text is labeled with tilt events, and the same set of linguistic-prosodic features (as in the training set) are extracted. The related tilt parameters are

calculated from the extracted features using the descriptions obtained from training. The tilt parameters are then plugged into predetermined quadratic or linear functions to model the pitch accents and boundary tones or connections, respectively.

2.1.5 The IPO Approach

The IPO approach ('t Hart et al. 1990, Cohen and 't Hart 1967) was developed at the Institute of Perception Research (IPO) at Eindhoven, Netherlands in the 1960s. It was originally used to model Dutch intonation. The IPO model is often classified as a *perceptual* intonation model because of the assumptions underlying the model: Not all changes in F_0 are perceived by the human ear; only F_0 changes that are perceived by the human ear need to be modeled; and finally, the human ear perceives tone variations (rise vs. fall) and not tone intensities (high vs. low).

Given these assumptions, the IPO approach models the raw F_0 contour as a piece-wise linear approximation of the original contour, known as a *close copy contour*. It is called a *close copy contour* because upon resynthesis, it is perceptually indistinguishable from the original F_0 contour. Generating the close copy contour also includes specifying the *declination line* (a line that represents the overall trend of the F_0 contour).

Close copy contours are classified into discrete, phonetically defined types of F_0 rises and falls. The classification parameters describe the deviation of the close copy from the declination line and include descriptive factors such as its height and slope relative to the declination line, its span relative to the span of the declination line, its timing in relation to associated syllable(s) duration, and its rate of change. The particular parameters used for classification differ from language to language.

Once an inventory of F_0 rises and falls covering the entire combinatorial space of the classification parameters has been collected, a *grammar* specifying the possible and permissible combinations of the F_0 rises and falls is written in terms of the parameters. When the IPO intonation model is used in speech synthesis systems, this grammar is used for F_0 contour generation. F_0 contours predicted by this grammar must be perceptually equivalent to and as acceptable as natural F_0 contours (Sproat 1997). IPO has been implemented in speech synthesis systems for Dutch (Terken 1993), English (Willems et al.

1988), and German (van Hemert et al. 1987).

2.2 Intonation models in the superpositional approach

Models based on the superpositional approach to intonation modeling consider the F_0 contour to be a superposition (or overlay) of simpler components. The three superpositional models reviewed here are the Fujisaki model, the linear alignment model, and the superposition of function contours (SFC) model.

2.2.1 The Fujisaki Intonation Model

The Fujisaki intonation model, developed by Fujisaki (Fujisaki 1983, Fujisaki and Hirose 1982) is the best known intonation model based on the superpositional approach. In the Fujisaki model, the F_0 contour is considered to be an addition (in the log domain) of two components: the *phrase command* and the *accent command*.

The *phrase command* characterizes the overall trend of the intonation of an utterance, represented by the global movement of the associated F_0 contour. The *accent command* on the other hand, characterizes particularly extreme excursions of intonation used to stress certain syllables or words in the utterance, represented by the local peaks and valleys in the F_0 contour. The phrase command is modeled by pulses, while the accent command is modeled by step functions. The discontinuities in the two commands are then smoothed using separate filters to output phrase and accent components that appear continuous. The phrase and accent components are then added in the log domain to produce an *additive F_0 contour* — the defining characteristic of the Fujisaki model.

To use the Fujisaki model for F_0 prediction in text-to-speech synthesis, the pulses are placed at intonational phrase boundaries, while the step functions are associated with other key phonological units such as accent groups. Linguistic and other properties of the text determine the amplitude of each of the commands and the width of the accent commands. This model has been successfully used for intonation modeling of many languages (Fujisaki 2002, Mixdorff 2004).

2.2.2 The Linear Alignment Model

The linear alignment model was developed by van Santen and Möbius (1999) at Bell Laboratories. The distinguishing characteristic of this model is that it pays particular attention to the *alignment* between the pitch contour and the segmental stream underlying it.

Its concern with alignment is most effectively expressed in its modeling of the accent curve component. The accent component represents the same aspects of the F_0 contour as the accent component in the Fujisaki model, though it is modeled differently in this model. The accent curve is modeled by parameterized time warps of an accent curve *template*. The template can be defined as a sequence of anchor values $T_p =< P_1, P_2, ...P_n >$ that describe the archetypical shape of the associated accent curve type P. Also associated with P is an *alignment parameter matrix*, an ensemble of regression weights that describe the alignment of P to the segmental region underlying it. All accent curves of type P have the same template and the same alignment parameter matrix; they differ from each other only in terms of their duration.

Besides the accent curve, two other components of the additive F_0 contour are specified by the linear alignment model: the phrase curve and the segmental perturbation curve. As in the Fujisaki model, the phrase curve illustrates the long-term shape of the F_0 contour. The phrase curve is modeled by a piecewise quasi-linear function spanning three time points. The segmental perturbation curve describes the segmental influences on the pitch contour such as pitch increase in vowels following voiceless plosives and pitch lowering in nasals and glides. The segmental perturbation curves are modeled by exponential decay functions.

The linear alignment model has been used for generating intonation in the Bell Labs multilingual text-to-speech system (van Santen et al. 1998). To synthesize speech from text, each of the three components of F_0 specified by the linear alignment model have to be related to linguistic entities. The phrase curve is anchored at three points: the start of the utterance, the start of the syllable that carries the nuclear pitch accent, and the end of the utterance. The accent curve is tied to a left-headed foot, which is defined as a sequence

consisting of an accented syllable followed by all unaccented syllables that precede the next accented syllable or a phrase boundary. The degree of emphasis at a particular foot is obtained by multiplying the accent curve by a height factor. Segmental perturbation curves are anchored at vowel onset. The amplitude of this function is determined by the broad class of the onset consonant; it has a maximal value for voiceless consonants, a smaller value for voiced obstruents, and a zero value for sonorants.

2.2.3 The SFC Model

The superposition of functional contours (SFC) model of intonation was developed at the Institute for Speech Communication. It was proposed by Aubergé (1993) and implemented by Holm and Bailly (2000). Like other superpositional models, the principal assumption of the SFC model is that the pitch contour is obtained by a superposition of simpler contours. In case of the SFC model, the simpler contours are multiparametric contours called *functional contours* (FCs).

Functional contours form the core of the distinguishing assumption of this model. They are assumed to directly encode specific *metalinguistic functions* tied to various discourse units, without any intermediate representation (Bailly and Holm 2005). *Metalinguistic functions* refer to intonation functions that delimit phonological units and convey propositional and interactional information about these units within the discourse. Examples of metalinguistic functions are hierarchy, segmentation, emphasis, and speaker attitude.

Every functional contour has the following three properties:

1) It is *function-specific*, i.e. tied to a particular metalinguistic function.

2) It spans the extent of the unit(s) tied to the function it encodes; this extent is called the *scope* or domain of the FC.

3) The FC shape is a function of the metalinguistic function it encodes and its scope. However, it is important to note that the FC shape is not specified a priori in this model; rather it emerges in the training phase of the model's implementation.

The SFC model has been implemented for pitch prediction in TTS systems for German, French, and Mandarin Chinese. The metalinguistic functions encoded by the functional contours are defined. One *contour generator* per function is implemented as a neural

network. Each contour generator generates a family of functional contours that encode the same metalinguistic function and hence have the same shape, differing only in terms of their time domains. The input to each contour generator is information relating to the scope of the associated function and the position of each syllable within the scope. The output is four output parameters per syllable (three F_0 values and a lengthening factor).

Training the contour generators to generate a particular pattern of functional contour is not a straight-forward process because recovering the unique contributions of the contour generators to their sum (i.e. the F_0 contour) is an ill-posed problem. To determine the individual contributions of each contour generator and the particular pattern of the FC, an *analysis-by-synthesis loop* (Holm and Bailly 2000) is used.

2.3 Other intonation models

There are some intonation models that are based neither on the sequential nor on the superpositional approach, though they contain some features that may be considered superpositional or sequential. Here we review two such models: the Kiel model, which is referred to as a prosodic model, and the STEM-ML model, which is referred to as a physiological model.

2.3.1 The Kiel Intonation Model (KIM)

The Kiel intonation model was developed by Kohler and his colleagues (Kohler 1991, Kohler 1997) to model intonation patterns in German. In KIM, the F_0 contour is modeled as a sequence of *global intonational units*, each unit linked to one emphasized word. The global intonational units are considered to be produced and perceived holistically, and cannot be split. These global units are either peaks or valleys or peak-valley combinations, and differ from each other in term of their pragmatic, semantic, and syntactic functions. They were determined in KIM by means of function-oriented phonetic experiments. KIM postulates that there is a prototypical intonational unit associated with a particular pragmatic-semantic-syntactic function combination. However, KIM does not ignore the *microprosodic phenomena* (e.g., F_0 shifts at the obstruent-vowel boundaries and

F_0 changes in nasals and glides; effects of intrinsic pitch are also included in this category) observed in the F_0 contour; microprosodic phenomena is also a part of the model.

Since KIM was developed with a focus on TTS synthesis, the F_0 prediction rules are well specified. KIM applies two sets of rules, namely, *symbolic feature rules* and *parametric rules*, for pitch prediction in TTS systems. The symbolic feature rules are applied to *phonological units* which have been annotated with syntactic, pragmatic, and semantic markers. The *phonological units* are either segmental (vowels and consonants) or non-segmental (morphological and phrase boundaries). The symbolic feature rules output the global intonational units associated with the phonological units, encoded as binary features (such as +/-terminal, +/-valley, +/-quest, +/-early, +/-late). These feature values are then used by the parametric rules to generate the F_0 contour of the target utterance. The parametric rules include rules for aligning the global intonational units with the segmental structure of the target utterance, downstepping of accent peaks, speech rate, prosodic boundaries and, finally, articulation-induced microprosody (Kohler 1995).

2.3.2 The STEM-ML Intonation Model

STEM-ML (Soft TEMplate Markup Language) is a physiological model of intonation. It was developed by Kochanski and Shih (2000) to investigate the deviation of Mandarin Chinese tones from their expected canonical shape when occurring in natural sentences. However, this model has been designed to be language-independent, and thus, can be applied to non-tone languages like English.

The STEM-ML intonation model is founded on three key assumptions:

1. Human speech is pre-planned several syllables in advance.
2. Humans produce speech that optimally balances the physiological effort required to speak against unambiguity of the spoken message. The speaker expends maximal effort to produce correct prosody at prosodically crucial events because the cost of ambiguity is high at these points. However, he minimizes effort between such events because the cost of ambiguity is low.
3. Speech prosody is continuous and smooth over small time periods.

STEM-ML includes a tagging system (see Kochanski and Shih (2003) for a complete description of the tagset) for intonation mark-up and specification and a quantitative model to generate the F_0 contour. Two important building blocks of the STEM-ML model are parameters and *soft templates*. The parameters are associated with the tags in the tagging system. The soft templates are a part of the quantitative intonation generation model. The parameters and the soft templates together generate the F_0 contour.

In this model, the F_0 contour is considered a concatenation of the local accents. The local accents are represented by the soft templates. The soft templates are "soft" in the sense that the accent templates allow substantial distortion caused by neighboring accents. The concept of soft templates arises from the previously stated pre-planning assumption. An accent template is affected by past as well as future templates. The degree of distortion is controlled by a parameter called *strength*. The strength parameter reflects the cost of ambiguity in the previously stated assumption regarding optimally balanced speech. Therefore, if strength (hence cost of ambiguity) is large, the template shape remains unchanged to reflect maximal articulatory effort, whereas if it is low, the accent shape is compromised to reflect minimal articulatory effort.

Besides local tags that control local accent shapes, there are global tags that control speaker-specific information. Thus, a STEM-ML model is built on a particular speech corpus. The implementation of the STEM-ML model involves two phases: the learning phase and the generation phase. In the learning phase, the values of the parameters are determined iteratively by minimizing the difference between the actual F_0 of every STEM-ML tagged utterance in the corpus and the F_0 predicted by the model. In the generation phase, when faced with the task of generating a target utterance, the model first tags the text underlying the utterance, then uses the pre-determined values of the parameters associated with the tags to modify the soft templates, and finally, concatenates the modified accent templates to produce the F_0 contour. STEM-ML has been used to model Mandarin (Kochanski and Shih 2001), Cantonese (Lee et. al 2002), and English (Shih and Kochanski 2003) speech.

2.4 Advantages and disadvantages

As shown in the preceding descriptions, the different intonation models are implemented in TTS systems using different computational mechanisms: linguistic-prosodic rules, templates, neural networks, decision trees, and linear regression. Each of these computational mechanisms has its advantages and disadvantages. The advantages of the rule-based approaches are that they are easy to implement and that they produce consistent intonational contours. The disadvantage of the rule-based approaches is that intonational contours lack the richness and variability of natural intonation contours. The machine learning approaches (i.e., those using decision trees, neural networks, and linear regression), on the other hand, have the advantage of being able to produce natural-sounding intonational contours, because they learn the mapping between annotated text and corresponding natural pitch contours in the training phase. However, the machine learning approaches have the disadvantage of requiring large amounts of training data to cover the combinatorial space of phoneme sequences and prosodic contexts. Since it is not feasible to cover this entire combinatorial space, machine learning approaches are often beset with data sparsity problems. The reason for capturing the contours related to the same phoneme sequence in different prosodic contexts is that the shape of the contour changes depending on context. One of the main advantages of the template-based approach is its ability to handle the changing shape of an intonation contour depending on context. The template-based approach posits a reasonable number of templates (or prototypical shapes) related to different prosodic events that can be distorted by means of parameters. However, one of the drawbacks of this approach is that the template and the parameters may fit the data in a way that is phonologically absurd.

Chapter 3

PRISM: Procedure for Representing Intonation in the Superpositional Model

Given an F_0 contour, the outcome of our decomposition algorithm is a set of component curves that constitute the F_0 contour in accordance with the general assumptions of the superpositional model. Our decomposition algorithm is therefore called PRISM, which stands for "procedure for representing intonation in the superpositional model". Besides being an acronym, the name PRISM also symbolizes the key idea of decomposition.

PRISM is described in full detail in Section 3.3 of this chapter. We lead up to the description of PRISM with a discussion of our motivation for using a superpositional approach to intonation modeling (Section 3.1) and the limitations of existing decomposition algorithms based on the superpositional approach (Section 3.2).

3.1 Motivation for a superpositional approach

The superpositional approach and the sequential approach to intonation modeling differ in the assumptions they make in order to develop a mapping between the linguistic factors and the F_0 contour. In the superpositional approach, the F_0 contour is assumed to be determined by local factors (such as word emphasis and syllable stress) as well as global factors (such as utterance modality and overall phrasing structure). In the sequential approach, it is assumed that the F_0 contour is completely determined by local factors while remaining generally unaffected by global factors (Pierrehumbert 1980).

There is, however, sizable empirical evidence showing that the F_0 contour is affected

not just by local factors but also by non-local factors. A well-known example is the outcome of Liberman and Pierrehumbert's "list experiment" (Liberman and Pierrehumbert 1984). The results of this experiment showed that the F_0 contours of list-type utterances downstepped at each list item in relation to the item prior to it. Since then, the constrained nature of the F_0 contours of lists has also been observed by Selting (2007). Kutik et al. (1983) showed that an F_0 contour interrupted by a parenthetic remark resumes right afterwards such that the contour is similar to the contour in the equivalent utterance without the parenthetic remark. This shows that the speaker pre-plans the global shape of the utterance, depending on the utterance type (e.g., lists versus parenthetical remarks). A third example can be found in Klabbers' work (2004). She found that peaks in a monosyllabic foot generally occur later when the foot is phrase-medial compared to when it is phrase-final or utterance-final. This shows that the shape of the F_0 contour at a particular time point is affected by the phrasing structure of the entire utterance. Other examples are cited in Möbius (1995). In light of such examples that clearly show the effect of global factors on the F_0 contour, assuming that the F_0 contour is determined entirely at a local level is unreasonable. Thus, in the presented research, we take the superpositional approach to intonation modeling.

We are further persuaded in favor of the superpositional approach due to the capability of superpositional models to determine the individual effects of various factors on the F_0 contour. Every component contour in the superpositional model is in principle orthogonal to, or independent of, all other contours; thus it can elegantly capture the effect of the factor with which it is associated (Möbius 1995). In contrast, it is unlikely that a sequential model will capture the individual effects of the various factors, given its disregard for the effect of global factors on the F_0 contour. Determining the effect of any local factors in the sequential approach will probably be confounded by the actual effect of one or more global factors that were ignored.

3.2 Limitations of existing decomposition algorithms based on the superpositional approach

Existing decomposition algorithms based on the superpositional approach have certain limitations. We discuss these limitations in terms of two key issues: balancing constraint with coverage in a decomposition algorithm, and the relevance of microprosodic modeling.

3.2.1 Balancing constraint with coverage in a decomposition algorithm

The main difficulty of F_0 contour decomposition in the superpositional framework is that there is no unique solution to the decomposition of a single F_0 contour into multiple components. There are an infinite number of component contour shapes that can add up to form the given F_0 contour. The problem is essentially under-determined because there are more unknowns than data. An important issue therefore is to regularize the decomposition problem by making theoretical assumptions about the shapes of the component curves that are realistic from a physiological, linguistic, or paralinguistic point of view.

To be physiologically founded, the assumptions must stem from what is physiologically possible (e.g., the velum can only block the nasal cavity at one end); to be linguistically founded, the assumptions must be rooted in what is allowed in a language; and to be paralinguistically-based, the assumptions must be conditioned by those aspects of speech "that do not belong to the arbitrary conventional code of language, but which nevertheless are meaningful and important in speech communication" (Schotz 2003), such as speaker emotions and attitudes. The relative importance of the different classes of assumptions in shaping the surface manifestations of the F_0 contour is without consensus. For example, in the Fujisaki model, the assumptions about the shapes of the component curves are primarily based on the physiological and physical mechanisms of the larynx (Fujisaki 1983). In contrast, the SFC model (Bailly and Holm 2005) has no physiological motivations. Rather, the assumptions regarding component shapes are entirely motivated by higher-order (para) linguistic constraints (Raidt et al. 2004).

The importance of different classes of assumptions aside, the main goal of these theoretical assumptions about the shapes of the component contours is to constrain the

decomposition problem. However, this goal is not easily achieved because constraint conflicts with coverage, the other goal of decomposition. A robust decomposition algorithm should have *wide coverage*, i.e., it should describe as many perceptually and generatively distinct intonational events as possible (Taylor 2000).

Overly strict assumptions imply limited coverage, whereas unrestrained coverage predicates limited assumptions. None of the existing superpositionally-based decomposition algorithms are able to adequately balance the two due to the unsuitability of the underlying assumptions. We present three examples to illustrate this point. The first example involves the Fujisaki model (Fujisaki 1983). As mentioned earlier, this model makes assumptions about the shapes of the component curves on the basis of the physiological and physical mechanisms of the larynx. The assumptions are quantitatively described in the model as mathematical equations that specify the particular shapes of the F_0 contour components. However, standard intonation analysis methods based on the Fujisaki model assume *overly strict equation-specified shapes of component curves* and thus lack the generality to fit all observed F_0 contours (Taylor 1992, 2000). When the fit is accurate (such as for algorithms described in Mixdorff 2000 and Narusawa et al. 2002), the linguistic meaningfulness of the parameters is questionable (Aguero et al. 2004).

A second example illustrating the constraint versus coverage problem encountered in component curve extraction involves the SFC model (Bailly and Holm 2005). In marked contrast to the Fujisaki model, the SFC model specifies no low-level shape assumptions, relying instead *only* on high-level (para)linguistic assumptions (Raidt et al. 2004). The shapes of the component curves are determined by *contour generators* implemented as neural networks; each generator generates one prototypical shape (Bailly and Gorisch 2006) that describes a family (or a class) of component curves. However, the results of intonation analysis based on this model indicate that the complete lack of shape constraints may not be the ideal solution either. In an objective evaluation (Raidt et al. 2004), it was found that prosody generated by the SFC model had certain shortcomings when there was not enough statistical information available, i.e. there was a lack in pertinent training data. This dependency of performance on training data is a classic problem encountered by data-driven systems that may be alleviated by specifying some well-founded assumptions

about the shapes of the component curves.

The need for shape-related assumptions is further illustrated by another evaluation performed by Bailly and Gorisch (2006). It was found that the concept of using a single prototypical shape per component class in the SFC model is unable to satisfactorily model the F_0 deviations in German. The authors conjecture that either a finer classification of the component classes or a better understanding of whether the deviations are contextual or arbitrary may solve the problem. We, however, speculate that problems arise from the lack of any low-level assumptions about the component contour shapes. As a result, the neural networks learned a sort of "averaged" prototypical shape per curve class, with little consideration of the relative importance of different aspects of the particular shape. Also, given the admittedly simplistic method of separating microprosodic phenomena from the prosodic phenomena (Bailly and Holm 2005), it is possible that segmental perturbations may be skewing the "averaged" shape away from the "true" shape of the component curves.

The aforementioned examples illustrate the shortcomings of intonation models (and the associated decomposition strategies) that specify overly strict or overly lenient assumptions about the component contour shapes. A promising compromise between the two is provided by van Santen and Möbius's Linear Alignment Model (van Santen and Möbius 1999) that delineates fewer shape-related assumptions than the Fujisaki model, though more than the SFC model. In the Linear Alignment Model, component curves in the same phonological or perceptual class are generated from a common template using non-linear time-warp functions. Components differ either by having different templates or different time-warp functions (van Santen and Möbius 1997). A significant claim of this model is that the shape of the phrase curve in this model is essentially unconstrained, except for the assumption that it remains smooth over long stretches of time (van Santen 2002), in order to facilitate considerable and meaningful variability in the phrase curve. This is certainly a very desirable feature of the intonation model and has the potential to remedy the lack of generality encountered by intonation analysis algorithms based on the Fujisaki model.

However, this property of the phrase curve model was not exploited for decomposing

natural F_0 contours. Rather, the phrase curve was obtained by drawing a line from the start of a pitch accent to the end of the phrase (van Santen 2002). Crucially, this method of decomposing F_0 contours has been shown to be successful only in decomposing F_0 contours of very simple declarative carrier phrases in which one or two words were systematically varied (van Santen and Möbius 1997). But the success of this decomposition strategy hinges on the fact that for these simple utterances, each containing a single accent curve, it is safe to assume that the observed local minima correspond to the phrase curve. In the case of more complex utterances, this assumption will not hold true, either due to segmental perturbations or *overlapping* accent curves. Within the context of the superpositional approach, accent curves are considered to be overlayed on the phrase curve to form the F_0 contour. It is conjectured that sometimes neighboring accent curves overlap in time. When accent curves overlap, the observed minima correspond to the point of overlap between the accent curves and not to the phrase curve; thus the decomposition strategy built on the assumption that observed local minima correspond to the phrase curve will fail. This decomposition strategy also appears to overlook the occurrence of gaps in the F_0 contour due to unvoiced phones. These gaps may obscure the true phrase curve and thus thwart the decomposition strategy.

Studying the issue of constraint versus coverage for different existing decomposition algorithms reveals that a decomposition algorithm that can balance the two conflicting goals properly (1) must not be based on overly strict assumptions about the shapes of the component curves, yet must not be without any shape-related assumptions; (2) must be independent of training data; and (3) must address important considerations regarding the F_0 contour such as overlapping accent curves, gaps in the F_0 contour, and effects of *microprosody* (i.e., segmental perturbations) in order to be able to handle both simple and complex utterances. We have handled all three of these features in our decomposition algorithm PRISM (as outlined in Section 1.3.1) in an attempt to properly balance constraints with coverage in the decomposition of natural F_0 contours.

3.2.2 The relevance of microprosodic modeling

In the previous section, microprosody was briefly mentioned. Microprosody, which is another term used to refer to segmental perturbations, includes vowel intrinsic pitch effects, perturbations of post-obstruent vowels, presence of voiceless regions, and nasality effects on the F_0 contour (van Santen and Möbius 1999). So far, most existing superpositional intonation models and decomposition strategies pay scant attention to microprosody. The Fujisaki model completely ignores microprosody, and because of that, intonation analysis algorithms based on the Fujisaki model smooth out microprosodic variations in the F_0 contour as a preprocessing step to the actual analysis (Mixdorff 2000, Aguero et al. 2004). The SFC model and the related intonation analysis use a microprosodic model that is admittedly "too simple" and needs improvement (Bailly and Holm 2005). The same neglect of microprosodic modeling marks the sequential models as well.

However, modeling microprosody should not be neglected in intonation modeling for the following three reasons. First, microprosodic variations of the F_0 contour are not random Gaussian noise; rather, they represent highly systematic effects of certain segmental classes on the F_0 contour; for example, the F_0 contour is shifted upward in vowel regions during the first 50-100 ms after the offset of an obstruent (van Santen and Hirschberg 1994). Combined together, several microprosodic variations may produce spurious local maxima and minima on the F_0 contour, or perturb what otherwise might have been a linear contour, or artificially create a linear contour (van Santen and Möbius 1999). Microprosodic variations, thus, can obscure the "true" shapes of the phrase curves and accent curves, the two components of the F_0 contour that are of greater interest in intonation modeling. Therefore, a decomposition algorithm must include explicit models of microprosody so that during decomposition, microprosodic variations can be factored out from the F_0 contour to uncover the "true" shapes of the phrase curves and accent curves.

The idea of factoring out microprosodic variations from the F_0 contour follows easily from human auditory perception. In Silverman (1987), a group of British English listeners were asked to judge the prominence of two accented words that differed only in terms of the vowel; one word contained a high vowel such as /i/ while the other, a low vowel such

as /a/. The listeners judged the word containing the high vowel to be less prominent even though the F_0 peak in each of the two words was at the same height. This occurred due to the vowel intrinsic pitch effect. Vowels, even in unaccented syllables, are associated with variations in the fundamental frequency contour; this is called the *intrinsic* F_0 of vowels. Intrinsic F_0 is positively correlated with vowel height; other things being equal, a higher vowel has a higher intrinsic F_0 than a lower vowel. The reason that listeners in the Silverman study assessed the same F_0 peak as less prominent when produced with the high vowel than when produced with the low vowel was because they perceptually factored out the intrinsic F_0 of each vowel. This perceptual phenomenon is called the vowel intrinsic pitch effect. Factoring out microprosodic variations from the F_0 contour during F_0 curve decomposition thus simulates the vowel intrinsic pitch effect in perception.

The second reason for modeling microprosody is that microprosodic variations pose a challenge for prosodic signal modification methods (van Santen et al. 2000). For example, *creaking* causes pitch halving. Simple smoothing and interpolation (two techniques often used to deal with microprosodic effects) would still leave behind an uncharacteristic bulge in the F_0 contour, which would require such severe prosodic modification during synthesis that degradation in the quality of the resultant speech would be inevitable. Van Santen suggests microprosodic modeling as a solution to the problem of speech quality degradation due to prosodic modification (van Santen et al. 2000). He suggests developing an explicit model of microprosody to separate the natural F_0 contour into a "residuals" contour containing the microprosodic variations and an underlying smooth contour that contains no unusual artifacts. During synthesis, the target contour would replace the smooth natural contour, and the same time-warping operations could be applied to the spectral representation and the residuals contour. Thus, prosodic signal modification is minimized, and degradation in the quality of synthesized speech is minimized as an ultimate result.

The third reason for modeling microprosody is that it has been found that microprosodic effects serve as important cues to segmental identity (Kohler 1990, Silverman 1987, Silverman 1990) and contribute to the perception of "naturalness" of intonation (Silverman 1987). The lack of microprosodic modeling is perhaps more easily perceived in the presence of constant improvement in the quality in present day synthesizers, even

if it could be ignored in the early days of single-pulse LPC synthesis.

Given the aforementioned three reasons for microprosodic modeling, it seems likely that proper modeling of microprosodic variations can improve intonation analysis and synthesis. To drive towards this goal, an important type of microprosodic variation, namely, segmental perturbations associated with post-obstruent vowels, is explicitly modeled in our decomposition algorithm. (The other types of microprosodic variations are not insignificant but it is beyond the scope of this thesis to model all of them.) The additivity assumption of the general superpositional approach underlying our decomposition algorithm can easily accommodate an explicit model of segmental perturbations. This is a key point of departure of our decomposition algorithm from the decomposition algorithm presented in Sakai (2004). Sakai's algorithm is also based on the more general principles of the superpositional approach, but it does not exploit the potential of the general superpositional assumptions to model segmental perturbations.

3.3 Decomposition of F_0 contours using PRISM

To overcome the aforementioned limitations of existing decomposition algorithms, we have developed a new algorithm called PRISM (Procedure for Representing Intonation in the Superpositional Model) to decompose given F_0 contours into component curves using the superpositional approach. A detailed description of the PRISM algorithm, including the assumptions underlying the algorithm, the implementation of the algorithm, and the tests used to evaluate the performance of the algorithm, is presented in this section.

3.3.1 Assumptions underlying the decomposition algorithm

PRISM is based on four assumptions. The first two are the core assumptions of the general superpositional approach. The next two are judiciously developed problem-specific assumptions that help regularize the ill-posed problem of F_0 contour decomposition.

1. **Additive decomposition of the F_0 contour**: Additive decomposition is the central assumption of the general superpositional approach (van Santen and Möbius

1999, van Santen et al. 2004:ii) and, consequently, of the PRISM algorithm. According to this assumption, an observed F_0 contour is obtained by "generalized addition" of various component curves of different classes, as shown in Equation 3.1.

$$F_0(t) = \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t) \tag{3.1}$$

where C is a set of curve classes (e.g., {phrase, accent, perturbation}), c is a particular curve class, and k is a particular curve from a particular class. The operator $\bigoplus$ represents an addition-like or multiplication-like function, in that it satisfies the properties of monotonicity (if $a \geq b$ then $a \bigoplus x \geq b \bigoplus x$) and commutativity ($a \bigoplus b = b \bigoplus a$). The reason the operator $\bigoplus$ is used in Equation 3.1 instead of an addition operator like $\sum$ is because decomposition can be performed either in the log domain or the linear domain. In our decomposition algorithm, C = {phrase, accent, perturbation} and $\bigoplus$ is addition in the linear domain.

Based on the argument that the perception of frequency is logarithmically-based (Silverman 1987, 't Hart and Cohen 1973), it might be objected that decomposition in the logarithmic domain would be more appropriate. However, we can find little consensus in literature regarding the most appropriate domain for representing intonation. In fact, several studies have shown that frequencies below 1 kHz — which we are mainly concerned with since the average F_0 is 125 Hz for males and 200 Hz for females, and the maximum F_0 would not go much above 700 Hz — are perceived linearly. The shape of the Mel scale (Stevens and Volkmann 1940), a psychoacoustic scale developed to measure the perception of pitch, is approximately linear below 1 kHz and logarithmic above. Since the Mel scale, other psychoacoustic scales such as the Bark scale (Zwicker 1961) and the ERB scale (Hermes and van Gestel 1991) have also been proposed. Both these scales suggest that for frequencies below 500 Hz a linear scale is more appropriate. Also, in an experiment set up to find out whether prominence-lending pitch movements should be expressed in Hz or semitones, Rietveld and Gussenhoven (1985) concluded that prominence judgements were in better agreement with a Hz scale than with a scale of semitones. However,

regardless of whether the ideal scale for representing intonation is logarithmic, linear or otherwise; transforming from a linear to another numerical domain is not essential to the idea of superposition and therefore, in practice, not required for our decomposition algorithm to work.

2. **Different temporal scopes of the component curves**: Another key assumption of the general superpositional approach is that each of the component curves is tied to a distinct phonological entity and follows a distinct time course (van Santen 2002). Based on this assumption, in the PRISM algorithm, it is assumed that (1) the phrase curve is tied to the *intonational phrase*[1] and spans the phrase length, (2) the accent curve is tied to a distinct left-headed foot and is left-aligned with the start of the foot (based on van Santen (2002) which showed that pitch accents are associated with feet and not accented syllables), and (3) the segmental perturbation curves are associated with initial parts of vowels following a transition from an obstruent (as shown by van Santen and Hirschberg (1994)).

3. **Smooth unconstrained phrase curve**: The shape of the phrase curve is assumed to be essentially unconstrained except for the broad assumption that it should be smooth over long stretches of time. Arguments in favor of this constraint can be based on quasi-physical concepts similar to those that have been mentioned in support of the Fujisaki model (Fujisaki 1983) and the Stem-ML model (Kochanski and Shih 2003). This assumption about the shape of the phrase curve is a departure from standard assumptions about the shape of the phrase curve such as the oft-used linear declination line, or the log-timescale exponential decay of the Fujisaki model. This assumption allows the possibility of phrase curves displaying considerable and meaningful variability (van Santen 2002). For example, one may be able to account for a plateau-like F_0 contour ("hat pattern") that spans two feet by positing a phrase curve that bulges upward and downward somewhat more strongly than allowed in the Fujisaki model, and positioning two accent curves that overlap in time at the

[1] An intonational phrase consists of one or more phonological phrases, where a phonological phrase is any syntactically determined phrase structure, e.g., a noun, verb or prepositional phrase. Phonological phrases mark the "locations of optional or obligatory intonation breaks" (Kiparsky 1975)

rise and fall locations of the phrase curve.

4. **Component curves from common templates**: Shapes of component curves are unspecified in the general superpositional approach. However, a corollary of Assumption 2 is that component curves can be classified in terms of the phonological entities they span and their time courses. Based on this idea of curve classes, van Santen and Möbius conceptualized the generation of accent curves from class-specific templates (e.g., declarative template, continuation rise template, interrogative template) via parameterized time-warp functions (van Santen 2002). This template-based generation of component curves is a promising alternative to overly strict or overly lenient shape constraints described in Section 3.2.1. Thus, in the PRISM algorithm, it is assumed that for every estimated accent curve A_i, there exists a template curve E_i (shared with similarly-shaped accent curves) of norm 1, such that A_i equals a scaled (by a height factor, h_i) and time-warped E_i (Equation 3.2).

$$A_i(t) = h_i \times time\_warped(E_i(t)), \qquad |E_i(t)| = 1 \tag{3.2}$$

Based on van Santen and Möbius (1997), the accent curve template corresponding to a declarative utterance is assumed to be a single-peaked curve that rises from 0 to a peak value and again descends to 0 (implemented by a Gaussian curve[2]); the template corresponding to a continuation rise consists of a dual motion in which an early peak is followed by a valley and a final rise (implemented by the summation of a Gaussian curve and a rising exponential curve); and, the template for the interrogative contour is a curve that rises steeply in the nucleus of the accented syllable (implemented by a rising exponential curve). To address the idea of overlapping accent curves, it is also assumed that an accent curve starts at the beginning of the associated left-headed foot and finishes at *or beyond* the end of the foot but, conservatively, no later than the point where the next accent curve reaches its peak.

[2]Note: Exactly which particular function is chosen to implement a class-specific template is not very important. What is important is that the function used to implement a class-specific template adheres to the general shape constraints of the template. For example, instead of a Gaussian curve, a cosine curve can be used just as well to implement the declarative accent curve template. Regardless of the curve used, the time-warping procedure renders templates with similar shape constraints equivalent.

Similar to accent curves, segmental perturbation curves are also assumed to be generated from class-specific templates. The template corresponding to segmental perturbations associated with post-obstruent vowels — the specific type of segmental perturbation that is modeled in this work — is assumed to be a decay function, which is implemented by a negative exponential curve. This assumption is guided by the research presented in van Santen and Möbius (1997).

The template-based generation of the phrase curve is more complicated. Notice that in Assumption 3, we specified a smooth unconstrained phrase curve. Hence, we cannot directly specify a class-specific template of the phrase curve. Thus, in the two-phase implementation of PRISM (described in Sections 3.3.2 and 3.3.4), we first obtain an initial estimate of the phrase curve, Phr_{init_est}, using the wavelet decomposition phase, and then use Phr_{init_est} as the phrase curve template in the template decomposition phase.

3.3.2 Implementation of PRISM: Wavelet Decomposition

We have implemented PRISM as a two-phase decomposition algorithm. The first phase of the algorithm is the wavelet decomposition phase. In this phase, a partial decomposition of the F_0 contour into a phrase curve and a *combined* accent and segmental perturbation curve, as shown in Equation 3.3, is performed. $P(t)$ denotes phrase curve, $A(t)$ denotes accent curve, $S(t)$ denotes segmental perturbation curve, $R(t)$ denotes the summation of the accent and segmental perturbation curves, and the superscripts m and n denote that there can be several accent curves and segmental perturbation curves.

$$F_0(t) = P(t) + R(t), \quad where\ R(t) = \Sigma_i^m A_i(t) + \Sigma_i^n S_i(t) \tag{3.3}$$

An obvious method for recovering the phrase curve would be to determine local minima and draw a smooth curve through these points (as in van Santen 2002). However, this fails when accent curves overlap in time, because overlapping accent curves cause all points between successive pitch accents to be above, and not on, the phrase curve. Another equally obvious solution is the *filtering method* that consists of computing the Fourier transform, removing the higher frequency components that presumably correspond

to the accent curves (and possibly the segmental perturbation curves), and computing the inverse transform to generate an estimate of the phrase curve. The linearity of this method would seem to be particularly appropriate given the quasi-linearity of the superpositional approach. And indeed, this method is used as an important component in a Fujisaki parameter estimation system developed by Mixdorff (2002) and was proposed earlier by Sakurai and Hirose (1996). Our experiments showed, however, that this method provided undesirable results. Depending on the frequency cutoff, the result was either a smooth but poorly fitting phrase contour, or a "wobbly" phrase contour containing the negative lobes of the frequency components that carried the accent curves. In fact, this observation was also made by Sakurai and Hirose (1996), though their work on phrase boundary detection was not hampered by this less than optimal phrase curve shape. As an aside, applying Mixdorff's Fujisaki parameter estimation system to curves generated by a simplified version of the Linear Alignment Model yielded inaccurate estimates of the phrase curves (van Santen et al. 2004:ii).

Thus, as an alternative to the filtering method and the local minima based method of obtaining the phrase curve, we developed a wavelet-based decomposition method. Wavelets share with filtering the features of linearity and frequency specificity. The key difference is the temporal "locality" of the wavelet transform. That is, like the Fourier transform, the wavelet transform of an input signal (here, the F_0 curve) represents the signal as a weighted sum of basis functions. However, these basis functions consist of dilations (time scaling) and translations (moving along the time axis) of the wavelet function. Thus, unlike the Fourier transform, the behavior of same-sized (or, equivalently, same-frequency) basis functions at different locations is independent. This basic feature might be critical in addressing the "wobbly" phrase contour problem encountered with simple filtering.

Description of the wavelet decomposition algorithm

We have implemented an algorithm for the partial decomposition of the F_0 contour into a phrase curve and a combined accent and perturbation curve. Assumptions 1, 2, and 3 outlined in Section 3.3.1 are used by this algorithm. The algorithm comprises five steps

(shown in Figure 3.1) that are described below:

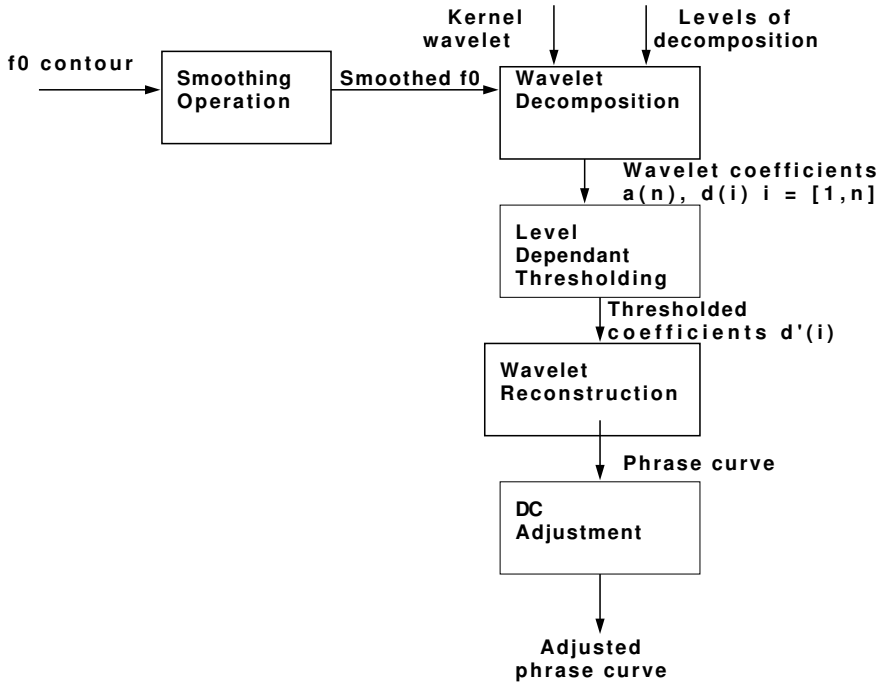

Figure 3.1: Wavelet decomposition algorithm to extract the phrase curve.

1. **Smoothing the F_0 contour**: Remove high frequency values, while preserving the global shape of the F_0 contour. Based on this requirement, the Savitzky-Golay (Savitzky and Golay 1964) filter of order 3 and length 5 is used for smoothing the F_0 contour. The Savitzky-Golay filter determines the smoothed value of each data point by performing a local polynomial regression. This filter has been found to be good in preserving pertinent high-frequency information (Orfanidis 1996), such as peak height and width, which are often washed out by other standard smoothing filters such as an n-sample average or an order-n Butterworth low-pass filter.

2. **Wavelet decomposition**: Discrete wavelet transform (DWT) is used to decompose the smoothed F_0 contour. DWT requires the choice of a kernel wavelet $\psi(t)$ and level of decomposition n. Which kernel wavelet should be chosen for a given analysis task is a perplexing question. The most appropriate kernel is one that minimizes

reconstructed error variance and maximizes the signal to noise ratio (SNR). Keeping this objective in mind, we systematically experimented with several different kernels and obtained a set of kernel wavelets that gave minimal reconstructed error and maximized SNR. This set of possible kernels contained both short and long *support* wavelets. Support refers to the length of the kernel wavelet. Longer support wavelets detect slow long term changes well, but they may smooth over local anomalies, whereas short support wavelets are highly sensitive to local variations but may be unable to encode the higher-order polynomial behavior of a highly smooth phrase curve. For the decomposition algorithm, two wavelets were selected: Daubechies db3, a short support wavelet, for detecting phrase curves with sharp inflections, and db10, a long support wavelet, for detecting higher-order smooth phrase curves that changed slowly over time. Admittedly, other wavelets could have been considered, such as wavelets specifically designed to optimally filter accent curves of a certain shape; however, it was felt that this would reduce the generality of the method.

The level of decomposition chosen for the DWT usually depends on the signal being analyzed and the outcome desired. For our purposes, the total number of levels best suited for the purpose of extracting the phrase curve from the F_0 contour was empirically established to be $log_2(n) + 1$, where n is the length of the signal in 10-millisecond units.

Every kernel wavelet $\psi(t)$ is associated with a dilation function $\phi(t)$ that is orthogonal to it. The relationship between $\psi(t)$ and $\phi(t)$ is shown by the following equations:

$$\phi(t) = \sum_k c(k)\phi(2t - k) \tag{3.4}$$

$$\psi(t) = \sum_k (-1)^k c(1 - k)\phi(2t - k) \tag{3.5}$$

Essentially, Equation 3.4 represents a low-pass filter that defines the low frequency information called the *approximation coefficients a* and Equation 3.5 represents a high-pass filter that defines the high frequency information called the *detail coefficients d* (Mallat 1989). This is the basic cell for one-level wavelet decomposition as

shown in Figure 3.2.

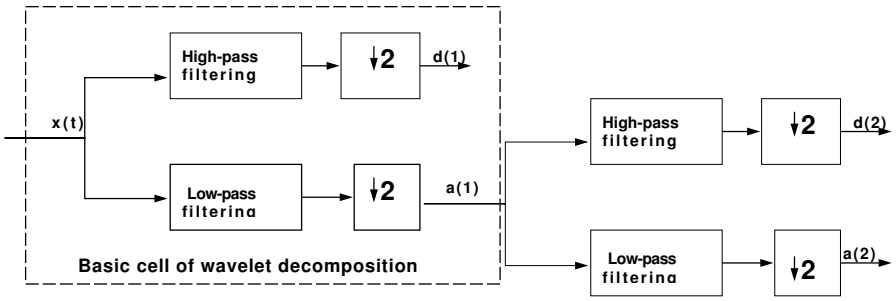

Figure 3.2: Two-scale filtering operation of the DWT (Mallat 1989).

For multi-level decomposition, the basic cell is iterated over the downsampled output of the low-pass filtering operation. In our case, the basic cell is iterated n times, where n is the maximum number of levels of decomposition, to obtain one set of approximation coefficients $a(n)$ and n sets of detail coefficients $d(i)$, where $i = [1, n]$ levels. The reason why there is only one set of approximation coefficients is that at the m-th iteration of the basic cell, the $a(m-1)$ set of approximation coefficients is decomposed (see Figure 3.2).

3. **Level-dependent thresholding:** Extracting the phrase curve from the F_0 contour implies *removing* the combined accent and perturbation curves to unveil the underlying phrase curve. Given that earlier results indicated that simply removing higher-frequency components (as in low-pass filtering) does not work well, a different technique for modifying the wavelet transform was explored. Specifically, a technique called level-dependent thresholding (Hall et al. 1997, Härdle et al. 1998) was used. This technique involves discarding certain detail coefficients that fall below a threshold $t(m)$ at each level of decomposition m. The threshold $t(m)$ is not a global threshold; rather, it is computed at each level of decomposition.

   A form of the aforementioned technique that is represented by the following rules was used. Given level m,

   - if $mean(abs(idwt(a, d(m), db3)) < mean(F_0\ contour)$, set all detail coefficients

$d(m) = 0$, where a is the approximation coefficient, 'db3' is the kernel wavelet and $idwt$ is the inverse discrete wavelet transform (more on this in 'Wavelet reconstruction').

- if $mean(abs(idwt(a, d(m), db3)) \geq mean(F_0\ contour)$, set only certain details coefficients $\hat{d}(m) = 0$, where $abs(\hat{\mathrm{d}}(\mathrm{m})) < median(abs(d(m))$.

There are two principles implicit in these rules:

(a) Zero out all coefficients that might contain information about accent curves or segmental perturbations, thus ensuring that we capture the right phrase curve in spite of aliased local minima resulting from overlapping accent curves.

(b) Of the remaining detail coefficients, truncate all coefficients that have magnitudes smaller than the median magnitude, thus retaining information about any sharp inflections in the phrase curve.

Obviously, these rules are a mixture of theory and experimentation. The rule regarding which level of detail is set to zero was discovered empirically, while the rule to remove detail coefficients whose magnitude is less than zero has been found in well-established wavelet theory (Donoho and Johnstone 1994).

4. **Wavelet reconstruction**: Once the detail coefficients have been manipulated to obtain a new set of detail coefficients $d'(i)$ where $i = [1, n]$ levels, the inverse discrete wavelet transform (IDWT) is used to reconstruct the signal. IDWT involves the use of reconstruction filters that are the inverse of the decomposition filters shown in Figure 3.2. The reconstruction yields the phrase curve from the F_0 contour.

5. **DC adjustment**: The estimated phrase curve obtained is often found to have a 5-10 Hz shift along the y-axis. In order to fit it appropriately to the F_0 contour, the phrase curve is adjusted by a value DC, such that $DC = mean(x1$ and $x2)$, where $x1$ is the difference in the means of the estimated phrase curve and the original F_0 contour, and $x2$ is the difference in the values of the end-points of the estimated phrase curve and the original F_0 contour. This method of DC adjustment is only applicable for simple declarative utterances that end in voiced phonemes and needs

to be revised in case of more complex utterances such as those that either contain continuation rises or interrogatives, or end in unvoiced phonemes.

We observed that the vertical shift in the phrase curve occurred more often in higher-order smooth phrase curves that were generated using the Fujisaki model. From this we conjecture that the vertical shift occurred when some of the detail coefficients related to the vertical height of the phrase curves were (unintentionally) discarded in the thresholding step in order to estimate the smooth phrase curves.

This step, admittedly, is somewhat ad hoc. However, this step is eliminated from the wavelet decomposition phase in the linked PRISM algorithm (described in Section 3.3.6) because there the output of the wavelet decomposition phase is used as an initial estimate (or template) of the phrase curve, Phr_{init_est}, that is fed to the template decomposition phase of the linked PRISM algorithm. Since it is a template, Phr_{init_est} has to match the underlying "true" phrase curve only in terms of the general shape; consequently, the vertical shift is unimportant and no DC adjustment is required.

3.3.3 Objective evaluation of the wavelet decomposition algorithm

To test the performance of the wavelet decomposition algorithm, a set of 56 declarative F_0 contours was synthesized and the wavelet decomposition algorithm was used to extract the phrase curve from each of the contours. Two examples of phrase curves extracted by the algorithm can be seen in Figure 3.4. Of the 56 F_0 contours, 28 were based on the Fujisaki intonation model and 28 were *loosely based* on the Bell Labs intonation model. (The qualifier *loosely based* is used because we have avoided using the Bell Labs patented accent peak placement mechanism. Also, the use of perturbation curve templates has been ignored for the purposes of this experiment. Nevertheless, the superpositional quality of the intonation model remains intact.)

These F_0 contours had one phrase component and 1-3 accent components. The accent components were systematically varied in height and placement (as shown in Table 3.3) to create F_0 curves with one, two, and three accents. The single phrase component was

also varied in its slope, while keeping its mean magnitude roughly the same.

Other than the accent parameters, the other Fujisaki parameters (Fujisaki 1983) that were used in constructing the F_0 contours were defined as follows: $\alpha = 2$, $\beta = 20$ and $\gamma = 0.9$. The minimum frequency Fb was set at 70 Hz. Two different values for the phrase command magnitude Ap were used, namely, 0.3 and 0.5, to get two different slopes.

For the Bell Labs-inspired model, besides the variation in the accent parameters, the only other variation was in the magnitude and slope of the phrase curve. The phrase curve was anchored at three points: the beginning of the phrase t_{start}, the end of the phrase t_{end}, and the beginning of the last set of consecutively occurring accented syllables $t_{nucleus}$ (for more information see van Santen and Hirschberg 1994). In order to keep the mean magnitude more or less constant but vary the slope of the phrase curve, the height of the phrase curve at t_{start}, $t_{nucleus}$, and t_{end} was varied by two sets of values, namely (110 Hz, 90 Hz, 70 Hz) and (110 Hz, 90 Hz, 50 Hz).

$2x$	x	$1.5x$
✓	✓	✓
✓	✓	
✓		✓
✓		
	✓	✓
	✓	
		✓

Figure 3.3: Accent height and placement table. $x = 10, 20$ for the Bell Labs inspired model; $x = 0.225, 0.45$ for the Fujisaki model. Check marks indicate accent curves present.

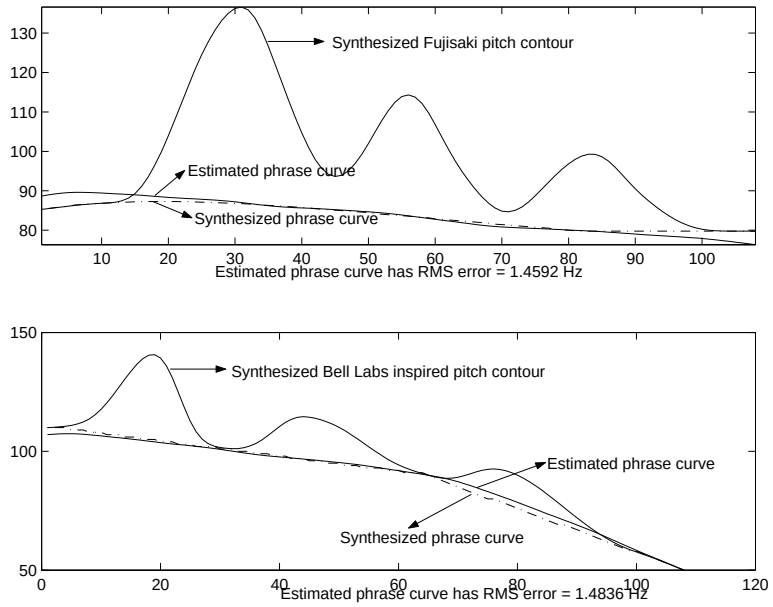

Figure 3.4: Phrase curves estimated by wavelet decomposition.

Results of phrase curve extraction using wavelet decomposition

To measure the performance of the algorithm, the Root Mean Squared Error (RMSE) between the estimated phrase curve and the known synthesized phrase curve was measured. The results are shown in Table 3.1.

	Bell Labs inspired model	Fujisaki model
Mean RMSE (Hz)	3.1536	2.2688
Standard Deviation (Hz)	1.7614	1.0334

Table 3.1: Performance of the wavelet decomposition algorithm.

To test the limits of this wavelet decomposition algorithm, we also compiled the results obtained by making the optimal DC adjustment. The optimal DC adjustment is the difference between the mean of the estimated phrase curve and the mean of the synthesized phrase curve. The results are shown in Table 3.2.

	Bell Labs inspired model	Fujisaki model
Mean RMSE (Hz)	2.3723	1.9789
Standard Deviation (Hz)	1.2034	0.9734

Table 3.2: Performance of the wavelet decomposition algorithm with optimal DC adjustment.

Discussion of results

The results of this objective test are encouraging, especially in light of the simplicity of the method. However, the curves analyzed were exceedingly simple smooth synthetic curves, representing single-phrase utterances with up to three pitch accents. It is unlikely that this algorithm all by itself will be able to reliably extract the phrase curves of more complicated natural F_0 contours containing (1) gaps due to the presence of stops and fricatives, or (2) yes/no-question rises or (3) continuation rises. Nevertheless, this wavelet-based partial decomposition algorithm is certainly a useful first step of the multi-step decomposition algorithm that is PRISM. In PRISM, decomposition of more complex F_0 contours containing gaps, yes/no-question rises, continuation rises, or segmental perturbations requires template decomposition.

3.3.4 Implementation of PRISM: Template decomposition

Template decomposition is the second phase of the PRISM implementation. The core concept of template decomposition is that the component curves contained in any F_0 contour can be estimated by applying parameterized time-warp functions to class-specific templates (van Santen 2002).

In our implementation of this algorithm, accent curves are generated by parameterized time-warp functions applied to three different templates: the template corresponding to a declarative utterance is assumed to be a single-peaked curve that rises from 0 to a peak value and again descends to 0 (implemented by a Gaussian curve); the template corresponding to a continuation rise consists of a dual motion in which an early peak is followed by a valley and a final rise (implemented by the summation of a Gaussian curve and a rising exponential curve); and the template for the interrogative contour is a curve that rises steeply in the nucleus of the accented syllable (implemented by a rising exponential curve).

Segmental perturbations associated with post-obstruent vowels are the specific class of segmental perturbations that we are considering in this thesis. The template corresponding to a segmental perturbation curve of this type is a fast decaying function that spans the length of a post-obstruent vowel; it is implemented by a negative exponential function in our algorithm.

The template for a phrase curve is the initial estimate of the phrase curve that was obtained by the wavelet decomposition phase of PRISM. However, to observe the performance of the template decomposition phase confounded by the wavelet decomposition phase, we have modified Assumption 3, the assumption about the unconstrained smooth phrase, so that the phrase curve is assumed to be piecewise linear. Under the modified assumption, the phrase curve can be approximated by n line segments p_i, where n is the number of feet in the phrase (Equation 3.6). Each p_i spans the length of a foot. The points at which the phrase curve changes direction are called *inflection points*.

$$p_i(t) = \beta_i(t) + \gamma_i, \qquad P = concatenate(p_i) \tag{3.6}$$

This assumption is amended once the correctness of the template decomposition phase of

the PRISM implementation has been established.

Description of the template decomposition algorithm

We have implemented a template decomposition algorithm to decompose given F_0 contours into phrase curves, accent curves, and segmental perturbation curves. Assumptions 1, 2, and 4 outlined in Section 3.3.1, and the modification of Assumption 3 (defined by Equation 3.6) underlie this algorithm. Another assumption that is held within the context of this algorithm is that the locations of the left-headed feet are known[3]. The algorithm consists of the following three steps (shown in Figure 3.5).

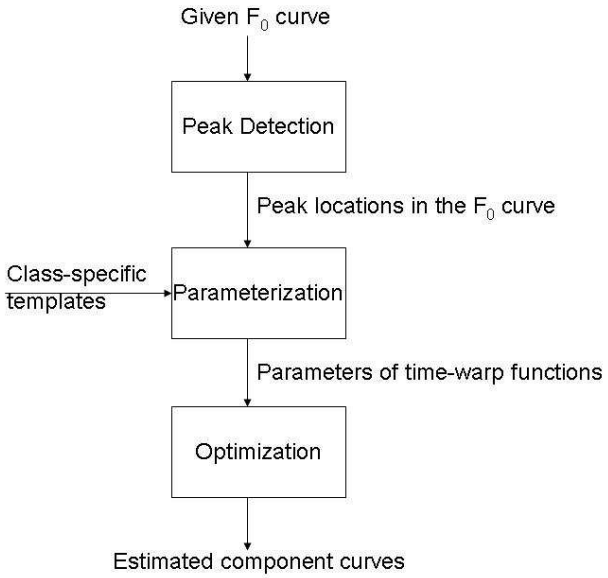

Figure 3.5: Template decomposition algorithm to extract component curves.

1. **Peak detection**: The first step is the detection of the peak associated with each accent curve in a given F_0 contour. Intuitively, it may seem like a straightforward

[3]Assuming the availability of labeled data is a prerogative of speech synthesis, where a fully labeled and segmented speech corpora is easily available. Automatic labeling of large corpora remains a problem due to inaccurate prediction of pitch accents. However, it is beyond the scope of this thesis to address this problem.

process to find where the maximum (or minimum) of an accent curve occurs. However, the presence of segmental perturbations and continuation rises makes the data noisy and automatic detection of peaks or valleys based on this intuitive concept error-prone. Standard smoothing algorithms do not produce the desired result either, because accent curves smoothed by such methods still contain uncharacteristic bulges where the segmental perturbations were located, and the continuation rises remain unchanged.

Exploiting the idea that each accent curve has exactly *one* true peak (or valley) in it, we use *unimodal smoothing* (Härdle 1990) to solve the peak detection problem. Within the context of peak detection, unimodal smoothing is defined as follows. Given an accent curve $A_i \in \mathbb{R}^n$ and a weight vector $w \in \mathbb{R}^n$ (the product of the ESPS *get_f0* (Talkin 1993) voicing flag and energy), the goal is to find a smoothed version of the accent curve, $\hat{A}_i \in \mathbb{R}^n$, that minimizes $E = \sum w_i(\hat{A}_i - A_i)^2$ subject to two constraints. Given that $\hat{A}_i$ has a mode at $i = p$, where p is the location of the true peak of the accent curve A_i, then

$$\forall(i,j). \quad i < j < p \quad \Rightarrow \quad \hat{A}_i \leq \hat{A}_j \leq \hat{A}_p \tag{3.7}$$

$$\forall(k,l). \quad p < k < l \quad \Rightarrow \quad \hat{A}_p \geq \hat{A}_k \geq \hat{A}_l \tag{3.8}$$

The function $f : A_i \mapsto \hat{A}_i$ is greatly dependent on the location of segmental perturbations, continuation rises, and gaps in the F_0 contour, and it differs greatly from one accent curve to another. Thus, function f cannot be prespecified, and parametric unimodal smoothing approaches cannot be used. Consequently, a nonparametric approach called *isotonic regression* (Robertson et al. 1988, Barlow et al. 1972, Wegman 1980) was used to predict $\hat{A}_i$. Isotonic regression is regression in which the only assumption about the function relating the predictor variable (A_i) and response variable ($\hat{A}_i$) is that it is isotonic, i.e., either *always* non-decreasing or *always* non-increasing.

However, the unimodality constraints defined above are not isotonic because

some of the constraints are non-decreasing (inequality 3.7) while others are non-increasing (inequality 3.8). But because unimodal constraints can easily be transformed into isotonic ones, isotonic regression can be used to perform unimodal smoothing. Consider the constraints specified by inequality 3.8. From this inequality, it follows that

$$\forall(k,l). \quad p < k < l \quad \Rightarrow \quad -\hat{A}_p \leq -\hat{A}_k \leq -\hat{A}_l \tag{3.9}$$

$$\Rightarrow \quad 2(\hat{A}_p) - \hat{A}_p \leq 2(\hat{A}_p) - \hat{A}_k \leq 2(\hat{A}_p) - \hat{A}_l \tag{3.10}$$

$$\Rightarrow \quad \hat{A}_p \leq 2(\hat{A}_p) - \hat{A}_k \leq 2(\hat{A}_p) - \hat{A}_l \tag{3.11}$$

The inequalities 3.7 and 3.11 together represent a set of isotonic constraints on $Pf(\hat{A}_i)$, where $Pf(x_i)$ is defined as:

$$Pf(x_i) = \begin{cases} 2x_p - x_i & \text{if } i \geq p \\ x_i & \text{if } i < p \end{cases}$$

$Pf(x_i)$ is a simple linear transformation of x_i that reflects the observations beyond the peak location p about the horizontal line passing through p. $Pf(x_i)$ has an inverse, $Pf^{-1}(x_i)$.

Thus, peak detection by unimodal smoothing via isotonic regression involves the following five steps:

(a) A peak location p is posited.

(b) All points on the accent curve beyond the peak location p are reflected about the horizontal line passing through p to obtain $Pf(A_i)$.

(c) The monotone smooth $Pf(\hat{A}_i)$ that minimizes $\sum w_i(Pf(\hat{A}_i) - Pf(A_i))^2$ subject to monotonicity constraints is predicted by isotonic regression.[4]

(d) $\hat{A}_i$ is obtained by applying the $Pf^{-1}(x_i)$ to $Pf(\hat{A}_i)$. The root weighted mean

[4]The isotonic regression was implemented using the pool adjacent violators algorithm (PAVA) (Barlow et al. 1972). PAVA is a simple iterative algorithm that checks a given sequence of values for monotonicity. If consecutive values that violate monotonicity constraints are found, the values are 'pooled' together, i.e., replaced by their weighted average. The process continues until no violators are left.

square error between $\hat{A}_i$ and A_i, RWMSE$(A_i, \hat{A}_i)$, is calculated, where

$$\text{RWMSE}(X_i, Y_i) = \sqrt{\frac{\sum w_i.(X_i - Y_i)^2}{\sum w_i}} \tag{3.12}$$

(e) The previous four steps are repeated for $p = \{1, 2, ..., n\}$, where n is the length of the accent curve in frames. The p value that yields the least RWMSE$(A_i, \hat{A}_i)$ is considered to be the true peak location and $\hat{A}_p$, the true peak value.

In case the accent curve contains a valley rather than a peak, the same five-step process is used to find it, except that the transformation function used is $-Pf(\hat{A}_i)$. If it is not known ahead of time, whether the accent curve contains a peak or a valley, the five-step process is repeated for each condition, and the condition that yields the lowest RWMSE reveals the true shape of the accent curve.

2. **Parametrization of component curves**: Parameterizing the component curves (i.e., describing the component curves in terms of certain parameters) is the second step of the algorithm. Accent curves are parameterized in two steps. First, for any accent curve A_i, the template E_i, (from which A_i is assumed to be generated) is characterized by n values sampled at equal time points. Then, the accent curve is characterized by a piecewise linear time-warp function consisting of n time points that correspond to the n time points in the template, and a height parameter h_i.

An example of the accent curve parametrization is shown in Figure 3.6. If A_i is a declarative accent curve, the template E_i is characterized by m pitch values that approximate a Gaussian sampled at equal time intervals with the first and last values rounded down to 0, as shown in Figure 3.6a. The piecewise linear time-warp characterizing the shape of A_i is akin to the one shown in Figure 3.6b. Multiplying the height parameter h_i by the time-warped E_i (shown in Figure 3.6c) fully characterizes the estimated accent curve A_i (shown in Figure 3.6d).

Segmental perturbations curves are parameterized in a manner similar to accent curves. First, for any segmental perturbation curve SP_i, the template F_i, (from which SP_i is assumed to be generated) is characterized by k values sampled at equal time points. Then, the segmental perturbation curve is characterized by a piecewise

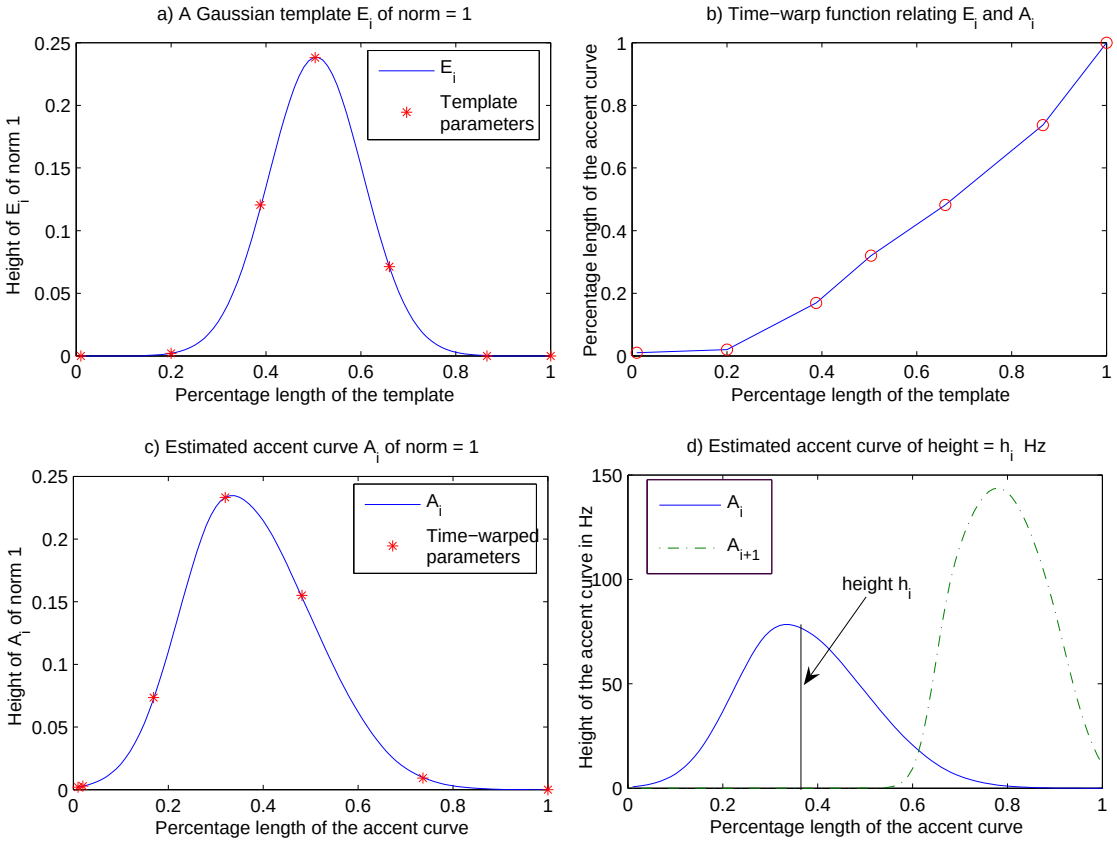

Figure 3.6: Accent curve parametrization.

linear time-warp function consisting of k time points that correspond to the k time points in the template, and a height parameter g_i.

The phrase curve is characterized by m interconnected line segments. Each segment begins at the start of a distinct foot and spans the length of the foot. Because the phase curve is continuous and because the location of the feet are known, $m+1$ parameters are needed to describe the interconnected line segments making up the phrase curve. However, depending on the utterance type (e.g., declarative versus interrogative) and the phones used in the utterance, some other assumptions about the phrase curve can be made, which can reduce the number of parameters needed to describe the phrase curve. For example, if an utterance starts with a sonorant phone, it can be assumed that the value of the first point on the related phrase curve

equals the value of the first point on the raw F_0 contour of the utterance[5], and only m parameters are needed to characterize the phrase curve. The same reduction in the number of parameters also applies if the utterance is a declarative utterance that ends with a sonorant phone, because in this case, the value of the last point on the phrase curve can be assumed to be equal to the value of the last point on the raw F_0 contour (see footnote 5). So, if the utterance is declarative, and it starts and ends with a sonorant phone, then only $m-1$ parameters are needed to describe the phrase curve.

3. **Optimization**: Finally, the parameters are estimated using the Nelder-Mead (also known as the down-hill simplex) method (Lagarias et al. 1998) as implemented in the routine *fminsearch* in MATLAB (Matlab 2002). The root weighted mean squared error (RWMSE) between the predicted and the observed F_0 contours is used as the objective function that has to be minimized. The weights are given by the product of the voicing flag and energy. This choice of weights ensures a good fit between the predicted and the observed F_0 contours in the F_0 contour regions with reliable F_0 values, such as regions corresponding to vowels and sonorants.

   The Nelder-Mead method does not require any derivative information. This makes it useful for finding parameters of discontinuous functions, such as an F_0 contour. However, a problem encountered by this optimization method is that it occasionally gets stuck in a rut and the parameters produced are not optimal. Thus, to encourage the optimizer to converge to an optimal solution, judiciously chosen initial values of the parameters are provided. We generate initial guesses for accent curve parametrization in this way: the first time-warp parameter is set equal to the foot start time since it is assumed that accent curves start at the beginning of the foot; the central time-warp parameter is set equal to the location of the peak (or valley) in the accent curve; the last time-warp parameter is set equal to the peak location of the accent curve next to the one under consideration since it is assumed

[5]This is based on the assumption that accent curves have a starting value of zero. If the accent curve is declarative, it also ends at zero.

that an accent curve may finish *beyond* the end of the foot but, conservatively, no later than the point where the next accent curve reaches its peak. The remaining time-warp parameters are evenly spaced between the first, central, and last parameters. The height parameters are initialized to the height of the peak (or valley) found in the corresponding accent curves. The initial guesses for segmental perturbation curve parametrization are generated similarly to accent curve parametrization. The phrase curve parameters are initialized with F_0 values corresponding to the start times of the left-headed feet in the utterance.

3.3.5 Objective evaluations of template-based decomposition algorithm

To evaluate the performance of the template decomposition algorithm, we conducted four objective tests: (1) to test the accuracy with which the algorithm can estimate declarative accent curves that overlap in time; (2) to test the accuracy with which the algorithm can estimate continuation rise accent curves; (3) to test the accuracy with which the algorithm can estimate interrogative accent curves; and (4) to test the accuracy with which the accent curves can estimate segmental perturbation curves. Each of these objective tests and their results are described in the following sections.

Objective Test 1: Estimating overlapping declarative accent curves

In this test, a set of 75 declarative F_0 contours with single-peaked pitch accents was synthesized, and the template decomposition algorithm was used to extract the phrase curve and the accent curves. The F_0 contours were generated using a simplified version of the Linear Alignment model (van Santen and Möbius 1999), henceforth referred to as SLAM. In this model, the F_0 contour is a summation of a phrase curve and n accent curves. The phrase curve is created by linear interpolation over three points: start of the phrase, start of the last foot, and end of the phrase. An accent curve is created by cosine interpolation over the start of the associated foot, the location of the peak in the foot, and the end of the foot. Peak location is a function of foot duration and the number of syllables in the foot.

A set consisting of 32 F_0 contours with two accent curves and 43 F_0 contours with

three accent curves was created. For every contour, the accent curves were asymmetric in shape, and the non-phrase-final accent curves overlapped with the next accent curve. The other specifications of these F_0 contours were as follows:

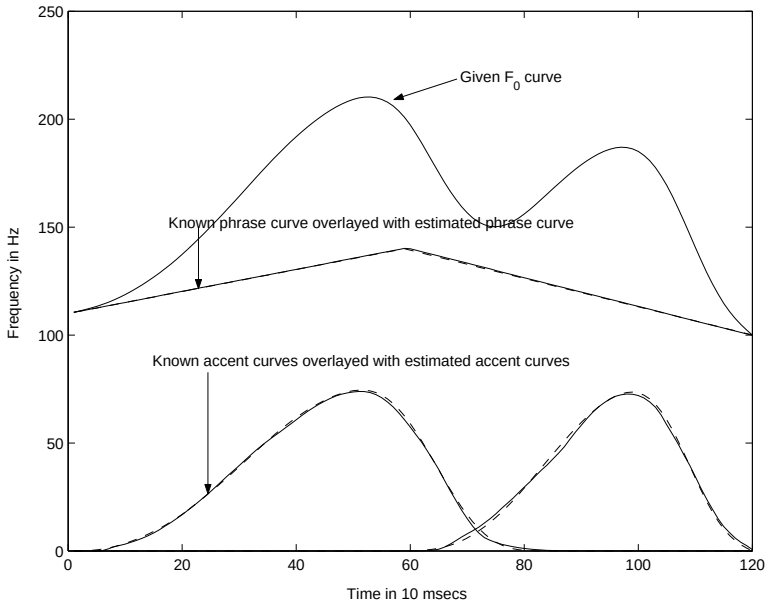

Figure 3.7: Example of a two-accent synthetic F_0 contour used to test the template decomposition algorithm.

1. *Two-accent case*: Accent curve heights were set at 50 Hz and 75 Hz; the slope of the pre-nuclear segment of the phrase curve was set at ±50 Hz/s; the slope of the nuclear segment of the phrase curve was set at ±70 Hz/s; the initial boundary tone was always 110 Hz; and successive accent curve overlap was set at 10% and 20%. An example is shown in Figure 3.7.
2. *Three-accent case*: Similar to the two-accent case, except that accent curve heights were set at 50 Hz, 75 Hz, and 95 Hz; the slopes of the two pre-nuclear segments of the phrase curve were set at ±50 Hz/s and ±70 Hz/s; and the slope of the nuclear segment of the phrase curve was set at ±70 Hz/s. From the 432 possible combinations of such curves, only a random subset (43 out of 432) was used for testing the algorithm. An example is shown in Figure 3.8.

Overlapping accent curves illustrate a key strength of the template decomposition

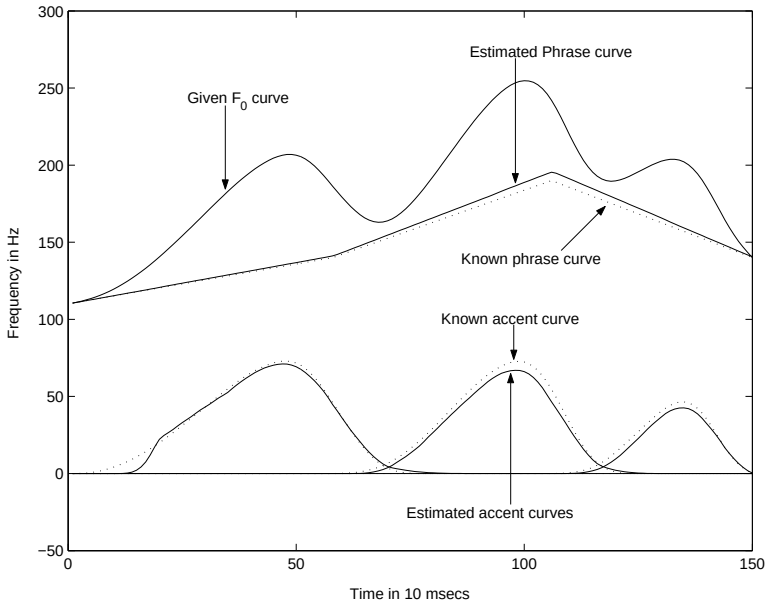

Figure 3.8: Example of a three-accent synthetic F_0 contour used to test the template decomposition algorithm.

algorithm over other decomposition approaches. In the *filtering method* (Sakurai and Hirose 1996, Mixdorff 2002), overlapping accent curves can only be retrieved by assuming that the phrase curve is completely smooth with no possible inflection points, whereas in the wavelet decomposition algorithm described in section 3.3.2, the decomposition is only partial: it returns the phrase curve and *a summation* of the accent and segmental perturbation curves.

Results of objective test 1

The accuracy with which the template decomposition algorithm estimates component curves was measured by the RMSE between the estimated and the known component curves for the 75 instances. The results are shown in Table 3.3.

Objective Test 2: Estimating continuation rise accent curves

In this test, a set of 32 F_0 contours was synthesized using SLAM, and the template decomposition algorithm was used to extract the phrase curve and the accent curves. Each of the F_0 contours consisted of two accent curves. The single-peaked accent curves

	Mean RMSE (Hz) for 2-accent contours	Mean RMSE (Hz) for 3-accent contours
Phrase	4.16	9.11
Accent-1	3.99	4.58
Accent-2	2.21	6.76
Accent-3	N/A	3.20

Table 3.3: Performance of the template decomposition algorithm. Accent-i is the i-th accent curve in the contour.

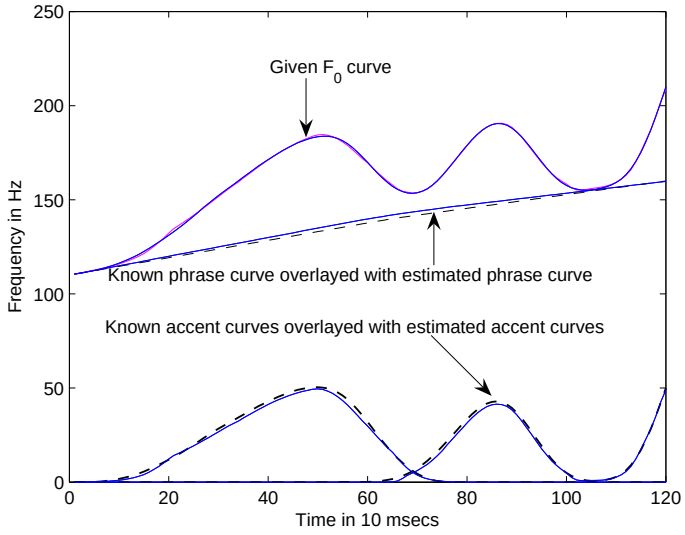

Figure 3.9: Example of a synthetic F_0 curve containing a continuation rise curve used to test the template decomposition algorithm.

were asymmetric in shape, and the non-phrase-final accent curves overlapped with the next accent curve. The first accent curve was a declarative accent curve generated by cosine interpolation over the start of the associated foot, the location of the peak in the foot, and the end of the foot. The second accent curve was an accent curve that corresponded to a continuation rise; it was created by cosine interpolation over four time points: the start of the associated foot, the location of the peak in the foot, a time point 75% into the foot, and the end of the foot. An example of such a curve is shown in Figure 3.9. The shape-related parameters of the phrase curve and the accent curves were set to the same values as in the two-accent case of objective test 1.

Results of objective test 2

The accuracy with which the template decomposition algorithm estimates component curves was measured by the RMSE between the estimated and the known component curves for the 32 instances. The results are shown in Table 3.4.

	Mean RMSE (Hz)	Standard Deviation (Hz)
Phrase	4.61	1.72
Accent-1	3.25	1.50
Accent-2	3.52	0.84

Table 3.4: Performance of the template decomposition algorithm. Accent-i is the i-th accent curve in the contour.

Objective Test 3: Estimating interrogative accent curves

In this test, a set of 32 F_0 contours was synthesized using SLAM, and the template decomposition algorithm was used to extract the phrase curve and the accent curves. Each of the F_0 contours consisted of two accent curves. The first accent curve was a declarative accent curve generated by cosine interpolation over the start of the associated foot, the location of the peak in the foot, and the end of the foot. The second accent curve was an accent curve that corresponded to an interrogative. It was generated by cosine interpolation over three points: the start of the foot, a time point 80% into the foot, and the end of the foot. An example of such a curve is shown in Figure 3.10. The shape-related parameters of the phrase curve and the accent curves were set to the same values as in the two-accent case of objective test 1, though in this case, there was no overlap between the accent curves.

Results of objective test 3

The accuracy with which the template decomposition algorithm estimates component curves was measured by the RMSE between the estimated and the known component curves for the 32 instances. The results are shown in Table 3.5.

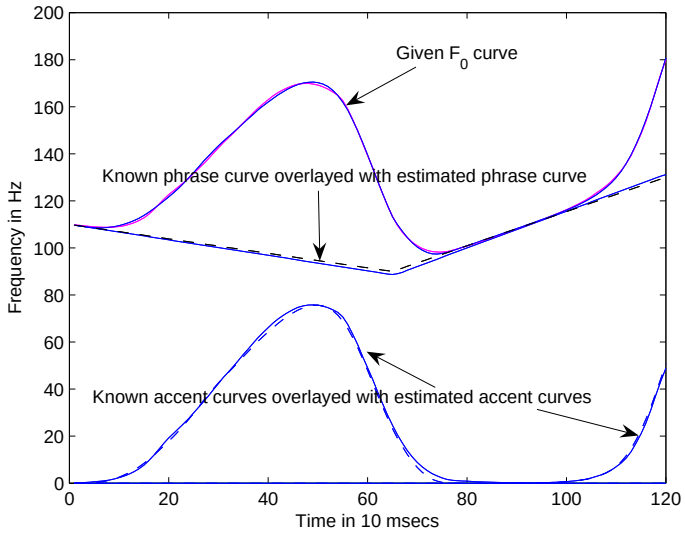

Figure 3.10: Example of a synthetic F_0 curve containing an interrogative curve used to test the template decomposition algorithm.

	Mean RMSE (Hz)	Standard Deviation (Hz)
Phrase	1.68	1.09
Accent-1	1.41	0.60
Accent-2	1.17	1.01

Table 3.5: Performance of the template decomposition algorithm. Accent-i is the i-th accent curve in the contour.

Objective Test 4: Estimating segmental perturbation curves

In this test, a set of 32 F_0 contours was synthesized using SLAM, and the template decomposition algorithm was used to extract the phrase curve, the accent curves, and the segmental perturbation curves. Each of the F_0 contours consisted of two accent curves. The single-peaked accent curves were declarative accent curves generated by cosine interpolation over the start of the associated foot, the location of the peak in the foot, and the end of the foot. The accent curves were asymmetric in shape, and the non-phrase-final accent curves overlapped with the next accent curve. The shape-related parameters of the phrase curve and the accent curves were set to the same values as in the two-accent case of objective test 1. Each of the F_0 contours also contained a segmental perturbation curve that was generated by a negative exponential function. The height, width, and location

of the segmental perturbation curve were assigned randomly. An example of such a curve is shown in Figure 3.11.

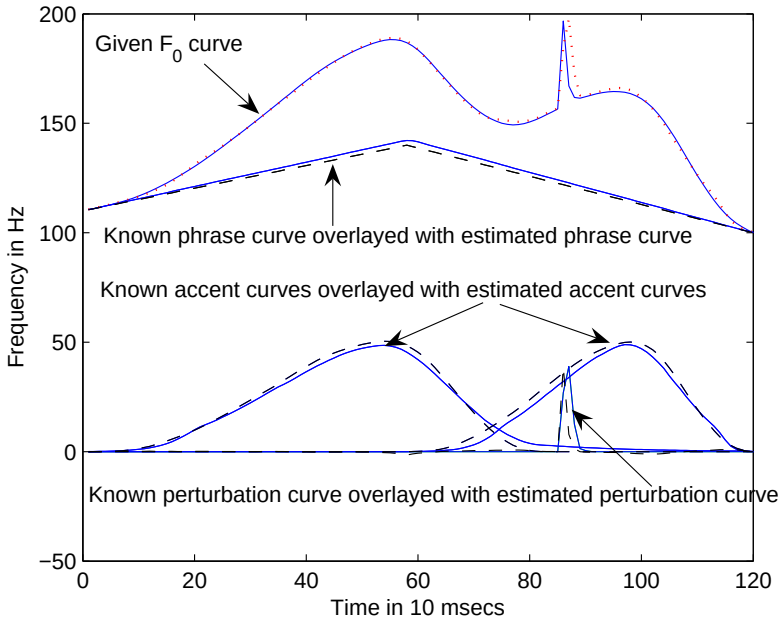

Figure 3.11: Example of a synthetic F_0 contour containing a segmental perturbation curve used to test the template decomposition algorithm.

Results of objective test 4

The accuracy with which the template decomposition algorithm estimates component curves was measured by the RMSE between the estimated and the known component curves for the 32 instances. The results are shown in Table 3.6.

	Mean RMSE (Hz)	Standard Deviation (Hz)
Phrase	3.99	2.3
Accent-1	4.16	1.58
Accent-2	3.47	1.24
Perturbation	3.29	0.33

Table 3.6: Performance of the template decomposition algorithm. Accent-i is the i-th accent curve in the contour.

Discussion of results

The mean RMSE between the known component curves and the estimated component curves obtained from the four aforementioned objective tests ranges from 1 Hz to 9 Hz approximately (a histogram is shown in Figure 3.12). This range of mean RMSE values is well under the frequency difference required for *just noticeable difference* (JND) in pitch for connected speech (illustrated in Figure 3.12).

A study by 't Hart (1981) showed that for connected speech containing complex F_0 contours, people require a frequency difference of at least 1.5-2 semitones (which is a difference of ≈10-12 Hz at 100 Hz) to be able to reliably detect a noticeable difference in pitch. For some people, the frequency difference needed for reliable detection may be as great as 4 semitones (which is a difference of ≈25 Hz at 100 Hz). The experiment by 't Hart was performed using pairs of stimuli with synthesized rising or falling pitch accents that corresponded to four-syllable Dutch number names. Given this stimuli, listeners were asked to judge which of the two accents in each stimulus pair had a larger pitch movement, i.e., F_0 excursion from the declining *baseline* (which loosely corresponds to a phrase curve). To make it unambiguous that subjects were judging the excursion from the baseline and not peak height, members of each stimulus pair were frequently in different "registers" (i.e., of different overall pitch); further, subjects were warned about a certain type of stimulus pair in which one accent had a higher peak while the other had a larger pitch movement. Subject responses to this experiment varied widely; either (1) they were not able to discriminate differences of less than 4 semitones (these were the non-discriminators), or (2) they were able to discriminate differences of about 1.5-2 semitones (these were the discriminators), or (3) they wrongly tried to discriminate between a given stimulus pair in terms of the final pitch value (these were the final pitch discriminators). From this, 't Hart concluded that only differences of more than 3 semitones (which is a difference of ≈20 Hz at 100 Hz) play a part in speech communication. While these JND values for the connected speech are not as well established as for sustained pure tones, these values, however, have clear implications for the accuracy criteria that should be aimed for during the intonation analysis of connected speech.

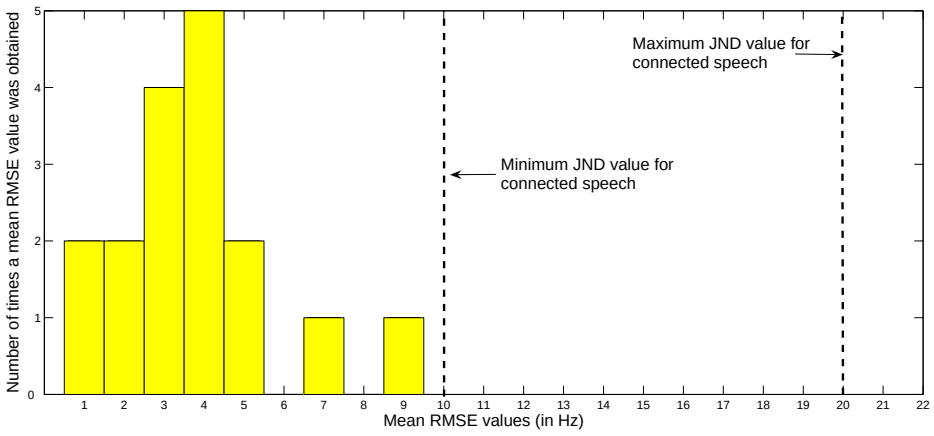

Figure 3.12: Histogram showing the distribution of mean RMSE values obtained from the objective evaluations of the template decomposition algorithm.

The favorable comparison between the range of mean RMSE values obtained from the objective tests and the JND values for connected speech shows that the template decomposition algorithm can recover the underlying component curves with sufficiently high accuracy. Evidence for robustness is provided by the fact that the synthetic accent curves (asymmetric curves cobbled together via cosine interpolation) were different in shape from the templates (Gaussian curves and rising exponential curves). Therefore, extrapolating from these results, we can reason that this algorithm can be used to decompose natural F_0 contours into component curves.

3.3.6 PRISM Implementation: Linking the two phases

Having ascertained by individual objective evaluations that both wavelet decomposition and template decomposition performed the specified partial decomposition subtasks quite well, we linked the two phases together. The link was established as follows: Given an F_0 contour, the wavelet decomposition phase is used to obtain an initial estimate of the smooth, unconstrained phrase curve, Phr_{init_est}. Just as there are templates for the accent curves and segmental perturbation curves, we can assume that Phr_{init_est} is a template for the phrase curve associated with the given F_0 contour. Thus, Phr_{init_est} is given as an input to the template decomposition phase. Because it was observed that the phrase curve

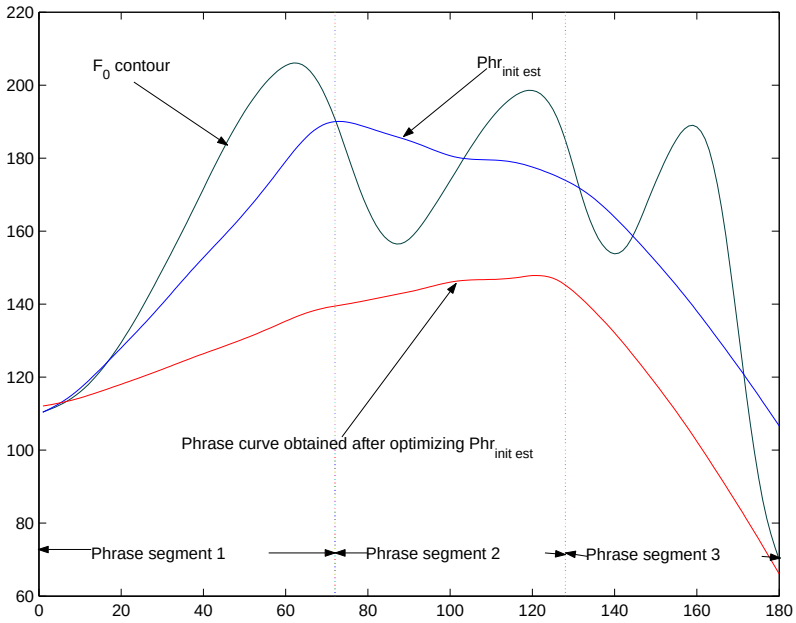

Figure 3.13: Illustrating the concept of linking the wavelet-based algorithm to the template-based algorithm to enable the full decomposition of F_0 contours.

estimate obtained from the wavelet decomposition phase was linearly shifted away from the "true" phrase curve (Section 3.3.2), in the template decomposition phase, a transformation function is applied to the phrase curve estimate in order to remove the linear shift. Hypothesizing that different segments of the phrase curve spanning different feet may have different linear shifts, a piecewise linear transformation is applied to Phr_{init_est} to obtain the final smooth estimated phrase curve, as illustrated in Figure 3.13. The parameters of the piecewise linear transformation function are obtained along with the time-warp parameters needed for describing the accent curves and segmental perturbation curves using the optimization method described in Section 3.3.4.

3.3.7 Objective evaluation of the linked PRISM algorithm

To test the performance of the linked PRISM algorithm, we applied it to the same set of synthetic F_0 contours that were used in objective test 2 of the template decomposition phase of PRISM in Section 3.3.5. Each F_0 contour belonging to this set contained a declarative accent curve and a continuation rise accent curve. While the set did not

explicitly contain interrogative accent curves, the shape of a continuation rise accent curve may be considered to be a declarative accent curve overlayed with an interrogative accent curve. Thus, by using this set to test the performance of the linked PRISM algorithm, we get a fairly good idea of its performance on F_0 contours containing any of the three types of accent curves (declaratives, continuations rises, and interrogatives).

Results of component curve extraction using linked PRISM

The accuracy with which the linked PRISM algorithm estimates component curves was measured by the RMSE between the estimated and the known component curves for the 32 instances. The results are shown in Table 3.7.

	Mean RMSE (Hz)	Standard Deviation (Hz)
Phrase	4.88	1.82
Accent-1	3.37	1.47
Accent-2	3.78	1.29

Table 3.7: Performance of the linked PRISM algorithm. Accent-i is the i-th accent curve in the contour.

Discussion of results

The mean RMSE values obtained by estimating component curves with the linked PRISM algorithm is within the range of mean RMSE values obtained by estimating component curves with either wavelet decomposition or template decomposition. This shows that the linked PRISM algorithm has a similar performance accuracy as the two individual phases of PRISM and that no significant noise was introduced in the process of linking the two phases of PRISM together.

Comparing the results of this objective test with the results of the objective test 2 in Section 3.3.5, we observe that the mean RMSE values obtained by estimating component curves using the linked PRISM algorithm are not significantly different from the mean RMSE values obtained by estimating component curves from the same set of synthetic F_0 contours using template decomposition alone. This is most likely because the phrase curve underlying each of these synthetic F_0 contours was a piecewise linear function and

not a smooth higher-order function. Hence, the template decomposition algorithm was able to extract it just as well as the wavelet decomposition algorithm. No additional benefit was gained in this case from the initial estimate of the phrase curve, Phr_{init_est}. If the phrase curve is a smooth higher-order function, the template decomposition algorithm individually will not be able to estimate the phrase curve as well. The linked PRISM algorithm is likely to estimate the phrase curve more accurately because it has a smooth function Phr_{init_est} (obtained from the wavelet decomposition phase) as the phrase curve template.

3.4 Application of PRISM to natural F_0 curves: Demonstrating the importance of modeling segmental perturbations

An important distinction between the PRISM decomposition algorithm and many other existing decomposition algorithms is that PRISM explicitly models segmental perturbations while others do not. Our main reason for modeling segmental perturbations is based on the hypothesis that segmental perturbations distort the surface realization of F_0 contours and thus obscure the "true" shapes of the underlying phrase curves and accent curves. To estimate these accent and phrase curves, segmental perturbations must be modeled and removed.

Other important reasons for modeling segmental perturbations are that (1) segmental perturbations are not random but highly predictable, (2) smoothing cannot remove perturbations because they are locally systematic (as opposed to random frame-to-frame Gaussian noise) and positively biased (i.e., their integral over an appropriate time interval is positive), (3) segmental perturbations have been shown to be perceptually important (Silverman 1987), and (4) modeling segmental perturbations minimizes signal modification related to pitch modification in speech synthesis[6](van Santen et al. 2004:i).

[6]During speech synthesis, if we impose a smooth perturbation-less target F_0 contour on, say, an obstruent-vowel region, then the initial part of the vowel has to be modified by quite a bit. However, if the target F_0 contour also contains an appropriate segmental perturbation curve, then the pitch modification is considerably less.

We explored the validity of the hypothesis that segmental perturbations distort the surface realization of F_0 contours by analyzing the difference in F_0 contour shape between minimal pairs of utterances that differed only in terms of one consonant. This analysis involved the use of PRISM to decompose the F_0 contours of the minimal pair utterances. A possible objection against the previous analyses shown in this chapter may be that the analyzed curves were synthetic, which may have made the problem too easy for our decomposition algorithm PRISM. A problem with natural curves, however, is that the ground truth is unknown; we partially get around this problem by the aforementioned minimal pairs analysis using PRISM. This particular application of PRISM is one instance of a more general class of applications in which PRISM is used to analyze natural F_0 contours. To this end, PRISM is an analysis tool that is useful for critically testing theories and assumptions regarding intonation.

A description of the corpus from which the minimal pairs were obtained, the analysis used to compare the minimal pairs in terms of their F_0 contours, and the results of the analysis are given in the following sections.

3.4.1 Description of the corpus

The corpus from which we extracted the minimal pairs consisted of 108 utterances. The utterances were recorded by a female native speaker of American English. Each utterance contained one target word embedded in a carrier phrase. Each target word is a three-phoneme sequence: consonant-vowel-consonant. The consonants choices were /k/, /b/, and /l/, and the vowel choices were /i:/, /u/, and /@/. All possible combinations of these phonemes yielded 27 target words.

Each target word was embedded in two carrier phrases: "Please say the word BILL again." and "Please DON'T say the word bill again." The word "bill" is an example target word. The capitalized word in each of the two carrier phrases indicates the word emphasized in the phrase. The carrier phrases were so structured to elicit the target word from the speaker once in the emphasized form and once in the unemphasized form. This yielded $27 \times 2 = 54$ utterances. Each of the 54 utterances were recorded at two speaking rates: normal but slow and normal but fast. Thus, a total of 108 utterances were recorded.

3.4.2 Minimal pair analysis

From the aforementioned corpus, we obtained the set of 54 utterances that were recorded at the normal but slow speaking rate. The F_0 contours of these 54 utterances were extracted (at 10 ms intervals) using Praat (Boersma and Weenink 2008). For each of the utterances, the F_0 contour corresponding to the target word was excised.

The F_0 contours of the target words were divided into two sets of 27 F_0 contours each. One set contained the F_0 contours related to the emphasized version of the target words and the other set contained the F_0 contours related to the unemphasized version of the target words.

/k i: l/ vs. /b i: l/	/k i: l/ vs. /l i: l/	/k u l/ vs. /b u l/
/k u l/ vs. /l u l/	/k @ l/ vs. /b @ l/	/k @ l/ vs. /l @ l/
/k i: b/ vs. /b i: b/	/k i: b/ vs. /l i: b/	/k u b/ vs. /b u b/
/k u b/ vs. /l u b/	/k @ b/ vs. /b @ b/	/k @ b/ vs. /l @ b/
/k i: k/ vs. /b i: k/	/k i: k/ vs. /l i: k/	/k u k/ vs. /b u k/
/k u k/ vs. /l u k/	/k @ k/ vs. /b @ k/	/k @ k/ vs. /l @ k/
/b i: k/ vs. /l i: k/	/b u k/ vs. /l u k/	/b @ k/ vs. /l @ k/
/b i: l/ vs. /l i: l/	/b u l/ vs. /l u l/	/b @ l/ vs. /l @ l/
/b i: b/ vs. /l i: b/	/b u b/ vs. /l u b/	/b @ b/ vs. /l @ b/

Table 3.8: Minimal pairs of target words.

From each set, we obtained 27 pairs of F_0 contours corresponding to the minimal pairs of target words that differed only in terms of the first consonant (shown in Table 3.8). One can assume that since both words in each minimal pair were recorded in the same prosodic context (i.e., same phrase boundaries, same emphasis condition, same or close to the same speaking rate and vowel duration, and embedded in the same carrier phrase), they have very similar F_0 contours. However, we conjecture that this assumption will not be borne out by simply comparing the raw F_0 contours (or even the smoothed F_0 contours) of the minimal pairs. This conjecture follows from the hypothesis that segmental perturbations caused by vowel-initial obstruents distort the surface realization of F_0 contours and thus obscure the "true" shapes of the underlying phrase curves and accent curves. Phrase curves and accent curves are the components of the F_0 contour that map to prosodic features (phrase curve to phrase level prosody and accent curve to foot or word level prosody);

thus, the similarity within each pair of F_0 contours corresponding to each minimal pair is really in terms of these component curves. However, if the shapes of these component curves are obscured by the presence of segmental perturbations, comparisons between each F_0 contour pair corresponding to each minimal pair will yield large differences. For instance, if one of the two words in the minimal pair contains a vowel-initial obstruent while the other does not, (e.g., /b u b/ vs. /l u b/), the shape of the raw F_0 contour of the former will be far different from the shape of the raw F_0 contour of the latter (see panels a and b of Figure 3.14 for an example). If the two words in the minimal pair contain different vowel-initial obstruents (e.g., /b u b/ vs. /k u b/), their corresponding F_0 contours may still differ from each other because the segmental perturbations caused by different obstruents may have different amplitudes and spans even though their general shape is roughly the same.

We propose that the similarity within each F_0 contour pair corresponding to each minimal pair can be established if we remove the segmental perturbations from the F_0 contours using PRISM. First, PRISM is used to decompose each F_0 contour into a phrase curve, an accent curve, and a segmental perturbation curve; then, only the phrase curve and the accent curve are recombined using the superpositional approach to obtain the "clean" F_0 contour, i.e., a contour containing no segmental perturbations. Since the shape of each clean F_0 contour will correspond more closely to the shapes of the underlying phrase curve and accent curve, we expect that each minimal pair will show a high degree of similarity when compared in terms of the corresponding pair of clean F_0 contours.

To test whether the assumption that minimal pairs produced in the same prosodic context have similar F_0 contours is valid, and also to test whether this similarity is hidden due to the presence of segmental perturbations, we compared each pair of F_0 contours corresponding to each minimal pair in the following five ways:

C1: In terms of their raw F_0 contours: The raw F_0 contours are normalized so that their lengths and norms are equal, and their peak locations are the same. Then, the RMSE between the pair of raw F_0 contours is calculated. (See Panels a and b in Figure 3.14 for an example.)

C2: In terms of their smoothed F_0 contours: The gaps in the normalized raw F_0 contours are removed using linear interpolation, and the resulting contours are smoothed using moving averages. Then, the RMSE between the pair of smooth F_0 contours is calculated. (See Panels c and d in Figure 3.14 for an example.)

C3: In terms of their accent curves: Accent curves are obtained by PRISM decomposition from the normalized F_0 contours. Then, the RMSE between the pair of accent curves is calculated. (See Panels e and f in Figure 3.14 for an example.)

C4: In terms of their phrase curves: Phrase curves are obtained by PRISM decomposition from the normalized F_0 contours. Then, the RMSE between the pair of phrase curves is calculated. (See Panels g and h in Figure 3.14 for an example.)

C5: In terms of their clean F_0 contours: The phrase curve and accent curve obtained by PRISM decomposition from each normalized F_0 contour are combined using the superpositional approach to obtain clean F_0 contours. Then, the RMSE between the pair of clean F_0 contours is calculated. (See Panels i and j in Figure 3.14 for an example. In each of these two panels, the corresponding raw F_0 contours are shown by the dashed lines. Notice that the PRISM analysis not only removed the segmental perturbations caused by the vowel-initial obstruents but also removed the voicing irregularities at the end of the F_0 contour.)

3.4.3 Results of minimal pair analysis

The mean of the RMSE differences within each pair of F_0 contours (raw, smooth, or clean) or components of the F_0 contours (phrase curves or accent curves) are computed across each set (emphasized and unemphasized). They are shown in Table 3.9. The column labels C1 through C5 correspond to the five different types of comparisons enumerated in the preceding section.

3.4.4 Discussion of results

From the mean RMSE values obtained from each of the five types of comparisons, we see that the difference within each pair of the raw F_0 contours (C1) is the greatest. The difference within each pair of smooth F_0 contours (C2) is considerably less. And the

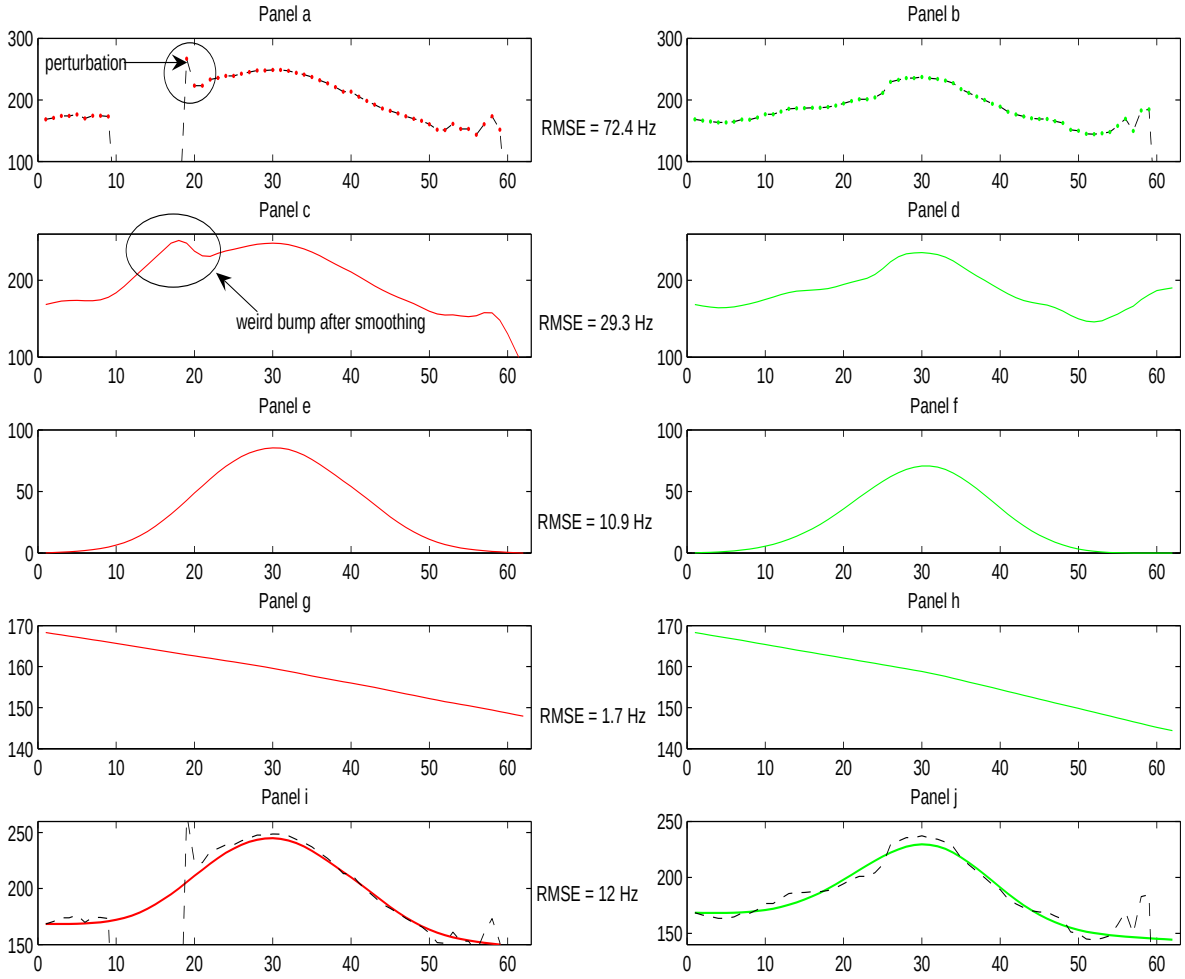

Figure 3.14: Illustrating the five different comparisons between the pair of F_0 contours corresponding to minimal pair /b u b/(left panels) vs. /l u b/(right panels). In each panel, time (in 10 ms intervals) is on the horizontal axis, while frequency (in Hz) is on the vertical axis. RMSE indicates difference between F_0 contour pairs. The dashed lines in the bottom set of panels indicates the raw F_0 contours.

	C1 (Raw)	C2 (Smoothed)	C3 (Accent)	C4 (Phrase)	C5 (Clean)
Emphasized Pairs	101.7095	53.4332	11.3227	6.7147	13.2538
Unemphasized Pairs	78.6524	19.2642	3.7104	5.6680	6.3729

Table 3.9: Mean RMSE (in Hz) obtained by comparing the F_0 contours of each minimal pair in five different ways.

difference within each pair of clean F_0 contours (C5) is the least of the three. We also see that part of the within-pair difference in clean F_0 contours is due to the corresponding within-pair difference in accent curves (C3) and part is due to the corresponding within-pair difference in phrase curves (C4). This was expected since a clean F_0 contour is a summation of a non-negative phrase curve and a non-negative accent curve.

The results trend similarly for both the emphasized and the unemphasized set of target words. However, the paired differences in the case of the unemphasized set of target words is considerably less than that of the emphasized set. This is due to the fact that when words are unemphasized, the related pitch excursions are smaller.

These results indicate that (a) gaps and segmental perturbations in the F_0 contour caused by the presence of obstruents preceding vowels in the associated speech stream do distort the shape of the F_0 contour; (b) smoothing the F_0 contour does not undo the distortion completely because smoothing may still leave behind uncharacteristic bulges in the F_0 contour (shown in Panel c of Figure 3.14); and (c) removing the segmental perturbations from the F_0 contour using PRISM is very effective at producing the clean shape of the F_0 contour, in which the "true" shapes of the underlying phrase curve and accent curve are unobscured.

Conclusively proving that the phrase curve and the accent curve that make up each clean F_0 contour are indeed the "true" component curves is not possible. However, given the sufficiently high accuracy with which PRISM estimated component curves in the previously described objective evaluations, we can accept that these curves closely approximate the "true" curves. This position is further bolstered by the results that show that each minimal pair of prosodically equivalent target words is very similar in terms of the corresponding pair of phrase curves, accent curves, and clean F_0 contours (as shown by the

low mean RMSE difference produced by comparisons C4, C3, and C5, respectively).

The within-pair similarity of phrase curves, accent curves, and clean F_0 contours corresponding to each minimal pair also supports the assumption that utterances that are recorded in the same prosodic context have very similar F_0 contours. These results have positive implications for intonation generation via recombinant synthesis. In recombinant synthesis, we generate a target F_0 contour by superpositionally combining phrase curves and accent curves that match the target in terms of prosodic context. Thus, validation of the assumption that utterances produced in the same prosodic context have very similar F_0 contours implies that our idea of using prosodic context as a search criterion to produce target F_0 contours is reasonable.

Overall, this analysis provides strong empirical evidence in favor of the hypothesis that segmental perturbations distort the surface realization of F_0 contours and thus obscure the "true" phrase curves and accent curves. In addition, this analysis supports the thesis that PRISM can be used to analyze natural F_0 contours so that critical tests of theories and assumptions regarding intonation can be performed. Another such application of PRISM is demonstrated in Chapter 5. In that chapter, PRISM is used to test assumptions regarding perceived prominence and the shapes of F_0 contour component curves.

3.5 Conclusions

In this chapter, we presented a detailed description of our decomposition algorithm PRISM that can be used to decompose any given F_0 contour into phrase curves, accent curves, and segmental perturbation curves. Using several sets of synthetic F_0 contours whose component curves were known, we objectively evaluated the accuracy with which PRISM can estimate component curves. In these evaluations, the RMSE between the known and estimated component curves were under the frequency difference required for just noticeable difference in pitch for connected speech. This indicates that PRISM can decompose F_0 contours with sufficiently high accuracy.

We also showed an application of PRISM to natural F_0 contours through the particular problem of analyzing the difference in F_0 contour shape between minimal pairs of

utterances that differed only in terms of one consonant. The performance of PRISM on natural F_0 contours cannot be objectively evaluated because the ground truth is unknown; however, the strong within-pair similarity of estimated phrase curves and accent curves corresponding to each minimal pair does suggest that the component curves estimated by PRISM are close approximations of the "true" component curves.

Chapter 4

Intonation Generation in Recombinant Synthesis

4.1 Background

Two different approaches to speech synthesis are currently prevalent: *traditional concatenative synthesis* and *unit selection synthesis*. Traditional concatenative synthesis (e.g., diphone synthesis) involves generating target intonation contours using statistical, linguistic, and heuristic rules, which are then imposed on stored speech units using signal modification methods. In contrast, unit selection synthesis uses neither synthetic contours nor any signal modification. It involves simply searching a large tagged corpus for units that match the target in terms of phonemic sequence and prosodic context, and concatenating the units together.

Each of the two approaches has its strengths and weaknesses. Unit selection synthesis preserves the richness of natural speech but it struggles to cover the large combinatorial space of possible phoneme sequences and prosodic contexts that a given phone sequence can occur in. In fact, van Santen et al. (2005) claim that the probability is near-certain that a given input text will require phoneme sequence and prosodic context combinations that the speech corpus does not have. This sparsity problem is less than apparent in unit selection synthesis because unit selection synthesis often uses corpora that are optimized to capitalize on the extremely uneven frequency distribution of phone sequence-prosodic context combinations, and because not all prosodic variations are audibly perceptible.

However, these band-aid solutions do not address the fundamental challenge of unit selection synthesis. Speech quality improvement, demanded by current interest in emotional speech, concept-to-speech, and human-machine dialogue applications, is possible only by building bigger corpora, which in turn increases the combinatorial complexity — since each acoustic unit has to be recorded in many different prosodic contexts — and amplifies the sparsity problem of unit selection synthesis.

The combinatorial problem is not encountered by traditional concatenative synthesis because it relies on statistical, linguistic, and heuristic rules to generate target contours that are imposed on selected acoustic unit sequences. However, this method is limited by two factors: the lack of "naturalness" of the generated F_0 contours and the speech quality degradation resulting from signal modifications (van Santen et al. 2004:i). The reason for both is that the mapping between the linguistic control factors (e.g., word stress, proximity to phrase boundaries) and the multiple acoustic dimensions such as intonation, duration, spectral balance, and spectral dynamics is poorly understood. Thus, both the task of generating F_0 contours and imposing them on phone sequences is difficult.

An alternative to these two approaches, called *recombinant synthesis* (also called *multi-level unit sequence synthesis*), has recently been proposed by van Santen et al. (2005). The recombinant synthesis method involves searching one corpus for acoustic units, while searching another corpus for components of the F_0 contour that have been extracted from natural speech. The component curves are added together according to the general superpositional approach to construct the target F_0 contour, which is imposed on the acoustic unit sequences using standard signal modification methods. The two key concepts of this approach are (1) decoupling the prosodic and the phonemic requirements of the target units, and (2) generating quasi-natural target F_0 contours using F_0 components that were extracted from natural speech.

These two concepts help combine the strengths of unit selection synthesis and traditional concatenative synthesis, while minimizing their weaknesses. Decoupling the prosodic and the phonemic requirements of the target units reduces the combinatorial problem from a quadratic to a linear problem because unlike unit selection synthesis, a single unit in

the corpus does not have to simultaneously satisfy both the phonemic and prosodic requirements. The target contour generation method addresses the naturalness problem. Based on the superpositional framework, the target contour is generated by the superposition of appropriate phrase curves, accent curves, and segmental perturbation curves that were extracted from natural speech. Thus, the naturalness of the target contour is enabled at multiple time scales, and it sounds far more natural than a synthetic target contour that is generated by rule in traditional concatenative synthesis. Also unlike unit selection synthesis, in which larger pitch excursions may cause audible discontinuities at the concatenation points of the target contour, the target contour generated in the recombinant synthesis method is smooth and continuous. Finally, speech degradation due to signal modification problems is also handled by the recombinant approach. First, since the F_0 curve is generated such that the synchrony between the segmental perturbations of the target contour and the segmental frames of the target unit are preserved, signal processing artifacts that arise from modifying the target contour to the original contour of the acoustic unit can be predicted and minimized (as shown in Klabbers and van Santen 2003). Secondly, the modification artifacts that develop when the F_0 difference between the target and the source is large is also minimized by generating the target F_0 contour from component curves that match the source closely. Thus, the three main problems affecting the quality of speech synthesis — namely, data sparsity, the lack of naturalness of target contours, and output speech degradation due to signal modification — are handled by the recombinant synthesis method.

But a major problem that needs to be solved before this synthesis method can be implemented is the automatic decomposition of natural F_0 curves into component curves. This problem is challenging because (1) few assumptions about the shapes of the component curves can be made, (2) segmental perturbations may produce spurious peaks and valleys in the F_0 contours, (3) successive accent curves may overlap in time, and (4) natural F_0 contours may have gaps due to silences or unvoiced sounds. None of the existing decomposition algorithms are able to address all of these concerns (van Santen et al. 2005). For example, the algorithms based on the Fujisaki model (e.g., Mixdorff

2000, Narusawa et al. 2002) make strong assumptions about the shapes of the component curves, while the decomposition algorithm outlined in Sakai (2004) does not model segmental perturbations. PRISM aims to meet all these concerns and more (as outlined in Section 1.3.1), and thus it should enable the successful generation of intonation in the recombinant synthesis method. To test whether PRISM is indeed able to do so, we developed a proof-of-concept (POC) implementation of recombinant synthesis that can generate meaningful and expressive intonation of fairly long and complicated sentences (described in Section 1.3.2).

4.2 Proof-of-concept implementation of recombinant synthesis

As outlined in Section 1.3.2, the POC implementation involves extracting component curves of natural F_0 contours using PRISM, creating a corpus of the component curves, selecting appropriate component curves from the corpus according to phonological/prosodic rules, and additively recombining the extracted curves to generate target F_0 contours. This implementation does not include creating an acoustic unit corpus or searching it for appropriate acoustic units; rather, the target F_0 contours are imposed on pre-recorded natural speech so that the strengths and weaknesses of this intonation generation method can be observed, unconfounded by the effects of acoustic unit search and concatenation.

An incremental approach has been used in the development of the POC implementation, beginning with intonation generation for simple utterances and then moving to more complex utterances. For the purposes of this implementation, the complexity of utterances is considered along two dimensions: types of phonemes used (sonorant and non-sonorant) and phrase modality (declarative and continuation rise). Phonemic categorization into sonorant versus non-sonorant is important in this context. Sonorant phonemes produce reliable F_0 values whereas non-sonorants either cause gaps in the F_0 contour (due to unvoiced phones) or produce unreliable F_0 values. The latter situation increases the complexity of the decomposition algorithm and consequently, of recombinant synthesis. The two different phrase modalities are considered for similar reasons: the shapes of accent

curves corresponding to continuation rises are more complicated than that of accent curves corresponding to declaratives.

Adhering to the idea of incremental development, the POC implementation of recombinant synthesis is divided into two stages of increasing complexity: (1) generating intonation of fully sonorant declarative utterances; and (2) generating intonation of phonemically unrestricted declarative utterances containing continuation rises. A detailed description of each stage of implementation and of the perceptual tests used to evaluate each stage is presented in the following sections.

4.2.1 POC implementation of recombinant synthesis to generate intonation of fully sonorant declarative utterances

In this stage of the proof-of-concept implementation, we have developed a corpus of F_0 component curves and generated the intonation of 13 pre-determined simple declarative sentences using recombinant synthesis. The corpus was created by the decomposition of natural F_0 contours using the template decomposition stage of the PRISM algorithm described in Section 3.3.4.

The first step in developing the corpus was to record the natural utterances from which the component curves needed to generate the intonation of the 13 target utterances would be extracted. The recordings were elicited from one native male speaker of American English. Each utterance set had the following format, where x denotes utterance set number and $x = 1, 2, ..., 13$:

a_x: **Leena**$_a$ Roy$_a$.

b_x: **Leena**$_b$ Weller$_b$ and **Ann**$_b$ Roy$_b$.

c_x: Leena$_c$ **Weller**$_c$ and Ann$_c$ **Roy**$_c$.

The speaker was instructed to put a relatively higher emphasis on the highlighted words and to pronounce utterances of type b_x and c_x in a "list-like" manner, in order to elicit variability in pitch range more typical of natural speech used in unit selection synthesis than of the stilted speech in diphone synthesis. A total of 39 utterances were

recorded. In each of the 13 utterance sets described above, utterance a_x was considered the target utterance. It was *not* used for building the intonation corpus. The corpus was built by decomposing utterances of type b_x and c_x.

Every utterance was manually segmented into left-headed feet using Wavesurfer (Sjölander and Beskow 2000). F_0 contours were extracted (at 10 ms interval) using the ESPS *get_f0* utility (Talkin 1993); based on Orfanidis (1996), high-frequency noise was removed using the Savitzky-Golay filter of order 3 and length 5.

A Gaussian curve was used as an accent curve template. It is important to note that the exact shape of the template is not critical as long as it is single-peaked and has initial and final values of 0, because the time warping procedure renders different templates with these features equivalent. Given these inputs, the template decomposition stage of PRISM was used to decompose every natural F_0 contour into an estimated phrase curve and a number of estimated accent curves. An example of the natural F_0 contour decomposition is given in Figure 4.1. The average warping function obtained by the decomposition of the utterance set b_x is shown in Figure 4.2. Note that for the first three accent curves, the average warping function goes beyond 100% of the associated foot duration, indicating that each of the first three accent curves overlapped with the following accent curve.

Once the corpus was prepared, two versions of the target F_0 contours (i.e. of type a_x: **Leena**$_a$ Roy$_a$) were generated using methods DECOM$_1$ and DECOM$_2$, respectively. Both methods generated target contours by selecting appropriate component curves from the corpus and adding them together according to the general superpositional approach. However, the methods differed in terms of the criteria used to select accent curves, as outlined below:

- "Semi-natural" F_0 with accent curves from the right prosodic context (DECOM$_1$): For every utterance set, the accent curves associated with the first and the last units of b_x (e.g., **Leena**$_b$ and Roy$_b$ from "**Leena**$_b$ Weller$_b$ and **Ann**$_b$ Roy$_b$") were obtained. The obtained accent curves were scaled by a suitable height factor and added to a synthetic phrase curve to generate the target F_0 curve. The phrase curve was generated to approximately mimic the average phrase curve obtained by decomposition of a_x.

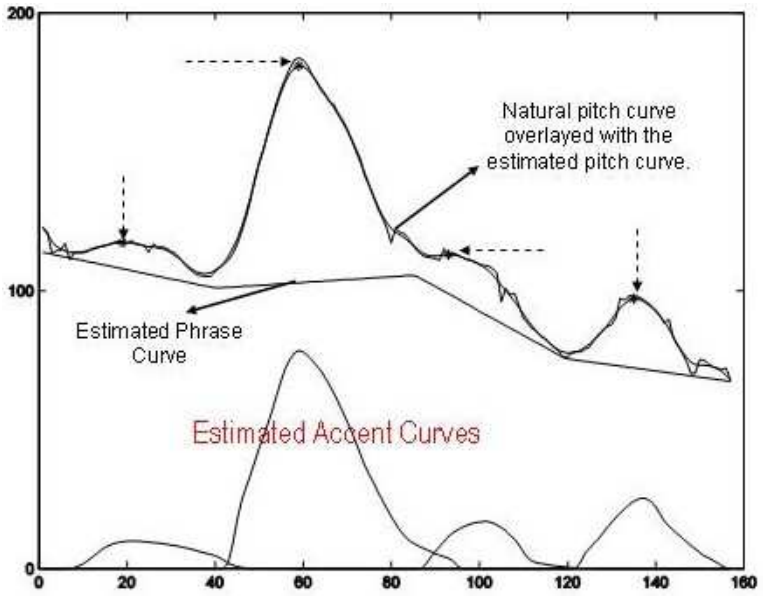

Figure 4.1: Decomposition of a natural F_0 contour of type b_x. The dashed arrows indicate the peaks detected by the peak detection algorithm. The x-axis is time in 10 ms intervals, and the y-axis is frequency in Hz.

- "Semi-natural" F_0 with accent curves from the wrong prosodic context (DECOM_2): In the previously described method of F_0 generation, the accent curves selected from the corpus appear in the same prosodic context as the target accent curves. However, one may not be able to find accent curves in the corpus that match the target accent curves completely in terms of prosodic context. Method DECOM_2 presupposes such a prosodic mismatch between the candidate accent curves and the target accent curves. Accordingly, for every utterance set, the accent curves associated with the first and last units of c_x (e.g., Leena_c and $\mathbf{Roy}_c$ from "Leena_c $\mathbf{Weller}_c$ and Ann_c $\mathbf{Roy}_c$") were selected to produce the target contour. The relative emphasis on Leena_c and $\mathbf{Roy}_c$ were the reverse of the relative emphasis on $\mathbf{Leena}_a$ and Roy_a, respectively. However, the position in phrase and the number of syllables for the matched pairs, ($\mathbf{Leena}_a$, Leena_c) and (Roy_a, $\mathbf{Roy}_c$), were identical. Given the matching position in phrase and the identical number of syllables, we conjectured that the difference in emphasis between the matched pairs of accent curves translated

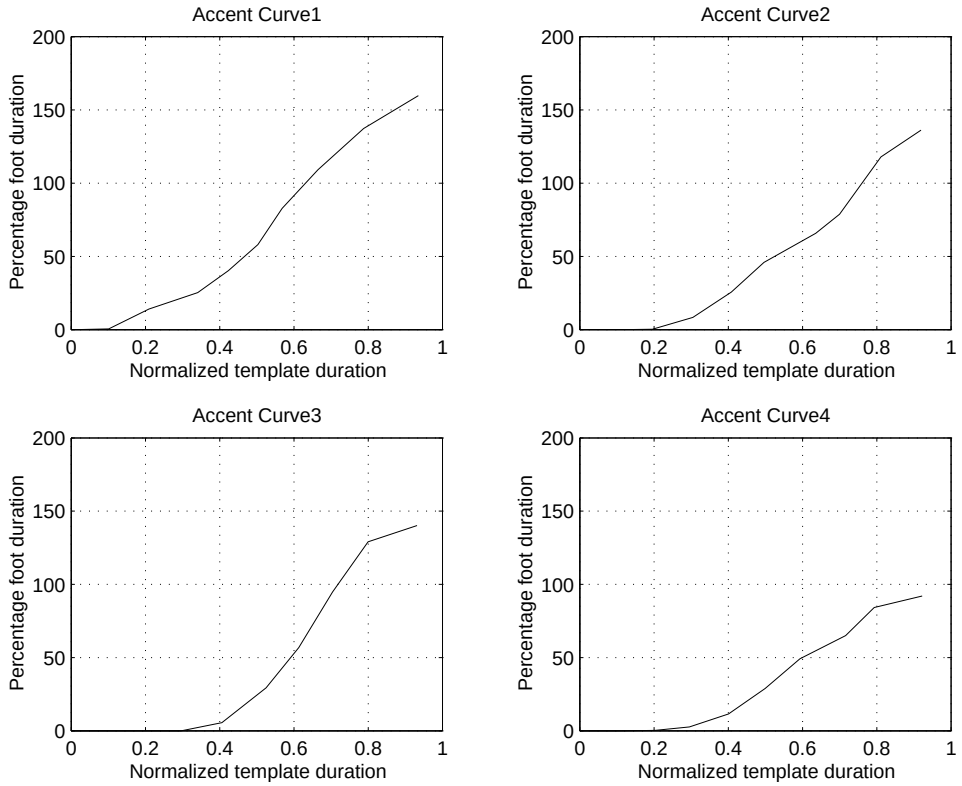

Figure 4.2: Average warping function related to the decomposition of utterance set b_x.

to only a difference in the height factor of the accent curve[1]. Based on this conjecture, the accent curves were appropriately scaled and added to a synthetic phrase curve to generate the target contour. The phrase curve was synthesized in the same manner as in method DECOM$_1$.

Each of the two versions of the target contours were imposed on the original utterance of type a_x using STRAIGHT (Kawahara et al. 2001). Duration mapping between the generated F_0 and the original F_0 was done at the foot level. It is important to note that in the described implementation, the phrase component used for the target contour generation was not natural; rather, it was a synthetic curve that mimicked the target. This limitation was rectified in the next stage of the POC implementation (Section 4.2.4).

[1]This conjecture is generally supported by our finding, presented in Chapter 5, that accent curve height is the strongest correlate of perceived prominence (or emphasis).

4.2.2 Perceptual test to compare intonation generation in recombinant synthesis to other methods

In van Santen et al. (2004:i, 2005), it is claimed that intonation generated by the recombinant synthesis method will produce more natural-sounding intonation than that produced either by synthetic target F_0 contours (as in traditional concatenative synthesis) or by concatenation of raw F_0 contour fragments (as in unit selection synthesis). Verification of this claim is not only needed to demonstrate the purported superiority of the recombinant synthesis concept but also to confirm the effectiveness of the PRISM decomposition algorithm in extracting "true" component curves; if the extracted component curves are not at least close approximations of the "true" curves, then the target contours constructed from them will not sound natural. Thus, a comparative perceptual test was conducted to test this claim.

For the purpose of the comparative perceptual test, the same 13 utterances were generated using three other methods described below:

- Raw F_0 concatenated (CONCAT): For each utterance set, the raw F_0 curve was extracted from the first and the last units of utterance b_x (because $\mathbf{Leena}_a$ and $\mathbf{Leena}_b$ were in the same prosodic context, as are Roy_a and Roy_b) and concatenated to generate the target F_0 contour. We initially tried to match the peaks of the concatenated F_0 to the peaks of the natural F_0 of the target utterance by multiplying each of the units by a height factor. However, doing so sometimes resulted in more drastic jumps at the unit boundaries, as shown in Figure 4.3; therefore, we decided not to perform the peak-matching.

- SLAM F_0 (BY-RULE): The target F_0 contours were generated according to the intonation model called SLAM (Simplified Linear Alignment Model, described in Section 3.3.5), implemented by the OGI version of the Festival Speech Synthesis system (Black and Taylor 1997). The phrase curve parameters were set to approximate the speaker's phrase curve characteristics (such as average phrase start, phrase end, and phrase height) and the accent curve parameters were set to produce peaks at approximately the same heights as the peaks in the natural utterances.

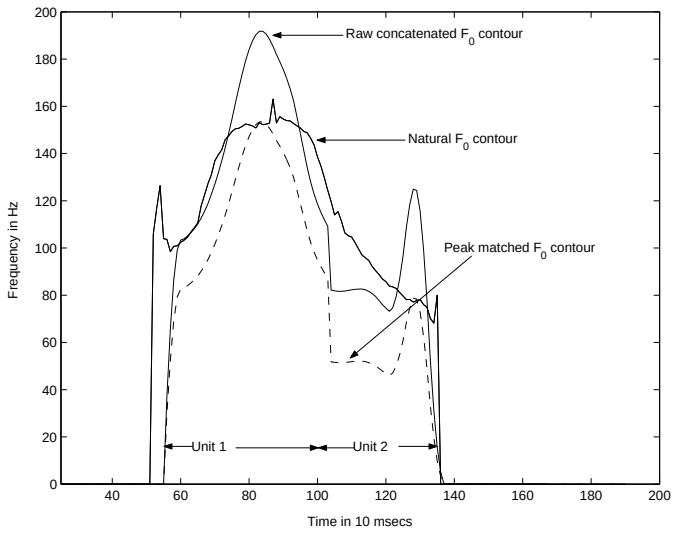

Figure 4.3: Peak-matching concatenated F_0 contours can make the pitch mismatches at unit boundaries more drastic.

- Resynthesized F_0 (ORIG): To compare the target intonation contours generated by DECOM_1, DECOM_2, CONCAT, and BY-RULE to a gold standard, the original F_0 curve of utterance type a_x was extracted using ESPS and reimposed on the same utterance, thus creating a version called ORIG.

Just like DECOM_1 and DECOM_2, these target contours were imposed on the utterance of type a_x using STRAIGHT, and duration mapping between the generated F_0 and the original F_0 was performed at the foot level.

Thus, 13 utterance sets, each comprising 5 target F_0 contours applied to the same utterance, were created to conduct the comparative perceptual test. In the perceptual test, the ordering of the 13 sets was randomized, and within each set, the order of presentation of the 5 target utterances was randomized as well. The perceptual test was presented to six listeners using a CGI-based script, WWStim (Veenker 2001). The five target utterances of each randomized set were presented on the same page, with each set on a different page. The test was performed on one computer with a M-Audio Duo USB audio interface and a high quality AKG headset. The listeners were asked to listen to each of the 65 target utterances, one at a time, and to rate the naturalness of the intonation on a five-point scale. Of the six listeners, four were students from our university and two were staff

members. All listeners were fluent in American English, and four are involved in speech research. The results formed a 65 stimuli $\times$ 6 listeners score matrix S.

Results of the perceptual test comparing intonation generation in recombinant synthesis to other methods

Repeated measures ANOVA: The first step was to find out if the six listeners could indeed differentiate between the five methods of intonation generation in terms of naturalness. We employed repeated measures ANOVA to establish the difference between the five methods. Since all six subjects were measured under all five intonation generation methods, exposed to exactly the same stimuli, it was expected that the scores obtained from each subject would be correlated. Correlated data violates the independence assumption of simple one-way ANOVA, and thus the latter could not be used to test differences among the five intonation generation methods. Instead, repeated measures ANOVA was used.

In the repeated measures framework, this perceptual test has one factor — Intonation Methods — which has five levels: ORIG, CONCAT, BY-RULE, DECOM_1, and DECOM_2. Thus, we have a one-way repeated measures design with 5 levels and 6 subjects. Under the repeated measures design, the following assumptions are made about the data. First, we assume normality of the distribution of the dependent variable at each level of the factor. Second, we assume constancy in the dependent variable's variance across levels of the factor. Third, we assume sphericity, i.e. the correlation between all possible pairs of repeated measures (e.g., BY-RULE vs. DECOM_1, DECOM_1 vs. DECOM_2) is identical.

Before performing the repeated measures ANOVA, the null hypothesis $H0$ and the α value were set as follows:

- $H0$ = there is no difference between the different types of intonation methods.
- $\alpha = 0.05$.

The one-way repeated measures analysis was performed using the statistical package R. The score matrix S was reshaped as shown in Table 4.1.

Because Intonation Methods was crossed with the random factor Subject in this repeated measures design, an error term for Intonation Methods was also specified. A summary of the results obtained by the repeated measures ANOVA is shown in Table 4.2.

Observations	Subject	Intonation Methods	Scores
1	S1	RESYNTH	3
2	S1	RESYNTH	4
...	...	...	...
64	S1	$DECOM_2$	3
65	S1	$DECOM_2$	1
....	...	...	...
158	S3	BY-RULE	2
159	S3	BY-RULE	3
...	...	...	...
389	S6	$DECOM_2$	2
390	S6	$DECOM_2$	3

Table 4.1: Data for repeated measures ANOVA.

Source	Sum of Squares	degrees of freedom	Mean square	F	p-value
Model	42.990	4	10.747	11.252	6.018e-05
Error	19.103	20	0.955		

Table 4.2: Summary of results of one-way repeated measures ANOVA.

The results of the repeated measures ANOVA, presented in Table 4.2, show that there is a significant difference (p-value = 6.018e-05) in naturalness among the five methods of intonation generation.

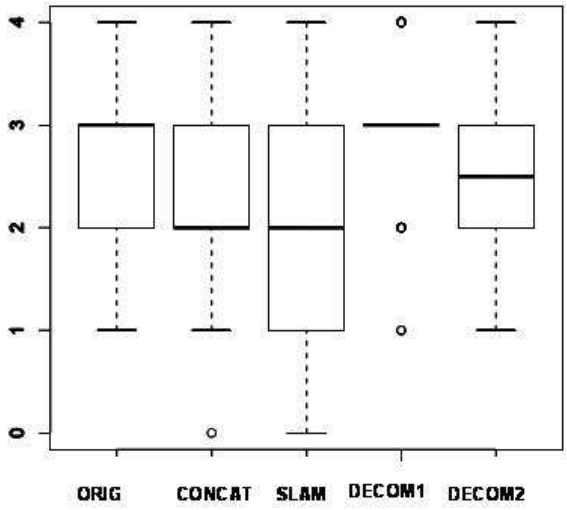

Figure 4.4: Boxplot depicting the difference in the perceived naturalness of the five intonation generation methods.

A boxplot visually depicting the difference between the five methods is shown in Figure 4.4. Two points are apparent from the boxplot. First, the interquartile range of DECOM_1 is extremely small, meaning that most of the subjects scored DECOM_1 similarly. Second, the median value of DECOM_1 is equal to the median value of ORIG, meaning that the naturalness of DECOM_1 approaches the naturalness of re-synthesized natural speech, encoded as ORIG in this experiment. Together, these two points hint at the notion that most of the subjects considered speech produced by method DECOM_1 to be as natural as re-synthesized natural speech. However, to firmly establish this notion further tests were performed.

Planned t-tests: The second step was a more direct attempt to establish our claim that the recombinant intonation generation method produces more natural-sounding intonation than that produced either by synthetic target F_0 contours (as in diphone synthesis) or by concatenation of raw F_0 contour fragments (as in unit selection synthesis). Three planned t-tests were performed to test three key predictions, according to which we expected DECOM_1 to be better than CONCAT, BY-RULE, and DECOM_2. Since we had a directional sense about this analysis, a one-tailed t-test was performed. All t-tests were performed using MATLAB. For each t-test, the null hypothesis $H0$ and the alternative hypothesis HA were set as follows:

- $H0$: mean naturalness of DECOM_1 = mean naturalness of the other method.
- HA: mean naturalness of DECOM_1 > mean naturalness of the other method.

Since only a few planned t-tests were performed rather than the 10 possible t-tests, a particular variation of t-tests called the Bonferroni t-test was used. In keeping with the Bonferroni t-test, a stricter criterion for rejecting the null hypothesis was used. Using a family-wise α of 0.05, the criterion to reject the null hypothesis was $0.05/3 = 0.0167$.

The results of the planned t-tests are shown in Table 4.3. Each of the t-tests has 5 degrees of freedom. Since the p-values for all three t-tests are less than 0.0167, we can reject the null hypothesis and conclude that DECOM_1 produces more natural-sounding speech than CONCAT, BY-RULE, and DECOM_2. This establishes the claim that recombinant synthesis produces more natural-sounding speech than other synthesis methods quite convincingly.

	t-test	p-value	Significant?
$DECOM_1$ - CONCAT	4.70	0.0027	Yes
$DECOM_1$ - BY-RULE	4.61	0.0029	Yes
$DECOM_1$ - $DECOM_2$	3.58	0.008	Yes

Table 4.3: Planned t-tests showing that the subjects judged intonation generated by $DECOM_1$ to be significantly more natural-sounding than intonation generated by other methods.

One way between subjects ANOVA: Though we established through the previous two tests that the five methods were significantly different, the means of the five methods (ORIG = 2.846, CONCAT = 2.397, BY-RULE = 1.987, $DECOM_1$ = 2.897, and $DECOM_2$ = 2.462) do not reveal that information very well. This led us to explore whether the different subjects used different range of ratings to score the perceptual test. Using Subjects as a factor, we performed a one-way between-subjects ANOVA using MATLAB.

Three assumptions underlie the one-way between-subjects ANOVA: First, each subject's scores are normally distributed. Second, the variance (of scores) between subjects is constant. Third, there is no correlation between subjects. The null hypothesis is that there is no difference between the subjects' rating range and the α value is 0.05. The ANOVA results are shown in Table 4.4. The F-score equals 8.64 and the p-value equals

Source	Sum of Squares	degrees of freedom	Mean square	F	p-value
Between subjects	36.359	5	7.27179	8.64	8.88389e-8
Error	323.015	384	0.84119		

Table 4.4: Summary of results of one-way ANOVA on subjects.

8.88e-8. Since the F-score is greater than 1 and p-value < 0.05, the null hypothesis is rejected, and we conclude that there is a difference between the subject ratings. A boxplot showing the rating range of the different subjects is shown in Figure 4.5. Note that subject 5 uses a very different range of ratings than all the other subjects, while subject 2 uses a wider range than other subjects.

Principal component analysis: Finally, to obtain a picture of the combined scores corrected for some subjects' using a different range of ratings or not being in line with the majority of subjects, we performed a principal components analysis (PCA) on the score matrix S, after its columns were transformed into z-scores. This analysis produces a

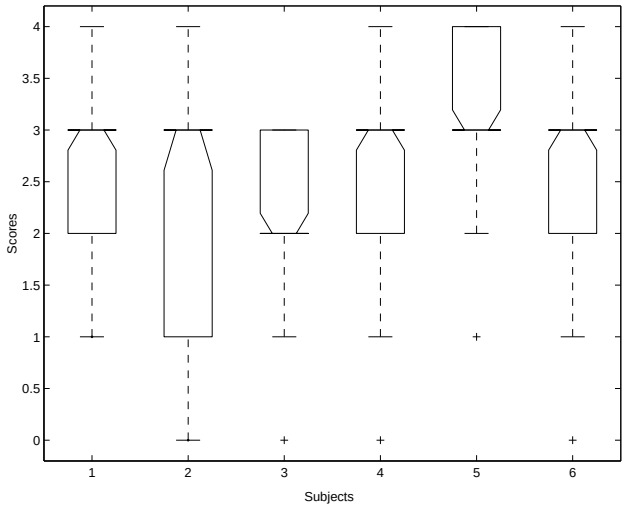

Figure 4.5: Boxplot showing the difference in subject ratings of naturalness.

weighted combination of the ratings, assigning larger weights to mainstream listeners and eliminating any differences in individual usage of the rating scales by the z-transformation. PCA was performed using MATLAB. The resulting averages are shown in Figure 4.6. It illustrates DECOM$_1$'s superior performance over BY-RULE, CONCAT, and DECOM$_2$ in terms of naturalness. It also illustrates that speech produced by DECOM$_1$ is as natural as re-synthesized natural speech.

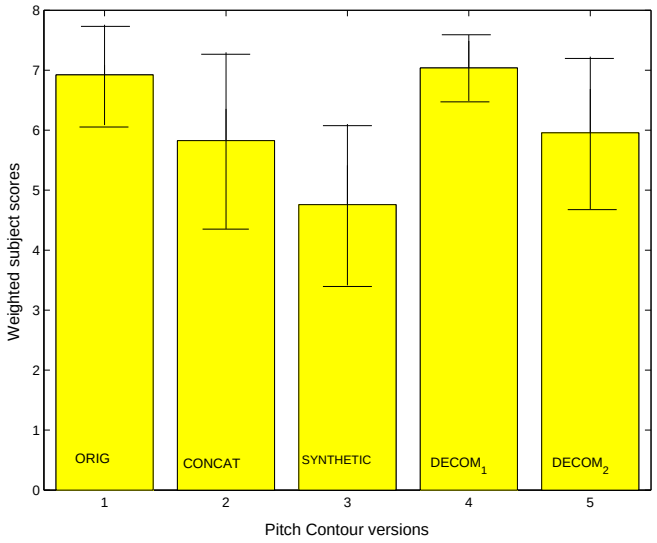

Figure 4.6: Means and standard deviations of different methods of F_0 generation. The bar graph shows the means, and the line above each bar indicates the standard deviation.

4.2.3 Discussion of results

From these four statistical tests, we can conclude that method DECOM$_1$ produces more natural-sounding F_0 compared to methods BY-RULE, CONCAT, and DECOM$_2$. We hypothesize that the superior performance of DECOM$_1$ over the other methods can be attributed to the following key differences.

DECOM$_1$ vs. BY-RULE: The key difference between DECOM$_1$ and BY-RULE is that DECOM$_1$ generates F_0 contours containing details of the natural pitch, whereas BY-RULE does not. An example illustrating this difference is shown in Figure 4.7 (set a). This difference arises from the fact that DECOM$_1$ uses component accent curves extracted from natural F_0 contours to generate target F_0 curves, whereas BY-RULE generates target F_0 curves using statistically-based rules.

DECOM$_1$ vs. CONCAT: The main difference between DECOM$_1$ and CONCAT is that DECOM$_1$ *always* generates continuous F_0 curves whereas CONCAT sometimes generates F_0 curves that have audible pitch discontinuities, as shown in Figure 4.7 (set b). In the illustrated example, the discontinuity between the two parts of the F_0 curve predicted by CONCAT occurred in spite of the fact that each of the two F_0 curve fragments were chosen from appropriate prosodic contexts. Though chosen from appropriate contexts, the two F_0 curve fragments were chosen from a list-type utterance. List-type utterances are often produced with a lively intonation and hence have extreme variations in pitch. Thus, even two F_0 curve fragments excised from the same list utterance and concatenated together can result in a large pitch jump at the boundary between the units. One might argue that this discontinuity can be smoothed; however, most smoothing techniques will cause the natural details inherent in each of the F_0 curve fragments to be lost in the smoothed F_0 contour. Since a desired goal of intonation generation is to preserve the details of natural pitch, we have refrained from performing such smoothing. Another intuitive argument may be that the F_0 curve fragments should be extracted from less expressive speech. Doing so will perhaps result in no (or small) pitch discontinuities; however, the resultant F_0 contour will also be less expressive.

DECOM$_1$ and DECOM$_2$: Finally, the main difference between DECOM$_1$ and

DECOM_2 is this: while in DECOM_2 the component accent curves are extracted from sub-optimal prosodic contexts, in DECOM_1 the component accent curves are extracted from the optimal prosodic contexts. Consequently, the F_0 contour generated by DECOM_2 might not match the target F_0 contour as well as the F_0 contour generated by DECOM_1. This key difference is illustrated in Figure 4.7 (set c).

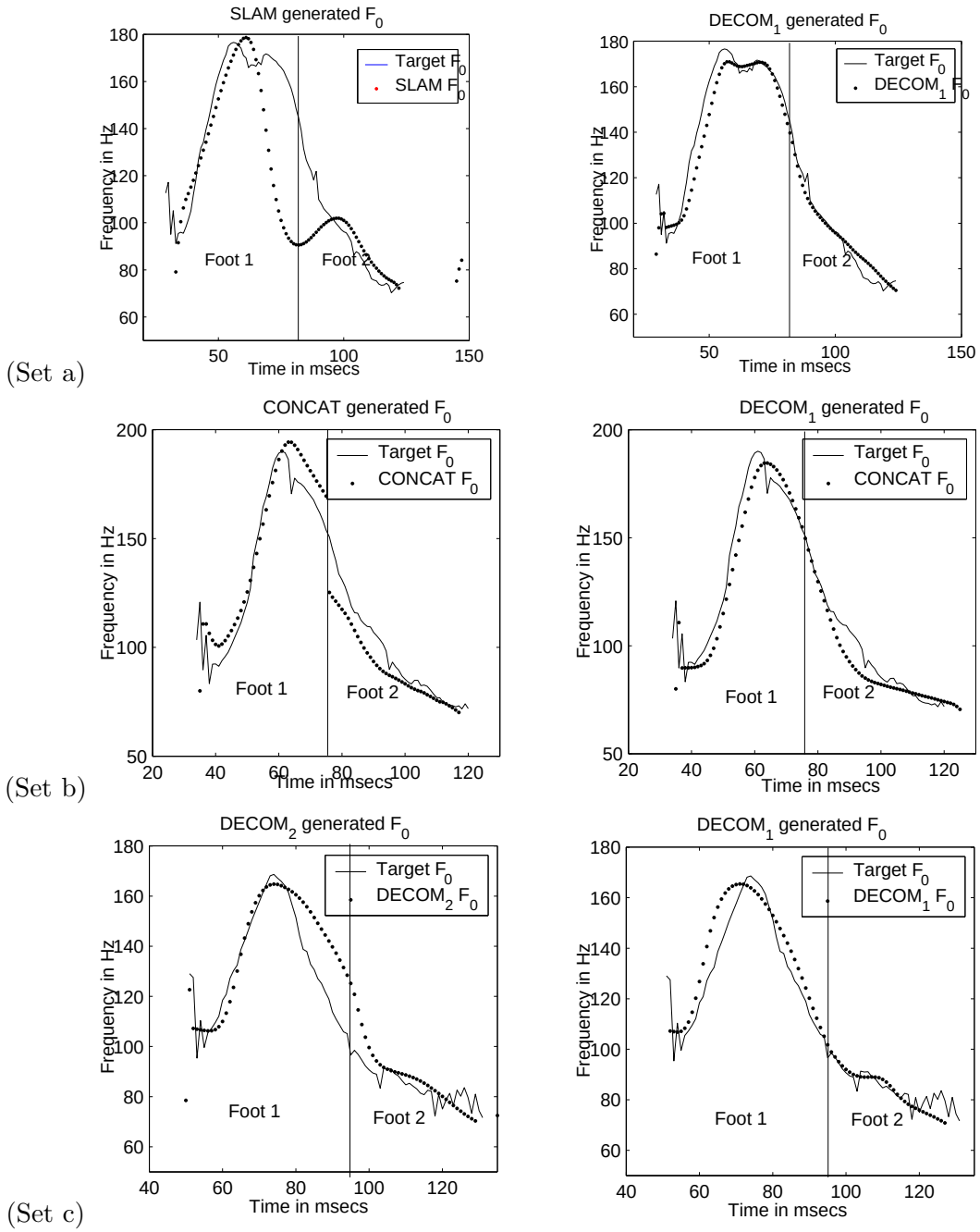

Figure 4.7: Main differences in F_0 prediction between DECOM_1 and each of BY-RULE, CONCAT, and DECOM_2.

4.2.4 POC implementation of recombinant synthesis to generate intonation of phonemically unrestricted declarative utterances containing continuation rises

In this stage of the proof-of-concept implementation, we have developed a corpus of F_0 component curves and generated the intonation of six pre-determined complicated sentences using recombinant synthesis. The target sentences were hand selected from Grimm's fairy tales. All the sentences were of the list-type. They were unrestricted in terms of the phonemes that occurred in them, i.e., they contained sonorants as well as non-sonorants. They contained both declaratives and continuation rises. They differed from each other in terms of the number of phrases in each sentence (between 3 and 5) and the number of feet in each phrase (between 1 and 4). The target sentences were thus selected to demonstrate that the decomposition algorithm PRISM can properly decompose F_0 contours obtained from utterances containing non-sonorants and continuation rises.

The six target utterances were the following:

1. He put two bushels in his right pocket, two in his left, four in a wallet, half on his back, and half on his breast.
2. He gave to the first a cock, to the second a scythe, and to the third a cat.
3. Hans arrived before the king with money and jewels, gold and silver, and cows, sheep, and goats.
4. The kings, princes, and councillors ridiculed and mocked him.
5. The man took out the bread, wine, fruit, and meat.
6. They shot hares, wild deer, birds and pigeons, and whatsoever there was to eat.

Working backwards from the target sentences, a minimal corpus of F_0 component curves needed to generate the intonation of the target utterances was designed. Unlike the implementation described in Section 4.2.1, in this stage, both the accent curves *and* the phrase curves were selected from the corpus. For any target utterance, x, the phrase curves needed were extracted from an utterance in the corpus, y, according to two rules:

P1: The number of phrases in x is equal to the number of phrases in y.

P2: The i-th phrase in x, $phr(x)_i$, and the i-th phrase in y, $phr(y)_i$, must belong to the same equivalence class, EP_n, where n indicates the number of feet in the phrase, and $n = 1, 2$, and ≥ 3.

These rules were met by the set of six utterances listed in Section A.1 of Appendix A.

Rules for selecting the accent curves from the corpus were different from the rules for selecting the phrase curve. Unlike the selection of the phrase curve, all accent curves in a target utterance need not be obtained from the same utterance in the corpus. So, an accent curve in the target utterance, $acc(x)_i$, was obtained from an accent curve in the corpus, $acc(y)_i$, according to the following four rules:

A1: $acc(x)_i$ and $acc(y)_i$ must belong to the same equivalence class, EA_n, where n indicates the number of syllables in the foot spanned by the accent curve, and $n = 1, 2, 3$, and > 3.

A2: The number of phonemes in the first syllable of $acc(x)_i$ must be equal to the number of phonemes in the first syllable of $acc(y)_i$.

A3: The j-th phoneme in the first syllable of $acc(x)_i$ and the j-th phoneme in the first syllable of $acc(y)_i$ must belong to the same broad phonemic class, defined as vowel, sonorant consonant, or obstruent. The term j ranges from 1 to n, where n is the number of phonemes in the first syllable.

A4: The relative position of the accent curve $acc(x)_i$ in the phrase (i.e., phrase-initial, phrase-medial, and phrase-final) must be the same as the relative position of the accent curve $acc(y)_i$ in the phrase.

Governed by these rules, the accent curves for the target utterances were obtained from the set of 13 utterances listed in Section A.2 of Appendix A.

Just like the six target sentences, the sentences used for creating the component curve corpus were hand selected from Grimm's fairy tales. A female native speaker of American English rendered the target utterances as well as the corpus utterances. She was instructed to read them in the expressive manner of reading stories. The corpus-related utterances were decomposed using the template decomposition stage of the PRISM algorithm described in Section 3.3.4. As in Section 4.2.1, a Gaussian curve was used as a template for

the declarative type of accent curves, and a summation of a Gaussian curve and a rising exponential curve was used as a template for the continuation rise. The piecewise linear phrase curve obtained by the template decomposition algorithm was smoothed using penalized least squares (Green and Silverman 1994).

Following decomposition of the F_0 contours corresponding to the utterances in the corpus, the intonation for the target utterances was generated using five steps:

1. Phrase curves and accent curves were selected according to the aforementioned rules.
2. The heights of the accent curves spanning over a target phrase x_i were obtained from the heights of the accent curves spanning over the matching corpus phrase y_i. Rule P2 allowed the possibility that the number of accent curves in x_i and y_i were not the same; in such a case, as many accent curves as possible were matched up; if there were accent curves in x_i that did not have a match in y_i, their height was obtained by interpolating over the heights of accent curves that had found matches in y_i.
3. Duration mapping of the accent curves and the phrase curves was performed at the foot and phrase level, respectively.
4. Location of the peaks in the accent curves was decided according to the peak placement rules developed by Klabbers and van Santen (2004).
5. The accent curves were overlayed on the phrase curve according to the general superpositional approach to produce quasi-natural target F_0 contours. An example of such an F_0 contour is shown in Figure 4.8.

The aforementioned steps together make up the intonation generation method in recombinant synthesis. We refer to this method as RECOM in the following sections. The six quasi-natural target F_0 contours obtained by RECOM were imposed on the corresponding natural target utterances using Praat (Boersma and Weenink 2008)[2]. These target F_0 contours are called "quasi-natural" because each of these contours is composed of phrase curves and accent curves that were obtained from natural F_0 curves.

[2]We used Praat instead of STRAIGHT because during resynthesis, STRAIGHT caused severe distortions to the speech obtained from our female speaker even though it worked perfectly on the speech obtained from our male speaker in the previous stage of the POC implementation.

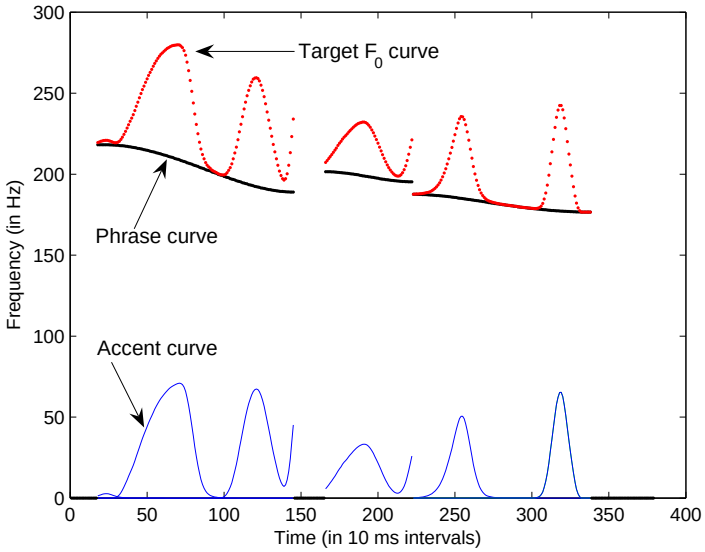

Figure 4.8: Intonation generated for the target sentence, "The man took out the bread, wine, fruit, and meat." using recombinant synthesis.

4.2.5 Perceptual test to compare intonation generation in recombinant synthesis to other methods

In the first stage of the POC implementation, we conducted a comparative perceptual test to evaluate how the intonation generated by recombinant synthesis compared to other intonation generation methods in terms of naturalness. However, in addition to naturalness, we are also concerned about the expressiveness of intonation in speech synthesis — the ideal output of a speech synthesis system is as natural and as expressive as natural speech. With this in mind, in this stage of the POC implementation of recombinant synthesis, we conducted a perceptual test to evaluate how the intonation produced by recombinant synthesis compared to natural speech and to other intonation generation methods in terms of expressiveness.

For the purpose of the comparative perceptual test, intonation of the 6 target utterances were generated using four other methods, which are described below:

- Raw F_0 concatenated (CONCAT1): The F_0 contour of each target utterance was generated by concatenating raw F_0 curve fragments. The raw F_0 curve fragments

were concatenated at the foot level, i.e., each raw F_0 curve fragment spanned a foot. Correspondingly, each raw F_0 curve fragment was selected using rules A1 through A4 that were used to select accent curves in recombinant synthesis. Since these raw F_0 curve fragments were obtained from phonemically unrestricted utterances, a raw F_0 curve fragment may have contained gaps in it. If such a gap occurred at the edge of the raw F_0 curve fragment, its concatenation with another raw F_0 curve fragment, even with post-concatenation smoothing, would have produced weird artifacts in the target F_0 contour. To prevent the occurrence of such artifacts, we removed gaps in each raw F_0 curve fragment using "sagging transitions" implemented via quadratic spline functions. The concept of sagging transitions was established by Pierrehumbert (1981). We obtained the particular rules for when to use sagging versus sustained transitions from Jilka et al. (1999).

- Raw F_0 concatenated and peak-matched to recombinant synthesis (CONCAT2): The target F_0 contours were produced in a manner similar to CONCAT1. However, in this method, each raw F_0 curve fragment was also vertically shifted so that the height of the peak in the fragment matched the peak height of the corresponding fragment in the target contour produced by recombinant synthesis. The vertical shift was done by multiplying each of the raw F_0 curve fragments by an appropriate height factor. Vertical shifting, however, sometimes resulted in drastic jumps at the fragment boundaries, as previously shown in Figure 4.3. To produce smooth target F_0 contours, we smoothed over these drastic jumps by spline interpolation. (See footnote [3] for our reason for performing peak-matching).

- SLAM F_0 (BY-RULE): The target F_0 contours were generated according to the previously discussed intonation model SLAM, implemented by the OGI version of the Festival Speech Synthesis system. The component curve parameters were set to

[3] When we published the results of the perceptual test discussed in Section 4.2.2, not peak-matching the target contours produced by raw pitch concatenation to those produced by recombinant synthesis was considered "unfair" by some readers. It was conjectured that if peak-matched, the target contours produced by raw pitch concatenation may sound as good as those produced by recombinant synthesis. To test this conjecture and to address the concern about unfairness, we performed peak-matching in this perceptual test.

mimic the general characteristics of the speaker's intonation.

- Resynthesized F_0 (ORIG): To compare the target intonation contours generated by RECOM, CONCAT1, CONCAT2, and BY-RULE to a gold standard, the original F_0 curve of each target utterance was extracted using ESPS and reimposed on the same utterance, thus creating a version called ORIG.

Each of these target contours was imposed on the corresponding natural target utterances using Praat; duration mapping was performed at the foot level; location of the peaks in the accent curves was decided according to the peak placement rules developed by Klabbers and van Santen (2004), as in the case of the target F_0 contours produced by recombinant synthesis (RECOM).

Thus, 6 utterance sets, each comprising 5 target F_0 contours applied to the same utterance, were created to conduct the comparative perceptual test. In the perceptual test, the ordering of the 6 sets was randomized, and within each set, the order of presentation of the 5 target utterances was randomized as well. The perceptual test was presented to 10 listeners using WWStim. The five target utterances of each randomized set were presented on the same page, with each set on a different page. The test was performed on one computer with a M-Audio Duo USB audio interface and a high quality AKG headset. The listeners were asked to listen to each of the 30 target utterances, one at a time, and to rate the intonation of each target utterance using the following seven-point scale: (a) very unnatural, (b) unnatural, (c) natural but barely expressive, (d) natural and slightly expressive, (e) natural and moderately expressive, (f) natural and very expressive, and (g) natural and extremely expressive. Our reasoning behind using such a scale is based on the idea that an utterance cannot be both unnatural and expressive. The term "expressive" in the context of synthesized speech means expressing as a rational human would naturally; consequently, an utterance cannot be unnatural and expressive at the same time.

Of the 10 listeners included in this perceptual test, 4 were individuals from our university and 6 were individuals from outside the university. All listeners were native speakers of American English. All ten were naive subjects. The results formed a 30 stimuli $\times$ 10 listeners score matrix T.

Results of the perceptual test comparing intonation generation in recombinant synthesis to other methods

	Very unnatural/ Unnatural	Natural		
		Barely/Slightly expressive	Moderately expressive	Very/Extremely expressive
ORIG	0	0	26.7	73.3
RECOM	8.3	8.3	60.0	23.3
CONCAT1	45.0	8.3	38.3	8.3
CONCAT2	23.3	6.7	55.0	15.0
BY-RULE	6.7	28.3	51.7	13.3

Table 4.5: Percentages calculated from subject score matrix.

In the score matrix T, each intonation generation method has 60 (6×10) scores associated with it. A summary of these scores is presented in Table 4.5. Each entry in the table represents a percentage of the 60 scores associated with each intonation generation method.

From this table we see that the subjects perceived the intonation of resynthesized natural speech (ORIG) to be either moderately expressive or very/extremely expressive. Intonation generated by RECOM was perceived as moderately expressive a majority of the time (60%). It was also perceived as very/extremely expressive a substantial percent of the time (23.3%). Compared to CONCAT1, CONCAT2, and BY-RULE, intonation generated by RECOM is closest to ORIG. Intonation generated by CONCAT1 was considered unnatural significantly more often than intonation generated by other methods. Intonation generated by CONCAT2 was perceived as less unnatural than CONCAT1 (23.3% for CONCAT2 versus 45% for CONCAT1). Intonation generated by CONCAT2 was also perceived as moderately expressive more than 50% of the time. This shows the positive effect of peak-matching CONCAT2 to RECOM, and thus indirectly affirms the method of specifying peak height in RECOM. Intonation generated by BY-RULE was perceived as unnatural least often compared to RECOM, CONCAT1, and CONCAT2. This speaks to the reliability of rule-based intonation generation. However, intonation generated via BY-RULE was perceived as moderately or very/extremely expressive less often than intonation generated by either RECOM or CONCAT2. This shows that the reliability of rule-based intonation generation comes at the price of expressiveness.

Source	Sum of Squares	degrees of freedom	Mean square	F	p-value
Model	234.553	4	58.638	7.087	0.0003
Error	297.847	36	8.274		

Table 4.6: Summary of results of one-way repeated measures ANOVA.

Further analyses of these subject scores using the same four statistical tests that were used to analyze the subject scores obtained from the perceptual test comparing different intonation generation methods in terms of naturalness (Section 4.2.2) is presented below. Since both perceptual tests were structurally equivalent, the suitability of the statistical tests in terms of their assumptions and implications (discussed in depth in Section 4.2.2) holds for this perceptual test as well.

Repeated measures ANOVA: In the repeated measures framework, this perceptual test has one factor — Intonation Methods — which has five levels: ORIG, RECOM, CONCAT1, CONCAT2, and BY-RULE. Thus, we have a one-way repeated measures design with 5 levels and 10 subjects. Under the repeated measures design, we make the same assumptions about this dataset as were made for the dataset obtained by the perceptual test discussed in Section 4.2.2.

Before performing the repeated measures ANOVA, the null hypothesis $H0$ and the α value were set as follows:

- $H0$: there is no difference between the different types of intonation methods.
- $\alpha = 0.05$.

The one-way repeated measures analysis was performed using the statistical package R. The score matrix T was reshaped in the same way the score matrix S was reshaped in Table 4.1.

Since Intonation Methods was crossed with the random factor Subject in this repeated measures design, an error term for Intonation Methods was also specified. A summary of the results obtained by the repeated measures ANOVA is shown in Table 4.6. These results show that there is a significant difference (p-value = 0.0003) in expressiveness among the five methods of intonation generation.

A boxplot visually depicting the difference between the five methods is shown in Figure 4.9. Two points stand out from this boxplot: (1) the interquartile range of RECOM

is distinctly narrower than that of of CONCAT1, CONCAT2, and BY-RULE; and (2) the median value of RECOM is higher than that of CONCAT1, CONCAT2, and BY-RULE, though less than the median value of ORIG (re-synthesized natural speech). These two points together indicate that most of the subjects considered the intonation produced by recombinant synthesis to be more expressive than intonation produced by synthetic target F_0 contours (as in diphone synthesis) and intonation produced by concatenation of raw F_0 curves (as in unit selection synthesis). However, intonation produced by recombinant synthesis was perceived as less expressive than the intonation of natural speech.

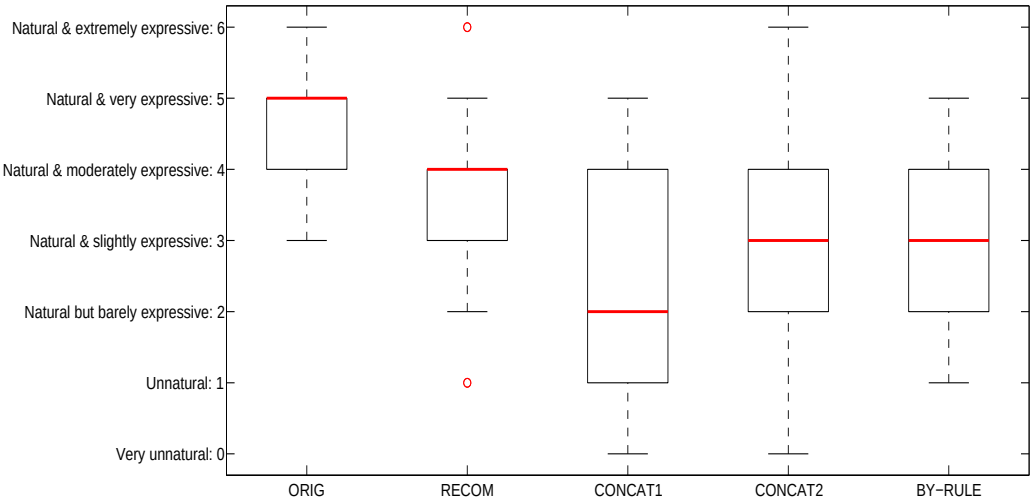

Figure 4.9: Boxplot depicting the difference in the perceived expressiveness of the five intonation generation methods.

Planned t-tests: As in the previous perceptual test, we performed three planned t-tests to test three key predictions, according to which we expected RECOM to be better than CONCAT1, CONCAT2, and BY-RULE. Since we had a directional sense about this analysis, one-tailed t-tests were performed. All t-tests were performed using MATLAB. For each t-test, the null hypothesis $H0$ and the alternative hypothesis HA were set as follows:

- $H0$: mean expressiveness of RECOM = mean expressiveness of the other method.
- HA: mean expressiveness of RECOM > mean expressiveness of the other method.

Again, as before, we used the Bonferroni t-test. Thus, using a family-wise alpha of 0.05, the criterion to reject the null hypothesis was $0.05/3 = 0.0167$. The results of the planned t-tests are shown in Table 4.7. All three t-tests have 9 degrees of freedom. Since the p-values for all three t-tests are less than 0.0167, we can reject the null hypothesis and conclude that the subjects judged the intonation generated by RECOM to be significantly more expressive than the intonation generated by CONCAT1, CONCAT2, and BY-RULE. This analysis thus strongly supports our expectation that recombinant synthesis will produce more expressive speech than other synthesis methods.

	t-test	p-value	Significant?
RECOM - CONCAT1	5.5739	0.0002	Yes
RECOM - CONCAT2	3.8739	0.0019	Yes
RECOM - BY-RULE	2.8567	0.0094	Yes

Table 4.7: Planned t-tests showing that the subjects judged intonation generated by RECOM to be significantly more expressive than intonation generated by other methods.

One way between subjects ANOVA: Though the previous tests revealed that subjects perceived a significant difference among the five intonation generation methods, this information is not very well highlighted by a simple comparison of the mean ratings corresponding to the different methods (ORIG = 4.9333, RECOM = 3.6667, CONCAT1 = 2.2667, CONCAT2 = 3.0500, and BY-RULE = 3.1167). This is perhaps because different subjects used different range of ratings to score the perceptual test. To examine this idea, we performed a one-way between subjects ANOVA using MATLAB. The null hypothesis is that there is no difference between the subjects' rating range and the α value is 0.05. The ANOVA results are shown in Table 4.8. The F-score equals 12.35 and the p-value equals 1.11e-16. Since the F-score is greater than 1 and p-value < 0.05, we can reject the null hypothesis and conclude that there indeed is a difference between the subject ratings. A boxplot showing the rating range of different subjects is shown in Figure 4.10.

Source	Sum of squares	Degrees of freedom	Mean square	F	p-value
Between subjects	203.9867	9	22.6652	12.3458	1.1102e-016
Error	532.4	290	1.8359		

Table 4.8: Summary of results of one-way ANOVA on subjects.

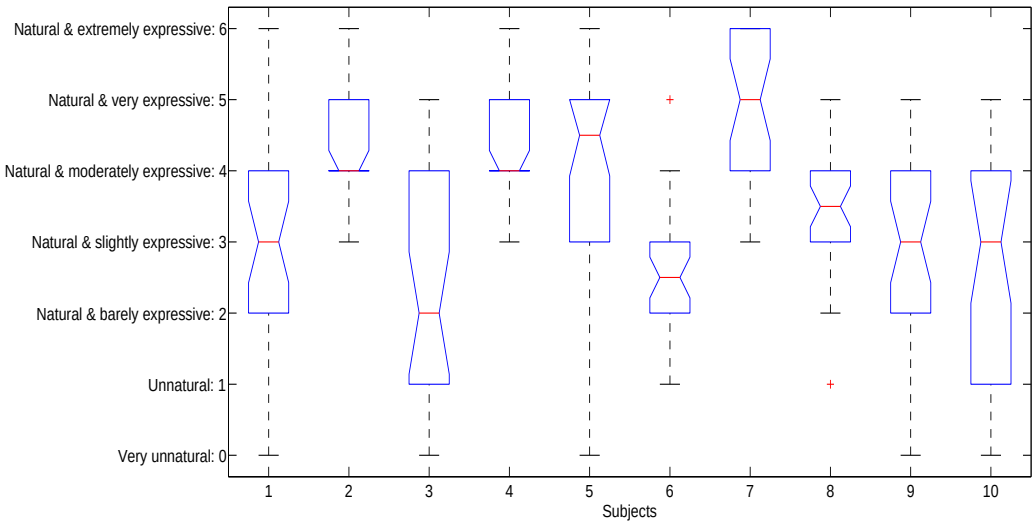

Figure 4.10: Boxplot showing the difference in subject ratings.

Principal component analysis: To obtain a picture of the combined scores corrected for some subjects using a different range of ratings or not being in line with the majority of subjects, we performed a principal components analysis (PCA) on the score matrix T, after its columns were transformed into z-scores. MATLAB was used to conduct this analysis. The resulting averages are shown in Figure 4.11. It shows that intonation generated by RECOM was considered to be more expressive than intonation generated by BY-RULE, CONCAT1, and CONCAT2. However, it also shows that the subjects did not consider intonation generated by RECOM to be as expressive as re-synthesized natural speech (ORIG).

4.2.6 Discussion of results

Based on the analyses in the previous section, we can conclude that our intonation generation method RECOM produces more expressive intonation than methods BY-RULE, CONCAT1, and CONCAT2. The better performance of RECOM over the other methods can be attributed to the key differences outlined in the following paragraphs.

RECOM vs. BY-RULE: A key difference between an F_0 contour generated by

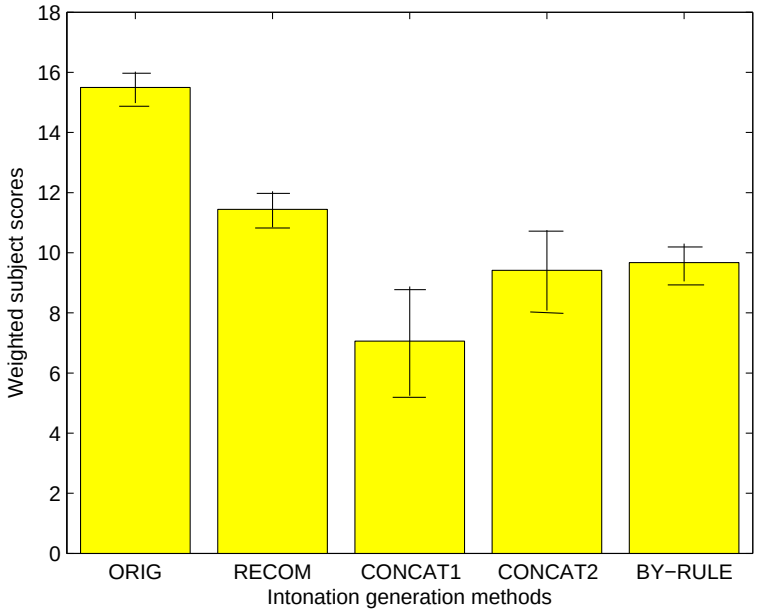

Figure 4.11: Means and standard deviations of different methods of F_0 generation. The bar graph shows the means, and the line above each bar indicates the standard deviation.

RECOM and a corresponding F_0 contour generated via BY-RULE is that in the former, accent curves of the same utterance type (i.e., declarative, continuation rise, or interrogative) may vary in terms of their detailed shape, while in the latter, accent curves of the same utterance type have exactly the same shape. The term "detailed shape" refers to shape particulars such as slopes of the rise and fall of an accent curve, width of the accent curve, and presence (or absence) of a plateau in the peak, and not the general shape of the accent curve. An example of this difference is seen in Figure 4.7 (set a). Both accent curves in the F_0 contour generated by SLAM (which is a BY-RULE method) have the same shape though the second accent curve is smaller in size. The two accent curves in the F_0 contour generated by recombinant synthesis (in the figure denoted DECOM_1) do not have the same shape — the first accent curve has saddle-like shape at the peak but the second accent curve does not; rather, the second accent curve is so flat that its shape is almost obscured by the first accent curve overlapping with it.

The reason for this key difference is that accent curves in BY-RULE are generated by statistically-based rules, whereas accent curves in RECOM (and in DECOM_1 and

DECOM$_2$) are obtained from natural F_0 contours using PRISM. The additional variability in the detailed shape of the accent curves helps F_0 contours generated by RECOM to approximate natural F_0 contours more closely. Consequently, compared to BY-RULE generated F_0 contours, RECOM generated F_0 contours approximate the expressiveness of natural F_0 contours better (as indicated by Figure 4.11). However, this additional variability needs to be handled carefully. If the candidate accent curve selected from the accent curve corpus has a variability that is inappropriate for the target, it can cause the F_0 contour to sound unnatural. See Table 4.5; notice that though intonation produced by RECOM was rated moderately expressive and very/extremely expressive more often than intonation produced via BY-RULE, it was also rated as unnatural slightly more often than intonation produced by BY-RULE (8.3% for RECOM versus 6.9% for BY-RULE). Thus, the expressiveness of recombinant synthesis comes with some risk of unnaturalness.

RECOM vs. CONCAT1: The key difference between an F_0 contour generated by RECOM and an F_0 contour generated by CONCAT1 is that the former is a superposition of accent curves and phrase curves — extracted from natural F_0 curves using PRISM — while the latter is a concatenation of raw F_0 curve fragments. In CONCAT1, we do not have any explicit control over the phrase curves and the accent curves. As a result, concatenation of the raw F_0 curve fragments may produce F_0 contours whose phrase curves and accent curves may be inappropriate for the target. Consider Figure 4.12. The F_0 contours shown in this figure correspond to the phrase "and whatsoever there was to eat." in the utterance "They shot hares, wild deer, birds and pigeons, and whatsoever there was to eat." The F_0 contour produced by CONCAT1 has a gently rising phrase curve while the F_0 contour produced by RECOM has a falling phrase curve. Given that it has been shown that utterance-final phrase curves in natural F_0 contours have a declining slope with the declination becoming steeper in the last foot of the phrase (Klabbers and van Santen 2004), the RECOM generated F_0 contour is a more appropriate target contour than the CONCAT1 generated F_0 contour. The latter contour may be perceived as unnatural.

RECOM vs. CONCAT2: F_0 contours were generated by CONCAT2 using the same process as CONCAT1, except that the F_0 contours produced by CONCAT2 were peak-matched to the F_0 contours produced by RECOM. In the process of peak-matching,

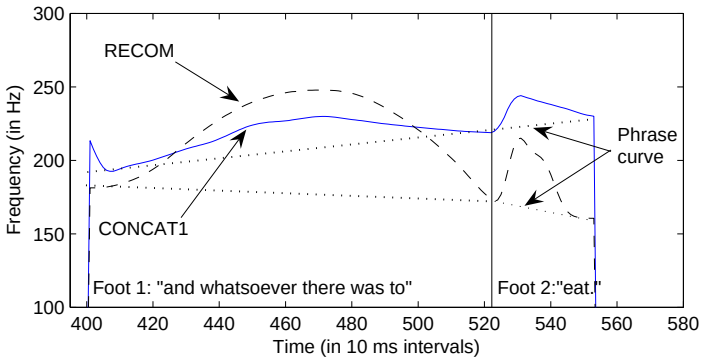

Figure 4.12: Comparison of F_0 contours produced by RECOM and CONCAT1.

we may have indirectly caused the phrase curves and accent curves underlying each CONCAT2 generated F_0 contour to be more like the phrase curves and accent curves underlying each corresponding RECOM generated F_0 contour. Consequently, the problem of inappropriate component curves found in the case of F_0 contours generated by CONCAT1 may have been ameliorated. However, in CONCAT2, there is still no way to control the accent curve separately from the phrase curve. Thus, during peak-matching, we have to vertically shift each F_0 curve fragment in its entirety. As a result, pitch excursions in F_0 contours produced by CONCAT2 may be different — generally smaller — than pitch excursions in F_0 contours produced by RECOM. This is a key difference between the F_0 contours generated by RECOM and the F_0 contours generated by CONCAT2. As an example, consider Figure 4.13. The F_0 contours shown in this figure correspond to the phrase "and whatsoever there was to eat." in the utterance "They shot hares, wild deer, birds and pigeons, and whatsoever there was to eat." Notice that both F_0 contours have a falling phrase curve. However, the accent curves underlying the F_0 contour generated by CONCAT2 have smaller amplitudes compared to the corresponding accent curves underlying the F_0 contour generated by RECOM. As a result, even though both F_0 contours may sound natural, the F_0 contour generated by CONCAT2 sounds less expressive than the F_0 contour generated by RECOM. The other concern with CONCAT2 is that the difference in mean height of two peak-matched F_0 curve fragments may still negatively affect

perception of naturalness, in spite of our attempt to connect the two fragments seamlessly via interpolation and smoothing.

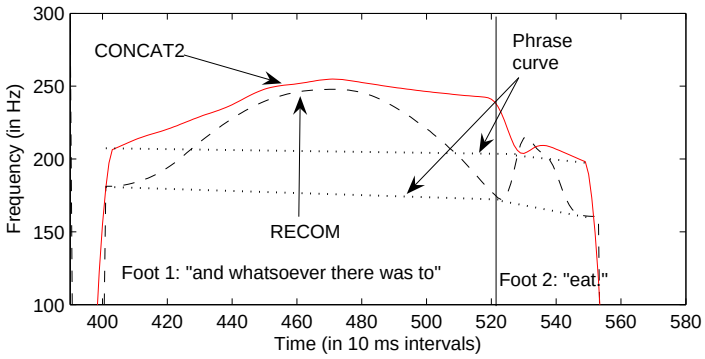

Figure 4.13: Comparison of F_0 contours produced by RECOM and CONCAT2.

4.3 Conclusion

In this chapter, we demonstrated intonation generation in recombinant synthesis via a proof-of-concept implementation. We conducted two perceptual tests to compare this intonation generation method to other intonation generation methods. The perceptual tests showed that the intonation generated by recombinant synthesis is more natural-sounding and more expressive than intonation generated by concatenation of raw F_0 curves (as in unit selection synthesis) and intonation generated by statistical, linguistic, and heuristic rules (as in traditional concatenative synthesis). Thus, our intonation generation method is a promising alternative to the latter two methods. Comparing recombinant intonation generation with raw pitch concatenation, we can conclude that the continuity of the target F_0 contours provided by the recombinant intonation generation method clearly outweighed any pitch modification distortion.

Chapter 5

Intonational Correlates of Perceived Prominence From A Superpositional Perspective

5.1 Background

Intonation serves several communicational functions. One of the primary functions of intonation is to aid *perceived prominence*, "a psycholinguistic measure of the degree of perceived salience of a syllable and consequently of the word or larger unit in which that syllable is the most prominent" (Monaghan 2002, p.99). The effect of intonation on perceived prominence has been the subject of much research. Several researchers (Gussenhoven et al. 1997, Rietveld and Gussenhoven 1985, Terken 1991, and Pierrehumbert 1979) found that the position and height of peaks in the F_0 contour influence the perception of prominence. One of their most crucial findings is the *effect of declination* according to which a peak occurring later in an F_0 contour must have a lower height than a peak that occurs earlier in the contour in order to be perceived as equally prominent.

The effect of intonation on the perception of prominence has also been demonstrated — though slightly less directly than the aforementioned studies — by research on *focus* perception. The focus of an utterance, semantically its most salient part, is generally the most prominent part of the utterance; thus, the effect of intonation on perception of focus can be extrapolated to its effect on perception of prominence. A study by Rump and Collier (1996) showed that in utterances containing two F_0 peaks, perception of focal condition (i.e., *neutral* focus, *early* or *late single contrastive* focus, and *double contrastive*

focus) was signalled by relative peak heights, all other things being equal. Parts of an utterance that are marked by neutral focus convey new information. Parts of an utterance that are marked by contrastive focus convey information that is in contrast to other information presented either in the same utterance (double contrastive focus) or in a preceding utterance (single contrastive focus). In the single contrastive focus condition, early or late indicate the position of the focussed part of the utterance relative to the rest of the utterance. In the Rump and Collier study, it was observed that a high first peak and a much lower (or absent) second peak was perceived as early single contrastive focus, whereas a high second peak and a lower first peak was perceived as late single contrastive focus. Neutral focus and double contrastive focus were each signalled by relatively equal peaks — though in some cases, the second peak was slightly lower. Neutral focus and double contrastive focus were distinguished from each other by the height of the first peak; it was higher in the case of double focus than in the case of neutral focus.

Similar studies of the relation between focus conditions and intonation were also conducted by Cooper et al. (1985), Eady and Cooper (1986), and Welby (2003). Eady, Cooper, and their colleagues observed that different focal conditions were related to different F_0 patterns. Welby found that listeners were most attentive to the presence or absence of pitch accents in determining focal structure, though phrase boundaries were perhaps also important. These studies thus demonstrated the influence of intonation on the perception of focus, and by extrapolation, on the perception of prominence.

Besides intonation, other aspects of speech that are considered to influence the perception of prominence are loudness and duration. Early studies of prominence (Fry 1955, 1958) asserted that the prominent syllable was marked, in decreasing order of importance, by duration, F_0, and loudness. However, these studies were performed only on minimal pairs of single, isolated, synthesized English words that were distinguished by a difference in stress placement rather than focus, so the results need to be interpreted accordingly. More recent studies (Beckman 1986, Silipo and Greenberg 1999, Silipo and Greenberg 2000, Turk and Sawusch 1996, and Kochanski et al. 2005) also found duration and loudness to be good detectors of prominence. Thus, it is clear that F_0 is not the only acoustic correlate of perceived prominence.

Nevertheless, most studies have found perceived prominence to be primarily a function of F_0; see Terken and Hermes (2000) for a comprehensive review of the literature on prominence. Most of the prominence studies, however, have been conducted either in a sequential framework of intonation or in a quasi-superpositional framework (as in Gussenhoven et al. 1997); only a few studies have been performed in the superpositional framework (such as Mixdorff and Widera 2001 and Vainio et al. 2003). The studies in the sequential framework have considered - and found to be significant - the influence of several intonational parameters, such as pitch peak height (absolute or relative), pitch accent shape, rate of baseline declination, and pitch onset height. However, studies in the superpositional framework (almost all of which were based on the well-known Fujisaki model) have so far investigated the influence of only one intonational parameter, namely, the height of the accent component of the F_0 contour.

Broadly speaking, accent component height in the superpositional framework corresponds to pitch peak height relative to the baseline in the sequential framework. The other intonational parameters that have been examined in the sequential framework also have corresponding parameters in the superpositional framework, but they have not been studied yet.

We aim to fill this gap by investigating how perceived prominence is influenced by the following four parameters of intonation obtained in the superpositional framework: (1) accent component height, (2) accent component shape (corresponding to pitch accent shape[1]), (3) phrase component onset height (corresponding to pitch onset height[1]), and (4) phrase component slope (corresponding to rate of baseline declination[1]). The importance of studying these parameters together is that it will help us to determine (a) the individual effects of these parameters, (b) their combined effect, and (c) their relative importance to perceived prominence. Our investigation is conducted through a perception experiment described in Section 5.3. A formal description of the aforementioned four parameters that are varied in the perception experiment is given in Section 5.2.

[1]Pitch accent shape, pitch onset height, and rate of baseline declination are three intonational parameters in the sequential framework.

5.2 Parameter definitions

For the purposes of this investigation, we consider the F_0 contour to be a summation of a phrase curve component $P(t)$ and one or more accent curve components $A_i(t)$ (Equation 5.1).

$$F_0(t) = P(t) + \sum_i A_i(t) \tag{5.1}$$

$$P(t) = \text{concatenate}(p_i), \quad p_i(t) = \beta_i(t) + \gamma \tag{5.2}$$

$$A_i(t) = h_i \times time\_warped(E_i(t)), \quad |E_i(t)| = 1 \tag{5.3}$$

Here the phrase curve $P(t)$ is represented by a smoothed piecewise linear curve (Equation 5.2). The slope of $P(t)$ (β in the simplest case) is called the *phrase curve slope*. The value of P(0) (γ in the simplest case) is called the *phrase curve onset height*.

Each accent curve $A_i(t)$ is represented by a parameterized time-warp of a class-specific (e.g., declarative versus continuation rise) template curve $E_i(t)$ that has been multiplied by a height parameter h_i (Equation 5.3). We call this height parameter the *accent curve height*. *Accent curve shape* is characterized by both the time-warp function and the template curve E_i. It is a non-controversial fact that accent curves that differ in terms of their underlying templates differ perceptually; e.g., the difference between a declarative accent curve and an interrogative accent curve is easily perceived. However, whether accent curves that differ in terms of their time-warp functions are also perceptually differentiable is still an open question. In the current study, we examine the difference between two such accents: normal (default) accents versus contrastive accents. In terms of accent curve shape, these two accents can be distinguished in terms of the location of the peak in the accent curve. Thus, accent curve shape will be represented by the location of the peak in the accent curve within the context of this investigation.

5.3 Perception experiment

A stimulus set of 144 utterances was generated for the perception experiment using the following three steps:

1. Recording the base set of natural utterances: Three pairs of utterances $U_{a,b}$ were

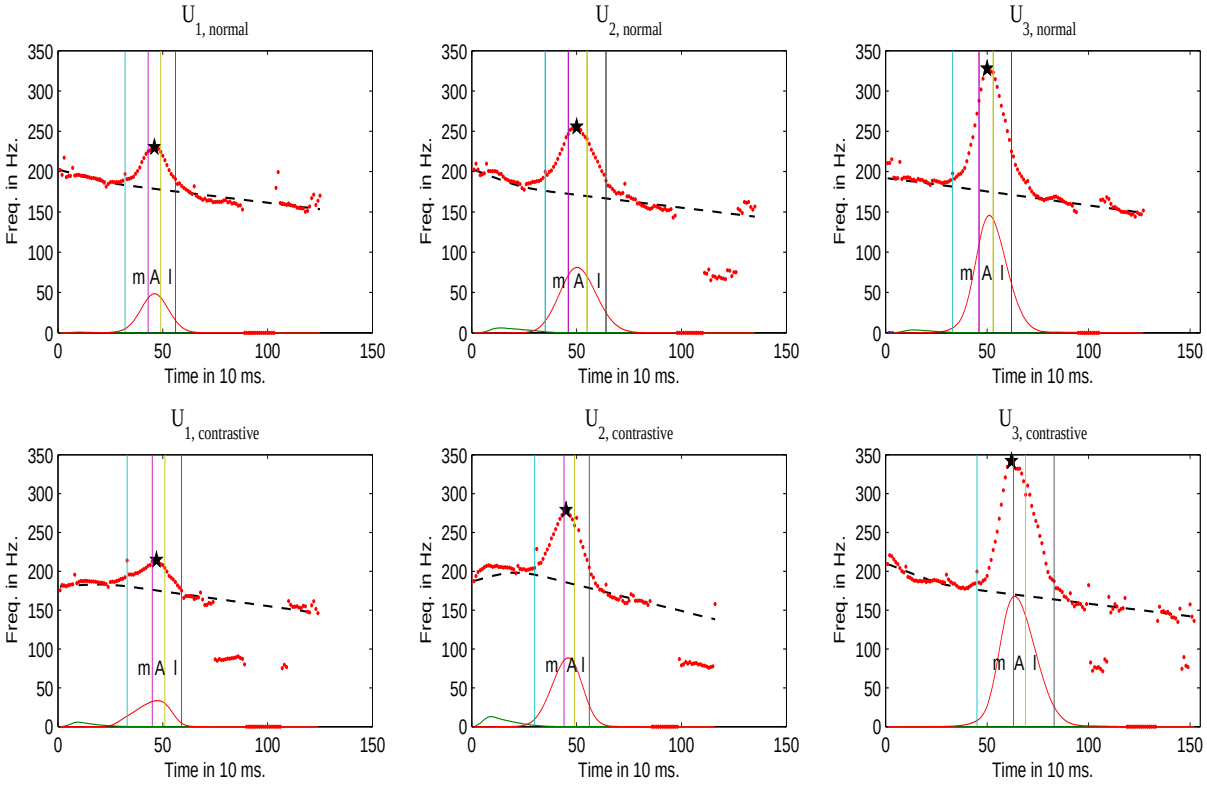

Figure 5.1: F_0 contours (and associated phrase curves and accent curves) of the 6 utterances in the base set. Each utterance is labeled $U_{a,b}$, where a indicates pair number ($a = 1, 2, 3$), and b indicates accent type ($b = normal, contrastive$). The dashed lines represent the phrase curves, the continuous lines represent the accent curves, and the dotted lines represent the natural F_0 curves.

recorded from one adult female native speaker of American English, where a indicates pair number ($a = 1, 2, 3$), and b indicates accent type ($b = normal, contrastive$). The text of each pair of utterances was "Now MOLLY Mel is here." and "No, now MOLLY Mel is here." The speaker was instructed (i) to emphasize the capitalized word "MOLLY", (ii) to keep the degree of emphasis in both utterances of the same pair approximately the same, and (iii) to increase the degree of emphasis from one pair to the next.

The word "No" in the second utterance of each utterance pair was meant to put the speaker in a corrective/contrastive mindset so that she produced a contrastive utterance. However, after recording, the part of the waveform corresponding to "No" was removed,

so that all six utterances had exactly the same textual content. Thus, a base set of six natural utterances was created, such that within-pair, the utterances varied in terms of accent type (normal versus contrastive) while between-pair, the utterances varied in terms of accent curve height. The stimulus set was generated by manipulating the F_0 contours of the utterances in the base set. The F_0 contours of the utterances in the base set are shown in Figure 5.1.

To create natural sounding intonation, F_0 contour manipulation must be accompanied by duration manipulation. However, in the context of this investigation, duration manipulation will introduce additional factors that may confound the observed effects of intonational parameters on perceived prominence. So, to observe the effects of intonational parameters on perceived prominence unconfounded by the effects of duration manipulation, we created the base set described above. During stimulus set generation, each natural F_0 contour from the base set was only manipulated so far as to not require any corresponding duration manipulation. However, because the three pairs of utterances in the base set varied substantially in the degree of emphasis, we were able to obtain a sufficiently large number of stimuli F_0 contours.

2. Decomposing the natural F_0 contours: The F_0 contours of the utterances in the base set were decomposed into phrase curves and accent curves using PRISM. The phrase curves and the accent curves produced by decomposing the F_0 contours of the utterances in the base set are shown in Figure 5.1. Notice that the estimated component curves look utterly reasonable: each phrase curve has a declining slope and each accent curve is a single-peaked function that rises from 0 to a peak value and again descends to 0, as is expected of the phrase curve and the accent curve underlying a single-phrase declarative utterance containing one emphasized word.

3. Generating the stimulus set of quasi-natural F_0 contours: From the natural F_0 contour of each utterance in the base set, 24 quasi-natural F_0 contours were generated as follows. First, a set of values called $Equipeaks_{a,b}$ was computed. $Equipeaks_{a,b} = \{OrigPeak_{a,b} - 25Hz, OrigPeak_{a,b}, OrigPeak_{a,b} + 25Hz\}$, where $OrigPeak_{a,b}$ is the absolute peak height in the natural F_0 contour associated with the utterance $U_{a,b}$. $OrigPeak_{a,b}$ is denoted by a star in Figure 5.1.

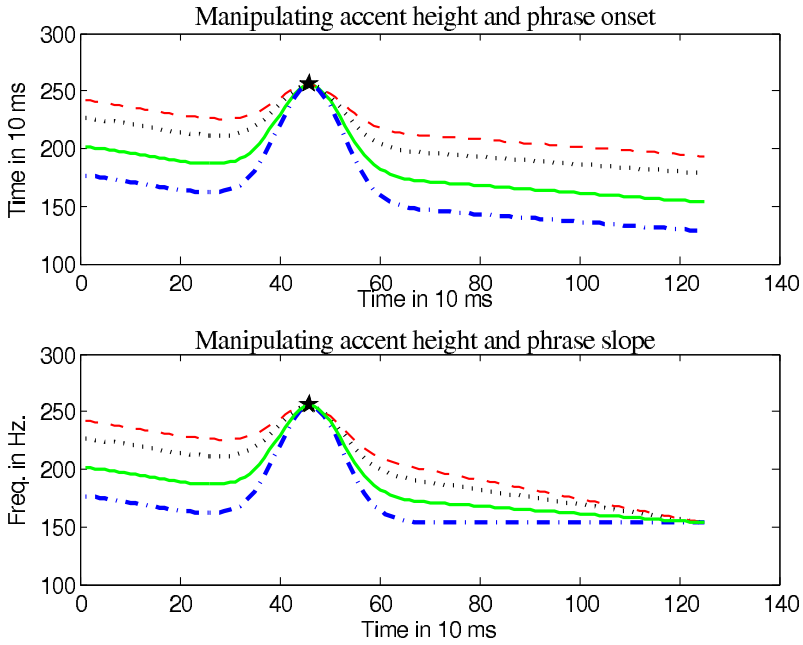

Figure 5.2: The manipulation of the base set to obtain the stimulus set is illustrated here. The star indicates the equipeak value. In the upper panel, each equipeak is attained by four different combinations of accent curve height and phrase curve onset height. In the lower panel, each equipeak is attained by four different combinations of accent curve height and phrase curve slope.

The set $Equipeaks_{a,b}$ was so named because each element of this set, $Equipeaks_{a,b}(i)$, represents the absolute peak height of eight different quasi-natural F_0 contours, as shown in Figure 5.2. Each of these eight contours was obtained by combining the curves $\hat{P}_{a,b}$ and $\hat{A}_{a,b}$ according to the superpositional approach outlined in Equations 5.1 through 5.3. The curve $\hat{P}_{a,b}$ represents a modified version of the phrase curve $P_{a,b}$, which was obtained by the decomposition of the natural F_0 contour of the utterance $U_{a,b}$. Similarly, $\hat{A}_{a,b}$ represents a modified version of the accent curve $A_{a,b}$, which was obtained by the decomposition of the natural F_0 contour of $U_{a,b}$.

$$\hat{A}_{a,b} = A_{a,b} \times (1 + \frac{d}{h}) \tag{5.4}$$

$$\hat{P}_{a,b} = P_{a,b} + d \tag{5.5}$$

$$\hat{P}_{a,b} = P_{a,b} + f(t) \tag{5.6}$$

$A_{a,b}$ was modified only in terms of height to produce $\hat{A}_{a,b}$, as shown in Equation 5.4. In

this equation, h denotes the height of $A_{a,b}$ and d denotes the difference in height between $\hat{A}_{a,b}$ and $A_{a,b}$. Four different d values were used ($d = \{-40, -25, 0, 25\}$) to produce four different $\hat{A}_{a,b}$ curves. The peak location in $\hat{A}_{a,b}$ was the same as the peak location in $A_{a,b}$. Thus, the shape of $\hat{A}_{a,b}$ was identical to the shape of $A_{a,b}$, because in this study accent curve shape was parameterized by peak location.

For each $\hat{A}_{a,b}$, two different types of corresponding $\hat{P}_{a,b}$ curves were produced. (S1) $\hat{P}_{a,b}$ was produced by changing the onset height of $P_{a,b}$ by d, as shown in Equation 5.5. There was no modification in the original slope of $P_{a,b}$ but its onset height was systematically modified. (S2) $\hat{P}_{a,b}$ was produced by adding a line $f(t)$ to $P_{a,b}$, as shown in Equation 5.6. The line $f(t)$ was constrained so that $f(t = T) = 0$ and $f(t = pl) = d$, where T denotes the end time-point of the $P_{a,b}$ and pl denotes the time-point corresponding to the peak location in $A_{a,b}$. By adding the constrained line to $P_{a,b}$, we systematically modified the original slope of $P_{a,b}$. However, as a consequence of this slope modification method, the onset height of $P_{a,b}$ was also modified, while its offset height remained unmodified. See the lower panel of Figure 5.2; $\hat{P}_{a,b}$ was produced by pivoting $P_{a,b}$ about the offset. The reason $P_{a,b}$ was not pivoted about the onset (i.e., holding onset height constant while slope was modified) was to avoid the inadvertent generation of a $\hat{P}_{a,b}$ curve with a rising slope.

The two types of phrase modifications were so formulated to ensure that superpositionally combining either type of $\hat{P}_{a,b}$ curve with the $\hat{A}_{a,b}$ curve produced F_0 contours of the same absolute peak height $Equipeaks_{a,b}(i)$, even though different phrase-level intonational parameters were modified for each type. The first type of $\hat{P}_{a,b}$ curve was produced by modifying the phrase curve onset height. Superpositionally combining this type of $\hat{P}_{a,b}$ curve with the $\hat{A}_{a,b}$ curve produced four F_0 contours — corresponding to the four d values — that differed from each other in terms of accent curve height and phrase curve onset height, as shown in the upper panel of Figure 5.2. The second type of $\hat{P}_{a,b}$ curve was produced by modifying phrase curve slope. Superpositionally combining this type of $\hat{P}_{a,b}$ curve with the $\hat{A}_{a,b}$ curve produced four F_0 contours — corresponding to the four d values — that differed from each other in terms of accent curve height and phrase curve slope, as shown in the lower panel of Figure 5.2. Thus, eight quasi-natural F_0 contours with the same absolute peak height $Equipeaks_{a,b}(i)$ were produced.

Because per utterance $U_{a,b}$, we had specified three values of $Equipeaks_{a,b}$ (i.e., $OrigPeak_{a,b}-25Hz$, $OrigPeak_{a,b}$, $OrigPeak_{a,b}+25Hz$), a set of 24 ($= 8 \times 3$) quasi-natural F_0 contours per utterance was generated. Each of these 24 quasi-natural F_0 contours was reimposed on $U_{a,b}$ using STRAIGHT to obtain 24 stimulus utterances. Since there are 6 natural utterances in the base set, a total of 144 ($= 6 \times 24$) stimulus utterances were thus obtained.

The 144 stimulus utterances can be subdivided into two subsets, S1 and S2. S1 consists of the 72 stimulus utterances whose F_0 contours are systematically varied in terms of accent curve height and phrase curve onset height, while S2 consists of the 72 stimulus utterances whose F_0 contours are systematically varied in terms of accent curve height and phrase curve slope. A third intonational parameter that is systematically varied in both sets is accent curve shape. The variation in accent curve shape ensues from the fact that though corresponding $\hat{A}_{a,b}$ and $A_{a,b}$ have identical shapes, the $A_{a,b}$ related to each of the 6 utterances in the base set differs in shape from the others due to a difference in peak location. (Observe from Fig. 5.1 that the peak location relative to the phones labels /m/, /A/, and /l/ for each of the 6 utterances in the base set is different.)

This quasi-natural stimulus set differs from the completely synthetic (as in Gussenhoven et al. 1997 and Terken 1991) or completely natural (as in Mixdorff and Widera 2001) stimulus sets that are commonly used in prominence studies. For example, the experiments reported in Gussenhoven et al. (1997) were performed on a stimulus set consisting of synthetic F_0 contours, each of which comprised a baseline with one or two rise-and-fall combinations superimposed on it. These rise-and-fall combinations were linear segments that either joined together to form a sharp peak or were linked by a horizontal segment at the peak to form a somewhat flat-hat contour. A criticism against such contours shapes is that they do not occur in natural speech (van Santen et al. 2002). Although some earlier studies reported that such stylized pitch contours cannot be distinguished from natural pitch contours, van Santen et al. argue that many of these experiments were carried out with low-quality TTS systems whose more glaring synthesis errors obscured the prosodic differences between stylized pitch contours and natural pitch contours. Van Santen et al. also suggest that performing the same experiments using current TTS technology would

show different results.

On the other hand, using synthetic pitch contours also have an up-side. One can systematically vary synthetic pitch contours in terms each of the parameters of interest; this is something that is practically not possible for completely natural F_0 contours because even very skilled recording artists may not be able to vary one intonational parameter while holding all others constant. This is an important argument against using completely natural F_0 contours.

The quasi-natural stimuli that we have generated for this study of perceived prominence sidesteps each of the aforementioned issues related to the other two kinds of stimuli. Each quasi-natural F_0 contour is composed of phrase curves and accent curves that were obtained from natural F_0 contours; thus, visual comparison of these quasi-natural F_0 contours and natural F_0 contours show a high degree of similarity. However, because the phrase curves and the accent curves were parameterized according to Equations 5.1 to 5.3, we were able to systematically vary the intonational parameters — though the parameters for each base utterance were only varied to the extent that they did not require any change in duration of the base utterance.

5.3.1 Listening protocol

To obtain the prominence judgements on the stimulus set of 144 utterances, a listening test was administered to ten listeners using a CGI-based script, WWStim (Veenker 2001), that automatically presents stimuli to the listeners and transfers their results to a database. Each listener was presented with an unique randomized ordering of the stimulus set. Each utterance in the randomized set was presented on a different page. The test was performed on one computer with a high quality AKG headset. To prevent listener fatigue, the test was performed in two 15 minute sessions, in which the listeners listened to 80 and 64 utterances, respectively.

The listeners were asked to listen to each utterance and rate the emphasis on the word “Molly” on a five-point MOS (Mean Opinion Score) scale marked with the following labels, indicating increasing levels of emphasis (or prominence): not emphasized, slightly

	X1	X2	X3	Y
X1	1.000	-0.389***	-0.180	0.899***
X2		1.000	-0.069	-0.591***
X3			1.000	0.114
Y				1.000

Table 5.1: Correlation matrix for S1[2].

	X1	X2	X4	Y
X1	1.000	-0.389***	-0.243*	0.891***
X2		1.000	-0.088	-0.590***
X4			1.000	0.106
Y				1.000

Table 5.2: Correlation matrix for S2[2].

emphasized, moderately emphasized, very emphasized, and extremely emphasized. Before beginning the listening test, an example set of five utterances, each with a different emphasis level, was played for the listener to make the listener aware of the range of emphasis that he/she expected to hear. The same example set was repeated after every 20 utterances to remind the listener of the emphasis range. Of the ten listeners, nine were native speakers of American English and one was a non-native fluent speaker. Eight of the ten listeners were "prosodically naive", while two were involved with speech research.

5.4 Results

The results of the listening test formed a 144 stimuli $\times$ 10 listeners score matrix. From this score matrix, we computed the mean prominence scores obtained from all subjects for each of the 144 utterances.

5.4.1 Combined and individual influence of accent curve height, accent curve peak location, and phrase curve onset height

A multiple linear regression model was fitted on the mean prominence scores obtained from the 72 stimulus utterances belonging to the subset S1. Mean prominence score (Y) was the response variable, while accent curve height (X1), accent curve peak location (X2), and phrase curve onset height (X3) were the predictor variables. The zeroth-order correlations between these variables are shown in Table 5.1. This table indicates that (1) the accent curve height (X1) has a strong positive correlation with perceived prominence; (2) accent curve peak location (X2) has a moderate negative correlation with perceived

[2]In the correlation matrix, *** indicates that correlation is significant at alpha $\leq$ 0.0001; ** shows correlation is significant at alpha = 0.01; * shows correlation is significant at alpha = 0.05.

prominence; and (3) phrase curve onset height (X3) has a slight, statistically insignificant (p-value > 0.05) positive correlation with perceived prominence.

The linear function resulting from the multiple regression analysis is given by Equation 5.7.

$$Y = 0.0173\ X1 - 0.0162\ X2 + 0.0108\ X3 \tag{5.7}$$

Notice that phrase curve slope is not included in this regression model. The exclusion of phrase curve slope from the regression model should come as no surprise because the phrase curve slopes of the utterances in S1 were not varied from the original slopes of the utterances in the base set, which in turn were not significantly different from each other.

R-Squared	Adjusted R-Squared	Std. Error of Estimate	F-Value	p-value
0.9373	0.9336	0.0664	250.5201	< 0.0001

Table 5.3: Summary of results of multiple linear regression on subset S1.

How well the regression model denoted by Equation 5.7 fits the data is illustrated by the model summary presented in Table 5.3. The R-squared and adjusted R-squared values indicate that over 93% of the variance in mean prominence scores can be predicted by accent curve height, accent curve peak location, and phrase curve onset height.

The beta weights (also known as normalized regression weights) corresponding to each of the predictors are shown in Figure 5.3. Beta weights indicate how much influence each of the independent variables exert on the dependent variable. There are two points of note in this figure. First, the beta weight corresponding to phrase curve slope straddles zero, thus illustrating its insignificant contribution to the regression model. Secondly, and more importantly, the beta weight corresponding to phrase curve onset height (0.25) is greater in magnitude than its zeroth-order correlation with mean prominence scores (0.1140). This indicates that phrase curve onset height is in a *suppression* relationship with another predictor variable in this regression model; that is, the inclusion of phrase curve onset height in the regression equation increases the predictive validity of another variable (or set of variables). Since phrase curve onset height is fairly independent of accent curve peak location (correlation = 0.06), phrase curve onset height must be suppressing

the error in accent curve height. This conclusion is further supported by the observation that the beta weight of phrase curve onset height has the opposite sign of its zeroth-order correlation with accent curve height. The conclusion that phrase curve onset height is in a suppression relation with accent curve height in this regression model is useful in interpreting the squared semi-partial correlations of the predictor variables with respect to the response variable.

The squared semi-partial correlation coefficient of a predictor variable represents the proportion of the total variance in the response variable that is uniquely explained by the predictor variable. For example, accent curve height is correlated with both accent curve peak location (-0.3893) and phrase curve onset height (-0.1803). Thus, the variance common to accent curve height and accent curve peak location is -0.3893^2, and the variance common to accent curve height and phrase curve onset height is -0.1803^2, while the variance specific to accent curve height is $(1 - (-0.3893^2) - (-0.1803^2) =)$ 0.8159. The squared semi-partial correlation coefficient of accent curve height with respect to the mean prominence scores represents how much of the total variance in mean prominence scores is explained by the variance specific to accent curve height.

The squared semi-partial correlation coefficients of accent curve height, accent curve peak location, and phrase curve onset height with respect to mean prominence scores are shown in Figure 5.4. All coefficients are statistically significant ($p < 0.0001$). Here, 60.81% of the variance in mean prominence scores is uniquely explained by accent curve height, 5.11% of the variance is uniquely explained by accent curve peak location, and 21.25% is explained by the variance common to accent curve height and accent curve peak location, while 6.2% of the variance is explained by phrase curve onset height in suppression with accent curve height.

The contribution of phrase curve onset height is interesting. Observe from Table 5.1 that individually, phrase curve onset height only accounts for a non-significant 1.3% $(= 0.114^2 \times 100)$ of the variance in mean prominence scores; however, from its squared semi-partial correlation coefficient, it appears that it explains 6.2% of the variance. This apparent conflict can be interpreted easily if one takes the suppression relation between accent curve height and phrase curve onset height into account. Phrase curve onset height

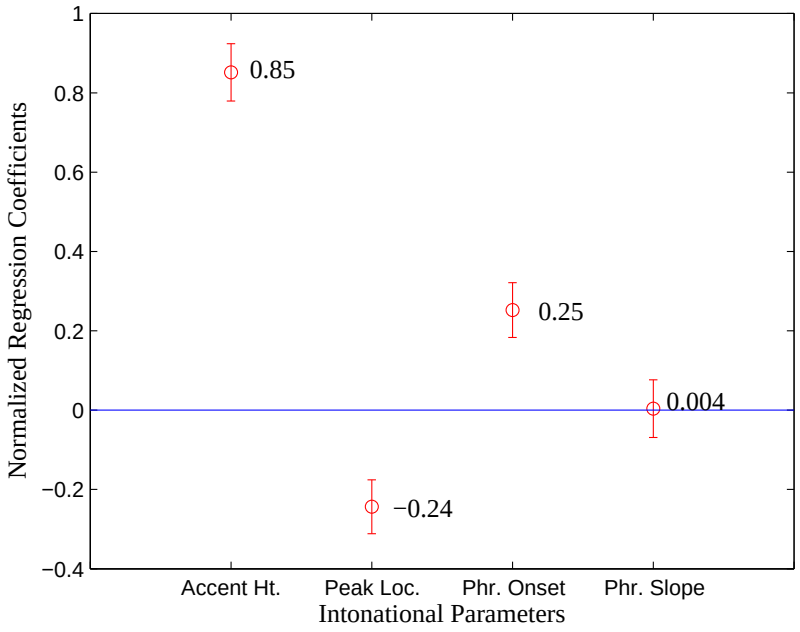

Figure 5.3: Normalized regression coefficients (or beta-weights) of each of the four predictors: accent curve height, accent curve peak location, phrase curve onset height, and phrase curve slope. The circle indicates the beta-weight and the bar represents the confidence interval around it.

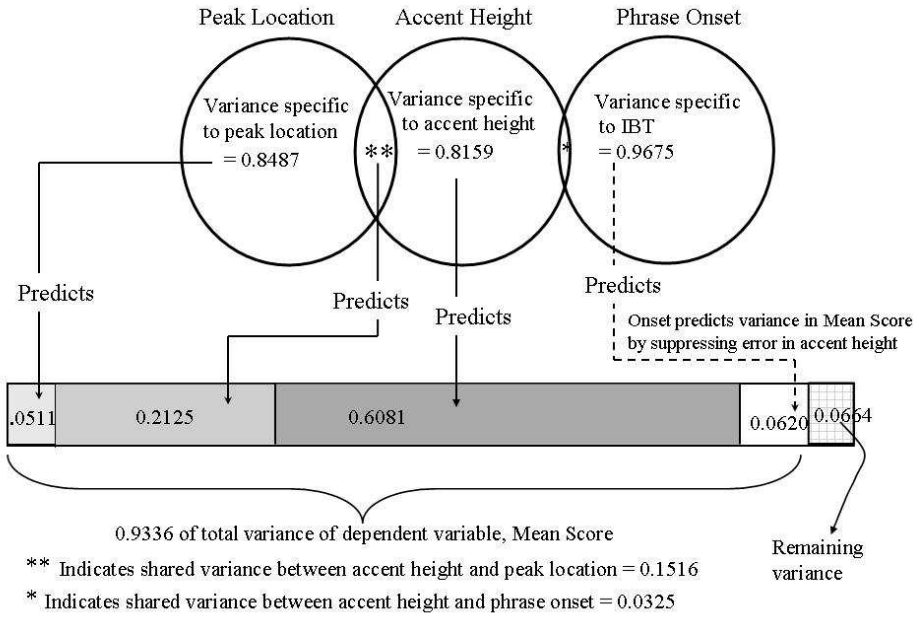

Figure 5.4: Illustration of the relationship between the independent variables and the dependent variable. The independent variables are represented by a Venn diagram and the dependent variable is represented by a bar.

and accent curve height are negatively correlated with each other, while they each are positively correlated with mean prominence scores. So, when the effect of accent curve height is partialled out from the effect of phrase curve onset height, the correlation between phrase curve onset height and mean prominence scores becomes stronger (i.e. from an insignificant 1.3% to a significant 6.2%). Another way of thinking about this phenomena is this. While individually phrase curve onset height barely explains any variance in mean prominence scores, when it is put together with accent curve height in a regression model, it correlates with the source of error (i.e. the variance that is irrelevant to mean prominence scores) in accent curve height and thus increases the prediction by 6.2%. Excluding phrase curve onset height from the regression model causes the explained variance to fall from 93.66% to 87.44%.

From the regression analysis on subset S1 described in the preceding paragraphs, we can draw the following three conclusions:

1. Accent curve height, accent curve peak location, and phrase curve onset height together can predict perceived prominence with a high degree of accuracy, as indicated by the high adjusted R-squared value.
2. Of the three predictor variables, accent curve height is the strongest intonational correlate of perceived prominence, as indicated by its higher zeroth-order correlation coefficient, higher beta weight, and higher semi-partial correlation coefficient. The correlation between accent curve height and perceived prominence is strongly positive; that is, an increase in accent curve height is perceived as an increase in prominence.
3. Of the other two predictor variables, we would argue that accent curve peak location is a stronger correlate of perceived prominence than phrase curve onset height. The reason is that while the unique contribution of accent curve peak location is only 5.11% of the total variance in mean prominence scores, it explains another 21.25% of the total variance together with accent curve height. In comparison, phrase curve onset height explains 6.2% of the total variance but only when accent curve height is present in the regression model. By itself, phrase curve onset height only accounts

for a statistically non-significant 1.3% of the variance in mean prominence scores.

Accent curve peak location and perceived prominence are negatively correlated; that is, the earlier the peak occurs in the accent curve and hence, closer to the start of the vowel in the first syllable, the greater the perceived prominence of the accented foot.

5.4.2 Combined and individual influence of accent curve height, accent curve peak location, and phrase curve slope

So far, the influence of phrase curve slope has not been established. To investigate the influence of phrase curve slope, a regression model was fitted on the mean prominence scores obtained from the 72 stimulus utterances belonging to the subset S2. Mean prominence score (Y) was the response variable, while accent curve height (X1), accent curve peak location (X2), and phrase curve slope (X4) were the predictor variables; for phrase curve slope, we only considered the magnitude, since all utterances had a falling phrase curve, and hence, the same slope sign (i.e., negative). (Phrase curve onset height is not included in this regression model because in the set S2 phrase curve onset height is collinear with phrase curve slope (correlation = 0.96), as a consequence of slope modification.) The zeroth-order correlations between the four variables are shown in Table 5.2. This table indicates that (1) the accent curve height (X1) has a strong positive correlation with perceived prominence; (2) accent curve peak location (X2) has a moderate negative correlation with perceived prominence; and (3) phrase curve slope (X4) has a slight, statistically insignificant (p-value > 0.05) positive correlation with perceived prominence.

Equation 5.8 shows the linear function resulting from the multiple regression analysis.

$$Y = 0.0178\ X1 - 0.0148\ X2 + 1.6846\ X4 + 1.1603 \tag{5.8}$$

The fit of the regression model denoted by Equation 5.8 is illustrated by the model summary presented in Table 5.4. The R-squared and adjusted R-squared values indicate that over 94% of the variance in mean prominence scores can be predicted by accent curve height, accent curve peak location, and phrase curve slope.

The beta weights corresponding to each of the predictors are shown in Figure 5.5. Note

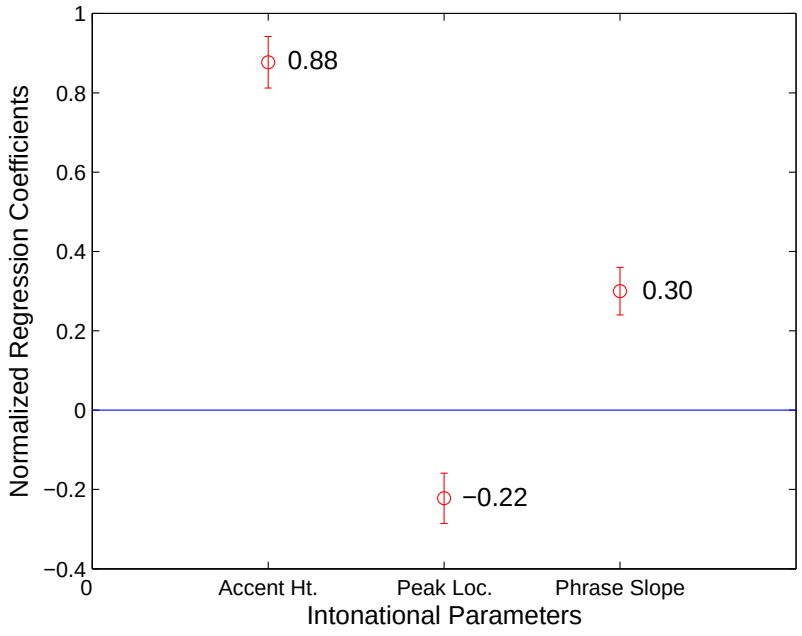

Figure 5.5: Normalized regression coefficients (or beta-weights) of each of the three predictors: accent curve height, accent curve peak location, and phrase curve slope. The circle indicates the beta-weight and the bar represents the confidence interval around it.

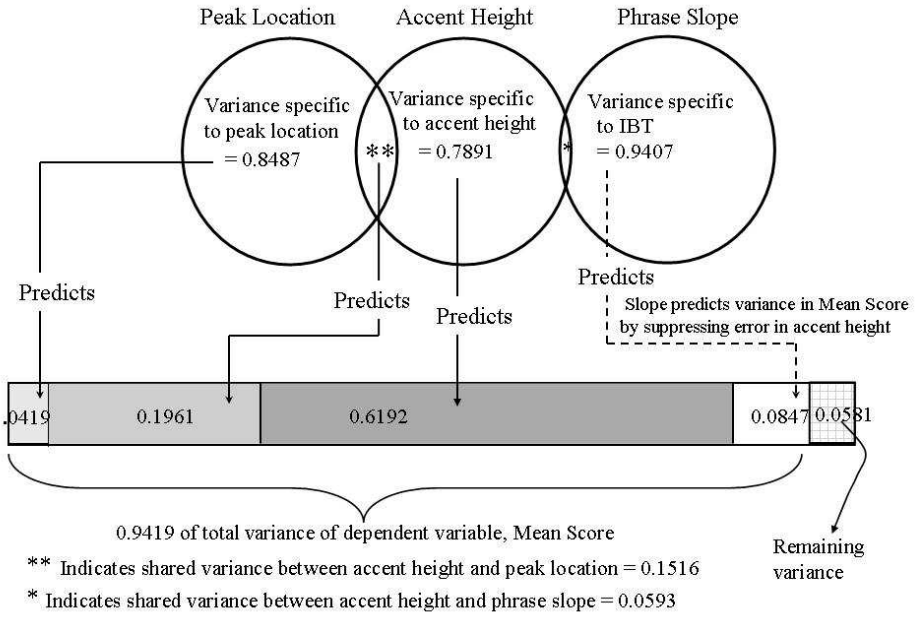

Figure 5.6: Illustration of the relationship between the independent variables and the dependent variable. The independent variables are represented by a Venn diagram and the dependent variable is represented by a bar.

R-Squared	Adjusted R-Squared	Std. Error of Estimate	F-Value	p-value
0.9444	0.9419	0.0581	384.8744	< 0.0001

Table 5.4: Summary of results of multiple linear regression on subset S2.

that the beta weight corresponding to phrase curve slope (0.30) is greater in magnitude than its zeroth-order correlation with mean prominence scores (0.1060). By the same reasoning as in the case of phrase curve onset height, we can conclude that phrase curve slope is in a suppression relationship with accent curve height.

The squared semi-partial correlation coefficients of accent curve height, accent curve peak location, and phrase curve slope with respect to mean prominence scores are shown in Figure 5.6. All coefficients are statistically significant ($p < 0.0001$). Here, 61.92% of the variance in mean prominence scores is uniquely explained by accent curve height, 4.19% of the variance is uniquely explained by accent curve peak location, and 19.61% is explained by the variance common to accent curve height and accent curve peak location, while 8.47% is explained by phrase curve slope in suppression with accent curve height. The role of phrase curve slope in this regression model is similar to that of phrase curve onset height in the regression model denoted by Equation 5.7. In this regression model, individually phrase curve slope accounts for a non-significant 1.12% ($= -0.106^2 \times 100$) of the variance in mean prominence scores; however, it explains 8.47% of the total variance in mean prominence scores by correlating with the source of error in accent curve height.

From the regression analysis on subset S2 described in the preceding paragraphs, we can draw the following three conclusions:

1. The three intonation parameters, accent curve height, accent curve peak location, and phrase curve slope, together can predict perceived prominence with a high degree of accuracy, as indicated by the high adjusted R-squared value.
2. Of the three predictor variables, accent curve height is the strongest intonational correlate of perceived prominence, as indicated by its higher zeroth-order correlation coefficient, higher beta weight, and higher semi-partial correlation coefficient. Accent curve height and perceived prominence are strongly positively correlated.
3. Using the same argument as in the case of phrase curve onset height previously, we

contend that accent curve peak location is a stronger correlate of perceived prominence than phrase curve slope. As before, we find accent curve peak location and perceived prominence to be negatively correlated.

5.4.3 Phrase curve serves as line of reference for evaluation of F_0 peak prominence

Besides determining if and how the four intonational parameters influence perceived prominence, our experimental setup also allows us to investigate the following key questions. Is there a line of reference that listeners use to determine the prominence of peaks in the F_0 contour? If so, is the reference line abstract or observable?

We can infer that there indeed is a reference line that listeners use while making prominence judgements from the following. When different accent curve heights were used to attain the same absolute peak $Equipeaks_{a,b}(i)$ in the F_0 contour (as in Figure 5.2), prominence scores were systematically higher for the higher accent curve heights. This observation is illustrated by Figure 5.7 and Figure 5.8. The connected lines in these graphs are called *equipeak lines*. Each equipeak line is associated with an absolute peak height $Equipeaks_{a,b}(i)$. The circles on the equipeak lines represent the change in mean prominence scores with respect to the changes in accent curve height and phrase curve onset height (or phrase curve slope in Figure 5.8) for the same absolute peak height. In 30 of the 36 equipeak lines, the slope is positively correlated with accent curve height. A two-tailed sign test indicated that this result is significant at $p < 0.0001$. This result implies that perceived prominence is not simply a correlate of absolute peak value; rather, listeners are able to perceive a difference in prominence corresponding to the different accent curve heights that are used to attain the same absolute peak value. This in turn implies that the phrase curve serves as the reference line for the evaluation of prominence, because by definition, accent curve height is the difference between absolute peak height and its projection on the phrase curve.

The inference that the phrase curve is the reference line for the evaluation of prominence is further strengthened by two analytical results. (1) In subset S1, where phrase curve slope is unvaried among the utterances, adding phrase curve onset height to the

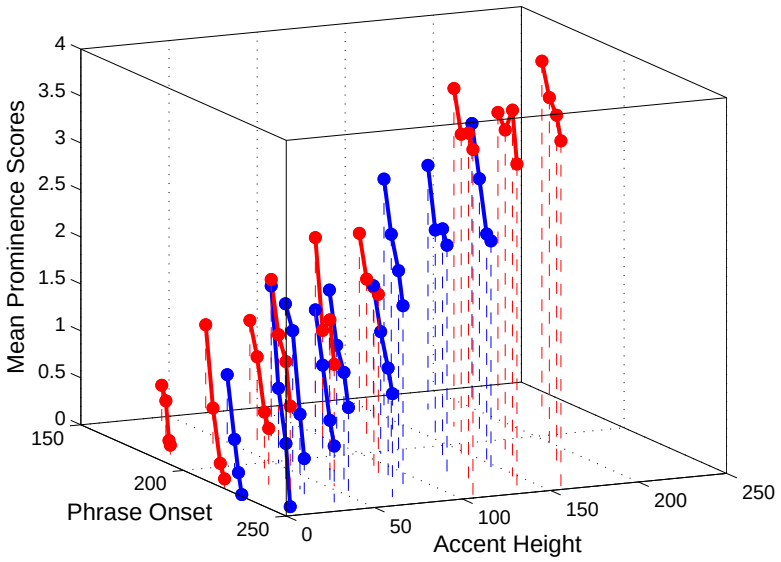

Figure 5.7: Displaying the mean prominence scores with respect to phrase curve onset height and accent curve height. Each line in the graph represents an equipeak that was attained with different accent curve height and phrase curve onset height combinations.

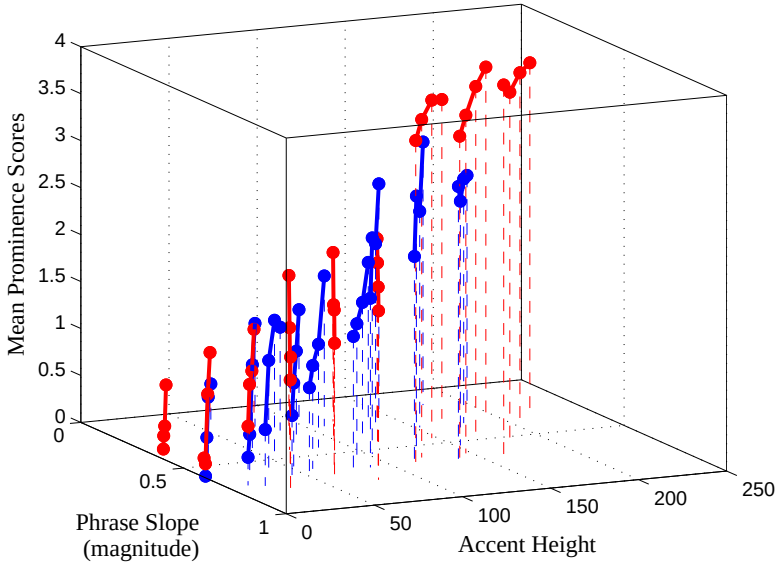

Figure 5.8: Displaying the mean prominence scores with respect to phrase curve slope and accent curve height. Each line in the graph represents an equipeak that was attained with different accent curve height and phrase curve slope combinations.

regression model containing the accent curve parameters increases the explained variance significantly from 87.44% to 93.66%. (2) In subset S2, where phrase curve slope is systematically varied among the stimuli utterances, adding phrase curve slope to the regression model containing the accent curve parameters increases the explained variance significantly from 85.72% to 94.19%. Since in the current experimental design, we are unable to evaluate a regression model that includes both phrase curve onset height and phrase curve slope, an argument can be made that it is not the phrase curve but some linear transformation of it that is the reference line. While possible, this alternative theory is unlikely, given the very low error rates in the two regression models (6.5% and 5.8%).

5.4.4 Contrastive versus normal accents

Last but not the least, these results can be used to answer another question: Do contrastive accents influence perceived prominence differently than normal accents? Our results show that the influence of contrastive accents on perceived prominence can indeed be differentiated from that of normal accents. Listeners perceive contrastive accents to be more prominent than normal accents of comparable height. We came to this conclusion for three reasons. First, while recording the base set our speaker consistently produced contrastive accents with the accent curve peak located closer to the start of the vowel in the stressed syllable than in the case of corresponding normal accents. Second, our regression analysis established that accent curve peak position is significantly negatively correlated with perceived prominence; that is, the closer the peak is to the start of the vowel in the stressed syllable, the higher the perceived prominence. Third, when we paired comparable utterances that differed only in terms of their accent curve shape (i.e., normal versus contrastive) and compared the mean prominence scores of the utterances in each pair, we found that in 16 of the 18 such pairs, the prominence scores related to the contrastive accents were significantly (indicated by a sign-test, p=0.002) higher than the prominence scores related to the normal accents. The utterance pairs were comparable in terms of phrase curve onset height (i.e., $<$ 15 Hz absolute difference), phrase curve slope (i.e., $<$ 2 Hz/sec absolute difference), and accent curve height (i.e., the height of normal accent $\geq$ the height of the corresponding contrastive accent).

5.5 Discussion

Our finding that accent curve height and perceived prominence are strongly positively correlated is well in line with most previous prominence studies. Our definition of accent curve height matches the definition of accent component height in the studies reported in Mixdorff and Widera (2001) and Vainio et al. (2003), both of which were prominence studies performed in the superpositional context. Most other prominence studies were not conducted in the superpositional framework but in either a sequential framework or a quasi-superpositional framework (as in Gussenhoven et al. 1997). As a result, our definition of accent curve height does not find an exact match there. Rather, in such studies, our idea of accent curve height corresponds to the pitch peak height relative to a linear baseline. In spite of the framework (and thus definitional) difference, our finding regarding the strong correlation between accent curve height and perceived prominence was borne out by these studies as well.

Our finding that contrastive accents are perceived as more prominent than normal accents is supported by the experimental results reported in Krahmer and Swerts (2001) but with a caveat. The results of this study on Dutch intonation indicate that contrastive accents are perceived to be more prominent, but only when they are presented within the utterance context. The caveat is that the prominence difference disappears when the accents are presented in isolation to the subjects.

The Krahmer and Swerts (2001) study also substantiates our observation that contrastive accents are marked by early peaks. This study reports that contrastive accents have a "nuclear" shape (i.e., with early peaks) while normal accents have a "prenuclear" shape (i.e., with delayed peaks). This shape difference, however, was only found to be well-pronounced for adjectives, which always occurred in a *non-nuclear position* in the phrase (i.e., not the last pitch accent in a phrase); the shape of the normal and the contrastive accents were essentially identical on the noun, which always occurred in the *nuclear position* in the phrase (i.e., last pitch accent in a phrase). We conjecture that the difference in shape was lost for the nouns because peak alignment of the accent curve ranging over the noun was more strongly influenced by position in the phrase (i.e. nuclear) than by

accent type (i.e., contrastive versus normal). Taken together, the reported difference in shape and prominence of the contrastive accents versus the normal accents in Krahmer and Swerts (2001) corroborates the results of our regression analysis that show that early accent curve peak location corresponds to higher perceived prominence[3].

A difference in shape between contrastive and normal accents is also reported by Pierrehumbert and Hirschberg (1990); using the ToBI scheme, they report that contrastive accents have an early rise L+H* shape while normal accents have a simple high H* shape. Watson et al. (2004) report that though the interpretations of the L+H* and the H* shape overlap, L+H* creates a strong bias towards contrastive information, whereas H* is compatible with both new and contrastive information. Bartels and Kingston (1994) and Ladd (1993) do not find differences in shape between normal and contrastive accents, but they report that contrastive accents have a greater peak height than corresponding normal accents.

Overall, accent curve height and accent curve peak location are two of the most important correlates of perceived prominence. We draw this conclusion based on our analysis results that show that accent curve height and accent curve peak location together can explain over 85% (see Figure 5.4 and Figure 5.6) of the variance in the mean prominence scores. However, phrase curve onset height and phrase curve slope also assist in the perception of prominence; this finding is based on our analysis results that show that their presence in the regression equation helps explain an additional 6-8% of the variance in the mean prominence scores.

The idea that phrase curve onset height and phrase curve slope contribute to the perception of prominence is in line with earlier findings. Gussenhoven et al. (1997) report that both onset height and slope of the declining baseline affect perceived prominence. In their study, three stretches of the declining baseline are treated piecemeal: one spanning the beginning of the utterance to the beginning of the pitch accent, the second spanning the pitch accent, and the third spanning the end of the pitch accent to the end of the utterance. Given this, changing the onset height of the baseline only vertically shifts

[3]The authors do not draw this conclusion.

the stretch from the beginning of the utterance to the beginning of the pitch accent. In our study, on the other hand, the phrase curve cannot be partitioned into individual pieces that are treated independently, and thus change in the phrase curve onset height vertically shifts the entire contour. The slope of the baseline in Gussenhoven et al. (1997) corresponds to the phrase curve slope in our study.

Gussenhoven et al. (1997) show that if the stretch spanning the beginning of the utterance to the beginning of the pitch accent is greater than 400 ms, onset height is negatively correlated with perceived prominence. However, the relationship is reversed if the stretch spanning the beginning of the utterance to the beginning of the pitch accent is less than 400 ms. The authors explain these results by reasoning that when the span between the utterance beginning and the accent beginning is greater than 400 ms, the listener has enough time to determine the reference line and thus use this information to determine the prominence of the F_0 peaks; for a given F_0 peak, a high baseline corresponds to a low pitch accent and hence, lower prominence, while a low baseline corresponds to a high accent and hence, higher prominence. However, when the span is shorter than 400 ms, the listeners are uncertain about the reference line. In such a case — based on 't Hart et al.'s study of Dutch intonation ('t Hart et al. 1990) — the listeners interpret high onset height as more lively and hence more prominent and low onset height as less lively and less prominent.

In our study, phrase curve onset height is found to assist in the perception of prominence. The direction of the relationship cannot be conclusively determined since the zeroth-order correlation between phrase curve onset height and perceived prominence is statistically insignificant ($p > 0.05$); however, if we go by the beta weights, the direction of the relationship may be considered to be positive. This result may be interpreted using Gussenhoven's reasoning since for all stimuli, the span between the utterance beginning and the accent beginning is less than 400 ms.

Gussenhoven et al. (1997) also report that the slope of the baseline line is positively related to perceived prominence. In the current study, the slope of the phrase curve (in conjunction with the accent curve parameters) is found to affect the perception of prominence. Like phrase curve onset height, the direction of the relationship between

phrase curve slope and perceived prominence cannot be conclusively determined since the zeroth-order correlation between phrase curve slope and perceived prominence is statistically insignificant ($p > 0.05$). However, if we go by the beta weights, the direction of the relationship may be considered to be positive. This result is similar to the one reported in Gussenhoven et al. (1997).

Our determination that the phrase curve serves as a reference line that listeners use to evaluate the prominence of F_0 peaks is supported by Fujisaki and Hirose (1984) and 't Hart et al. (1990). Terken (1991) also shows that the observed minima in the F_0 contour are indeed relevant for the perception of F_0 peaks, but the results of the experiment argue against a direct correlation between perceived prominence and the observed excursion of the F_0 peaks from a given baseline. Gussenhoven et al. (1997), on the other hand, determine that the reference line is abstract — a widely-held position in the sequential intonation framework.

The reasoning behind Gussenhoven et al.'s determination is as follows. In their study, the stimuli consisted of synthetic F_0 contours, each of which comprised a baseline and a pitch accent superimposed on it. The accent was a rise-and-fall combination developed from two linear segments that joined together at the top to form a sharp peak. As mentioned earlier, Gussenhoven et al. treated the stretch of the baseline spanning the beginning of the utterance to the beginning of the accent independently of the other parts of the baseline. In two of their experiments, Gussenhoven et al. used sets of F_0 contours in which the former span was kept completely horizontal and systematically varied in terms of onset height, while the rest of the baseline was held constant. In both these experiments, listeners were asked to evaluate the perceived prominence of each of the stimuli F_0 contours. Analysis of the experimental results indicated that the listeners perceived a declining baseline. From this, the authors reasoned that since "the actual F_0 in the stimuli ... did not decline", this implies that the reference line is abstract rather than being induced by the observable F_0 declination.

We, however, do not agree with Gussenhoven et al.'s statement "the actual F_0 in the stimuli ... did not decline" because their description of the experimental stimuli shows that though the stretch spanning the beginning of the utterance to the beginning of the

accent was kept horizontal, the remainder of the baseline declined — albeit to the same final F_0 value. We further conjecture that since such an F_0 contour with a completely horizontal stretch leading up to the beginning of the accent is quite unlikely to occur in natural speech, listeners used the remainder of the baseline — which was distinctly longer than the horizontal stretch — to extrapolate a declining baseline. In addition, if such an F_0 contour was analyzed in the superpositional framework, it would be determined to contain a declining phrase curve and an accent curve with a long rise.

5.6 Conclusions

From our investigation of the combined and individual influence of four intonational parameters obtained in the superpositional framework on perceived prominence, we can draw the following six conclusions:

C1: Accent curve height, accent curve peak location, and phrase curve onset height (or phrase curve slope) together can predict perceived prominence with a high degree of accuracy.

C2: Accent curve height is the strongest intonational correlate of perceived prominence. The correlation between accent curve height and perceived prominence is strongly positive.

C3: The influence of the accent curve parameters on perceived prominence is much greater than the influence of the phrase curve parameters. Although the latter increase the predictive power of the accent curve parameters, they predict perceived prominence non-significantly in isolation.

C4: Perceived prominence is not simply a correlate of absolute peak value; rather, listeners are able to perceive a difference in prominence corresponding to the different accent curve heights that are used to attain the same absolute peak value.

C5: The reference line that listeners use to make prominence judgements is observable, and it corresponds to the phrase curve.

C6: Contrastive accents are perceived as more prominent than normal accents of equal or even somewhat greater height.

The significance of these conclusions is substantial. Conclusion C2 validates the basic assumption of the superpositional approach that prominence is strongly tied to the accent curve height. However, conclusions C1 and C3 show that this assumption, though necessary, is not sufficient. Since accent curve height is not the only determiner of prominence, superpositional models need to include additional constraints that relate prominence to the other relevant intonational parameters. Conclusion C4 indicates that accent curve height — and not absolute peak height — is the true correlate of prominence; any observed relationship between absolute peak height and prominence is due to the strong correlation between absolute peak height and accent curve height. Conclusion C5 contradicts the position that the reference line that listeners use to evaluate the prominence of F_0 peaks is abstract. Conclusion C6 has important implications for speech synthesis: the common strategy of using normal accents to intonationally highlight all crucial information in an utterance is not sufficient; given the difference in prominence between normal and contrastive accents, generation of contrastive accents is necessary to achieve the goal of producing expressive speech.

Chapter 6

Conclusions and Future Directions

The central goal of our research is the development of an algorithm for decomposing any arbitrary natural F_0 contour into its component curves using only the general assumptions of the superpositional approach, in order to enable the testing of assumptions made by intonation models based on the superpositional approach and to enable the implementation of the recombinant synthesis method. We divided this central goal into three subgoals:

1. Development of the decomposition algorithm for extracting the component curves of natural F_0 contours using only the general assumptions of the superpositional approach.

2. Implementation of a proof-of-concept of the recombinant synthesis method in order to demonstrate the essential role of the decomposition algorithm in this synthesis method.

3. Testing of assumptions regarding the perception of prominence and the shapes of F_0 contour components in order to show the decomposition algorithm's utility in testing assumptions about intonation in the superpositional framework.

Chapters 3, 4, and 5 describe the ways in which we met sub-goals 1, 2, and 3, respectively. Here we present a summary of each of these chapters, the overall significance of this body of research, its limitations and constraints, and the directions that this research can take in the future.

6.1 Chapter summaries

6.1.1 Chapter 3

In chapter 3, we presented a detailed description of our decomposition algorithm PRISM that can be used to decompose any given F_0 contour into component curves, in accordance with the superpositional approach. The general assumptions of the superpositional approach that underlie PRISM are as follows: (1) an observed F_0 contour is obtained by the "generalized addition" of component curves of different classes, and (2) each of the component curves is tied to a distinct phonological entity and follows a distinct time course.

The three classes of component curves that are posited in the PRISM framework are phrase curves, accent curves, and segmental perturbation curves. Each of these classes of component curves is considered to be generated from a corresponding class-specific template by parameterized time-warp functions. Adhering to the idea of using only the general assumptions of the superpositional approach, no equations governing the shapes of these templates are specified; instead some general shape constraints, based on observations in previous research, are outlined. In fact, exactly which particular function is chosen to implement a class-specific template is not very important. What is important is that the function used to implement a class-specific template adheres to the general shape constraints of the template. For example, instead of a Gaussian curve, a cosine curve can be used just as well to implement the declarative accent curve template. Regardless of the function used, the time-warping procedure renders templates with similar shape constraints equivalent.

The PRISM decomposition algorithm has two phases: (1) a wavelet decomposition phase that partially decomposes a given F_0 contour into a phrase curve and a *combined* accent curve and segmental perturbation curve; and (2) a template decomposition phase that completely decomposes the F_0 contour into a phrase curve, one or more accent curves, and one or more segmental perturbation curves. The reason for the two-phase PRISM implementation is that in the PRISM framework, a smooth unconstrained phrase curve is assumed and hence, a template of the phrase curve cannot be directly specified. Thus,

first an initial estimate of the phrase curve, Phr_{init_est}, using the wavelet decomposition phase is obtained, and then Phr_{init_est} is used as the phrase curve template in the template decomposition phase.

Using several sets of synthetic F_0 contours — corresponding to declaratives utterances, interrogatives utterances, and utterances containing continuation rises — whose component curves were known, we objectively evaluated the accuracy of PRISM in estimating component curves. In these evaluations, the RMSE differences between the known and estimated component curves were under the frequency difference required for just noticeable difference in pitch for connected speech. From this, we conclude that PRISM can decompose F_0 contours of declaratives utterances, interrogatives utterances, and utterances containing continuation rises with sufficiently high accuracy.

In this chapter, we also showed an application of PRISM to natural F_0 contours through the particular problem of analyzing the difference in F_0 contour shape between minimal pairs of utterances that differed only in terms of one consonant. The performance of PRISM on natural F_0 contours cannot be objectively evaluated because the ground truth is unknown; however, the strong within-pair similarity of estimated phrase curves and accent curves corresponding to each minimal pair does suggest that the component curves estimated by PRISM are close approximations of the "true" component curves.

6.1.2 Chapter 4

Recombinant synthesis is a speech synthesis method that involves searching one corpus for acoustic units, while searching another corpus for F_0 component curves that have been extracted from natural speech. The selected F_0 component curves are additively recombined according to the general superpositional approach to construct the quasi-natural target F_0 contour, which is imposed on the selected acoustic unit sequences using standard signal modification methods.

The generation of quasi-natural target F_0 contours using F_0 component curves that have been extracted from natural speech is a key concept of recombinant synthesis. However, to realize the benefits of this concept in practice, the existence (and usage) of a robust

decomposition algorithm is essential. To demonstrate that PRISM can fulfil this essential role in the implementation of recombinant synthesis, we developed a proof-of-concept (POC) implementation of the recombinant synthesis method. This POC implementation is presented in Chapter 4 of this dissertation. This implementation was restricted to intonation generation in the recombinant synthesis method. It did not include creating an acoustic unit corpus or searching it for appropriate acoustic units; rather, the target F_0 contours were imposed on pre-recorded natural speech so that the strengths and weaknesses of the intonation generation method could be observed unconfounded by the effects of acoustic unit search and concatenation.

The POC implementation of recombinant synthesis was divided into two stages of increasing complexity: (1) intonation generation of fully sonorant declarative utterances; and (2) intonation generation of phonemically unrestricted declarative utterances containing continuation rises. In each stage of implementation, we conducted perceptual tests to compare the recombinant intonation generation method to other intonation generation methods. The results of the perceptual tests show that the intonation generated by recombinant synthesis is more natural-sounding and more expressive than intonation generated by concatenation of raw F_0 curves (as in unit selection synthesis) and intonation generated by statistical, linguistic, and heuristic rules (as in traditional concatenative synthesis). Thus, the recombinant intonation generation method can be considered a promising alternative to the latter two methods. The results of the comparison between the recombinant intonation generation method and the concatenation of raw F_0 curves indicate that the continuity of the target F_0 contours provided by the recombinant intonation generation method clearly outweighed any pitch modification distortion.

6.1.3 Chapter 5

In Chapter 5, we described our investigation of the intonational correlates of perceived prominence from a superpositional perspective. This is a particular application of a more general class of applications in which PRISM is used to analyze F_0 contours so that theories and assumptions regarding intonation can be critically tested. Specifically, we explore three questions in this chapter. (1) What are the combined and individual influences of

accent curve height, accent curve peak location, phrase curve onset height, and phrase curve slope on perceived prominence? (2) Is there a line of reference that listeners use to determine the prominence of peaks in the F_0 contour? If so, is the reference line abstract or observable? (3) Does contrastive accent influence perceived prominence differently from normal accent?

To answer these questions, we conducted a perceptual test, in which 10 listeners were asked to make prominence judgements on a controlled stimulus set of intonationally varied quasi-natural F_0 contours. Building this stimulus set involved (1) development of a base set of natural utterances that were systematically varied in emphasis level and emphasis type (i.e., normal versus contrastive emphasis); (2) PRISM analysis of the base set to extract natural phrase curves and accent curves; and (3) generation of quasi-natural F_0 contours by first, systematically varying the parameters of interest in the natural component curves and then, recombining them according to the general principles of the superpositional approach.

The results of the perceptual test indicate the following:

C1: Accent curve height, accent curve peak location, and phrase curve onset height (or phrase curve slope) together can predict perceived prominence with a high degree of accuracy.

C2: Accent curve height is the strongest intonational correlate of perceived prominence. The correlation between these two variables is strongly positive.

C3: The influence of the accent curve parameters on perceived prominence is much greater than the influence of the phrase curve parameters. Although the latter increase the predictive power of the accent curve parameters, they predict perceived prominence non-significantly in isolation.

C4: Perceived prominence is not simply a correlate of absolute peak value; rather, listeners are able to perceive a difference in prominence corresponding to the different accent curve heights that are used to attain the same absolute peak value.

C5: The reference line that listeners use to make prominence judgements is observable, and it corresponds to the phrase curve.

C6: Contrastive accents are perceived as more prominent than normal accents of equal

or even somewhat greater height.

Most of these conclusions are in agreement with prior work; the only one that disagrees with some previous studies is C5.

6.2 Overall significance

6.2.1 Addresses concerns regarding the superpositional approach

The superpositional approach to intonation modeling has been the subject of much controversy. The most comprehensive critique on this subject has been presented in Ladd (1996). One of the main issues brought up by Ladd is that estimating component curves posited in the superpositional approach from observed F_0 contours is not straightforward and is often left unspecified, as in Thorson (1983) and Gårding (1983).

We believe that Ladd's critique is valid. One cannot claim to have a complete model unless one can specify exactly what the components of the model are and how they can be estimated; it is not enough to say that the components exist. Thus, an algorithm that can estimate the component curves of observed F_0 contours is needed to fill in the details necessary to fully define a model of intonation in the superpositional approach.

We addressed this need in our research. We clearly outlined the algorithm (i.e., PRISM) needed to estimate phrase curves, accent curves, and segmental perturbation curves from given F_0 contours corresponding to declarative utterances, interrogative utterances, and utterances containing continuation rises. By objective evaluations using several sets of synthetic F_0 contours whose component curves were known, we demonstrated that the outlined PRISM algorithm estimates component curves from given F_0 contours of the aforementioned types with sufficiently high accuracy. Since such objective evaluations are not possible for natural F_0 contours — because the ground truth is unknown — we evaluated how well PRISM can estimate component curves from natural F_0 curves indirectly, based on the results of the minimal pair analysis described in Section 3.4.2 and the perceptual tests described in Sections 4.2.2 and 4.2.5.

The high within-pair similarity of the phrase curves, accent curves, and clean F_0 contours of each minimal pair attests to the ability of PRISM in estimating component curves

from natural F_0 contours. The effectiveness of PRISM in estimating "true" component curves is also indirectly confirmed by the results of the perceptual tests that showed that the intonation generated by recombinant synthesis is more natural-sounding and more expressive than intonation generated by the other synthesis methods. In our implementation of the recombinant synthesis method, the component curves that were used to construct the target F_0 contours were extracted using PRISM; if these component curves extracted by PRISM were not at least close approximations of the "true" curves, then the target F_0 contours constructed from them would not sound natural or expressive. Since our implementation of recombinant synthesis involved extracting component curves from several sets of natural F_0 contours that ranged in complexity from fairly simple (such as fully sonorant declarative utterances) to somewhat complex (such as declarative utterances containing a few obstruents) to very complex (such as very long, phonemically unrestricted utterances containing declaratives and continuation rises), it indicates that PRISM can robustly estimate component curves from arbitrary natural F_0 contours containing declaratives and continuation rises — and perhaps even interrogatives.

Another criticism leveled by Ladd in the context of the Fujisaki model — the best known superpositional model — is the model's apparent inability to model certain contour shapes observed in English. In their analysis of this criticism, van Santen and Möbius (1999) attribute this inability to the very strong constraints on the shape of commands and contours, especially on the shape of the phrase curve. Following their suggestion to remain "open to the possibility that phrase curves exhibit considerable and meaningful variability" (van Santen and Möbius 1999), we specified few constraints on the shape of the phrase curve besides smoothness. The phrase curves estimated by PRISM can thus be expected to capture certain important characteristics of F_0 contours, such as phrase-final lowering[1], which could not be captured by the negative impulse used as a shape constraint for the phrase curve in the Fujisaki model, as demonstrated by Liberman and Pierrehumbert (1984). In this way, we attempted to address this important concern raised by Ladd regarding the superpositional approach to intonation modeling.

[1]Phrase final lowering is a sentence-level phenomenon whereby the final peak of an utterance undergoes a more drastic lowering in F_0 than that of the previous F_0 peaks.

6.2.2 Demonstrates the importance of modeling microprosody

Besides attempting to address these previously raised concerns about the superpositional approach, we also examined the importance of explicitly modeling microprosody in intonation analysis. Microprosody includes perturbations in the F_0 contour due to vowel intrinsic pitch effects, post-obstruent vowels, nasality effects, and the presence of voiceless regions. Most superpositionally-based intonation models and consequently, the intonation analysis algorithms based on them, ignore microprosody. However, research has shown that microprosodic modeling is important for several reasons. From the perspective of estimating component curves, it is important because microprosodic variations may distort the surface realization of F_0 contours and thus obscure the "true" shapes of the underlying phrase curves and accent curves. Based on this hypothesis, we explicitly modeled a particular type of microprosodic variation, namely, segmental perturbations due to post-obstruent vowels, in the PRISM algorithm. The minimal pair analysis using PRISM showed that the within-pair similarity of each minimal pair that differed in terms of a single consonant was significantly higher in terms of "clean" F_0 contours (i.e., without segmental perturbations) than in terms of raw (or even smoothed) F_0 contours. These results are strong empirical evidence in support of the above-mentioned hypothesis, and thus establish the importance of modeling segmental perturbations in intonation analysis.

6.2.3 Pushes recombinant synthesis ahead

Another important topic that this research has addressed is recombinant synthesis. The recombinant synthesis concept and its relative strengths and weaknesses in relation to other synthesis methods was presented earlier by van Santen et al. (2004:i and 2005); however, lacking a robust decomposition algorithm that could be used to populate the corpus of natural F_0 component curves, this synthesis method had not been implemented. As a result, actual evaluation of how this synthesis method compares to other synthesis methods was not possible. Development of the PRISM decomposition algorithm in our research made it possible for us to implement the first proof-of-concept of the recombinant synthesis method, and thus it enabled us to evaluate the legitimacy of the expectation

that intonation generated by recombinant synthesis will be more natural-sounding and more expressive than intonation generated by concatenation of raw F_0 curves (as in unit selection synthesis) and intonation generated by rule (as in traditional concatenative synthesis). Results of the comparative perceptual tests indicate that this expectation is well warranted.

6.2.4 Enables testing of assumptions about intonation

Prominence marking is an important consideration in speech synthesis. In most superpositional intonation models, prominence is assumed to be only a function of accent curve height. However, previous research — mostly in the sequential framework — has posited that perception of prominence is also related to other intonational parameters. Thus, in our research, we investigated how perceived prominence is influenced by four parameters of intonation obtained in the superpositional framework, namely, accent curve height, accent curve shape, phrase curve onset height, and phrase curve slope.

The results of our investigation show that perceived prominence is affected not only by accent curve height — though its effect is the greatest — but also by the other intonational parameters under consideration. This finding implies that the assumption that prominence is a function of accent curve height, though necessary, is not sufficient. Superpositional models therefore must include additional constraints that relate prominence to the other relevant intonational parameters. The results of the investigation also indicate that the reference line that listeners use to evaluate the prominence of F_0 peaks is observable, and it corresponds to the phrase curve. This finding contradicts the widely-held position in the sequential intonation framework that the reference line is abstract. We conjecture that the difference in their position about the reference line versus ours arises because within the sequential framework, there is no way to relate the reference line to a particular component of intonation, given the assumptions of this framework; however, in the superpositional framework, we can easily relate the reference line to a component of intonation, namely, the phrase curve. This finding that the phrase curve is important to the perception of prominence also helps refute the idea that F_0 minima have no importance in perception.

Finally, the results of our analysis also show that listeners can perceive a difference

between normal and contrastive accents. This finding implies that the common strategy in speech synthesis of using normal accents to intonationally highlight all crucial information in an utterance is not sufficient. Given the difference in prominence between normal and contrastive accents, generation of contrastive accents is necessary to achieve the goal of producing expressive speech. Considering these findings and their implications, the significance of this investigation of the intonational correlates of perceived prominence can be regarded as fairly substantial.

The investigation of the intonational correlates of perceived prominence and the minimal pair analysis used to examine the importance of modeling microprosody both required the use of PRISM to extract the F_0 component curves and parameters thereof. This demonstrates PRISM's significance as an analysis tool that is useful for critically testing theories and assumptions regarding intonation in the superpositional framework.

6.2.5 Long-term significance

The long-term significance of the research presented in this dissertation is this. The PRISM decomposition algorithm can be used to estimate component curves of natural F_0 contours using only the general assumptions of the superpositional approach, and thus it will enable the accurate determination of (i) the effects of linguistic control factors on the component curves, (ii) their time courses, and (iii) the interactions between curve classes. Barring the discovery of fatal flaws in the superpositional concept along the way, these determinations may ultimately lead to the development of a — potentially complex — superpositional intonation model that contains a richer collection of curve classes than current models, specifies a clear and phonologically well-justified mapping between each class and the F_0 contour, and generalizes well to fit across a wide range of speech materials. Such a model can help us achieve the elusive goal of generating synthetic speech that has as meaningful and as expressive intonation as human speech.

6.3 Constraints and limitations

In developing the PRISM algorithm, we have assumed that the natural F_0 contours it encounters are labeled and segmented. This is a valid assumption in the area of speech synthesis where a fully labeled and segmented speech corpus is easily available. However, labeling and segmentation are not fully automatic procedures; either the transcription output (such as phone labels) from labeling and segmentation systems have to be manually checked for accuracy by labeling experts, or the labels (such as foot and phrase labels) have to be completely manually assigned because there are no reliable systems to automatically assign such labels. This makes PRISM analysis of intonation a fairly time-consuming process and prevents users from being able to PRISM analyze intonationally-interesting utterances on-the-fly. Much work therefore remains to be done to automate the tedious manual labeling process — especially the foot and phrase-level labeling — and thus speed up the PRISM analysis procedure.

Another limitation of the current implementation of PRISM is that it only models three common types of accent curves, namely, those corresponding to declaratives, interrogatives, and continuation rises, and only one type of microprosodic variation, namely, segmental perturbations due to post-obstruent vowels. Given the established presence of other types of accent curves and microprosodic variations, we need to extend PRISM to be able to estimate these additional component curves as well. We believe that this can easily be accomplished in the future owing to the modular implementation of PRISM; the main challenge, however, will be determining the general shape constraints of the templates corresponding to these additional component curves.

Aside from the aforementioned limitations pertaining to the PRISM algorithm, our research is limited in two other ways. One limitation is that the proof-of-concept implementation of recombinant synthesis presented in this dissertation is limited to intonation generation only. To implement a full-fledged recombinant synthesis system, other important issues such as accent prediction, duration generation, and acoustic unit selection need to be considered. In limiting ourselves to intonation generation in recombinant synthesis, we make the subtle assumption that intonation generation and acoustic unit selection are

mutually independent; in practice, however, flaws in one component can limit the success of the other.

Another limitation of our research is that in the experimental design set up to investigate the effect of different intonational correlates on perceived prominence, phrase curve onset height, and phrase curve slope were strongly correlated. As a result, we were unable to evaluate a regression model that included both phrase curve onset height and phrase curve slope. Consequently, an argument can be made that it is not the phrase curve but some linear transformation of it that is the reference line that listeners use to evaluate the prominence of F_0 peaks. While this alternative theory is unlikely given the low error rates of the regression models (6.5% and 5.8%), it cannot be completely discarded in the current experimental design.

6.4 Future directions

The many analyses presented in this dissertation have empirically established the reliability of PRISM as an intonation analysis tool that can be used to decompose arbitrary natural F_0 contours into component curves posited in the superpositional framework. Given this, in the future, we can use PRISM to investigate some larger questions in speech synthesis including the following:

- What is the prosodic difference between newscaster-like speech — that is often used in speech synthesizers — and spontaneous speech obtained from real people engaged in real tasks, speaking for their own communicative purposes? Answers to this question will improve the naturalness of synthesized speech.
- What are the intonational correlates of different emotions? Answers to this question will help the synthesis of emotional speech.
- What are the prosodic differences among people depending on their ethnic background, socio-economic situation, emotional center, energy level, age, etc.? Answers to this question will help generate voices that are person/character-specific rather than generic male or female.

- What are the prosodic characteristics of conversational speech events such as hesitations and disfluencies (e.g., utterance restarts with words such as "well" and "yeah", fillers such as "um", "uh", and "ah", and hedges such as "like") and backchannel responses (e.g., "yeah", "un-huh", and "mmm")? Answers to this question will help speech synthesizers generate more conversationally appropriate dialogue.
- What are the prosodic aspects of the synchrony between speech and gestures or task-related performances? Answers to this question will help improve the coordination between the speech and gestures of computer-generated animated characters.

Finding answers to these questions will require careful development of prosodic corpora, extraction of component curves from natural F_0 contours using PRISM, and an in-depth analysis of the parameters of the component curves. It is quite likely that addressing these questions will require us to extend PRISM so that it can estimate other types of component curves besides the ones currently handled. We may also have to add to, remove, or modify the current constraints on the component curves. Given these considerations, we believe addressing these larger questions of speech synthesis will certainly be challenging, though not nearly as challenging as without PRISM as a tool in our tool-kit.

Besides speech synthesis, the PRISM decomposition algorithm may be useful for speech recognition. So far, most speech recognition systems make no use of the communicative information contained in the intonation of speech, instead relying solely on the textual information. This is due to the lack of an easy yet robust way to analyze intonation of arbitrary utterances - a problem that may potentially be solved by appropriately incorporating the PRISM decomposition algorithm into speech recognition systems. Last but not least, the PRISM decomposition algorithm may be beneficial for developing diagnostic tools for detecting neurodevelopmental and neurodegenerative disorders whose symptoms include compromised intonation.

Bibliography

Aguero, P.D., Wimmer, K., and Bonafonte, A. (2006). Automatic analysis and synthesis of Fujisakis intonation model for TTS. *Proceedings of the International Conference on Speech Prosody, March 23-26, 2004, Nara, Japan.*

Aubergé, V. (1993). Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis. *Proceedings of the ESCA Workshop on Prosody, September 27-29, 1993, Lund, Sweden*, 62–65.

Bailly, G., and Holm, B. (2005). SFC: A trainable prosodic model. *Speech Communications, 46*, 364–384.

Bailly, G., and Gorisch, I. (2006). Generating German intonation with a trainable prosodic model. *Proceedings of Interspeech-ICSLP, September 17-22, 2006, Pittsburgh, PA.*

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions.* New York, NY: John Wiley & Sons.

Bartels, C., and Kingston, J. (1994). Salient pitch cues in the perception of contrastive focus. *The Journal of the Acoustical Society of America, 95(5)*, 2973.

Beckman, M.E. (1986). *Stress and Non-stress Accent (Netherlands Phonetic Archives No. 7).* Dordrecht, The Netherlands: Foris.

Black A., and Hunt, A. (1996). Generating F_0 contours from the ToBI labels using linear regression. *Proceedings of the 4th International Conference on Spoken Language Processing, October 3-6, 1996, Philadelphia, PA* (Vol. 3), 1385–1388.

Black, A. and Taylor, P. (1997). Festival speech synthesis system: System documentation (1.1.1). *Human Communication Research Centre Technical Report, HCRC/TR-83*, Edinburgh, UK.

Boersma, P. and Weenink, D. (2008). *Praat: Doing phonetics by computer.* Retrieved on April 26, 2008 from: http://www.fon.hum.uva.nl/praat/.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees.* Monterey, CA: Wadsworths and Brooks.

Campbell, N., and Venditti, J. (1995). J-ToBI: An intonation labelling system for Japanese. *Proceedings of the Autumn 1995 Meeting of the Acoustical Society of Japan, Utsunomiya, Japan, Vol. 1*, 317–138.

Cohen, A., and 't Hart, J. (1967). On the anatomy of intonation. *Lingua, 19*, 177–192.

Cooper, W.E., Eady, S.J., and Mueller, P.R. (1985). Acoustical aspects of contrastive stress in question/answer contexts. *The Journal of the Acoustical Society of America, 77(6)*, 2142–2156.

Crystal, D. (1987). *Cambridge Encyclopedia of Language*. Cambridge, UK: Cambridge University Press.

Donoho, D.L., and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*, 425–455.

Dusterhoff, K., and Black, A. (1997). Generating F_0 contours for speech synthesis using the Tilt intonation theory. *Intonation: Theory, Models and Applications, (Proceedings of the ESCA Workshop, September 18-20, 1997, Athens, Greece)* (pp. 107–110). Ed. A. Botinis, G. Kouroupetroglou, and G. Carayiannis. ESCA and the University of Athens.

Eady, S.J. and Cooper, W.E. (1986). Speech intonation and focus location in matched statements and questions. *The Journal of the Acoustical Society of America, 80(2)*, 402–415.

Fry, D.B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*, 765–768.

Fry, D.B. (1958). Experiments in the perception of stress. *Language and Speech, 1*, 126–152.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. *The Production of Speech*. (pp. 39–55). Ed. P.F. MacNeilage. New York, NY: Springer.

Fujisaki, H. (2002). Modeling in the study of tonal features of speech with application to multilingual speech synthesis. *Proceedings of the Joint Conference of SNLP and Oriental COCOSDA Workshop, May 9-11, 2002, Hua Hin, Thailand*.

Fujisaki, H., and Hirose, K. (1982). Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. *Preprints of the Working Group on Intonation, (Proceedings of the 13th International Congress of Linguists, Tokyo, Japan)*, 57–70.

Fujisaki, H., and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *The Journal of the Acoustical Society of Japan (E), 5*, 233-242.

Gårding, E. (1983). A generative model of intonation. *Prosody – Models and Measurements*. (pp. 11–25). Ed. A. Cutler and D.R. Ladd. Heidelberg, Germany: Springer Verlag.

Green, P.J., and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London, UK: CRC Press.

Goldsmith, J.A. (1990). *Autosegmental and Metrical Phonology*. Oxford, UK: Blackwell Publishers.

Gussenhoven, C., Repp, B.H., Rietveld, A., Rump, H.H., and Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *The Journal of Acoustical Society of America, 102(5)*, 3009–3022.

Hall, P., Penev, S., Kerkyacharian, G., and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statistics and Computing, 7*, 115–124.

Härdle, W. (1992). Monotonic and unimodal smoothing. *Applied Nonparametric Regression*. (Chapter 8.1). Cambridge, UK: Cambridge University Press.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). Wavelets, approximation and statistical applications. *Lecture Notes in Statistics, 129*, Heidelberg, Germany: Springer Verlag.

Hermes, D.J., and van Gestel, J.C. (1991). The frequency scaling of speech intonation. *The Journal of the Acoustical Society of America, 90*, 97–102.

Holm, B., Bailly, G. (2000). Generating prosody by superposing multi-parametric overlapping contours. *Proceedings of the International Conference on Speech and Language Processing, October 16-20, 2000, Beijing, China*, 203–206.

Jans, L. (2000). *Use of assistive technology: Findings from national surveys*. Retrieved on December 13th, 2006 from Assistive Technology Data Collection Project, an InfoUse Web site: http://www.infouse.com/atdata/csun.pdf.

Jilka, M., Mohler, G., and Dogil, G. (1999). Rules for the generation of ToBI-based American English intonation. *Speech Communication, 28*, 83–108.

Jones, D. (1960). *An Outline of English Phonetics* (9th edition). Cambridge, UK: Cambridge University Press.

Jones, D. (2002). *Selected Works* (Vol. 4). Ed. B. Collins and I. Mees. London, UK: Routledge.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. *Proceedings of MAVEBA, Sept.13-15, 2001, Firenze, Italy.*

Keller, E., and Zellner Keller, B. (2000). New uses for speech synthesis. *The Phonetician, 81*, 35–40.

Kiparsky, P. (1975). Stress, syntax, and meter. *Language, 51(3)*, 576–616.

Klabbers, E., van Santen, J., and Wouters, J. (2002). Prosodic factors for predicting local pitch shape. *Proceedings of the 2002 IEEE Workshop on Speech Synthesis, September 11-13, 2002, Santa Monica, CA.*

Klabbers, E. and van Santen, J. (2003). Control and prediction of the impact of pitch modification on synthetic speech quality. *Proceedings of the 8th European Conference on Speech Communication and Technology, September 1-4, 2003, Geneva, Switzerland*, 317–320.

Klabbers, E. and van Santen, J. (2004). Clustering of foot-based pitch contours in expressive speech. *Proceedings of the 5th ISCA Speech Synthesis Workshop, June 14-16, 2004, Pittsburgh, PA.*

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118(2)*, 1038–1054.

Kochanski, G.P., and Shih, C. (2000). Stem-ML: Language independent prosody description. *Proceedings of the International Conference on Spoken Language Processing, October 16-20, 2000, Beijing, China*, 239–242.

Kochanski, G.P., and Shih, C. (2001). Automated modeling of Chinese intonation in continuous speech. *Proceedings of the 7th European Conference on Speech Communication and Technology, September 3-7, 2001, Aalborg, Denmark*, 911–914.

Kochanski, G.P., and Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication, 39(3-4)*, 311–352.

Kohler, K.J. (1990). Macro and micro F0 in the synthesis of intonation. *Papers in Laboratory Phonology I.* (pp. 115–138). Ed. J. Kingston and M.E. Beckman. Cambridge, UK: Cambridge University Press.

Kohler, K.J. (1991). Studies in German intonation. *Arbeitsberichte 25 des Instituts fr Phonetik und digitale Sprachverarbeitung* (pp. 295–360). Universität Kiel.

Kohler, K.J. (1995). The Kiel intonation model (KIM), its implementation in TTS synthesis, and its application to the study of spontaneous speech. Retrieved on July 15th, 2006 from the University of Kiel Web site: http://www.ipds.uni-kiel.de/kjk/forschung/kim.en.html.

Kohler, K.J. (1997). Parametric control of prosodic variables by symbolic input in TTS synthesis. *Progress in speech synthesis* (pp. 459–475). Ed. J. van Santen, R. Sproat, J. Olive, and J. Hirschberg. New York, NY: Springer.

Krahmer, E., and Swerts, M. (2001). On the alleged existence of contrastive accents. In *Speech Communication, 34*, 391–405.

Kutik, E.J., Cooper, W.E., and Boyce, S. (1983). Declination of fundamental frequency in speakers' production of parenthetical and main clauses. *The Journal of the Acoustical Society of America, 73(5)*, 1731–1738.

Ladd, D.R. (1993). Phonological features of intonational meaning. *Language, 59*, 721–759.

Ladd, D.R. (1996). *Intonational Phonology.* Cambridge, UK: Cambridge University Press.

Ladefoged, P. (1993). *A Course in Phonetics*, (3rd edition), Orlando, FL: Harcourt Brace College Publishers.

Lagarias, J.C., Reeds, J.A., Wright, M.H., and Wright, P.E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal on Optimization, 9(1)*, 112–147.

Laver, J. (1980). *The Phonetic Description of Voice Quality*, Cambridge, UK: Cambridge University Press.

Laver, J. (1994). *Principles of Phonetics*, Cambridge, UK: Cambridge University Press.

Lee, T., Kochanski, G., Shih, C., and Li, Y. (2002). Modeling tones in continuous Cantonese speech. *Proceedings of the International Conference on Spoken Language Processing, September 16-20, Denver, CO*, 2401–2404.

Liberman, M., and Pierrehumbert, J. (1984). Intonational invariance under changes of pitch range and length. *Language Sound Structure* (pp. 157–233). Cambridge, Massachusetts: MIT Press.

Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions Pattern Analysis and Machine Intelligence, 11*, 674–693.

MathWorks, Inc. (2002). FMINSEARCH. MATLAB function. From the MATLAB command 'type fminsearch'.

Mayo, C., Aylett, M., and Ladd, D.R. (1997). Prosodic transcription of Glasgow English: An evaluation study of Glatobi. *Intonation: Theory, Models and Applications, (Proceedings of the ESCA Workshop, September 18-20, 1997, Athens, Greece)* (pp. 231–234). Ed. A. Botinis, G. Kouroupetroglou, and G. Carayiannis. ESCA and the University of Athens.

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of fujisaki model parameters. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, June 9, 2000, Istanbul, Turkey*, 1281–1284.

Mixdorff, H., (2002), "Speech technology, ToBI, and making sense of prosody', In *Proceedings of the International Conference on Speech Prosody, April 11-13, 2002, Aix-en-Provence, France*, 31–37.

Mixdorff, H. (2004). Quantitative tone and intonation modeling across languages. *Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, March 28-31, 2004, Beijing, China*, 137–142.

Mixdorff, H., and Fujisaki, H. (2000). Symbolic versus quantitative descriptions of F0 contours in German: Quantitative modeling can provide both. *Proceedings of Prosody 2000: Speech Recognition and Synthesis, October 2-5, 2000, Krakau, Poland*.

Mixdorff, H., and Widera, C. (2001). Perceived prominence in terms of a linguistically motivated quantitative intonation model. *Proceedings of the European Conference on Speech Communication and Technology, September 3-7, 2001, Aalborg, Denmark*, 403–406.

Möbius, B. (1995). Components of a quantitative model of German intonation. *Proceedings of ICPhS '95, August 13-19, 1995, Stockholm Sweden*.

Monaghan, A. (2002). State-of-the-art summary of European synthetic prosody R & D. *Improvements in Speech Synthesis. Cost 258: The Naturalness of Synthetic Speech* (pp. 93–103). Ed. E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale. Chichester, UK: John Wiley and Sons Ltd.

Murray, I., and Arnott, J. (1996). Synthesizing emotions in speech: Is it time to get excited? *Proceedings of the 4th International Conference of Spoken Language Processing, October 3-6, 1996, Philadelphia, PA*, 1816–1819.

Narusawa, S., Minematsu, N., Hirose, K., and Fujisaki, H. (2002). A method for automatic extraction of model parameters from fundamental frequency contours of speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, May 13-17, 2002, Orlando, Florida*, 509–512.

Orfanidis, S.J. (1996). *Introduction to Signal Processing*. Englewood Cliffs, NJ: Prentis-Hall.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *The Journal of the Acoustical Society of America, 66(2)*, 363-369.

Pierrehumbert, J. (1980). *The Phonetics and Phonology of English Intonation*, PhD dissertation, MIT.

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America, 70*, 985–995.

Pierrehumbert, J., and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication* (pp. 271–311). Ed. P.R. Cohen, J. Morgan, and M.E. Pollack.

Raidt, S., Bailly, G., Holm, B., and Mixdorff, H. (2004). Automatic generation of prosody: comparing two superpositional systems. *Proceedings of the International Conference on Speech Prosody, March 23-26, 2004, Nara, Japan.*

Reyelt, M., Grice, M., Benzmuller, R., Mayer, J., and Batliner, A. (1996). Prosodische Ettikettierung des Deutschen mit ToBI. *Natural Language and Speech Technology* (pp. 144–155). Ed. D. Gibbon. Berlin, Germany: Mouton de Gruyter.

Rietveld, A.C.M., and Gussenhoven, C. (1985). On the relation between pitch excursions and prominence. *The Journal of Phonetics, 13*, 299-308.

Robertson, T., Wright, F., and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. New York, NY: John Wiley & Sons.

Rossi, M. (2004). Intonation: Past, present and future. *Acoustical Imaging* (Vol. 27). Ed. W. Arnold and S. Hirsekorn. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rump, H., and Collier, R. (1996), Focus conditions and the prominence of pitch-accented syllables, *Language & Speech, 39*, 1–17.

Sakai, S. (2004). F0 modeling with multi-layer additive modeling based on a statistical learning technique. *Proceedings of the 5th ISCA Speech Synthesis Workshop, June 14-16, 2004, Pittsburgh, PA*.

Sakurai, A., and Hirose, K. (1996). Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours. *Proceedings of the 4th International Conference on Spoken Language Processing, October 3-6, 1996, Philadelphia, PA*, 817–820.

Savitzky, A., and Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*, 1627–1639.

Schötz, S. (2003). Paralinguistic phonetics in NLP models & methods (Term paper for course in Natural Language Processing, Lund University, Sweden). Retrieved on September 25th, 2006 from http://www.ling.lu.se/persons/Suzi/downloads/NLP_paper_Susanne.pdf.

Schroeter, J., Conkie, A., Syrdal, A., Beutnagel, M., Jilka, M., Strom, V., Kim, J.K., Kang, H.G., and Kapilow, D. (2002). A perspective on the next challenges for TTS research. *Proceedings of the IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA.

Selting, M. (2007). Lists as embedded structures and the prosody of list construction as an interactional resource. *Journal of Pragmatics, 39(3)*, 483–526.

Shih, C., and Kochanski, G. (2003). Modeling Intonation: Asking for Confirmation in English. *Proceedings of the 15th International Congress of Phonetic Sciences, August 3-9, 2003, Barcelona, Spain*.

Silipo, R., and Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS99), August 1-7, 1999, San Francisco, CA*, 2351–2354.

Silipo, R. and Greenberg, S. (2000). Prosodic stress revisited: Reassessing the role of fundamental frequency. *Proceedings of the NIST Speech Transcription Workshop, May 16-19, 2000, College Park, MD*.

Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours*. PhD dissertation, University of Cambridge.

Silverman, K. (1990). The separation of prosodies: comments on Kohler's paper. *Papers in Laboratory Phonology I* (pp. 139–151). Ed. J. Kingston, and M. Beckman. Cambridge, UK: Cambridge University Press.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling

English prosody. *Proceedings of the 1992 International Conference on Spoken Language Processing, October 12-16, 1992, Banff, Alberta, Canada*, 867–870.

Sjölander, K., and Beskow, J. (2000). Wavesurfer - An open source speech tool. *Proceedings of the International Conference on Speech and Language Processing, October 16-20, 2000, Beijing, China*.

Sproat, R. (Ed.). (1997). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Sproat, R., Ostendorf, M., and Hunt, A. (Ed.). (1999). *The Need for Increased Speech Synthesis Research: Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis*.

Stevens, S.S., and Volkmann, J. (1940). The relation of pitch of frequency: A revised scale. *The American Journal of Psychology, 53*, 329-353.

t'Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America, 65*, 811–821.

't Hart, J., and Cohen, A. (1973). Intonation by rule: A perceptual quest. *Journal of Phonetics, 1*, 309-327.

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

Talkin, D. (1993). *ESPS.*. Entropic Research Lab Inc.

Taylor, P. (1992). *A Phonetic Model of English Intonation*, PhD dissertation, University of Edinburgh, UK.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(3), pp. 1697–1714.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America, 89(4)*, 1768–1776.

Terken, J. (1993). Synthesizing natural-sounding intonation for Dutch: rules and perceptual evaluation. *Computer Speech and Language, 7*, 27–48.

Terken, J., and Hermes, D.J. (2000). The perception of prosodic prominence. *Prosody: Theory and Experiment, Studies presented to Gösta Bruce* (pp. 89-127). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Thorson, N. (1983). Two issues in the prosody of Standard Danish. *Prosody – Models and Measurements* (pp. 27–38). Ed. A. Cutler and D.R. Ladd. Heidelberg, Germany: Springer-Verlag.

Traber, C. (1992). F0 generation with a database of natural F0 patterns and with a neural network. *Talking Machines: Theories, Models, and Designs* (pp. 287–304). Ed. G. Bailly, C. Benoit, and T. Sawallis. Amsterdam, the Netherlands: Elsevier.

Traber, C. (1993). Syntactic processing and prosody control in the SVOX TTS system for German. *Proceedings of the 3rd European Conference on Speech Communication and Technology, September 22-25, 1993, Berlin, Germany*, 2099–2102.

Traber, C. (1995). *SVOX: The Implementation of a Text-to -Speech System for German.* PhD dissertation, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich, Zurich, Switzerland.

Turk, A.E., and Sawusch, J.R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America, 99(6)*, 3782–3790.

Vainio, M., Mixdorff, H., Järvikivi, J. and Werner, S. (2003). The production and perception of focus in Finnish. *Proceedings of the 15th International Congress of Phonetic Sciences, August 3-9, 2003, Barcelona, Spain.*

van Hemert, J., Adriaens-Porzig, U., and Adriaens, L. (1987). Speech synthesis in the SPICOS project. *Analyse und Synthese gesprochener Sprache* (pp. 34–39). Ed. H. Tillmann, and G. Willee. Hildesheim, Germany: Georg Olms Verlag.

van Santen, J. (1997). Combinatorial issues in text-to-speech synthesis. *Proceedings of the 5th European Conference on Speech Communication and Technology, September 22-25, 1997, Rhodes, Greece*, 2511–2514.

van Santen, J. (2002). Quantitative modeling of pitch accent alignment. *Proceedings of the International Conference on Speech Prosody, April 11-13, 2002, Aix-en-Provence, France*, 107–112.

van Santen, J., and Hirschberg, J. (1994). Segmental effects on timing and height of pitch contours. *Proceedings of the Third International Conference on Spoken Language Processing, September 18-22, 1994, Yokohama, Japan*, 719-722.

van Santen, J., Kain, A., and Klabbers, E. (2004:i). Synthesis by recombination of segmental and prosodic information. *Proceedings of the International Conference on Speech Prosody, March 23-26, 2004, Nara, Japan.*

van Santen, J., Kain, A., Klabbers, E., and Mishra, T. (2005). Synthesis of prosody using multi-level unit sequences. *Speech Communication, 46(3-4)*, 365–375.

van Santen, J., Macon, M., Cronk, A., Hosom, P., Kain, A., Pagel, V., and Wouters, J. (2000). When will synthetic speech sound human: Role of rules and data.

Proceedings of the International Conference on Speech and Language Processing, October 16-20, 2000, Beijing, China.

van Santen, J., Mishra, T., and Klabbers, E. (2004:ii). Estimating phrase curves in the general superpositional intonation model. *Proceedings of the 5th ISCA Speech Synthesis Workshop, June 14-16, 2004, Pittsburgh, PA*.

van Santen, J., and Möbius, B. (1997). Modeling pitch accent curves. *Intonation: Theory, Models and Applications, (Proceedings of the ESCA Workshop, September 18-20, 1997, Athens, Greece)* (pp. 321–324). Ed. A. Botinis, G. Kouroupetroglou, and G. Carayiannis. ESCA and the University of Athens.

van Santen, J., and Möbius, B. (1999). A quantitative model of F_0 generation and alignment. *Intonation: Analysis, Modeling and Technology* (pp. 269–288). Ed. A. Botinis. Dordrecht, the Netherlands: Kluwer Academic Publishers.

van Santen, J., Möbius, B., Venditti, J., and Shih, C. (1998). Description of the Bell Labs intonation system. *Proceedings of the 3rd ESCA Speech Synthesis Workshop, November 26-29, 1998, Jenolan Caves, Australia*, pp. 293–298.

van Santen, J., Wouters, J., and Kain, A. (2002). Modification of speech: A tribute to Mike Macon. *Proceedings of the 2002 IEEE Workshop on Speech Synthesis, September 11-13, 2002, Santa Monica, CA*.

Veenker, T. (2001). WWStim: A CGI script for presenting web-based questionnaires and experiments. Retrieved September 19, 2005, from Utrecht University: http://www.let.uu.nl/~Theo.Veenker/personal/projects/wwstim/doc/en/.

Watson, D., Tanenhaus, M., and Gunlogson, C. (2004). Processing pitch accents: Interpreting H* and L+H*. Presented at the 17th Annual CUNY Conference on Human Sentence Processing, Cambridge, MA.

Wegman, E. (1980). Two approaches to nonparametric regression: splines and isotonic inference. *Recent Developments in Statistical Inference and Data Analysis* (pp. 323-334). Ed. K. Matusita. Amsterdam, the Netherlands: North-Holland Publishing Company.

Welby, P. (2003). Effects of pitch accent position, type and status on focus projection. *Language & Speech, 46(1)*, 53–81.

Willems, N., Collier, R., and 't Hart, J. (1988). A synthesis scheme for British English intonation. *Journal of the Acoustical Society of America, 84(4)*, 1250–1261.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America, 33(2)*, 248–249.

Appendix A

Intonation Corpus Generation

A.1 Utterances from which phrase curves were obtained for the POC implementation of recombinant synthesis

1. One played the bass viol, the other the fiddle, and the third put the trumpet to his lips, and blew out his cheeks as much as he possibly could.
2. The eldest determined to be a blacksmith, the second a barber, and the third a fencing-master.
3. He bestowed on him much gold, and lands, and meadows, and herds, and made him immensely rich.
4. Ivan killed the lamb, lit a fire, threw the meat into the kettle, and boiled it.
5. The princess arose, living, healthy, and beautiful.
6. They consume much, pay for nothing, play mischievous tricks, and steal his belongings.

A.2 Utterances from which accent curves were obtained for the POC implementation of recombinant synthesis

1. The first, a powerful king, the second, a prince, the third, a count, the fourth, a knight, the fifth, a nobleman, the sixth, a merchant, and the seventh, a teacher.
2. Let me be your companion and play-fellow, and sit by you at your little table, and eat off your little golden plate, and drink out of your little cup, and sleep in your little bed.

3. One hid himself beneath the hay, another under the roof, a third in the straw, the fourth in the stove, the fifth in the cellar, the sixth under a tub, the seventh beneath the wine-cask, the eighth under an old fur cloak, the ninth and tenth beneath the cloth out of which she always made their clothes, and the eleventh and twelfth under the leather out of which she cut their shoes.
4. Twenty miller's men sat cutting, hick hack, hick hack, hick hack, and the mill went klipp klapp, klipp klapp, klipp klapp.
5. Come to me not clothed, not naked, not riding, not walking, not in the road, and not off the road, and if you can do that I will marry you.
6. I must have three dresses, one as golden as the sun, one as silvery as the moon, and one as bright as the stars.
7. She took three different things from her treasures, a golden ring, a golden spinning-wheel, and a golden reel.
8. No wind stirred, no brook murmured, no bird sang, and no sunbeam forced its way.
9. One put his stockings on for him, one tied his garters, one brought his shoes, one washed him, and one dried his face with her tail.
10. There were four candles, four loaves of bread, and four bottles of wine on the table.
11. The flowers were made of green silk, the bed of ivory, the canopy of red velvet, and the slippers of glass.
12. There the child fared well, ate sugar-cakes, and drank sweet milk, and her clothes were of gold, and the little angels played with her.
13. My young nephew has a snug berth, a nice bit of money in hand, and plenty of bread to break; besides, he has quite as many patches as I have.

Biographical Note

Taniya Mishra was born on June 25, 1977 in Calcutta, India. She received her Bachelor in Arts degree in Computer Science and Mathematics from Saint Mary's University of Minnesota, Winona, Minnesota, in May 2001, graduating magna cum laude. She joined the OGI School of Science and Engineering at OHSU in September 2001 as a graduate student. While at OGI, her area of expertise was speech synthesis with special focus on intonation modeling. Taniya is a scholarship winner to the Grace Hopper Celebration for Women in Computing. She is the author of eight conference publications, two journal articles, and one book chapter.