NORTHWESTERN UNIVERSITY

The Determination, Analysis, and Synthesis of Fundamental Frequency

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Communication Sciences and Disorders – Speech and Language Pathology

By

Xuejing Sun

EVANSTON, ILLIOIS

December 2002

ABSTRACT

The Determination, Analysis, and Synthesis of Fundamental Frequency

Xuejing Sun

Intonation modeling is important for understanding the human speech communication process and developing sophisticated speech technology applications. The goal of this research is to finding ways to improve the quality of determination, analysis, and synthesis of fundamental frequency (F0) for intonation modeling.

The problem of F0 determination was approached by developing an accurate pitch determination algorithm. The algorithm originated from an effort to find solutions for processing alternate cycles in speech, which is notoriously difficult to handle in intonation research. The algorithm determined F0 by computing Subharmonic-to-Harmonic Ratio (SHR). Evaluation results showed that the proposed method outperformed several state-of-the-art algorithms for pitch determination. Towards the so-called "intonation contours", the algorithm suppressed the subharmonics, manifested by alternate cycles in the time domain, by adjusting the SHR threshold to yield continuous and smooth F0 curves.

The problem of F0 analysis and synthesis was approached by applying the notion of underlying pitch target in intonation modeling. An automatic procedure for extracting underlying pitch target from a surface F0 contour was developed. The model parameters were evaluated by (1) examining their correlations, (2) predicting their values from high-

level linguistic features, and (3) using their values to predict pitch accent with machine learning methods. Results showed that the parameters were statistically independent and linguistically predictable.

Finally, based on both theoretical understanding and computational considerations, an F0 generation system for speech synthesis was proposed. This system is characterized by (1) a hierarchical structure with four tiers, which allows for modeling different intonation components in different tiers, (2) the underlying pitch target as the phonetic representation of intonation, which contains as few as two parameters, and (3) the ensemble decision tree training, which provides better learning performance and makes the system adaptable to other speaking styles and languages. Both objective and subjective evaluation results indicated that the proposed system performed better than other approaches, which confirmed (1) the necessity of an intonation hierarchy and an appropriate phonetic representation, and (2) the importance of decoding underlying pitch target from the surface F0 form for intonation modeling.

Acknowledgments

First, I would like to thank my advisor, Yi Xu, for directing my research on intonation modeling and supporting me both mentally and financially in the past several years. His critical comments and insights greatly strengthened this work. He sets high standard on both research and writing. I also would like to thank my dissertation committee members, Janet Pierrehumbert, Bruce Smith, and Mark Randolph, for valuable discussion and suggestions on the thesis.

I owe many thanks to Kimberly Fisher and Charles Larson for providing many suggestions to the pre-dissertation projects. I am grateful to Jean-Claude Junqua for giving me an internship opportunity to work in Panasonic Speech Technology Laboratory. Ted Applebaum in Panasonic Speech Technology Laboratory was my mentor during the internship, who helped me to conduct data-driven prosody modeling experiments. I am thankful to Zheng Shen, who is the advisor of my bachelor's thesis - my first research project in speech.

I wish to thank all the staff members and students in our department. Particularly, Jay Bauer, Nina Capone, Pan Chen, Ciara Leydon, Danielle Lodewyck, Li Sheng, and Xueqing Xu read the manuscript. Xueqing Xu provided many insights on intonation modeling throughout this research.

Many thanks are also to numerous speech researchers. Alan Black and Paul Taylor provide the Festival speech synthesis system and Edinburgh Speech Tool library for free. Paul Boersma makes the Praat software freely available. Without these tools, this thesis would remain to be a remote possibility. I thank George Meyer for Keele database setup, and Herbert Griebel for bug fixing in the pitch determination program.

Finally, I must thank my family and friends, without whom I could never have begun this research. Especially I would like to thank Jing Yang, my wife, for endless love, support, and patience during the whole course of Ph.D. study.

This work is supported in part by NIH grant DC03902. The graduate school of Northwestern University awards me the University Fellowship and provides me a research grant and travel fund. The department of Communication Sciences and Disorders awards me a teaching assistantship and reimburses part of my travel expenses. Part of the travel funds is from International Speech Communication Association (ISCA) and Acoustical Society of America (ASA).

Contents

List of Tables

List of Figures

Figure Page

Chapter 1

Introduction

This is a dissertation about intonation modeling. Intonation modeling refers to a process in which intonation is analyzed, represented, predicted, and synthesized through certain methods so that various intonational phenomena can be processed and reconstructed in a systematic and manageable way.

1.1 The Role of Intonation

Intonation plays an important role in the human speech communication process. Intonation is realized mainly by varying fundamental frequency (F0), which in turn is the result of changing the rate of vocal fold vibration. Controlling vocal fold vibration is a complex and elaborate process, in which many physiological apparatus are involved (Titze, 1994). This advanced F0 production system makes rich intonation variations possible. The perceptual correlate of F0 is pitch. The human hearing system is very sensitive to variations of pitch (Moore, 1989), which provides the physical foundation of effective intonation perception. From the speaker's point of view, intonation is used to convey pragmatic information, emotion, etc. Syntactically similar sentences with different intonation patterns can convey dramatically distinct information. From the listener's point of view, intonation plays an

important role in: (1) segmenting utterances; (2) resolving syntactic ambiguity; (3) serving as a continuity guide in noisy environments; and (4) providing cues to the state of the speaker (O'Shaughnessy, 2000). In addition to the above functions, pitch contours also carry lexical meaning in tone languages, such as Mandarin Chinese.

Because intonation plays an important role in the human speech communication process, understanding intonation would be crucial to speech technology applications, such as speech synthesis and speech recognition/understanding. One of the goals of these technologies is to seamlessly bridge the human-computer speech communication process. In speech synthesis, the naturalness of intonation directly affects the overall quality of synthetic speech. As a result, intonation modeling has been one of the central areas in speech synthesis research. Figure 1.1 illustrates the schematic architecture of a speech synthesis system, in which intonation prediction is an intermediate step between text analysis and waveform generation. Intonation also provides valuable information for speech recognition systems. As Figure 1.2 shows, one way of utilizing intonation in speech recognition is combining F0 with spectral information to identify individual speech units, such as a phoneme, word, etc. Intonation can also be applied at later stages. Certain prosodic events, such as pitch accent and prosodic boundaries, can be predicted from acoustic utterances using features like F0, duration, and energy. A speech recognition/understanding system can use these symbolic prosodic markers to identify some syntactic, semantic, or emotional information.

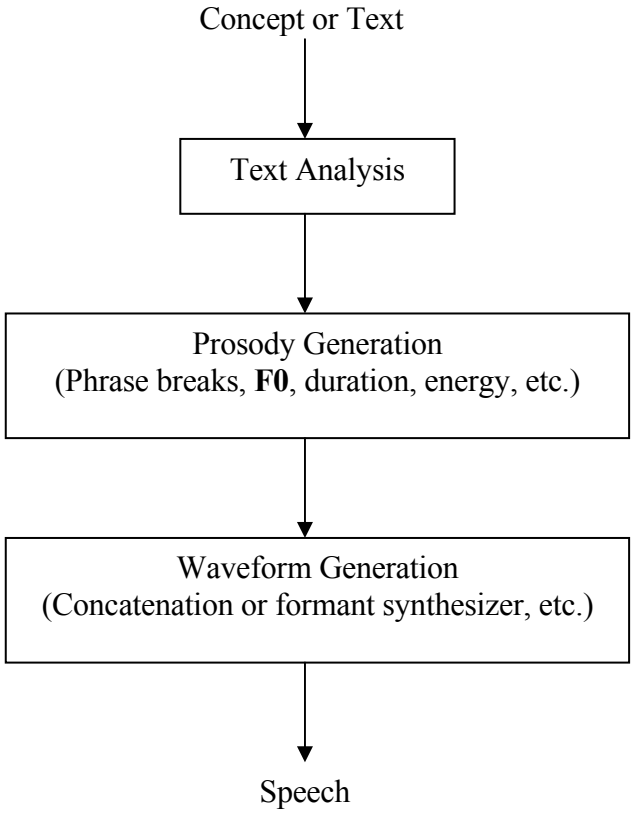


Figure 1.1: Simplified block diagram for a speech synthesis system.

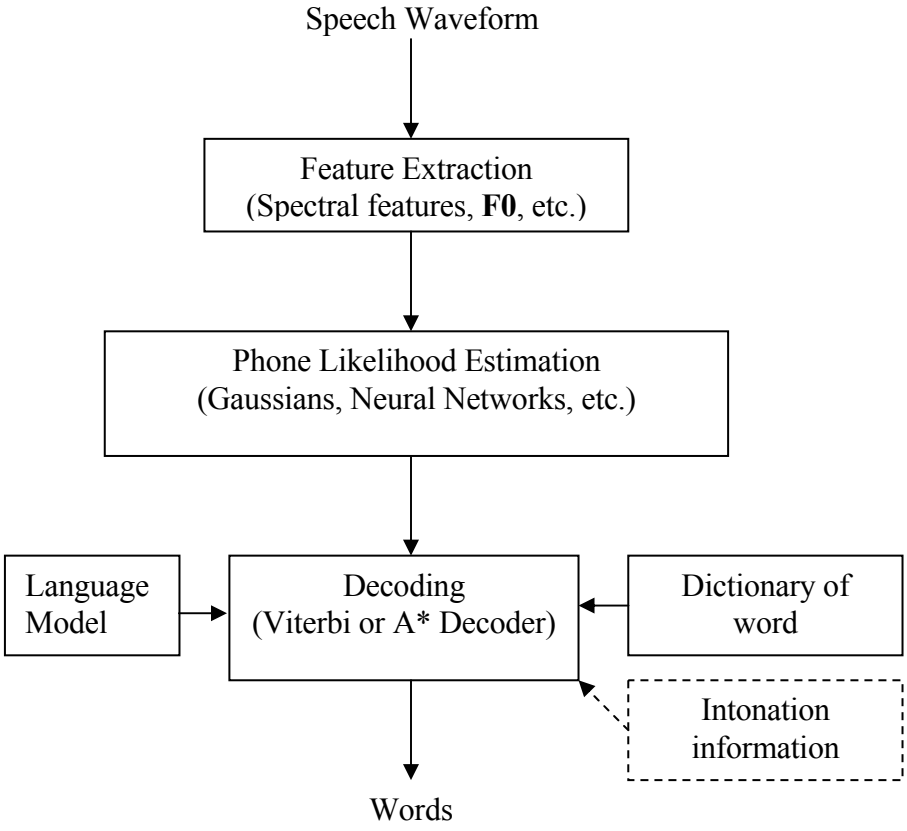


Figure 1.2: Simplified block diagram for a speech recognition system.

1.2 Theoretical Issues Concerning Intonation Modeling

Intonation has been intensively studied in the past decades (Botinis, 2001). Although significant advancements have been made, many issues remain unresolved. This is because our theoretical understanding of intonation is still incomplete. Debate continuous over some of the old issues and in the meantime, new challenges have arisen that have not yet been explored.

Incomplete knowledge and various specific issues limit the success of speech technology application. For example, the application of intonation in speech recognition has not been very successful. Taylor (2000) attributed this to the usage of inappropriate intonation models, which often underutilize acoustic F0 information. The rest of this section addresses some important issues concerning intonation modeling as well as F0 generation for speech synthesis. Figure 1.3 shows a simplified block diagram of an F0 generation system, which demonstrates that F0 synthesis closely depends on the quality of the underlying intonation model.

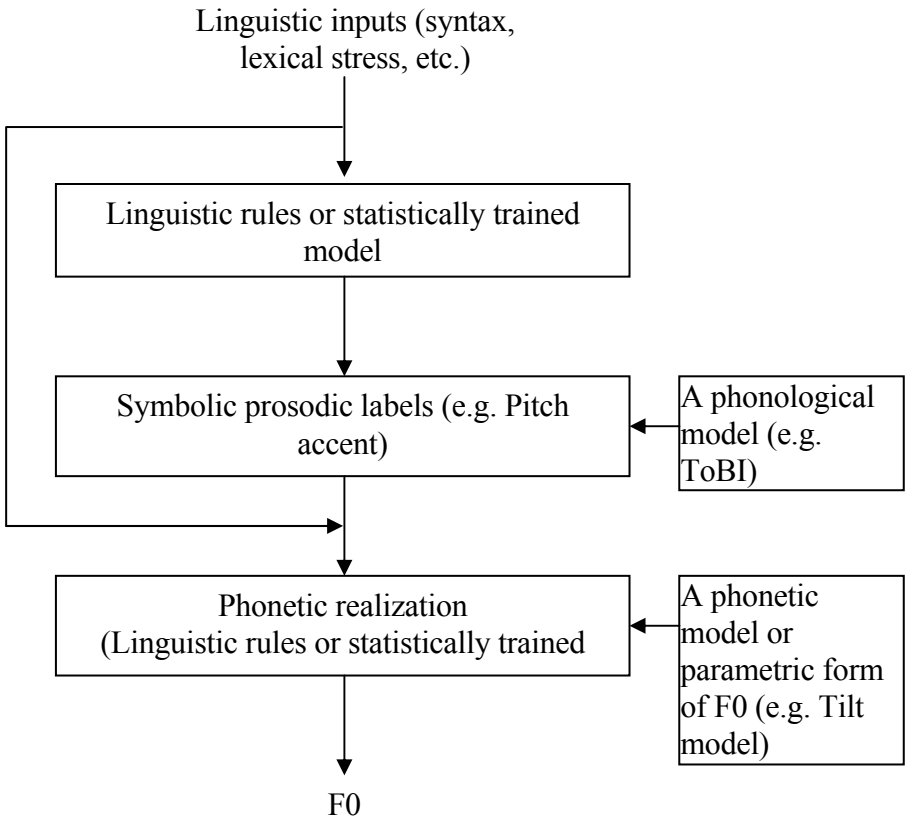


Figure 1.3: Simplified block diagram for an F0 generation system.

1.2.1 The Intonation Hierarchy

Among various theories, it is generally acknowledged that the communicative functions of intonation have many levels (Botinis et al., 2001; Ladd, 1996). At the lexical level, F0 can be used for tone distinction, as in Mandarin Chinese and for stress distinction, as in Greek. At sentence or phrase level, intonation may signal the location of focus. At discourse level, intonation can convey information such as, a new topic. All these functions are realized simultaneously, which makes the final form of intonation very complicated. Furthermore, the final form intonation will take also integrates other physiological, linguistic, paralinguistic, and extralinguistic factors (Botinis et al., 2001), etc. Different from the intonation functions, which are intended by the speaker, these factors contribute to intonation in the form of constraints. For example, intonation variations are limited by the maximum speed of pitch change that can be achieved by the human speech production apparatus (Xu and Sun, 2002).

The multi-level communicative functions imply a hierarchical structure of intonation (Botinis, 2001; Ladd, 1996, Chapter 6). That is, larger prosodic units are composed of smaller prosodic units. The hierarchical structure is usually represented using a tree diagram. The following is a simple tree diagram from Ladd (1996) which shows Halliday's four English prosodic structure types: Utterance (Utt), Tone Group (TG), Foot (F), and Syllable (Syl).

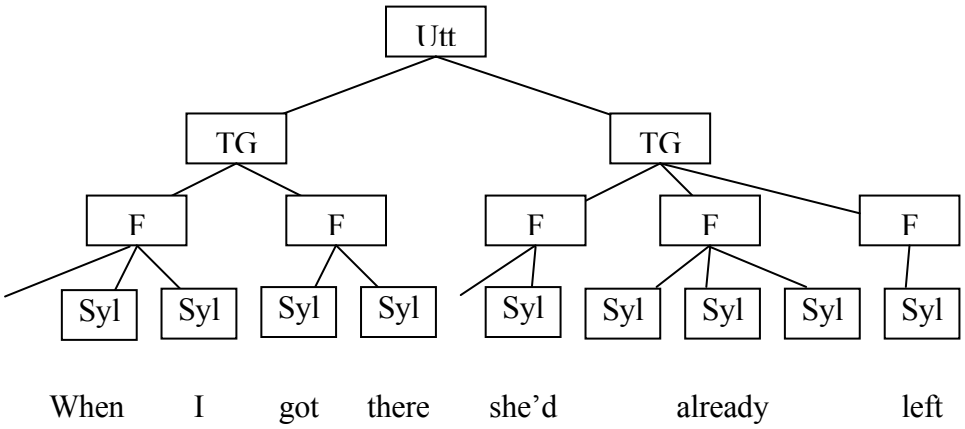


Figure 1.4: A tree diagram of hierarchical prosodic structure.

Various intonation theories propose quite different inventories of category or domain types. For example, Pierrehumbert's model (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986) contains intonational phrase (IP) and intermediate phrase (ip). Furthermore, there are arguments on whether the hierarchical structure should be strictly non-recursive (Ladd, 1996). Nonetheless there seems to be a clear consensus on the existence of a hierarchy.

This hierarchical structure has implication for intonation modeling. Intonation has multiple levels, and each level may have its unique patterns and possible interactions with other layers. The lowest level in the tree constitutes the smallest prosodic unit, which forms the basis of the larger domain types. Syllable is usually regarded as the smallest type in this prosodic structure. In other words, F0 contour of a syllable is treated as the smallest unit, and not separated into several parts. However, there are exceptions. In English, some

syllables could be very long and contain complex intonation contours similar to multi-syllabic units. In such cases, viewing the syllable as a single unit may result in the loss of information.

Although many researchers agree upon the hierarchical structure of intonation, it remains controversial whether there are interactions among the different levels. The superpositional approach (e.g. Fujisaki's model (Fujisaki,1983, 1988)) argues that there are strong interactions among different layers, where different layers are superimposed on each other. As shown in Chapter 2, superpositional models usually contain an explicit global intonation component at the phrase level, which determines the overall trend of F0 contour. In contrast, linear views (Beckman and Pierrehumbert, 1986; Pierrehumbert, 1980) contend that having such modules implies that speakers are pre-planning intonation across the entire utterance, which would seem to be implausible as it greatly increases the cognitive load. On this basis, proponents of the linear views hold that the interaction is minimal. The global "declination" trend may actually be the result of phonologically triggered *downstep* that occurs iteratively or the paralinguistic signaling of newness. Larger scale intonational patterns are seen as the results of concatenation of smaller intonation units. Xu (2001) attributes downstep, non-final focus and new topic as three separate sources of declination. He further identifies two separate components of downstep, anticipatory raising and carryover lowering, and two separate components of focus, on-focus pitch range expansion and post-focus pitch range suppression. Xu (1999) shows both numerically and graphically how the effect of downstep and focus contribute separately but additively to the overall

down trend in Mandarin. Taylor (2000) argues that adding such a phrase component constrain the model in accounting for the wide range of observed long term contour shapes in English. Kochanski and Shih (2002) explicitly assume the existence of pre-planning in intonation production, but do not enforce a phrase component in their system.

The approach taken in this work assumes that there are interactions among different layers, but does not include an explicit function to model global trends. The phrase level F0 contour is the concatenation result of word F0 contours, which in turn are the joints of syllable F0 contours. A syllable F0 contour is the smallest unit in this intonation hierarchy. Inside each syllable, different intonation components are vertically superimposed rather than concatenated horizontally. For example, in Mandarin, the F0 shape of a syllable may well be determined mainly by its lexical tone, but the overall height may be further determined by a larger scope intonation pattern. In this hierarchy, different layers cover different intonation domains. The prosodic units at the top of the tree represent more global intonation patterns while the units at the bottom are responsible for more local effects. Therefore, various factors at different levels should be applied to the corresponding prosodic units. Higher-level factors may determine the global intonation variation patterns, whereas factors at the segmental level may affect local pitch movements. Many factors have been found to contribute to intonation variation (Botinis, 2001; Xu, 2001). Xu (2001) classifies some known factors into two categories: voluntary vs. involuntary.

Voluntary factors include:

- Lexical tone

- Prosodic structure
- Syntax
- Pragmatics
- Emotion

Involuntary factors include:

- Overall pitch range of a speaker
- Vowel intrinsic pitch
- F0 perturbation by consonants
- Maximum speed of pitch change
- Maximum speech pitch direction shift
- Coordination of laryngeal and supralaryngeal movements

Unfortunately, properly correlating these factors to different layers is not an easy task. The proposed F0 generation system in this thesis represents an initial attempt to apply various factors to different layers.

1.2.2 The Relevancy of Intonation Components

In an ideal situation, with available information we want to generate exactly the same intonation patterns as would be produced by a speaker. Unfortunately the current status of intonation modeling indicates that this goal is probably unrealistic. Therefore we may need to focus on the most important component first. So what is the criterion for defining "important" components? Perception is probably the criterion that most people would agree upon. It is known that certain F0 variations are not perceptible intonationally, which implies that certain intonation components can be ignored when using perception as the criterion ('t Hart et al., 1990).

So what should we model and what should we ignore? Clearly researchers have quite different views on this issue. One group of models, such as Pierrehumbert's model (Pierrehumbert, 1980; Pierrehumbert, 1981), Tilt model (Taylor, 2000), and the PaIntE model (Mohler and Conkie, 1998), put enormous emphasis on modeling pitch accent, and specify little phonetic properties for the transition between pitch accents. For example, Pierrehumbert (1981) proposes a "sagging transition" between consecutive H* accents. The Tilt model (Taylor, 2000) uses linear transitions to interpolate between intonation events. The essential idea behind this is that the F0 between two accents is just a function of distance and does not represent a meaningful target (Ladd and Shepman, in press). On the other hand, other approaches (e.g. Black and Hunt, 1996; Ross and Ostendorf, 1999) predict F0 values for every syllable. Empirical experiments conducted in this thesis on the Tilt model have indicated that using interpolation to connect accented syllable sometimes yields unsuccessful perception results. This seems to imply that some perceptually significant information exists between accented syllables which cannot be ignored or replaced by simple interpolation. Recently emerging evidence has shown support for this assumption. In English, Ladd and Shepman (in press) present converging results from intonation production and perception experiments that an F0 valley between two H* peaks is a phonologically specified tonal target and should therefore be treated as the reflex of an L tone. In Mandarin, Chen and Xu (2002) show that a neutral tone does have a phonetic target, which is conventionally assumed to be phonologically unspecified as judged by the contrast criteria.

The above discussion focuses on whether an intonation model should consider only pitch accent or both accented and unaccented components. Suppose we have already solved this problem, now the question is what needs to be modeled for a pitch accent (or nonaccent). First, does a phonetic realization model for a pitch accent matter much? Some researchers (e.g. Monaghan, 1992) hold that models for predicting abstract prosodic labels (e.g. phrase break, pitch accent) from the text need the most work while available phonetic realization models of these labels are adequate. Other researchers believe that a phonetic model is crucial (e.g. Taylor, 2000). Moreover, Silverman (1987) argues that not only is such a phonetic model necessary, but a detailed modeling of segmental intonation variation (i.e., micro-intonation) is also important. One of the segmental effects he examined is the consonant perturbation effect. He found that preceding consonants can raise the pitch of the following vowel while the following consonants often lower the preceding vowel pitch, and voiceless consonants appear to have greater effects. Moreover, the perturbation effects seem to be amplified by the stress placed on a syllable. Xu et al. (2002) also found similar results for Mandarin Chinese. Van Santen and Hirschberg (1996) stress the importance of segmental composition on intonation by showing the close relationships between syllable onset, coda types and the peak height and alignment. Dusterhoff (2000) presents a comprehensive review and discussion of segments and intonation. By reviewing previous studies, he identified that listeners do perceive and use micro-intonation, but for segmental, rather than intonation perception. He then concludes that “it seems logical, therefore, that it

is more important in intonation synthesis to correctly model macro-intonation than the micro-intonation over intonational 'connective tissue'." (Dusterhoff, 2000, p. 37)

This thesis acknowledges the importance of predicting abstract prosodic labels but also argues that current phonetic realization is not adequate and a proper model is still needed. The key point is to identify the underlying form of intonation. For example, it is well known that the four lexical tones in Mandarin have canonical forms as High, Low, Rise, and Fall. It has also been reported that in connected speech these canonical forms often go through significant changes. Thus to model tones, it is necessary to decode tones from surface F0 contours. Such a decoding process requires knowledge about the source of contextual tonal variations. Xu and Wang (2001) argue that a significant portion of these contextual tonal variations are due to articulatory constraints that speakers cannot avoid. Speakers can only approximate the underlying pitch targets (lexical tones) asymptotically during production. This is contrary to the traditional view that these contextual variations are *phonologized*. Due to the articulatory constraints, a speaker cannot reach the targeted canonical form in a timely fashion. Time is needed to make the transition from the previous tone. The effect from the previous tone diminishes gradually after the transition point. Assuming this articulatory constraint is language independent, one may also view English intonation as a result of asymptotic approximation to some underlying pitch targets. By referring to certain intonation component as an *underlying* target, we assume that it is an intended goal that is deliberately controlled by the speaker. Thus, it has greater connection with higher-level linguistic information, such as syntax, semantics, and pragmatics. The

implication of this is that such underlying targets should be treated differently from other components which are mostly determined by local or segmental factors.

1.2.3 Level Tone vs. Pitch Movement

The debate over level tone vs. pitch movement has a long history (cf. Ladd, 1996 and the references therein). Some researchers (Ladd, 1996; Pierrehumbert, 1980) argue that level tones are the primitive unit of intonation, whereas others believe that pitch movement is of the utmost importance ('t Hart et al., 1990). In Xu and Wang (2001), the high and low tones in Mandarin are referred to as static targets whereas the rise and fall tones are dynamic targets. In this thesis, level tones are referred to as static targets and pitch movements are dynamic targets. It is proposed that speakers employ both static targets and dynamic targets even in non-tonal languages, such as English. It is hypothesized that these targets are operated at different levels and implemented simultaneously. At a higher level, with more cognitive factors involved, speakers are trying to reach a static (level) target. This one-dimensional variation, high or low, may be sufficient to express most of the voluntary intentions and brings minimal cognitive burden to speakers. The level tone is thus seen as a phonological target. The variation of a level tone target may be seen as the pitch range variation. Global functions, such as focus, operate on pitch range rather than on local contours. Pitch range includes both overall level and span (Ladd 1996, chap. 7; Xu and Wang, 2001). At lower levels, pitch movement becomes dominant, which greatly enriches the inventory of intonation variation patterns. Intuitively, pitch movement is more complex

than a level tone in terms of both production and perception because it has the freedom on a two-dimensional plane. It is assumed that some pitch movement patterns may be conventionalized because they are at a low level and are repeated more frequently. Thus, the cognitive load caused by using pitch movement as an intonational unit may be reduced. Pitch movement carries detailed intonation information. Hence, it is considered as a phonetic target.

1.3 Practical Issues Concerning Intonation Modeling

In addition to the theoretical issues discussed above, practical considerations should be made to build a successful F0 generation system for speech synthesis.

1.3.1 Parametric Representations of F0

Finding a proper parametric representation of F0 is another challenging task that is often unavoidable. For illustration purposes, let's consider the model in the syllable domain. The representation forms can be put on a continuum. At one end of the continuum, the simplest model may only contain one F0 target for each syllable, which may be the raw F0 value, or a transformed value. At the other end, the most complex model may contain all the F0 values in one syllable or a large set of parameters. Most models, such as the Tilt model (Taylor, 2000), fall somewhere in between. Consequently, there is a dilemma in describing F0: (1) simple parametric forms may result in low accuracy; and (2) complex forms may be low in predictive power or require significant amount of data to train. Building an efficient

yet accurate and predictable parametric model is not easy. First, we need to choose parameters that are linguistically meaningful. One may argue that this is not necessary since in theory a sophisticated machine learning algorithm (e.g., Neural networks) can learn any nonlinear relationship between the input and output targets. However, in practice this is hardly true due to insufficient training data. Thus, a linguistically motivated parameter will always be beneficial if only limited data are available which is often the case. Second, in order to make the model compact, the correlation between parameters should be low. Highly correlated parameters contain redundant information. Using orthogonal parameters will make a model more compact without losing much information. Furthermore, this would allow us to choose a specific and compact feature set based on existing linguistic knowledge since each parameter focuses on different aspect of intonation. Keeping these issues in mind, the proposed system in this thesis employs underlying pitch targets as the basic parametric representation of F0 and strives to maintain simplicity, accuracy, and predictive power.

1.3.2 Adaptability

Intonation can vary substantially across different speakers, speaking styles, or languages. For example, a news-reading style speech characterized by declarative statements can contain quite different intonation patterns from conversational speech. This requires that the system be adaptable to accommodate the new variation patterns. To make a system adaptable to different languages, the system should be relatively language neutral and

capture some mechanisms that are universal to many languages. It should also be flexible enough to allow for fine tuning of a particular language. In a rule-based system, manual adjustments need to be made to add new rules or delete old rules. This process could be very labor intensive. Furthermore, it requires the designer be familiar with the intonation patterns of both the source and target domains. On the other hand, if a system has a structure that is statistically trainable, it can be re-trained for each new domain. Rules are derived automatically by the learning algorithms used for training. The premise for training a system is there is certain amount of data available from the target domain. Moreover, the database should contain all information required by the system. In practice, annotating the database is often a more challenging task than acquiring the raw data. For instance, many systems rely on ToBI (Silverman et al., 1992) labels, which can be difficult to obtain efficiently and reliably. The constraint posed by feature availability is discussed in the next section. In this thesis, the parametric representation of F0 is based on the principle that is assumed to be language independent. The hierarchical structure of the system is also applicable to many languages and speaking styles. To make it trainable, the system is built upon several decision trees which can be derived from a corpus automatically.

1.3.3 Feature Availability

One of the practical constraints in F0 generation for speech synthesis is that not all intonation-relevant information is readily available. For example, as is known, semantic or pragmatic information is important for intonation prediction. Unfortunately, such

information is often difficult to extract from unrestricted text. This implies that an F0 generation system should exploit the available features to a maximum extent and be open to new features when they become available. Effective utilization of the input features requires that the designer understand the relationship between various factors and intonation phenomena and use the most appropriate features for the corresponding intonation components. One solution to the openness of new features is to make the system trainable as was suggested in the previous section. That is, when new features are added, the mapping between input features and the output F0 values would be adjusted automatically with prespecified learning algorithms. It can be seen later that the proposed system tries to link different features to different intonation patterns in order to use features more effectively. The decision tree trained system makes it flexible to add new features.

1.3.4 Pitch Determination

Another difficulty often faced by intonation researchers is how to extract *suitable* F0 values from acoustic utterances. Due to the nature of intonation research, an accurate pitch determination algorithm (PDA) is probably not enough. Intonation research often requires a continuous and smooth F0 contour, namely "intonation contour" (Dusterhoff, 2000). Note that such a claim does not necessarily mean that the abrupt jumps, resulting from effects like glottalization are meaningless in intonation. In fact, some researchers argue that they might signal linguistic information under certain circumstances (Redi and Shattuck-Hufnagel, 2001). However, at present, most intonation research focuses on the pitch

movement patterns on a larger domain, which is also the emphasis of this thesis. Also, in some cases, these variations may not be intended by the speaker to convey intonation-related information. Conventional approaches usually try to eliminate these local pitch jumps by using methods such as smoothing, dynamic programming, EGG signal comparison, and manual correction, etc. Smoothing is probably the most widely used, particularly median smoothing, which is quite effective in removing local outliers. However, when a "bad" segment is long, using a median smoother cannot solve the problem. This thesis attempts to tackle the problem by developing a novel pitch determination algorithm (see Chapter 3).

1.4 Thesis Goal and Outline

The goal of this thesis is to improve the quality of determination, analysis, and synthesis of fundamental frequency. Specifically, by addressing the above issues, this thesis attempts to show theoretically: (1) the importance for a multi-tier hierarchical representation of intonation; and (2) the importance of appropriate phonetic formation that allows modeling various intonation components selectively. Practically, this thesis intends to demonstrate: (1) the importance of using predictable and orthogonal parameters to represent F0 curves; (2) the advantage of employing a data-driven approach; and (3) the impact of proper pitch determination.

The remainder of this thesis is organized into six chapters. Chapter 2 reviews various existing theories, models, and techniques developed for intonation modeling. In

Chapters three through six, the proposed solutions to the issues mentioned earlier are presented. Specifically, Chapter Three introduces an intonation oriented pitch determination algorithm, which targets at certain F0 tracking problems that trouble intonation research. In Chapter Four, the analysis and synthesis of intonation using underlying pitch target is presented. First, an automatic procedure to extract underlying pitch targets from the surface F0 contours is described. Then several experiments are conducted to show the eligibility of underlying pitch targets for intonation modeling. Based on the parametric form described in Chapter Four, an F0 generation system for speech synthesis is proposed in Chapter 5. This system employs a multi-tier hierarchical intonation structure and is statistically trained via multiple decision trees. Chapter 6 describes a perception experiment to evaluate the F0 generation system. Finally, Chapter 7 concludes the main contributions of this thesis and suggests potential directions for future work.

Chapter 2

A Review of Intonation Models and F0 Generation Methods

To address various theoretical issues and build successful applications, many intonation models have been proposed in the past. These models are often compared from certain angles, such as whether the system is phonological or phonetic; whether the system views intonation production as a linear or superpositional process; whether the basic intonational unit of the system is level tone or pitch movement (Botinis et al., 2001; Ladd, 1996; Taylor, 2000). This chapter describes intonation models along the axis of phonological vs. phonetic. It should be noted that using such a dichotomy to describe intonation models does not mean that there is no phonetics in a phonological model or vice versa. A detailed comparison between the models is presented after the review of intonation models. Beside intonation models, several commonly used F0 generation methods are also reviewed in this chapter.

Before digging into the details of each model, it is helpful to picture an ideal intonation model. In Crystal (1969), six principles were proposed: high accuracy, high consistency, automatic applicability, simplicity of symbol set, degrees of complexity for the symbols which reflect the significant differences in the data, and a restriction to cover only those aspects of intonation which are linguistically significant. Taylor (2000) defines three

desired properties of intonation representation: (1) constrained: the representation should be compact and contain least the amount of redundancy; (2) wide coverage: the model should be able to distinctively describe various intonation patterns; (3) linguistically meaningful: the model parameters should be predictable from high-level linguistic information. Hirst et al. (2000) believe that a satisfactory global theory of intonation will require four levels of analysis: (i) physical (acoustic, physiological) (ii) phonetic (iii) surface phonological and (iv) deep phonological. It can be seen that researchers share some common views about an ideal intonation model, that is, it should be accurate, efficient, and predictable. Moreover, automatic conversion between the model and F0 contours is also desirable, especially for a practical system used by TTS applications.

2.1 Phonological Models

The goal of a phonological model is to study the universal organization and underlying structure of intonation. Complex intonation patterns are compressed into a set of highly succinct and abstract vocabulary with wide coverage. These (categorical) symbols are regarded as the primitive entities in representing intonation. The development of such an inventory is often based on phonetic analysis of F0 curves either from a production perspective or from a perception perspective.

2.1.1 Autosegmental-Metrical (AM) Approach

One of the dominant views on intonational phonology is the Autosegmental-Metrical (AM) approach (Goldsmith, 1990; Ladd, 1996). Ladd (1996) states four principles of the AM approach to intonation:

- Linearity of tonal structure
- Distinction between pitch accent and stress
- Analysis of pitch accents in terms of level tones
- Local sources for global trends

The most influential example is Pierrehumbert's model for American English (Pierrehumbert, 1980). The original model has been extended by Beckman and Pierrehumbert (1986), and has evolved into a standard for transcribing intonation of American English - Tone and Break Indices (ToBI) (Silverman et al., 1992).

The main idea of Pierrehumbert's model, or AM approach in general (Ladd, 1996), is that English intonation can be represented by two types of tones: High and Low. The model describes intonation with multiple levels. The largest prosodic unit is called *intonational phrase*, which can contain several *intermediate phrases*. A phrase is represented as a sequence of high (H) or low (L) *tones*. The tone inventory is described below:

Pitch Accents can be either single tones (H*, L*) or bitonal (H*+L, H+L*, L*+H, L+H*). The "*" symbol represents the alignment of a tone with the accented syllable. All these symbols have corresponding phonetic interpretations. For example, H* is a local peak

aligned with the accented syllable and L* is a local valley, and L*+H is an accent contour that is low for a good portion of the accented syllable and then rises sharply (Ladd, 1996). The rise can extend into the following unstressed syllable if there is one.

Boundary tones, denoted by the "%" symbol, are single High or Low tones aligned with the edges of a phrase (H%, L%). Thus, they indicate the onset and offset pitch of an intonational phrase, respectively.

Phrase accents, indicated by the "-" symbol, are used to represent the pitch movement between a pitch accent and a boundary tone (H-, L-).

In addition to the tone labels described above, the ToBI (Silverman et al., 1992) transcription system incorporates break indices ranging from 0 to 4, where a larger number represents a higher degree of de-coupling between each pair of words. Specifically, break index 4 refers to intonational phrase boundary (i.e. major break); 3 refers to intermediate phrase boundary (i.e. minor break).

Pierrehumbert's model has a linear structure in describing intonation in that intonation is solely determined by a local component, which is in contrast to the superpositional approach which treats intonation resulting from the addition of several components, such as a local pitch accent and a global phrase contour. Also, in the Pierrehumbert model, level tones (High or Low) are assumed to be the basic intonational units.

The mapping from phonology onto acoustics and physiology is a *dynamic interpretative process* (Pierrehumbert and Beckman, 1988). In this process, a set of

phonetic realization rules are applied to convert abstract tonal representation into F0 contours by considering the metrical prominence of the syllables and the temporal alignment of the tones with the accented syllables (e.g. Pierrehumbert, 1981). This direct phonology-F0 mapping implies that a distinct layer of phonetic model is not necessary.

2.1.2 The IPO Perception Based Approach

The IPO model, developed at Institute of Perception Research at Eindhoven, is probably the best-known perceptual model of intonation ('t Hart et al., 1990). The essence of this model is that only those *perceptually relevant pitch movements* are important to intonation and that natural F0 contours can be reduced to perceptually relevant straight lines by a *stylization* procedure. To realize this, three steps are taken in the intonation analysis. First, the *perceptually equivalent* approximations or *close-copy stylizations* are obtained with an interactive procedure. A listener replaces an F0 contour with a minimum number of straight lines in the logarithmic F0 domain, with which the re-synthesized intonation is perceptually identical to the one synthesized using original F0 contour. In the second stage, a *standardization* procedure is performed to construct a pitch movement inventory. The pitch movements are categorized into discrete, phonetically defined types of F0 rises and falls, according to the values of some common features, such as direction, range, and duration. In the third step, an *intonation grammar* is defined to regulate possible and permissible combinations of pitch movements into longer-range contours, as well as the sequencing of the contours across prosodic boundaries. It can be seen that compared with the AM

approach, the IPO model regards pitch movement, rather than level tones, as the basic intonational unit.

Originally developed for Dutch, the IPO model has been applied to other languages. It was also implemented in speech synthesis systems for Dutch and English (Willems et al., 1988).

The core of the IPO model is stylization. However, since stylization depends on the perception of human listeners, there could be inter- and intra-listener variation. The stylization procedure is also time-consuming and labor intensive. To overcome these issues, automatic stylization of intonation contours has been proposed (d'Alessandro and Mertens, 1995), in which a psychoacoustic criterion *glissando* rate is applied to select perceivable pitch movement.

The experimental results of IPO model suggest that not all aspects of intonation are perceptually relevant. Some intonation details that are not *important* can be ignored. This is an important implication for modeling intonation, especially when the goal is to generate perceptually natural intonation, for instance, in speech synthesis.

2.2 Phonetic Models – Parametric Approach

Phonetic models use a set of continuous parameters to describe intonation patterns observable in an F0 contour (Taylor, 2000). An important goal is that the model should be capable of reconstructing F0 contours faithfully when appropriate parameters are given. However, to make it functional, a phonetic model should also be linguistically meaningful.

In fact, using certain functions, such as polynomial equations, to accurately represent F0 contour is not a difficult task. What is more challenging is developing a model whose parameters are predictable from available linguistic information. To be more specific, the mapping from various linguistic factors, which could affect intonation, to the model parameters, or vice versa, is more critical. Partly due to the availability of large corpora, more phonetic models, mainly for text-to-speech systems, have been proposed in recent years. Among the existing phonetic models, two lines of approaches will be distinguished: parametric vs. non-parametric.

In parametric models, the original F0 values are usually transformed into some parametric forms. And after the parameters are predicted, the intonation can be re-synthesized with these parameters.

2.2.1 The Fujisaki Model

Fujisaki and his colleagues (1983, 1988) developed an intonation model for Japanese, which has also been applied to other languages (e.g., German (Mixdorff, 2000)). The model additively superimposes a basic F0 value (Fmin), a *phrase component*, and an *accent component*, on a logarithmic scale. The control mechanisms of the two components are realized as critically damped second-order systems responding to impulse commands in the case of the phrase component, and rectangular commands in the case of the accent component. The Fujisaki model attempts to relate the functions to the human speech production apparatus. It is a typical superpositional approach in that it assumes different

intonation components to be superimposed on top of each other. This is different from the AM approach described above, which is predominantly sequential. One of the critical steps towards application of the Fujisaki model is the extraction of model parameters from F0 contours. Fully automatic extraction of model parameters has been developed (Mixdorff, 2000), which would facilitate researchers working with the model on large corpora.

2.2.2 INTSINT

INTSINT (INternational Transcription System for INTonation), proposed by Hirst, Di Cristo, and coworkers (e.g., Hirst et al., 2000), describes intonation with a limited set of abstract tonal symbols, which is designed such that the inventory of pitch patterns of a given language is not required.

The input to the INTSINT system is a series of target points, which is estimated by a low-level F0 modeling technique called MOMEL. The MOMEL algorithm aims to analyze and synthesize F0 curves automatically. There are four basic stages:

(1) preprocessing of F0

(2) estimation of target-candidates

(3) partition of candidates

(4) reduction of candidates

Briefly, to estimate target candidates, first a moving window is applied to a segment of F0 values. Then, within each window, a quadratic regression is performed. The sequence of target candidates is partitioned by using another moving window in which the average values of two half windows are compared. The partition boundaries are determined where

the values correspond to local minima and are greater than the overall average value. A final reduction procedure is taken to eliminate those outlying candidates by certain definitions.

In INTSINT, abstract symbols are defined to represent the target points derived from MOMEL procedure. They are T, M, B, H, S, L, U, D, which stand for Top, Mid, Bottom, Higher, Same, Lower, Upstepped, Downstepped respectively. Among these symbols, tones {T, M, B} are regarded as absolute tones, which refer to the speaker's overall pitch range; tones {H, S, L, U, D} are relative with respect to the value of the preceding target point. The relative tones are further distinguished between non-iterative {H, S, L} and iterative {U, D} tones. An automatic coding scheme was also proposed to relate the target points and the abstract symbols through a set of rules (Hirst et al., 2000). This system has been applied to a number of languages and shown to be quite robust.

INTSINT can be viewed as a hybrid phonetic/phonological model. It starts with a low-level phonetic analysis. Then a phonological description system is derived from the phonetic analysis results.

2.2.3 The Tilt Model

In the Tilt model (Taylor, 2000), the basic intonational unit is the so-called *intonation event.* Similar to Pierrehumbert's model, the intonation event contains pitch accent and boundary tone, but is described by several continuous *tilt* parameters. First, as shown in Figure 2.1, each event is parameterized by measuring the amplitudes and durations of the

rises and falls, denoted by A_{rise}, A_{fall}, D_{rise}, and D_{fall}, respectively. These parameters constitute the rise/fall/connection (RFC) model, which is the prequel of the Tilt model. Then the four parameters are converted into three Tilt parameters, namely *tilt*, A_{event}, and D_{event}.

The tilt parameter is calculated as follows:

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

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

$$tilt = \frac{1}{2} tilt_{amp} + \frac{1}{2} tilt_{dur} = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})}$$

$$A_{event} = |A_{rise}| + |A_{fall}|$$

$$D_{event} = |D_{rise}| + |D_{fall}|$$

Together with two additional ones, the five parameters of the Tilt model are:

- Event amplitude (A_{event}): The amplitude of an intonation event relative to starting F0 (in Hertz).
- Event duration (D_{event}): the duration of an intonation event
- *tilt*: a dimensionless parameter in the range of [-1, 1] describing the shape of an intonation event. -1 represents a pure fall, 1 is pure rise, and 0 indicates that the event contains equal portions of rise and fall.
- *Position*: the peak location of an intonation event, which is usually defined as the distance between the vowel starting time and the peak location.

- *Start F0*: the absolute F0 at the start of an intonation event

Taylor (2000) argues that the tilt parameters are linguistically meaningful and can be regarded as phonological entities. Taylor (2000) also shows that the Tilt model and AM/ToBI modes are similar in many respects. (1) Both models view the intonational representation of an utterance as a linear sequence of events based intonational entities rather than as a superpositional organization. These events are associated with syllables/segments in an autosegmental structure. (2) Both models agree the existence of an abstract phonological representation of intonation.

Given known locations of intonation events and intonational phrase boundaries, an automatic procedure for extracting tilt parameters has been developed (Taylor, 2000). Taylor (2000) also describes an intonation event detector using Hidden Markov Model (HMM) technique.

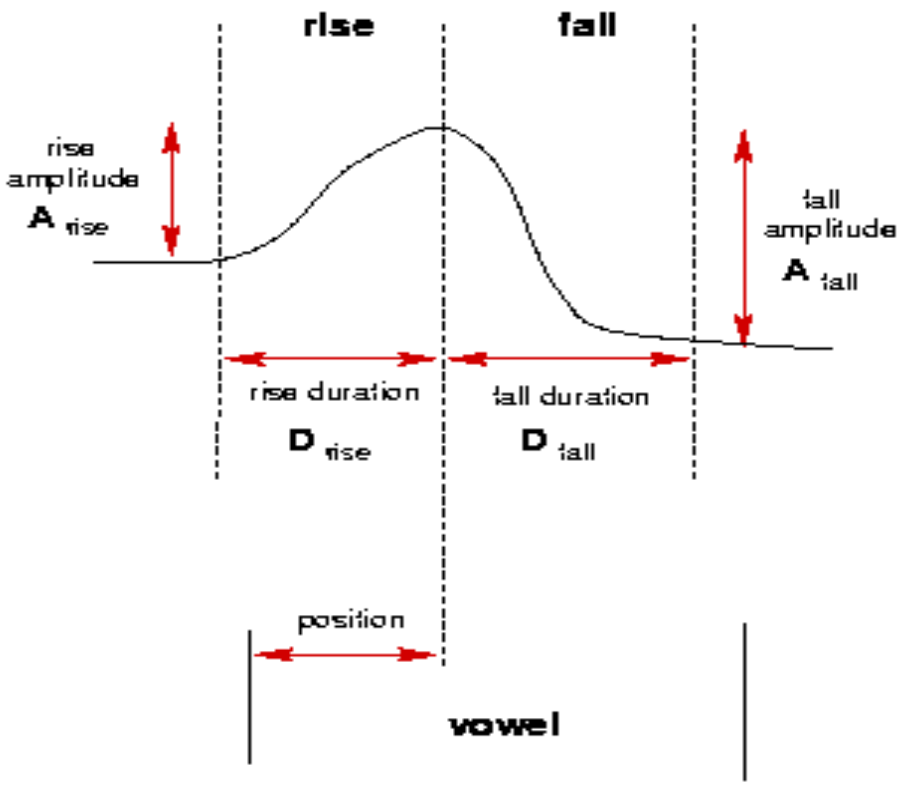


Figure 2.1: Tilt parameters for describing an intonation event.

2.2.4 PaIntE

Mohler and Conkie (1998) describe an intonation event using the sum of a rising and a falling sigmoid functions with a fixed time delay. The parametric intonation event (PaIntE) contains six parameters:

- Two parameters (*a1* and *a2*) represent the steepness of the rising and falling sigmoids, respectively.

- Parameter *b* stands for the alignment of the function. The syllable length is defined as unity.
- Two other parameters (*c1* and *c2*) model the amplitudes of the rising and falling sigmoids, respectively.
- Parameter *d* corresponds to the function's peak

This model emphasizes intonation events, which is similar to the Tilt model in this regard. Thus, the parameterization of F0 contours is not limited by syllable boundary but uses the accented syllable as an anchor point. The approximation is performed within a three-syllable window around the syllable carrying accent.

Mohler and Conkie (1998) further applied Vector Quantization (VQ) to the PaIntE parameters. They argue that (1) intonation can be described by a number of distinct shapes; (2) reducing data can improve machine learning performance; (3) VQ allows predicting all six parameters together rather than individually.

2.2.5 The Pitch Target Approximation Model

A pitch target approximation model for generating F0 contours in Mandarin Chinese was proposed by Xu and Wang (1997, 2001) and quantified in Xu et al. (1999). In this model, the surface F0 contour is viewed as the result of asymptotic approximation to an underlying pitch target, which can be a static target ([high] or [low]) or a dynamic target ([rise] or [fall]). A pitch target is defined as the smallest unit that is articulatorily operable. The host unit of a pitch target is assumed to be the syllable (for Mandarin, at least). This model is primarily based on evidence from Mandarin, where four basic pitch targets ([high], [low],

[rise] and [fall]) are associated with the lexical tones in Mandarin: H (High), L (Low), R (Rising) and F (Falling), respectively. Rather than simply simulating the surface F0 contours, this model aims at some deeper speech production mechanism, which is similar to the Fujisaki model in this regard. Importantly, it emphasizes the role of articulatory constraints in intonation modeling. Articulatory or physiological constraints have been paid much attention to lately by a number of intonation researchers (e.g. Kochanski and Shih, 2002; Xu and Sun, 2002). Xu and Sun (2002) conducted experiments to investigate the maximum speed of pitch change produced by English and Mandarin speakers and found that physiological limitation on the speed of pitch movement is greater than has been recognized. They further argue that the maximum speed of pitch change is often approached in speech, and that the role of physiological constraints in determining the shape and alignment of F0 contours in speech is probably greater than has been appreciated.

Xu et al. (1999) applied the pitch target approximation model to Mandarin and showed that it is able to reconstruct F0 contours accurately. This model forms a basis of the F0 generation system proposed in this thesis. More detailed discussion is deferred to later chapters.

2.3 Phonetic Models – Nonparametric Approach

Without doubt, F0 values themselves are good indicators of high-level linguistic information. Therefore, if the goal is not to define a set of abstract symbols to describe

intonation as in phonological models, but to generate intonation based on available linguistic information, one could just use original F0 contours as the targets. As can be seen above, the effort made by researchers in developing parametric models is to find a better representation of F0 contours. However, if inappropriate forms are used, the predictions can be significantly different from the original F0 contours. Instead of devoting effort to finding such a representation, some researchers use the original F0 contours directly or F0 with some trivial modifications as the output targets. Such systems are referred to as non-parametric models in this thesis.

In Black and Hunt (1996), three F0 target points were selected at the beginning, middle, and end of a syllable, respectively. For each point, they built a multivariate regression model. ToBI labels were incorporated in the feature set. In synthesis, three target points were predicted for each syllable. And the final F0 contours were interpolated from the three points and smoothed. This model is referred to as the Three-Target model hereafter.

In Traber (1992), a neural network with two hidden layers was trained to generate F0 for German. His system was also syllable based, as that in Black and Hunt (1996). Eight target F0 points were selected from each syllable. A syllable was first split into two segments at the point of maximum energy in the vowel. Then each segment was further split into two segments with equal length. A linear regression was performed for each of the four segments and the two end points for each line were saved for predicting targets. These eight F0 points were used as the output of a neural network.

Based on Traber (1992), Buhmann et al. (2000) built an intonation generation system for six languages. In their work, a recurrent neural network (RNN) of the Elman type (Elman, 1990) was trained with backpropagation-through-time algorithm. The output targets of the network were five equidistant F0 points selected from each syllable. These points were transformed into F0 contours by interpolation. The evaluation results indicated the system works well for all six languages.

Ross and Ostendorf (1999) developed a dynamical system model to predict F0 values in a syllable. The model generates normalized F0 contours, whereas the discourse-level F0 range is predicted by regression trees. Linguistic knowledge was applied in selecting proper model parameter dependencies. Various linguistic features at three levels, including phoneme level, syllable level, and phrase level, were considered. Among the input features, ToBI-related information constituted an important part. All the F0 points were modeled, and the system parameters were estimated via expectation-maximization (EM) algorithm.

2.4 Comparisons Between Different Models

Phonological models use categorical symbols to describe intonation, which can be seen as a representation at the cognitive level. A significant difference between the two phonological models described above – the AM approach and the IPO approach is their view on the primitive intonational unit in this cognitive structure of intonation. In AM approach, level tone (High or Low) is the basic unit while the IPO approach favors pitch movement. When

generating F0 for speech synthesis using the AM approach, phonetic realization rules are used to map phonological entities to F0 values (e.g. Pierrehumbert, 1981).

Contrary to phonological models, phonetics models use continuous parameters to describe intonation, which constitute a distinct level between phonological representation of intonation and F0 curves (Hirst et al., 2000). Thus, mapping from phonology and other high-level representation (e.g. semantic and syntactic representation) onto acoustics become a two-step process. First, phonetic realization rules are used to map phonological entities or other linguistic features to the parameters of the phonetic models. Then the parameters are converted into F0 values according to the formulation of the model. For the first step, similar challenge is faced as that in phonological models described above. The second step is usually easy to implement since such a conversion is fixed and does not vary across conditions. A phonetic model can evolve into a phonological model by (1) categorizing the continuous parameters as in the INTSINT model (Hirst et al., 2000) or (2) finding direct links between phonological entities and model parameters as in the Tilt model (Taylor, 2000). A phonetic model can also find connections between its model parameters and some physiological entities as in the Fujisaki model and the pitch target approximation model (Xu and Wang, 2001). These models attempt to explain intonation phenomena using some universal mechanisms at the physiological level. Such a motivation is attractive as it can make the model more language independent and linguistically more interpretable.

Among phonetic models, non-parametric approaches, such as the Three-Target model (Black and Hunt, 1996), are often more language independent and easier to implement. Although seemingly simple, non-parametric models can frequently achieve very competitive results compared with parametric models. This is because F0 values themselves are meaningful linguistic entities and thus predictable from linguistic features. However, even though input features at different levels are used, non-parametric models often lack of a clear internal hierarchy. This could be a problem since it is known that various components in intonation are caused by factors at different levels and have different perceptual weights. In addition, in non-parametric models the smoothness of the predicted F0 curve is not treated inherently and must be realized through some post-processing techniques. By contrast, parametric models, while placing more constraints on intonation, can selectively model intonation components based on theoretical assumptions or input features. The smoothness of F0 contours is controlled by choosing appropriate parameters and function forms. However, transforming F0 values into parametric representations has the risk of losing linguistic meaning and in turn predictive power. Also, certain parametric forms, such as polynomial functions, Fourier series, etc., may approximate intonation very accurately, but the mapping between their parametric space and linguistic features may be too complex for a statistical algorithm to learn successfully.

2.5 F0 Generation Methods

Once an intonation model is developed to describe F0 contours, an appropriate way is needed to predict model parameters and generate F0 contours from input linguistic information. This constitutes another crucial part of the whole intonation modeling process. In the past, many methods have been proposed, including rule-based and data-driven approaches.

Rule-based approaches require sets of heuristic rules written by linguistic experts. In contrast, data-driven approaches derive rules or some relations automatically by machine from a corpus. They are less dependent on heuristic rule writing. When appropriately trained, such systems can generate quite natural intonation with rich variation patterns since statistical algorithms can learn very complex relationships between input features and output targets (e.g. F0 or model parameters). Manually exploring these quantitative interactions is too laborious and even impossible when a large amount of factors are involved.

An automatically trained system can be less language dependent and can be adapted to different speaking styles or speakers. However, data-driven systems critically rely on prosodically labeled corpora, which could be very labor-expensive to build. This hinders the development of data-driven systems using very large databases, while more training data are normally very beneficial to capture more F0 variation patterns and thus synthesize

more natural intonation. Training the model on different speech styles or different speakers could also be limited if proper databases are not available.

Finally, a data-driven approach requires an appropriate machine learning algorithm, and selecting such an algorithm is a nontrivial task. It should be noted that statistically trained models seem to be less robust than rule-based systems. Although they can produce intonation with more variations, sometimes very bad F0 contours can also be generated. To overcome this, hybrid systems that combine data-driven and rule-based approaches have been investigated (e.g. Meron, 2002).

2.5.1 Linguistic Rules

With existing linguistic knowledge, it is natural to define a set of rules to generate F0 contours. Pierrehumbert (1981) developed a rule system to generate English neutral declarative intonation primarily based on her theory presented in Pierrehumbert (1980). In her system, two categories of target values, high and low, are defined with respect to the current F0 range. F0 contours are generated by connecting these target points. When two high targets are sufficiently farther apart, a sagging contour is used to do the connection, which is realized through a quadratic function. In other cases, target points are connected via monotonic curves.

Jilka et al. (1999) define complex rules for generating ToBI-based American English intonation. The system contains the following five components:

(1) ToBI labels are associated with specific target values. All the target positions are defined relative to the pitch range and the voiced part of the labeled syllable, which can roughly be viewed as a way to describe F0 in a two dimensional plane.

(2) For pitch accents (e.g. H*), target points' position in pitch range is first defined, which is represented in percentage between topline and baseline. For position in the voiced part of the syllable, four cases are considered: first syllable of ip (intermediate phrase), last syllable of ip, one syllable ip, and normal case.

(3) For phrase tones of both intermediate and intonational phrases, different treatments were developed for phrase initial and phrase final tones, respectively. The corresponding target points are also described in terms of position in pitch range and in the voiced part of a syllable.

(4) Special distance rules are defined when target points that are opposed in pitch are more than three syllables apart. Generally new target points are added between the opposing target points.

(5) Valleys usually appear between two high target points. Depending on the distance between the two points, 0-2 target points are added in between. The valleys are deliberately set to be not very deep.

After they are all in place, the target points are connected via linear interpolation. Note that, for TTS applications, this system is not a complete F0 generation system in the sense that ToBI style pitch accents and phrase accents are used as input, which would need to be

predicted by separate modules in a real system. The authors tested this rule system on a portion of Boston Radio Corpus (Ostendorf et al., 1995) and conducted both objective and subjective evaluations. For objective evaluation, they computed root-mean-square-error (*RMSE*) and correlation coefficient (*r*) and get 32.4 Hz and 0.605, respectively. Perception tests yielded satisfactory results.

Rule-based systems can be very efficient and usually can produce consistent intonation contours due to explicit constraints posed by the designers. However, constructing a set of sophisticated rules is a tremendously challenging task. Many rule-based systems tend to produce intonation lacking in rich variations, which can be attributed to the insufficiency of the rule set. Nevertheless, rule-based systems are direct applications of intonation research findings, which can verify various linguistic theories and research results. This in turn can strengthen our understanding of the human speech communication process. Along with better understanding of intonation, more sophisticated rules can be developed in the future.

2.5.2 Template Approach

There are several studies adopting the template approach (e.g. Huang et al., 1996, 1997; Meron, 2002; Tao and Cai, 2002). The following discussion will use Microsoft's whistler (Huang et al., 1996, 1997) TTS system as an example. In this system, *Clause* contains several syllables; *Clause Specification Vector S* is an N-dimensional vector of the tones and other context features for each syllable in the clause; *Clause Pitch Vector P* is an N-

dimensional vector of the F0 values for each syllables. Thus, the prosody pattern database contains a very large set of *S:P* pairs. To generate F0 contours, firstly, an input sentence is transformed into a clause specification vector *S(i)* by the text analysis module. Then a most similar vector S in the database to the input clause specification vector *S(i)* is selected. The corresponding prosody vector *P* is used to generate the pitch contour anchor points. The selection of vector *S* is realized through a dynamic programming algorithm to minimize a cost function.

2.5.3 Neural Networks

Neural networks are very popular machine learning algorithms utilized in many domains (Bishop, 1995; Haykin, 1999). They simulate the cognitive process of human brain by employing parallel computing. It has been proved that with enough training data and appropriate structure, a neural network is capable of approximating any nonlinear function. One of the attractive features of neural networks is that they can learn temporal structures by employing some particular topology, such as time-delay window, recurrent connection, etc. Neural networks have been applied to prosody modeling extensively (e.g. Buhmann et al., 2000; Chen et al., 1998; Sun, 2001; Traber, 1992). In this line of approach, a neural network is used to learn the mapping between the input linguistic/acoustic features and output targets, which are usually F0 values or some transformed parameters. With appropriate structures, a neural network can capture the complex relationship between linguistic features and intonation. However, the trained model is not human readable,

which is undesirable if one wants to gain more theoretical insight into the issue. Nevertheless, for many practical applications, since the goal is merely to generate natural intonation, the ease of design and good learning properties of neural network make it a reasonable choice.

2.5.4 Decision Trees

Decision trees are widespread machine learning algorithms that have been applied to prosody modeling extensively, including F0 generation (e.g. Dusterhoff et al., 1999; Sun, 2002b), duration prediction (Ripley, 1992), phrase prediction (Wang and Hirschberg, 1992; Sun and Applebaum, 2001), and pitch accent prediction (Hirschberg, 1993; Sun, 2002c). With decision trees, Dusterhoff et al. (1999) achieved relatively good results within the Tilt model framework. They first built a classification tree to predict intonation event location, which, however, was not integrated into the whole system in the evaluation. Then five regression trees were built for each of the tilt parameters, namely, start F0, amplitude, duration, tilt, and peak position. Their numerical results for Boston Radio Corpus, Speaker F2b are *RMSE* = 34.3; $r = 0.6$.

Using a tree to represent data structure can be found in almost every Data Structure textbook. In a decision tree, a series of questions is often asked, which could be binary or non-binary. Based on the answer, an appropriate descendent path will be traversed along, which can be another tree or a terminal (leaf) node. For example, one wants to know

whether the syllable / n ey /[1] in the word "intonation" bears a pitch accent in a particular sentence. Available information includes lexical stress, number of syllable in the word, part-of-speech, etc. A simple binary tree can be constructed as follows (Figure 2.2):

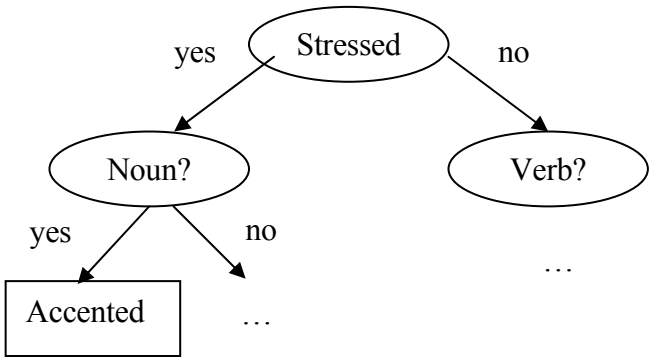


Figure 2.2: A sample decision tree.

The first node is usually called root node. Between the nodes are links or branches. The terminal node is the leaf node.

The challenging part is how to construct such trees. In rule-based systems, a set of human-crafted rules are defined based on existing knowledge. On the other hand, in a data-driven system, rules are automatic derived from existing data. There are many decision tree algorithms. Classification and Regression Trees (CART) (Breiman et al., 1984), is one of the most widely used, which will be briefly described below.

[1] ARPAbet transcription

Given some input features (predictor), one wants to predict the values of certain variables (predictee) by building a decision tree. If the predicted variable is categorical, CART constructs a classification tree, in which the leaf node contains the most likely category along its corresponding path or a probability distribution of each category. On the other hand, if the variable is continuous, a regression tree is constructed, in which the leaf node contains the mean and standard deviation, similar to multi-linear regression.

The core idea of CART tree construction is to find an appropriate question about some input feature that makes the best split of the data. The "best" is in the sense that after the split, the two partitions have the lowest mean "impurity" value. The "impurity" is some mathematical measure that can represent how similar the samples are to each other in a group. A group containing more similar samples has lower "impurity" value. After a split is made, the algorithm continues to make new partitions recursively by finding a question that minimizes the impurity value. The recursion stops when some criteria are met, such as the minimum number of samples in a partition. It can be seen that CART is a *greedy algorithm* in that at each stage it finds the locally best question that makes a split. Once the partitions are determined, the algorithm continues the same process and never looks back to make changes. This is suboptimal, but a global optimization would be very computationally expensive. Fortunately, in practice it seems this suboptimality doesn't pose serious problems in most cases.

Decision tree algorithms have several attractive features (Huang et al., 2001; Sun, 2002b):

- The algorithm is applicable to any data structure.
- The binary tree structure allows for compact storage, efficient prediction.
- The decision tree contains interpretable questions and structure that are helpful for better understanding of the problem.
- It requires less hand-tuning of the parameters than algorithms such as neural networks.
- It can handle missing data and also is robust to outliers and mislabeled data samples.
- The derived tree model can be integrated into existing systems very easily.

2.5.5 Ensemble Learning

Ensemble learning has received much attention lately in the machine learning community and has been shown to be superior to single-classifier systems in many real world problems (Dietterich, 2002). However, to the author's knowledge, applications of ensemble learning to prosodic modeling problems have been scarce, with the exception of the study by Lee and Oh (2001), in which ensemble decision trees were applied to F0 generation for Korean. In this thesis, ensemble learning is applied to several prosodic modeling problems. The details of ensemble learning are described below.

For classification problems, ensemble learning algorithms construct a set of classifiers and then classify new data by taking a (weighted) vote of their predictions. For regression problems, a (weighted) mean or median is often used. Dietterich (2002) classifies various ensembles construction methods into two categories: (1) Methods for

Independently Constructing Ensembles; (2) Methods for Coordinated Construction of Ensembles.

Methods for Independently Constructing Ensembles

There are several methods falling into this category. In one approach, one can manipulate the training examples. A well-known algorithm, Bagging (Bootstrap Aggregation) (Breiman, 1996), generates multiple classifiers by following this route. Each time a different training set is presented to the learning machine. The new training set is constructed by drawing samples from the original training set randomly with replacement. The final results are obtained usually by voting for classification or taking averages for regression. For bagging to be successful, the learning machine should be *unstable*, that is, a small change in the training set would result in large changes in the training output. Decision tree and neural network are typical unstable learners.

Instead of presenting different training data to each base learner, the input feature set can be split into different subsets to train multiple learning machines. In a study to identify volcanoes on Venus, Cherkauer (1996) trained 32 neural networks upon different feature subsets and achieved performance comparable to human experts. Finally, another method is to manipulate the output targets. In Bakiri and Dietterich (2001), an error-code correction scheme applying to output labels was developed for letter-to-sound conversion.

Methods for Coordinated Construction of Ensembles

Boosting, specifically AdaBoost (Freund and Schapire, 1997), also combines multiple classifiers by presenting different training set to the base learner. However, instead

of using random selection as in bagging, the construction of a new training set depends on a weight distribution, which is updated over iterations. Initially all the training samples have the same weight. After each iteration, the weight distribution is updated such that misclassified samples have more weight. With the updated weight distribution, there are two ways of generating new training samples. In *reweighting*, the original training set is used, but each sample is associated with a new weight. This method is applicable to the learners that can handle weighted samples. In *resampling*, the new training set is constructed according to the weight distribution, where samples with more weight are more likely to be selected. Note that although it might be suboptimal, resampling is used in this work mainly because the base learner - CART does not handle weighted samples. Finally, to classify a new sample, a weighted combination of multiple classifiers is used. The mathematical descriptions of two typical AdaBoost algorithms are included in the Appendix A, where Figure A.1 illustrates the AdaBoost.M1 algorithm for multi-class problems (Freund and Schapire, 1997), and Figure A.2 illustrates a boosting algorithm for regression problems described by Drucker (1997).

Extensive theoretical studies have been conducted to explore why ensemble learning works so well. A common explanation is the bias-variance decomposition (Breiman, 1996; Dietterich, 2002; Haykin, 1999). It has been shown that the prediction error of a classifier can be decomposed into two components: bias and variance. A low bias and high variance classifier is characterized by low training error and poor generalization ability. On the other hand, if a classifier has higher training error, it is regarded as highly

biased. Ensemble methods like bagging can reduce the amount of variance. Boosting can reduce both bias and variance. An individual decision tree can achieve very high accuracy on the training set, but tend to give poor generalization on the testing set. Thus, applying ensemble learning on decision trees can effectively improve performance by lowering variance. A more detailed discussion on this topic can be found elsewhere, such as Dietterich (2002).

2.6 Summary

In this chapter, several existing intonation models are briefly reviewed. Models are presented as either phonological or phonetic. The Autosegmental-Metrical and the IPO approaches are described as examples of phonological systems. For phonetic models, parametric and non-parametric approaches are further distinguished. Example parametric models include the Fujisaki model, INTSINT, Tilt model, PaIntE model, and Pitch Target Approximation Model. Several typical nonparametric systems are also discussed, which in general select some F0 target points from each syllable and predict their values via machine learning algorithms, such as neural network and dynamical system. The review presented here is far from complete. For more comprehensive treatment, readers are referred to other sources (e.g., Botinis et al., 2001; Ladd, 1996).

This chapter also discusses several practical techniques used in F0 generation for speech synthesis, including linguistic rules, template approach, neural network, decision trees, and ensemble learning. Machine learning and pattern recognition are rapidly growing

areas, and new algorithms emerge quickly. For instance, recent work on support vector machine and belief network have led to many successful examples. It is believed that intonation modeling will benefit enormously from machine learning research. It should be noted that selecting an algorithm for intonation modeling is not hard; rather, finding a suitable algorithm and adapting it to the target domain is more challenging.

Chapter 3

Pitch Determination Algorithm Towards Intonation Modeling

Accurate pitch tracking is critical to various speech processing tasks. This is especially true for intonation related research, where F0 values themselves are the research target. This chapter thus aims to tackle the "intonation contour" issues addressed in Chapter 1 by proposing a new pitch determination algorithm (PDA) towards intonation modeling (Sun, 2002a).

Pitch determination is one of the most difficult problems in speech analysis. Automatic pitch determination algorithms make various errors, such as pitch doubling and pitch halving. One of the reasons for the errors is the occurrence of alternate cycles of amplitude and/or period in speech signals. For normal speech, alternate cycles usually appear in creaky voice or voice with laryngealization associated with low F0, which are often characterized as perceptually rough voices. In pathological voice, alternate cycles can be found even in normal mode of production. Traditional methods for dealing with this issue of alternate cycles include both automatic and manual procedures. In automatic procedures, most of the current algorithms rely on fine-tuning threshold parameters based on particular databases, or post-processing techniques such as linear/nonlinear smoothing, or dynamic programming. On the other hand, manual correction relies on visual inspection

of the waveform, pitch contours or vocal pulses with or without the aid of EGG signals. Since alternate cycles constitute inherently ambiguous vibration patterns and perceived pitch, the above corrective methods often fail to yield reasonable F0 values. Manual correction can be highly effective with the aid of EGG signals, but when the database is large, as in the case of many data-driven modeling work, manual correction is often time consuming and labor expensive. Therefore, automatic methods are usually preferred, provided that they are reliable.

3.1 Background

3.1.1 Perceived Pitch of Alternate Cycles

According to Titze (1994, 1995), alternate cycles of amplitude and/or period in speech waveform primarily reflect the vibratory patterns of the vocal folds. These amplitude and period alternations can be viewed as the result of amplitude modulation (AM) and frequency modulation (FM), respectively. That is, the slowly varying component modulates the faster component. In a simple case, the ratio between the two components can be one half. The low frequency component is often called subharmonic, which can be any integer fraction of the fundamental frequency (e.g. 1/2, 1/3, 1/4, …, 1/n) (Titze, 1994). Titze (1994) suggests that subharmonic generation can occur when there is left-right asymmetry in the mechanical or geometric properties of the vocal folds. Svec et al. (1996)

offer an alternative explanation - the subharmonic vibratory pattern of vocal folds could result from a combination of two vibration modes whose frequency ratio is 3:2.

The amount of amplitude modulation can be defined as a percentage in the following form (Titze, 1995):

$$M = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} 100 \tag{3.1}$$

where A_i and A_{i+1} are the amplitudes of consecutive pulses (see Figure 3.1(a)), and M is the modulation index which can vary from 0 to 100 percent. Similarly, frequency modulation can be defined as a percantage in the following form (Titze, 1995):

$$M = \frac{T_i - T_{i+1}}{T_i + T_{i+1}} 100 \tag{3.2}$$

where T_i and T_{i+1} are the periods of consecutive pulses (see Figure 3.1(b)).

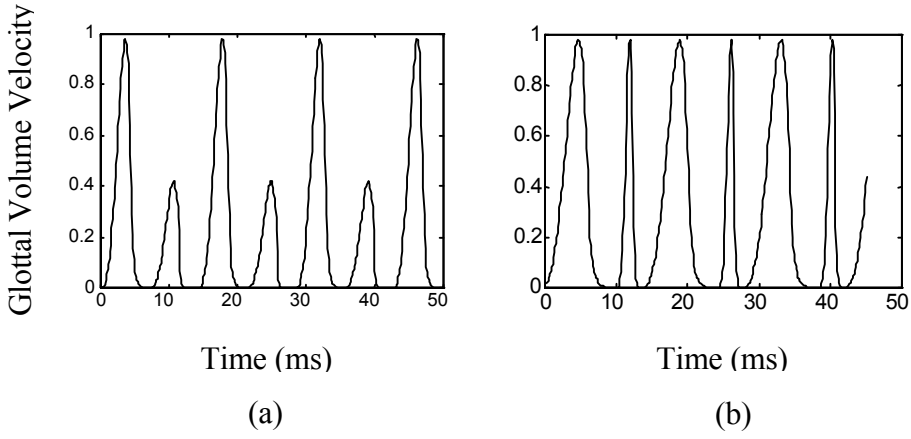


Figure 3.1: Schematic representations of glottal pulses with alternate cycles. (a) amplitude alternation (b) period alternation.

In the frequency domain, the representation of modulation is the appearance of subharmonic components between the harmonics. Figure 3.2 shows an example, which contains spectra of synthetic vowel /a/ with different amounts of amplitude modulation.

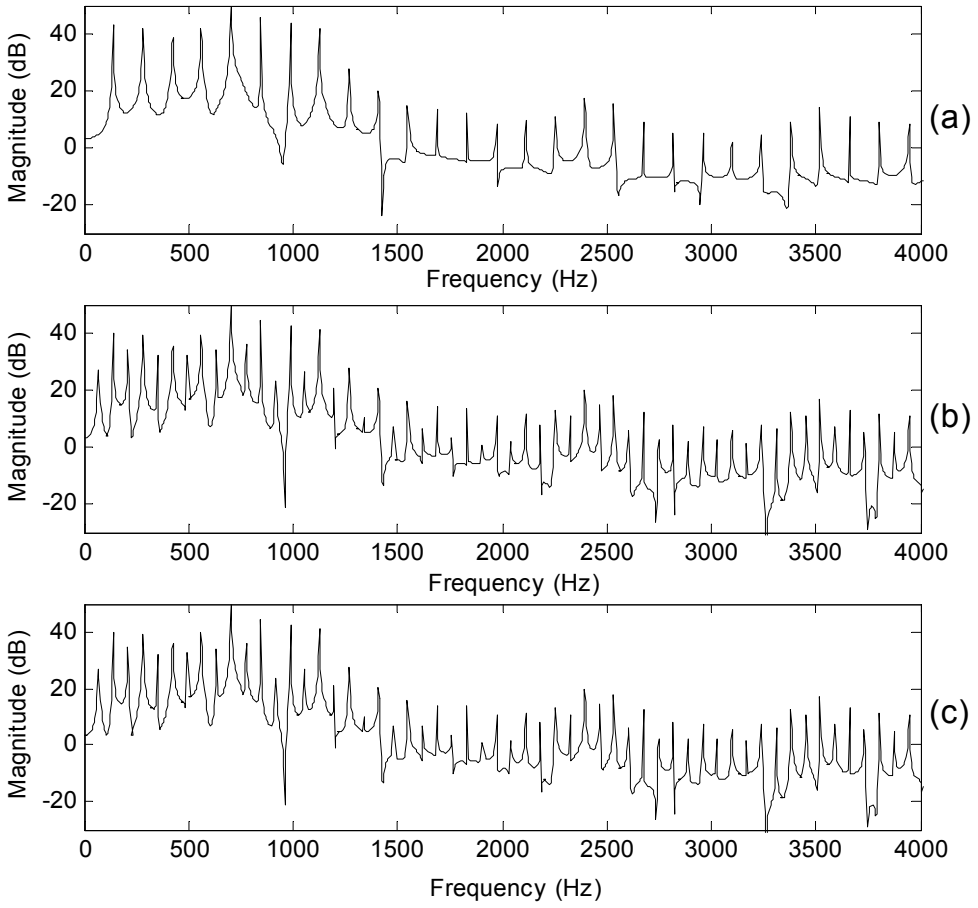


Figure 3.2: Spectrum of synthetic vowel /a/. (a) without modulation (b) amplitude modulation on the glottal source signal with index *M*=50% (c) amplitude modulation on the glottal source signal with index *M*=90%.

Physically determining the fundamental frequency of alternate cycles is often difficult due to the ambiguous vibration patterns resulted from subharmonic frequency. In speech perception, however, pitch must have been perceived. Otherwise, perception of intonation patterns involving alternate cycles would be impossible. In such a case, human perception may help. It could be the case that even the perceived pitch is also somewhat ambiguous. But at least the circumstance under which it is ambiguous can be experimentally determined.

Recall that the manifestation of amplitude modulation and/or frequency modulation in the frequency domain is the presence of subharmonics. This means that one parameter can be used to describe both amplitude modulation and frequency modulation. When the amplitude of the subharmonics is low, or more exactly, when the amplitude ratio between the subharmonics and harmonics is low, the subharmonics have no effect on pitch perception. When the amplitude ratio is sufficiently high, pitch may be perceived as one octave lower. To formally describe this relationship, a new parameter named Subharmonic-to-Harmonic Ratio (SHR) is proposed to describe the amplitude ratio between subharmonics and harmonics (Sun and Xu, 2002).

Sun and Xu (2002) conducted a pitch perception study in an attempt to quantify the relationship between perceived pitch and SHR. In their experiment, vowels with alternate cycles were synthesized through amplitude and frequency modulation, which generates subharmonics with lowest frequency of 0.5F0. Listeners were asked to judge the pitch these synthesized vowels. The result shows that pitch perception is closely related to SHR (i.e.,

the amplitude ratio between subharmonics and harmonics). Generally, when the ratio was smaller than 0.2, the subharmonics did not have an effect on pitch perception. As the ratio increased approximately to above 0.4, the pitch was mostly perceived as one octave lower, which corresponded to the lowest subharmonic frequency. When SHR was between 0.2 and 0.4, the perceived pitch was ambiguous. These findings suggest that pitch can be determined by computing SHR and comparing it with the pitch perception data.

Beside pitch determination, SHR can be also used as a parameter for describing voice quality. It is known that different degrees of "irregularity" can elicit different degrees of "roughness" sensation (Titze, 1994). Therefore, it is desirable to have an objective measure to quantify this relationship, which can be used as an index to classify voice production mode for a particular speaker or compare voice quality for different speakers. Since SHR can describe alternate cycles quantitatively, it can be used as a predictor of the perceptual quality of voice.

3.1.2 Related Pitch Determination Algorithms

The procedure for computing SHR falls in the general category of spectrum compression techniques. Since finding the lowest harmonic directly in the spectrum has shown to be unreliable for pitch determination (Hess, 1991), researchers usually try to estimate the frequency of the lowest harmonic indirectly by taking advantage of the harmonic structure. A group of PDAs based on spectrum compression have been developed, in which the spectrum is compressed along the frequency axis at different ratios and the compressed spectra are added together to make the F0 peak more prominent (e.g. Hermes, 1988;

Schroeder, 1968). Such an approach significantly improves the algorithm's reliability in locating the correct peak. However, wrong peaks can be selected when subharmonics are present since plain summation of harmonics does not provide a mechanism to control for the effect of subharmonics. In the current algorithm, instead of looking for one single peak that actually represents the summation of the harmonics and subharmonics, the effects of the harmonics and subharmonics are decomposed. Then subharmonics are examined to determine whether they are strong enough to be regarded as pitch candidates.

3.2 The Algorithm

For each short-term signal, let $A(f)$ represent the amplitude spectrum, and f_0, and f_{max} be the fundamental frequency and the maximum frequency of $A(f)$, respectively. The *sum of harmonic amplitude* (*SH*) is therefore defined as:

$$SH = \sum_{n=1}^{N} A(nf_0) \tag{3.3}$$

where N is the maximum number of harmonics contained in the spectrum, and $A(f) = 0$ if $f > f_{max}$. If we confine the pitch search range to $[F0_{min}\ F0_{max}]$, then $N = \text{floor}(f_{max} / F0_{min})$. In practice, only part of the spectrum is used to reduce computational cost. Following Hermes (1988), we set f_{max} =1250 Hz.

Assuming the lowest subharmonic frequency is one half of f_0, the *sum of subharmonic amplitude* (*SS*) is defined as:

$$SS = \sum_{n=1}^{N} A((n-1/2)f_0) \tag{3.4}$$

Note that the current algorithm is extendable to other subharmonic frequencies, such as 1/3F0 and 1/4F0. Consequently, the *subharmonic-to-harmonic ratio* (*SHR*) can be obtained by dividing *SS* with *SH*:

$$SHR = \frac{SS}{SH} \tag{3.5}$$

In practice, however, estimating *SHR* directly using Eq. (3.5) is not trivial. Thus an alternative way is proposed. First, the linear frequency scale is transformed into a logarithmic scale as described in Hermes (1988). Let *LOGA*(•) denote the spectrum with log frequency scale, then *SH* and *SS* can be represented as

$$SH = \sum_{n=1}^{N} LOGA(\log(nf_0)) = \sum_{n=1}^{N} LOGA(\log(n) + \log(f_0)) \tag{3.6}$$

$$SS = \sum_{n=1}^{N} LOGA(\log(n-1/2) + \log(f_0)) \tag{3.7}$$

To obtain *SH*, the spectrum is shifted leftward along the logarithmic frequency abscissa at even orders, i.e., log(2), log(4),… log(4*N*). These shifted spectra are added together and denoted by

$$SUMA(\log f)_{even} = \sum_{n=1}^{2N} LOGA(\log f + \log(2n)) \tag{3.8}$$

Since $LOGA(\log f) = 0$ when $f > f_{max}$, from Eqs. (3.6)–(3.8) it follows:

$$SUMA(\log(1/2\,f_0))_{even} = SH \tag{3.9}$$

$$SUMA(\log(1/4\,f_0))_{even} = SH + SS \tag{3.10}$$

Similarly, the spectrum is shifted leftward at log(1), log(3), log(5), … log(4*N*-1)

$$SUMA(\log f)_{odd} = \sum_{n=1}^{2N} LOGA(\log f + \log(2n-1)) \tag{3.11}$$

$$SUMA(\log(1/2\,f_0))_{odd} = SS \tag{3.12}$$

$$SUMA(\log(1/4\,f_0))_{odd} = \Delta \tag{3.13}$$

where Δ represents the sum of the values at $\log(nf_0) \pm \log(1/4\,f_0)$.

Next, the difference between $SUMA_{even}$ and $SUMA_{odd}$ is denoted by

$$DA(\log f) = SUMA(\log f)_{even} - SUMA(\log f)_{odd} \tag{3.14}$$

From Eqs. (3.9)(3.10)(3.12)(3.13), it can be derived

$$DA(\log(1/2\,f_0)) = SH - SS \tag{3.15}$$

$$DA(\log(1/4\,f_0)) = SH + SS - \Delta \tag{3.16}$$

Figure 3.3 gives an example for each function defined above, i.e., *LOGA,* $SUMA_{even}$, $SUMA_{odd}$, *DA*.

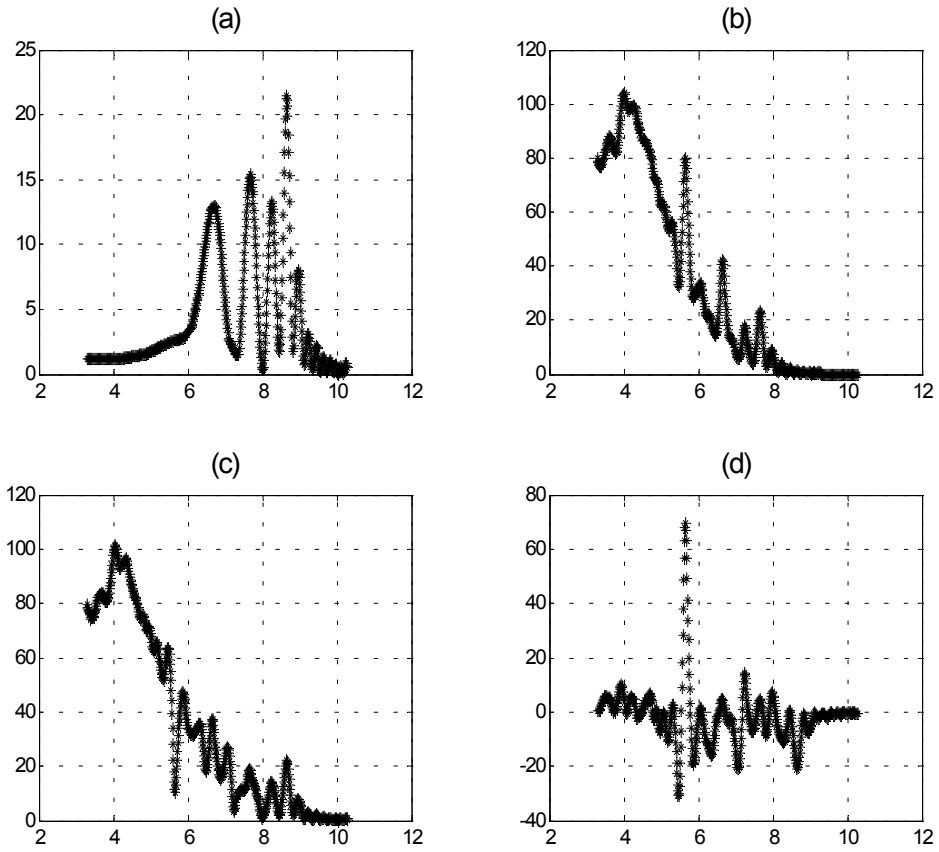


Figure 3.3: Schematic representations of four functions for calculating *SHR*. (a) *LOGA* (b) $SUMA_{even}$ (c) $SUMA_{odd}$ (d) *DA*.

In "regular" speech where *SS* ≈ 0, the maximum value of *DA*(•) would be at $\log(1/2\,f_0)$. On the other hand, if the magnitude of subharmonics becomes substantial, the maximum value of *DA*(•) would be most likely at $\log(1/4\,f_0)$ as Δ≈0 and the second maximum value is at $\log(1/2\,f_0)$. *SHR* can therefore be calculated approximately using Eqs. (3.15)(3.16):

$$\frac{DA(\log(1/4\,f_0)) - DA(\log(1/2\,f_0))}{DA(\log(1/4\,f_0)) + DA(\log(1/2\,f_0))} = \frac{SS - 1/2\,\Delta}{SH - 1/2\,\Delta} \approx SHR \tag{3.17}$$

In searching for the maximum value, the position of the global maximum is located first and denoted as $\log(f_1)$. Then, starting from that point, the position of the next local maximum denoted as $\log(f_2)$ is selected in the range of $[\log(1.9375f_1, \log(2.0625f_1]$. However, some special cases need to be treated differently:

- If $DA(\log(f)) \leq 0$ for all f, the frame is regarded as unvoiced.
- If $1.9375f_1 > F0_{max}$, only f_1 is returned.
- If $DA(\log(f_1)) > 0$ and $DA(\log(f_2)) \leq 0$, only f_1 is returned.

After the two peaks are located, *SHR* can be easily derived following Eq. (3.17):

$$SHR = \frac{DA(\log f_1) - DA(\log f_2)}{DA(\log f_1) + DA(\log f_2)} \tag{3.18}$$

If *SHR* is less than a certain threshold value, it indicates that subharmonics are weak in amplitude and we should favor the harmonics. Thus, f_2 is selected and the final pitch value is $2f_2$. Otherwise, f_1 is selected and the pitch is $2f_1$. Based on the pitch perception results (Sun and Xu, 2002) mentioned in the Background section, 0.2 is selected as the threshold value, although other values in [0.2, 0.4] would also work. For specific tasks, threshold values outside of this range can also be used. For example, using higher threshold values will favor the harmonics in the algorithm, even though the pitch of this frame would be represented by the subharmonics when heard in isolation. This is usually desirable in intonation modeling for speech synthesis where a globally smooth contour is preferred.

To detect voicing status, the noise floor is estimated, which is defined to be three times the energy of the first frame. If the energy of a frame is higher than the noise floor, it is passed into the pitch determination module. Using the estimated F0 value, the fundamental period is derived and two periods of signal are selected from this frame: one before the middle point and one after it. If the correlation between the two segments is higher than 0.2 and the zero-crossing rate of either segment is higher than 3500 Hz, the frame is classified as voiced.

3.3 Evaluation

3.3.1 Databases

Two databases were used for the evaluation. The first is the CSTR database[1], which contains five minutes of speech from one male and one female speaker. The speech signal is sampled at 20 KHz with 16-bit resolution. The reference pitch values are provided which are estimated from simultaneously recorded EGG signals. The other database used is the Keele pitch extraction reference database[2]. This database contains speech from ten speakers (five males and five females). The sampling rate is 20 KHz and the resolution is 16 bits. The reference pitch values provided by the Keele database are determined by an autocorrelation method at 10 ms frame rate.

[1] Available at *http://www.cstr.ed.ac.uk/~pcb/*
[2] Available at *ftp://ftp.cs.keele.ac.uk/pub/pitch/*

3.3.2 Results

To solely evaluate pitch determination, the voicing detection and post-processing modules were disabled as much as possible for all algorithms. This is because most algorithms make voicing estimation errors, such as misclassifying many "difficult" voiced segments as unvoiced, which could result in a seemingly better pitch determination. Specifically, the present algorithm - Subharmonic-to-Harmonic Ratio based pitch determination (SHRP) was compared with the following:

eSRPD: Enhanced super resolution pitch determinator has shown to be superior to seven other algorithms in (Bagshaw, 1994).

PDA: "PDA" program is included in Edinburgh Speech Tool Library <*http://www.cstr.ed.ac.uk*>, which is described as an implementation of super resolution pitch determinator (SRPD) (Medan et al., 1991). An example command is "pda m1nw0000.wav -otype ascii -o cstr_f0/m1nw0000.f0 -fmin 50 -fmax 550 -shift 0.010 -length 0.040 –L"

GET_F0: "GET_F0" is included in the ESPS package, which is an implementation of RAPT algorithm (Talkin, 1995). We use the default setting of the algorithm specified in the ESPS document except that we set "voice_bias=1" to encourage the algorithm to make more voicing hypotheses.

PRAAT: Praat software <*http://www.praat.org*> includes several PDAs, and we use the one described in (Boersma, 1993) which usually gives very good results according to our experience. An example command is "To Pitch (ac)... 0.01 50 15 1 0 0 0.01 0.35 0 550".

Note that even though an effort was made to make the algorithms yield an F0 value for each frame, in some cases voicing detection could not be removed completely. For program "PDA", a switch to disable voicing diction is provided. As a result, the default

setting with minor adjustments was used, which includes a full voicing determination module. For “GET_F0” and “PRAAT” a smaller number of frames were classified as unvoiced. To be safe, for these three algorithms analyses were only performed for the frames classified as voiced in both reference and the estimated data. This may make their results better than in reality. For the present algorithm, a pitch value for each frame was computed and only those classified as unvoiced by the reference data were discarded.

A commonly used criterion for evaluating pitch tracking performance is *gross error rate* (GER) (Bagshaw, 1994). A gross error is identified when the estimated F0 value is 20% (or 10%) higher or lower than the reference F0 value. For the CSTR database, the same evaluation settings were used as described in Bagshaw (1994): 38.4 ms for frame length, 6.4 ms for frame interval, 50Hz-250Hz for male speaker, and 120Hz-400Hz for female speaker for F0 range. The evaluation programs are also from the original database. Since the original results in Bagshaw (1994) contain both voicing and pitch estimation, evaluation results with voicing detection for the current algorithm (SHRPv) are also reported.

Tables 3.1 and 3.2 show the evaluation results for male and female speech, respectively, which include unvoiced error rate (UER), voiced error rate (VER), GER, and absolute deviation. The GER is further classified into High and Low, which represent the percentage of incorrect F0 values that are higher or lower than the reference F0, respectively. Due to some mislabeling in the database, the voiced error rate (VER) was not zero even though all the frames were set as voiced. After correction, improvements were

observed. However, for a fair comparison, only the results using the original database are presented. Note that the results of several other PDAs on this database are also listed, which are obtained from Bagshaw (1994) and Ying et al. (1996).

PDAs	Male					
	Voice (%)		GER (%)		Absolute Deviation (Hz)	
	UER	VER	High	Low	Mean	SD
CPD	18.11	19.89	4.09	0.64	2.94	3.60
FBPT	3.73	13.90	1.27	0.64	1.86	2.89
HPS	14.11	7.07	5.34	28.2	3.25	3.21
IPTA	9.78	17.45	1.40	0.83	2.67	3.37
PP	7.69	15.82	0.22	1.74	2.64	3.01
SRPD	4.05	15.78	0.62	2.01	1.78	2.46
mAMDF	N/A	N/A	3.35	5.19	N/A	N/A
mAMDFp	N/A	N/A	1.94	2.33	N/A	N/A
eSRPD	4.63	12.07	0.90	0.56	1.40	1.74
GET_F0	155.55	0.22	1.14	2.50	2.76	3.54
PRAAT	206.47	0.15	1.23	1.53	2.22	3.42
SHRP	**218.78**	**0.02**	**0.58**	**1.21**	**1.90**	**3.07**
SHRPv	**18.24**	**6.46**	**0.26**	**0.70**	**1.64**	**2.40**

Table 3.1: PDA evaluation for CSTR male speech.

PDAs	Female					
	Voice (%)		GER (%)		Absolute Deviation (Hz)	
	UER	VER	High	Low	Mean	SD
CPD	31.53	22.22	0.61	3.97	6.39	7.61
FBPT	3.61	12.16	0.60	3.55	5.40	7.03
HPS	19.10	21.06	0.46	1.61	4.59	5.31
IPTA	5.70	15.93	0.53	3.12	4.38	5.35
PP	6.15	13.01	0.26	3.20	6.11	6.45
SRPD	2.35	12.16	0.39	5.56	4.14	5.51
mAMDF	N/A	N/A	1.22	14.8	N/A	N/A
mAMDFp	N/A	N/A	0.63	2.93	N/A	N/A
eSRPD	2.73	9.13	0.43	0.23	4.17	5.13
GET_F0	94.88	1.63	1.09	0.86	5.43	7.10
PRAAT	249.11	0.04	1.21	0.88	4.82	6.76
SHRP	**250.58**	**0.02**	**0.89**	**0.85**	**4.95**	**6.72**
SHRPv	**5.64**	**10.17**	**0.39**	**0.40**	**4.16**	**5.24**

Table 3.2: PDA evaluation for CSTR female speech.

- Cepstrum pitch determination (CPD) (Noll, 1967)
- Feature-based pitch tracker (FBPT) (Phillips, 1985)
- Harmonic product spectrum (HPS) (Schroeder, 1968)
- Integrated pitch tracking algorithm (IPTA) (Secrest and Doddington, 1983)
- Parallel processing method (PP) (Gold and Rabiner 1969)
- Super resolution pitch determinator (SRPD) (Medan et al., 1991)
- Enhanced version of SRPD (eSRPD) (Bagshaw, 1993)
- Modified AMDF-based PDA without error correction (mAMDF) (Ying et al., 1996)
- Modified AMDF-based PDA with probabilistic error correction (mAMDFp) (Ying et al., 1996)
- Subharmonic-to-Harmonic Ratio based pitch determination (SHRP)

For the CSTR male speech (Table 3.1), without voicing detection the present algorithm (SHRP) achieves 0.58% and 1.21% for GER High and GER Low, respectively. By enabling a voicing detection module, the algorithm (SHRPv) yields 0.26% and 0.70%, respectively. For the CSTR female speech, the results are 0.89%, 0.85% (for SHRP) and 0.39% and 0.40% (for SHRPv). Among other algorithms listed in the tables, eSRPD performs the best. By taking the average of GER High and GER Low for both male speech and female speech, the overall mean GER for eSRPD is 0.53%, whereas the current (SHRPv) is 0.44%. This shows that the performance of current algorithm is better than that of eSRPD algorithm when the voicing estimation error is within the same scale.

For the Keele database, the configuration is: 40 ms for frame length, 10 ms for update interval, and [50 550] for F0 range, which is the default setting of "GET_F0" program. Tables 3.3 and 3.4 list GER for each speaker and mean GER for each gender.

PDAs	GER (%)					
	M1	M2	M3	M4	M5	Mean
PDA	5.17	10.22	3.40	3.16	5.15	5.42
GET_F0	1.49	11.36	2.74	2.59	1.59	3.95
PRAAT	3.36	8.32	1.30	2.96	1.59	3.30
SHRP	**4.29**	**4.49**	**0.41**	**0.55**	**0.68**	**2.08**

Table 3.3: PDA evaluation for the Keele male speakers.

PDAs	GER (%)					
	F1	F2	F3	F4	F5	Mean
PDA	7.28	4.97	4.22	14.06	4.48	7.00
GET_F0	11.23	6.15	6.62	7.15	2.74	6.78
PRAAT	4.31	2.21	2.98	4.66	1.08	2.99
SHRP	**2.22**	**1.63**	**1.66**	**2.61**	**0.59**	**1.74**

Table 3.4: PDA evaluation for the Keele female speakers.

As demonstrated in Table 3.3 and Table 3.4, the mean GERs of the present algorithm (SHRP) for male speakers and female speakers are 2.08% and 1.74%, respectively. The Praat algorithm is the best among the other three algorithms, which has 3.30% and 2.99%, respectively.

The above evaluation results from the two databases show that the present approach indeed yields better pitch estimation. Note that the present algorithm does not employ any post-processing techniques in this evaluation, whereas others usually have an integrated post-processing module. Although it could be argued that the algorithm is developed and evaluated based on the same databases so the results may be biased, it is believed that the bias effect is not significant. This is because the parameter values used in the pitch determination module are not derived by fine-tuning using the databases but rather from perception results (Sun and Xu, 2002) or commonly used by other algorithms.

3.4 Application to Intonation Research

The motivation for developing this algorithm was to find a better way to handle alternate cycles in speech, and consequently deliver reliable F0 values for intonation research. The above evaluation demonstrates that the present algorithm performs very well when using the F0 values of EGG signals as the reference. This constitutes an important step towards the original goal of generating “intonation contours”, which suppresses the local pitch jumps and yield globally continuous F0 contours. In the context of the present algorithm, the subharmonic frequency would be suppressed, while the harmonic frequency would be

favored more frequently by increasing the SHR threshold. Figure 3.4 is an illustration of applying SHR threshold analysis to pitch determination on a real speech sample. The utterance produced by a Mandarin speaker contains syllables with the Low tone, which is known to be associated with creaky voice. It can be seen that alternate cycles appear at several regions, such as [0.54s, 0.6s], and [0.7s, 1s]. Figure 3.4(b) shows SHR curves for each short speech frame in 40ms increments. It can be seen in the figure that the SHR curve indicates the locations of alternate cycles quite precisely where typical pitch tracking errors might occur. The higher SHR values in the range of [0s, 0.2s] are due to the unvoiced segments. When using the default threshold (i.e., 0.2) there is a discontinuity from 0.54s to 0.6s due to the alternate cycles and subsequently a lot of pitch octave jumps are observed within the latter part of the utterance (see Figure 3.4(c)). As shown in Figure 3.4(d), application of standard solutions, such as a five-point median filter to smooth outliers, does not help much. To alleviate this situation, larger threshold values are used, such as values greater than 0.5. In doing so, the harmonic frequency is chosen as the F0 even though the subharmonic component is very strong. The result of using a threshold 0.8 (shown in Figure 3.4(e)) demonstrates that by increasing the threshold, some pitch values become one octave higher while the numbers of "bad" points are reduced. Then median smoothing can be applied. Figure 3.4(f) shows that the discontinuity between 0.54s and 0.6s is now removed. Nonetheless, the longer segment between 0.7s and 1s remains to be a problem. For this kind of severe situation, some additional information, such as simultaneously recorded EGG signals, is probably needed. Nevertheless, the current algorithm can solve

some moderate problems, if not all, towards extracting pitch contours suitable for intonation modeling.

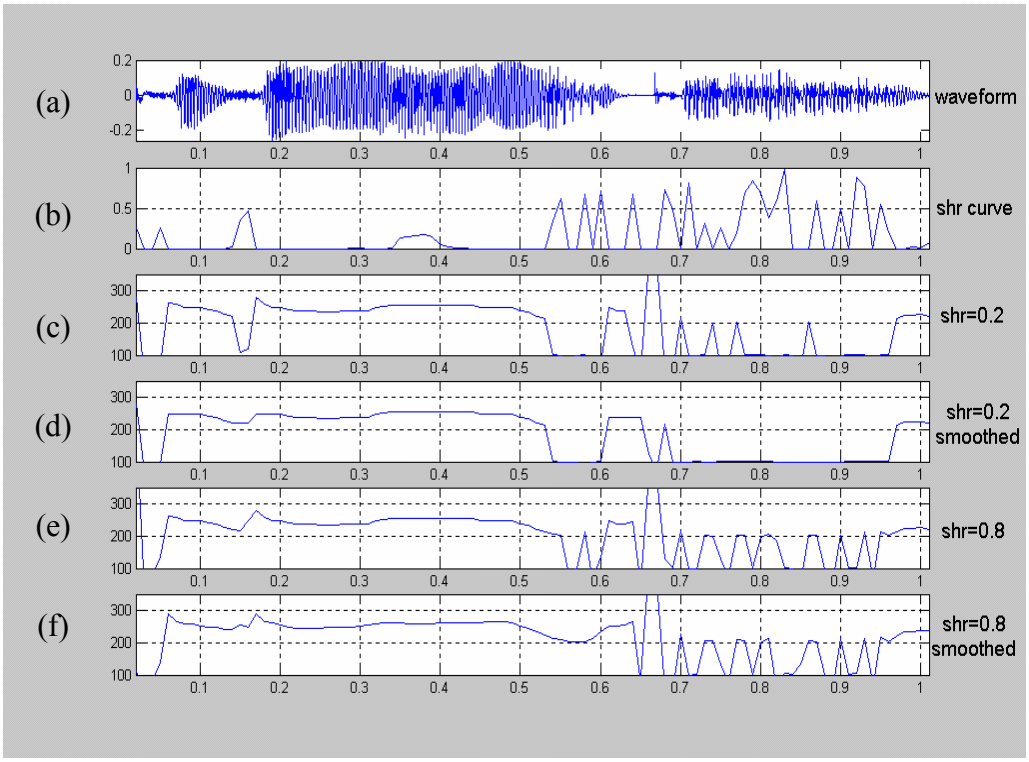


Figure 3.4: An illustration of the pitch determination results on a segment of speech by the current SHR based algorithm. (a) speech waveform; (b) SHR values of each short frame (40 ms); (c) raw pitch contour with SHR Threshold = 0.2; (d) smoothed (five-point median filter) pitch contour with SHR Threshold = 0.2; (e) raw pitch contour with SHR Threshold = 0.8; (f) smoothed (five-point median filter) pitch contour with SHR Threshold = 0.8.

It should be noted that above analysis concerns F0 contours mostly for speech synthesis. Different approaches might be needed when applying the algorithm to other domains, such as speech recognition or speaker recognition. For example, as mentioned

earlier, pitch lowering resulted from alternate cycles may be useful in identifying certain sentence boundaries as glottalization often occurs at then end of the sentence (Klatt and Klatt, 1990; Redi and Shattuck-Hufnagel, 2001). Consequently, a high SHR threshold as employed above may thus become undesirable.

3.5 Summary

In this chapter, a pitch determination algorithm based on Subharmonic-to-Harmonic Ratio (SHR) estimation is proposed and tested[3]. The motivation of the algorithm is to extract "intonation contours" from speech signals, which should be continuous and smooth. Pitch tracking errors such as pitch doubling and halving are undesirable for intonation modeling yet frequently occur. In this work, alternate cycles in speech are identified as a source of pitch tracking errors. An algorithm is therefore developed to handle alternate cycles. Based on the properties of alternate cycles in the frequency domain, subharmonic-to-harmonic ratio is proposed as a parameter to quantify the variation patterns of alternate cycles. A perception study reveals a close relationship between SHR and perceived pitch. The core of the algorithm is calculating SHR and applying the results of the perception study to determine pitch. Performance analysis indicates that it is superior to other algorithms being evaluated. Finally, application of this algorithm to intonation research is addressed. SHR

[3] The source code and evaluation routines are available at

<*http://mel.speech.nwu.edu/sunxj/pda.htm*>

can also potentially be used in voice quality research. Appendix B presents a preliminary discussion on voice quality analysis using SHR.

Chapter 4

Analysis and Synthesis of Intonation Using Underlying Pitch Targets

This chapter addresses issues related to parameterization of F0 contours. One of the key ideas of the present approach is to model the underlying form of F0 contours rather than the surface form. The parametric form of intonation adopted in this thesis is based on the pitch target approximation model described in Xu and Wang (2001) and Xu et al. (1999). Although the model was originally developed for Mandarin, the intonation production mechanism from which the model is motivated is believed to be universal. In this chapter an application of the model to American English is presented.

4.1 Underlying Pitch Target Analysis

In order to apply a parametric intonation model to tasks like speech synthesis, speech recognition, or theoretical analysis, an automatic, robust, and efficient procedure to extract model parameters from F0 contours is highly desirable. As shown in Chapter 2, such a procedure plays an important role in intonation models, especially for parametric approaches, such as the Fujisaki model and the Tilt model. Therefore, the rest of this section is devoted to developing an automatic routine to extract underlying pitch targets

from F0 contours. This procedure is based on the formulation described by Xu et al. (1999) with some modifications.

First, for each syllable in the range of [0, D], we define

$$T(t) = at + b \tag{4.1}$$

$$y(t) = \beta \exp(-\lambda t) + at + b \tag{4.2}$$

$$0 \le t \le D\ ,\ \lambda \ge 0$$

where $T(\cdot)$ represents the underlying pitch target, and $y(\cdot)$ represents the surface F0 contour. Parameters a and b are the slope and intercept of the underlying pitch target, respectively. These two parameters describe an intended intonational goal by the speaker, which can be very different from the surface F0 contour being observed. Coefficient β is a parameter measuring the distance between F0 contour and the underlying pitch target when t = 0. Parameter λ is a positive number representing the rate of decay of the exponential part. In other words, it describes how fast the underlying pitch target is approached. The greater the value of λ is, the faster the speed. Due to the physiological limits of human speech apparatus, these parameters are all subject to certain constraints, such as the maximum pitch range and maximum speed of pitch change (Xu and Sun, 2002).

The estimation of these parameters can be done through nonlinear regression. However, as pointed out by Ratkowsky (1990), the above model or alike does not consistently have good estimation properties. One of the solutions is replacing some of the parameters with so-called *expected-value* parameters (Ratkowsky, 1990). Let (t_0, y_0) denote

the first point on the F0 contour. By plugging this point into Eq. (4.2), we can replace parameter β, and have:

$$y(t) = (y_0 - b)\exp(-\lambda t) + at + b \tag{4.3}$$

Note that here we assume $t_0 = 0$.

Next, let (t_1, y_1) denote a point where the exponential component becomes zero, i.e., the position where underlying pitch target is reached or almost reached. Note that an exponential function can never be zero but approximates zero indefinitely. Here we force the exponential part to be zero in order to simplify the model and make the regression analysis more robust. In doing so, we in fact deviate from the original assumption proposed in Xu and Wang (2001) and Xu et al. (1999) that the underlying pitch target is asymptotically approximated. Instead, the present formulation assumes the target is reached or almost reached within the corresponding syllable. Thus Eq. (4.3) becomes:

$$y_1 = at_1 + b \tag{4.4}$$

With Eq. (4.4) we can replace either a or b in Eq. (4.3). That is:

$$y(t) = (y_0 - b)\exp(-\lambda t) + (\frac{y_1 - b}{t_1})t + b \tag{4.5}$$

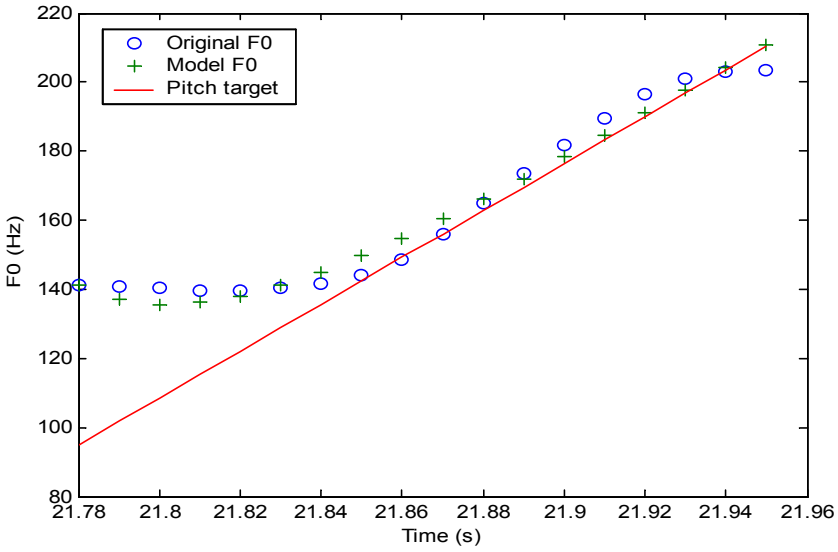
or

$$y(t) = (y_0 - y_1 + at_1)\exp(-\lambda t) + at + y_1 - at_1 \tag{4.6}$$

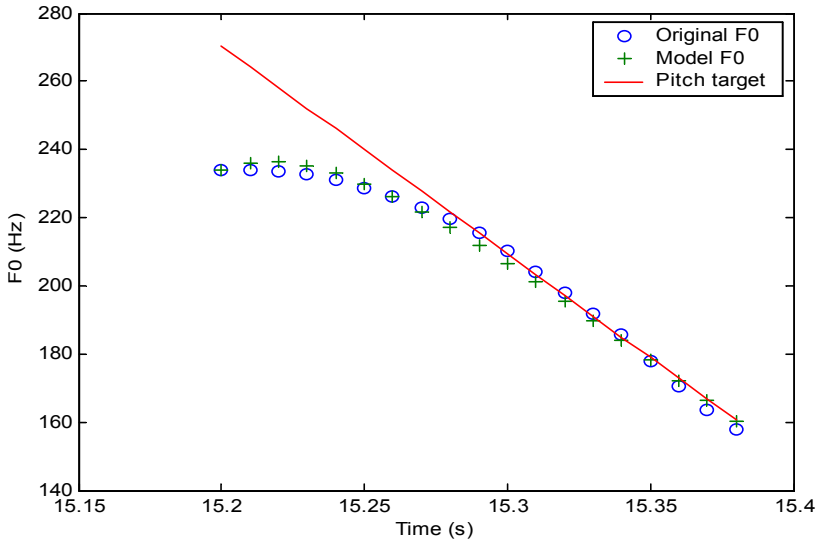
Empirical evidence shows that Eq. (4.6) has better estimation properties. We can estimate λ and a using Eq. (4.6), and derive b using Eq. (4.4).

The nonlinear regression routine used for estimation is an implementation of the widely used Levenberg-Marquardt algorithm. When nonlinear regression fails, linear regression is performed and parameters β and λ are set to zero. In practice, for (t_0, y_0), an average of the first two F0 values is used in estimation because the very first point sometimes can be aberrant. For (t_1, y_1), the middle point is empirically selected. This is based on the assumption that generally the pitch target should be approached around the middle point of the syllable. Note that this analysis procedure is performed on the vocalic part of a syllable. Thus, each syllable is associated with one pitch target. This appears to be a nontrivial assumption, which, unfortunately, could be too simplistic under certain conditions. Nevertheless, such an assumption brings us consistency and efficiency. The above procedure not only assumes one pitch target in each syllable, but also constrains the target to be implemented all the way to the end of the syllable (the vocalic part). This could be a flaw as the ending portion of F0 contour may be affected by the following consonants (Silverman, 1987) while the current model does not have the mechanism to factor them out. Some preliminary work aiming at this issue is presented later in this chapter.

To illustrate the concept of the model and the underlying pitch targets, two analysis results are illustrated in Figure 4.1. Figure 4.1a shows a syllable with rising pitch movement. The algorithm identifies an underlying pitch target with a positive slope, which seems to be quite reasonable. The Model F0 in the figure, referring to the reconstructed F0 using estimated parameters, matches the original F0 very closely. Similarly, Figure 4.1b illustrates an underlying pitch target with falling shape.



(a)



(b)

Figure 4.1: Examples of underlying pitch targets (lines) and both the original surface F0 contours (circles) and that generated by the model (pluses) with semitone scale: (a) a target with positive slope; (b) a target with negative slope.

4.2 Properties of the Model Parameters

This section discusses several aspects of the model parameters by applying the model to real world data.

4.2.1 The Speech Corpus

The Boston University Radio Speech Corpus, speaker F2B (Ostendorf et al., 1995) is used throughout this thesis. The database, which consists of about 40 minutes of speech read aloud by a female professional announcer, is labeled using the ToBI (Silverman et al., 1992) system. In total, there are 14377 syllables. The database also contains text information, such as part-of-speech, and acoustic information such as phone duration. F0 curves were determined by the SHRP algorithm introduced in Chapter 3 with intervals of 10 ms, and smoothed by de-step filter (Bagshaw, 1994), five-point median and linear filters (Hanning window). No manual correction was performed on F0 values. The mean and standard deviation of the F0 contours are 167.38 Hz and 45.65 Hz, respectively. It can be seen that this female speaker has a relatively low pitch voice. Also there is nontrivial amount of creaky voice in the database, which makes intonation modeling work quite difficult.

4.2.2 Statistics of the Model Parameters

Table 4.1 provides a quick look over the parameters of the pitch target approximation model. The mean and standard deviation of parameter *a*, which represents the slope of an

underlying pitch target, are -94.10 and 367.69, respectively. These values correctly indicate the style of the current corpus, that is, a news-reading style with declarative statement and significant pitch variations. The values of parameter *b* are close to the global values listed above. The differences are reasonable because (1) there are more pitch targets with falling slopes, thus, the average F0 height of starting points is higher than the global mean; (2) the starting F0 of a pitch target (parameter *b*) is a hypothesized point, which could have very high or very low F0 values. This implies that the surface F0 would have a smaller pitch range than parameter *b* since the implementation of an underlying target is subject to physiological constraints, such as the speaker's pitch range and the maximum speed of pitch change (Xu and Sun, 2002). For comparison, the mean and standard deviation of the middle F0 (Mid F0) of underlying pitch targets are 167.78 Hz and 45.07 Hz, respectively, which are almost identical to the global values shown in the previous section (167.38 Hz and 45.65 Hz). The mean and standard deviation of the end F0 (End F0) value of the underlying pitch targets are 162.40 Hz and 51.11 Hz, respectively, which represent an opposite pattern in contrast with the starting F0.

	a (Hz/s)	*b* (Hz)	λ (1/s)	β (Hz)
Mean	-94.10	173.15	73.32	-0.08
Std	367.69	51.37	133.29	30.05

Table 4.1: The mean and standard deviation (Std) of the pitch target approximation model parameters (*a*: Pitch target slope; *b*: Pitch target intercept; λ: Target approximation rate; β: Initial distance to the target) for the F2B corpus.

Parameter λ represents the target approximation speed, which is defined to be a positive number. As shown in Figure 4.2, the values of parameter λ are not normally

distributed. A quick inspection of Figure 4.2 reveals that the majority parameter values fall in the region of [30, 40], which is much lower than the average. Parameter β measures the distance between the starting surface F0 and the underlying starting F0. The almost zero mean shown in Table 4.1 indicates that the starting points of surface F0 are symmetrically distributed with respect to the underlying pitch targets.

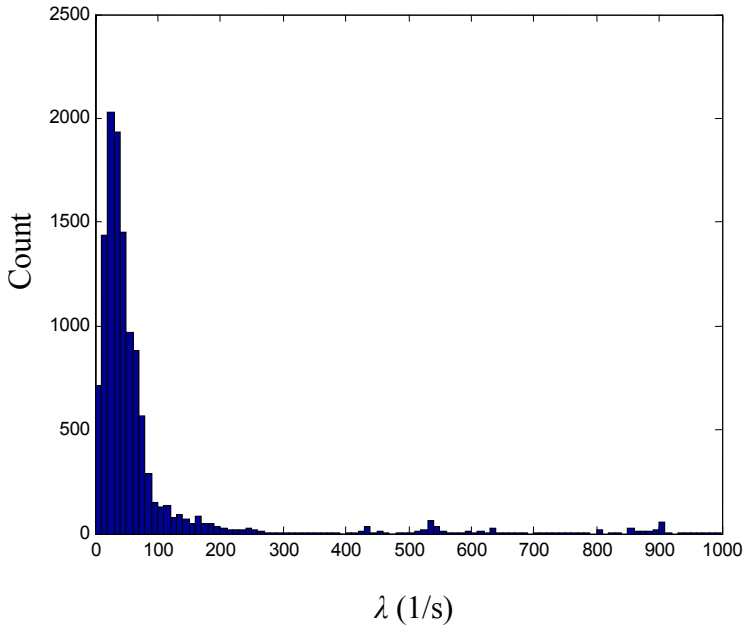


Figure 4.2: Frequency distribution of parameter λ for the F2B corpus. The values of λ are greater than 0 but less than 1000.

4.2.3 Synthesis Accuracy

When using a parametric model to represent intonation, it is imperative to prove that the model can reconstruct F0 contours accurately given appropriate parameter values. Even

though high accuracy is not equal to a good model, a model with poor accuracy is certainly not desirable.

In the following, the model parameters derived from nonlinear regression analysis are plugged into equation 4.2 to generate F0 contours (referred to as Model F0 hereafter). For objective comparison between the original F0 and Model F0, root mean square error (*RMSE*) and Pearson's correlation coefficient (*r*) are adopted, which are commonly used in the literature (e.g. Dusterhoff et al., 1999; Ross and Ostendorf, 1999) They have been shown to be better than other available metrics (Hermes, 1998). Given two F0 contours *X* and *Y*, *RMSE* is defined as:

$$RMSE = \sqrt{\frac{\sum_{i}^{N}(x_i - y_i)^2}{N}}$$

where *N* is the total number of points in each F0 contour.

It is known that listeners are more sensitive to some parts of intonation than others. Thus, if using perception as the accuracy criterion, the above objective measures would be insufficient as they treat all F0 points equally. Nevertheless, as a "perfect" comparison technique is still not available, these measures give us quick and acceptable solutions.

Table 4.2 shows the present results as well as those obtained by Dusterhoff et al. (1999) using the Tilt model. Note that only F0 values at voiced portion are considered. It can be seen that the current approach yields closer F0 values numerically, which confirms the eligibility of the parametric form.

Model	*RMSE* (Hz)	*r*
Present	6.5	0.99
Tilt model (Dusterhoff et al., 1999)	14.5	0.93

Table 4.2: Synthesis accuracy of the present model and the Tilt model. *RMSE* and correlation coefficient (*r*) are calculated between the original F0 values and the reconstructed F0 values.

4.2.4 Orthogonality of the Model Parameters

To represent and generate F0 contours efficiently, a compact model is highly desirable (Taylor, 2000). This requires the model parameters to be less correlated with each other. If the model parameters are highly correlated, it implies that redundant information is encoded. Another advantage of using independent parameters to model intonation is that a compact and proper feature set is more likely to be selected to predict these parameters.

Table 4.3 shows the correlation matrix for the model parameters. Pitch target duration is also included for the purpose of reference. It can be seen that the magnitude of correlation between the two core parameters *a* and *b* is not low. This could be explained by the fact that when a pitch target has a significant rise, it tends to start at a very low F0 value. Similarly a sharp fall implies a higher starting point. To compare this point with values at other positions, Table 4.4 presents the correlations between *MidF0*, *EndF0* and other model parameters. It is evident that *MidF0* has much lower correlations with other model parameters. This indicates that *MidF0* may be a better model parameter when compared with parameter *b* and *EndF0*.

	a	*b*	*λ*	*β*	*Duration*
a	1.00				
b	-0.41	1.00			
λ	-0.13	-0.02	1.00		
β	0.32	-0.51	0.00	1.00	
Duration	0.07	0.03	-0.37	-0.13	1.00

Table 4.3: Correlation matrix for the pitch target approximation model parameters (*a*: Pitch target slope; *b*: Pitch target intercept; *λ*: Target approximation rate; *β*: Initial distance to the target) and the duration of the pitch targets for the F2B corpus.

	a	*λ*	*β*	*Duration*
Mid F0	-0.01	-0.01	-0.33	-0.04
End F0	0.39	-0.01	-0.07	-0.10

Table 4.4: Correlation coefficient values between the pitch target approximation model parameters, Mid F0 and End F0, respectively.

To compare the current results with others, the following two tables list the correlation matrices for the Tilt parameters (Taylor, 2000) and the Three-Target model parameters (Black and Hunt, 1996), respectively. The parameter values of the two models were also extracted from the aforementioned F2B corpus.

Tables 4.3-4.6 demonstrate that in general the parameters of the pitch target approximation model have lower correlations when using *MidF0* in place of parameter *b*. This indicates that the pitch target approximation model not only can reconstruct more faithful F0 contours, but also are more compact than the other two models.

	Position	*Start F0*	*Amplitude* (A_{event})	*Duration* (D_{event})	*tilt*
Position	1.00				
Start F0	0.11	1.00			
Amplitude	0.08	0.36	1.00		
Duration	-0.05	-0.06	0.32	1.00	
tilt	0.64	0.32	0.08	-0.15	1.00

Table 4.5: Correlation matrix for the Tilt model parameters calculated from the F2B corpus.

	Start F0	*Mid F0*	*End F0*
Start F0	1.00		
Mid F0	0.82	1.0000	
End F0	0.64	0.86	1.00

Table 4.6: Correlation matrix for the Three-Target model parameters calculated from the F2B corpus.

4.3 Predicting Underlying Pitch Targets

Having a model to describe intonation accurately is never our last stop. Rather, whether the model parameters are predictable from high level linguistic features is crucial to the success of intonation modeling. In Sun (2001), a recurrent neural network was employed to predict the parameters of underlying pitch targets given some linguistic features. Trained on a small corpus, the system achieved promising results. To further illustrate the predictive power of the model, this section conducts an experiment to predict model parameters using the CART algorithm described in Chapter 2. Training and testing data were taken from the above F2B corpus with approximately a 9:1 ratio similar to that in Dusterhoff et al. (1999).

"WAGON" (Taylor et al., 1999), an implementation of standard CART algorithm, was used to build regression trees.

4.3.1 Methods

First, for each syllable in the database, the following features were extracted.

- Vowel type
- Coda type
- Syllable Stress
- Syllable position in a word
- Number of syllable in a word
- Pitch accent type
- Phrase accent type
- Number of syllables to the major phrase break
- Part of speech
- Word position in a sentence
- Number of words to the major phrase break

Contextual information was also included for some features, such as adjacent pitch accent, syllable stress, etc.

The input features were applied to the decision tree program to train two regression trees, whose outputs are parameters *a* and *MidF0*, respectively. Note that all the syllables are considered here. This is different from those "accent-based" approaches such as Dusterhoff et al. (1999), in which only accented syllables were modeled.

The trained trees were used to predict parameter *a* and *MidF0* on the testing data given the input features. To generate F0 contours from pitch targets, many routes can be

taken. One way could be training two additional trees for parameter λ and β, which, however, have turned out to be less fruitful on the current database. Directly connecting underlying pitch targets as that in Sun (2001) ignores too much segmental information. In the current work, a more complex interpolation scheme was employed: For a given pitch target, a cubic-spline interpolation starting from the end of the previous target to the middle point of the current target was performed. This ensures the smoothness of pitch trajectory. It should be noted that this interpolation scheme is still quite simple and even incorrect in some cases. Possible extensions include defining rules based on existing linguistic knowledge. For example, the starting F0 value of a syllable may be raised to a certain extent when the syllable starts with certain consonants (Silverman, 1987; Xu et al., 2002).

In the above each parameter of an underlying pitch target is predicted using a single decision tree. To explore the effectiveness of ensemble machine learning on pitch target prediction, Bagging and AdaBoost, described in Chapter 2, were tested on the top of CART. For both Bagging and AdaBoost, the maximum number of trees was set to be 50. The stop value for each tree was 5. Here the prediction is a regression problem since continuous pitch target parameters are the targets. Therefore, the regression version of AdaBoost algorithm shown in Figure A.2 in Appendix A was employed. The final result was the (weighted) mean of the ensemble tree outputs rather than a majority vote.

4.3.2 Results

Table 4.7 presents comparison results between predicted F0, original F0, and model F0 of several systems. Since the present approach is a parametric approach, the model F0, which is the reconstructed F0 using the original parameter values, represents the upper limit for the predicted F0. To compare with other approaches, Table 4.7 also list the results of the Three-Target model (Black and Hunt, 19996), the Tilt model (Dusterhoff et al., 1999), and the Dynamical System model (Ross and Ostendorf, 1999). Note that the results of some other similar studies (e.g. Buhmann et al., 2000; Molher and Conkie, 1998; Traber, 1992) are not included since they were obtained from different databases.

	RMSE (Hz)	*r*
Original F0 – Predicted F0	33.1	0.72
Model F0 – Predicted F0	32.2	0.73
Original F0 – Predicted F0 (Black and Hunt, 1996)	34.8	0.62
Model F0 – Predicted F0 (Dusterhoff et al., 1999)	34.3	0.60
Original F0 – Predicted F0 (Ross and Ostendorf, 1999)	34.7	N/A

Table 4.7: F0 prediction results of the present approach and other systems. *RMSE* and correlation coefficient (*r*) are calculated between the original F0 values and the reconstructed model F0 values, and the predicted F0 values.

Table 4.7 shows that in terms of *RMSE*, the present approach achieves 33.1 Hz (compared with Original F0) and 32.2 Hz (compared with Model F0), respectively. Both of them are lower than the *RMSEs* obtained by other approaches. In terms of correlation coefficient, the proposed model yields much higher values, which are 0.72 and 0.73, respectively. These results demonstrate that the proposed model parameters have higher

predictive power than those of other systems. It should be noted that different ways of computing *RMSE* and correlation coefficient might affect the results. In this work, a *RMSE* and a correlation coefficient value were calculated for each file (in the current database, each file contains one paragraph, which usually consists of several sentences), and the final results were the average of all the files in the testing set. Other authors might have used different schemes.

Informal listening indicates that the utterances synthesized with the predicted F0 contours are fairly close to the original. Although this is not completely surprising as ToBI accent features provide important intonation information, it is surely encouraging because it shows the good predictive power of underlying pitch targets. Note that all the systems compared in Table 4.7 were trained with hand-labeled accent information.

Table 4.8 lists the numerical results of ensemble learning in terms of *RMSE* and correlation coefficient. Based on these objective criteria, clearly both Bagging and Boosting outperform the single CART system. However, informal listening on the several testing utterances did not reveal audible perceptual difference. It is speculated that ensemble learning might yield more robust intonation in the long run. That is, it may behave more consistently across various conditions, which, however, can only be proved with more extensive and systematic evaluations.

	RMSE (Hz)	*r*
CART	33.1	0.72
Bagging	31.9	0.75
AdaBoost	31.7	0.75

Table 4.8: Comparison between CART, Bagging, and AdaBoost for F0 generation.

4.4 Pitch Accent Prediction Using Underlying Pitch Targets

Predicting model parameters from high-level linguistic features for F0 generation is by no means an easy task. However, as pointed out by Hirst et al. (2000), the inverse problem is even more challenging, that is, "Given an F0 curve, how can we recover a symbolic representation?" Figure 4.3 illustrates a simple diagram of this two-way process. Functioned as a bridge between symbolic representation and actual F0 curve, a good model should perform equally well in both coding and synthesis.

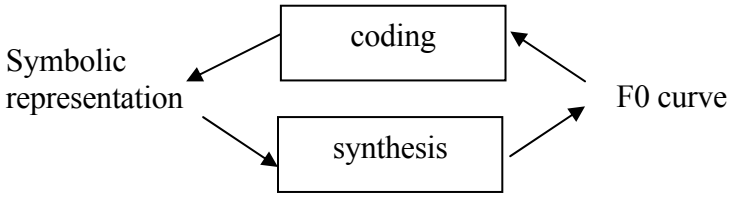


Figure 4.3: Relations between symbolic representation, phonetic F0 model, and F0 curve.

In the previous section, parameters of underlying pitch targets are predicted from high-level symbolic linguistic information and indicate good predictive power. In this section, to illuminate its capability of recovering symbolic representation, underlying pitch targets are used as input to predict ToBI style pitch accent – a type of prosodic symbols. This work has previously been reported in Sun (2002c).

Automatic prediction of prosodic patterns such as pitch accents can find applications in text-to-speech, automatic speech recognition, and corpus development.

Depending on the application, prosodic event recognition systems can utilize acoustic information, text information, or both. A system capable of recognizing pitch accent from acoustic information can be applied to speech recognition or understanding systems. For example, Niemann et al. (1997) successfully applied pitch accent and other prosodic information to their speech-to-speech translation system and achieved improved performance.

A variety of algorithms have been investigated for predicting prosodic patterns, including Hidden Markov Model (HMM) (e.g. Taylor, 2000), neural network (e.g. Muller and Hoffmann, 2001), dynamical system (Ross and Ostendorf, 1995), and decision trees (e.g. Hirschberg, 1993). However, to the author's knowledge, the use of ensemble learning has not been investigated before. Thus, in addition to examining the eligibility of underlying pitch target in predicting pitch accent, this section also explores the application of ensemble learning to pitch accent prediction.

Three experiments were conducted based on the input feature type: (1) pitch accent prediction using only acoustic features; (2) pitch accent prediction using only text features; (3) pitch accent prediction using both acoustic and text features. These correspond to three applications of pitch accent prediction, i.e., speech recognition, speech synthesis, and corpus development, respectively.

Training and testing data were again taken from Boston University Radio Speech Corpus, speaker F2B. Similar to Ross and Ostendorf (1995), the ToBI pitch accent labels were grouped into four types: High, Low, Down-stepped high, and Unaccented. The labels

were aligned with syllables. Syllable boundaries were assumed known during pitch accent prediction. The distribution of pitch accent types in the database is shown in Table 4.9. The data set was split into training and testing sets with approximately a 4:1 ratio as that in Ross and Ostendorf (1995).

	Pitch accent type			
	Unaccented	High	Downstep	Low
Training set	7804	2717	853	151
Testing set	1929	677	211	35

Table 4.9: Pitch accent distribution in the F2B database.

4.4.1 Methods

In each experiment, models were built using single CART, bagging with CART, and AdaBoost with CART. Since this is a multi-class classification problem, the AdaBoost.M1 algorithm shown in Figure A.1 in Appendix A was used. The number of iterations for bagging and boosting was limited to 50. Guided by the theory of bias and variance decomposition, the following procedures were adopted: Overtrain CART to generate a tree with low bias by using a small stop value, which refers to the minimum number of samples in the leaf nodes; Use bagging or boosting to reduce variance. The results of ensemble trees were combined by taking an unweighted vote (for Bagging) or weighted vote (for AdaBoost).

When predicting pitch accent from acoustic utterances, many acoustic features are thought to be correlates of pitch accent. Only F0, energy, and segmental duration were

considered in this study. The F0 related features were derived from underlying pitch targets. Several parameters from each pitch target were extracted, including middle F0 value (MidF0), the slope, and the change of F0 and slope between pitch targets, i.e. ∆ MidF0 and ∆ Slope. Together with syllable energy and duration, the feature set contains:

- MidF0 of the current, previous, and next pitch target
- ∆ MidF0 with respect to the previous and next pitch target
- Slope of the current, previous, and next pitch target
- ∆ Slope with respect to the previous and next pitch target
- Syllable duration
- Syllable energy

Stop value 30 was chosen for single CART since it yielded lowest error on the testing set. For bagging and boosting, stop value 5 was used in order to generate overtrained trees with low bias.

Prediction of pitch accent from text has been studied extensively in the past due to its important application in speech synthesis. It has been shown that many factors can affect pitch accent placement. In this work, however, choices were limited to those that could be derived from unrestricted text without much difficulty. The feature set contains:

- Vowel identity
- Syllable stresses of the preceding and following syllables
- The positions of the current, previous, and next syllables in a word
- Number of syllables in the current and previous words
- Part-of-speech of the previous and next words
- A composite feature made up by part-of-speech and stress for the current syllable

- Number of words from the beginning of the sentence and to the end of the sentence

The stop value was 20 for single CART, and 5 for bagging. For boosting, however, stop value 20 was used, which gave better results than a smaller value.

In the third experiment, the acoustic and text features listed above were combined to predict pitch accent. The stop value was 20 for single CART, and 5 for both bagging and boosting.

4.4.2 Results

For a quick comparison, Table 4.10 lists the overall correct rate regardless of pitch accent type for all the experimental conditions. Detailed evaluation results in the form of confusion matrix are shown in Appendix C. The same evaluation method used by Ross and Ostendorf (1995) was adopted since their study and the present work are very similar with respect to experimental configuration. In the tables, each column represents the prediction results for each pitch accent type with percentage and frequency count.

	Overall correct rate (%)
Text – CART	80.47
Text – Bagging with CART	80.64
Text – AdaBoost with CART	80.50
Acoustic-CART	82.89
Acoustic - Bagging with CART	84.71
Acoustic - AdaBoost with CART	84.71
Both – CART	84.26
Both - Bagging with CART	86.89
Both - AdaBoost with CART	87.17

Table 4.10: The overall correct rate for pitch accent prediction using CART, Bagging, and AdaBoost. Input features include text and acoustic features. The acoustic features are derived from the pitch target parameters, syllable duration, and syllable energy.

Table 4.10 demonstrates that ensemble learning, bagging and AdaBoost, yields favorable results over a single decision tree. The best results were obtained in the third task, in which the overall correct rates for bagging and AdaBoost are 86.89% and 87.17%, respectively. For the same task, 84.26% was achieved with a single CART. Note that the third task includes both text and acoustic features. This implies that when more input features are available, they might be better exploited by combining multiple classifiers. The gain of ensemble learning in the second task is less impressive. One of the possible reasons could be that the text-based input features used in the second task were insufficient to predict pitch accent. The consequence of this insufficiency is that some patterns are extremely difficult to learn, which could not be remedied even by combining multiple trees. Therefore, better feature sets are needed in future studies. For example, since the pitch

accents are predicted at syllable level, it may be more reasonable to convert part-of-speech from a word-level feature to a syllable-level feature.

It is usually difficult to compare results obtained from different studies directly because the corpus, prosodic labeling scheme, input feature set, and many other important experimental conditions could be different. Nevertheless, the present work shares many similarities with Ross and Ostendorf (1995, 1996), and hence the results may be comparable. In Ross and Ostendorf (1995), a dynamical system with a bigram tone sequence model was developed to predict pitch accent using acoustic features and an 84.61% (calculated from Table 1 in their paper) overall correct rate was achieved. In this work, both bagging and boosting yield 84.71% overall correct rate. In Ross and Ostendorf (1996), decision trees combined with Markov sequence models were used to predict pitch accent using text-based features and an 80.17% (calculated from Table VI in their paper) overall correct rate was obtained. Correspondingly, in the second experiment of the present study, bagging and boosting achieve 80.64% and 80.50% overall correct rates, respectively. Note that simpler feature sets were used in this work, whereas in Ross and Ostendorf (1995) the intermediate phrase information was assumed to be available, which could affect the results in a nontrivial way. Thus, it is believed that using underlying pitch targets and ensemble learning represent a promising direction to pursue in predicting pitch accent as well as other prosodic events.

The encouraging results indicate that underlying pitch target demonstrates itself as an eligible representation form for higher-level intonation information. It can be used for

both F0 generation and prosodic event recognition. The results also show that by combining multiple decision trees one can consistently improve system performance without adding much complexity. Even better results could be obtained with more sophisticated input features and ensemble learning algorithms, but those experiments remain to be done. In order to see more applications of ensemble learning, preliminary work has been done to apply Bagging and AdaBoost to other prosodic modeling problems, such as duration modeling and phrase break prediction, which are included in Appendices D and E, respectively. Both experiments show encouraging results.

It has been illustrated by many studies that boosting usually performs better than bagging (e.g. Dietterich, 2002). The results of bagging and boosting in this work, however, seem to be quite similar. Moreover, bagging seems to be faster in reducing error rate. In other words, to achieve similar performance, bagging needs less iterations or fewer classifiers. Additionally, boosting is essentially sequential, whereas bagging can be executed in parallel. Thus, to build a prosodic event recognition system, bagging seems to be a better choice to begin with. It should be noted that the current boosting implementation is the simplest one for multi-class problems. It is expected that better results could be achieved by using more sophisticated versions, such as AdaBoost.M2 (Freund and Schapire, 1997).

4.5 Possible Extensions

The above experiments have yielded some success in verifying the significance of underlying pitch targets for intonation modeling. This indicates that the underlying pitch target analysis procedure presented earlier is sound to begin with. However, as pointed out before, there are some inherent problems with the procedure. This section attempts to address these issues by making certain modifications on the model.

In the model, it is assumed that a pitch target is implemented throughout the corresponding syllable. This assumption becomes problematic when the duration is long enough, so that a pitch target may have been fully implemented well before the syllable boundary. The rest of the F0 trajectory may belong to the domain of next pitch target or simply be the result of some involuntary effects, such as post-consonantal perturbation (Silverman, 1987). Although these effects could exist regardless of whether the target ends early or not, it is possible that they have greater effects on the analysis results when the syllable duration is long and the target is approached early regarding the syllable offset. Without considering these effects, the analysis procedure developed above indeed sees some negative outcome. For example, Figure 4.4 illustrates an example processed by the original analysis procedure described earlier. Due to the stringent requirement, the pitch target needs to be implemented through the final part, where a substantial pitch drop has been observed. As a result, the slope of the derived pitch target seems to be shallower than the implemented pitch trajectory.

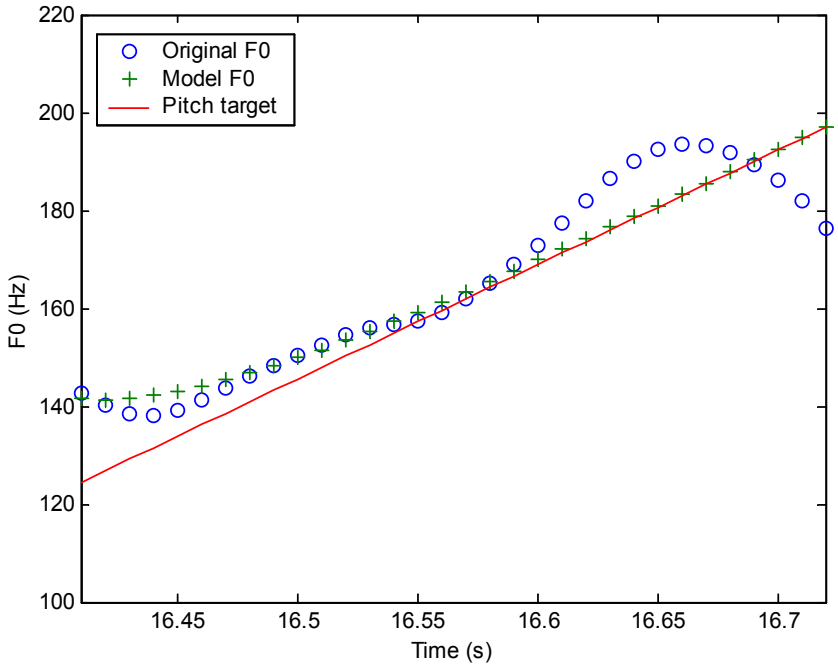


Figure 4.4: An example of problematic results using thee original pitch target analysis.

In Xu et al. (1999), the F0 implementations of two adjacent pitch targets are connected by a pair of hyperbolic tangent functions, which allows a smooth F0 transition and a movable transition point with respect to the syllable boundary. In this work, a similar approach is taken except that the ending point of a pitch target implementation is explicitly modeled by an extra parameter.

4.5.1 Modified Underlying Pitch Target Analysis

First, for a syllable with duration *D*, we denote the point where the implementation of a pitch target ends as t_p, and limit its range in [0.5*D*, *D*]. Hence the definitions of the pitch target and the corresponding surface F0 contour become:

$$T(t) = at + b \tag{4.7}$$

$$y(t) = \beta \exp(-\lambda t) + at + b \tag{4.8}$$

where $0 \le t \le t_p$, $\lambda \ge 0$

Note that Eqs. 4.7 and 4.8 imply that both the underlying pitch target *T(t)* and its implementation *y(t)* are limited by t_p. This should be viewed as a result of a practical simplification rather than a theoretical assumption in that whether the target covers only a portion of a syllable remains an open question, and thus should be left for future research.

For simplicity, the portion of F0 contour between the ending point of pitch target t_p and syllable boundary is described by a first-order linear function:

$$g(t) = ct + d \tag{4.9}$$

where $t_p < t \le D$

Then the surface F0 values in the range of [0, *D*] can be represented by concatenating Eqs. 4.8 and 4.9.

$$Y(t) = \begin{cases} \beta \exp(-\lambda t) + at + b & 0 \le t \le t_p \\ ct + d & t_p < t \le D \end{cases} \tag{4.10}$$

Similar to Xu et al. (1999), a pair of hyperbolic tangent functions is used to connect the two parts. Thus, the total surface F0 contour can be approximately represented by:

$$Y(t) = 0.5[1 - \tanh(m(t - t_p))]y(t) + 0.5[1 + \tanh(m(t - t_p))]g(t) \qquad (4.11)$$

where $0 \le t \le D$, $0.5D \le t_p \le D$, and m is a constant.

The basic idea is when t is less than t_p, the surface F0, i.e. $Y(t)$, is represented by the first function $y(t)$, and when t is greater than t_p, the surface F0 is represented by the second function $g(t)$. The value of m is chosen so that the hyperbolic tangent functions make sharp transitions from -1 to 1 around point t_p. Considering that the current F0 contour is sampled every 10 ms, values greater than 500 should be large enough to meet the requirement. Similar to the previous procedure, nonlinear regression analysis was applied to estimate parameter values.

4.5.2 Preliminary Results

Preliminary analysis was performed on the F2B corpus. Results show that allowing the pitch target to finish early can correct some problems. Figure 4.5 illustrates the analysis result on the same syllable shown in Figure 4.4 with the modified procedure. By comparing Figure 4.5 and Figure 4.4, it can be seen that by explicitly modeling the ending point, the current procedure seems to have generated a more "realistic" pitch target. That is, the slope of the pitch target becomes steeper. Such a correction on pitch target slope could have an impact on the tasks like F0 generation or pitch accent prediction when a pitch movement is perceptually significant. With further inspection of the two figures, we may hypothesize

that if the situation in Figure 4.4 is not uncommon in the database, one would expect that the variation of the slope would be increased over the previous results while the variation of parameter λ reduced. The statistics shown in Table 4.11 support such a hypothesis. The standard deviation of parameter a is larger than that in Table 4.1 while the standard deviation of parameter λ is smaller.

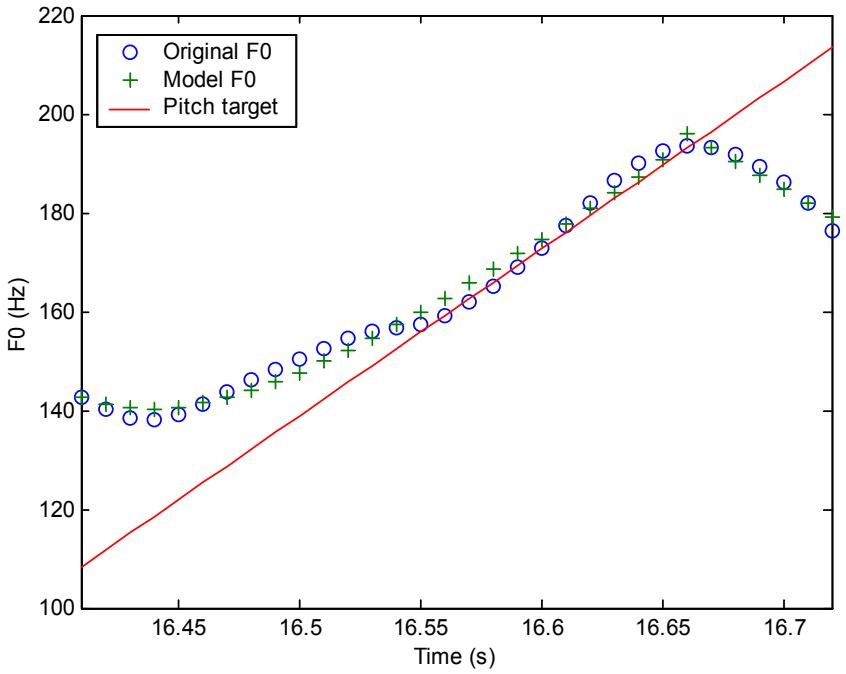


Figure 4.5: An example of the pitch target analysis results using the modified procedure.

	a (Hz/s)	*b* (Hz)	λ (1/s)	β (Hz)	*MidF0* (Hz)
Mean	-86.10	173.46	68.23	0.76	168.59
Std	401.68	49.46	85.66	27.07	45.12

Table 4.11: The mean and standard deviation (Std) of the pitch target approximation model parameters (*a*: Pitch target slope; *b*: Pitch target intercept; λ: Target approximation rate; β: Initial distance to the target) derived with the modified procedure for the F2B corpus.

Note that the above analysis is yet preliminary. More detailed work should be conducted before any specific conclusions can be made. Consequently, the F0 generation system introduced in the next chapter is still based on the original procedure of pitch target analysis.

4.6 Summary

To summarize, this chapter presents a procedure to identify underlying pitch target from surface F0 contours. Experiments were conducted to show the connections between underlying pitch targets and high-level linguistic information. The encouraging results suggest that underlying pitch target is an eligible parametric form of F0, and thus can be used in an F0 generation system. The online demo is available at http://mel.speech.northwestern.edu/sunxj/prosody.htm.

Chapter 5

An F0 Generation System for Speech Synthesis

The previous chapter reports tests regarding the predictability of the underlying pitch target and shows that pitch targets have the potential to be used in an F0 generation system. In this chapter, an F0 generation system based on underlying pitch target is developed for text-to-speech synthesis. This system employs a hierarchical structure to model different components of intonation in different tiers. Guided by existing linguistic knowledge, models for each tier are built with selected linguistic features. The parametric forms of F0 contours are represented by underlying pitch targets. The system is statistically trainable in that the parameters of underlying pitch targets are estimated and predicted via multiple decision trees. The goal of building such a system is twofold. On the one hand, more natural synthetic intonation for speech synthesis is pursued. On the other hand, the development process itself will shed light on theoretical issues and deepen our understanding of intonation production and perception. It will be seen later that many issues discussed in Chapter 1 will be addressed in the F0 generation system. This work has been previously described in Sun (2002b).

To meet the requirement for speech synthesis, an F0 generation system should be effective, that is, it should be capable of generating natural intonation contours. Second, a system should be adaptable to different speaking styles and languages. Third, a system

should be able to synthesize intonation efficiently without adding an excessive computational load to the whole speech synthesis system. It is worth noting that with the rapid increase of computing power, the last requirement may become less critical over time. The proposed F0 generation system strives to meet these requirements. As described in Chapter 2, there are rule-based and data-driven approaches. The F0 generation system proposed in this thesis belongs to the latter category. The recent advancement of corpus-based speech synthesis makes data-driven systems more favorable, as more natural intonation patterns can be statistically learned from the same database used for unit selection synthesis.

5.1 A Multi-Tier F0 Generation System

The model consists of four tiers, with each tier addressing a different aspect of intonation. The basic intonational unit is the *syllable*. In the following discussion first the system structure and the functionality of each tier are described, and then the parametric representations are discussed.

5.1.1 Hierarchical Structure

Tier One models intonation variations beyond the word level. It has been pointed out that in considering the communicative function of prosody, separating prosodic structure within the word from prosodic structure beyond the word level is beneficial (Pierrehumbert, 1999). F0 variation patterns above the word level reflect syntax, semantics, pragmatics, and

discourse information and should be separated from those affected by other local factors. Commonly observed phenomena, such as focus effect and new topic effect will be modeled in this tier. This process is similar to predicting pitch accent for each word (Hirschberg, 1993), except that in this case continuous F0 values, rather than discrete categorical labels, need to be predicted. In other words, phonological target – level tone, is modeled. It is expected that this tier will be the most critical part of the system since it is responsible for modeling the most perceptually significant pitch variations. Intonation at this level is also the most difficult to model as it requires high-level information which is hard to extract from unrestricted text.

Tier Two considers intonation at the word level, i.e., internal F0 variations in a multi-syllablic word. For example, unstressed syllables often have less dramatic pitch variations than stressed syllables. When a word is associated with a focus, the focus usually aligns with the syllable bearing primary stress. And other syllables in the same word often exhibit intonation patterns such as post-focus pitch range suppression. Consequently, modeling unstressed syllables by looking at adjacent stressed syllables may capture these patterns. The F0 variation patterns at this level should be relatively easier to model than those in Tier One, as presumably they would be less affected by high-level information.

The two tiers described above model intonation variations within or beyond the word level. The phonetic details, or phonetic target, at the syllable level must still be accounted for. To fill in this gap, F0 variations within a syllable are modeled in Tier Three using pitch movements, e.g., whether the contour is rising or falling. It should be noted that

not all information within a syllable is modeled at this level. Instead, only the most significant part, namely, a voluntary phonetic target, is addressed.

Tier Four models segmental effects, such as some tonal coarticulation and consonant perturbation, which are not modeled by Tier Three. Although exceptions exist, these effects are usually local, with scope that does not extend beyond the syllable level.

5.1.2 Parametric Representation Using Underlying Pitch Target

The four tiers described above describe the system structure. To generate F0 contours, parametric representations of intonation should be defined. This section will show how the parameters of an underlying pitch target are related to each tier.

For Tier One, one F0 target point (referred to as *anchor F0* hereafter) is chosen from each word (for multi-syllablic words, the point is chosen from the syllable bearing the primary stress). For single-syllable words, the middle F0 point of a syllable's pitch target (MidF0) is selected; for multi-syllablic words, the MidF0 of the syllable bearing primary stress is selected. It should be noted that the database used in the present study only provides information on stress versus non-stress. Thus, the first stressed syllable is arbitrarily selected as the primary stressed syllable. This is a suboptimal solution and should be addressed in future work. It is well known that syllables bearing primary stress usually receive pitch accents (e.g. Botinis et al., 2001; Ladd, 1996), thus motivating the choice of an anchor F0 point from syllables bearing primary stress to represent the whole word. The reasons for using the middle point of a pitch target have been described in

Chapter 4 and are reiterated here. (1) It is assumed that a majority of underlying pitch targets would be reached around the middle of a syllable; (2) *MidF0* is statistically independent from other parameters; (3) Empirical evidence shows that this point performs better than others (e.g. start and end points). It is worth noting that the single F0 point with fixed alignment relative to the syllable boundary can be seen as a one-dimensional variable, which describes intonation patterns beyond the word level. In AM theory, phrasal intonation is described by two level tones (High and Low) (Ladd, 1996, p 60). The level tone is essentially a one-dimensional variable, too. The difference is that the variable in the present work is treated as a continuous variable, rather than a categorical variable. Thus, the variable may take continuous values rather than being allocated to one of the two categories.

For Tier Two, one F0 point from each syllable that does not bear primary stress is picked to represent F0 variations within a word. This tier concerns all the syllables without primary stress in multi-syllablic words. Two phonetic realities are considered: the absolute F0 value of the syllable and the F0 difference with respect to the anchor F0 determined in Tier One. Specifically, for each syllable, the MidF0 value of the pitch target and its difference from that of the syllable bearing primary stress are two entities of concern.

As stated above, the purpose of Tier Three is to fill in pitch movement details at the syllable level. The slope of an underlying pitch target (parameter a) is thus used to characterize such F0 variations, which can be superimposed on the overall F0 height of a syllable provided by Tier One and Tier Two.

In Tier Four either parameters β or λ in Eq. (4.2) in Chapter 4 or some alternative post-processing schemes can be used to model segmental effects. As mentioned in Chapter 4, empirical evidence has revealed that β and λ exhibit complex behaviors and are difficult to model. Hence the same interpolation scheme used in the previous chapter is employed. It is not anticipated that this simplification will cause serious problems as this tier is assumed to be less important in terms of intonation perception as discussed in Chapter 1. Practically speaking, in a unit-selection based speech synthesis system, segmental effects could already be included if an appropriate unit is selected.

5.1.3 Learning Parameter Values

The final step in building the current system is to determine the appropriate parameter values from input features. Since a data-driven approach is taken in this work, parameter values will be learned from a corpus using machine learning algorithms. Specifically, decision tree algorithm – CART and ensemble learning described in Chapter 2, were employed to train the system.

In forming the feature set, efforts were made to use features that are readily available in a TTS system. Therefore, some features were deliberately excluded, such as duration and some phrasal information, even though we recognized that these features were useful for generating more accurate intonation patterns. Of all the features discussed below, phrase break is probably the most difficult feature to obtain from text. To make the task simpler, only the major phrase break, also known as the intonational phrase break, was

considered. This feature can be predicted by systems such as the one described by Sun and Applebaum (2001). In general, the feature sets used in this work are quite simple, and most of them are commonly used in other studies.

In Tier One, the input features can be loosely classified into three groups:

- Syntactic, semantic, pragmatic features or other discourse information that determines pitch accent distribution: Currently, only part-of-speech is included in this group due to its availability.
- Positional features: word position in a sentence and in an intonational phrase.
- Other features: the vowel contained in the syllable, syllable stress and syllable position in a word. These features may not be as important as those in the first two groups with respect to predicting the overall intonation pattern, but they could be helpful in predicting minute F0 variations.

Three regression trees were designed based on these features or on features derived from them as illustrated in Figure 5.1. The input feature set for the first tree primarily includes the positional features. In the second tree, in addition to positional features, part-of-speech tags of current, previous and next word were used. In the third tree, the length of the context window was expanded by incorporating the part-of-speech of previous and next two words. The choice of which feature to include from the third groups in each tree was determined experimentally. The final F0 values were obtained by averaging the results predicted by three regression trees. It is worth noting that building trees with different feature sets represents an attempt to capture the specific intonation patterns in each tree.

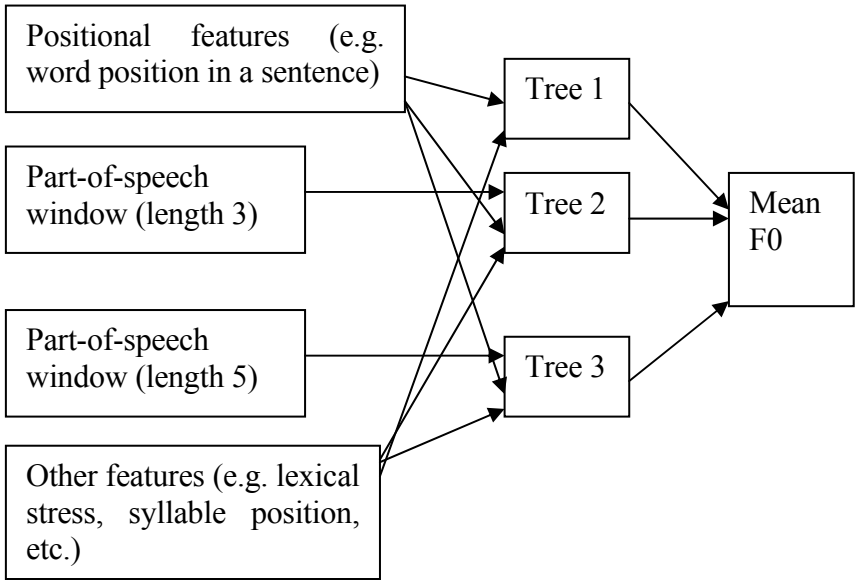


Figure 5.1: Block diagram for Tier One.

In Tier Two, the following features were collected for each syllable:

- Lexical stress (1 or 0) of the current, previous, and next syllable
- Syllable position in a word (initial, medial, final, and single) of the current, previous, and next syllable
- Number of syllables in the current word and previous word
- Position of the word in a sentence and an intonational phrase.

A significant difference in the feature set between Tier Two and Tier One is that in Tier Two more local context information is included while part-of-speech information is removed. As shown in Figure 5.2, two regression trees were built upon these input features. In the first tree, the MidF0 difference between the current syllable and the syllable bearing

primary stress in the same word was predicted. The second tree was trained to predict the absolute MidF0 value of the current syllable. The predicted values obtained from the first tree was converted to F0 values by adding these values to the corresponding anchor F0 values obtained from Tier One. Then the final F0 values were estimated by taking the average of reconstructed F0 from the first tree and the F0 predicted values from the second tree.

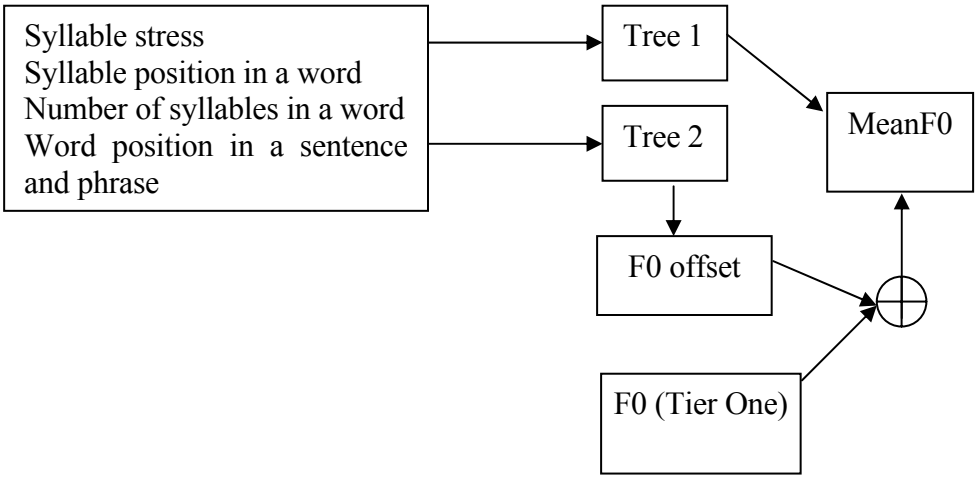


Figure 5.2: Block diagram for Tier Two.

The input features for Tier Three were derived from the following information:

- Vowel contained in the current syllable
- Final voiced consonant of the current syllable
- Lexical stress (1 or 0) of the current, previous, and next syllable
- Syllable position in a word (initial, medial, final, and single) of the current, previous, and next syllable

- Number of syllables in the current and previous word
- Part-of-speech of the current, previous, and next word
- Position of the syllable in an intonational phrase
- Position of the word in a sentence and an intonational phrase.

Two regression trees (Tree 1 and Tree 2) were built upon the input features and their outputs are subsequently averaged (see Figure 5.3). The main difference between Tree 1 and Tree 2 is whether part-of-speech is included in the feature set. This construction of feature set reflects the assumption that pitch movements at the syllable level are affected by factors at different levels.

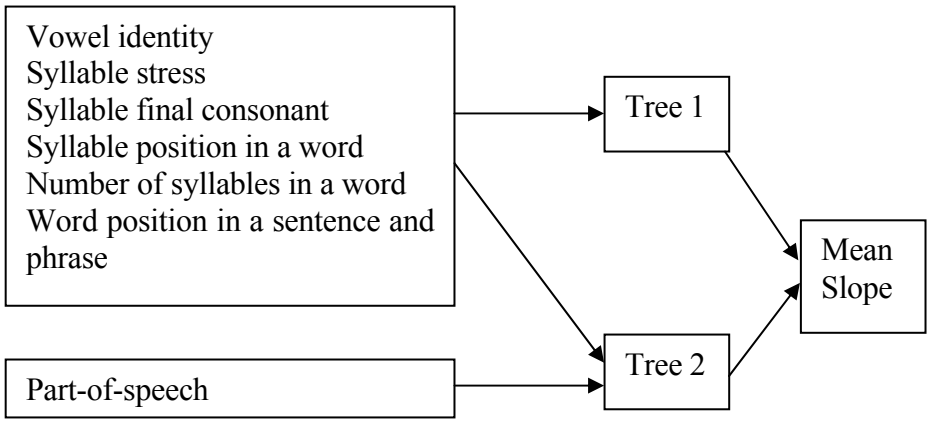


Figure 5.3: Block diagram for Tier Three.

The results of above learning processes are several binary decision trees. To generate F0 contours from an arbitrary text, additional procedures need to be taken to

collect features for each syllable. Figure 5.4 illustrates the complete process for generating F0 contours for a sentence.

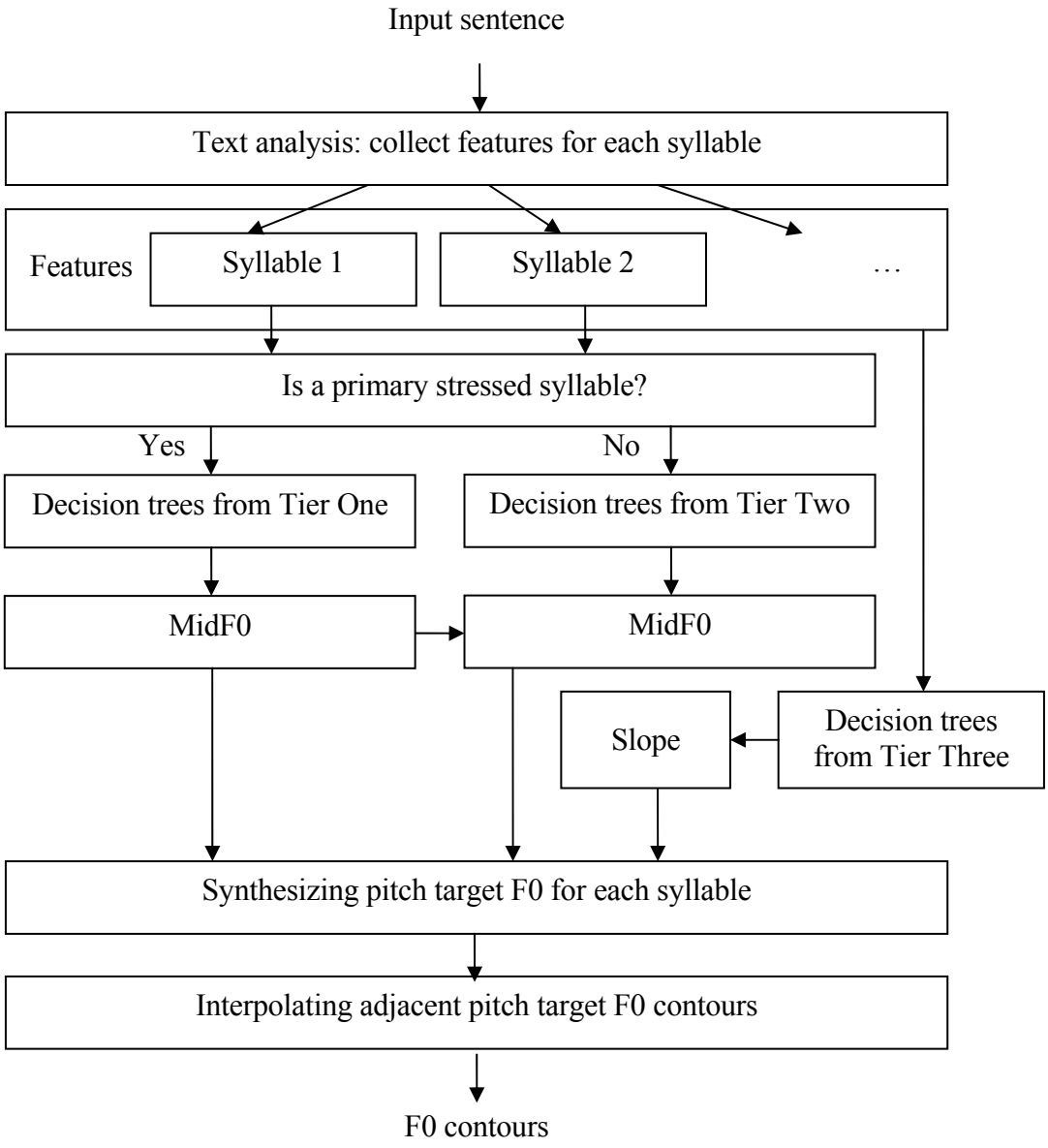


Figure 5.4: Block diagram for F0 generation with the present system.

5.2 Numerical Evaluation Results

To test the proposed system, experiments were conducted on the same training and testing sets used in section 4.3 of the previous chapter. The results were evaluated both objectively and subjectively. A review of objective results is presented here and a discussion of the subjective results is deferred to the next chapter. To compare the system's performance with that of other systems, two alternative systems were evaluated. The Three-Target model (Black and Hunt, 1996) represents a typical non-parametric approach. The Tilt model (Taylor, 2000) is an example of parametric models. The implementation details of the two systems are described in Appendices F and G. It is known that the same model can generate quite different results with different learning methods, training corpus, input features, etc. To minimize the effects of these factors, the Three-Target model and the Tilt model were implemented with:

- Same training and testing corpus
- Same decision tree program - WAGON
- Same smoothed F0 contours
- Similar feature set

Table 5.1 presents the evaluation results, which show that the proposed system produces very encouraging results even without ToBI labels (*RMSE* = 36.2; *r* = 0.66). The results are much better than those of the Tilt model and the Three-Target model.

	RMSE	*r*
Original F0 – Predicted F0	36.2	0.66
Model F0 – Predicted F0	35.3	0.67
Original F0 – Predicted F0 (Tilt model)	41.9	0.52
Original F0 – Predicted F0 (Three-Target model)	41.2	0.52

Table 5.1: F0 generation comparison between the proposed system, the Tilt model, and the Three-Target model.

Numerical results only give us a partial picture of how good the proposed system is. Since the ultimate goal of an F0 generation system is to synthesize perceptually natural intonation, a perceptual evaluation of synthetic intonation is indispensable. Such a test not only can examine the system performance, but can also help us gain insight into important intonation modeling issues. Due to the importance of a perceptual evaluation of the system, Chapter 6 will be devoted to this topic.

5.3 Discussion

The encouraging results show that the proposed approach provides a viable direction for modeling intonation. Several important aspects of the system are discussed below.

First, by describing intonation using a multi-tier approach, the system was built step by step, in a roughly hierarchical way. This approach allows emphasis to be placed on the more important components of intonation (e.g. Tier One in the model). This structure allows input features to be used more efficiently and allows the most appropriate model to be chosen for each tier. Using this structure, promising results were obtained without using accent labels. Many other approaches require pre-labeled pitch accents (e.g. Dusterhoff et

al., 1999) to predict intonation. The use of accents, however, may have limitations. (1) To train an accent prediction system, a pre-labeled database is necessary. Such a labeling process can be labor intensive and time-consuming and can thus limit the use of larger corpora. (2) The trained model may not be readily transferable across different domains, corpora, speaking styles, and speakers. (3) Labeling usually involves labelers' subjective interpretation of sentences. As a result, some acoustic units may be marked as accented due to their linguistic strength rather than to a strong acoustical manifestation. For example, a less familiar word to the listener may sound more like an accented word than a more familiar word, even if the less familiar word does not have a strong acoustic representation. In other words, human labeled accents usually reflect both linguistic effects and acoustic effects which makes to the prediction of pitch accent difficult. In the present study, it is assumed that underlying pitch targets reflect the intonation patterns that the speaker tries to convey and the listener hears. In other words, high-level linguistic information is mapped onto the acoustic domain through pitch height, pitch movement magnitude and direction. It should be noted that although pitch accent information derived from text is not emphasized in the proposed F0 generation system for text-to-speech, such information can be very useful in concept-to-speech (McKeown and Pan, 1999) and Prosody Markup languages (Huang et al., 2001; Kochanski and Shih, 2002) for speech synthesis. For those applications, an accurate pitch accent prediction system is highly desirable.

Second, by using a parametric form, the system can ensure the consistency and smoothness of the contour at the domain covered by the model, which is the syllable in the

present work. In non-parametric approaches (e.g. Black and Hunt, 1996; Buhmann et al., 2000; Traber, 1992), F0 contours are generated by connecting several target F0 values that are predicted either separately or together. Although problems can be alleviated by employing interpolation or post-smoothing techniques, those models have no inherent ability to prevent an abnormal F0 trajectory.

Third, compared with other parametric models (e.g. Dusterhoff et al., 1999; Mohler and Conkie, 1998; Taylor, 2000), the present parametric representation is very simple. Only two parameters need to be predicted in the present implementation. The simple parameter set used in this work reflects efforts to avoid using a complex model that may be less predictable and less efficient.

Finally, in the system two or three regression trees were built for each tier either by splitting the input feature set (Tier One and Tier Three) or by predicting different output values (Tier Two). This design has both linguistic and mathematical motivations. For example, from a linguistic perspective, different subsets of features were used in different trees in Tier One in order to model certain intonation patterns more accurately. In Tier Two, predicting different F0 forms reflects the assumption that the F0 values of unstressed syllables depend both on the input features and on the F0 of the syllable with primary stress within the same word. From a machine learning perspective, as described in Chapter 2, these methods are specific types of ensemble learning techniques, which have proved to be superior to single learning machine. However, the results of using ensemble learning for regression problems need to be treated with caution, especially when the outputs of a large

number of trees are averaged. It is possible that this averaging process is devastating even though improvements are achieved with respect to certain numerical measures. Therefore, more systematic evaluations are necessary for ultimate verification.

Chapter 6

Perceptual Evaluation of Synthetic Intonation

Intonation as a communicative function involves both production and perception. An intonation model without support from perception is not complete (Hirst et al., 2000). A vigorous analysis from a perception point of view not only will verify many assumptions made in the model but also will identify new issues. Consequently, a better intonation model could be developed. Practically speaking, perceptual evaluation of an F0 generation system constitutes an important component in developing speech synthesis systems. Different from speech recognition, where objective evaluation usually offers a sufficient solution, the nature of speech synthesis requires user listening evaluation as the ultimate standard. Objective measures, such as *RMSE* and correlation coefficient used in Chapter 5, can only be regarded as secondary criteria for evaluating synthetic intonation. Therefore, motivated by both theoretical and practical reasons, in this chapter, a formal perception experiment was designed and conducted to evaluate the proposed F0 generation system.

6.1 Related Research

Evaluating an intonation generation system subjectively is a challenging task. Although no systematic evaluation methods have been documented, several previous studies have

conducted perception evaluation test, which describe some general frameworks for evaluating synthetic intonation.

In Syrdal et al. (1998), three intonation models were compared: (1) a rule-based system (Jika et al., 1999); (2) the Tilt model trained with decision trees (Taylor, 2000; Dusterhoff et al., 1999); and (3) a Vector Quantized model based on a parametric model – PaIntE (Mohler and Conkie, 1998). Twelve test utterances were used, each less than 10 seconds long. These utterances contained two prosodic styles, i.e., news-reading and interactive prompt. The mean number of words of news-reading sentences was 17.3, with a range of 10-25 words. Prompt style utterances ranged from 12 to 19 words in length, with a mean of 15.7 words. The utterances were synthesized using the AT&T unit-selection based synthesizer within the Festival Speech Synthesis System (Black et al., 1998). Two synthesis techniques, the Harmonic plus Noise Model (HNM) (Schroeter et al., 1997) and Pitch-Synchronous Overlap and Add (PSOLA) technique (Moulines and Charpentier, 1990) were used.

During the test, the subject was asked to rate the quality of the synthesized utterances on a 5-point Mean Opinion Score (MOS) scale, in which 5=excellent, 4=good, 3=fair, 2=poor, and 1=bad. The results showed that the third system (a Vector Quantized model based on PaIntE) received the highest rating, whereas the Tilt model received the lowest. However, whether the differences were statistically significant was not reported.

In Ross and Ostendorf (1999), two perceptual tests were conducted using 15 sentences from the Boston University Radio Corpus (Ostendorf et al., 1995). They used

the 1994 AT&T TTS synthesizer with a female voice. In their test, three different versions of each sentence were presented to the listeners: (1) The dynamic system F0 model with hand-labeled prosodic markers (Ross and Ostendorf, 1999); (2) the hybrid target/filter F0 model (TTS default) (Anderson et al., 1984) with hand-labeled prosodic markers; and (3) the hybrid target/filter F0 model and predicted prosody labels. The subjects were asked to compare the naturalness of the three utterances by rating them on a scale of 1 to 5, with 1 representing the most natural. The results showed that the dynamical system model was rated the most natural whereas the third system was the least. The differences were significant at 0.01 level.

In Mixdorff and Jokisch (2001), a perception experiment was conducted to evaluate their proposed prosodic model for German. 12 sentences with length ranging from 12 to 44 syllables were used as the testing material. Subjects were asked to rate the utterances on a five-point scale. The authors noted that this grading approach could be less accurate than the pairwise comparison method. However, since there were too many stimuli (400), using pairwise comparison was unrealistic.

As discussed in Dusterhoff (2000), previous studies on subjective intonation evaluation can be classified into two categories, namely pairwise comparison and acceptability ranking. In the first technique, listeners are asked to compare two utterances with different F0 contours, which can be original or predicted by intonation models. Other aspects of the sentences should be kept equal. In the second technique, subjects' task is to rate an utterance on a scale, such as from 1 to 5. Among the above three studies, Syrdal et

al. (1998) and Mixdorff and Jokisch (2001) used the second technique, whereas Ross and Ostendorf (1999) adopted a scheme which was a combination of the two paradigms.

Both techniques suffer from some limitations. In a pairwise comparison, two utterances are compared without giving a rating score. The result can only tell us which one is the better of the two. Detailed information about the magnitude of the difference and their absolute naturalness is usually not available. It could be the case that both utterances are equally bad. On the other hand, in acceptability ranking, by rating an utterance in isolation without any reference, the inconsistency of a subject's judgment could weaken the power of the rating scores. Thus, in the current study, a hybrid paradigm similar to Ross and Ostendorf (1999) was adopted.

6.2 Experimental Questions

To examine its performance compared with other approaches, two additional systems were also evaluated. The Three-Target model (Black and Hunt, 1996) was selected as a typical non-parametric approach. The Tilt model (Taylor, 2000) was chosen as an example of a parametric model. The implementation of the two systems was described in the Appendices F and G.

By comparing with these models, this study aimed to examine if the proposed model had greater predictive power and produced more natural intonation by using a multi-tier structure and underlying pitch target as the phonetic representation of intonation. The

original intonation was used as the reference for evaluating the naturalness of the synthesized intonation.

One may ask whether using the original intonation as the *truth* is too limited since for a single sentence many versions of intonation are acceptable. It is believed that this is necessary within the current paradigm. The three F0 generation systems under investigation in the present study are all data-driven and trained on the same single-speaker corpus. The goal of these trainable systems is to learn the prosodic patterns of this particular speaker. Thus, as there are no other criteria, the closer to the original a system can get, the higher predictive power the system has. Obviously the above reasoning is based on the premise that the speaker's intonation patterns are relatively stable and predictable. That is, under similar linguistic conditions, the speaker will use similar intonation patterns. In Ross and Ostendorf (1999), the utterance with original F0 contour was not presented. This is plausible because in their comparison a rule-based system was included. Due to the nature of a rule-based system, forcing the subjects to evaluate its intonation with the intonation of a particular speaker as reference does not look fair.

6.3 Methods

6.3.1 Subjects

24 native speakers of American English (10 males and 14 females) between the ages of 18 and 42 (Mean = 26.1; Std = 6.45) participated in the experiment. Among the 24

participants, 22 were undergraduate or graduate students at Northwestern University, and two were staff members in the Department of Communication Sciences and Disorders. All subjects reported to have normal hearing, vision, and language ability, and had no or minimal experiences with synthetic speech. Prior to the experiment, subjects were asked to sign an informed consent form. Subjects were paid for their participation.

6.3.2 Stimuli/Apparatus

All the material used in the experiment was taken from the testing set of the Boston University Radio Corpus used in Chapter 5. Because of the news-reading style, the sentences tended to be very long. This could pose problems to the subjects due to memory limitation, which has been noted in previous studies. Therefore, some long sentences were split into shorter parts, which yielded 51 utterances in total (See Appendix H for the sentences). The length of these utterances ranged from 2.81 seconds to 7.44 seconds, and the average duration was 5.05 seconds, with a standard deviation of 1.42 seconds. These numbers were roughly in line with those in Syrdal et al. (1998), in which most of the test utterances were less than 10 seconds.

For each test sentence, the original recording was used as the reference. Three synthetic versions, with intonation predicted by three different systems were generated using the TD-PSOLA technique (Moulines and Charpentier, 1990). Specifically, the *PSOLA Resynthesis* function in Praat[1] was used to replace the original pitch contour with

[1] Praat software is available for free from http://www.praat.org

the generated ones. Similar procedures have also been employed by Mixdorff and Jokisch (2001) to generate stimuli for their intonation evaluation experiment. It has been observed that for the segment where the vocal epoch detection algorithm in Praat does not yield a value, the segment is kept untouched by Praat in the resynthesized sentence. Apparently this is not desirable, as synthesized F0 values are not used for these segments. Nevertheless, this situation applies to all three systems. So the final results should not be affected much since the main goal here is to compare these approaches. For each test utterance, a reference version was generated using TD-PSOLA with the original F0 contour.

The signals were generated by standard audio hardware and software on an Apple G4 machine. The utterances were presented to the subject via HD 265 linear Sennheiser headphones. Sound volume was adjusted to a comfortable level for each subject.

6.3.3 Procedure

All the tests were conducted in a sound-treated booth in the Speech Perception Laboratory at Northwestern University[2]. The whole experiment lasted for about 50 minutes.

The subject was seated comfortably in the booth facing a computer monitor with the headphones on. The test utterances were presented in a random order through a series

[2] The experiment can be accessed online at http://mel.speech.nwu.edu/prosody/scripts/default.asp

of web pages. The randomized sequence was generated dynamically each time the experiment program started. Each subject therefore had a different sequence of stimuli.

In each trial, three utterances synthesized using different intonation systems as well as the reference utterance were presented to the subject. The reference utterance was labeled as "reference natural speech", whereas the synthetic utterances were randomly labeled as "synthetic speech 1", "synthetic speech 2", and "synthetic speech 3", respectively. Subjects could play the stimuli as many times as they wanted. They were asked to rate the quality of the stimuli on the 5-point Mean Opinion Score (MOS) scale described earlier, in which 5=excellent, 4=good, 3=fair, 2=poor, and 1=bad. They were informed that the reference was the original recorded speech while the synthesized utterances were generated by different intonation modeling systems, and the goal of the experiment was to compare their naturalness with respect to the reference. Therefore, they should pay more attention to the intonation of the utterances rather than other qualities. After the experiment was completed, all the results were written into a file for later analysis.

6.4 Results

The experiment was a repeated measure design, where each subject participated in all the experimental conditions. The independent variable was intonation system, and the dependent variable was the Mean Opinion Score (MOS).

For each subject, the average MOS and the standard deviation for each system over all 51 utterances were computed. The results as presented in Table 6.1 show that the proposed system received the highest average rating (3.60), while the Tilt model received the lowest (3.04). When examining the subjects individually, all but one of them rated the current system as the best among the three. That one subject rated the proposed system second to the Three-Target model. Individual differences among the subjects seemed to be nontrivial. Some subjects tended to be very strict, while others tended to be lenient across the board. This can be seen from the overall standard deviations for each system averaged over all subjects (0.34, 0.48, and 0.51). Nevertheless, the present system presented the lowest standard deviation, indicating that it was rated more consistently.

A repeated measure Analysis of Variance (ANOVA) revealed that the differences among the three systems were significant ($F=59.47$, $p< 0.0001$). *Scheffe post hoc* tests were followed to examine pairwise differences. Significant differences at $\alpha=0.01$ level were observed between the present system and the Three-Target approach, and between the present system and the Tilt model. The difference between the Three-Target model and the Tilt model was also significant, but at 0.05 level. These results confirmed that the present approach was sound and could generate more natural intonation than the other two systems. As discussed in Chapter 2, one of the drawbacks of data-driven approaches is performance inconsistency. The overall standard deviations shown in Table 6.1 only provided a general trend of agreement among subjects. Therefore, to examine the consistency issue further, a more detailed analysis of the results would be very meaningful. A basic criterion used here

was that a more consistent system should yield a smaller standard deviation. For each subject, a standard deviation over the 51 sentences for each system was obtained. Then the mean standard deviation over all the subjects was calculated. The results are listed in Table 6.2.

Subject ID	The present system	The Three-Target model	The Tilt model
1	3.87	3.45	3.43
2	4.01	3.83	3.66
3	3.14	2.57	2.22
4	4.14	3.71	3.47
5	3.48	2.50	2.85
6	3.61	3.31	3.08
7	3.31	3.37	2.75
8	3.43	2.84	2.69
9	3.46	3.28	3.25
10	3.43	2.65	2.57
11	3.46	2.69	2.77
12	3.29	2.45	2.02
13	3.63	2.87	2.70
14	3.22	2.78	2.69
15	3.85	3.30	3.29
16	2.92	2.78	2.55
17	3.63	3.61	3.12
18	4.48	4.37	4.28
19	3.71	3.42	3.49
20	3.78	3.67	3.44
21	3.49	3.22	2.91
22	3.59	3.18	2.96
23	3.83	3.41	3.25
24	3.66	3.51	3.64
Mean	**3.60**	**3.20**	**3.04**
Std	**0.34**	**0.48**	**0.51**

Table 6.1: The average Mean Opinion Score for the three F0 generation systems.

Subject ID	The present system	The Three-Target model	The Tilt model
1	1.17	1.37	1.12
2	1.06	1.09	1.03
3	0.84	0.90	0.99
4	1.24	1.31	1.16
5	0.65	0.61	0.67
6	0.84	0.98	0.96
7	0.95	1.09	0.98
8	0.96	0.86	1.01
9	1.00	0.90	1.01
10	0.27	0.32	0.34
11	0.69	0.68	0.78
12	0.68	0.77	0.76
13	1.00	1.04	1.13
14	0.54	0.56	0.51
15	0.89	0.84	0.89
16	0.55	0.44	0.53
17	0.69	0.83	0.96
18	1.06	1.21	1.12
19	0.58	0.62	0.71
20	0.53	0.56	0.60
21	0.80	1.19	0.90
22	0.94	1.10	1.12
23	1.03	0.94	0.97
24	0.96	0.84	0.77
Mean	**0.83**	**0.88**	**0.88**

Table 6.2: The consistency of the three F0 generation systems measured by the mean standard deviation over the 51 sentences.

Table 6.2 shows that the present system is slightly better in terms of consistency averaged over all subjects. Again, the subjects exhibited quite different behaviors. Subject 10 had the smallest standard deviations, the numbers being 0.27, 0.32, and 0.34, respectively. Subject 4 had the largest standard deviations of 1.24, 1.31, and 1.16,

respectively. For convenience, it is often desirable to use one score to represent both the absolute MOS and the consistency of the system. Thus, a weighted MOS for each subject was derived by dividing the raw MOS in Table 6.1 by the corresponding standard deviation in Table 6.2. Table 6.3 shows that in terms of the weighted MOS, the rank sequence of the three systems remains the same as before. However, the magnitude of difference has changed with the current system standing out much better than the other two.

Since the three systems were all trained and tested with the same corpora, the same decision tree program, and similar feature sets, it was expected that they should exhibit some kind of similarities. One way of showing this could be to examine how the three systems behaved for each individual sentence. Hence, the MOS of all the subjects were averaged for each sentence. A scatter plot of these scores is shown in Figure 6.1. The plot in Figure 6.1 reveals that the three systems indeed show some similar trends. That is, when the first system received a higher rating score for a particular sentence, most likely the other two were also rated higher. Likewise, if the first system got a low rating, the other two were also low. Figure 6.1 also clearly shows that the present system outperforms the other two because most of its points are in the upper half of the plot.

Subject ID	The present system	Three-Target model	Tilt model
1	3.32	2.51	3.06
2	3.80	3.53	3.57
3	3.72	2.85	2.23
4	3.33	2.83	2.98
5	5.31	4.08	4.24
6	4.31	3.37	3.21
7	3.48	3.11	2.82
8	3.58	3.29	2.65
9	3.46	3.66	3.21
10	12.48	8.23	7.52
11	4.97	3.98	3.56
12	4.86	3.19	2.65
13	3.62	2.77	2.38
14	6.01	4.97	5.25
15	4.33	3.92	3.70
16	5.30	6.26	4.77
17	5.27	4.32	3.24
18	4.25	3.62	3.83
19	6.38	5.50	4.92
20	7.18	6.50	5.75
21	4.35	2.71	3.23
22	3.80	2.88	2.64
23	3.73	3.61	3.35
24	3.80	4.17	4.71
Mean	**4.78**	**3.99**	**3.73**

Table 6.3: Weighted rating scores for the three F0 generation systems by dividing the Mean Opinion Score and the corresponding standard deviation for each subject.

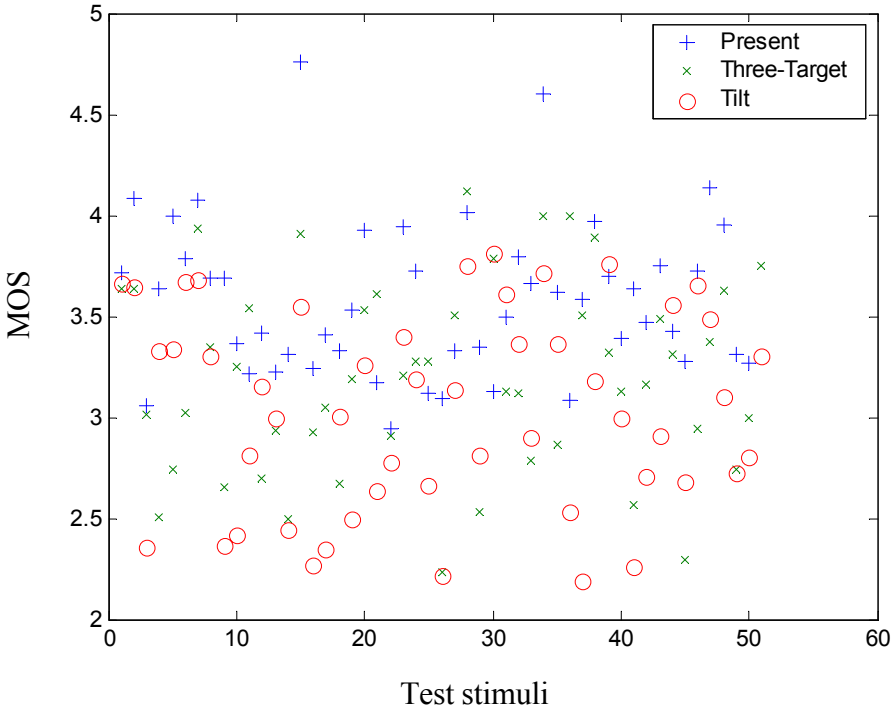


Figure 6.1: Averaged Mean Opinion Score for the three systems for each sentence.

6.5 Discussion

The perception test yielded some interesting results. First, it showed that the proposed system was favored by the subjects over the other two systems. This confirmed that the proposed system could generate more natural intonation. Specifically, the utilization of a hierarchical structure, underlying pitch targets, and ensemble trees, was effective. The results also indicated that synthesizing natural intonation without pre-labeled pitch accents

was possible. This may open the door to using very large corpus, which presumably can cover more prosodic patterns and improve the performance of a data-driven system.

It was not without surprise to see the Three-Target model rank the second. As described in Chapter 2, the Three-Target model is a simple non-parametric approach. The Tilt model, on the other hand, received the lowest score. This could partially be due to the present implementation, which may not be as successful as that in Dusterhoff et al. (1999). However, it was suspected that some inherent limitations of the model might be more responsible for the results. One of the issues could be that in the Tilt model each intonation event is not limited by syllable boundary, that is, it describes intonation spanning several syllables. Although this is consistent with autosegmental phonology and gains flexibility in intonation description (Taylor, 2000), it may pose problems in practice. Specifically, during training and testing of such a data-driven system, an event usually needs to be aligned with a syllable bearing primary stress. Consequently the information derived from this syllable is used as input features to predict the parameters of this event, which, however, may in fact describe intonation patterns of several syllables. This mismatch could cause problems for statistical learning systems, in that the input information may be insufficient or even incorrect. Another possible reason might be the uneven distribution of parameters. Each intonation event is described by five tilt parameters, which are capable of encoding quite detailed information. The scale of the current corpus seems to be not large enough to let the learning algorithm model all the five parameters successfully. In contrast, no parameters are used to describe F0 contours between intonation events besides interpolation. This

seems to underutilize the information in the corpus. Indeed, informal observation has found some unnatural F0 contours resulting from simple interpolations. As discussed in Chapter 1, this shows that meaningful intonation information may exist between intonation events (Chen and Xu, 2002; Ladd and Shepman, in press).

Although the test was not designed to evaluate the F0 smoothness of the three systems, informal observation has revealed that the proposed model normally generated smoother F0 trajectories than the Three-Target model. This smoothness produced better segmental sound qualities. As mentioned earlier, the Three-Target model is a typical non-parametric approach. Thus, it shows that a proper parametric model of local pitch movement is indeed important for better synthetic intonation.

It is worth mentioning that when using original pitch accent labels, both the present system and the Three-Target model can do a reasonably good job. Although the sentences synthesized with original pitch accents were not included in the perception test, informal listening by several speech researchers confirmed the good quality. Interestingly, the numerical results shown in Table 4.7 in Chapter 4 still exhibit a vast difference between the original F0 and the synthesized F0. This supports the idea discussed in Chapter 1 that although intonation is complex, it contains redundant information. To generate natural intonation, it is not necessary to model all the details in F0 contours, instead, modeling the most relevant components is more important.

The results of the perception test not only demonstrate that the proposed system can generate more natural intonation than other two systems, but also confirm the importance

of perception study in intonation research. Perception studies can provide very valuable information, which is critical for understanding intonation production mechanisms and building better intonation models. Therefore, more work on intonation perception is undoubtedly called for in future research.

Chapter 7

Conclusions

7.1 Summary

The main body of the research in this thesis concerns intonation modeling. The work on the determination, analysis, and synthesis of fundamental frequency has resulted in several important findings regarding the issues raised in Chapter 1. The rest of this section concludes these findings and points out the contributions which distinguishes the present work from others.

7.1.1 A Multi-Tier Hierarchical Representation of Intonation

This thesis shows the importance of a hierarchical representation of intonation. A new F0 generation system was proposed which accepts linguistic features derived from text and outputs continuous F0 contours. This system employs a multi-tier hierarchical structure to represent intonation and trains each level statistically using ensemble decision trees. A hierarchical structure has shown to be effective in representing intonation in many phonological systems. Such an organization allows better explorations of intonation patterns at different scales. In the present system, the general tendency of different levels is related to different scales of effect in the contour. For example, the choice of pitch accent

sequence (phonological) operates at the phrasal level, and shows up in the general placement of the F0 contour for one word versus a later word. Segmental effects (plainly phonetic) are subsyllabic and be treated as an adjustment on a more idealized contour. In doing so, the input feature can be utilized more effectively.

7.1.2 Relevant Intonation Components

This thesis indicates that certain intonation components are more relevant than has been recognized. Specifically, the proposed F0 generation system focuses more on the underlying pitch targets instead of treating everything in the surface F0 contours equally. The encouraging objective and subjective evaluation results demonstrate that the underlying pitch targets capture a majority amount of intonation information. What is left can be simulated reasonably well by simple interpolations. Such decomposition makes the system different from many previous approaches. The importance of modeling underlying pitch targets is partially confirmed by less favorable results obtained using the Three-Target model, which models the surface F0 contours. Mixing everything together poses difficulties in selecting an appropriate feature set. Simply including all the features in one set results in underutilization of certain features.

The F0 generation system described in Chapter 5 models intonation for every syllable, which is similar to the Three-Target model and some previous approaches in this regard. It is different from the event-based systems, such as the Tilt model, which only models intonation events. As revealed by the evaluations, using simple interpolations for

F0 contours between intonation events seems to miss important information sometimes, which leads to unnatural synthetic intonations. This observation supports some other studies' results (Chen and Xu, 2002; Ladd and Shepman, in press).

7.1.3 A Parametric Representation of F0

This thesis employs underlying pitch target as the parametric representation of F0 contours. Chapter 4 presented a robust underlying pitch target analysis procedure. Specifically, a simple linear target is used to represent most significant intonation component in a syllable. The target is represented by its middle F0 value (MidF0) and slope. These two parameters are shown to be statistically independent with a correlation coefficient of -0.01, which is critical to build a compact model. Automatic prediction of underlying pitch targets from linguistic inputs and predicting pitch accents using underlying pitch targets as the input were performed. The results indicate that the parameters are linguistically predictable. Two parameters correspond to intonation at different levels. The MidF0 can be viewed as the phonetic realization of a level tone – a phonological target, while the slope can be viewed as the phonetic realization of a pitch movement – a phonetic target. Thus, this thesis assumes that both level tones and pitch movements exist but are implemented at different levels. Such an organization has been realized in the proposed F0 generation system and has yielded promising results.

7.1.4 Decision Tree, Ensemble Learning and Adaptability

In this thesis, the mapping between input linguistic features and F0 contours is derived automatically by training the system statistically on a corpus. CART, a decision tree algorithm, and ensemble learning were employed as the machine learning algorithms. A statistically trainable system generally can produce more natural intonation when appropriate amount of data are available. It is also more adaptable to new speaking styles or languages. The multi-tier F0 generation system trained with multiple decision trees also distinguishes itself from previous multi-level phonological systems. The experiments conducted in this thesis demonstrate that decision tree algorithms have many advantages over other approaches for F0 generation. Ensemble learning techniques, such as Bagging and AdaBoost, consistently outperform a single CART, which prove themselves as suitable methods for intonation modeling and other prosodic modeling problems.

7.1.5 Extraction of Proper Intonation Contours

Chapter 3 described a novel pitch determination algorithm and its evaluation results. The motivation of the algorithm was to find some solutions to alternate cycles in speech, which is notoriously difficult to deal with in intonation research. "Intonation contours" should be continuous and smooth. Alternate cycles in speech signals often result in abrupt pitch drops if using conventional pitch determination algorithms, which is unsuitable for intonation research. If the segment with the undesirable low pitch is long, normal smoothing can not eliminate the wrong points. The proposed pitch determination algorithm tackles the

problem by handling alternate cycles in a unique way. First, alternate cycles are viewed as the results of amplitude modulation or frequency modulation. The variation patterns of alternate cycles are represented by subharmonics in the frequency domain. A new parameter named subharmonic-to-harmonic ratio (SHR), describing the amplitude ratio between the subharmonics and harmonics, is defined to quantify variations of alternate cycles in the time domain. The main body of this algorithm is to compute the subharmonic-to-harmonic ratio. And the pitch is determined by evaluating the ratio against certain threshold values. The threshold values are derived from a pitch perception experiment. Evaluation results on two public available databases revealed that the proposed method outperformed several state-of-the-art pitch determination algorithms. Toward the so-called "intonation contours", the algorithm suppresses the subharmonics by increasing the SHR threshold. Besides intonation modeling, another potential application of this algorithm is for voice quality research. Preliminary voice quality analysis using subharmonic-to-harmonic ratio is shown in Appendix B.

7.2 Limitations and Future Directions

Although this thesis has achieved some success in determining pitch and synthesizing natural intonation, some caveats need to be mentioned. The following sections address these issues and also point out potential directions for future research.

7.2.1 Single Target, Multiple Targets, or Composite Target?

In the phonological systems described in Chapter 2 (e.g. Pierrehumbert, 1980), there is not a distinct layer for phonetic representation of intonation. The phonetic realization rules make the system flexible enough to map various phonological entities to F0 values, which could implement one target, multiple targets, or a composite target in each syllable. On the other hand, a phonetic model employs an explicit function to parameterize contours. The capability of including one or more targets depends on the definition of the model.

The current parametric representation of intonation allows only one pitch target in each syllable. A further constraint posed by the system is the target is in the form of a simple linear function rather than a complex one, such as a quadratic function. Such constraints make the system efficient and yield some promising results on the news-reading corpus. It should be kept in mind that news-reading corpora only cover a particular subset of intonation patterns. Thus testing the system on databases with more speaking styles is certainly needed in future work. Although the system is statistically trainable and adaptable to new speaking styles, the constraints mentioned above may be too stringent and limit the coverage of the system for synthesizing more expressive speech. For example, in Mandarin, using a simple linear target to represent the third tone can be problematic. Although the final rise is often negligible in connected speech, it is clearly a meaningful target when the tone is spoken in isolation. Therefore, the whole concave curve might be the target, which cannot be represented by a linear function. In English, similar cases can happen particularly when the duration of a single-syllable word is long. Speakers can

voluntarily implement more than one targets or a complex quadratic target in order to articulate expressive speech. In a phonological system (e.g. ToBI) this can be realized by using as a bi-tonal or even tri-tonal pitch accent. In the present approach, simply allowing more pitch targets in each syllable or using a complex target will harm the consistency and efficiency gained by the system. A modified procedure was developed at the end of Chapter 4, which allows the target to be finished before the syllable boundary. If there are two targets in a syllable, this procedure most likely will throw away the second one. Such a solution is obviously not perfect but may work if the second target is not perceptually significant. While the work is still in its preliminary form, it could be elaborated by future work. In any case, how to develop a framework that can accommodate these issues constitutes an important step towards a more realistic intonation model for English.

7.2.2 Application of Subharmonic-to-Harmonic Ratio

The pitch determination algorithm introduced in Chapter 3 yields encouraging results on F0 tracing. In order to make the algorithm more useful for intonation or voice quality research, more extensive experiments are needed. For example, performing more comprehensive analysis on subharmonic-to-harmonic ratio for a particular speaker could be useful in developing a voice profile for this speaker. Such a profile has potential applications in clinical work (Titze, 1995) as well as other fields like speaker recognition. Appendix B also laid out an initial comparison between subharmonic-to-harmonic ratio and harmonic-to-noise ratio. More comprehensive studies are needed to gain deeper understanding on this

issue. Another application of pitch determination algorithm is extracting F0 values for speech recognition. This often requires the algorithm be noise robust (Wang and Seneff, 2000). Therefore, testing and improving the current algorithm under noisy speech condition is another potential direction for future work.

Bibliography

Anderson, M. D., Pierrehumbert, J. B., and Liberman, M. Y. (1984). Synthesis by rule of English intonation patterns. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pp 281-284.

Bagshaw, P. C. (1994). *Automatic prosody analysis.* University of Edinburgh, Scotland, UK.

Bakiri, G., Dietterich, T. G. (2001). Achieving high-accuracy text-to-speech with machine learning, In B. Damper (Ed.) *Data Mining in Speech Synthesis*. Kluwer Academic Publishers, Boston, MA.

Beckman M. and Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonol. Yearbook*, vol. 3, pp. 255–309.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press, UK.

Black, A. and Hunt, A. (1996). Generating F0 contours from ToBI labels using linear regression. *Proc. of ICSLP*, Philadelphia, Penn.

Black, A., Taylor, P., and Caley, R. (1998). *The festival speech synthesis system*. http://www.cstr.ed.ac.uk/projects/festival.html.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences 17*, pp. 97-110.

Botinis, A., Granström, B. and Möbius, B. (2001). Developments and paradigms in intonation research. *Speech Commun*. *33*, 263-296.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *26*(2), 123-1401.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, P. J. (1984). *Classification and regression trees* (Wadsworth and Brooks).

Buhmann, J., Vereecken, H., Fackrell, J., Martens, J. and Coile, B. (2000). Data driven intonation modelling of 6 languages, *Proc. of ICSLP*, Beijing, China, 3, pp. 179-182.

Campbell, W. N. and Isard, S. D. (1991). Segment durations in a syllabic frame, *J. Phonetics 19*, 37–47.

Chen, S. H., Hwang, S. H., and Wang, Y. R. (1998). An RNN-based prosodic information synthesizer for Mandarin text-to-speech, *IEEE Trans. Speech and Audio Proc*., 6(3):226-239.

Chen, Y. and Xu, Y. (2002). Pitch target of neutral tone in standard Chinese. *presented at the 8th Laboratory Phonology*, New Haven, Connecticut.

Cherkauer, K.J. (1996) Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In Chan, P. (Ed.) *Working Notes of the AAAI Workshop on Intergrating Multiple Learned Models*, pp. 15-21.

Conkie, A., Riccardi, G. and Rose, R. C. (1999). Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events, *Proceedings of EUROSPEECH '99*, Budapest, Hungary, 1, pp. 523-526.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge University Press. Cambridge.

d'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9, 257-288.

Dietterich, T. G. (2002). Ensemble learning. In *The Handbook of Brain Theory and Neural Networks, Second edition,* (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press.

Drucker, H. (1997). Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 107-115. Morgan Kaufmann.

Dusterhoff, K.E. (2000). *Synthesizing fundamental frequency using models automatically trained from data*, Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

Dusterhoff, K. E., Black, A. W. and Taylor, P. A. (1999). Using decision trees within the tilt intonation model to predict f0 contours. *Proceedings of Eurospeech'99*, Budapest, Hungary.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, vol. 14, pp. 179-211.

Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119-139.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech,* edited by P. F. MacNeilage (Springer-Verlag, New York) pp. 39-55.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour, In: Fujimura, O. (Ed.), *Vocal Physiology: Voice Production, Mechanisms and Functions*. Raven, New York, pp. 347-355.

Gold, B., and Rabiner, L.R. (1969). Parallel processing technique for estimating pitch period of speech in the time domain. *J. Acoust. Soc. Am.*, 46(2, part 2):442-448.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology* (Blackwell Publishers, Oxford).

Goubanova, O. (2001). Predicting segmental duration using Bayesian belief networks, *Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis,* Perthshire, Scotland, pp. 47-51.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* second edition Prentice-Hall, Upper Saddle River, N J.

Hermes, D. J. (1988). Measurement of pitch by subharmonic summation. *J. Acoust. Soc. Am*., 83, 257-264.

Hermes, D. J. (1998). Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41:73-82.

Hess, W. J. (1991). Pitch and voicing determination. in *Advances in Speech Signal Processing*, S. F. a. M. M. Sondhi, Ed. New York, NY: Marcel Dekker, Inc., pp. 3-48.

Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text, *Artificial Intelligence* 63, 305-340.

Hirst, D.J., Di Cristo, A. & Espesser, R. (2000). Levels of representation and levels of analysis for intonation. in M. Horne (ed) *Prosody : Theory and Experiment Studies Presented to Gösta Bruce*. (Kluwer, Dordrecht).

Huang, X., et al. (1996). Whistler: A trainable text-to-speech system, *Proceedings of the Fourth International Conference on Spoken Language Processing*, (IEEE, Philadelphia, USA), 4, pp. 2387-2390.

Huang, X., et al. (1997). Recent improvements on microsoft's trainable text-to-speech system – whistler. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, (IEEE, Munich, Germany), 2, pp. 959-963.

Huang, X., Acero, A., and Hon, H-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, NJ.

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

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J. Acoust. Soc. Am., 59*(5), 1208-1221.

Klatt, D. M. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers, *J. Acoust. Soc. Am. 87*, 820-857.

Kochanski, G. and Shih, C. (2002). Prosody modelling with soft templates. *Speech Communication*, in print.

Koehn, P., Abney, S., Hirschberg, J., and Collins, M. (2000). Improving intonational phrasing with syntactic information. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol 3, pp. 1289-1290, Istanbul.

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

Ladd, R. & A. Shepman. (in press). Sagging transitions between high accent peaks in English: experimental evidence. *Journal of Phonetics*.

Lee, S. and Oh, Y-H, (2001). Tree-based modeling of intonation. *Computer Speech and Language*, Vol. 15, No. 1, pp. 75-98.

McKeown, K. and Pan, S. (1999). Prosody modeling in concept-to-speech generation: methodological issues. *Philosophical Transactions of the Royal Society, Series A.*

Medan, Y., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signals. *IEEE Trans. ASSP*, 39:40-48.

Meron, J. (2002). Applying fallback to prosodic unit selection from a small imitation database. *Proc of ICSLP2002*, Denver, Colorado, pp. 2093-2096.

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings of ICASSP 2000*, vol. 3, pages 1281-1284, Istanbul, Turkey.

Mixdorff, H. and O. Jokisch (2001). Implementing and evaluating an integrated approach to modeling German prosody. In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perth Atholl Palace Hotel, Scotland, pp. 211-216.

Mohler, G. and Conkie, A. (1998). Parametric modeling of intonation using vector quantization, *Proc. of Third International Workshop on Speech Synthesis*.

Monaghan, A. I. C. (1992). Heuristic strategies for the higher-level analysis of unrestricted text, In G. Bailey, C. Benoit, and T. R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp. 143-161. Elsevier Science Publishers B. V.

Moore, B.C.J. (1989). *An introduction to the psychology of hearing*. 3rd edition, San Diego, CA: Academic Press.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453-467.

Muller, A.F. and Hoffmann, R. (2001). A neural network model and a hybrid approach for accent label prediction. *Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis,* Perthshire, Scotland.

Murphy, P. J. (1999). Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *J. Acoust. Soc. Am*. 105, 2866-2881.

Niemann, H., Nöth, E., Kiebling, A., Kompe, R., Batliner, A., (1997). Prosodic Processing and its use in Verbmobil, *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. München, Germany, pp. 75-78.

Noll, A.M. (1967). Cepstrum pitch determination, *J. Acoust. Soc. Am.*, 41(2):293-309.

O'Shaughnessy D. (2000). *Speech commucation: human and machine*, 2nd edition, Piscataway, NJ: IEEE Press.

Ostendorf, M., Price, P. J. and Shattuck-Hufnagel, S. (1995) The Boston University Radio News Corpus, Boston University Technical Report No. ECS-95-001.

Ostendorf, M. and Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary locations. *Computational Linguistics*, 20 (1), pp. 27-54.

Phillips, M.S. (1985). A feature-based time domain pitch tracker, *J. Acoust. Soc. Am.*, 77:S9-S10 (A).

Pierrehumbert, J. (1980). *The phonology and phonetics of english intonation*, Ph.D. dissertation, MIT, Cambridge, MA.

Pierrehumbert, J. (1981). Synthesizing intonation, *J. Acoust. Soc. Am.* 70, 985-995.

Pierrehumbert, J. and Beckman, M. (1988). *Japanese Tone Structure*. MIT Press. Cambridge, Mass.

Pierrehumbert, J. (1999). Prosody and Intonation, *MIT Encyclopedia of Cognitive Science*.

Plante, F., Meyer, G. and Ainsworth, W.A. (1995). A pitch extraction reference database, *Eurospeech'95*, Madrid, Spain, pp.837 –840.

Rapp, S. (1998). Automatic labelling of german prosody, *Proceedings of 5th Int. Conf. on Spoken Language Processing*, Sydney, Australia.

Ratkowsky, D. A. (1990). *Handbook of nonlinear regression models* (Marcel Dekker, Inc., New York).

Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers, *Journal of Phonetics*, Vol. 29, No. 4, pp. 407-429

Riley, M. D. (1992) Tree-based modeling of segmental durations. In G. Bailey, C. Benoit, and T. R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp 265-273. Elsevier Science Publishers B. V..

Ross, K. and Ostendorf, M. (1995). A dynamical system model for recognising intonation patterns, *Proc. of Eurospeech*, Madrid, pp. 993–996.

Ross, K. and Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis, *Computer, Speech and Language* 10, 155-185.

Ross, K. and Ostendorf, M. (1999). A dynamical system model for generating fundamental frequency for speech synthesis, *IEEE Trans. Speech and Audio Proc*. 7, 295-309.

Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamental frequency measurement, *J. Acoust. Soc. Am.*, vol. 43, pp. 829-834.

Schroeter, J., Stylianou, Y., and Dutoit, T. (1997). Diphones concatenation using a harmonic plus noise model of speech, *Proc. Eurospeech97*.

Secrest, B.G., and Doddington, G.R. (1983). An integrated pitch tracking algorithm for speech systems, In *Proc. IEEE ICASSP,* 1352-1355, Boston.

Silverman, K., (1987). *The Structure and Processing of Fundamental Frequency Contours*. Ph.D. Dissertation, University of Cambridge, Cambridge.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992). Tobi: A standard for labelling english prosody, *Proceedings of ICSLP 92*, Banff, Alberta), pp. 867-870.]

Sun, X. (2001). Predicting underlying pitch targets for intonation modeling, *Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis,* Perthshire, Scotland, pp. 143-147.

Sun, X. (2002a). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio, *Proc. of ICASSP*, Orlando, Florida.

Sun, X., (2002b). F0 generation for speech synthesis using a multi-tier approach, *Proc of ICSLP2002*, Denver, Colorado, pp. 2077-2080.

Sun, X. (2002c). Pitch accent prediction using ensemble machine learning, *Proc of ICSLP2002*, Denver, Colorado, pp. 953-956.

Sun, X. and Applebaum, T.H. (2001). Intonational phrase break prediction using decision tree and n-gram model, *Proc. of Eurospeech2001,* Aalborg, Denmark, Sept. 3-7, Vol 1, pp. 537-540.

Sun, X. and Xu, Y. (2002). Perceived pitch of synthesized voice with alternate cycles, *Journal of Voice*, in print.

Svec, J. G., Schutte, H. K., and Miller, D. G. (1996). A subharmonic vibratory pattern in normal vocal folds, *J. Speech Hear Res.* 39(1):135-43.

Syrdal, A., Mohler, G., Dusterhoff, K., Conkie, A., and Black, A.W. (1998). Three methods of intonation modeling, In *Proceedings 3rd ESCA Workshop on Speech Synthesis*, pp. 305-310.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds.: Elsevier Science, pp. 495 –518.

Tao, J. and Cai, L. (2002). Clustering and feature learning based f0 prediction for Chinese speech synthesis, *Proc of ICSLP2002*, Denver, Colorado, pp. 2097-2100.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model, *J. Acoust. Soc. Am.* 107, 1697-1714.

Taylor, P., and Black, A. (1998) Assigning phrase breaks from part of speech sequences, *Computer Speech and Language*. Vol. 12, pp. 99-117.

Taylor, P., Black, A., and Caley, R. (1999). *Introduction to the Edinburgh Speech Tools*. http://www.cstr.ed.ac.uk/ projects/speech_tools/.

't Hart, J., Collier, R., Cohen, A., (1990). *A perceptual study of intonation.* Cambridge University Press, Cambridge.

Titze, I. R. (1994). *Principles of Voice Production* (Prentice-Hall, Inc., Englewood Cliffs, NJ).

Titze, I. R. (1995). *Workshop on Acoustic Voice Analysis- Summary Statement* (National Center for Voice and Speech, Denver).

Traber, C. (1992). F0 generation with a database of natural f0 patterns and with a neural network, In *Talking machines: Theories, models, and designs,* edited by C. Benoit, T. R. Wawallis and G. Bailey (Elsevier Science, Amsterdam,The Netherlands: pp. 287-304.

van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8, 95-128, 1994

van Santen, J. P. H. and Hirschberg, J. (1994). Segmental effects on timing and height of pitch contours. In *Proc. Of ICSLP*, vol. 2, pp 719-722, Yokohama.

Wang, C., and Seneff, S., (2000). Robust Pitch Tracking for Prosodic Modeling in Telephone Speech, In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey.

Wang, M. Q. and Hirschberg, J. (1992). Automatic classification of intonational phrasing boundaries. *Computer Speech and Language*, 6 (2), pp. 175 - 196.

Willems, N., Collier, R., t'Hart, J., (1988). A synthesis scheme for British English intonation, *J. Acoust. Soc. Am. 84*, 1250-1261.

Xu, C. X., Xu, Y. and Luo, L.-S. (1999). A pitch target approximation model for f0 contours in Mandarin, *Proceedings of The 14th International Congress of Phonetic Sciences*, San Francisco), pp. 2359-2362.

Xu, C. X., Xu, Y. and Sun, X. (2002). Consonant perturbation on F0 in English and Mandarin, *Presented at the 9th Meeting of ICPLA*, Hong Kong.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *J. Phonetics 27*, 55-105.

Xu, Y. (2001). Sources of tonal variations in connected speech, *Journal of Chinese Linguistics monograph series #17*: 1-31.

Xu, Y, and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech, *J. Acoust. Soc. Am.,* Vol. 111, No. 3, pp. 1399-1413.

Xu, Y. and Wang, Q. E. (1997). What can tone studies tell us about intonation? *Intonation: Theory, Models and Applications*, *Proceedings of an ESCA Workshop*. A. Botinis, G. Kouroupetroglou and G. Carayannis. European Speech Communication Association, Athens, Greece: 337-340.

Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese, *Speech Communication 33*, 319-337.

Ying, G. S., Jamieson, L. H., and Mitchell, C. D. (1996). A probabilistic approach to AMDF pitch detection, *Proceedings of the 1996 International Conference on Spoken Language Processing*, Philadelphia, PA, Oct. 1201-1204.

Yumoto, E., Gould, W. J., and Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness, *J. Acoust. Soc. Am.* 71, 1544–1549.

Appendices

Appendix A

AdaBoost Algorithms

Input: sequence of N training examples $((x_1, y_1), \ldots, (x_N, y_N))$

Initialize weight distribution $W_i=1/N$, where $i=1,\ldots, N$.

Do for t=1, … , T where T specifies the total number of iterations

1. Train classifier using weight distribution W_i
2. Get back a hypothesis $h_t : X \rightarrow Y$
3. Calculate the error of h_t:

$$\varepsilon_t = \sum_{i=1}^{N} p_i^t \left|\text{sgn}[h_t(x_i) - y_i]\right|$$

$$\text{where } \text{sgn}(x) = \begin{cases} 1, x > 0 \\ 0, x = 0 \\ -1, x < 0 \end{cases}$$

if $\varepsilon_t > 0.5$, then set $T = t - 1$ and abort loop.

4. Set $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$
5. Set the new weights vector to be

$$w_i^{t+1} = w_i^t \beta_t^{1-|\text{sgn}[h_t(x_i)-y_i|}$$

Output the hypothesis

$$h_f(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} (\log \frac{1}{\beta_t}) \left|\text{sgn}[h_t(x) - y]\right|$$

Figure A.1: Boosting algorithm AdaBoost.M1 for multi-class problems.

Input: sequence of N training examples $((x_1, y_1), \ldots, (x_N, y_N))$

Initialize weight distribution $W_i=1/N$, where $i=1,\ldots, N$.

Do for t=1, … , T where T specifies the total number of iterations

1. Train classifier using weight distribution W_i
2. Get back a hypothesis $h_t : X \rightarrow Y$
3. Calculate a loss for each h_t:

$$L_i = \frac{| h_t(x_i) - y_i |}{D} \qquad \text{(linear)}$$

$$L_i = \frac{| h_t(x_i) - y_i |^2}{D^2} \qquad \text{(square law)}$$

$$L_i = 1 - \exp\left[\frac{-| h_t(x_i) - y_i |}{D}\right] \qquad \text{(exponential)}$$

where $D = \sup | h_t(x_i) - y_i |$

4. Calculate an average loss: $\overline{L} = \sum_i^N L_i w_i$
5. Set $\beta = \overline{L} /(1 - \overline{L})$
6. Set the new weights vector to be

$$w_i^{t+1} = w_i^t \beta_t^{[1-L_i]}$$

Output the hypothesis

$$h_f(x) = \inf\left\{ y \in Y : \sum_{t:h_t \le y} \log(1/\beta_t) \ge 0.5 \sum_t \log(1/\beta_t) \right\}$$

Figure A.2: Boosting algorithm for regression problems.

Appendix B

Application of SHR to Voice Quality Analysis

Voice quality research is critical for a better understanding of underlying voice production/perception mechanisms as well as developing many practical speech applications. In speech-language pathology, describing certain types of voices whether pathological or normal are often needed. In unit-selection based speech synthesis, the pre-recorded voice is crucial to the success of the synthesizer. Therefore, having an objective criterion to select "good" speakers is very desirable. For speaker recognition, developing detailed and accurate voice quality profiles could provide important information for differentiating speakers. SHR seems to have the potential to serve as a quantitative descriptor in this regard.

To relate SHR to voice quality, some simple analyses were performed using the Keele database. For each speaker, all voiced frames were selected and SHR frequency distributions were calculated (see Tables B.1 and B.2).

Speakers	SHR distribution (%)					
	0	(0,0.2)	[0.2,0.4]	(0.4,0.6]	(0.6,0.8]	[0.8,1]
M1	38.39	0.88	2.53	12.65	24.37	21.18
M2	72.50	4.27	1.95	2.82	8.40	10.06
M3	85.69	0.48	0.34	1.10	3.70	8.69
M4	89.53	0.12	0.12	0.37	2.89	6.96
M5	41.67	0.58	1.26	10.28	23.90	22.31

Table B.1: SHR distribution for the Keele male speakers.

Speakers	SHR distribution (%)					
	0	(0,0.2)	[0.2,0.4]	(0.4,0.6]	(0.6,0.8]	[0.8,1]
F1	62.38	14.24	1.11	1.37	7.84	13.06
F2	71.14	12.72	1.10	2.47	3.68	8.89
F3	72.85	7.68	0.93	1.79	5.76	10.99
F4	52.52	10.37	5.16	10.15	11.31	10.48
F5	56.30	16.85	0.48	6.78	9.74	9.85

Table B.2: SHR distribution for the Keele female speakers.

As discussed in the Background section of Chapter 3, when SHR is in the medium range, especially [0.2, 0.4], perceived pitch becomes ambiguous. For the current data, in Table 3.3 (Chapter 3), speakers M1 and M2 have higher GERs whereas speaker M3 has the lowest. Correspondingly, in Table B.1, M1 and M2 have higher SHR percentage in the range of [0.2, 0.4] among five speakers, whereas M4 has the lowest. For the female speakers, speakers F4 and F5 represent the worst and best cases for pitch determination, respectively. Visual inspection and listening to the speech waveform suggest that M1 and M2 indeed contain more “irregular” speech cycles and appear to have low and rough voices in general. On the other hand, speaker M3’s speech seems to be much more “regular”. Similarly F4 seems to have more creaky voice than F5 despite the average pitch of F4 is

much higher. Tables B.1 and B.2 also show that the female speakers have greater number of SHRs in the range of (0, 0.2) compared with male speakers. This indicates that female speech might have greater amount of small amplitude or period fluctuations.

It would be interesting to compare SHR with Harmonic-to-Noise Ratio (HNR) (Yumoto et al., 1982), which defines the noise level in speech with respect to harmonics. Due to the potential application in voice quality analysis, HNR has been widely investigated (c.f. Murphy, 1999). HNR and SHR share some similarities: (1) both try to quantify the energy level of some "aperiodic" components in regard to the harmonic components, and (2) both can indicate some distinct perceptual voice quality. However, significant differences exist between the two as discussed below.

Depending on the definition of "noise", HNR and SHR may have very different targets. In many popular software packages, HNR is provided as a parameter different from jitter and shimmer. This implies that only some kind of random noise component in the voice, like aspiration noise, is treated (Murphy, 1999). Following this definition, HNR targets different phenomenon from that described by SHR. However, the HNR calculated by commonly used methods (e.g. Yumoto et al., 1982) can be affected by many factors, such as Gaussian noise, or jitter and shimmer (Murphy, 1999; Titze, 1995). This makes direct applications of HNR problematic. As pointed out by Titze, "the measure correlates best with an overall perception of 'noisiness and roughness' in the signal, regardless of what the source might be." Nevertheless, researchers attempt to restrict HNR to the narrow definition. A lot of research has been devoted lately to developing algorithms computing

perturbation (e.g., jitter and shimmer) free HNR (e.g. Murphy, 1999), which describes noise components only. In any case, SHR seems to be more specifically defined. Furthermore, the value of SHR has a more straightforward connection with voice quality and waveform type (Sun and Xu, 2002). It is bounded in the ranges from 0 to 1, in which medium values correspond to the "rough" voice. On the other hand, HNR normally is not bounded, and its relationship with voice quality is not clearly defined.

Appendix C

Pitch Accent Prediction Results

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	93.73%(1808)	17.87%(121)	38.86%(82)	91.43%(32)
High	5.29%(102)	77.55%(525)	46.45%(98)	8.57%(3)
Downstep	0.98%(19)	4.58%(31)	14.69%(31)	0%(0)
Low	0%(0)	0%(0)	0%(0)	0%(0)

Table C.1: Results of pitch accent recognition using acoustic features with single CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	95.08%(1834)	15.51%(105)	37.91%(80)	88.57%(31)
High	4.35%(84)	82.42%(558)	50.71%(107)	8.57%(3)
Downstep	0.52%(10)	2.07%(14)	10.90%(23)	0%(0)
Low	0.05%(1)	0%(0)	0.47%(1)	2.86%(1)

Table C.2: Results of pitch accent recognition using acoustic features with bagging CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	94.82%(1829)	13.59%(92)	31.28%(66)	85.71%(30)
High	4.56%(88)	80.35%(544)	48.82%(103)	2.86%(1)
Downstep	0.62%(12)	5.76%(39)	19.43%(41)	5.71%(2)
Low	0%(0)	0.30%(2)	0.47%(1)	5.71%(2)

Table C.3: Results of pitch accent recognition using acoustic features with AdaBoost CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	90.82%(1755)	19.79%(134)	24.64%(52)	17.14%(6)
High	7.47%(144)	75.48%(511)	60.19%(127)	77.14%(27)
Downstep	1.71%(33)	4.73%(32)	15.17%(32)	5.71%(2)
Low	0%(0)	0%(0)	0%(0)	0%(0)

Table C.4: Results of pitch accent prediction using text features with single CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	92.43%(1783)	24.08%(163)	26.07%(55)	20%(7)
High	6.22%(120)	71.20%(482)	57.35%(121)	71.43%(25)
Downstep	1.35%(26)	4.73%(32)	16.59%(35)	8.57%(3)
Low	0%(0)	0%(0)	0%(0)	0%(0)

Table C.5: Results of pitch accent prediction using text features with bagging CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	91.45%(1783)	21.57%(163)	26.07%(55)	22.86%(7)
High	6.58%(120)	72.53%(482)	54.03%(121)	71.43%(25)
Downstep	1.97%(26)	5.91%(32)	19.43%(35)	5.71%(3)
Low	0%(0)	0%(0)	0.47%(1)	0%(0)

Table C.6: Results of pitch accent prediction using text features with AdaBoost CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	94.56%(1824)	14.03%(95)	27.96%(59)	74.29%(26)
High	4.30%(83)	78.43%(531)	49.29%(104)	20%(7)
Downstep	1.09%(21)	7.39%(50)	22.27%(47)	2.86%(1)
Low	0.05%(1)	0.15%(1)	0.47%(1)	2.86%(1)

Table C.7: Results of pitch accent prediction using both acoustic and text features with single CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	96.84%(1868)	11.96%(81)	28.44%(60)	85.71%(30)
High	2.85%(55)	83.90%(568)	52.13%(110)	5.71%(2)
Downstep	0.31%(6)	4.14%(28)	19.43%(41)	5.71%(2)
Low	0%(0)	0%(0)	0%(0)	2.86%(1)

Table C.8: Results of pitch accent prediction using both acoustic and text features with bagging CART.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	96.79%(1867)	9.31%(63)	24.17%(51)	85.71%(30)
High	2.54%(49)	83.46%(565)	50.24%(106)	5.71%(2)
Downstep	0.47%(9)	7.09%(48)	24.17%(51)	0%(0)
Low	0.21%(4)	0.15%(1)	1.42%(3)	8.57%(3)

Table C.9: Results of pitch accent prediction using both acoustic and text features with AdaBoost CART.

Appendix D

Duration Prediction Using Ensemble Learning

For duration modeling, many approaches have been proposed previously, including rule-based systems (e.g. Klatt, 1976), decision trees (Riley, 1992), sum-of-product model (van Santen, 1994), neural networks (Campbell and Isard, 1991), belief networks (Goubanova, 2001), etc. In this work, using ensemble learning with decision tree, some preliminary work was conducted on vowel duration prediction. Table D.1 lists all the input features. The same training and testing set as described in Chapter 4 were used. It can be seen from Table D.2 that both Bagging and AdaBoost can improve the results over a single CART.

Feature name	Feature value
Vowel	aa ae ah ai ao aw ax axr ay eh el em en er ey ih iy ow oy uh uw
PhonePrev PhoneNext	OTHER UNVOICED VOICED_STOP
PhoneCur.Loc	SYL_INITIAL_LOC SYL_MEDIAL_LOC SYL_FINAL_LOC SYL_SINGLE_LOC
SylCur.NumPhones	1 2 3 4 5 6 7 8 9
SylCur.Stress, SylPrev.Stress, SylNext.Stress	0 or 1
SylCur.Loc, SylPrev.Loc SylNext.Loc	WORD_INITIAL, WORD_MEDIAL WORD_FINAL, WORD_SINGLE
WordCur.NumSyllables	1 2 3 4 5 6 7 8 9
PitchAccent PitchAccentPrev PitchAccentNext PitchAccentPrevPrev	NONE HIGH LOW DOWNSTEP_HIGH
PhraseAccent, PhraseAccentPrev PhraseAccentNext	NONE HIGH LOW DOWNSTEP_HIGH HIGH_BOUNDARY LOW_BOUNDARY
SylCur.DistSylsPreBreak SylCur.DistSylsNextBreak	_other_ 0 1 2 3 4 5 6 7 8 9 10)
WordCur.PartOfSpeech	NONE CC TO EX IN DT CD NN_NNP_NNPS_NNS JJ_JJR_JJS VB_VBD_VBP_VBZ_VBG_VBN WRB MD PRP RB_RBR_RBS_RP WP PRP$ WDT PDT WP$

Table D.1: Input features for duration prediction.

	RMSE (ms)	*r*
CART	28.9	0.80
Bagging	26.0	0.84
AdaBoost	26.8	0.83

Table D.2: Comparison between CART, Bagging, and AdaBoost for vowel duration prediction.

Appendix E

Phrase Break Prediction Using Ensemble Learning

Intonational phrase break prediction is an important step in text-to-speech synthesis, leading to prediction of prosodic boundaries. The information borne in phrase breaks is crucial to both intonation and duration modeling.

There have been several studies of stochastic phrase break prediction for text-to-speech systems (Koehn et al., 2000; Ostendorf and Veilleux, 1994; Sun and Applebaum, 2001; Taylor and Black, 1998; Wang and Hirschberg, 1992), which employ algorithms such as decision trees, Markov models, etc. The features that have previously been used in phrase-break prediction include part-of-speech (POS), pitch accent, syntactic structure, duration, etc. POS has proved to be an effective yet easily derived feature for predicting phrase breaks.

In the current work, using the same training and testing data described in Chapter 4, a single CART was trained as the baseline system with the input features listed in Table E.1. The stop value was 10. In both Bagging and Boosting, the maximum number of trees was 50 and the stop value for each single tree was 5.

Table E.2 shows the overall correct prediction rates of single CART, Bagging CART, and AdaBoost CART. Bagging seems to achieve more gains while AdaBoost only improves the results minimally.

Feature name	Feature value
WordCur.PartOfSpeech WordPrev.PartOfSpeech WordNext.PartOfSpeech WordPrevPrev.PartOfSpeech WordNextNext.PartOfSpeech	NONE CC TO EX IN DT CD NN_NNP_NNPS_NNS JJ_JJR_JJS VB_VBD_VBP_VBZ_VBG_VB N WRB MD PRP RB_RBR_RBS_RP WP PRP$ WDT PDT WP$
WordCur.NumSyllables	1 2 3 4 5 6 7 8 9
WordCur.Punctuation	NONE COMMA SENT
WordCur.DistWordsBOS	_other_ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Table E.1: Input features for intonational phrase break prediction.

	Overall correct rate (%)
CART	88.23
Bagging	89.86
AdaBoost	88.34

Table E.2: Comparison between CART, Bagging, and AdaBoost for intonational phrase break prediction.

Appendix F

Implementation of the Three-Target Model

The implementation of the Three-Target model is straightforward. It consists of two steps. First, a ToBI-style pitch accent prediction model is built. Then, an F0 target model is built using pitch accent as major input features.

For the first step, the original ToBI accents were grouped following Ross and Ostenforf (1999). Therefore, pitch accent contains "NONE HIGH LOW DOWNSTEP_HIGH", whereas phrase accent contains "NONE, HIGH, LOW, DOWNSTEP_HIGH, HIGH_BOUNDARY, LOW_BOUNDARY". Each syllable has two labels, pitch accent and phrase accent. Two separate decision trees are grown based on the input features shown in Table F.1.

For each syllable, three points were selected at the start, middle and end of the voiced portion. Table F.2 lists the features for building regression trees for the three target points.

Feature name	Feature value
Vowel	aa ae ah ai ao aw ax axr ay eh el em en er ey ih iy ow oy uh uw
SylCur.Stress, SylPrev.Stress	0 or 1
WordCur.NumSyllables WordPrev.NumSyllables	_other_ 1 2
SylCur.Loc, SylPrev.Loc SylNext.Loc	WORD_INITIAL, WORD_MEDIAL WORD_FINAL, WORD_SINGLE
SylCur.DistSylsPreBreak SylCur.DistSylsNextBreak	_other_ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20)
WordCur.PartOfSpeech WordPrev.PartOfSpeech WordNext.PartOfSpeech	NONE CC TO EX IN DT CD NN_NNP_NNPS_NNS JJ_JJR_JJS VB_VBD_VBP_VBZ_VBG_VBN WRB MD PRP RB_RBR_RBS_RP WP PRP$ WDT PDT WP$
WordCur.DistWordsBOS WordCur.DistWordsEOS	_other_ 0 1 2 3 4 5

Table F.1: Input features for pitch accent and phrase accent prediction decision trees.

Feature name	Feature value
Vowel	aa ae ah ai ao aw ax axr ay eh el em en er ey ih iy ow oy uh uw
SylCur.Stress, SylPrev.Stress, SylNext.Stress	0 or 1
SylCur.Loc, SylPrev.Loc SylNext.Loc	WORD_INITIAL, WORD_MEDIAL WORD_FINAL, WORD_SINGLE
PitchAccent PitchAccentPrev PitchAccentNext PitchAccentPrevPrev	NONE HIGH LOW DOWNSTEP_HIGH
PhraseAccent, PhraseAccentPrev PhraseAccentNext	NONE HIGH LOW DOWNSTEP_HIGH HIGH_BOUNDARY LOW_BOUNDARY
SylCur.DistSylsPreBreak SylCur.DistSylsNextBreak	_other_ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20)
WordCur.Break	_other_ 3 4

Table F.2: Input features for regression trees to predict three F0 targets.

During the prediction phase, the predicted pitch accents and phrase accents were incorporated into the input feature set to predict F0 target values using the three targets trees. The results of accent prediction and F0 prediction are presented below.

Recognized	Hand-labeled			
	Unaccented	High	Downstep	Low
Unaccented	90.82%(1755)	19.79%(134)	24.64%(52)	17.14%(6)
High	7.47%(144)	75.48%(511)	60.19%(127)	77.14%(27)
Downstep	1.71%(33)	4.73%(32)	15.17%(32)	5.71%(2)
Low	0%(0)	0%(0)	0%(0)	0%(0)

Table F.3: Results for pitch accent prediction using text features with single CART.

Recognized	Hand-labeled					
	Unaccented	High	Low	Downstep-High	High-boundary	Low-boundary
Unaccented	98.64% (1160)	84.38% (27)	82.61% (19)	94.12% (16)	0% (0)	0% (0)
High	0.43% (5)	15.63% (5)	8.70% (2)	0% (0)	0% (0)	0% (0)
Low	0.94% (11)	0% (0)	4.35% (1)	5.88% (1)	0% (0)	0% (0)
Downstep-high	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
High-boundary	0% (0)	0% (0)	0% (0)	0% (0)	58.33% (35)	22.86% (32)
Low-boundary	0% (0)	0% (0)	0% (1)	0% (0)	41.67% (25)	77.14% (108)

Table F.4: Results for phrase accent prediction using text features with single CART.

Original F0 – Predicted F0	RMSE	*R*
With hand-labeled ToBI accents	33.9	0.70
With predicted ToBI accents	41.2	0.52

Table F.5: Comparison between original F0 and predicted F0 generated by the Three-Target model.

Appendix G

Implementation of the Tilt Model

The implementation of the Tilt model is a nontrivial task as its analysis/synthesis procedure is not as straightforward as that of the Three-Target model. Fortunately, automatic Tilt analysis/synthesis programs are freely available (Taylor et al., 1999). This greatly facilitates the implementation process and also makes the results more comparable and reproducible. Note that according to the Tilt theory only the phonetic detail of the so-called "intonation event" is modeled, which makes an intonation event prediction module necessary. On the other hand, in Three-Target model, every syllable is modeled and pitch accents are used as input features. Following the Tilt model definition, ToBI pitch accents are mapped to intonation event "a", and boundary tones are mapped to intonation event "b". For each intonation event, the basic procedure of implementing the Tilt model is as follows: (1) Convert F0 values into tilt parameters; (2) Build decision trees for each tilt parameters based on some input features from training set; (3) Given a new set of data, predict tilt parameters using the decision trees; (4) Convert these tilt parameters into F0 values.

A decision tree was built to predict intonation events for each syllable using features such as POS, which are basically the same as those used by the Three-Target model (See Table G.1). The feature set was optimized based upon the classification rate on

the testing set. The overall correct prediction rate is 86.21% (1000/1160). Detailed results are shown in Table G.2.

Feature name	Feature value
SylCur.NumPhones	_other_ 1 2 3 4 5 6
SylCur.Stress SylPrev.Stress SylNext.Stress SylPrevPrev.Stress SylNextNext.Stress	0 or 1
SylCur.Loc, SylPrev.Loc SylNext.Loc	WORD_INITIAL, WORD_MEDIAL WORD_FINAL, WORD_SINGLE
WordCur.NumSyllables WordPrev.NumSyllables	_other_ 1 2 3 4
SylCur.DistSylsPreBreak SylCur.DistSylsNextBreak	_other_ 0 1 2 3 4 5 6 7 8 9 10)
WordCur.PartOfSpeech WordPrev.PartOfSpeech WordNext.PartOfSpeech	NONE CC TO EX IN DT CD NN_NNP_NNPS_NNS JJ_JJR_JJS VB_VBD_VBP_VBZ_VBG_VBN WRB MD PRP RB_RBR_RBS_RP WP PRP$ WDT PDT WP$
WordCur.Break WordPrev.Break WordNext.Break	_other_ 3 4

Table G.1: Input features for intonation event prediction decision tree.

Predicted	Hand-labeled		
	a	b	None
a	83.73% (283/338)	0% (0/157)	15.64% (104/665)
b	0% (0/338)	100.00% (157/157)	0.15% (1/665)
None	16.27% (55/338)	0% (0/157)	84.21% (560/665)

Table G.2: Results for intonation event prediction.

The raw F0 contours are the same as those used by the proposed system and the Three-Target model. The Tilt model requires the F0 contours to be interpolated throughout the unvoiced region. Initially cubic-spline method was used, but the results were not as satisfactory as those obtained using simple linear interpolation because many negative F0 values were generated by spline extrapolation.

After obtaining intonation events and F0 values, the "tilt_analysis" program from Edinburgh Speech Tool Library (Taylor et al., 1999) was used to extract tilt parameters from each intonation event. The command line was “tilt_analysis f0_file -w1 0.05 -w2 0.05 -limit 0.01 -range 0.1 -event_names “a b” -e event_file -otype tilt -o tilt_file”. Then a decision tree was built for each tilt parameter, in which the stop value was empirically set to 30. The original intonation event labels were included in the input feature set. Table G.3 shows all the input features. Finally, on the testing set, tilt parameters were predicted based on predicted intonation events. Then, "tilt_synthesis" program was used to synthesize F0 contours using these predicted tilt parameters. The command line was “tilt_synthesis tilt_file -event_names “a b” -s 0.01 -o f0_file”. In the generated F0 contours, all values less than 50 Hz were set to 50 Hz. Note that besides tilt parameters, the “tilt_synthesis” program also needs to know the F0 values and the time of the phrase start and end points. In a practical system, these values need to be predicted as well. The phrase starting and ending time can be derived once the phrase break and segmental duration information are available. For starting and ending F0 of the phrase, separate trees can be trained to predict

these values. In the present work, however, original F0 values and time of the phrase start and end points were used for simplicity. Table G.4 shows the objective evaluation results.

Feature name	Feature value
Vowel	aa ae ah ai ao aw ax axr ay eh el em en er ey ih iy ow oy uh uw
Coda	NONE l r w y m n ng nx eng
SylCur.Stress, SylPrev.Stress, SylNext.Stress	0 or 1
SylCur.Loc, SylPrev.Loc SylNext.Loc	WORD_INITIAL, WORD_MEDIAL WORD_FINAL, WORD_SINGLE
WordCur.NumSyllables WordPrev.NumSyllables	_other_ 1 2 3 4
IntonationEvent	NONE a b
SylCur.DistSylsPreBreak SylCur.DistSylsNextBreak	_other_ 0 1 2 3 4 5 6 7 8 9 10
WordCur.PartOfSpeech WordPrev.PartOfSpeech WordNext.PartOfSpeech	NONE CC TO EX IN DT CD NN_NNP_NNPS_NNS JJ_JJR_JJS VB_VBD_VBP_VBZ_VBG_VBN WRB MD PRP RB_RBR_RBS_RP WP PRP$ WDT PDT WP$
WordCur.Break	0 3 4

Table G.3: Input features for regression trees to predict tilt parameters.

Original F0 – Reconstructed F0	*RMSE* (Hz)	*r*
Model F0	28.8	0.81
Predicted F0 (Original intonation event)	38.5	0.62
Predicted F0 (Predicted intonation event)	41.9	0.52

Table G.4: Comparison between original F0 and predicted F0 generated by the Tilt model.

Appendix H

Sentences Used in the Experiment

f2bs32p3_1: Another Sierra club spokeswoman says Weld will listen to both sides of the story before making a judgment
f2bs32p3_2: but she doesn't get that feeling from Silber
f2bs32p3_3: But while Weld may be long on people skills, he may be short on money to implement his proposals
f2bs32p3_4: Weld supports the C.L.T. tax roll back petition, unlike those who endorsed him yesterday who are working to defeat question three.
f2bs32p3_5: But even if the roll back passes, they still say they'd rather see Weld in the corner office, than Silber.
f2bs32p3_6: When told of the endorsement, Silber said environmentalists picked the wrong man.
f2bs32p4_1: Several prominent Democrats in the environmental movement, including former U.S. Senator Paul Tsongas who are backing Silber, agree.
f2bs32p4_2: Attorney Douglas McDonald, a specialist in environmental law,
f2bs32p4_3: says Silber's got his vote because of his anti-C.L.T. stance.
f2bs32p4_4: McDonald also appreciates what he calls Silber's environmental pragmatism,
f2bs32p4_5: and says the quote beaver thing was unfortunate.
f2bs32p5_1: But Weld's endorsers say Silber's involved in a masquerade.
f2bs32p5_2: Being against C.L.T. they say, doesn't automatically make Silber pro-environment.
f2bs32p5_3: They say the only true environmentalist in the gubernatorial race is William Weld.
f2bs32p5_4: For WBUR, I'm Margo Melnicove.
f2bs33p2_1: Town officials say, given that Rockport's already heavily in debt to pay for a new school complex,
f2bs33p2_2: there's no new library in sight unless Denghausen's money can be used for construction.
f2bs33p2_3: The present library opened in nineteen- oh- six,
f2bs33p2_4: director, Steven Rask, apologizes for piles of books on the floor, saying there's no more room on the shelves.
f2bs33p2_5: If allowed to move to the former school house, Rask says the library could accommodate nearly three times as many volumes,
f2bs33p2_6: an expanded children's room, a community meeting hall and a reading room named after its benefactor.

f2bs33p2_7: Now there's no place to read, says Rask, except for a few chairs squeezed into corners.
f2bs33p3_1: Smithsonian spokesperson, Madeleine Jacobs, insists the institution's not being greedy.
f2bs33p3_2: The Smithsonian already inherited nearly four million dollars from Mrs. Denghausen who died a year before her husband.
f2bs33p3_3: It's due four to five million more once Mister Denghausen's estate is settled.
f2bs33p3_4: Being entrusted with most of the couple's fortune, says Jacobs,
f2bs33p3_5: gives the Smithsonian a responsibility to ensure that none of it's misspent.
f2bs33p4_1: Rockport residents find that argument especially galling,
f2bs33p4_2: saying to the Smithsonian, Denghausen's nothing more than a generous stranger.
f2bs33p4_3: The case will be heard in Essex's probate court August fifteenth,
f2bs33p4_4: meanwhile several of Rockport's well connected citizens are cancelling their Smithsonian memberships
f2bs33p4_5: and lobbying politicians to throw their weight behind the town.
f2bs33p4_6: If Rockport prevails, the Smithsonian's Jacobs says the institution's likely to let the matter rest in peace.
f2bs33p4_7: For WBUR, I'm Margo Melnicove.
f2bs34p1_1: Wakefield businessman, Raymond Marshall, says his property on the edge of Grafton is an ideal location for the Asbestos Conversion plant.
f2bs34p1_2: Marshall says it would take about ten million dollars to equip the building for glass making,
f2bs34p1_3: with six furnaces operating round the clock to melt down truckloads of asbestos waste each day.
f2bs34p1_4: Marshall says the venture would be good for Grafton, asbestos haulers would pay a fee for disposal,
f2bs34p1_5: and Marshall's agreed to give ten percent to the town which he says could amount to more than three million dollars annually,
f2bs34p1_6: nearly one- third of Grafton's total budget.
f2bs34p1_7: The operation would also be good for Marshall's profit margin, although he refuses to be specific.
f2bs34p1_8: And Marshall says, it would be good for the environment to have cancer causing asbestos fibers removed from it forever.
f2bs34p3_1: Grafton residents say they're worried about what could happen even before the material gets to the plant.
f2bs34p3_2: En route asbestos filled trucks would have to drive through residential areas.
f2bs34p3_3: Only one accident, says the town's public health nurse, could subject countless homes to an asbestos shower.
f2bs34p3_4: Town planner, Peter Lowett, says more research is needed to determine if that fear is well-founded.

f2bs34p3_5: Meanwhile, he says town officials will not rubber-stamp the project
f2bs34p3_6: even though Grafton could certainly use the millions of dollars it's been promised.
f2bs34p4_1: Beyond town zoning and Public Health Board's review a state environmental impact report is likely to be required.
f2bs34p4_2: Developer, Raymond Marshall, says if the permitting process drags on, he'll take his business out of state.
f2bs34p4_3: For WBUR, I'm Margo Melnicove.

Vita

Xuejing Sun

Department of Communication Sciences and Disorders, Northwestern University
2299 North Campus Drive, Evanston, IL 60208
Phone: (847) 491-2429
Fax: (847) 467-2776
E-mail: sunxj@northwestern.edu
http://mel.speech.nwu.edu/sunxj

Education

Ph.D. in Speech Science, Northwestern University, Evanston, IL, 2002
Bachelor of Science in Psychology, Peking University, Beijing, China, 1997

Research Experience

Research Assistant Northwestern University, Evanston, IL, 1998 - Present

Research Intern Panasonic Speech Technology Laboratory, Santa Barbara, CA, summer, 2000

Research Assistant Peking (Beijing) University, Beijing, China, 1995 – 1997

Computer Experience

Software Developer Kellogg School of Management, Northwestern University, 1998 - 1999

Software Engineer Wood&Stone Software, Inc., Beijing, China, 1994 - 1997

Web Developer Yuansun Software, Inc., Beijing, China, Summer, 1996

Awards and Professional Activities

- Graduate Research Grant award , Northwestern University, 2000-2001
- University Fellowship, 1997 -1998
- Microsoft Certified Systems Engineer, 1997
- Excellent Student of the Class award, 1996
- Third-place award in the "Challenge Cup" research & invention contest of Peking University, 1996
- Member of IEEE, ASA, ISCA.

Publications

Sun, X. "Pitch accent prediction using ensemble machine learning," *Proc of ICSLP2002*, Denver, Colorado, Sept. 16-20, 2002.

Sun, X. "F0 generation for speech synthesis using a multi-tier approach," *Proc. of ICSLP2002*, Denver, Colorado, Sept. 16-20, 2002.

Sun, X. and Xu, Y. "Perceived pitch of synthesized voice with alternate cycles," *Journal of Voice*, 2002.

Sun, X., "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *Proc. of ICASSP2002,* Orlando, Florida, 1, pp 333-336, 2002.

Xu, C. X, Xu, Y., and Sun, X. "Consonant Perturbation on F0 in English and Mandarin," 9th Meeting of the International Clinical Phonetics and Linguistics Association, Hong Kong, May 1-4, 2002 (Abstract).

Xu, Y, and Sun, X. "Maximum speed of pitch change and how it may relate to speech," *J. Acoust. Soc. Am.,* Vol. 111, No. 3, pp. 1399-1413*,* March 2002.

Sun, X. "Predicting Underlying Pitch Targets for Intonation Modeling," *Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis,* Perthshire, Scotland, pp. 143-148, 2001.

Sun, X. and Applebaum, T.H. "Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model," *Proc. of Eurospeech2001,* Aalborg, Denmark, Vol 1, pp. 537-540, 2001.

Sun, X., Xu, C.X., and Xu, Y. " Experiment on pitch target approximation model for generating Mandarin F0 contour," *J. Acoust. Soc. Am*. 109 (5), Pt. 2 of 2, 2414, May 2001 (Abstract only).

Sun, X. "A pitch determination algorithm based on subharmonic-to-harmonic ratio," *Proc. of ICSLP2000*, Beijing, China, 4, pp. 676-679, 2000,

Xu, Y, and Sun, X. "How fast can we really change pitch? – Maximum speed of pitch change revisited," *Proc. of ICSLP2000,* Beijing, China, 3, pp. 666-669, 2000.

Sun, X. "Voice Quality Conversion in TD-PSOLA Speech Synthesis," *Proc. of ICASSP2000,* Istanbul, Turkey, 2, 953-956, June, 2000.

Sun, X. "The perceived pitch of synthesized vowels with alternate pulse cycles," *J. Acoust. Soc. Am.* 107, 2907, 2000 (Abstract only).