Prosodic Modeling for Improved Speech Recognition and Understanding

by

Chao Wang

M.S., Massachusetts Institute of Technology (1997)
B.S., Tsinghua University, Beijing, China (1994)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2001

Author .....
Department of Electrical Engineering and Computer Science
June, 2001

Certified by .....
Stephanie Seneff
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

Prosodic Modeling for Improved Speech Recognition and Understanding

by

Chao Wang

Submitted to the Department of Electrical Engineering and Computer Science
in June, 2001 in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Abstract

The general goal of this thesis is to model the prosodic aspects of speech to improve human-computer dialogue systems. Towards this goal, we investigate a variety of ways of utilizing prosodic information to enhance speech recognition and understanding performance, and address some issues and difficulties in modeling speech prosody during this process. We explore prosodic modeling in two languages, Mandarin Chinese and English, which have very different prosodic characteristics. Chinese is a tonal language, in which intonation is highly constrained by syllable F_0 patterns determined by lexical tones. Hence, our strategy is to focus on tone modeling and account for intonational aspects within the context of improving tone models. On the other hand, the acoustic expression of lexical stress in English is obscure and highly influenced by intonation. Thus, we examine the applicability of modeling lexical stress for improved speech recognition, and explore prosodic modeling beyond the lexical level as well.

We first developed a novel *continuous* pitch detection algorithm (CPDA), which was designed explicitly to promote robustness for telephone speech and prosodic modeling. The algorithm achieved similar performance for studio and telephone speech (4.25% vs. 4.34% in gross error rate). It also has superior performance for both voiced pitch accuracy and Mandarin tone classification accuracy compared with an optimized algorithm in XWAVES. Next, we turned our attention to modeling lexical tones for Mandarin Chinese. We performed empirical studies of Mandarin tone and intonation, focusing on analyzing sources of tonal variations. We demonstrated that tone classification performance can be significantly improved by taking into account F_0 declination, phrase boundary, and tone context influences. We explored various ways to incorporate tone model constraints into the SUMMIT speech recognition system. Integration of a simple four-tone model into the first-pass Viterbi search reduced the syllable error rate by 30.2% for a Mandarin digit recognition task, and by 15.9% on the spontaneous utterances in the YINHE domain. However, further improvements by using more refined tone models were not statistically significant.

Leveraging the same mechanisms developed for Mandarin tone modeling, we incorporated lexical stress models into spontaneous speech recognition in the JUPITER weather domain, and achieved a 5.5% reduction in word error rate compared to a state-of-the-art baseline performance. However, our recognition results obtained with a one-class (including all vowels) prosodic model seemed to suggest that the gain was mainly due to the elimination of implausible hypotheses, e.g., preventing vowel/non-vowel or vowel/non-phone confusions, rather than by distinguishing the fine differences among different stress and vowel classes.

We also examined the use of prosodic cues in recognition confidence scoring. We conducted experiments to compare the accept/reject decision performance using only features derived from the recognizer outputs, such as normalized acoustic scores, with that obtained after prosodic features were included. The prosodic cues reduced the minimum classification error rate from 16.9% to 15.6% for utterance-level decisions, and from 10.9% to 10.2% for word-level decisions, a statistically significant result. We also explored the feasibility of characterizing directly the pitch contours of some selected common phrases in the JUPITER domain, without intermediate prosodic transcriptions. We achieved an accuracy of 58.8% in classifying these phrases on unseen data, based on F_0 features alone. We were able to identify some interesting typical F_0 patterns for these phrases, and use mutual information between these patterns to quantify interdependencies among phrases within an utterance. These phrase models can potentially be applied to score the intonation patterns of recognizer hypotheses, which can in turn be used to resort the N-best outputs for improved recognition/understanding accuracy or to support the rejection of erroneous hypotheses.

In this thesis, we make the following contributions to research in the area of prosodic modeling: (1) the development of a continuous pitch tracking algorithm that is particularly robust for telephone speech and prosodic modeling applications; (2) an empirical study of Mandarin tone and tonal variations, which analyzes the effects of tone coarticulation, tone sandhi, F_0 downdrift and phrase boundary, on the acoustic realizations of tone; (3) the development of a mechanism which is able to combine multiple classifiers and/or selectively score for a subset of phones in the recognition first-pass search; (4) the development and analysis of a preliminary framework for characterizing pitch contours of spontaneous English utterances without intermediate prosodic transcriptions; and (5) improvements in speech recognition and confidence scoring performance using prosodic information.

Thesis Supervisor: Stephanie Seneff
Title: Principal Research Scientist

Acknowledgments

Foremost, I would like to express my deepest gratitude to my thesis advisor Stephanie Seneff. I have enjoyed working with her since I joined the SLS group over five years ago. She has always been a great resource for ideas and solutions; and her encouragement and support have made difficult times during my thesis exploration less frustrating. I am also most grateful to her for helping me to overcome the language barrier as a non-native speaker of English. Her understanding, patience, encouragement, and being a role model have been the most important factors in increasing my confidence and improving my skills in oral communication and technical writing.

I would like to sincerely thank my other thesis committee members, Ken Stevens and Victor Zue, who provided valuable advice at individual discussions and committee meetings. I especially thank Victor for asking insightful questions and challenging me to think beyond facts and details, and Ken for offering different perspectives to this thesis. I would also like to thank Stafanie Shattuck-Hufnagal for sharing with me her expertise on English prosody.

SLS has been a great place in which to spend my graduate life at MIT. It is a privilege to have the opportunity to be exposed to intriguing problems and ideas, to have the best possible resources for conducting research, and to work with and learn from so many talented people. I am grateful to Victor Zue and Jim Glass for creating such a stimulating research environment, and for nurturing the comradeship among group members.

Everyone in the SLS group deserves sincere thanks. In particular, I would like to thank TJ Hazen for his help with the confidence scoring experiments and other recognition issues, Lee Hetherington for his help with FST and hunting down bugs, and Joe Polifroni for his help with data, demo systems and linguistics. I would also like to thank my past and present officemates Jane Chang, Jon Yi, and Eugene Weinstein for many fun hours spent in the office. In addition, I am grateful to Jane for teaching me how to write Sapphire objects, and to Jon for being a great resource for answering many technical and not-so-technical questions from Chinese computing to playing MP3. Thanks also to all present and many past members of SLS, Vicky, Sally, Michelle, Scott, Kenny, Ray, Drew, Giovanni, Mike, Grace, Xiaolong, Karen, Issam, Han, Ming, Lei, Ed, Ernie, for all of your help, and for making SLS a friendly and fun place.

I would like to thank my academic advisor Roger Mark for his encouragement and support during my first year at MIT, and for keeping me up with the schedules of department requirements.

I would like to thank my parents for instilling upon me the value of education at young age, and for always encouraging me to pursuit my aspirations.

Finally, I would like to thank my husband Jinlei for his patience and occasional impatience, for his love, for being a fun companion, and for always being there for me.

This dissertation is based upon work supported by DARPA under contract N66001-96-C-8526, monitored through Naval Command, Control and Ocean Surveillance Center, by the National Science Foundation under Grant No. IRI-9618731, and by DARPA under contract N66001-99-1-8904, monitored through Naval Command, Control, and Ocean Surveillance Center.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Prosody generally refers to the organization of *spoken* utterances; however, the term has been defined in many ways in the literature. The various definitions for prosody can be classified into three categories, after (Shattuck-Hufnagel and Turk 1996). One class defines prosody as the collection of its acoustic correlates, i.e., fundamental frequency (F_0), duration, amplitude, and "segment quality or reduction". A second class is based on the role of prosody in the linguistic structure, i.e., it defines prosody as the phonological organization of segments into higher-level constituents, such as utterances, intonational phrases, prosodic words, metrical feet, syllables, etc. A third class of definition combines the phonetic and phonological aspects of prosody, including both the higher-level organization, and the phonetic manifestation of this organization in the pattern of F_0, duration, amplitude, etc., within the utterance. We believe that the third class of definition is most appropriate for investigators of prosodic theory. In this thesis, however, we are mainly concerned with modeling the acoustic correlates of prosody to improve the performance of human-computer dialogue systems. Thus, we loosely specify prosody as the linguistic and extra-linguistic aspects of speech expressed predominantly by duration, energy, and F_0 cues. For a dialogue system, the segmental aspects of prosodic features, such as intrinsic phone duration, and the extra-linguistic aspects, such as prosodic cues of a speaker's emotional state, all contain useful information that can potentially be utilized to improve system performance.

In this chapter, we first motivate this thesis by describing the importance of prosody in human speech communication, the potential uses of prosody in human-computer dialogue

systems, and the difficulties of modeling prosody for automatic speech recognition and understanding (ASRU). Next, we briefly describe some related research in utilizing prosodic information in various aspects in a dialogue system. We then introduce the general goal and approach of this thesis. Finally, we give a chapter by chapter overview.

1.1 Motivations

1.1.1 Prosody in Human Speech Communication

Prosodic information plays an important role in human speech communication. We are able to infer a speaker's gender, age and emotion from his/her voice, regardless of what is said. Meanwhile, the same sequence of words can convey very different meanings with variations in *intonation*. For example, "next Monday" with a rising F_0 contour generally implies a question, while the same words with a falling F_0 contour are usually associated with a statement (Ladd 1996). Furthermore, the relative prominence of the two words can give rise to even richer meanings within context. When the prominence is on the word "next", the utterance could be emphasizing that the date is not "this" Monday or any other Monday; when the prominence is on the word "Monday", the utterance could be emphasizing that the date is not next "Tuesday" or "Wednesday", etc. Phrase grouping also contributes to the interpretation of spoken utterances (Price et al. 1991). "Old men and women" can mean "old men | and women" or "old | men and women", depending on where the phrase boundary is. Prosody also plays a part in defining words in most languages. In English, words are pronounced with certain stress patterns, and many minimal noun and verb pairs, such as "*per*mit" (noun) and "per*mit*" (verb), are disambiguated only by the stress position difference. In a tonal language such as Chinese, the syllable F_0 pattern is essential in determining the meaning of sounds, such that the syllable "*dà*" ("*da*" with a falling F_0 contour) is as different from "*dá*" ("*da*" with a rising F_0 contour), as it is different from "*bà*" ("*ba*" with a falling F_0 contour).

The preceding discussions have focused on the acoustic correlates of prosody, i.e., prosodic features. Prosody, as the higher-level organization of segments, is also an important factor in shaping the phonetics and phonology of sounds (Selkirk 1984). Studies have found

that segmental phonetic features are highly affected by the position of the segment in the prosodic hierarchy, such as syllable, metrical foot, intonational phrase, etc.

The various communicative functions of prosodic features are listed as follows:

- **Linguistic:**
  - Post-lexical: syntax, semantics, pragmatics, etc.
  - Lexical: word.
- **Extra-linguistic:**
  - Gender, age, emotion, attitude, etc.

The prosodic phenomena that serve the linguistic functions can also be divided accordingly, i.e., at the lexical level. In particular, the term *intonation* is used to refer to the prosodic component that carries the linguistic information above the lexical level (Ladd 1996). According to theories of intonational phonology, the intonation of an utterance can be described by categorical events, e.g., one or more intonational phrases, each consisting of one or more *pitch accents* followed by a *phrase tone* and then a *boundary tone* (Pierrehumbert 1980); and the phrase grouping and the relative prominence of these constituents within the utterance give rise to its "meaning" (syntactically, semantically, and pragmatically). The prosodic component at the lexical level varies for different languages. For example, it corresponds to lexical stress in English, tone in Mandarin Chinese, and accent in Japanese.

1.1.2 Potential Uses in Dialogue Systems

With the advancement in human language technologies, speech has been accepted as a natural method for humans to interact with computers. Many human-computer dialogue systems have been developed at MIT and elsewhere (Zue and Glass 2000), which allow users to access information (Goddeau et al. 1994; Glass et al. 1995; Zue et al. 2000), to conduct transactions (Seneff et al. 1999), or to perform other problem-solving tasks, using natural speech.

Several language-based technologies must be integrated to build systems that are able to interact with humans using speech. On the input side, *speech recognition* must be combined with *natural language processing* in order for the computer to derive an understanding of the spoken input. On the output side, *language generation* and *speech synthesis* capabilities must be developed in order for the system to deliver well-formed verbal responses. The system also needs to keep track of the *discourse* information and interpret an input utterance appropriately within the dialogue context. Overall, a *dialogue management* component controls the system's actions during its interaction with the user. Such actions usually include responding to a user query, asking for additional information, requesting clarification, or simply prompting the user to speak, etc.

The importance of prosody to the naturalness and intelligibility of speech is evident in speech synthesis (Klatt 1987). It is not surprising that much prosodic modeling work has been carried out on the prediction side for such applications. In a typical speech synthesis system, some linguistic analysis is usually performed, and prosodic tags (such as phrase boundaries, pitch accents, boundary tones, lexical stress labels, etc.) are attached to the input text. These prosodic markers are then used to control duration, fundamental frequency, energy, and segmental characteristics, through signal modification or unit selection, to achieve the acoustic realizations of the desired prosodic structure.

Overall, prosodic modeling for speech synthesis is somewhat "easier" than for speech recognition and understanding. On the analysis side, the speech data for synthesis are generally from a small number of speakers and more constrained in speaking style than those in most dialogue applications. On the modeling side, speech synthesis systems are only concerned with generating a single acceptable rendition of a given text; while speech on the input side can have multiple prosodic realizations and rich phonetic variations for conveying the same meaning.

Nevertheless, prosodic information can potentially contribute to many other aspects in a dialogue system besides speech synthesis, including speech recognition, understanding (e.g., syntactic/semantic analysis, etc.), and dialogue control, etc. On the speech recognition side, intrinsic phone duration and energy measurements can be utilized to provide acoustic constraints, in addition to spectral features, for phones; lexical stress information can be

used to train more accurate acoustic models (e.g., stress-dependent phone models), or to constrain lexical search; for a tonal language like Mandarin Chinese, tone pattern is essential in decoding the spoken words; more sophisticated prosodic models involving higher-level constraints (e.g., syntactic boundaries, etc.) can be used to resort the recognizer N-best list[1] for improved performance.

Prosody is also important to the analysis and interpretation of spoken utterances. Phrase boundary locations can be used to resolve syntactic ambiguities or to improve the efficiency of syntactic analysis; pitch accents can be detected to locate the focus of an utterance and assist semantic/pragmatic interpretation; the general shape of the pitch contour of an utterance (e.g., rising, falling, etc.) can be analyzed to determine the sentence mood (e.g., question, statement, etc.) (Daly 1994), and hence, the user intention; prosodic cues also reflect the structure of a message and play an important role in distinguishing dialogue acts (Jurafsky et al. 1998). In addition, prosody can aid the segmentation of a long speech recording into topics and sentences (Hakkani-Tür et al. 1999), and help locate speech disfluencies for improved parsing (Shriberg et al. 1997; Stolcke et al. 1998).

Prosody can also be used to explore robust dialogue strategies. The dialogue system can infer the speaker's emotion from verbal and prosodic cues in order to react properly. This is particularly useful during user-system error resolution, because a speaker tends to hyperarticulate when correcting system errors, while hyperarticulated speech subsequently causes even more recognition errors by the system (Shriberg et al. 1992; Oviatt et al. 1996). Another approach to solving this problem is using recognition confidence scores to reject incorrect recognition hypotheses (Hazen et al. 2000a; Hazen et al. 2000b). Instead of giving erroneous (and confusing) responses, the dialogue system can request clarification from the user, give more detailed instructions, or take more control of the interaction. Such accept/reject decisions in the confidence scoring framework can also potentially be improved by using prosodic cues (Hirschberg et al. 1999; Hirschberg et al. 2000).

[1] A recognizer N-best list consists of the N top-scoring sentence hypotheses for an input utterance.

1.1.3 Difficulties in Modeling Prosody

Prosodic modeling for speech recognition and understanding applications is a difficult problem due to many factors (Sagisaka et al. 1995; Shattuck-Hufnagel and Turk 1996; Kompe 1997):

- Prosody involves many linguistic and extra-linguistic aspects of a spoken utterance; however, the acoustic manifestation of such complex phenomena is more or less concentrated in three basic acoustic parameters: duration, F_0, and energy. When we model only selected prosodic components, the aspects unaccounted for by the modeling framework contribute to large variations in the measurements. For example, speaker F_0 differences and the sentential F_0 downdrift contribute to large variances in Mandarin tone features, if they are not taken into account by the tone modeling framework.

- The intonational aspects of prosody are not well understood. Although intonational phonology has proposed a relatively simple framework for describing the intonation of an utterance, i.e., as a sequence of intonational phrases, each consisting of certain categorical constituents, a suitable set of descriptive units has been elusive. In addition, automatic recognition of these constituents is a hard problem, and the correspondence of these constituents with the meaning of an utterance is very poorly understood. For example, although there is good correlation between prosodic phrase boundaries and syntactic boundaries, there are also plenty of cases in which they do not match; how to relate prominence to higher-level linguistic aspects is even more puzzling.

- There generally exist multiple prosodic realizations of a sentence to convey the same meaning. In addition, the acoustic expression of individual prosodic events is also highly variable. For example, a speaker can ask the question "What is the weather in Boston?" with a rising or a falling tune, to convey the same meaning. Individual prosodic events, such as lexical stress and pitch accents, can also be realized through different prosodic means.

- The extraction of prosodic features is error-prone. In particular, F_0 detection is likely

to suffer from pitch doubling/halving and voiced/unvoiced decision errors. Such errors make F_0 related measurements noisy and unreliable. Duration measurements are also susceptible to alignment errors.

The current approaches used in ASRU systems are also not optimal for incorporating prosodic constraints. For recognition systems based on hidden Markov models (HMMs), the inherent duration model has an exponential distribution, which is clearly not appropriate. More importantly, feature extraction in HMM systems is from a fixed-length frame; while it is often desirable to extract prosodic features from phones, syllables, or some higher-level constituents. A segment-based system has an advantage over frame-based systems in that prosodic features can be extracted from segments (phones) of variable lengths. However, to extract features from syllables or higher-level components, it is usually necessary to post-process the phone graph obtained from a first-stage recognizer. A probabilistic framework for syntactic and semantic analysis is also desirable for incorporating prosodic information, because hard decisions about phrase boundaries, pitch accents, etc., are not robust, due to high error rates in recognizing these prosodic attributes and variabilities in the realizations.

1.2 Prosody in Dialogue Systems

As mentioned previously, prosodic information can potentially contribute to many aspects in a dialogue system besides speech synthesis, such as speech recognition, syntactic/semantic analysis, dialogue management, etc. Many previous efforts have achieved limited success along these dimensions. In this section, we briefly review some prosodic research related to these aspects.

1.2.1 Prosody in Speech Recognition

Extensive work has been done on explicit duration modeling for HMM-based recognition systems. This is due to a limitation of the HMM framework, which imposes an exponential distribution on duration, determined by the state transition probabilities. Various techniques have been explored to introduce proper duration constraints into HMM recognizers for continuous speech. It was found in (Dumouchel and O'Shaughnessy 1995) that the

error rate on a single-speaker read corpus was reduced by 6% when constraints for minimum and maximum phone durations were imposed; however, a Gaussian mixture phone duration model did not improve recognition. For a segment-based recognition system, duration can be directly incorporated into the acoustic feature vector. Nevertheless, the use of explicit duration models, even when the duration feature was already included in the segment models as well, improved recognition accuracy on continuous Mandarin digits (Wang and Seneff 1998). It was also found that a sentence-level speaking rate normalization was important to the performance improvement. A hierarchical duration model was developed in (Chung 1997; Chung and Seneff 1999) to capture duration information at various sublexical levels. The model was able to quantitatively characterize phenomena such as speaking rate and prepausal lengthening. It also improved phonetic recognition and word spotting performances in the ATIS air travel information domain (Zue et al. 1993).

Lexical stress is an important property for the English language, which can be used to provide lexical constraints or to improve acoustic modeling for automatic speech recognition. Lexical stress pattern recognition has been explored to reduce the number of word candidates for large-vocabulary isolated word recognition (Aull 1984; Aull and Zue 1985; Waibel 1988), or to disambiguate stress-minimal word pairs (Freij and Fallside 1990). Acoustic models with different lexical stress properties were trained separately for more accurate modeling (Adda-Decker and Adda 1992; Sjölander and Högberg 1997). Prosodic models for syllables with different stress properties and syllable structures improved recognition when applied to resort recognizer N-best list (Jones and Woodland 1994). A stress-dependent cost matrix for phone to phoneme mapping was also able to improve recognition (Hieronymus et al. 1992). Tone recognition has traditionally been an integral part of Mandarin speech recognition in isolated syllable or word mode (Gao et al. 1991; Lee et al. 1993; Hon et al. 1994; Gao et al. 1995). In more recent years, tone models have also been incorporated into *continuous* Mandarin speech recognition systems (Wang et al. 1995; Cao et al. 2000; Huang and Seide 2000). We will review work on lexical stress modeling and Mandarin tone modeling in detail in Chapter 5 and Chapter 4 respectively.

It has also been proposed that prosodic models involving post-lexical constraints, such as syntactic boundaries, can be used to re-rank the recognizer N-best hypotheses for im-

proved recognition performance (Ostendorf and Veilleux 1994; Hunt 1994; Hirose 1995). However, we have not found any recognition results in support of this approach reported in the literature. A framework of similar flavor was developed for Japanese mora[2] recognition (Hirose and Iwano 2000). Prosodic word boundaries were detected based on statistical modeling of mora transitions of F_0 contours. The prosodic word boundary information was incorporated into the second stage of a two-stage mora recognition system, which slightly improved the recognition rates.

1.2.2 Prosody in Understanding

Despite the fact that the prosodic structure of an utterance exists independently of its syntactic structure, there is a general correspondence between the two. Several studies have found that prosodic phrase boundary locations can be utilized to assist syntactic analysis. Ostendorf *et al.* (1993) investigated the use of phrase boundaries to resolve syntactic ambiguities on a corpus of sentences read by radio announcers given two possible interpretations. Two scoring algorithms, one rule-based and one using a probabilistic model, rank the candidate parses by comparing the recognized prosodic phrase structure with the predicted structure for each candidate parse. It was found that the two algorithms achieved about 69% disambiguation accuracy, compared to a human perception accuracy of 84% on the same data. When the scoring algorithms used hand-labeled boundary breaks instead of acoustically detected boundaries, the performance became comparable to that of human subjects. Hunt (1995) tackled the same problem, using an approach which did not require hand-labeled prosodic data for training. Two prosody-syntax models were trained, without intermediate prosodic labels, using multi-variate statistical techniques. A 73% accuracy was achieved in resolving syntactic ambiguities for the same test as in (Ostendorf et al. 1993), and the performance could be improved to 75% if prosodic labels were used during training.

Kompe *et al.* (1997) implemented a framework to incorporate prosodic clause boundary information into the syntactic analysis of word hypothesis graphs. Syntactic boundaries

[2]The mora is a unit smaller than the syllable. A syllable contains at least one mora and normally contains no more than two. In Japanese, CV (C=consonant, V=vowel) and V syllables are considered monomoraic, whereas CVN (N= nasal consonant) and CVQ (Q= the first member of a geminated consonant) is considered bimoraic (Shattuck-Hufnagel and Turk 1996).

were marked according to whether they were likely to be expressed prosodically or not. These prosodic clause boundary labels were added into the parsing grammar, which was used to guide the search for the best word chain with prosodic boundary scores. This approach was tested on a German corpus collected for VERBMOBILE, a system for automatic speech-to-speech translation for meeting scheduling. The prosodic information reduced the parse time by 92% and the number of parse trees by 96%, with a very small degradation in parsing success rate.

1.2.3 Prosody in Dialogue Control

Prosody can also be used to improve the robustness of a dialogue system, and to enhance the quality of user experience. For example, error detection and handling is critical to the robustness of a dialogue system. Studies have shown that users often hyperarticulate when trying to correct system errors (Oviatt et al. 1996); while hyperarticulated speech subsequently led to even worse recognition performance, due to deviations from trained acoustic and language models (Shriberg et al. 1992). It is very important for the dialogue system to detect the presence of such trouble, and take measures to break the cycle of recognition failure. Oviatt *et al.* (1998) analyzed in detail the acoustic, prosodic, and phonological changes in user speech after different types of recognition errors. Results indicated that the primary hyperarticulate changes in speech were durational, with increased number and length of pauses most noteworthy. It was suggested that a dialogue system might evoke a recognizer specialized for error handling upon detection of hyperarticulation, or use multiple recognizers for different speaking styles. A more general solution is to use recognition confidence scores to reject any unreliable recognizer hypotheses, either caused by hyperarticulation, or due to other reasons such as out-of-vocabulary words, noise interference, etc. The dialogue system can take proper action, such as rejecting the utterance or requesting confirmation, to signal the user of system difficulty and guide the user in error correction (Hazen et al. 2000a; Hazen et al. 2000b). Hirschberg *et al.* (1999, 2000) have found that prosodic features are correlated with recognition performance, and prosodic information can be used to improve the rejection of incorrect recognition hypotheses.

A related issue is to detect the emotional state of a user, because humans sometimes

extend their inter-personal behavior into their interaction with computers. For example, users might express their satisfaction verbally or prosodically when a system works well for them; on the other hand, they tend to get frustrated and angry when a system keeps making errors. Thus, it is useful for a dialogue system to detect user emotion and react properly. Dellaert *et al.* (1996) explored the classification of utterances in terms of four emotional modes: happiness, sadness, anger, and fear, on a corpus of over 1000 utterances collected from a few speakers. The algorithm was able to achieve close-to-human performance using only F_0 related features and a majority voting technique on the simulated data. Huber *et al.* (1998) reported experiments on the detection of emotion (anger) using prosodic cues. Data were collected from 20 speakers, each speaking 50 neutral and 50 angry utterances. Word-based emotion classification was performed, with words in angry utterances labeled as "emotional", and words in neutral utterances labeled as "neutral". The acoustic modeling was by "brute-force": a total of 276 acoustic prosodic features were extracted from each word and its surrounding context. A precision of 94% and a recall of 84% were achieved on a test set with unseen speakers. Polzin (2000) explored using both verbal and non-verbal cues to classify three emotion modes: sad, angry, and neutral. The corpus consists of speech segments from English movies. Emotion-specific language and prosodic models were trained to classify the three modes. The prosodic model performed better than the language model, and both were well above chance. As summarized above, research on emotion detection is still largely preliminary. Finding good acoustic cues for various emotional modes remains an interesting research problem. It is also desirable to use data from real interactions between a user and a computer, rather than data in simulated or acted mood, to carry out the experiments. In addition, dialogue strategies incorporating user emotional state also need to be investigated.

1.3 Thesis Goals

The general goal of this thesis is to model the prosodic aspects of speech to improve human-computer dialogue systems. Towards this goal, we investigate a variety of ways of utilizing prosodic information to enhance speech recognition and understanding performance, and

address some issues and difficulties in modeling speech prosody during this process. We explore prosodic modeling in two languages: Mandarin Chinese and English, which have very different prosodic characteristics. Chinese is a tonal language, in which intonation is highly constrained by syllable F_0 patterns determined by lexical tones. Hence, our strategy is to focus on tone modeling and account for intonational aspects within the context of improving tone models. Lexical stress in English, on the other hand, is only obscurely expressed acoustically and highly influenced by intonation. Thus, we examine the applicability of modeling lexical stress for improved speech recognition, and explore prosodic modeling beyond the lexical level as well. Specifically, this thesis accomplishes the following tasks:

- Robust pitch tracking designed especially for telephone speech and prosodic modeling.
- Lexical tone modeling for Mandarin Chinese speech recognition.
- Lexical stress modeling for English speech recognition.
- Using prosodic cues to improve the utterance and word level confidence scoring of recognition hypotheses.
- Characterizing pitch contours of English phrases.

Pitch detection is an important first step in the analysis and modeling of speech prosody. The fundamental frequency is an important feature for many prosodic components, such as lexical stress, tone, and intonation. However, pitch estimation errors and the discontinuity of the F_0 space make F_0 related measurements noisy and undependable. Pitch detection algorithms also have inferior performance on *telephone speech*, due to signal degradation caused by the noisy and band-limited telephone channel. To address these problems, we will first implement a novel *continuous* pitch detection algorithm, which is designed explicitly to promote robustness for telephone speech and prosodic modeling.

We choose to focus our initial study of prosody on the F_0 contours of Mandarin Chinese for several reasons. First, unlike the obscure correlation of prosodic features with lexical stress in English, the syllable level F_0 contour in Mandarin clearly defines tone. Thus, tone modeling presents a well defined recognition problem. Furthermore, tones in continuous speech can vary to a great extent due to the influence from surrounding tones as well as

higher-level factors such as intonation. Thus, tone modeling presents an interesting and challenging research problem. It also provides us with a nice framework to address one of the difficulties in prosodic modeling, i.e., multiple prosodic components affecting the same acoustic signal. We can carry out the analysis and characterization of intonational influences within the context of improving tone recognition. Finally, tone modeling can potentially be used to improve recognition performance for continuous Mandarin Chinese speech. Traditionally, tone information has been used to assist lexical decoding for recognizing isolated syllables or words. Some researchers have argued that tone information is not critical in continuous speech recognition, because homophones are rare when multiple syllables are grouped into words. However, we think that speech recognition errors sometimes result in mismatched tone characteristics between the syllable hypotheses and the acoustics; thus, tone models can be used to improve speech recognition by discouraging such errors. To achieve this goal, we will explore and implement various methods to incorporate tone models into the speech recognition system.

Lexical stress in English can be viewed as analogous to tone in Mandarin Chinese. Leveraging the same mechanisms developed for Mandarin tone models, we explore the use of lexical stress information to assist speech recognition in spontaneous English utterances. The motivation is also similar to that for tone modeling, i.e., erroneous hypotheses will have worse "stress scores" than the correct hypothesis. However, unlike Mandarin tones, the acoustic manifestations of lexical stress are quite obscure. Thus, different issues are addressed in modeling lexical stress: first, what are the most informative acoustic correlates of stress; second, how well can the intrinsic stress properties of a vowel be determined from the acoustics in spontaneous speech; and third, can such information improve speech recognition performance?

Moving beyond improving speech recognition, we examine the use of prosodic cues in recognition confidence scoring for improved accept/reject decisions. Hirschberg *et al.* (1999, 2000) have found that there exist statistically significant differences in the mean values of certain prosodic features between correctly and incorrectly recognized user turns in interaction with a particular spoken dialogue system; and these prosodic cues can be used to improve accept/reject decisions on recognition outputs. However, it was also found that the

efficacy of the prosodic information was dependent on the quality of the recognition system. We first test if the approach of using prosodic cues in utterance-level confidence scoring can be generalized to the JUPITER system, which has been well-trained on a large corpus of speech data. We also examine if prosodic features can be used to better distinguish correctly and incorrectly recognized words.

Research on using prosody in syntactic and semantic analysis of spoken utterances has been sparse. Among the limited inquiries reported in the literature, most methods relied on an intermediate prosodic transcription to bridge the acoustics and the syntax/semantics of the utterance. Prosodic transcription is labor-intensive and time-consuming, which makes it impractical to transcribe large speech corpora. The prediction of these labels from acoustics is also error-prone. In addition, the mapping between the prosodic transcription and the syntax/semantics of an utterance is not obvious, except for the general correspondence of prosodic and syntactic boundaries. Thus, it is usually necessary to build sophisticated models for predicting prosodic labels from linguistic analysis, in addition to acoustic models for prosodic labels. We begin to explore the feasibility of characterizing directly the pitch contours of some selected English phrases in the JUPITER domain, without any intermediate prosodic labeling. These phrases are selected from common patterns in the JUPITER utterances, such as "what is", "tell me", etc. We seek to answer the following questions in our experiments: (1) can we identify phrases based on F_0 contour alone; (2) does the phrase F_0 pattern generalize across similar but not identical utterances; (3) does each phrase have some set of canonical patterns; (4) are there interdependencies among phrases in the utterance; and (5) will this information be useful to speech recognition or understanding?

1.4 Outline

The remainder of this thesis is organized into seven chapters. Chapter 2 describes the design principles and the implementation of a pitch tracking algorithm, which is particularly robust for telephone speech and prosodic modeling applications. Detailed evaluations of the algorithm are conducted on a labeled pitch extraction reference database under both studio and telephone conditions, and on a telephone quality Mandarin digit corpus. The

performances, in terms of pitch accuracy and tone classification accuracy, are compared with those of XWAVES, which is a commercially available product widely used by the speech community for prosodic research.

Chapter 3 presents an empirical study of a number of factors that contribute to the acoustic variations of Mandarin tones, including the overall F_0 declination of an utterance, the presence of a phrase boundary, tone coarticulation, and tone sandhi. Two telephone-quality Mandarin speech corpora used in our tone classification and speech recognition experiments are also described.

Chapter 4 presents the tone classification and speech recognition experiments on the two Mandarin corpora. We first describe the basic tone modeling framework, and compare the tone classification performance of various refined tone models. We then describe the implementation of two mechanisms in the recognition system for incorporating tone model constraints. A suite of speech recognition experiments are conducted to compare the contributions of using various tone models and different tone model integration methods.

Chapter 5 extends the framework described in Chapter 4 to modeling lexical stress for improved speech recognition in the English JUPITER weather information domain. We study the correlation of various F_0, energy, and duration measurements with lexical stress on a large corpus of spontaneous utterances, and identify the most informative features of stress using classification experiments. Stress classification and speech recognition experiments are conducted, and some analysis and interpretation of the speech recognition results are provided.

Chapter 6 describes experiments on using prosodic features to enhance the performance of speech recognition confidence scoring in the JUPITER domain. Various utterance-level and word-level prosodic features are compared, together with other features derived from the recognizer outputs, using the mutual information measure. Utterance and word confidence scoring performances with and without prosodic features are compared using standard classification and error detection criteria.

Chapter 7 describes a preliminary model for characterizing the pitch contours of some selected English phrases in the JUPITER domain. We perform classification experiments to examine how reliably these phrases can be distinguished by their F_0 contours alone. Data

clustering experiments are conducted to identify typical patterns for the F_0 contours of these phrases. A mutual information measure is used to quantify the correlation of various F_0 patterns of these phrases within an utterance.

Chapter 8 summarizes the main contributions of this thesis and suggests directions for future work.

Chapter 2

Robust Pitch Tracking

Robust pitch detection is a crucial first step to the analysis and modeling of speech prosody. The fundamental frequency (F_0) is an important feature for many prosodic attributes such as lexical stress, tone, and intonation. However, it is difficult to build reliable statistical models involving F_0 because of pitch estimation errors and the discontinuity of the F_0 space. Specifically, inaccurate voiced pitch hypotheses and erroneous voiced/unvoiced (V/UV) decisions can lead to noisy and very undependable feature measurements. This is especially the case for *telephone speech*, due to inferior pitch detection performance caused by the noisy and band-limited telephone channel.

This chapter describes a *continuous* pitch detection algorithm (CPDA), which is designed explicitly to promote robustness for telephone speech and prosodic modeling applications (Wang and Seneff 1998; Wang and Seneff 2000b). It is based on a robust pitch estimation method known as *harmonic matching* (Hess 1983). It also features a dynamic programming (DP) technique of extracting the F_0 value at every frame, regardless of the status of voicing, supplemented by a separate probability of voicing parameter. In the following sections, we first give an overview of our algorithm, emphasizing some design principles. Then we discuss some related work on harmonic matching pitch estimation and DP pitch tracking. Next we describe the pitch tracking and voicing probability estimation modules in detail. Then we evaluate the algorithm using the Keele pitch extraction reference database (Plante et al. 1995), under both studio and telephone conditions, as well as a telephone quality Mandarin digit corpus. We find that the CPDA is very robust

to channel degradation, and compares favorably to an optimized algorithm provided by XWAVES (Talkin 1995) for voiced pitch accuracy. It also significantly outperforms XWAVES when used for a Mandarin four tone classification task.

2.1 Design Principles

We are interested in developing a pitch determination algorithm (PDA) that is particularly robust for telephone quality speech and prosodic modeling applications. Pitch extraction for telephone speech is an especially difficult task, due to the fact that the fundamental is often weak or missing, and the signal to noise ratio is usually low. Frequency-domain pitch detection techniques have been used by previous research to achieve improved robustness for noisy and telephone speech (Schroeder 1968; Noll 1970; Seneff 1978; Martin 1982; Hermes 1988). Following these examples, we adopted a frequency-domain signal representation and developed a robust pitch extraction method which relied on the overall harmonic structure to identify pitch candidates. The pitch estimation method belongs to the category known as the harmonic matching approach, as discussed in (Hess 1983). We derive reliable estimations of both pitch and the *temporal change* of pitch using harmonic matching principles. Furthermore, we combine these constraints in a dynamic programming search to find a smooth and "continuous" pitch contour. In the following, we give a brief overview of the basic design principles of our algorithm.

Our pitch estimation method is based on the observation that harmonics will be spaced by a constant distance on a *logarithmic* frequency scale regardless of the fundamental. More formally, if a signal has harmonic peaks spaced by F_0, then, on a logarithmic scale, the peaks will occur at $\log F_0$, $\log F_0 + \log 2$, $logF_0 + \log 3$, ..., etc. The fundamental F_0 only affects the *position* of the first peak, and the subsequent harmonic peaks have fixed distances from the first peak. Thus, harmonic spectra with different fundamental frequencies can be aligned by simple *linear shifting*. By correlating a logarithmic spectrum with a harmonic template (with known fundamental frequency), we can obtain a robust estimation of the $\log F_0$ of the signal. By correlating two logarithmic spectra from the adjacent frames of a speech signal, we can obtain a very reliable estimation of the $\log F_0$ change.

Instead of determining an F_0 value for each frame by picking the correlation maximum, we use a dynamic programming search to combine the $\log F_0$ and $\Delta \log F_0$ estimations to find an overall optimal solution. We consider all values (quantized in a reasonable pitch range for human speech, e.g., $50Hz$ to $550Hz$) as possible F_0 candidates with different qualities. The quality of a pitch candidate P is indicated by the correlation of the spectrum and the template at the position corresponding to the difference of P and the template fundamental frequency; and the "consistency" of two consecutive pitch candidates is indicated by the correlation of the spectra of the adjacent frames at the position corresponding to the pitch difference. These two constraints are used to define a score function for the DP search, and the path in the quantized $(frequency, time)$ search space with the highest score gives the optimum pitch track.

To deal with *discontinuity* of the F_0 space for prosodic modeling, we believe that it is more advantageous to emit an F_0 value for each frame, even in unvoiced regions, and to provide separately a parameter that reflects the probability of voicing. This is based on the considerations that, first, voicing decision errors will not be manifested as absent pitch values; second, features such as those describing the shape of the pitch contour are more robust to segmental misalignments; and third, a voicing probability is more appropriate than a "hard" decision of 0 and 1, when used in statistical models. The *continuous* pitch tracking capability is implement within the DP search module, by disallowing unvoiced state in the search space. This is feasible also because of a favorable property of the $\Delta \log F_0$ estimation. We will address this point in detail in Sections 2.2 and 2.3.

The pseudo-code of our proposed pitch tracking algorithm is shown in Figure 2-1, and the implementation details will be discussed in Section 2.3. Although the CPDA is based on the same signal representation and pitch estimation framework as the subharmonic summation approach (Hermes 1988), both the signal processing and the tracking technique are quite different. The use of $\Delta \log F_0$ estimation in a dynamic programming search and the technique of "continuous" pitch tracking contribute to favorable properties for prosodic modeling.

N: the total number of frames in an input waveform
M: the total number of quantized pitch candidates
P_i: the quantized pitch candidates ($i = 0, \ldots, M-1$)
T: the harmonic template
X_t: the logarithmic-frequency spectrum at the t^{th} frame of the input waveform
$Score(t, i)$: the partial path score for the i^{th} pitch candidate at the t^{th} frame
begin
    compute T
    compute X_0
    compute the correlation of X_0 and T
    initialize $Score(0, i)$ for all P_i ($i = 0, \ldots, M-1$)
    for $t = 1, \ldots, N-1$
        compute X_t
        compute the correlation of X_t and X_{t-1}
        compute the correlation of X_t and T
        update the partial path score $Score(t, i)$ and
            save the back trace pointer for all P_i ($i = 0, \ldots, M-1$)
    end
    back trace to find the best pitch contour $P(t)$ ($t = 0, \ldots, N-1$)
end

Figure 2-1: The pseudo-code of the proposed continuous pitch tracking algorithm.

2.2 Related Research

A comprehensive study on various pitch determination algorithms is given in (Hess 1983). In this section, we give a brief introduction of two harmonic matching PDAs, namely, the spectral comb method (Martin 1982), and the subharmonic summation method (Hermes 1988), because their underlying principles for pitch estimation are very similar to our method. We also discuss some related work on applying dynamic programming techniques for pitch tracking (Secrest and Doddington 1983; Talkin 1995; Geoffrois 1996; Droppo and Acero 1998).

2.2.1 Harmonic Matching PDA

The two PDAs introduced here can be regarded as a direct implementation of the virtual-pitch theory of human pitch perception (Terhardt 1974; Terhardt 1979). The theory assumes that each spectral component activates not only those elements of the central pitch processor that are most sensitive to that frequency, but also those elements that have a lower harmonic

relation with this component. The contributions of the various components add up, and the element with the highest harmonic activation gives the perceived pitch.

The spectral comb method (Martin 1982) is based on the observation that harmonic peaks of a periodic signal occur at F_0, $2F_0$, $3F_0$, ..., etc. It uses an impulse sequence, called "spectral comb", to match with the signal spectrum. The distance between pulses equals the trial fundamental frequency P. The sum of the signal spectrum multiplied by the spectral comb is used as the harmonic estimation function. The value of P where this function reaches its maximum is taken as the fundamental frequency F_0. To account for the different number of components in the sum for different trial frequencies, the impulses in the spectral comb are weighted by a decaying factor, or the sum is normalized by the number of impulses. In addition, only the frequency components below $2KHz$ are considered, and the spectral values away from local peaks are set to zero.

The subharmonic summation method (Hermes 1988) is based on the observation that on a *logarithmic* frequency scale, harmonic peaks of a periodic signal occur at $\log F_0$, $\log F_0 + \log 2$, $\log F_0 + \log 3$, ..., etc. One can sum up the spectral values spaced by $\log 2$, $\log 3$, ..., from the pitch candidate $\log P$, and the value of P where the sum reaches its maximum gives the pitch estimation. Similar to the spectral comb method, the spectral values are weighted by an exponentially decaying factor in the summation. The logarithmically spaced spectrum is obtained by interpolating a regular FFT spectrum. More specifically, the waveform is first downsampled to $2500Hz$, and an FFT is applied to obtain the spectrum of $[0, 1250Hz]$. The spectral values away from local peaks are set to be zero, after which the spectrum is low-pass filtered. The spectral values on a logarithmic frequency abscissa are then obtained through cubic-spline interpolation.

Aside from their mathematical differences, the two methods are very similar in that they try to match the signal spectrum with an "ideal" harmonic template: in the first case, an evenly spaced impulse sequence; in the second case, a logarithmically spaced impulse sequence. We think that the impulse sequence is an overly simple model to represent a harmonic structure. Small perturbations in harmonic peaks can change the weighted sum, and thus, the outcome of the pitch estimation, which makes the two methods susceptible to noise interference. Furthermore, these two implementations can not avoid the pitch

doubling problem entirely, because harmonic peaks separated by $2F_0$ are also likely to give rise to a large peak in the weighted sum. This is especially the case when a few spectral components are more prominent due to formant energy, or when the fundamental is missing due to the telephone bandwidth. Our harmonic matching method will address these issues specifically, to achieve more accurate pitch estimation, and to improve robustness under adverse signal conditions.

2.2.2 DP Pitch Tracking

The advantage of using dynamic programming in pitch tracking is to incorporate continuity constraints across adjacent frames; thus, a smooth pitch contour can be obtained. It is natural that most of the score functions used for pitch tracking incorporate a transition cost to penalize large changes in neighboring F_0 hypotheses. The DP is usually combined with voiced/unvoiced decision (Secrest and Doddington 1983; Talkin 1995; Geoffrois 1996), because it is unreliable to impose continuity involving unvoiced frames, for which the outcome of frame based pitch extraction is random. For example, (Secrest and Doddington 1983) includes an unvoiced state in the DP search and defines three cost functions: the pitch deviation cost, which specifies the cost for the transition of pitch candidates from the previous frame to the current frame; the voicing state cost, which specifies the cost of the pitch candidates (including unvoiced state) for the current frame; and the voicing transition cost, which specifies the cost for changing voicing status from the previous to the current frame. The pitch deviation cost discourages large pitch changes with the exception of pitch doubling and halving across two frames. The voicing decision is a natural outcome of the DP search using this framework.

One can always force the DP to track pitch *continuously*, by disallowing the unvoiced state in the search space. The maximum *a posteriori* (MAP) pitch tracking method (Droppo and Acero 1998) essentially adopts such an approach. The algorithm uses the normalized auto-correlation function weighted by the energy of the current frame as a probabilistic indication for a pitch hypothesis, and the probability of the pitch change across two frames is modeled by a zero mean Gaussian distribution. The PDA utilizes two DP passes: the first DP pass forms pitch candidates for every frame by maximizing the total likelihood; then,

the second DP pass performs a voicing decision and sets unvoiced frames to be zero. The method compares favorably to two other algorithms on voicing decision errors and standard deviation of relative pitch errors. However, the results for voiced pitch accuracy, in terms of gross errors or the mean of absolute errors, are not reported in the paper. We suspect that the continuous pitch tracking in this case is likely to hurt the pitch tracking performance because of problems during unvoiced to voiced transitions, especially if rigorous continuity constraints are applied.

We prefer a *continuous* DP pitch tracking framework for prosodic modeling considerations. We also believe that such a framework with the appropriate transition cost can lead to improved pitch accuracy for voiced frames. This is based on the intuition that it is less likely to make errors when tracking the pitch for a long sequence of voiced frames, because the values are determined collectively through continuity constraints. When V/UV decisions are tied with DP, the voiced regions are likely to be more fragmented due to voiced to unvoiced decision errors; and the short voiced segments (with a small number of frames) are more likely to cause errors. The transition cost is critical to the performance of a continuous DP tracking approach. In our PDA, we use a transition weighting that is a robust estimation of pitch change in voiced regions, but only imposes very weak continuity constraints during unvoiced speech frames. We found that a continuous DP incorporating this cost function yielded superior performance to a DP search combined with V/UV in our experiments.

2.3 Pitch Tracking

2.3.1 Signal Representation

To obtain a logarithmically spaced spectrum, we directly sample a narrow band spectrum in the low-frequency region $[f_s, f_e]$ at linear intervals in the logarithmic frequency dimension. We define this representation as a *discrete logarithmic Fourier transform* (DLFT). Given a Hamming windowed speech signal $x_t(n)$ ($n = 0, 1, ..., N_w - 1$; and N_w is the window size), the DLFT is computed as follows:

$$X_t(i) = \frac{1}{N_w} \sum_{n=0}^{N_w-1} x_t(n) e^{-j\omega_i n} \qquad (i = 0, 1, \ldots, N-1) \tag{2.1}$$

$$\omega_i = 2\pi e^{(logf_s + i \cdot dlogf)} \cdot T_s \tag{2.2}$$

$$dlogf = (logf_e - logf_s)/(N-1) \tag{2.3}$$

where N is the size of the DLFT, and T_s is the sampling period of the waveform. $dlogf$ can be viewed as the frequency resolution in the logarithmic domain.

Notice that the DLFT formula is derived directly from the Fourier transform, with the frequencies sampled at logarithmic intervals. An alternative approach is to compute the spectrum using the standard FFT algorithm followed by a spline interpolation to resample the frequencies (Hermes 1988). However, the harmonic structure in the high frequency region is likely to be disturbed by the interpolation, due to more compressed spacing on a logarithmic scale at high frequencies.

In the default setting, the DLFT spectrum of the speech signal is sampled between $150Hz$ and $1500Hz$. The lower bound is chosen with consideration of the telephone bandwidth, and the upper bound is chosen to include a sufficient number of harmonic peaks (at least two for high-pitched voices). The spectrum is normalized by a μlaw conversion to reduce the dynamic range of harmonic peak height due to formant influences.

$$X_t(i) = log(1 + 50 \cdot X_t(i)/Max_t)/log(51) \quad (i = 0, 1, ..., N-1) \tag{2.4}$$

where Max_t is the maximum energy of the DLFT spectrum at the t^{th} frame.

$$Max_t = \max_i X_t(i) \tag{2.5}$$

This μlaw conversion holds the maximum component of the DLFT spectrum unchanged while promoting smaller values. A second μlaw is applied to reduce the dynamic range of energy throughout the utterance, for improved V/UV decision.

$$X_t(i) = X_t(i) \cdot log(1 + Max_t/100) \cdot 50 \quad (i = 0, 1, ..., N-1) \tag{2.6}$$

2.3.2 Harmonic Template

We construct the harmonic *template* from an ideal periodic signal, e.g., a $200Hz$ pulse train. The pulse train is first Hamming windowed, after which the DLFT spectrum is computed between $110Hz$ and $1100Hz$. The upper and lower frequency bounds are chosen such that the template includes an appropriate number of harmonic lobes and tapers off to zero at both ends. In our implementation, the harmonic template includes 5 complete harmonic lobes, which will match the DLFT spectrum of a voiced signal of about $270Hz$ (approximately in the middle of the F_0 search range of $50Hz$ to $550Hz$). Similar to the subharmonic summation approach, the energy of each harmonic lobe needs to be normalized. This is done by integrating over each lobe to find its area, followed by a scaling by the reciprocal of the area, subject to an exponential decay. The optimal decay factor is determined to be 0.85 empirically from development data. We also added negative lobes between the positive lobes in the template to discourage pitch doubling, as described in the next section. The negative lobes are obtained by computing the DLFT spectrum of the same pulse train at the following frequencies:

$$\omega_i = 2\pi e^{(logf_s + i \cdot dlogf - log100)} \cdot T_s \tag{2.7}$$

where $logf_s$, $dlogf$, and T_s are the same as in Equation 2.2. The frequency shift of $log100$ causes the spectrum to be shifted by $100Hz$ on the linear scale (half of the fundamental frequency of the pulse train); thus, the harmonic peaks in the new spectrum fall precisely in between those in the original one. This shifted DLFT spectrum is weighted by a negative factor and added to the original spectrum to form the template.

Figure 2-2 shows the waveform, Fourier transform and DLFT for a $200Hz$ pulse train and a voiced speech signal. The DLFT spectra of the signal and the pulse train are adjusted as described above. As illustrated in the figure, the harmonics in the two DLFT spectra can be aligned perfectly by linear shifting despite their F_0 differences; and the relative shift yielding the match is determined by the F_0 differences and the starting frequencies of the two DLFT spectra.

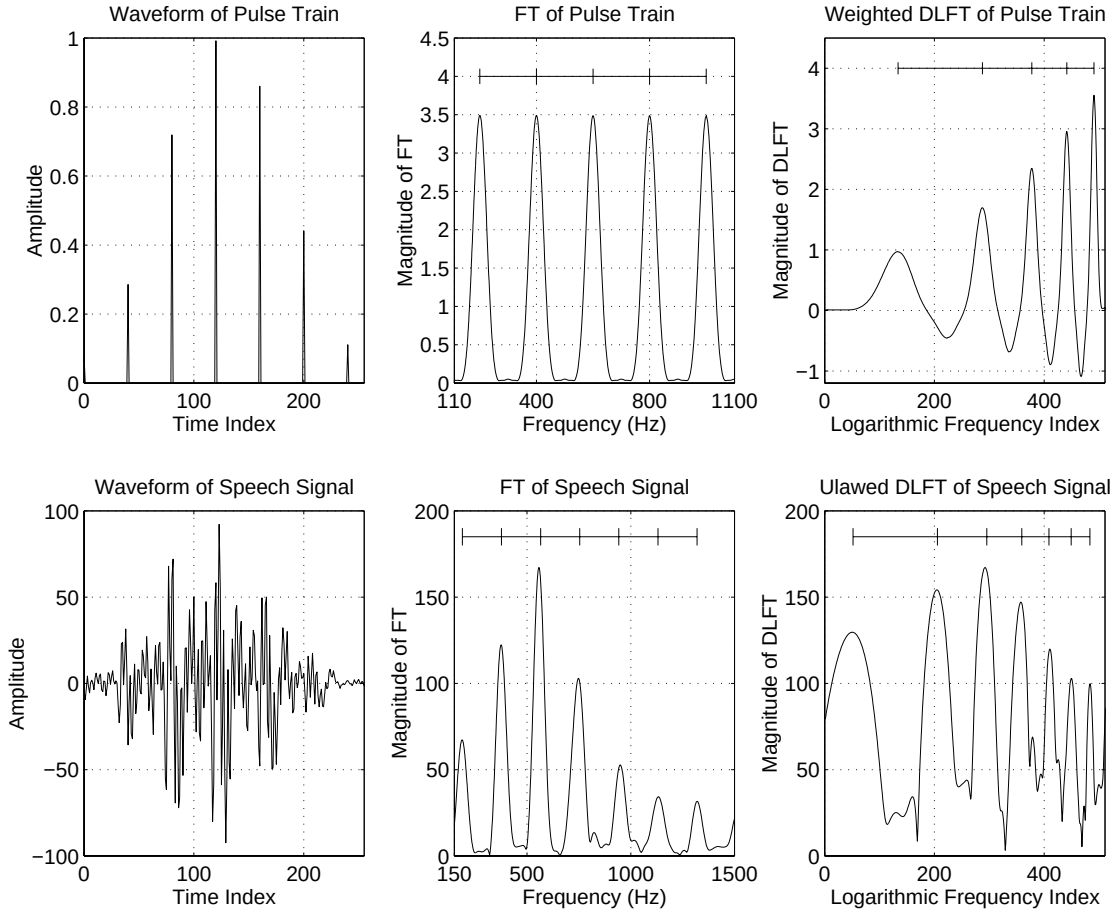

Figure 2-2: Windowed waveform, FT, and adjusted DLFT (refer to the text for details) for a pulse train and a voiced speech signal.

2.3.3 Two Correlation Functions

The "template-frame" correlation function provides a $\log F_0$ estimation by aligning the speech DLFT spectrum with the harmonic template, as shown in Equation 2.8.

$$R_{TX_t}(n) = \frac{\sum_i T(i)X_t(i-n)}{\sqrt{\sum_i X_t(i)^2}} \quad (N_L < n < N_H) \tag{2.8}$$

The template $T(n)$ is the modified DLFT spectrum of a Hamming windowed impulse train of $200Hz$, as described in Section 2.3.2. $X_t(n)$ is the μlaw converted DLFT of the signal at the t^{th} frame. The template is normalized to have unit energy in advance, so the correlation

is normalized by the signal energy only. The bounds for the correlation, $[N_L, N_H]$, are determined by the F_0 range $[F_{min}, F_{max}]$.

The mapping between pitch candidate P and the corresponding index in the template-frame correlation function can be derived from Equation 2.2. Assume the index of the trial pitch P in the signal DLFT spectrum is i_P. According to Equation 2.2, we have

$$\omega_{i_P} = 2\pi \cdot P \cdot T_s = 2\pi e^{(logf_s + i_P \cdot dlogf)} \cdot T_s \tag{2.9}$$

The relationship of P and i_P can be further simplified as:

$$logP = log150 + i_P \cdot dlogf \tag{2.10}$$

$$i_P = (logP - log150)/dlogf \tag{2.11}$$

where 150 is the low frequency bound for the signal DLFT spectrum, and $dlogf$ is the logarithmic frequency resolution. Similarly, the index of the fundamental frequency $(200Hz)$ in the template, i_{200}, can be determined as

$$i_{200} = (log200 - log110)/dlogf \tag{2.12}$$

where 110 is the low frequency bound for the template.

The relative shift to match the two frequencies in the template-frame correlation is the difference of these two indices:

$$I_P = i_{200} - i_P = (log200 - log110 - logP + log150)/dlogf \tag{2.13}$$

Conversely, we can also determine P from the correlation lag I_P by

$$P = \frac{200 \cdot 150}{110 \cdot e^{I_P \cdot dlogf}} \tag{2.14}$$

By substituting P in Equation 2.13 with the pitch range $[F_{min}, F_{max}]$, we obtain the bounds for template-frame correlation as

$$N_L = (log200 - log110 - logF_{max} + log150)/dlogf \quad (2.15)$$

$$N_H = (log200 - log110 - logF_{min} + log150)/dlogf \quad (2.16)$$

The position of the correlation maximum should correspond to the difference of $\log F_0$ between the signal and the template. However, as in all PDAs, frame based peak picking is susceptible to pitch doubling and halving problems. The correlation function has a relatively high peak when the harmonic lobes of the template align with $2F_0, 4F_0, 6F_0, ..., etc.$, of the signal spectrum, especially when the fundamental is missing. The negative lobes added between the positive lobes in the template can help reduce the tendency for pitch doubling, because such an alignment will be penalized by the negative contributions from the $3F_0, 5F_0, ...$ peaks. The weighting of negative lobes was optimized empirically to be 0.35.

The "cross-frame" correlation function provides constraints for $\Delta \log F_0$ by aligning two adjacent frames of the signal DLFT spectra, as shown in Equation 2.17.

$$R_{X_t X_{t-1}}(n) = \frac{\sum_i X_t(i) X_{t-1}(i-n)}{\sqrt{\sum_i X_t(i)^2}\sqrt{\sum_i X_{t-1}(i)^2}} \ (|n| < N) \quad (2.17)$$

The correlation is normalized by the energy of both signal frames. Because F_0 should not change dramatically across two frames, the correlation bound N is set to be about 10% of the number of samples in the DLFT spectrum. The maximum of the correlation gives a robust estimation of the $\log F_0$ difference across two voiced frames.

Figure 2-3 shows examples of the template-frame and cross-frame correlation functions in the voiced and unvoiced regions of a speech signal. For unvoiced regions, it is observed that the template-frame correlation is more-or-less random, and the cross-frame correlation stays fairly flat both within an unvoiced region and upon transition of voicing status. This feature of the cross-frame correlation function is critical for the DP based continuous pitch tracking algorithm.

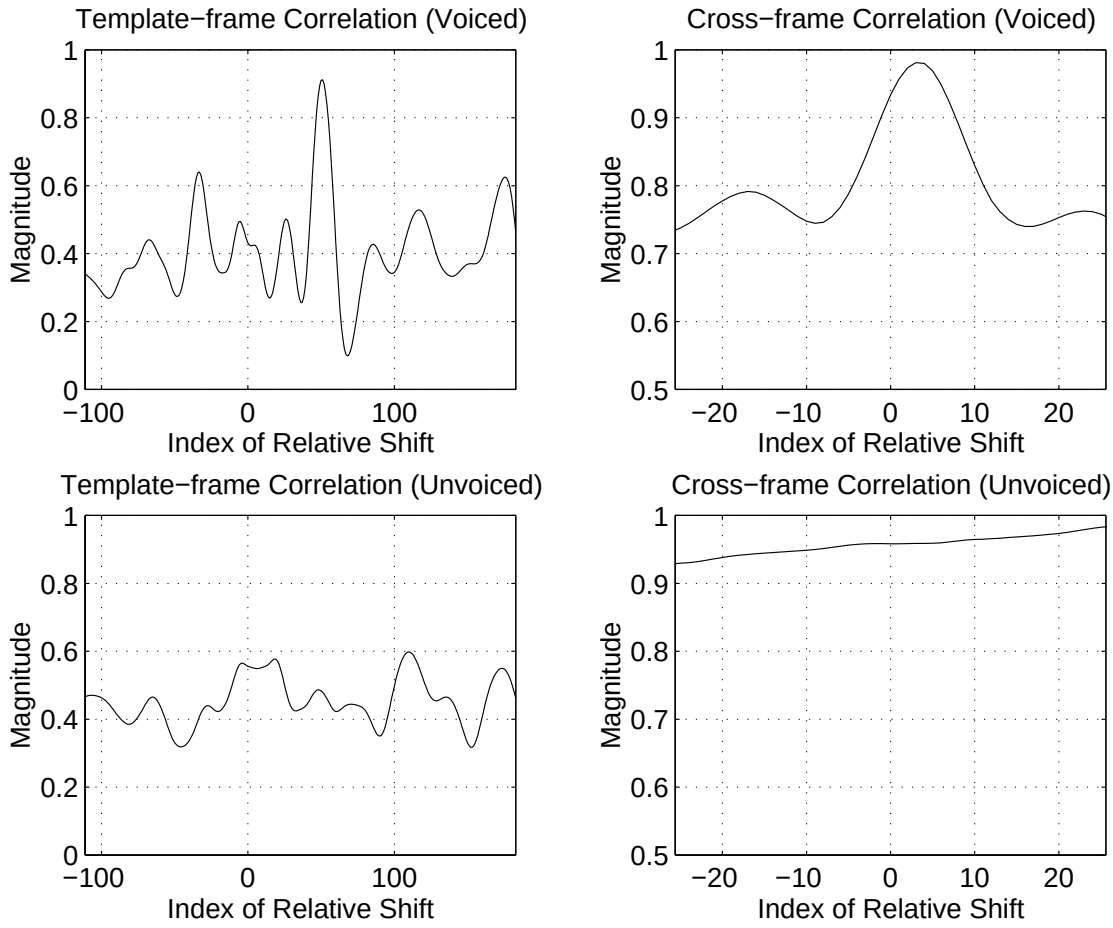

Figure 2-3: Examples of "template-frame" and "cross-frame" correlations for voiced and unvoiced DLFT spectra.

2.3.4 Dynamic Programming Search

Given the constraints for $\log F_0$ and $\Delta \log F_0$, we can easily formulate the problem of pitch tracking as a DP search. We define the target function in an iterative manner as

$$Score(t,i) = \begin{cases} R_{TX_0}(i) & (t=0) \\ \max_j \{Score(t-1,j) \cdot R_{X_t X_{t-1}}(i-j)\} + R_{TX_t}(i) & (t>0) \end{cases} \tag{2.18}$$

where i is the index in the template-frame correlation function. The pitch value P_i can be

converted from the index i by Equation 2.14 . The search is extended by inheriting the best past score as *weighted* by the cross-frame correlation plus the template-frame correlation for the current node. The pointer to the best past node is saved for back tracking upon arriving at the last frame. Due to the logarithmic sampling of the DLFT, the search space for pitch value is naturally quantized logarithmically, with constant $\Delta F_0/F_0$.

The target function ensures a very smooth pitch contour. An expansion of Equation 2.18 reveals that the internal score of a particular node on the path is weighted by a *series* of cross-frame weights from that node to the current node before contributing to the cumulative score. We also tried replacing the multiplication in Equation 2.18 with addition. This score function imposes constraints only across the neighboring frames. We obtained slight performance improvement in pitch accuracy, because the search is more flexible to follow abrupt changes in the pitch contour, such as those caused by glottalization. However, we think such sensitivity is less robust for prosodic modeling, and thus, did not pursue it further.

The DP search is forced to find a pitch value for every frame, even in unvoiced regions. We experimented with adding a node for unvoiced state in the search and incorporating the voicing probability into the target function. We found that this increased the number of pitch errors propagated from the voicing decision errors. It is observed that the cross-frame correlation stays relatively flat when at least one frame is unvoiced, as indicated in Figure 2-3. Thus, upon transition into unvoiced regions, the best past score will be inherited by *all* nodes; and the scores become somewhat random. However, once in voiced regions, the sequence of nodes corresponding to the true pitch values will emerge because of high internal scores enhanced by high cross-frame correlation coefficients.

Figure 2-4 shows the waveform, DLFT spectrogram, and phonetic and word transcriptions for a telephone utterance. The DLFT spectrum is computed in the $[150, 1200]\,Hz$ range. The search space for F_0 is from $50Hz$ to $550Hz$, part of which is overlayed with the DLFT spectrogram. As shown in the figure, the first harmonic of the spectrum is fairly weak; nevertheless, the DP search is able to track F_0 whenever there is clear harmonic structure. The pitch track in unvoiced regions is arbitrarily chosen by the search and probably does not have a meaningful interpretation.

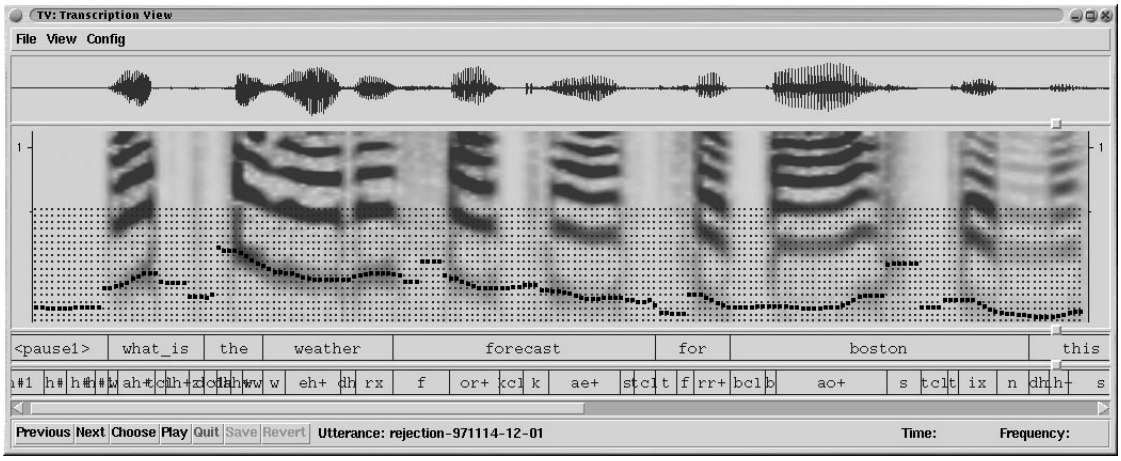

Figure 2-4: Waveform, DLFT spectrogram and transcriptions for the utterance "What is the weather in Boston this ...". Part of the quantized search space for F_0 and the chosen path are overlayed with the DLFT spectrogram.

2.4 Voicing Probability Estimation

To increase robustness for statistical modeling, the voicing decision module computes a voicing probability for each frame instead of making a hard decision.

The *posterior* probabilities of a frame being voiced/unvoiced can be obtained from the observation $\vec{O}$ by applying Bayesian Rules as shown in Equation 2.19, where V stands for voiced, and U for unvoiced. $P(V)$, $P(U)$, $P(\vec{O}|V)$ and $P(\vec{O}|U)$ can be obtained *a priori* from training data.

$$\begin{cases} P_V = P(V|\vec{O}) & = P(\vec{O}|V)P(V)/P(\vec{O}) \\ P_U = P(U|\vec{O}) & = P(\vec{O}|U)P(U)/P(\vec{O}) \\ P(\vec{O}) & = P(\vec{O}|U)P(U) + P(\vec{O}|V)P(V) \end{cases} \tag{2.19}$$

The observation vector includes two elements from the pitch tracking algorithm. One is the maximum of the unnormalized template-frame correlation, which can be interpreted as the "harmonic energy" of the signal. The second element is the minimum of the cross-frame correlation. It is small for voiced frames and close to 1 for unvoiced frames. We use the minimum of the forward and the backward cross-frame correlations to improve the prediction for the first and last frames of voiced regions, following the example in (Droppo

and Acero 1998). We also added the total signal energy and zero-crossing rate to the feature vector because they improve the voicing decision accuracy. Mixtures of diagonal Gaussian models were used to model the prior distributions $P(\vec{O}|V)$ and $P(\vec{O}|U)$.

2.5 Evaluation

A PDA is usually evaluated on two aspects: pitch estimation and voicing decision (Rabiner et al. 1976). Accuracy for voiced pitch estimation can be evaluated in terms of "gross error" rate (GER), which is the percentage of voiced hypotheses that deviate from the reference by a certain amount (often 10% or 20%), and the mean and variance of the *absolute* value of the error. The GER is a good indication of pitch doubling and halving errors, while the mean and variance of absolute error examines the deviation of hypothesized pitch from the reference. The voicing decision can be evaluated by the sum of voiced to unvoiced and unvoiced to voiced errors. Since the CPDA does not make an explicit voicing decision, we will focus the evaluation on *voiced frames*. Our final goal is to apply the CPDA in prosodic modeling. In this regard, we also evaluated telephone quality Mandarin tone classification performance using the CPDA for pitch tracking.

We compared the performance of the CPDA with an optimized algorithm provided by XWAVES in these aspects. The XWAVES PDA is based on the robust algorithm for pitch tracking (RAPT) method (Talkin 1995), which is a standard time-domain PDA. It relies on peaks in the normalized cross-correlation function to generate pitch candidates. A post-processing with dynamic programming is applied to select the best F_0 and voicing state candidates. The XWAVES PDA is already optimized using a large amount of hand labelled speech data. We will use the default setting for all internal parameters of XWAVES. To ensure similarity, both PDAs are set to have an F_0 search range of $50Hz - 550Hz$, and a frame rate of $100Hz$.

2.5.1 Voiced Pitch Accuracy

Keele Database

We use the Keele pitch extraction reference database (Plante et al. 1995) for this evaluation, because it provides reference pitch obtained from a simultaneously recorded laryngograph trace as "ground truth". There are five male and five female speakers[1], each speaking a phonetically balanced text of about 35 seconds. The speech data were recorded in a soundproof room using a head mounted microphone with a sampling rate of $20KHz$. In order to evaluate the PDAs under telephone conditions, we transmitted the waveforms through a noisy telephone channel and recorded at a sampling rate of $8KHz$. The transmitted waveforms were carefully calibrated with the originals to make sure that the pitch references are still valid.

The reference pitch in the Keele database is computed from the laryngograph trace using a floating autocorrelation of $25.6ms$ duration at $10ms$ interval, and is later manually validated. Besides "clearly unvoiced" and "clearly voiced" frames, there also exist some "uncertain" frames, where visual inspection of the laryngograph and the speech waveform reveals inconsistency in periodicity. When periodicity is observed in the laryngograph but not in the speech waveform, the uncertain frames are labelled with the negative of the reference pitch obtained from the laryngograph. When periodicity is observed in the speech trace but not in the laryngograph, the frame is labelled with "-1". The number and percentage of frames in each category are listed in Table 2-1. As reported in (Plante et al. 1995), most of the uncertain frames occur at voicing onsets and around plosive bursts. We use only the "clearly voiced" frames for evaluation as recommended by (Plante et al. 1995), because it is simply not clear what the pitch of the questionable frames should be. We also suspect that the performance at bursts and voicing onsets, where most of the uncertain frames occur, does not have a significant impact on the intended applications.

Since we do not have other verified data to optimize the parameters of the CPDA, we set aside two speakers (f1 and m1) as the development data, and tested on the remaining eight speakers. After optimization, the same parameters are used for the CPDA in all

[1](Plante et al. 1995) also reported data collected from 5 child speakers. However, the data were not present at the ftp site for downloading.

Category	Number of Frames	Percentage of Total (%)
Clearly Unvoiced (0)	15583	46.23
Clearly Voiced (*value*)	16960	50.31
Uncertain ($-value$)	659	1.95
Uncertain (-1)	509	1.51
Total	33711	100

Table 2-1: Number and percentage of frames in each category in the Keele database.

experiments including Mandarin tone classification.

Results and Analysis

Figure 2-5 illustrates some favorable features of the CPDA as compared to XWAVES. The display window shows (from top to bottom) the waveform, narrow-band spectrogram, DLFT spectrogram with pitch contour obtained by CPDA, reference pitch track provided by the Keele database, and the pitch track extracted using XWAVES, for a few speech segments taken from the Keele database. As observed from the figure, XWAVES is likely to make voiced to unvoiced decision errors at segmental transitions where change in the waveform shape is large (examples highlighted by arrows with "A" in the figure), or at the end of voiced regions where the signal is weak (example highlighted by an arrow with "B" in the figure). However, the CPDA is able to avoid these errors because it tries to track pitch for every frame, and the DLFT spectra at those frames still have sufficient information about the harmonic structure. It seems that the CPDA makes some "pitch doubling" errors where there appear to be some irregularities with the speaker's voicing (examples indicated by arrows with "C" the figure). This is due to the strong continuity constraints imposed in the DP tracking score function. We believe that a smooth pitch contour is more appropriate for prosodic modeling, so we did not try to relax the continuity constraints to correct for those errors.

Because XWAVES makes both gross errors and voicing decision errors on voiced frames, we divide the data into two subsets based on the outcome of XWAVES' V/UV decision, and summarize the performance for each subset separately, as shown in Table 2-2. The table gives both 20% GER and mean and standard deviation on absolute errors. The overall error

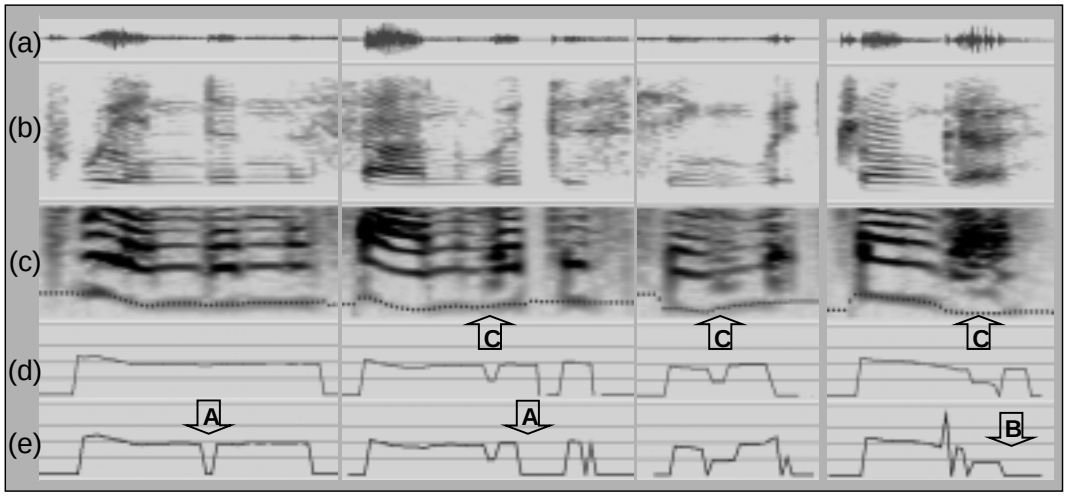

Figure 2-5: (a) Waveform, (b) narrow-band spectrogram, (c) DLFT spectrogram with pitch extracted using CPDA, (d) reference pitch track, and (e) pitch track extracted using XWAVES for some telephone quality speech segments in Keele database.

rate counts a voicing error as equivalent to a 20% GER. All percentages are with reference to the total number of voiced frames used in evaluation.

The performance of the CPDA is very robust to signal degradation, as indicated by the similar performances under studio conditions (4.25% total GER) and telephone conditions (4.34% total GER). This is achieved by using the band-limited frequency domain signal representation and relying on the overall harmonic structure to derive pitch estimates. The "missing fundamental" problem of telephone speech has virtually no impact because the CPDA ignores frequency information under $150Hz$. As expected, the CPDA is less accurate for the "XWAVES:V" subset under studio quality, because it does not utilize all available information, and favors a smooth contour. However, only 15% of the frames erroneously classified by XWAVES as unvoiced contain gross errors in the CPDA track (1.01% vs. 6.63% of all data). Furthermore, CPDA performs substantially better than XWAVES on both subsets for telephone speech.

We have observed that CPDA performs better for female speech than for male speech. When F_0 is low, the template-frame correlation suffers from missing low harmonics, and the cross-frame correlation suffers from compact spacing of higher order harmonics. This

Configuration		XWAVES:V			XWAVES:UV		Overall
		GER (%)	**Mean** (Hz)	**Std.** (Hz)	**V→UV** (%)	**GER** (%)	**Overall** (%)
Studio	XWAVES	1.74	3.81	15.52	6.63	-	8.37
	CPDA	3.24	4.61	15.58	-	1.01	4.25
Telephone	XWAVES	2.56	6.12	25.10	20.84	-	23.41
	CPDA	2.10	4.49	14.35	-	2.24	4.34

Table 2-2: Summary of performance on "clearly voiced" reference frames. Under each signal condition, the *voiced* reference data are divided into two subsets according to whether XWAVES determines them to be voiced, i.e., XWAVES:V and XWAVES:UV. All percentages are with reference to the *total number* of "clearly voiced" frames.

can potentially be improved by using gender-dependent parameters, or by using multiple frequency ranges for the DLFT.

2.5.2 Tone Classification Accuracy

We have demonstrated that CPDA has superior voiced pitch accuracy performance on telephone speech compared with XWAVES. We now examine if the advantage is carried over to prosodic modeling applications. In this regard, we compared CPDA with XWAVES on a tone classification task using a telephone-quality, Mandarin digit corpus.

The digit corpus contains a training set of 3900 utterances, and a test set of 355 utterances. The F_0 contour for each utterance was first normalized by its average to reduce cross-speaker differences. Tonal features are extracted from the syllable rhyme; they include 4 Legendre coefficients of the F_0 contour, and the duration. Two sets of experiments are conducted with and without an additional average probability of voicing feature. Refer to Chapters 3 and 4 for detailed descriptions of Mandarin tones and tone classification.

Results and Analysis

As summarized in Table 2-3, the result using CPDA (d) for pitch tracking is significantly better than that using XWAVES (a). We suspect that gaps in the pitch contours using XWAVES may be blamed for the inferior classification performance. We tried two approaches to dealing with the unvoiced frames when using XWAVES: (b) interpolate F_0 from the surrounding

Configuration	Error Rate w/o P_v (%)	Error Rate w/ P_v (%)
(a) XWAVES	25.4	25.6
(b) XWAVES (intp'd)	24.1	23.6
(c) XWAVES (biased)	24.9	25.4
(d) CPDA	19.2	18.2

Table 2-3: Summary of tone classification error rate.

Digit	Tonal Pinyin	CPDA	XWAVES
0	ling2	31.9	51.0
1	yi1	4.6	18.5
2	er4	18.5	24.4
3	san1	12.8	18.8
4	si4	24.4	27.9
5	wu3	38.0	39.4
6	liu4	10.9	16.0
7	qi1	9.5	18.6
8	ba1	6.5	15.3
9	jiu3	27.2	25.1

Table 2-4: Summary of tone classification error rate (in percentage) for each digit.

voiced frames, and (c) bias the V/UV decision threshold to greatly favor "voiced" decisions, followed by interpolation. As seen in the table, neither of them are particularly successful.

Table 2-4 summarizes the tone classification error rate for each digit by the CPDA system (d) and the best system of XWAVES (b). The largest performance gap between CPDA and XWAVES occurs for the digit "*yi*1", which is a front vowel with a low first formant in the region of the second harmonic. XWAVES is likely to make pitch halving errors because of the strong second harmonic, especially when the fundamental is mostly filtered out by the telephone bandwidth. Digits '*qi*1" and "*ling*2", which also have front vowels, cause similar problems for XWAVES. This explains the large degradation of performance of XWAVES on digits with tone 1 and 2. The error rates on tone 3 for both PDAs are similar.

Figure 2-6 shows the bubble plots of the confusion matrices from the two classification results, excluding the diagonal elements. It is observed that the confusion of tone 1 as tone

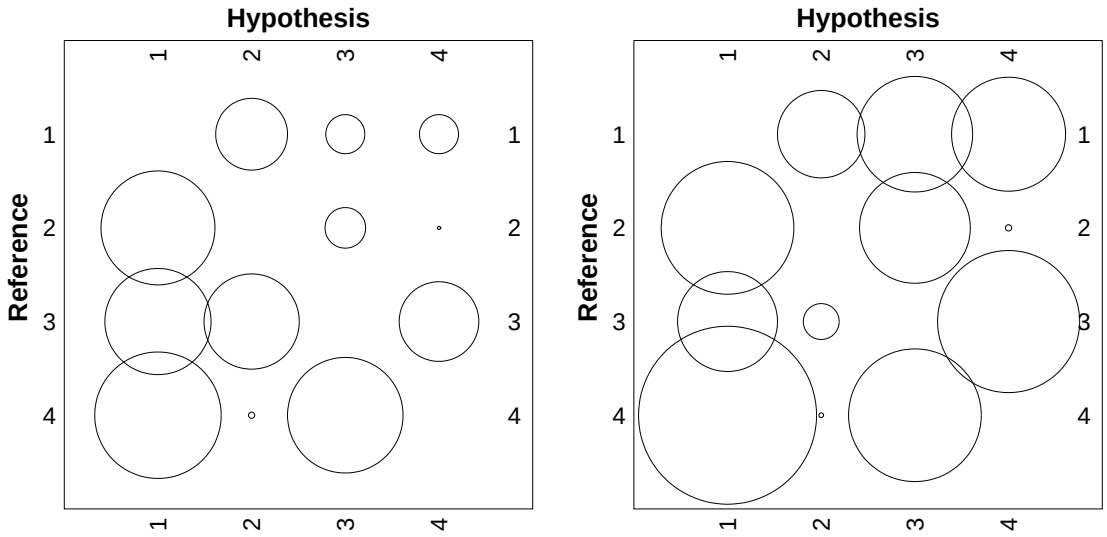

Figure 2-6: Bubbles of classification errors for CPDA (left) and XWAVES (right).

3 or tone 4, the confusion of tone 2 as tone 3, and the confusion of tone 4 as tone 1 are all much greater when using XWAVES. This is consistent with our previous analysis that high pitch halving errors may be blamed. High pitch halving errors for "yi1" (with tone 1) cause the model for tone 1 to be blurred with tone 3 (when the whole segment is halved) or tone 4 (when only the later portion of the segment is halved); and pitch halving errors for "ling2" (with tone 2) cause the model for tone 2 to be blurred with tone 3. The only case where the error rate for CPDA is higher is the tone 3 to tone 2 confusion. This is probably because glottalization often occurs at the end of tone 3 due to the low falling pitch. The pitch contour extracted by CPDA at those frames is likely to be random, which often results in an overall rising slope. The "33 $\rightarrow$ 23" tone sandhi rule might also contribute to the high error rate.

2.6 Summary

In this chapter, we presented a robust pitch tracking algorithm for telephone speech and prosodic modeling. The algorithm derives reliable estimations of pitch and the temporal change of pitch from the entire harmonic structure. The estimations are obtained easily

with a logarithmically sampled spectral representation, because signals with different F_0 can be aligned by simple *linear shifting*. The correlation of the DLFT spectrum with a carefully constructed harmonic template provides a robust estimation of the F_0. The correlation of two DLFT spectra from adjacent frames gives a very reliable estimation of the F_0 change. The constraints for both $\log F_0$ and $\Delta \log F_0$ are then combined in a dynamic programming search to find a very smooth pitch track. The DP search is able to track pitch continuously regardless of the voicing status, while a separate voicing decision module computes a probability of voicing per frame. We demonstrated that the CPDA is robust to signal degradation inherent in telephone speech. In fact, the overall GER for studio and telephone speech is nearly the same (4.25% vs. 4.34%). We also demonstrated that the CPDA has superior performance for both voiced pitch accuracy and Mandarin tone classification accuracy compared with the optimized algorithm in XWAVES.

Chapter 3

Analysis of Tonal Variations for Mandarin Chinese

For a tonal language such as Chinese, fundamental frequency plays a critical role in characterizing tone, which is an essential lexical feature. In this regard, we have focused our initial study of prosody on the F_0 contours of Mandarin Chinese. We believe that, unlike the obscure correlation of prosodic features with stress in English, the syllable level F_0 contour in Mandarin clearly defines tone; and other prosodic aspects, such as intonation, can be studied within the context of improving tone modeling.

There are four lexical tones in Mandarin Chinese, each defined by a canonical F_0 contour pattern: high-level (tone 1), high-rising (tone 2), low-dipping (tone 3), and high-falling (tone 4). The F_0 contour of a syllable spoken in isolation generally corresponds well with the canonical pattern of its tone, although there exists variability due to vowel intrinsic pitch, perturbation by the initial consonant, and the pitch range of a speaker, as well as other individual differences. In addition to these variations, tones in continuous speech undergo both phonological and phonetic modifications due to *tone sandhi*[1] and tone coarticulation, which can cause the F_0 contours to significantly deviate from the canonical forms. Tones can also be influenced by many other linguistic and paralinguistic commands, such as phrase grouping, sentence-level stress or focus, F_0 declination and downstep, sentence mode, emotion,

[1]Tone sandhi refers to the phenomenon that, in continuous speech, some lexical tones may change their tonal category in tonal context.

etc.

This chapter and the next chapter investigate the use of tone models to improve Mandarin Chinese speech recognition performance (Wang and Seneff 1998; Wang and Seneff 2000a). In this chapter, we present empirical studies of a number of factors contributing to Mandarin tone variations. The goal of this study is to gain some understanding of the phonology and phonetics of Mandarin tone and intonation, and to provide some guidance for improving statistical modeling of Mandarin tones as described in the next chapter. In the following sections, we first give some background knowledge of Mandarin Chinese. Then we present some related work on Mandarin tone and intonation studies. After that, we describe two Mandarin speech corpora used in our tone and speech recognition experiments: a Mandarin digit corpus, which consists of read random digit strings and phone numbers; and the YINHE corpus, which contains human-computer conversational speech. Finally, we analyze a number of factors that contribute to the phonetic variations of lexical tones, mainly using the Mandarin digit database. These factors include the overall F_0 declination of a sentence, as well as the presence of a phrase boundary, tone coarticulation, and tone sandhi. The relative contributions of these factors to tone recognition and speech recognition performances, when incorporated into tone modeling, will be presented in the next chapter.

3.1 Background on Mandarin Chinese

The Chinese language is ideographic and tonal-syllabic, in which each character represents a syllable with a particular tone, and one or more characters form a "word." For example, the Chinese word for "telephone" is made up of two characters: "电话". It is pronounced as "*diàn huà*" (*tonal pinyin*[2] transcription), with a falling tone for both the syllable "*dian*" and the syllable "*hua*". The tonal pinyin can also be written as "*dian*4 *hua*4", with the tone represented by a number. The syllable structure of Mandarin Chinese is relatively

[2] "Pinyin" is a phonetic transcription system widely used to describe Mandarin Chinese syllables. The nucleus vowel or diphthong is usually accented to signal the tone of a syllable, e.g., *dā, dá, dǎ, dà*. Alternatively, the tone can be marked by appending a numerical index at the end of the *pinyin* transcription, i.e., *da*1, *da*2, *da*3, *da*4. The pronunciation of a syllable is completely determined with the tonal pinyin representation.

<table>
<tr><td colspan="9">word
(电话 : telephone)</td></tr>
<tr><td colspan="4">character
(电 : electricity)</td><td rowspan="3"></td><td colspan="4">character
(话 : speech)</td><td rowspan="3"></td></tr>
<tr><td colspan="3">syllable
(dian)</td><td>tone
(4)</td><td colspan="3">syllable
(hua)</td><td>tone
(4)</td></tr>
<tr><td>[initial]
(d)</td><td colspan="3">final
(ian)</td><td>[initial]
(h)</td><td colspan="3">final
(ua)</td></tr>
<tr><td>[medial]
(i)</td><td>vowel
(a)</td><td>[nasal]
(n)</td><td>[medial]
(u)</td><td>vowel
(a)</td><td>[nasal]
-</td></tr>
</table>

Figure 3-1: An example illustrating the hierarchical relationship of words, characters, syllables, tones, syllable initials, syllable finals, and phonemes for Mandarin Chinese. Pinyin symbols are used to illustrate the decomposition of syllables, which do not always correspond to phonemic transcriptions. The optional components of the syllable structure are indicated by square brackets.

simple. It is commonly described by the syllable *initial* and $final$ (Wu and Lin 1989): the syllable initial can be a single consonant or null; the syllable final always consists of a vowel or diphthong, preceded by an optional medial glide and followed by an optional nasal ending (/*n*/ or /*ng*/). The entire sound inventory contains a total of 22 initials (23 if the null initial is included) and 38 finals in Mandarin Chinese. There are nearly 60,000 commonly used characters in the language, mapping to about 410 base syllables, or 1200 tonal syllables if distinctions in tone are considered. Disambiguation relies heavily on context, by identifying multi-syllable words or phrases from the string of tonal syllables. Figure 3-1 illustrates the hierarchical relationship of these elements with the Chinese word "电话" (*dian*4 *hua*4, telephone)[3] as an example.

There are four lexical tones and a neutral tone in Mandarin Chinese. As shown in Figure 3-2, the average F_0 contours of the four lexical tones match well with the canonical definition when spoken in isolation. The four lexical tones can be grouped according to the F_0 values at the beginning or the end of the tone contour, as summarized in Table 3-1. These groupings are useful for discussing the coarticulatory effects of tones. Unlike the

[3]We generally use tonal pinyin to represent Chinese words, with the English translations shown in parentheses. When the pinyin representation is insufficient, however, Chinese characters will be used, with both the pinyin and the English translation shown in parentheses.

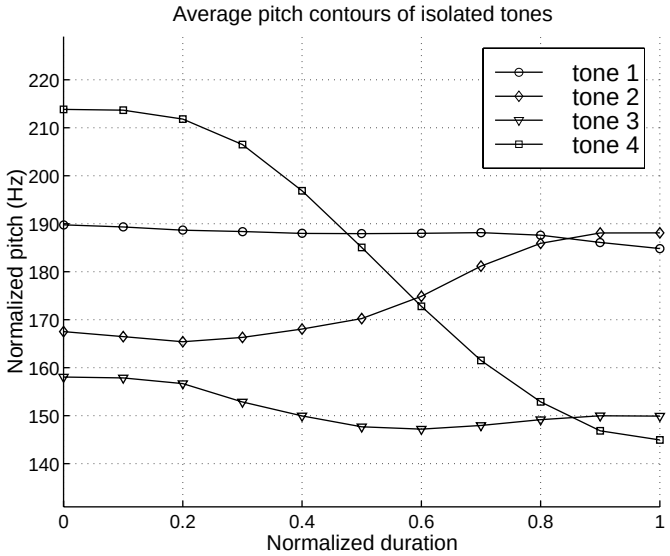

Figure 3-2: Average pitch contours of four lexical tones when pronounced in isolation. The time scale of each token is normalized by its duration. The data are selected from the Mandarin digit database. Since the database consists of multi-digit strings, we consider a digit to be "isolated" if it is bounded by silences or utterance boundaries on both sides.

	High	Low
Onset	tone 1, tone 4	tone 2, tone 3
Offset	tone 1, tone 2	tone 3, tone 4

Table 3-1: Groupings of lexical tones according to the onset and offset F_0 values.

lexical tones, the neutral tone does not have a stable F_0 contour pattern. It is usually associated with reduced duration and energy, while the F_0 contour depends largely on the surrounding tones (Wu and Lin 1989).

3.2 Related Research

3.2.1 Tone Sandhi

Tone sandhi refers to the *categorical* change of a tone when spoken in the context of other tones. The most well-known tone sandhi rule in Mandarin is the third tone sandhi rule,

which states that a low tone is changed to a rising tone when followed by another low tone (Chao 1968). It is supported by perceptual experiments that the changed tone is indistinguishable from the true rising tone (Wang and Li 1967). The third tone sandhi becomes quite complex when there are more than two third tones in a row, and the expression of the rule is found to depend on the prosodic structure rather than on the syntax (Shih 1986). For example, the sentence "*lao*3 *li*3 *mai*3 *hao*3 *jiu*3" (old Li buys good wine) can be turned into "*lao*2 *li*2 *mai*3 *hao*2 *jiu*3", giving a perceived phrase boundary between "*mai*3" (buy) and "*hao*3" (good). However, the surface realization for this utterance is not unique. As discussed in (Shattuck-Hufnagel and Turk 1996), Cheng found that the sentence was more likely to be read as "*lao*2 *li*3 *mai*3 *hao*2 *jiu*3" when spoken slowly.

Another sandhi rule concerning tone 2 is somewhat debatable, which states that a rising tone changes into a high tone when preceded by a high or rising tone and followed by any other tones (Chao 1968). In (Shih and Sproat 1992), however, it has been found that a rising tone surrounded by high tones still has different F_0 contours from the high tone. A perceptual study reported in (Xu 1994) further showed that most of the rising tones after a high tone were still perceived as the rising tone, even though the F_0 contours were flattened. These observations seem to suggest that the rising tone variation after high offset tones is due to tone coarticulation, rather than a phonological change of the intended tone category.

Some tone sandhi changes depend not only on the surrounding tones, but also on the lexical properties of the affected syllable. For example, the tonal variations for "不" (*bu*4, not) do not happen for its homophones such as "部" (*bu*4, part) or "步" (*bu*4, step). When "不" (*bu*4, not) is followed by a falling tone, it is changed to a rising tone; for its homophones, the falling tone remains unchanged (page 78 in (Modern Chinese Dictionary 1978)). Similarly, the tonal variations for "一" (*yi*1, one) do not happen for its homophones such as "医" (*yi*1, cure). The variation rules are quite complex: the high tone remains unchanged when "一" (*yi*1, one) is used as a word or as the last syllable of a word; otherwise, it is changed to a rising tone if the following syllable in the word has a falling tone, and a falling tone if the following syllable has a high, rising, or low tone (page 1337 in (Modern Chinese Dictionary 1978)).

In speech recognition, some tone sandhi rules, such as those regarding "一" (*yi*1, one)

and "不" (*bu4*, not), can be encoded directly in the lexicon by specifying the correct surface form in the word (if all the necessary conditions can be determined), or by allowing all possible forms as alternative pronunciations. The third tone sandhi rule is more difficult to incorporate, because it is not specific to a particular word, and the surface realization can not be uniquely determined from text analysis. Our solution is to use context dependent models to account for the behavior of tone 3 before another tone 3. This approach seems to be capable of capturing the majority case, i.e., the expression of third tone sandhi in the absence of a phrase boundary, as indicated by our analysis in Section 3.4.4.

3.2.2 Tone Coarticulation

Besides the phonological sandhi changes, tones in continuous speech are also influenced by the neighboring tones due to articulatory constraints.

Shen (1990) analyzed all possible combinations of Mandarin tones on "*ba ba ba*" tri-syllables embedded in a carrier sentence and found that both anticipatory and carry-over effects existed, and they were both assimilatory in nature. The coarticulatory effects changed not only the onset and offset F_0 values, but also the overall tone heights.

However, Xu (1997) studied F_0 contours of Mandarin bi-syllables "*ma ma*" embedded in a number of carrier sentences, and arrived at somewhat different conclusions. He found that anticipatory and carry-over effects differed both in magnitude and in nature: the carry-over effects were larger in magnitude and mostly assimilatory in nature, e.g., the onset F_0 value of a tone was assimilated to the offset value of a previous tone; the anticipatory effects were relatively small and mostly dissimilatory in nature, e.g., a low onset value of a tone raised the maximum F_0 value of a preceding tone.

It is possible that the discrepancies of these two studies arose from the relatively small databases used in the studies. In particular, (Shen 1990) used a total of 400 tri-syllables spoken by two speakers, with 4 tokens per tri-tone combination. We will conduct an empirical study of tone coarticulation using the Mandarin digit corpus, which contains much more data in terms of both the amount of speech and the number of speakers. The results might be slightly influenced by factors other than tone coarticulation, because of the less controlled linguistic and prosodic conditions. However, this is a more realistic scenario for

tone and speech recognition systems.

3.2.3 Tone and Intonation

The interaction of tone and intonation is still not very well understood. A study was conducted on a small set of read utterances to investigate if intonation can change the tone contours to beyond recognition (Shen 1989). It was found that both the shape and the scale of a given tone were perturbed by intonation. For example, interrogative intonation raises the tone value of the sentence-final syllable as well as the overall pitch level; tone 1 rises slightly in sentence initial position and falls slightly in sentence-final position under statement intonation; etc. However, it was concluded that the basic tone shape is preserved, e.g., the falling tone did not become falling-rising under question intonation.

A few studies have been conducted on the interaction between tone and sentence focus (Gårding 1987; Xu 1999). It has been found that the lexical tone of a syllable is the most important factor for determining the local F_0 contour, while the focus extensively modulates the global shape of the F_0 curve. The effects of focus are asymmetric: the F_0 range of words at a non-final focus is substantially enlarged, especially the high value; the F_0 range after the focus is both lowered and reduced; and the F_0 range before the focus remains similar to the case with no focus.

Downstep and declination are important aspects of intonation. Downstep refers to the phenomenon that a high (**H**) pitch target has lower F_0 height after a low (**L**) pitch target; while declination refers to the tendency for F_0 to gradually decline over the course of an utterance. A broad term "downtrend" is used to describe the combined effects of the two. It is argued in (Pierrehumbert 1980; Liberman and Pierrehumbert 1984) that the declination in English is the outcome of the downstep of subsequent **H** pitch accents throughout the utterance. However, a study of Mandarin tone 1 syllable sequences (Shih 1997) has found that Mandarin has a strong declination effect with or without any sentence focus. The F_0 decline can be modeled as an exponential decay, and the slopes of the decline are influenced by sentence focus as well as by the length of the sentence.

As indicated in (Shen 1989), tones at sentence initial and final positions seem to behave differently from at other positions. We will conduct a systematic study of the influence of

phrase boundaries (including the beginning and the end of an utterance) on the four lexical tones using the digit corpus. We will also try to characterize the *overall* F_0 downtrend on both the digit and the YINHE corpora. Since we can not restrict the tone sequence to be a particular pattern in "real" data, we designed a simple method to "remove" the local F_0 changes caused by tones, thus, revealing the global "intonational" contour. A few interesting observations are made from the data, including the relationship of the declination with phrase grouping, sentence length, sentence mode, etc. However, the effects of focus on tones will not be addressed in this thesis. This is because automatic detection of focus is a challenging task by itself, and its detection is likely to rely on tone information. This is a potential direction for future work as discussed in Chapter 8.

3.2.4 Domain of Tone

How tone contours align with other linguistic units in speech is an important issue for tone modeling. Specifically, which part of the F_0 contour in the syllable carries tone information?

Howie (1974) argued that tones in Mandarin were carried only by the syllable rhyme (vowel and nasal), while the portion of the F_0 contour corresponding to an initial voiced consonant or glide is merely an adjustment for the voicing of initial consonants. His argument was based on the observation that there was much F_0 perturbation in the early portion of a syllable. Xu confirmed in (Xu 1997) that the nasal part of a syllable carried tone information: the tone patterns remained consistent across syllables with and without a final nasal, and the movement of F_0 such as falling or rising continues all the way to the end of the syllable. However, he observed in (Xu 1998) that, in coarticulated tones, the F_0 contour of syllable "*ma*" immediately moved toward the first target of the next tone at the syllable onset instead of the rhyme. He thus argued that the syllable was the appropriate domain for tone alignment; and the large perturbation seen at the early portion of the F_0 contour of a syllable was the result of the carry-over effects from the preceding tone.

From a modeling point of view, it seems more advantageous to extract tone features from the F_0 contour of the syllable rhyme (vowel and nasal), because the large perturbation at the early portion of a syllable is less relevant to the current tone, and is likely to introduce "noise" that is dependent on the syllable structure. The syllable final is similar to the

syllable rhyme except for an optional medial (refer to Figure 3-1 for an illustration of the syllable structure of Mandarin Chinese), i.e., some syllable finals have a medial glide preceding the syllable rhyme. We have therefore decided to use the F_0 contour from the syllable final for tone modeling, especially since our speech recognition system uses syllable initials and finals as acoustic modeling units for Mandarin Chinese. Furthermore, the medial glide in null-initial syllables (e.g., "*wai*") is treated as a syllable initial in our Mandarin speech recognition system, so that the discrepancy between the syllable final and the syllable rhyme are limited to syllables with an initial and a medial (such as "*huai*"), in which case the presence of a syllable initial should have reduced the carry-over tone coarticulation effects.

3.3 Mandarin Chinese Corpora

Two Mandarin corpora of different complexity are used in our experiments: the Mandarin digit database, and the YINHE conversational speech database. We feel that continuous Mandarin digits form an excellent domain in which to carry out our initial study of Mandarin tone and intonation. First, digits cover all four lexical tones in Mandarin; thus, continuous digit strings provide an adequate domain in which to study tones and their contextual effects. Second, digit strings are usually spoken in phrase groups, especially long strings and phone numbers; thus, we have sufficient data to study the dependency of tone expression on the phrase structure. Third, those prosodic attributes that are not investigated in our study have very weak expression in digit strings. For example, each digit is a single syllable word, so that the variations due to lexical stress pattern differences should be minimal. We also observe that each digit string is likely to be spoken with a relatively plain intonation, so that the influence of focus should be small. The YINHE corpus is a linguistically rich database which contains spontaneous speech in addition to read utterances. We will use it to compare with the digit domain on tone and speech recognition performance, and to examine if the tone modeling approach we undertake is adequate for conversational speech.

DATA SET	TRAIN	TEST
# Utterances	3923	355
# Speakers	71	6

Table 3-2: Summary of the Mandarin digit corpus.

3.3.1 Mandarin Digit Corpus

The Mandarin digit corpus[4] was collected automatically by recording phone calls from native Chinese speakers, and the waveform was sampled at $8KHz$. A different list of 30 random phone numbers[5] (each containing 9 digits) and 30 random digit strings (each containing 5-10 digits) was given to each participant, and the subjects were instructed to read from the list in a naturally speaking way. The phone numbers are presented in the conventional phone number format, e.g., "(12) 345 - 6789", and the subjects are likely to follow the grouping when prompted to read each. The random digit strings are printed without any spacing, e.g., "12345678", so the phrase grouping for each string is up to the choice of each subject. Table 3-2 summarizes the number of utterances and speakers for the training and testing data sets.

3.3.2 YINHE Corpus

The YINHE database is associated with the YINHE system (Wang et al. 1997; Wang 1997), a Mandarin counterpart of the GALAXY conversational system (Goddeau et al. 1994). The user communicates with the computer in Mandarin Chinese, and the system is able to provide the user with information from three knowledge domains: the *city guide* domain answers questions about a large set of known establishments in the Boston area; the *flight* domain retrieves flight information worldwide from the Sabre reservations system; the *weather* domain provides world-wide weather information.

The YINHE corpus contains both read and spontaneous speech collected from native speakers of Mandarin Chinese. The spontaneous utterances were collected using a simu-

[4]The corpus is provided by SpeechWorks International, Inc. in Boston.

[5]The random phone numbers were generated to imitate the phone numbers in Taiwan, each of which consists of a 2-digit area code followed by a 7-digit number.

SET	TRAIN	TEST1	TEST2
Utterance Type	Spon. and read	Spon.	Read
# Utterances	4,953	194	209
# Speakers	93	6	6
Avg. # Syllables per Utterance	9.9	9.1	10.5

Table 3-3: Summary of the YINHE corpus.

lated environment based on the existing English GALAXY system. In addition, a significant amount of read speech data was collected through a Web data collection facility (Hurley et al. 1996). For each subject, 50 sentences within the YINHE domain were displayed in Chinese characters through the Web page. Both the spontaneous and read utterances were recorded through the telephone channel and digitized at an $8KHz$ sampling rate. The speech data were transcribed using tonal pinyin to simplify the input task. Because of the English based knowledge domains, a significant portion of the utterances contain English words such as city names and proper nouns, which are transcribed using English. Manual time-aligned phonetic transcriptions were not provided due to the tremendous effort required; instead, they were derived using a forced alignment[6] procedure during the training process.

Utterances containing English words or partial words are excluded from training or testing tone models. An exception is the English word "MIT", because it appears as the only English word in more than 600 utterances. Lexical tones are not defined for the word "MIT", so it is simply ignored by tone model training and testing. Speech data from 6 speakers are set aside to form a test set. Since we have both spontaneous and read utterances from these speakers, the data are further divided into a spontaneous test set and a read test set. The remaining utterances are used for training. A summary of the corpus is shown in Table 3-3.

[6]Forced alignment refers to the procedure of obtaining time-aligned phonetic transcriptions by running the recognizer in "forced" mode, in which the correct words are provided to the recognizer and the recognition system finds the corresponding sequence of phones and their time alignments given a lexicon and acoustic models.

3.4 Analysis of Mandarin Tonal Variations

As described in the beginning of this chapter, the phonetics of lexical tones can be influenced by many factors, which lead to large variances in statistical tone models. In this section, we conduct a number of empirical studies of our Mandarin data, to find regularities that can be utilized to reduce the variances of tone models and improve tone recognition performance. In the following, we examine the impact of F_0 downtrend, phrase boundary, tone coarticulation, and tone sandhi on the acoustic expression of Mandarin lexical tones. The F_0 contour for each utterance is first normalized by the average F_0 of the utterance as a preprocessing step, to account for speaker pitch range differences.

3.4.1 F_0 Downtrend

The effects of F_0 downtrend on tone are clearly demonstrated by the differences in the average tonal contour of each lexical tone at different syllable positions of the read phone numbers, as shown in Figure 3-3. For example, the top left plot in the figure shows the average tone 1 contour at each of the nine positions in the phone numbers. As shown in the plot, not only does tone 1 have an overall decreasing F_0 height throughout the utterance, but also the F_0 within each tone 1 contour falls slightly at most syllable positions. A closer examination of the data also reveals some details corresponding to the phrase structure of the phone numbers ("xx-xxx-xxxx"): the tone contour seems to rise slightly before a phrase boundary (or the utterance end); and there is a jump of F_0 level after each phrase boundary. However, the local reset of F_0 is relatively small compared to the declination, and the overall change of the F_0 level is predominantly decreasing. Similar trends can be observed for the other lexical tones as well, although the F_0 contour of each lexical tone is perturbed differently around phrase or sentence boundaries. It is clear that it is beneficial to compensate for the F_0 declination, because the F_0 level of tone 1 at the end of an utterance becomes similar to that of tone 3 at the utterance beginning, blurring their distinctions. We will focus on characterizing the *global* declination in this section; while the *local* modifications of tone contours around phrase boundaries will be discussed in the next section.

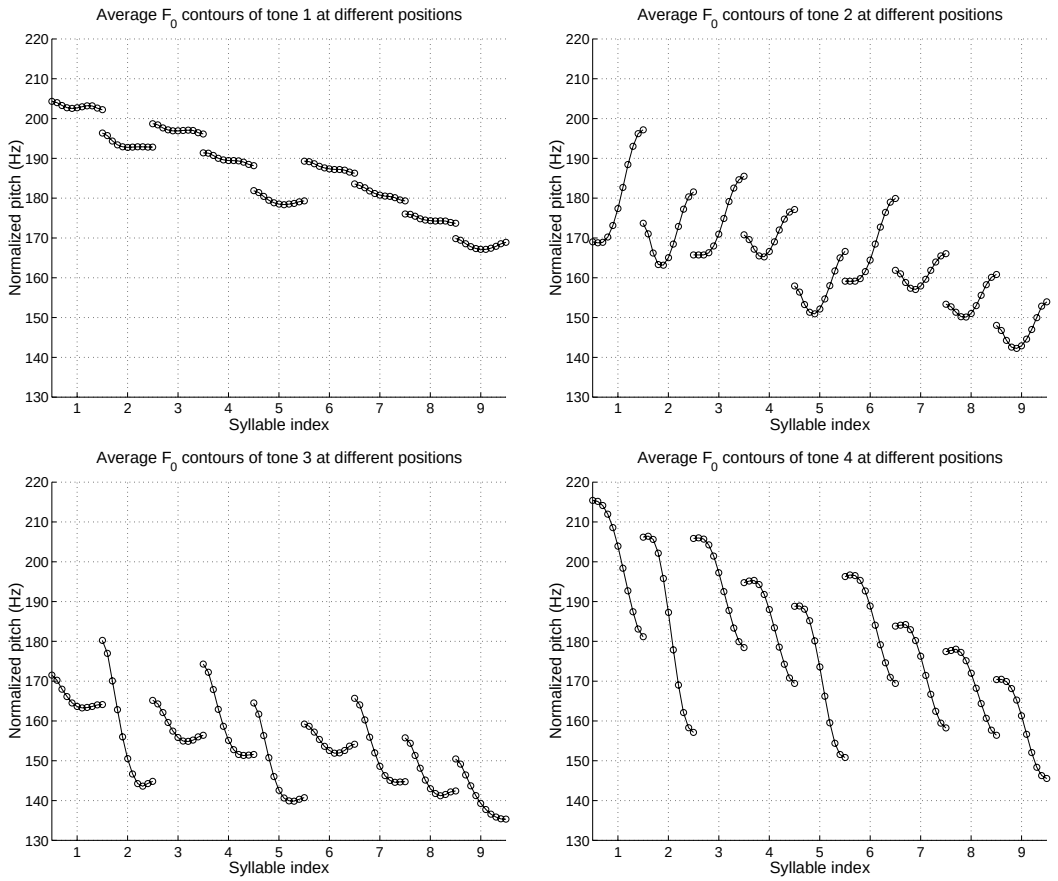

Figure 3-3: Average F_0 contours of four lexical tones at different syllable positions in the read phone numbers (xx-xxx-xxxx).

Because both tone and intonation are manifested as F_0 movements, it is difficult to separate the two aspects in the physical F_0 signal. Although Figure 3-3 offers a clear way to demonstrate the declination effects on tone, it is hard to quantify the declination factor due to the different tone contour shapes, and the approach can not be readily generalized to utterances with a different number of syllables. We want to design a method which removes the tonal contributions from the F_0 contour, thus, revealing the underlying "intonation" contour. Assuming that a set of utterances have similar F_0 declination in the intonation contour, we can then view the pitch contours of this set of data as a "constant" declination component with additive "random" perturbations caused by tones. Therefore, we can use an averaging approach to smooth out the "random" variations due to tones and obtain the

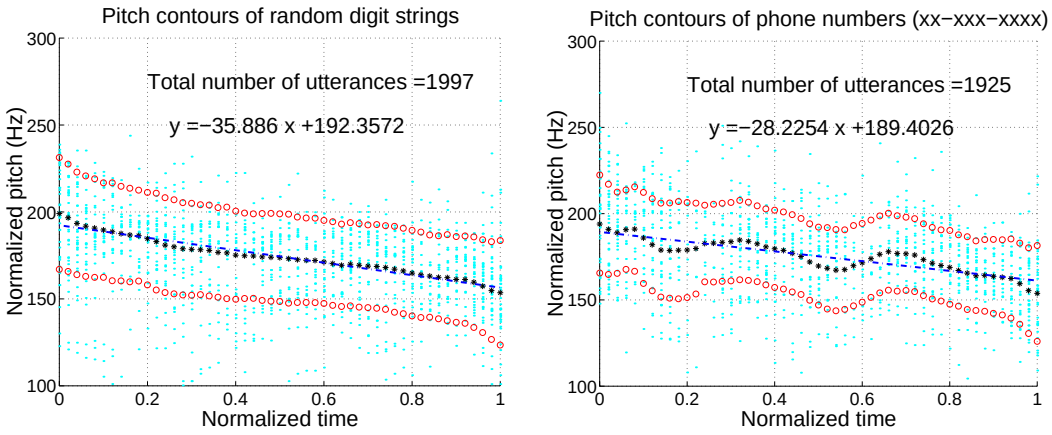

Figure 3-4: Pitch contours of random digit strings and phone numbers. The starred line represents the mean pitch contour, with the upper and lower circled lines indicating one standard deviation. The dashed line is the linear regression line for the average F_0 contour. The linear regression coefficients are also shown in the figures.

average as the underlying intonation contour.

We tested our method by plotting the F_0 contours of all digit data, grouped by random digit strings and phone numbers, in Figure 3-4. The time scale of each utterance is normalized by the utterance duration, so that utterances of different lengths can be aligned in time. It is obvious from the plot that there is a steady downdrift of the mean pitch contour, although the slope[7] for the downdrift trend is slightly different for random digit strings and phone numbers. The F_0 contour plot of phone numbers also reveals a more detailed phrase structure, which can be easily inferred from Figure 3-3. We believe that a random digit string also has similar behavior in its F_0 contour at phrase boundaries. The absence of such evidence from the plot is due to the "randomized" positions of the phrase boundaries in the time-normalized F_0 contour; thus, the "averaging" also smoothed out the phrase boundaries.

The overall F_0 declination of an utterance is likely to be dependent on the phrase structure. To examine carefully the relationship between the downdrift slope and the utterance duration or phrase structure, we grouped the random digit strings according to the number

[7]The slope here is actually the total declination in F_0, because the time scale of each utterance is already normalized by the utterance duration.

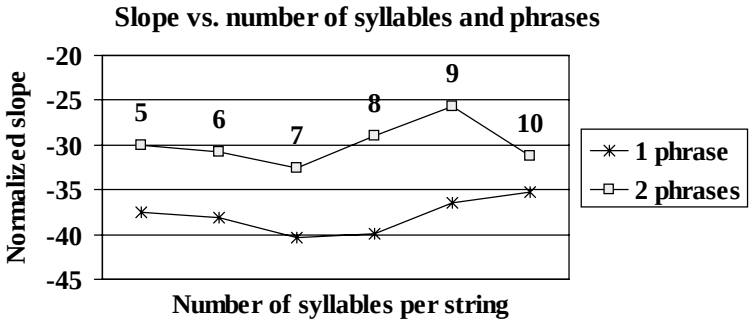

Figure 3-5: Downdrift slope of random digit strings grouped by the number of syllables and phrases.

of syllables and phrases, and obtained average F_0 slopes for each subset, as plotted in Figure 3-5. The phrase boundaries for random digits are detected automatically, by locating places where there is significant silence between two syllables. We realize that this is a very crude scheme, because those phrase boundaries not marked by a pause will not be labeled correctly. Automatic phrase boundary detection is another challenging task for Chinese, which will be discussed in more detail in Chapter 8. We were not able to obtain the slopes for utterances with 3 or more phrases, because of sparse data problems. From the plot we can see that the amount of total declination for two-phrase utterances is consistently smaller than that of their one-phrase counterparts, confirming that the F_0 base is raised after a pause. The slopes for digit strings with five to seven syllables seem to suggest that the amount of overall declination is larger for longer utterances (with the same number of phrases). The slopes for utterances with eight to ten digits do not form a clear trend. We think that this is caused by inaccuracies in the phrase boundary detection method, i.e., the long digit strings are likely to contain undetected phrase boundaries.

Figure 3-6 shows the average tone contour for each lexical tone at different syllable positions after a $uniform$ F_0 declination factor is removed from the F_0 contour of each utterance. As shown in the figure, the differences among position-dependent tone contours due to the F_0 declination are greatly reduced, with a noticeable over-compensation for tone 3. This seems to suggest that the declination factor is smaller for tone 3 than for other lexical tones. As expected, the F_0 declination within a phrase group is not corrected.

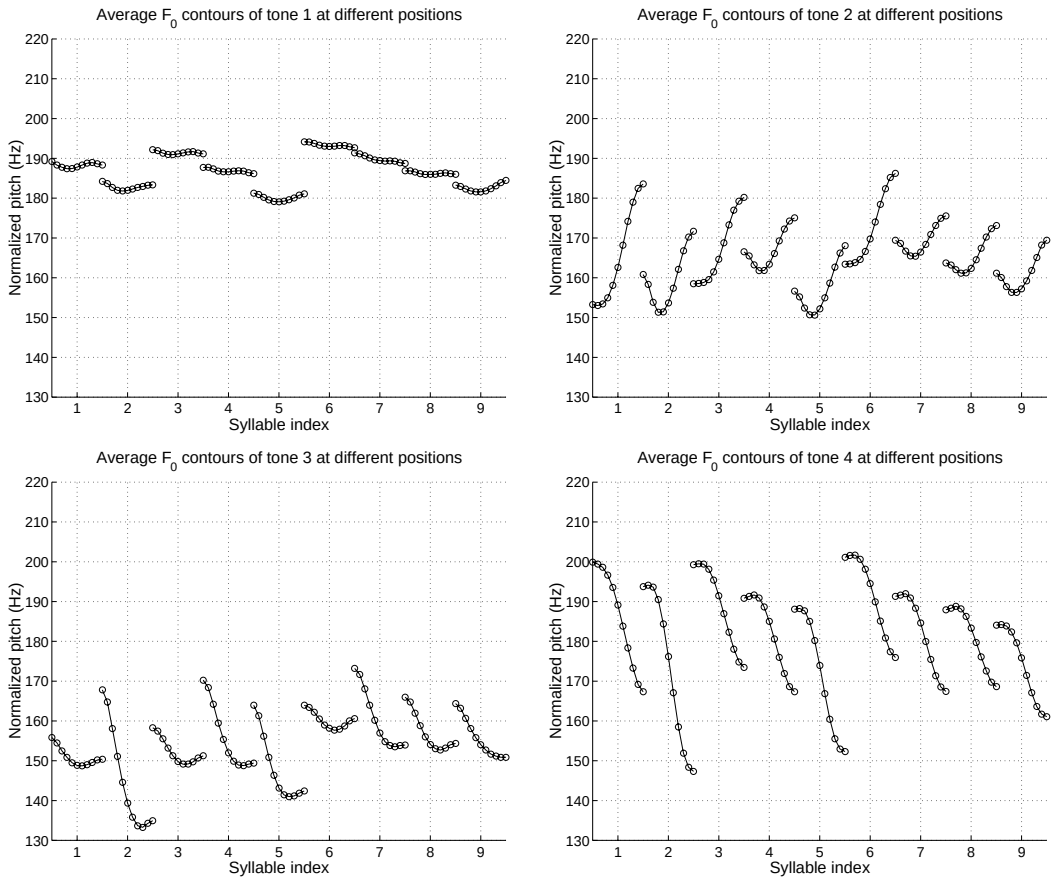

Figure 3-6: Average F_0 contours of four lexical tones at different positions in the read phone numbers (xx-xxx-xxxx) after a linear declination factor has been removed from the F_0 contour of each utterance.

However, the differences are relatively small compared to the inherent distinctions among tones. The behavior of tone contours at phrase initial and final positions will be addressed in the next section.

We are interested to know if F_0 downdrift is also affected by sentence mode. We examined that by comparing the mean F_0 contours for different types of utterances in the YINHE domain. The utterances were labelled manually using four categories, including declarative, command, wh-question, and particle-question, which is similar to the yes-no question in English. The average number of syllables for each set of utterances is shown in Table 3-4. As indicated in Figure 3-7, there are some differences in the F_0 slope, with wh-questions

Utterance Type	Average # Syllables
Declarative	9.5
Command	3.1
Wh-question	11.2
Particle-question	11.1

Table 3-4: Average number of syllables per utterance for each set of sentences in the YINHE domain.

having the sharpest drop and commands having the least. The small slope in the "command" utterances might be an artifact due to the biased tone content, i.e., a large portion of the data corresponds to "*fan*3 *hui*2" (go back), causing the F_0 contour to rise at the end; however, the relatively short duration of this type of utterance might also play a role. The data suggest that the particle-questions have smaller declination than wh-questions on average. This is consistent with the observation that a speaker tends to raise the F_0 towards the end of a yes-no question, which will result in a decrease in the overall F_0 declination. The declarative sentences also have smaller downdrift than the wh-questions. However, it is unclear if this is simply because the declarative sentences are shorter than the wh-questions on average.

3.4.2 Phrase Boundary

As indicated in Figure 3-3, the lexical tones behave differently around phrase boundaries. For example, tone 4 has much larger falling slopes at phrase-final positions, but not at sentence finals; tone 2 at phrase-initial positions has a rising shape instead of the fall-rise shape at other positions, and the F_0 excursion is also larger than in the other cases, etc. Figure 3-8 compares the average F_0 contours of four lexical tones at different phrase positions. The averages are computed using both random digits and phone numbers, and a uniform declination factor is removed from each utterance in addition to the normalization by the average F_0. The following observations can be made from the plots:

- The shape of tone 1 remains fairly flat at all phrase positions, except for a minor rising at phrase and utterance final positions. The F_0 level is the highest for phrase-

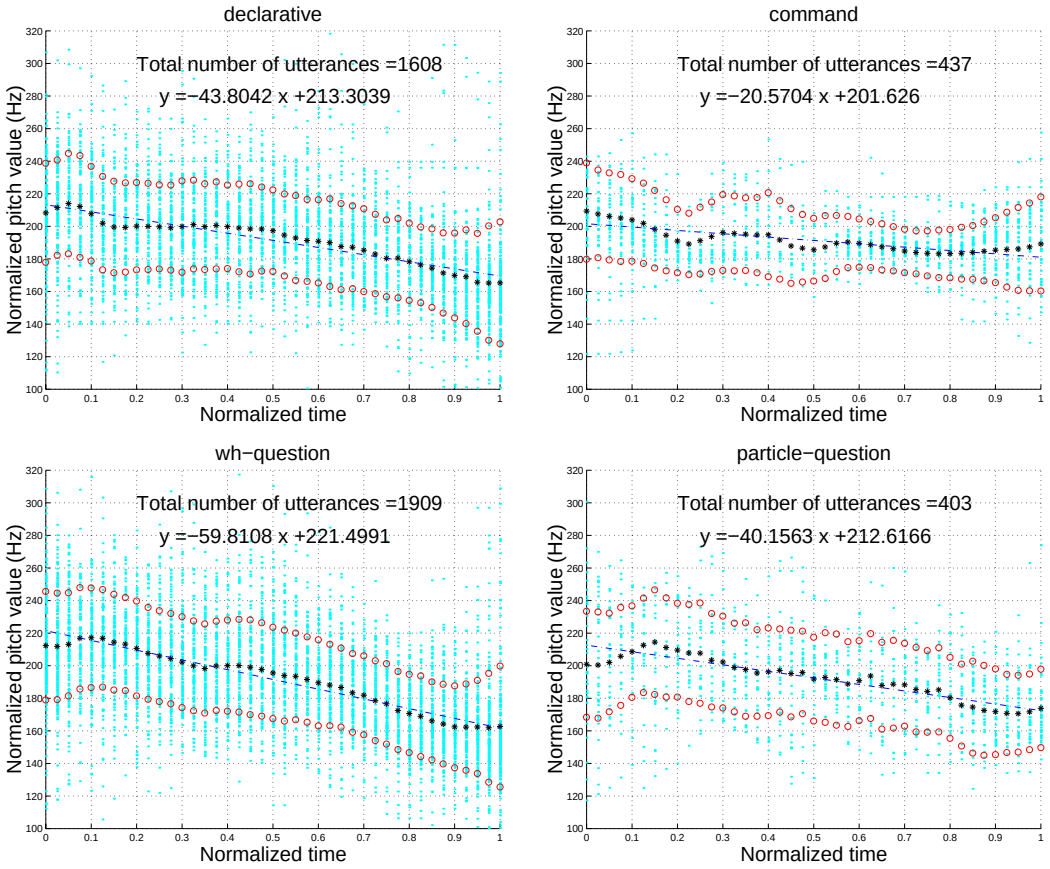

Figure 3-7: Pitch contours of YINHE utterances grouped by utterance type. The starred line represents the mean pitch contour, with the upper and lower circled lines for standard deviation. The linear regression coefficients are also shown in the figures.

initial syllables, lowest for phrase-final and sentence-final syllables, and intermediate for phrase-internal syllables. This is due to the sentence-based F_0 declination removal instead of a phrase-based removal.

- Tone 2 has an overall rising shape; however, there is a small falling portion preceding a larger rise at non-phrase-initial positions. The F_0 maximum is the largest for phrase-initial syllables, and the F_0 minimum is the smallest at internal phrase-final positions.

- Tone 3 has a low falling shape, with a low F_0 onset at phrase-initial positions, and higher F_0 onsets at other positions. The falling slope is much larger at the internal phrase-final positions.

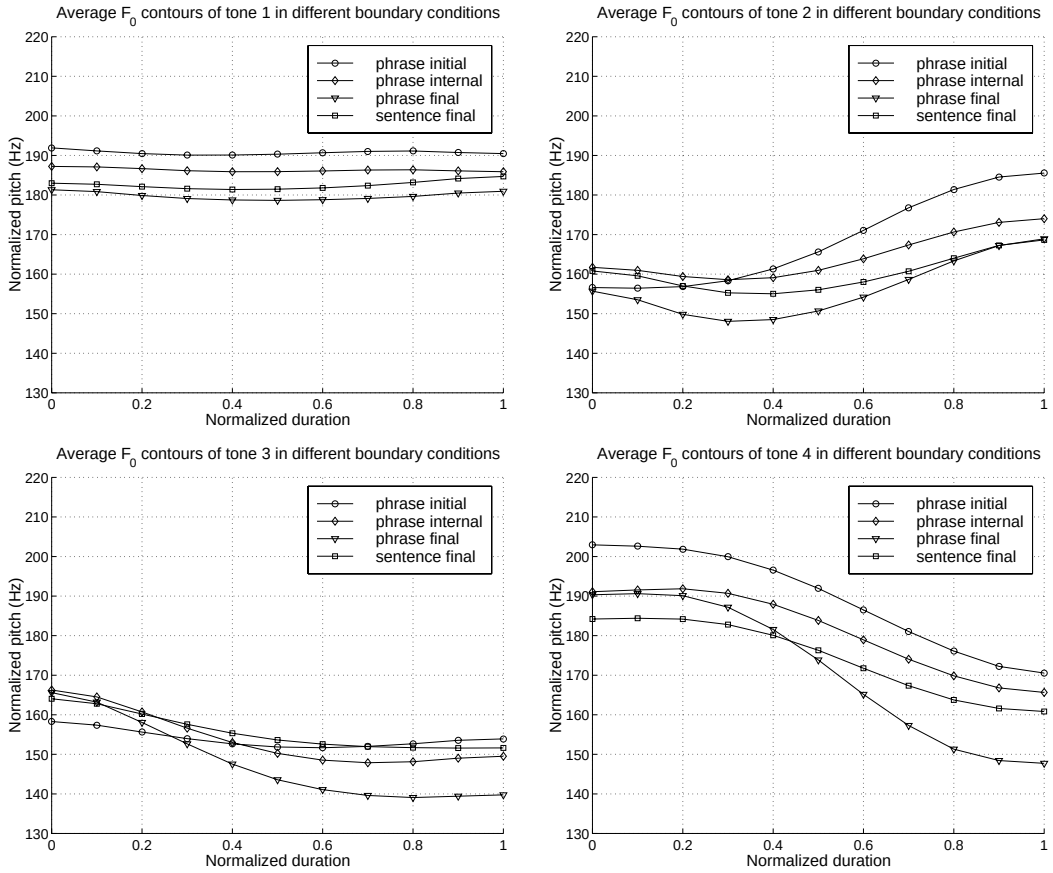

Figure 3-8: Average F_0 contours of four lexical tones at different phrase positions.

- Tone 4 has a high falling shape, with a much larger falling slope at internal phrase boundaries. The F_0 level differences are similar to those of tone 1.

It is surprising that there exist relatively large discrepancies between the tonal patterns of tone 2, tone 3 and tone 4 before internal phrase boundaries and those at the end of the utterances. This seems to suggest that the internal phrase boundaries differ somewhat from the end-of-utterance boundaries. One possible reason is that the sentence final does not need to be prosodically marked as strongly as the internal boundaries. Another possible reason is that F_0 at the end of an utterance is usually very low due to the overall declination, so it is not likely to go down much further. However, we also can not rule out the possibility that the results might be an artifact caused by pitch tracking errors at the utterance end

due to glottalization, which frequently resulted in a flat pitch contour with doubled pitch value. Regardless of the reasons, it seems more advantageous, from a modeling point of view, to treat the end of utterances and the internal phrase boundaries separately.

3.4.3 Tone Coarticulation

Tone coarticulation effects can be demonstrated by comparing the average F_0 contour of each lexical tone under different tone contexts. Since it is difficult to examine the tone contours in all 16 different combinations of left and right contexts at the same time, we performed the studies in two steps. First we study the influences of the left context, or the carry-over effects. Then we study the influences of the right context, or the anticipatory effects.

Carry-over effects

Figure 3-9 shows the average F_0 contour of each lexical tone in different left tone contexts for the Mandarin digit data. The influences of right contexts are washed out by the averaging. It is evident in the plots that carry-over effects are the largest at the onset of the tonal contours and taper off towards the end of the tone, so the F_0 contours of each tone in different left contexts approach similar target values (*on average*) at the end. Specifically, we can observe that all tones have a higher F_0 onset after tone 1 and tone 2 (high offset tones) than after tone 3 and tone 4 (low offset tones), and the F_0 onset is the lowest after tone 3. The seeming exception of tone 3 after tone 3 can be explained by the "$33 \rightarrow 23$" tone sandhi rule; the left context in this case is effectively tone 2, a high offset tone. The changes to the average tone contour shapes in different left contexts can be adequately explained by the interplay between the F_0 onset change and the inherent tone patterns. In particular, tone 2 falls after high offset tones to reach a relatively low F_0 value before rising, and tone 4 rises after tone 3 to reach a relatively high F_0 value before falling. In the other cases, the onset change does not significantly conflict with the canonical tone pattern; the difference in tone onset diminishes and the F_0 contour approaches the inherent tone pattern. From the data we conclude that the carry-over effects mainly change the F_0 onset of the following tone, and the change is assimilatory in nature. That is, the onset of the current tone is

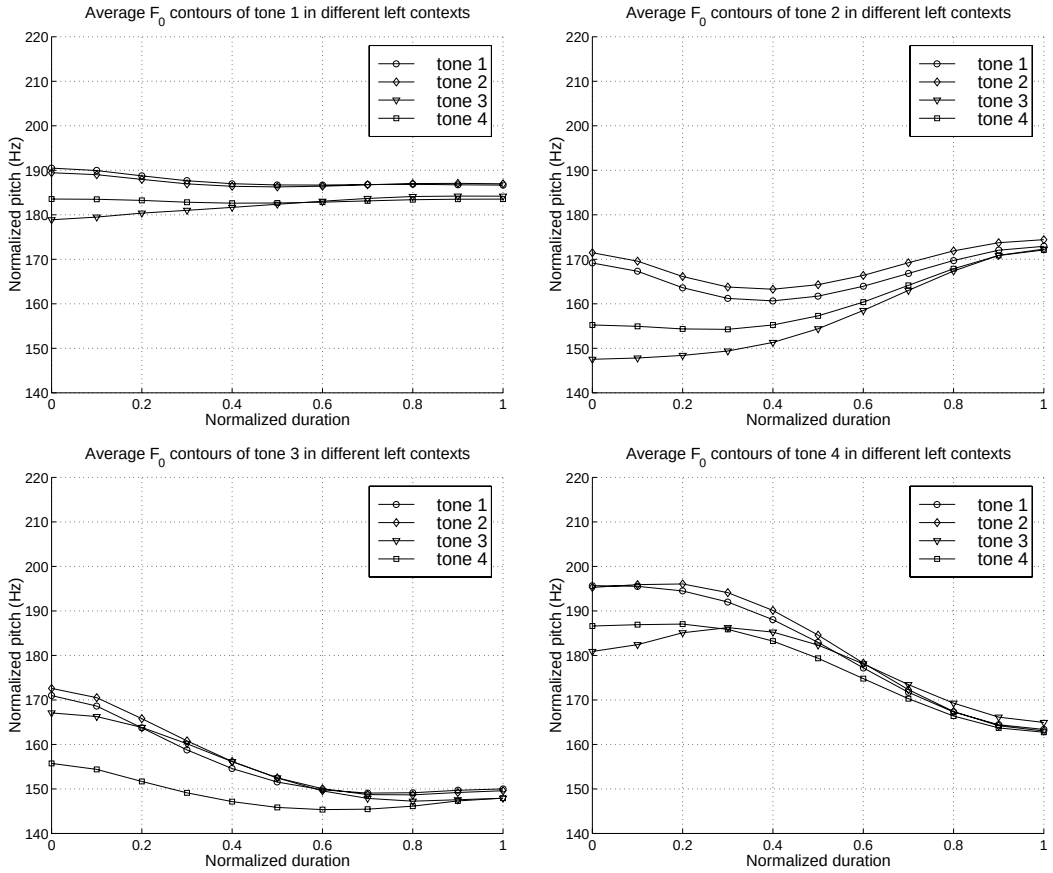

Figure 3-9: Average F_0 contours of four lexical tones in different left tone contexts.

assimilated by the offset of the previous tone. This is consistent with the findings in (Xu 1997).

Anticipatory effects

Figure 3-10 shows the average F_0 contour of each lexical tone in different right tone contexts for the Mandarin digit data. The influences of left contexts are washed out by the averaging. Unlike the carry-over effects, it seems that the anticipatory effects are not just limited to the tone offset. For example, the F_0 of tone 1 is on average higher before tone 3 than before tone 1 for the entire tone contour. In general, we find that the average contour *shape* for each lexical tone does not change much in different right contexts, with the exception of tone 3 before tone 3, which is due to the third tone sandhi rule. The anticipatory effects

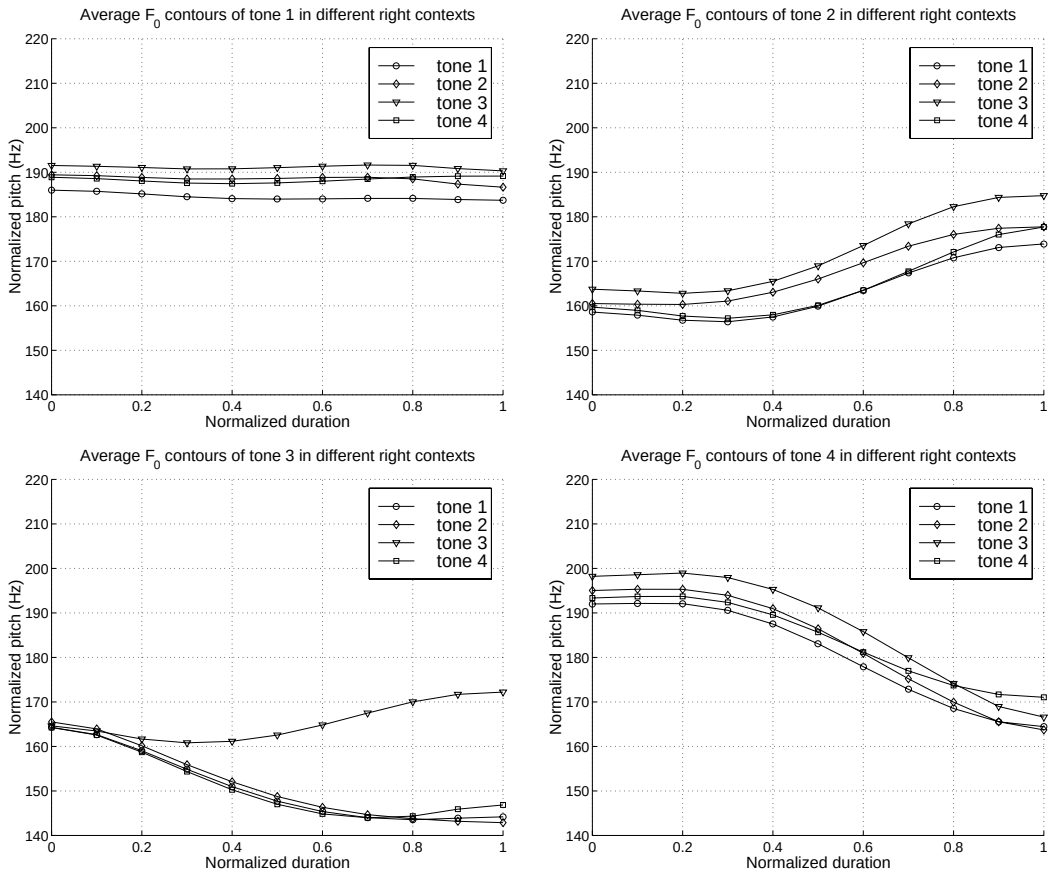

Figure 3-10: Average F_0 contours of four lexical tones in different right tone contexts.

seem to change the overall F_0 level. For example, similar to tone 1, tone 2 and tone 4 also have the highest F_0 level before tone 3 and the lowest F_0 level before tone 1. In addition, the slope of tone 4 seems to differ slightly depending on whether the following tone is tone 2 or tone 4. The falling slope of tone 4 is larger when followed by tone 2 and smaller when followed by tone 4, suggesting assimilation of the F_0 offset by the onset of the following tone. Tone 3 appears to be only slightly affected by the right context, except for the sandhi change. From the data, we are inclined to conclude that the anticipatory effects are mostly dissimilatory, and the entire tone contour is affected. With the exception of tone 3, all tones have higher F_0 level before tone 3, and lower F_0 before tone 1. However, there also seem to be some assimilatory effects, as indicated by the plot of tone 4. Overall, the anticipatory effects are smaller compared to the carry-over effects, as indicated by the small differences

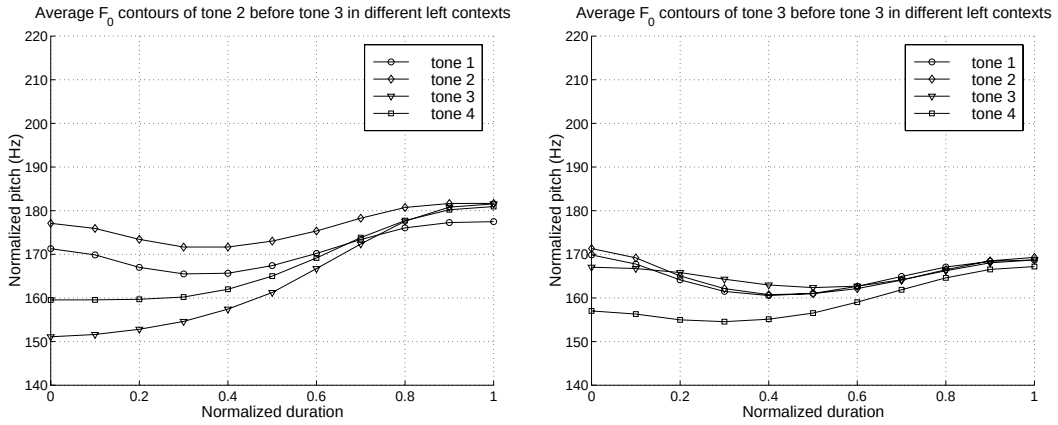

Figure 3-11: Average F_0 contours of tone 2 and tone 3 in different left tone contexts. The right tone context is fixed to be tone 3.

of average tone contours in various right contexts. These findings are generally consistent with those in (Xu 1997).

3.4.4 Tone Sandhi

As indicated in Figure 3-10, tone 3 before tone 3 has an average rising slope, similar to that of tone 2. We compare the sandhi-changed tone 3 with tone 2 more carefully, fixing the right context to be tone 3 while varying the left tone context, as shown in Figure 3-11. We can see that, aside from some F_0 level differences, tone 3 before tone 3 has very similar average contour shapes as tone 2 before tone 3, except when the left context is tone 3. This again can be explained by the third tone sandhi rule: most "333" tone sequences are changed to "223", so that tone 3 between two third tones should be similar to tone 2 between tone 2 and tone 3, which is the case as shown in the figure. In fact, both "333" and "233" are realized as "223" according to tone sandhi, which is supported by the figure. It seems that, although we did not try to infer the surface tone realization from tone sandhi rules, the context dependent models are able to capture that in the model statistics.

We are surprised by the consistent F_0 level differences between tone 2 and the sandhi-changed tone 3. As shown in the figure, tone 2 before tone 3 rises to a higher F_0 target than tone 3 before tone 3 in all left contexts. We suspect that this might be an artifact due

to differences in vowel intrinsic pitch or perturbation by the initial consonant; the YINHE data did not exhibit such a clear difference.

3.5 Summary

In this chapter, we presented empirical studies of Mandarin tone and intonation, focusing on analyzing sources of tonal variations. First we demonstrated the F_0 downtrend for Mandarin Chinese using both position-dependent tone statistics and the average F_0 contour of a set of aligned utterances. The data show that F_0 decreases consistently within a phrase; while there is a jump of F_0 level after each phrase boundary. However, the F_0 hike is relatively small compared to the declination, and the overall change of F_0 level is predominantly decreasing. We then characterized the effects of phrase boundary, tone coarticulation, and tone sandhi using a similar method, i.e., by comparing average tone contours in different immediate contexts. The most obvious effects of a phrase boundary seem to be on the tone excursion range. Tone 2, tone 3 and tone 4 at internal phrase-final positions reach a lower F_0 target than at other positions; tone 2 at phrase-initial positions also seems to rise to a higher F_0 target than at other positions. Tone coarticulation is manifested as both carry-over and anticipatory effects, with the carry-over effects appearing to be more significant. The carry-over effects mainly change the F_0 onset of the following tone, and the change is assimilatory in nature. The anticipatory effects are more complex, with both assimilatory and dissimilatory effects present in the data. The sandhi-changed tone 3 is similar to tone 2. It seems that a context dependent model using both left and right tone context should be able to capture the tone sandhi variation. In the next chapter, we try to account for these factors in tone modeling to reduce the variances of statistical tone models with the goal of improving tone and speech recognition performances.

Chapter 4

Mandarin Chinese Recognition Assisted with Tone Modeling

This chapter addresses the issues of applying tone modeling to improve Mandarin speech recognition performance (Wang and Seneff 1998; Wang and Seneff 2000a). This involves building statistical models to classify the lexical tones, and developing mechanisms in the recognizer to integrate these models into speech recognition.

We have explored two approaches to incorporating tone models into speech recognition. In the first approach, top N hypotheses are first obtained from a baseline recognizer using acoustic models only; tone models are then applied to resort the N-best outputs. The delay due to the resorting is fairly small, because there is minimal computation in the post-processing stage. The advantage of this approach is that context-dependent tone models can be applied based on the information available in the N-best list. Nevertheless, the effectiveness of this approach is dependent on the quality of the N-best list; i.e., the correct hypothesis can not be recovered if it is not in the N-best list. In the second approach, both the tone scores and the other acoustic scores are utilized in the first-pass Viterbi search. The motivation for this approach is that the tone models are utilized to explore the entire search space, so that the correct hypothesis has a better chance to benefit from the tone models. However, it is generally difficult and computationally expensive to incorporate more refined tone models into the first-pass search. A third possibility is to apply simple tone models in

the first-pass recognition process, to obtain a higher-quality N-best list than that obtained without tone models. More refined models can then be applied to the N-best list to achieve further improvements. However, we have found that the N-best list obtained with tone models is similar in quality to that of the baseline system, so the combined approach would have no advantage over the pure post-processing approach.

In the following sections, we first look at some related work on tone modeling and the incorporation of tone models into Mandarin Chinese speech recognition. Next we give some background information for the experiments presented in this chapter. This includes descriptions of the SUMMIT speech recognition system, and the baseline recognizers for the Mandarin digit domain and the YINHE domain configured from SUMMIT. We then describe the basic tone modeling framework, and compare the tone classification performance of various refined tone models. After that, we describe the implementation of two mechanisms in the SUMMIT system for incorporating tone models into speech recognition. We will present a suite of speech recognition experiments, comparing the contributions of using various tone models and different tone model integration methods to speech recognition performance. We found that the addition of tone models significantly improved speech recognition performance for both the digit domain and the YINHE domain. However, using more refined tone models only yielded small additional gains in speech recognition, even though the tone classification accuracies were improved significantly with the more refined models.

4.1 Related Research

Substantial work has been done on Mandarin Chinese tone recognition. Tone recognition is generally not considered as a difficult task, because there are only five tones (four lexical tones plus a neutral tone) in the language. Nevertheless, the performance of tone recognition systems is highly dependent on the test conditions. Very high tone recognition accuracy has been achieved for isolated syllables in both speaker-dependent and speaker-independent modes (Yang et al. 1988; Liu et al. 1989). An interesting study was conducted in (Liu et al. 1989) which demonstrated the effects of tone coarticulation on tone recognition accuracy.

The best average recognition rate for isolated mono-syllabic words was 97.9%. However, the performance on poly-syllabic words degraded, even with context-dependent tone model pairs. The best average recognition rate was 93.8% for the first syllable of di-syllabic words, 92.0% for the second syllable of di-syllabic words, and only 85.5% for the middle syllable of tri-syllabic words. In general, it is difficult to achieve high tone recognition accuracy for *continuous* Mandarin speech due to intonation and coarticulation interferences (Wang and Lee 1994; Wang et al. 1994; Wang and Cheng 1994; Chen and Wang 1995; Cao et al. 2000), and experiments on *spontaneous telephone* speech have rarely been reported in the literature.

Most tone recognition systems adopt a hidden Markov model (HMM) framework (Wang and Lee 1994; Wang et al. 1994; Cao et al. 2000; Huang and Seide 2000) or a neural network (NN) framework (Chang et al. 1990; Wang and Cheng 1994; Chen and Wang 1995). There are also a few systems that are based on other statistical or non-statistical classification methods (Wang et al. 1990; Wu et al. 1991). Tone features are generally obtained from F_0 and energy measurements. In the HMM approach, the tone feature vector is constructed at a fixed frame rate and usually consists of F_0 and short-time energy plus their derivatives. In the segment-based approaches, the tone feature vector is extracted from the entire tone segment. Parameters for describing the F_0 contour shape can be obtained by fitting the contour with a certain type of function (Wu et al. 1991), by projecting it onto some basis functions (Chen and Wang 1990), or by piece-wise linear fitting (Chang et al. 1990).

A few tone recognition systems tried to utilize intonation or context information to improve tone recognition performance. In (Wang et al. 1990), phrase components for both the F_0 declination and the pitch accent effects were estimated for Chinese four-syllable idioms based on Fujisaki's model (Fujisaki 1988). In (Wang et al. 1994), a two level HMM structure was used to compensate for the sentence F_0 declination effect, with a sentence HMM on the upper layer and state-dependent HMMs for tones on the lower layer. In (Wang and Cheng 1994), a prosodic model was developed based on a simple recurrent neural network (SRNN). The SRNN can learn to represent the prosodic state at each syllable of an utterance using its hidden nodes. The outputs of the hidden nodes then serve as additional features to a multi-layer perceptron (MLP) based tone recognizer. The recognition rate

was improved from 91.38% to 93.10% with the prosodic model under speaker-dependent mode. In (Cao et al. 2000), a decision tree clustering method was used to derive context-dependent tone model classes with considerations for the tone context, the syllable position in the word, and the consonant/vowel type of the syllable. The performance was improved from 60.9% to 70.1% for five tone classification on a dictation database.

Early work on Mandarin speech recognition has been restricted to isolated-syllable mode, in which the utterances are either single syllables or complete sentences spoken with a pause between syllables (Gao et al. 1991; Lee et al. 1993; Lin et al. 1996). In more recent years, Chinese speech recognition systems for isolated words (Hon et al. 1994; Gao et al. 1995) or *continuous* phrases and sentences (Lyu et al. 1995; Wang et al. 1995; Ho et al. 1995; Hsieh et al. 1996; Huang and Seide 2000; Shu et al. 2000) have also emerged. Aligning tone and syllable scores for isolated syllables is not an issue, because the mapping is one to one. However, it is generally necessary to synchronize acoustic and tone scores in multi-syllable word and continuous sentence recognition, except in systems which ignore tone information while relying only on lexical processing and language models to resolve homophone ambiguities (Hsieh et al. 1996; Wang et al. 1997; Shu et al. 2000). A straightforward method is to merge acoustic and tone features to form combined tone-acoustic models. This approach is adopted by (Huang and Seide 2000), in which the use of tone information greatly reduced the Chinese character error rates on a number of continuous Mandarin speech corpora, including a telephone speech database. However, this approach requires a large amount of training data, because syllable finals with different tones can not be shared in training, and vice versa for tones. In (Lyu et al. 1995), tone classifications are performed only on mono-syllabic and bi-syllabic words, while words containing more than two syllables are determined by base syllables only. In (Wang et al. 1995), HMMs for Mandarin sub-syllabic acoustic models and context-dependent tones are applied separately, and a *concatenated syllable matching* (CSM) algorithm is developed to match the base syllable and tone hypotheses. In (Ho et al. 1995), acoustic models are used in the forward search and right-context-dependent tone models are applied during the backward stack decoding. In (Cao et al. 2000), a special search algorithm is developed to integrate acoustic and tone scores. In this way, acoustic and tone features are used jointly to find an

optimal solution in the entire search space, while the models are trained separately to take advantage of data sharing. We will implement a similar approach in our segment-based speech recognition system, and compare that with a post-processing approach where tone models are only applied on recognizer N-best outputs.

Our domain-dependent Mandarin recognizers are built on the SUMMIT segment-based speech recognition system (Glass et al. 1996), developed at the Spoken Language Systems group of the MIT Laboratory for Computer Science. To be compatible with SUMMIT, the tone modeling framework is based on a segment-based approach as well. In addition, we adopt a statistical classification method using Gaussian probability density functions (PDFs), so that the tone scores are *probabilistic* and can be directly used in speech recognition. It has been argued in (Fu et al. 1996) that tone modeling would not help continuous speech recognition for Mandarin Chinese, because tone information is redundant with the language processing model. However, we think that tone information can be used to reduce speech recognition errors besides homophone disambiguation, as confirmed by a few other studies as well (Cao et al. 2000; Huang and Seide 2000). In this chapter, we conduct tone classification and speech recognition experiments on telephone speech of various degrees of linguistic complexity. We try to improve tone recognition performance for *continuous* speech by accounting for intonation and tone context influences, and to improve speech recognition for *spontaneous telephone* speech by including tone modeling.

4.2 Experimental Background

The digit and YINHE Mandarin speech corpora used for our tone and speech recognition experiments have been described in detail in Section 3.3. In this section, we give a brief introduction to the SUMMIT speech recognition system, as well as the baseline recognizers for the digit domain and the YINHE domain configured from SUMMIT.

4.2.1 SUMMIT Speech Recognition System

The SUMMIT system uses a *probabilistic segment-based* approach to speech recognition (Glass et al. 1996). It differs from the HMM-based framework in that acoustic *landmarks*[1] are first detected, and the acoustic features are extracted relative to the landmarks instead of at fixed frames. The features in a segment-based system usually span a much longer interval than a single frame, so that acoustic-phonetic correlations can be captured by the acoustic models. The SUMMIT recognizer can be configured to use *either* or *both* of segment and boundary models to provide acoustic constraints. The segment models focus on acoustic properties within segment units (usually phones[2]), while the boundary models capture acoustic information around landmarks (between two phones or within a phone).

The SUMMIT recognizer works as follows. Given a speech waveform, spectral features such as Mel-frequency Cepstral coefficients (MFCCs) are first computed at a constant frame rate. A segmentation graph is then derived using an *acoustic segmentation* algorithm (Glass 1988). The algorithm detects *landmarks* (boundaries) where spectral change between adjacent frames is large, and interconnects these boundaries to form a network of segments. Acoustic features are extracted within each segment as well as around boundaries. The current feature vector for segments typically has 40 measurements, consisting of three sets of MFCC averages computed over 3:4:3 portions of a segment, two sets of MFCC derivatives computed at the segment beginning and end, and the log duration of the segment. The boundary feature vector has 112 dimensions and is made up of MFCC averages computed over 8 time windows around the boundaries. Principal component analyses are performed on the acoustic feature vectors to reduce the feature dimensions as well as to "whiten" the observation space. The distributions of the feature vectors are modeled using mixtures of diagonal Gaussians. During training, the Gaussian mixture parameters are estimated from pooled training tokens with K-means clustering followed by iterative EM optimization. During recognition, the segment and boundary model scores for all segments and bound-

[1] Landmarks, as defined in the SUMMIT system, refer to places in the speech signal where there is significant spectral change.

[2] We define phones as acoustic units which are obtained by applying a set of phonological rules to the original phonemic form provided for each word in a lexicon. The notion of phone is used interchangeably with the set of basic context-independent acoustic model units in this thesis.

aries in the segmentation graph are evaluated in a computation-on-demand manner. A forward Viterbi module utilizes the acoustic scores and bigram language model probabilities to search a pre-compiled pronunciation network (lexicon). The search prunes unlikely paths which have poor total probability scores, and finds the most plausible word and phone sequences for the utterance. A backwards A^* search can then be applied on the pruned search space to generate the top N hypotheses. It can also apply higher-order language models to improve the speech recognition performance. More recently, the SUMMIT system has adopted a weighted finite state transducer (FST) representation of the search space (Glass et al. 1999). This new architecture allows a more flexible way to combine various types of constraints through FST manipulations, and the Viterbi search is only concerned with a single transducer encoding the entire search space. Under a typical recognizer configuration, the FST used by the search is composed from constituent FSTs specifying diphone to phone mappings, phonological rules, a lexicon, and a language model. A higher-order language model, also encoded as an FST, can be used in the backward A^* search to improve speech recognition.

Although SUMMIT was originally designed for recognizing English utterances, it has been applied successfully for various multilingual speech recognition tasks including Italian (Flammia et al. 1994), Japanese (Seneff et al. 2000), and Mandarin Chinese (Wang et al. 1997; Wang et al. 2000). The recognizer for Mandarin Chinese is very similar to that for English, except with Chinese-specific acoustic models, phonological rules, lexicon and language models. The lexicon and language models are also dependent on the application domain. The baseline recognizer for Mandarin Chinese does not include tone constraints, except perhaps implicitly in the lexicon and the language model due to the tonal pinyin representation of words.

Several enhancements need to be added to SUMMIT in order to incorporate *probability scores* of lexical tones into Mandarin speech recognition. Above all, a pitch tracking capability needs to be implemented, which has been described in Chapter 2. Tone classifiers can be trained in a similar manner as segment classifiers; however, new feature extraction methods need to be implemented to obtain parameters for tone contour shapes. This process is described in detail in Section 4.3. We also experiment with two mechanisms in

SUMMIT to incorporate tone scores into recognition, including a first-pass approach and a post-processing approach. The first-pass approach is compatible with both the FST and non-FST versions of SUMMIT, while the post-processing approach is currently implemented only for the non-FST architecture[3]. The implementations are described in detail in Section 4.4.

4.2.2 Mandarin Digit Recognizer

The baseline digit recognizer has 11 single-syllable words in its lexicon, including the alternative pronunciation "*yao*1" for digit "one". Chinese syllable initials and finals are chosen as the phone model units, with the addition of closure, inter-word pause, glottal stop, and nasal ending models, introduced by phonological rules. To fully exploit the small vocabulary size, digit-specific segment and boundary models are used. A bigram language model is trained from the training data. The language model constraints are fairly weak because the digit strings are randomly generated. We achieved 4.3% syllable/digit error rate (28.7% utterance error rate) on the test data with an A^* search.

We observed high error rates for single-vowel digits "*yi*1" (one), "*er*4" (two) and "*wu*3" (five). Segmentation of "vowel vowel" sequence such as "*wu*3 *wu*3" is difficult, especially due to frequent absence of glottal stops in continuous speech. The counter problem of vowel splitting also exists. This leads to high insertion and deletion error rates for these three digits. Digits "*yi*1" and "*wu*3" also tend to be obscured in coarticulation with other digits, such as in "*qi*1 *yi*1" (seven one), "*liu*4 *wu*3" (six five) and "*jiu*3 *wu*3" (nine five), which leads to even higher error rates for "*yi*1" and "*wu*3". There are also several confusing digit pairs, such as "*er*4/*ba*1" (two/eight), "*liu*4/*jiu*3" (six/nine), etc., partially due to the poor quality of telephone speech. Conceivably, these errors can be reduced by including tone models in recognition.

[3]This is due to the fact that the FST based system loses explicit references to time in the output phone graph.

4.2.3 YINHE Recognizer

We modified the YINHE recognizer described in (Wang 1997; Wang et al. 1997) to form a baseline system to carry out the tone modeling experiments. The original vocabulary for the YINHE domain contains about 1,000 words, with around 780 Chinese words and 220 English words. To better test the effects of tone models on Chinese speech recognition, we eliminated the English words from the vocabulary and excluded utterances that contain English words from training and testing. The data sets of the "cleaned-up" version of the YINHE corpus were introduced in detail in Section 3.3.

Only segment models are used by the recognizer[4]. They are inherited from the original YINHE recognizer, with unnecessary English-specific phone models omitted. Similar to those in the digit recognizer, the segment units are based on Chinese syllable initials and finals, supplemented with closure, silence, glottal stop, and nasal ending models. The model inventory covers all Mandarin syllable initials and finals except for the final "*ueng*", which occurs only in a few characters of the entire language. Syllable finals with only tone differences are pooled together to train one model, so that the acoustic models have no ability to distinguish tones. However, since words are represented using tonal pinyin, the lexicon and the language model provide some implicit tone constraints. The baseline recognizer achieved a syllable error rate of 8.2% on the spontaneous test data with a class bigram model.

The YINHE data complement the digit data in several ways. First, the YINHE corpus is linguistically richer than the digit domain and contains a significant amount of spontaneous speech, and thus, the phonetic variabilities for both segment and tone models are larger than those in the digit domain. Second, there are more lexical and language model constraints in the YINHE domain, while the digit utterances are simply "random" syllable strings.

Although it seems less advantageous to use tone models in the YINHE domain than in the digit domain, we think that tone information also has potential to reduce speech recognition errors for the more complex YINHE utterances. For example, substitution errors can produce incorrect tone transcriptions, and insertion and deletion errors usually result

[4]We felt that there were insufficient training data available for boundary models.

in abnormal tonal segments. These errors are likely to result in bad tone model scores. We will conduct experiments to see if tone modeling can improve speech recognition for the *spontaneous telephone* speech of the YINHE domain.

4.3 Tone Classification

In this section, we present the basic four-tone classification framework and explore the use of more refined tone models to improve tone classification performance on the digit and YINHE data. We first introduce the simple four-tone models, focusing on describing the features for tones. After that, we describe various refinements to the simple models, to reduce model variances with the goal of improving classification accuracy. The refinements are motivated by our empirical analysis of tonal variations presented in Section 3.4. Finally we summarize the tone classification performance of different tone models on the digit and YINHE test data. The statistical significance of the performance differences will be examined, and a detailed analysis of the tone confusion errors will be performed.

4.3.1 Simple Four Tone Models

The dominant component in tone expression is the F_0 contour pattern, i.e., its average, slope, and curvatures. There are various ways to quantify these features in a segment-based system, by either fitting the F_0 contour with a certain type of function, or projecting it onto some basis functions. We have chosen the first four coefficients of the discrete Legendre transformation to capture the tone contour pattern following the example in (Chen and Wang 1990), in which the Legendre coefficients were used to encode the pitch contour of Mandarin utterances. It has been found that the distortion of the reconstructed F_0 contour is fairly small if the decomposition is performed in a tone-by-tone manner, possibly because of the resemblance between the Legendre bases and the basic tone contour patterns.

The discrete Legendre bases can not be obtained by directly sampling the continuous Legendre polynomials, because the orthonormal properties of the bases are not preserved under the dot product definition in the discrete space. Similar to the dot product used to obtain Legendre polynomials in the continuous space, the dot product to derive the discrete

Legendre bases in the vector space $\mathcal{R}^{M+1}$ is defined as:

$$(\vec{u}, \vec{v}) = \frac{1}{M+1} \cdot \sum_{i=0}^{M} u_i \cdot v_i \tag{4.1}$$

where $\vec{u} = [u_0\ u_1\ u_2\ ...\ u_M]$ and $\vec{v} = [v_0\ v_1\ v_2\ ...\ v_M]$. The discrete Legendre basis vectors can be obtained as follows (Chen and Wang 1990):

$$\begin{aligned}
L_0(x_i) &= 1 && (4.2)\\
L_1(x_i) &= \sqrt{\frac{12M}{M+2}} \cdot (x_i - 0.5) && (4.3)\\
L_2(x_i) &= \sqrt{\frac{180 \cdot M^3}{(M-1)(M+2)(M+3)}} \cdot (x_i^2 - x_i + \frac{M-1}{6M}) && (4.4)\\
L_3(x_i) &= \sqrt{\frac{28000 \cdot M^5}{(M-1)(M-2)(M+2)(M+3)(M+4)}} \\
&\quad \cdot (x_i^3 - 1.5x_i^2 + \frac{6M^2 - 3M + 2}{10M^2} x_i - \frac{(M-1)(M-2)}{20M^2}) && (4.5)
\end{aligned}$$

where $x_i = \frac{i}{M}$ $(i = 0, 1, ..., M)$, and $M \geq 3$. We denote the basis vectors $[L_j(x_0)\ L_j(x_1)\ L_j(x_2)\ ...\ L_j(x_M)]$ as $\vec{l_j}$ $(j = 0, 1, 2, 3)$. Figure 4-1 displays these polynomial functions for the $M = 29$ case in the $[0, 1]$ interval. The discrete Legendre bases for the $\mathcal{R}^{30}$ space are obtained by sampling these functions at multiples of $\frac{1}{29}$ in the $[0, 1]$ interval, as illustrated in the figure.

For a discrete pitch contour segment $\vec{f} = [f_0\ f_1\ f_2\ ...\ f_M]$, the Legendre coefficients can be obtained as the dot product between the pitch contour $\vec{f}$ and the Legendre basis vectors:

$$a_j = (\vec{f}, \vec{l_j}) = \frac{1}{M+1} \cdot \sum_{i=0}^{M} f_i \cdot L_j(x_i) \tag{4.6}$$

A reconstruction of the pitch contour can be obtained from the coefficients as follows:

$$\vec{\hat{f}} = \sum_{j=0}^{3} a_j \cdot \vec{l_j} \tag{4.7}$$

The reconstructed contour $\vec{\hat{f}}$ is usually a smoothed version of the original contour $\vec{f}$, and

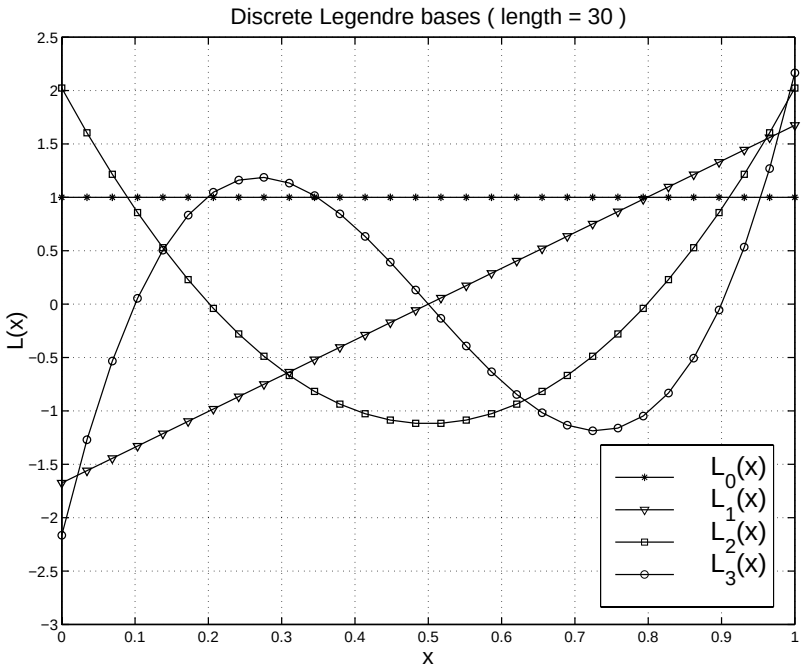

Figure 4-1: Discrete Legendre bases for the $\mathcal{R}^{30}$ vector space. The underlying polynomials for $M = 29$ are displayed with the discrete samples.

the distortion between the two vectors is found to be small for each of the four lexical tones (Chen and Wang 1990).

It can easily be shown that the first Legendre coefficient a_0 is the average value of the pitch segment, and the second Legendre coefficient a_1 happens to be the slope of the least square error regression line of the pitch segment. In addition to the four Legendre coefficients, we also included the average probability of voicing in the tone segment and the duration of the tone segment in the feature vector. We determined empirically that the average probability of voicing feature improved tone classification performance. The duration measure does not contribute significantly to tone discrimination, but it is helpful to limit insertion and deletion errors when the tone models are applied in speech recognition.

In a speaker-independent system, it is necessary to normalize the absolute F_0 with respect to the average over the entire utterance, to reduce across-speaker differences. We determined empirically whether to adjust by a ratio or a sum. Our data indicate that the ratio gives smaller variances for the pitch-related features we have chosen; thus, the

Legendre coefficients are *scaled* according to the average F_0 of the utterance.

A principle component analysis is applied to the six-dimensional tone feature vector, and mixtures of diagonal Gaussians are used to model the distributions of each of the four lexical tones. Except for the average F_0 normalization for each utterance, all other sources of tonal variations, such as the F_0 declination effects and the tone sandhi changes, are handled implicitly by the variances of the Gaussian probability density functions.

4.3.2 F_0 Downdrift Normalization

As shown in Section 3.4, the F_0 level at the end of an utterance is typically much lower than that at the utterance beginning, due to F_0 declination. It is clearly advantageous to compensate for the F_0 declination, so that the F_0 level distinctions between different lexical tones will be less smeared by their relative positions in an utterance.

We started by modeling the F_0 downdrift as a straight line for the digit and the YINHE utterances. The parameters were estimated by linear regression analysis of the mean F_0 contours of the training data for each domain. We then subtracted this downdrift from the F_0 contour of each utterance and re-trained tone models. This significantly reduced the tone classification errors for both the digit and the YINHE domains. A closer examination of the model parameters revealed that the variances of the F_0 related features for each lexical tone model were also greatly reduced.

We tried a number of ways to improve the declination model to achieve further performance improvement. One attempt is to use more refined phrase models instead of a sentence level model, because the declination is obviously dependent on the phrase structure of an utterance, as shown in the previous chapter. We tested this approach on phone numbers, which usually contain three phrases in each utterance. A regression line for each phrase is estimated from the mean F_0 contour of all training data. However, this refinement did not yield significant improvement over the simple sentence model. This is possibly because the declination within a phrase is relatively small compared to the distinctions between tones, so that the classification performance is not significantly affected. We also tried various regression analyses for each *individual* utterance's F_0 contour to approximate the intonation component. However, the tone classification performance degraded. We observed that the

resulting intonation curve using this approach follows the F_0 contour too closely, and hence, consumes part of the contribution from tones. This approach is also not robust to errors in pitch extraction and segmental alignment, which are inevitable in automatic systems.

4.3.3 Context Normalization

As analyzed in Section 3.4, the F_0 contour of each lexical tone is also systematically influenced by the phrase boundaries and tone contexts. For example, tone 2 after high offset tones has a relatively flat contour, making it less distinguishable from tone 1; tone 3 before tone 3 is changed completely to tone 2 due to tone sandhi; etc. It is conceivable that the tone classification performance can be improved by taking into account the contextual effects.

One can use context-dependent tone models to capture the differences of tone feature distributions under different contexts. However, the classification results from using context-dependent tone models can not be directly compared with the context-independent four tone case, because of the complications arising from the increased number of model classes and the splitting of the training data. We designed a classification experiment which uses the same number of tone models trained from the same amount of data as in the simple four lexical tone case, but with context effects "removed" from the tone measurements. This is realized by changing the F_0 contour of each tone according to its contexts to compensate for context effects. Specifically, we characterize the context effects as the differences between the mean of the Legendre coefficients of each context-dependent tone model and those of the corresponding context-independent tone model. We then alter the F_0 contour of each tone according to its contexts, by reconstructing an F_0 *difference contour* from the Legendre coefficient differences according to Equation 4.7, and combining that with the original tone F_0 contour. New context-independent models are then trained from those corrected F_0 contours. We performed correction for the first and second Legendre coefficients, i.e., the F_0 average and slope. We found that the variances of the retrained models were significantly reduced on those two dimensions, and the classification errors were further reduced.

During recognition, the tone context can be obtained from the recognizer N-best list. However, it is cumbersome to modify the tone feature vector according to the context at

run-time. Notice that the modification to the feature vector can be equivalently achieved by shifting the mean of the Gaussian PDF, because

$$P(\vec{x} + \vec{c} \mid \mathcal{N}(\vec{\mu}\ ,\ \Sigma)) = P(\vec{x} \mid \mathcal{N}(\vec{\mu} - \vec{c}\ ,\ \Sigma)) \tag{4.8}$$

where $\vec{x}$ is the observation vector, $\vec{c}$ is the adjustment due to the context, and $\mathcal{N}(\vec{\mu}\ ,\ \Sigma)$ is the Gaussian PDF with mean vector $\vec{\mu}$ and covariance matrix Σ. Thus, instead of modifying the input feature vector, we evaluate the original feature vector against the corresponding context-dependent model, which is obtained by shifting the mean of the retrained context-independent model accordingly. This is somewhat similar to the *tied model translation* technique used in speaker adaptation (Shinoda and Watanabe 1996; Kannan and Ostendorf 1997). Alternatively, we can group context-dependent models into similar classes through clustering techniques to handle the sparse data problem. We explored both of these methods to train robust context-dependent tone models and obtained comparable performance in speech recognition.

4.3.4 Summary of Tone Classification Results

Table 4-1 summarizes the tone classification results on the digit test data, and Table 4-2 summarizes the tone classification results on the read and spontaneous test sets of YINHE data. The classification error rate using simple four tone models is 18.6% for digit data, 32.7% for read YINHE data, and 35.4% for spontaneous data, consistent with the complexity of each data set. We excluded the neutral tone from tone models, because our initial experiments showed that the addition of neutral tone reduced the tone model contribution to speech recognition improvement. The refinements to the tone models achieve performance improvements for all three data sets. However, we notice that the relative error reduction also decreases from digit data to the more complex spontaneous YINHE data.

Statistical Significance

To examine whether the improvements in tone classification performance by using more refined tone models are statistical significant, we performed McNemar's test between each

System Configuration	**Classification Error Rate** (%)	**Relative Reduction** (%)
Simple	18.6	-
+ Intonation	16.0	14.0
+ Intonation & Context	13.3	28.5

Table 4-1: Four-tone classification results of simple and more refined tone models on the digit data.

System Configuration	**Read**		**Spontaneous**	
	ER(%)	**Rel.**(%)	**ER**(%)	**Rel.**(%)
Simple	32.7	-	35.4	-
+ Intonation	29.5	9.8	33.1	6.5
+ Intonation & Context	27.4	16.2	31.1	12.1

Table 4-2: Four-tone classification results of simple and more refined tone models on the read and spontaneous YINHE data (neutral tone excluded). "ER" is the tone classification error rate. "Rel." is the relative reduction of errors from the baseline performance.

pair of classification outputs (Gillick and Cox 1989). The McNemar significance level reflects the probability of the *hypothesis* that the differences between two classification results occur by chance. We set the threshold of the significance level to be 0.05, which means that the differences are considered as statistically significant if the probability of the differences occurring due to chance is less than 0.05. As summarized in Table 4-3 and Table 4-4, all classification performance differences are statistically significant.

	+ Intonation	+ Intonation & Context
Simple	*0.001*	*0.001*
+ Intonation	-	*0.001*

Table 4-3: Measure of statistical significance of tone classification performance differences on the digit data. Significant differences are shown in *italics*, while insignificant differences are shown in **boldface** (based on a threshold of 0.05). Significance levels less than 0.001 are mapped to 0.001 for simplicity.

	Read		Spontaneous	
	+ Intonation	+ Intonation & Context	+ Intonation	+ Intonation & Context
Simple	*0.001*	*0.001*	*0.015*	*0.001*
+ Intonation	-	*0.015*	-	*0.034*

Table 4-4: Measure of statistical significance of tone classification performance differences on the read and spontaneous YINHE data. Significant differences are shown in *italics*, while insignificant differences are shown in **boldface** (based on a threshold of 0.05). Significance levels less than 0.001 are mapped to 0.001 for simplicity.

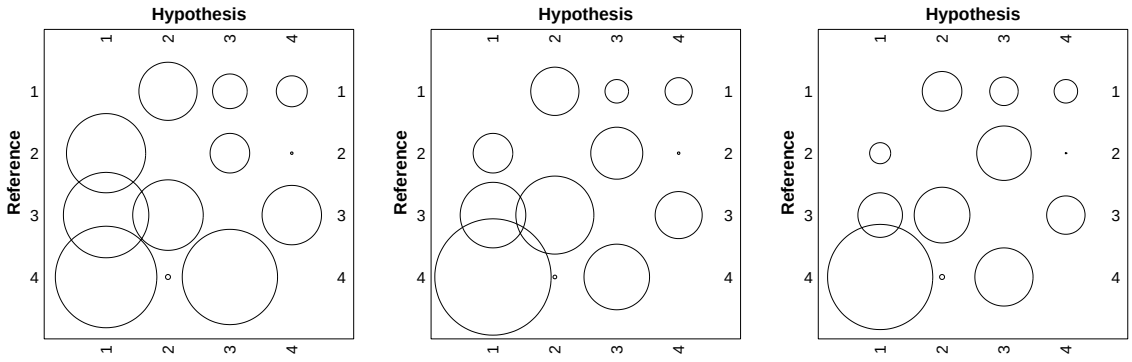

Figure 4-2: Bubble plots of classification errors for simple tone models (left), tone models normalized for F_0 declination removal (middle), and tone models normalized for both F_0 declination and context (right).

Classification Error Analysis

We have demonstrated that the tone classification performance is significantly improved after we account for the F_0 declination and context effects into tone modeling. To examine the details of the improvements, we study the change in confusion errors after each step of tone model refinement.

Figure 4-2 shows the bubble plots of the confusion matrices of the three sets of classification results on the digit data, excluding the diagonal elements. The *radius* of each bubble is proportional to the number of corresponding confusion errors, so that the size of the bubbles in the three plots can be compared. It is observed that most types of confusion errors are reduced after taking the F_0 declination and context into consideration. The most

prominent effects of F_0 declination removal seem to be the reduction of confusions between tones with different F_0 averages. For example, the confusion of tone 1 with tone 2 or tone 3, the confusion of tone 2 with tone 1, the confusion of tone 3 with tone 1 or tone 4, and the confusion of tone 4 with tone 3 are all greatly reduced. However, there is also a small increase in confusion between tones with similar F_0 averages. For example, the confusion between tone 2 and tone 3 and the confusion of tone 4 with tone 1 are slightly increased. The context normalization seems to reduce all types of confusion errors, except for a minor increase in confusion of tone 1 or tone 2 with tone 3. This is a small price to pay considering the greatly reduced confusions of tone 3 with tone 1 and tone 2. Similar trends of improvements with refined tone models are also observed on the YINHE data, although the distribution of confusion errors and the magnitude of improvements differ.

4.4 Incorporation of Tone Models into Speech Recognition

This section describes the implementation of two different mechanisms to incorporate tone models into Mandarin speech recognition. The post-processing approach applies the tone models to resort the recognizer N-best list, and the first-pass approach uses tone scores directly in the Viterbi search.

4.4.1 Post-Processing

We have found that the N-best list has great potential for improved speech recognition performance. For example, with a perfect post-selector, we could achieve less than 1% syllable error rate and 7% utterance error rate with a 10-best list for the digit domain, as compared to the 1-best performance of 4.3% syllable error rate and 28.7% utterance error rate. So we first tried to apply tone models to resort the recognizer 10-best outputs to improve speech recognition accuracy.

The post-processing scheme is similar to that proposed in (Serridge 1997). For each syllable final in an A^* path, the tone score is added to the total path score. The N-best hypotheses are then resorted according to the *adjusted total scores* to give a new "best" sentence hypothesis. We tried two methods to normalize the total adjustments to avoid

System	Sub.	Del.	Ins.	SER	UER
Baseline	1.0	2.2	1.0	4.3	28.7
+ Simple	0.9	1.5	0.8	3.2	20.3
+ Intonation	1.0	1.4	0.7	3.1	19.7
+ Intonation & Context	0.9	1.1	0.8	2.8	18.0

Table 4-5: Recognition error rates (in percentage) on digit data without tone models and with various tone models incorporated to resort the 10-best outputs. "SER" is the syllable error rate, which is the sum of the substitution, insertion, and deletion error rates. "UER" is the utterance error rate.

bias toward shorter strings: dividing the total tone score by the number of syllables in each path, or adding a *tone transition weight* to each added tone score. The rationale for the first method is to use the average tone score as an indication of the tone hypothesis quality of each path; the second method tries to adjust the tone likelihood scores to be zero-centered on average, similar to using a *segment transition weight* in the SUMMIT system. We also use a scaling factor to weight the scores contributed by the tone models, which can be optimized empirically based on recognition performance. Context-dependent model scores can also be applied simply by converting a tone hypothesis into its context-dependent form, with context obtained from its surrounding hypotheses.

We conducted speech recognition experiments on the digit and the YINHE data to see if applying tone models to resort the recognizer N-best outputs can improve speech recognition performance. We are also interested to know if the refined tone models can improve over the simple models, as implied by the encouraging tone classification results.

Table 4-5 summarizes the speech recognition performance on the digit domain with various tone models incorporated to resort the 10-best list; the baseline performance without tone models is also listed for comparison. As indicated in the table, the application of simple four tone models greatly reduced the syllable error rate from the baseline system. However, using more refined tone models further reduced the syllable and utterance error rates only slightly, compared to using the simple tone models.

We would like to evaluate the significance level of our recognition results, especially for those where differences are small. Unlike the classification results, the speech recognition

	+ Simple	+ Intonation	+ Intonation & Context
Baseline	*0.001*	*0.001*	*0.001*
+ Simple	-	**0.407**	**0.099**
+ Intonation	-	-	**0.171**

Table 4-6: Measure of statistical significance of speech recognition performance differences on the digit data. Significant differences are shown in *italics*, while insignificant differences are shown in **boldface** (based on a threshold of 0.05). Significance levels less than 0.001 are mapped to 0.001 for simplicity.

hypotheses from two systems can not be matched on a syllable-to-syllable basis due to insertion and deletion errors; furthermore, the adjacent syllables in a hypothesis are also likely to be dependent on each other because of lexical and language model constraints. A *matched pairs segment word error test* has been developed by (Gillick and Cox 1989) to address this problem. First, the output stream from a recognizer is divided into *phrase segments* (sentences or phrases bounded by long pauses) such that the errors in one segment are assumed to be statistically independent of the errors in any other segments. A *matched-pairs* test is then performed on the average difference in the number of errors in such segments made by two algorithms. Table 4-6 summarizes the significance levels of the differences between each pair of recognition results for the digit domain. As shown in the table, the improvements using various tone models are all significant compared to the baseline performance, but the relative differences among using different tone models are not statistically significant.

We also applied the tone models to resort the *10*-best outputs of the YINHE recognizer. The read YINHE data were used to optimize the relative weight of the tone score contribution, and the speech recognition performance is reported on the spontaneous data. Similar to the digit domain, application of various tone models all reduced the syllable and utterance error rates, as shown in Table 4-7. However, the improvements of more refined models over the simple model are very small and not statistically significant, as indicated in Table 4-8.

System	Sub.	Del.	Ins.	SER	UER
Baseline	5.9	1.2	1.1	8.2	29.9
+ Simple	5.2	0.9	1.1	7.2	27.8
+ Intonation	5.0	0.9	1.1	7.0	26.8
+ Intonation & Context	5.0	0.8	1.1	6.9	26.8

Table 4-7: Recognition error rates (in percentage) on spontaneous YINHE data without tone models and with various tone models incorporated to resort the 10-best outputs. "SER" is the syllable error rate, which is the sum of the substitution, insertion, and deletion error rates. "UER" is the utterance error rate.

	+ Simple	+ Intonation	+ Intonation & Context
Baseline	*0.033*	*0.015*	*0.010*
+ Simple	-	**0.276**	**0.522**
+ Intonation	-	-	**0.904**

Table 4-8: Measure of statistical significance of speech recognition performance differences on the spontaneous YINHE data. Significant differences are shown in *italics*, while insignificant differences are shown in **boldface** (based on a threshold of 0.05).

4.4.2 First-Pass

The post-processing approach can not recover the correct hypothesis if it is not in the N-best list. Here we examine if combining tone scores with segment and boundary model scores directly in the first-pass Viterbi search can lead to a performance that is superior to that obtained by the resorting method, because the tone models are utilized to explore the entire search space.

A straightforward way to incorporate tone information into speech recognition is to augment the acoustic features with tone features, and build tone-dependent syllable final models; thus, no change needs to be made to the recognizer. However, this approach has the disadvantage of splitting the training data, i.e., the same tone from different finals as well as the same final with different tones can not be shared in training. In addition, it is unclear how to interpret the use of F_0 features in the non-tonal units such as syllable initials and silences. Given that the segmental and tonal features are relatively independent of each

other, we can build separate segment and tone models, and combine the log probability scores for tonal segments as follows:

$$\log P(AT|\vec{x},\vec{y}) = \alpha \cdot \log P(A|\vec{x}) + \beta \cdot \log P(T|\vec{y}) \tag{4.9}$$

where $\vec{x}$ is the segment feature vector, $\vec{y}$ is the tone feature vector, A is a segment symbol (must be a syllable final), T is a tone symbol, and AT is the combined tonal segment symbol. This method is similar to the *committee-based classifier structure* described in (Halberstadt 1998) for combining heterogeneous acoustic features, where the feature vectors are assumed to be independent of each other. The scaling factors α and β are used to determine the relative weighting for the segment and tone scores, which can be adjusted empirically on held-out data to achieve optimal recognition performance. For the segments that do not have a well defined tone, such as the syllable initials and various silence models, the tone score is simply ignored. Similar to the post-processing approach, we add a transition weight to each tone score to avoid bias toward hypothesizing fewer tonal segments. The combined tonal segment scores are used to replace the pure segment scores, so that no change needs to be made to the Viterbi search module.

Table 4-9 summarizes the speech recognition performance on the digit domain and the YINHE domain, with the simple models for four lexical tones incorporated into the first-pass search or applied to resort the N-best list. As shown in the table, the performances using the two strategies are very similar, with a slight advantage for the first-path approach. We performed a significance test on the differences between the two methods on both domains. The significance level of the difference is 0.219 for the YINHE data and 0.242 for the digit data, which means that the differences are not statistically significant.

We found that the N-best list obtained using the first-pass approach was of similar quality to that of the baseline system on the digit domain. In addition, the results shown in the previous section suggest that more refined tone models did not improve over the simple four tone models significantly. Thus, we did not try to apply more refined tone models on the new N-best list to obtain further improvements.

Domain	Method	Sub.	Del.	Ins.	SER	UER
Digit	Baseline	1.0	2.2	1.0	4.3	28.7
	Post-processing	0.9	1.5	0.8	3.2	20.3
	First-pass	0.9	1.2	0.9	3.0	19.7
YINHE	Baseline	5.9	1.2	1.1	8.2	29.9
	Post-processing	5.2	0.9	1.1	7.2	27.8
	First-pass	5.1	0.9	0.9	6.9	26.8

Table 4-9: Speech recognition error rates (in percentage) on the digit and the YINHE data with simiple four tone models incorporated using first-pass and post-processing methods. The baseline performance in each domain is also included for reference. "SER" is the syllable error rate, which is the sum of the substitution, insertion, and deletion error rates. "UER" is the utterance error rate.

4.4.3 Performance Analysis

We examined the detailed differences of speech recognition results with and without tone models on the digit data. For the baseline recognizer, most of the errors are caused by insertions and deletions of the single-vowel digits "*yi*1" (one), "*er*4" (two) and "*wu*3" (five). There are 22 deletion errors for "*wu*3", 22 deletion errors for "*er*4", 10 deletion errors for "*yi*1", and 16 insertion errors of "*wu*3", making up 57% of the 123 total errors. We find that a large portion of the improvements obtained by using tone models are due to the reduction of insertions errors for "*er*4" and "*wu*3" and deletion errors for "*wu*3". For example, by applying tone models refined for both F_0 declination and context effects to resort the 10-best list, the recognizer makes only 7 deletion errors for "*wu*3", 11 deletions of "*er*4", and 12 insertions of "*wu*3". There are small reductions in other types of errors as well. The error reduction for digit "*yi*1" is relatively small because of the flat contour pattern of tone 1: the F_0 contour of two first tones will not differ much from that of a single first tone, except for duration differences; thus, the tone models are not very effective to resolve such cases. We suspect that a separate duration model with speaking rate considerations would be more effective to deal with such errors (Wang and Seneff 1998).

Figure 4-3 illustrates the correct recognition of the digit string "*ba*1 *er*4 *jiu*3 *wu*3 *wu*3" (eight two nine five five) with the assistance of tone models. As shown in the figure, the spectrogram has few acoustic cues to indicate the presence of two consecutive "*wu*3"'s in

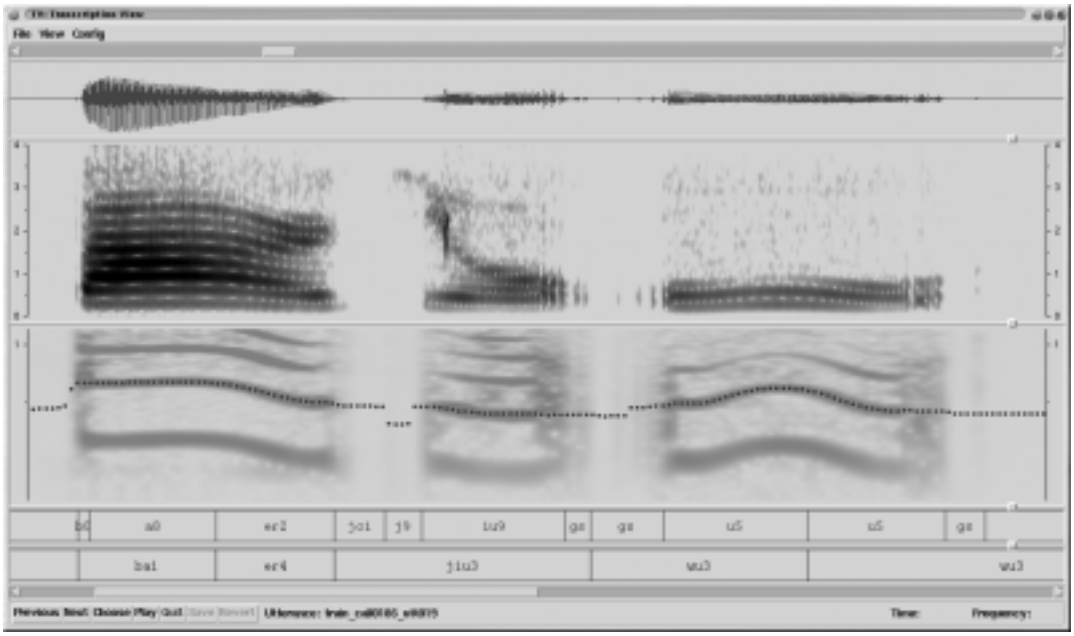

Figure 4-3: Waveform, spectrogram, pitch contour overlayed with DLFT spectrogram, phone transcription, and word transcription for the Mandarin digit utterance "ba1 er4 jiu3 wu3 wu3 ..." (eight two nine five five).

the signal, which is likely to result in a deletion of "*wu*3" if only segmental information is used. However, the pitch contour clearly contradicts with the hypothesis of a single "*wu*3", because the contour has a rise-fall shape instead of the typical low falling contour of tone 3. In fact, the contour supports the hypothesis of two "*wu*3"'s, taking into consideration the "33 → 23" tone sandhi rule. The correct hypothesis is likely to succeed with the addition of high tone model scores during speech recognition.

4.5 Summary

In this chapter, we presented tone classification and speech recognition experiments on Mandarin telephone speech of various degrees of linguistic complexity. We developed a segment-based tone classification framework, which uses discrete Legendre decomposition to parameterize tone F_0 contours and Gaussian classifiers to estimate tone probability scores. Using this basic framework, we demonstrated that tone recognition performance for *con-*

tinuous Mandarin speech can be significantly improved by taking into account sentence declination, phrase boundary, and tone context influences. We also developed two mechanisms to integrate tone models into speech recognition. Incorporation of a simple four-tone model into the first-pass Viterbi search reduced the speech recognition error rate by 30.2% for the digit domain and by 15.9% for the spontaneous utterances in YINHE domain. Using the simple four-tone model to resort the recognizer 10-best outputs yielded similar improvements for both domains. However, the more refined tone models fail to further improve the speech recognition performance significantly, even though the classification performances indicate otherwise. This seems to suggest that a simple and efficient strategy to utilize tone information can be achieved by integrating a simple four-tone model into the Viterbi search.

Chapter 5

Lexical Stress Modeling for Spontaneous English Speech Recognition

Lexical stress is an important property of the English language. It has been suggested in (Lea 1980) that stressed syllables provide *islands of phonetic reliability* in speech communication: the acoustic realization of a stressed vowel tends to be more distinctive than that of its unstressed counterpart (Lehiste 1970); phonological variations for stressed syllables are more limited compared to unstressed ones when the rate of speech changes (Lea 1980); human recognition of phones by visual examination of speech spectrograms achieved higher accuracies in stressed syllables than in unstressed or reduced syllables (Klatt and Stevens 1972; Lea 1973); and phoneme recognition by machine obtained much higher accuracy on stressed nuclei than on unstressed nuclei (Jenkin and Scordilis 1996). In addition, lexical studies have demonstrated that stressed syllables are more informative to word inference (Huttenlocher 1984; Carter 1987), and knowing the stress pattern of a word can greatly reduce the number of competing word candidates (Aull and Zue 1985). Clearly, lexical stress contains useful information for automatic speech recognition.

Early work on lexical stress modeling has focused on the recognition of stress patterns to reduce word candidates for large-vocabulary isolated word recognition (Aull and Zue 1985;

Waibel 1988), or to disambiguate stress-minimal word pairs (Freij and Fallside 1990). More recently, there have been attempts at utilizing stress information to improve *continuous* speech recognition. In (Adda-Decker and Adda 1992; Sjölander and Högberg 1997), the lexical stress property was used to separate phones during training to obtain more accurate acoustic models. In (Hieronymus et al. 1992), stress-dependent phonological rules were applied for phone to phoneme mapping. In (Jones and Woodland 1994), hidden Markov models for "weak/strong" and "stressed/unstressed" syllables were applied to resort the recognizer N-best outputs. A few studies also examined stress classification in continuous speech (Jenkin and Scordilis 1996; van Kuijk and Boves 1999); however, no speech recognition experiments were performed using the resulting stress models. In general, previous research on using stress models in continuous speech recognition has been limited, and we have not found any work on spontaneous English speech reported in the literature.

Encouraged by our results in using tone information to improve Mandarin speech recognition, we are inclined to apply the same approach to model English lexical stress for *spontaneous telephone* speech in the JUPITER domain (Zue et al. 2000). The motivation is a natural extension from the Chinese case, i.e., erroneous hypotheses will have worse *stress scores* than the correct hypothesis. We expect that substitution, insertion and deletion errors sometimes result in mismatched stress characteristics between the hypothesized syllable nucleus and its acoustics. By scoring for the stress pattern of a hypothesis, a system augmented with the additional constraints provided by stress models will perform better than a system which uses segmental constraints only. However, unlike Mandarin tones, the acoustic manifestations of English lexical stress are quite obscure. Although it has been found that prosodic attributes, i.e., energy, duration, and pitch, correlate with the stress property of a vowel, these features are also highly dependent on its segmental aspects (intrinsic values). To complicate things further, not all lexically stressed syllables are stressed in continuous speech. For example, mono-syllabic function words are often not stressed. In addition, a subset of lexically stressed syllables in a sentence also carry the pitch accents of the spoken utterance. Although "pitch accentedness" has been argued to be a more appropriate indication of "stress" in continuous speech, their occurrences can not be predicted from orthographical transcriptions, and hence, they are less useful to a recognizer.

On the other hand, lexical stress can easily be encoded in the lexicon of a segment-based recognizer. However, the question remains whether it can be determined from the acoustics in spontaneous speech with sufficient reliability to benefit recognition.

In this chapter, we test the approach of scoring the lexical stress patterns of recognizer hypotheses to improve automatic speech recognition performance. The research issues we want to address are: (1) how well can the underlying stress of a vowel be determined from the acoustics in spontaneous speech, and (2) can such information improve speech recognition performance? To answer these questions, we study the correlation of various pitch, energy, and duration measurements with lexical stress on a large corpus of spontaneous utterances, and identify the most informative features of stress using classification experiments. We also develop probabilistic models for various lexical stress categories, and combine the stress model scores with other acoustic scores in the recognition search for improved performance. We experimented with prosodic models of varying complexity, from only considering the lexical stress property to also taking into account the intrinsic differences among phones. We found that using prosodic models reduced the word error rate of a state-of-the-art baseline system in the JUPITER domain. However, the gain by using prosodic models seemed to be achieved mainly by eliminating implausible hypotheses, rather than by distinguishing the fine differences among various stress and segmental classes; thus, we found no additional gain by utilizing more refined modeling.

In the following sections, we first give a brief introduction to some previous research on lexical stress modeling for speech recognition. Then we provide some background knowledge for the experiments reported in this chapter, including the JUPITER corpus and a baseline JUPITER recognizer which incorporates stress markings in its lexicon. After that, we study the correlation of various prosodic measurements with lexical stress and identify the best feature set using classification experiments. Finally, we present speech recognition experiments using the basic lexical stress models and other prosodic models of varying complexity.

5.1 Related Research

There are usually two approaches to using lexical stress information in continuous speech recognition. One is to build separate models for lexically stressed and unstressed phones (usually only vowels), so that more accurate acoustic models can be obtained. The other method is to use stress information to assist lexical decoding, by using stress-dependent phonological rules, or by scoring explicitly for the stress patterns of the recognition hypotheses. In this section, we review some previous work in each category.

An apparent motivation for building stress-dependent acoustic models is that the acoustic realization of a stressed phone tends to be more distinctive than that of its unstressed counterpart (Lehiste 1970). By building separate models for stressed and unstressed phones, the models for stressed phones will be "cleaner" and provide more differentiation power. However, previous research following this approach has shown mixed results. In (Adda-Decker and Adda 1992), stress-dependent acoustic models were trained for continuous French and American-English speech respectively. It was found that the addition of stressed vowel models improved phonetic recognition performance on the French database, but not on the English database. One possible explanation for this discrepancy is that the two languages have very different lexical stress patterns. English has *free stress*, because the stressed syllable can assume any position in a word; while the French language has *bound stress*, in which the position of the stressed syllable is fixed in a word. In this study, the "stress" for French and English also seemed to be labeled differently. The stress markings for the French data were labeled by a human listener, and stressed words were added to the lexicon only if they occurred in the training data. Thus, the stress in this case takes into account sentence-level context. For the English data, it appeared that the stress labels were obtained from dictionary definitions. In (van Kuijk et al. 1996), stress-dependent models were trained for a continuous telephone speech database of Dutch, also a *free stress* language. The stress labeling was based on lexical stress, although two sets of models were trained depending on different treatments of monosyllabic function words. In either case, the stress-dependent models failed to show improvement over a baseline system with no stress markings. The authors suggested two possible reasons for the negative results. First, there were no explicit stress-related features in the acoustic models; and second, the

acoustic correlates of lexical stress were more susceptible to interference from higher-level prosodic phenomena in fluent speech. A third possibility might be that the number of training examples for stressed and unstressed vowels is usually very unbalanced. Thus, a blind separation of models by lexical stress lacks robustness due to many under-trained models. In (Sjölander and Högberg 1997), a set of phone-like units (PLUs) was derived using a decision tree algorithm, which considered phone identity, phone context, lexical stress, and function word affiliation as possible splitting criteria. It was found that PLUs derived using all of the above information outperformed PLUs derived using only phone information for a Swedish spontaneous speech corpus.

The approach of using stress information in lexical decoding usually requires classification of stress categories. Although stress can be reliably perceived by human listeners, its manifestations in the speech signal are very complex and depend on the language. In addition to syllable intensity, duration and pitch were also found to be correlated with lexical stress in English (Lea 1980). Some studies have also used spectral features, such as *spectral change*, measured as the average change of spectral energy over the middle part of a syllable (Aull and Zue 1985; Waibel 1988); *spectral tilt*, measured as spectral energy in various frequency sub-bands (Sluijter and van Heuven 1996; van Kuijk and Boves 1999); and Mel-frequency Cepstral coefficients (MFCCs) (Jones and Woodland 1994). The stress classification performance depends highly on the data, and also somewhat on the labeling of stress categories. High accuracy has been achieved for recognizing the complete stress patterns for isolated words (Aull and Zue 1985), and for distinguishing stress-minimal word pairs (Freij and Fallside 1990; Ying et al. 1996). The performance on unlimited continuous speech varies in the literature. In (Waibel 1988), a variety of feature combinations were tested on four relatively small English speech databases, and a best average error rate of 12.44% was reported with energy integral, syllable duration, spectral change and F_0 maxima features. In (Jenkin and Scordilis 1996), accuracies of about 80% were obtained on the English TIMIT database using various classifiers with 6 features from energy, duration and pitch. The stress labeling of the data was performed manually by two transcribers. In (van Kuijk and Boves 1999), the accuracy for a read telephone Dutch database was 72.6% at best. The stress labeling in this case was determined automatically from a dictionary. We

expect the classification accuracies to be higher if the stress labeling takes into consideration sentence-level effects, i.e., not all lexically stressed syllables are stressed (accented) in continuous speech (Shattuck-Hufnagel and Turk 1996).

A more important criterion is how much can the lexical stress models improve speech recognition performance. However, we are able to find only a few studies which reported speech recognition experiments using stress models (Hieronymus et al. 1992; Jones and Woodland 1994). In (Hieronymus et al. 1992), a phone lattice was first derived using hidden semi-Markov model (HSMM) based acoustic models, and a matrix of substitution, deletion and insertion costs between each phone and phoneme pair was used during lexical decoding. It was found that the word error rate was reduced by 66.1% when the cost matrix incorporated stress-dependent phonemes, mapping to phones that were not marked for stress. However, when the phones were also marked for stress, the performance gain was not as great. In (Jones and Woodland 1994), HMMs using energy and MFCCs as features were built for "weak/strong" (lexical-level stress) and "stressed/unstressed" (sentence-level stress) syllables and applied to resort the recognizer N-best outputs. The word error rate on the TIMIT database was reduced by 16% with the "strong-weak" models and by 23% with the "sentence-stress" models.

Although several studies have tackled the problem of stress classification in continuous speech, research on its application in speech recognition systems has been fairly limited. In our investigation, we will build acoustic models to capture the "stressedness" of syllable nucleus vowels and apply these stress models to score for the underlying stress pattern of a hypothesis for additional constraints in speech recognition. The motivation of this approach is similar to that of (Jones and Woodland 1994); however, the implementation aspects are quite different. The speech data and the recognition task in our study are also more challenging than those in most of the work introduced in this section.

5.2 Experimental Background

The speech recognition experiments are conducted on the JUPITER speech database, which was collected by recording conversations between users and the JUPITER weather informa-

User:	*What is the temperature of Green Bay Wisconsin?*
JUPITER:	In Green Bay today: high 11. Tonight: low minus 3 tonight. Is there anything else?
User:	*How much snow was on the ground in Green Bay Wisconsin?*
JUPITER:	Sorry, I'm not sure what you said.
User:	*What is the snowfall in Green Bay Wisconsin?*
JUPITER:	In Green Bay today: mostly cloudy with some snow flurries and snow showers. Is there something else you'd like to know?
User:	*Is there rain in Chicago?*
JUPITER:	In Chicago today: occasional rain. Tonight: rain, possibly a thunderstorm, near 100 percent chance of rain. Is there something else I can do for you?
User:	*What is the weather in El Paso Texas?*
JUPITER:	In El Paso today: clear and cool, high in mid or upper 50s. Tonight: clear, lows in mid 20s. What else would you like to know?
User:	*Is the sun shining in Fort Lauderdale?*
JUPITER:	In Fort Lauderdale today: periods of sunshine, scattered clouds. What other information can I give you?
User:	*Good bye!*
JUPITER:	Thanks for using Jupiter! Have a nice day!

Table 5-1: Example of user interaction with the JUPITER weather information system.

tion system through a telephone. In the following, we first introduce briefly the JUPITER corpus. Then we describe a baseline recognizer for the JUPITER domain, which was adapted from an existing JUPITER recognizer to facilitate lexical stress modeling experiments.

5.2.1 JUPITER Corpus

The JUPITER system (Zue et al. 2000) is a telephone-based conversational interface to online weather information, developed at the Spoken Language Systems group of the MIT Laboratory for Computer Science. A user can call the system via a toll-free number and ask weather-related questions using natural speech. JUPITER has real-time knowledge about the weather information for over 500 cities, mostly within the United States, but also some selected major cities world-wide. The system also has some content processing capability, so that it can give specific answers to user queries regarding weather actions, temperature, wind speed, pressure, humidity, sunrise/sunset times, etc. Table 5-1 shows an example dialogue between a user and JUPITER.

Data Set	# Utterances	Data Set	# Utterances
train	84,165	-	-
clean_test_1	1,819	all_test_1	2,468
clean_test_2	1,715	all_test_2	2,335
clean_test_3	1,313	all_test_3	1,861

Table 5-2: Summary of data sets in the JUPITER corpus.

A sizable amount of spontaneous telephone speech has been collected since the system was made publicly available via a toll-free number. To date, there have been over 756,000 utterances from over 112,000 phone calls recorded since May, 1997, and the data are still coming in. We use about 85,000 orthographically transcribed utterances in our experiments. Table 5-2 summarizes the number of utterances in the training and various test sets. The training set and the "clean" test sets contain only within-vocabulary utterances, while the "all" test sets also include utterances that contain out-of-vocabulary words and other artifacts, such as background speech, noise, etc. We use only "clean" data in the experiments described in this chapter: the "clean_test_1" set is used as development data for selecting stress features as well as for tuning various weights; the "clean_test_2" and "clean_test_3" sets are used for testing. The "all" test sets will be used for developing recognition confidence measures described in the next chapter.

5.2.2 Baseline JUPITER Recognizer

The baseline recognizer was adapted from an existing JUPITER recognizer, configured from the SUMMIT recognition system (Ström et al. 1999). Lexical stress markings were added to the 2,005-word lexicon to facilitate lexical stress modeling experiments. The initial stress labels were obtained from the LDC PRONLEX dictionary, in which each word has a vowel with primary stress and possibly a vowel with secondary stress. However, in continuous speech, the vowels of mono-syllabic function words, such as "*a*", "*it*", "*is*", etc., are likely to be unstressed or reduced. The JUPITER recognizer uses a few specific reduced vowel models as alternative pronunciations to account for these variations. Initially, the full vowels in mono-syllabic function words were marked with primary stress, as in PRONLEX. However, we

WHAT	is	the	TEMperature	in	GREEN	BAY	wisCONsin	?
(What	is	the	temperature	in	Green	Bay	Wisconsin	?)

Table 5-3: The lexical stress pattern of an example JUPITER utterance. Stressed syllables are indicated by capital letters.

found that too many vowels (more than 60%) were labeled with primary stress in the forced transcriptions derived with this lexicon. We thus labeled the full vowels in mono-syllabic function words as unstressed, with exceptions for a few wh-words such as "what", "when", "how", etc. We realize that this is only a coarse approximation, because function words can be stressed in continuous speech, while lexically stressed syllables in content words are not necessarily always stressed in spoken utterances. Table 5-3 illustrates the stress labeling of an example JUPITER utterance, with stressed syllables indicated by capital letters.

Secondary stress could be grouped with primary stress, be treated as unstressed, or be kept as a third stress category. We decided to defer the decision until after data analysis, so primary and secondary stress were marked distinctively in the lexicon. The reduced vowels were also distinguished from unstressed full vowels in our data analysis for more detailed comparison.

The baseline system uses boundary models only, because it was found that adding segment models did not improve recognition performance unless context-dependent segment models were used, in which case the speed of the recognizer was significantly slower (Ström et al. 1999). This is possibly because both models use features that are based on Mel-frequency Cepstral coefficients; thus, the context-independent segment models are somewhat redundant when boundary models are used. Our proposed approach is to reintroduce segment models, but to restrict them to prosodic aspects only. We hope that prosodic features can provide independent information to complement the boundary models, thus achieving improved recognition performance. Therefore, we did not try to retrain boundary models; the diphone labels in each boundary model class were simply expanded to cover variations in lexical stress. Both bigram and trigram language models were used by the recognizer, applied with the Viterbi and the A^* search passes. The modified recognizer achieved the same speech recognition performance as the original recognizer, which was the

Set	Sub.	Del.	Ins.	WER	UER
Development (clean_test_1)	4.3	1.6	1.7	7.6	20.2
Test (clean_test_2+clean_test_3)	5.8	2.9	2.2	10.9	24.8

Table 5-4: Baseline speech recognition performance on the JUPITER development data and test data. "WER" is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. "UER" is the utterance error rate. All numbers are in percentage.

state-of-the-art real-time performance on the JUPITER domain as reported in (Ström et al. 1999). The detailed baseline performances on the development and test data are summarized in Table 5-4. Various parameters in the recognizer have been optimized to achieve the lowest overall error rates on the development data.

5.3 Acoustic Correlates of Lexical Stress

The primary acoustic correlates of stress for English include all three prosodic attributes: energy, duration, and pitch. Stressed syllables are usually indicated by high sonorant energy, long syllable or vowel duration, and high and rising F_0 (Lea 1980). Some studies also used spectral features such as spectral energy change, sub-band spectral energy, and MFCCs (Aull and Zue 1985; Waibel 1988; Jones and Woodland 1994; Sluijter and van Heuven 1996; van Kuijk and Boves 1999). In this section, we study the distribution of various prosodic measurements for each lexical stress category, and determine the "best" feature set for lexical stress by performing classification experiments on development data. Some spectral features will also be examined in the classification experiments. All features are extracted from the nucleus vowel of a syllable, so that the resulting model can be directly incorporated into the recognition first-pass search.

Forced recognition[1] was used to generate phonetic alignments (with stress marks on vowels) for the training and development data, and alternative pronunciations in the lexicon were determined automatically by the boundary models. These automatically derived stress

[1]Forced recognition refers to the recognition mode in which the correct words are provided to the recognizer and the recognition system finds the corresponding sequence of phones and their time alignments given a lexicon and acoustic models.

Stress Category	# Tokens	Percentage
Reduced	35,991	22.6%
Unstressed	48,130	30.3%
Primary Stress	63,432	39.9%
Secondary Stress	11,523	7.2%
Total	**159,076**	**100.0%**

Table 5-5: Number of tokens in each lexical stress category for 20,000 training utterances in the JUPITER corpus.

labels will serve as the reference for both training and testing the stress models. This "forced alignment" approach can only provide a very coarse labeling of lexical stress due to two main sources of errors: the boundary models can make errors at choosing among alternative pronunciations, especially between reduced vowels and unstressed full vowels; more importantly, dictionary-defined stress will not always be expressed as acoustic stress, even though some considerations for function words have been incorporated in the lexicon. In practice, the forced alignment process is iterated, once the stress models are trained and incorporated into the recognizer, to improve the quality of the transcriptions. The stress models can also be better trained using more "distinctive" tokens of each lexical stress category, as described in (van Kuijk and Boves 1999). The results shown in this section are based on iterated forced transcriptions with three iterations. We observed that the phone boundaries and the alternative pronunciations in the forced alignments all appeared to be more accurately determined after one iteration with lexical stress models.

About 20,000 utterances in the training set are used in the analysis described in this section. The full training set will be used to obtain the final test results. Table 5-5 summarizes the number of tokens in each stress category. Vowels with primary stress form the largest group, because all words except some mono-syllabic function words are required to have a primary stress syllable.

5.3.1 Prosodic Feature Distributions

In this section, we study the histogram distributions of various prosodic features given each lexical stress class. These features are grouped by energy, duration, and pitch categories,

as described in the following. The main purpose of the histograms is to illustrate the "distances" among the stress categories compared to the variance for each category; thus, the units in the histograms are omitted due to their lack of importance and meaningful physical interpretations.

Energy

The energy signal used in our analysis is the root mean square (RMS) energy, which is computed by taking the square root of the total energy in the amplitude spectrum of the short-time Fourier analysis of the speech. The short-time Fourier transform is computed at every 5 ms, with a Hamming window of 25.6 ms applied to the speech signal. To reduce variance due to "volume" differences, the raw RMS energy contour is scaled so that the average energy of each utterance in non-silence regions is roughly equal. Three energy measurements are extracted from each syllable nucleus vowel: the average and maximum values of the RMS energy, and the integral of the RMS energy over the vowel duration. Notice that the integral of the RMS energy is equivalent to the multiplication of the average energy and the duration of a vowel.

Figure 5-1 shows histograms of each energy feature for the four lexical stress categories. As plotted in the figure, vowels with primary stress have the highest energy, while reduced vowels have the lowest energy. It is interesting to observe that vowels with secondary stress have very similar energy distributions as unstressed full vowels. This seems to suggest that secondary stress syllables are acoustically unstressed when judged by energy cues. The energy distributions of unstressed and secondary stress vowels are in between those of the reduced and primary stress vowels. Although there is good separation between the two extremes, the overlap considering all categories is significant.

Duration

Duration is measured for the syllable nucleus vowel. We tried to normalize the raw duration measure with a sentence-level speaking rate to reduce the variances caused by different speaking rates. This is for data analysis only, because speaking rate information is usually not available during the first-pass recognition. The speaking rate is estimated from the

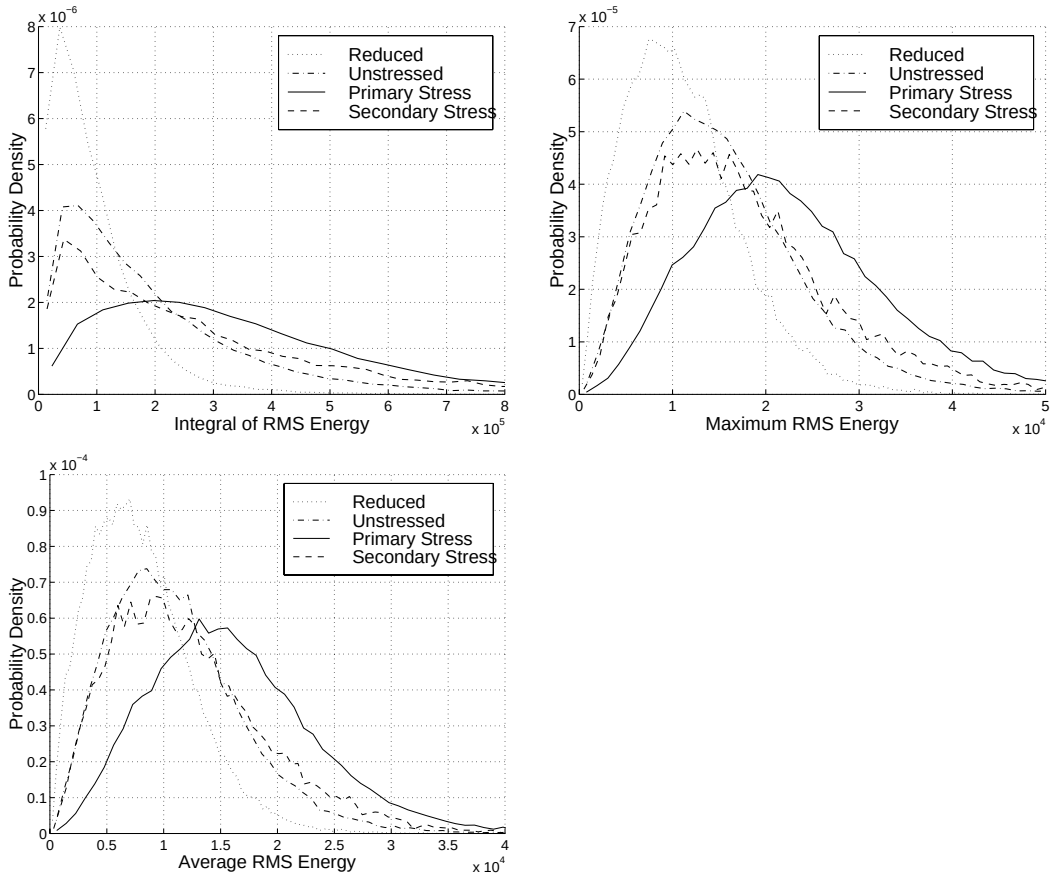

Figure 5-1: Histogram distributions of energy features for different stress classes in the JUPITER data.

vowels of an utterance as follows:

$$Speaking\ Rate = \frac{\sum \mu_{Dur}(V_i)}{\sum Dur(V_i)} \tag{5.1}$$

where $Dur(V_i)$ is the measured duration of the i^{th} vowel (V_i) in the sentence, and $\mu_{Dur}(V_i)$ is the average duration of V_i, computed from the entire corpus. The histograms of the raw duration, logarithmic duration, and the normalized duration for the four stress categories are plotted in Figure 5-2. The raw duration shows multiple peaks in the histograms. This is possibly due to speaking rate differences (slow vs. fast speech), because the normalized duration distribution does not show such characteristics. On average, stressed vowels have

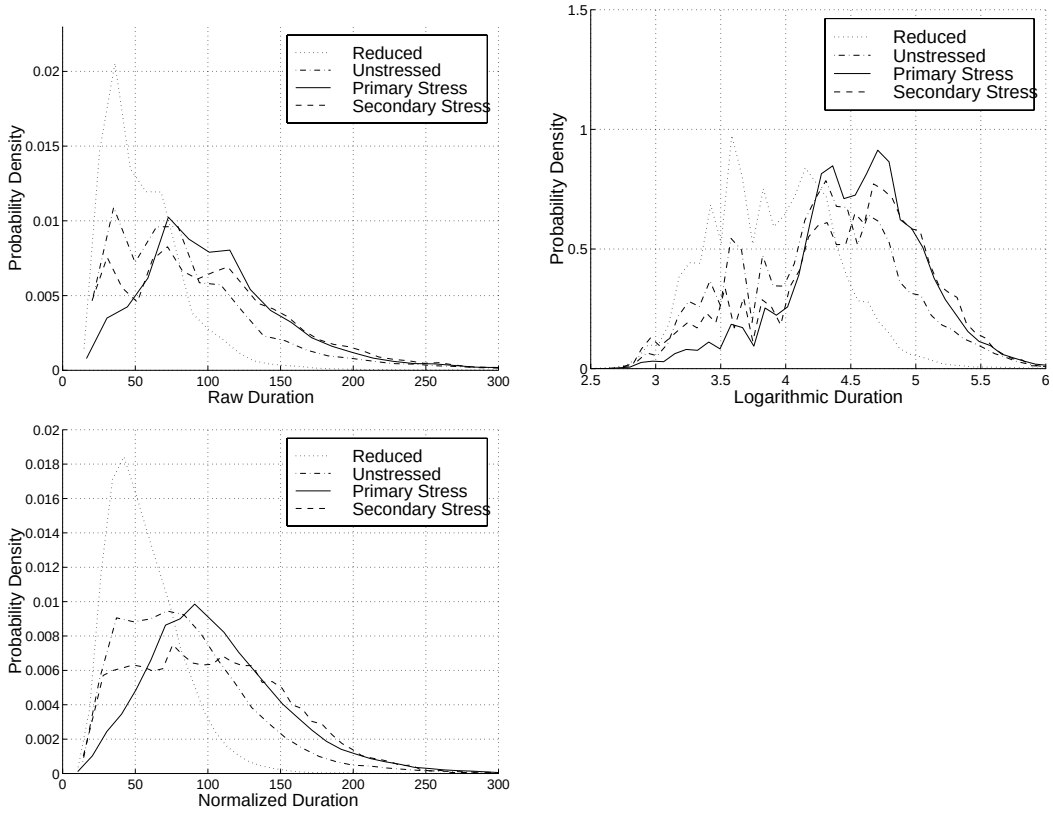

Figure 5-2: Histogram distributions of duration features for different stress classes in the JUPITER data. The *normalized duration* corresponds to the raw duration measure normalized with respect to a sentence-level speaking rate.

longer duration than unstressed full vowels, and reduced vowels are, as expected, the shortest. The large overlap among the full vowel categories is partially due to intrinsic duration interferences. We will address this issue in Section 5.4.

Pitch

The F_0 contour of each utterance is first normalized by a sentence-level average to reduce variations caused by speaker pitch differences. Four F_0-related measurements are examined, including the maximum F_0 value within the syllable nucleus, the average and slope of the F_0 contour of the nucleus vowel (similar to those of Mandarin tone contours described in Section 4.3.1), and the average probability of voicing, which is available via the voicing

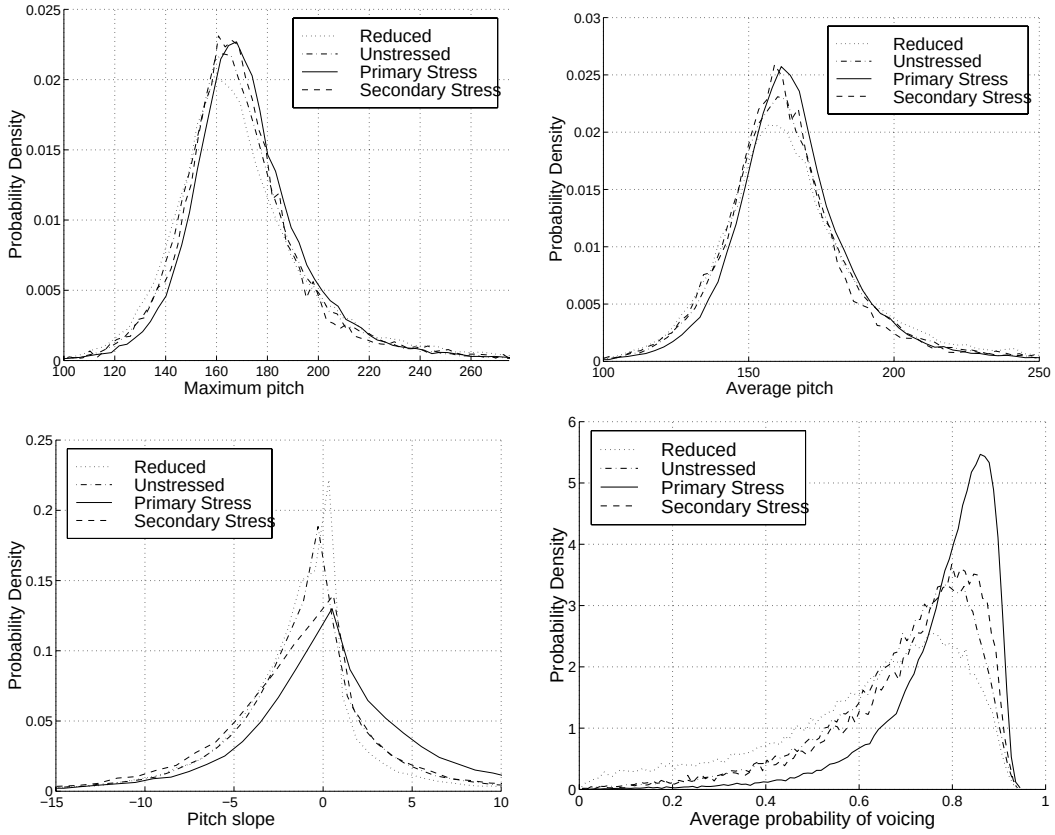

Figure 5-3: Histogram distributions of F_0 features for different stress classes.

estimation module of our pitch detection algorithm. We expect the average probability of voicing to be higher for stressed vowels than for unstressed and reduced vowels.

Figure 5-3 shows the histograms of these features for each lexical stress category. Contrary to some previous findings (Lea 1980; Waibel 1988), both the average and the maximum F_0 values show very small differences among different stress classes. The F_0 slope histograms show that the primary stress vowels are more likely to rise and less likely to fall than the other classes, possibly because of their function in signaling pitch accents. The voicing probability decreases with the "stressedness" of vowels as we expected. The overlap of pitch features is even more severe than that of energy and duration features.

5.3.2 Classification Experiments

The data analysis indicates that the distributions of most prosodic features differ for different lexical stress classes; however, the extent of overlap among classes is also significant. In this section, we use classification experiments to determine which set of features is most informative for modeling lexical stress. We will also include spectral tilt and MFCC features in the classification experiments, following the examples in (Sluijter and van Heuven 1996; Jones and Woodland 1994). *Spectral tilt* is characterized by the average logarithmic spectral energy in four frequency bands (in Hz): [0 500], [500, 1000], [1000, 2000], and [2000, 4000]. The MFCC features include 6 MFCCs averaged over the vowel duration.

For each stress feature vector, a principle component analysis is first applied, and mixtures of diagonal Gaussians are used to model the distribution of the feature vector for each lexical stress model. We trained models for all four lexical stress categories described in the previous section, because there seem to be some differences among all classes, and there are plenty of training data for each class. We listed both the four-class and the two-class classification accuracies for comparison. The two-class results are obtained by mapping the reduced, unstressed, and secondary stress classes into one "unstressed" class. Maximum likelihood (ML) classification is used, because we are interested to know how well the features can perform without the assistance of *priors*.

Table 5-6 summarizes the classification accuracy using each individual prosodic feature. As expected from the data analysis, the energy features performed the best, while the maximum and average pitch features yielded the poorest results. We noticed that the normalized duration (with respect to a sentence-level speaking rate) did not outperform the unnormalized duration measurements at stress classification, possibly due to large intrinsic duration interferences. We found the best-performing single feature to be the integral of energy over the syllable nucleus vowel, which combines the average energy and the duration information. This is consistent with the findings of a few previous studies (Waibel 1988; van Kuijk and Boves 1999).

Based on the results of individual features, we tried classification experiments using various combinations of features, including both the prosodic and the spectral measurements, as summarized in Table 5-7. The best set of *prosodic features* for stress classification consists

Feature	Accuracy (%) (four-class)	Accuracy (%) (two-class)
(1) integral of energy	**47.4**	**71.0**
(2) maximum energy	47.6	69.9
(3) average energy	45.7	70.3
(4) normalized duration	37.2	62.4
(5) raw duration	36.6	62.9
(6) logarithmic duration	41.8	61.1
(7) maximum pitch	32.8	56.2
(8) average pitch	33.1	52.9
(9) pitch slope	35.4	64.0
(10) avg. prob. voicing	43.9	62.2

Table 5-6: Vowel stress classification accuracy (in percentage) of each individual prosodic feature on the JUPITER development data.

of the integral of energy, raw duration, pitch slope, and the average probability of voicing. It is interesting to note that these features do not require any normalization with regard to measurements outside of the segment itself, nor do they require explicit knowledge of the phonetic identity. Adding spectral features improved stress classification performance, possibly because they capture the correlations between lexical stress and broad phone classes. The highest accuracy was achieved by combining MFCC features with the best prosodic feature set.

5.4 Speech Recognition Experiments

From our data analysis and classification experiments, it seems that prosodic features have only limited capability at distinguishing different lexical stress categories in spontaneous speech. In this section, we examine if scoring for lexical stress patterns in the recognizer hypotheses using these models can improve speech recognition performance on the JUPITER domain. We also try to refine the models by taking into account the intrinsic prosodic differences among nucleus vowels.

The four lexical stress models are incorporated into the first-pass Viterbi search using the same framework developed for Mandarin tones; except in this case, no phonetic segment

Feature Combination	Accuracy (%) (four-class)	Accuracy (%) (two-class)
(1)+(5)+(9)+(10)	48.5	73.0
(1)+(6)+(9)+(10)	48.3	72.9
(1-3)+(5-10)	49.4	72.6
(11) sub-band energy (4 features)	44.0	68.3
(12) MFCCs (6 features)	51.4	73.9
(1)+(5)+(9)+(10)+(11)	52.4	74.6
(1)+(5)+(9)+(10)+(12)	**55.9**	**77.0**
(1)+(5)+(9)+(10)+(11)+(12)	55.9	76.9

Table 5-7: Vowel stress classification accuracy (in percentage) of various combinations of features on the JUPITER development data. The combinations of features are described by feature indices as defined in Table 5-6 and this table.

Configuration	Sub.	Del.	Ins.	WER	UER
Baseline	4.3	1.6	1.7	7.6	20.2
+ Stress models	4.1	1.6	1.5	7.2	19.6

Table 5-8: Speech recognition error rates (in percentage) on the JUPITER development data. "WER" is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. "UER" is the utterance error rate.

models are used. Only syllable nucleus vowels are scored by the lexical stress model: for segments that do not carry lexical stress, such as consonants and silences, the stress scores are simply ignored (set to be zero). A transition weight is used with each applied stress score to avoid bias toward hypothesizing fewer stressed segments.

We found that the basic lexical stress model improved the *optimized* baseline speech recognition performance on the development data. In addition, the gain using only prosodic features was greater than when MFCC features were also used, even though the stress classification results implied otherwise. This is likely due to redundancy with the boundary models, in which MFCC features are already used. The baseline word error rate was reduced from 7.6% to 7.2%, a 5.3% relative reduction. The details are summarized in Table 5-8.

There also exist significant differences among the intrinsic prosodic values of different phones. For example, the duration of the vowel /*ih*/ (as in c*i*ty) is inherently shorter than

Configuration	Sub.	Del.	Ins.	WER	UER	Significance
Baseline	5.8	2.9	2.2	10.9	24.8	
+ Stress models	5.6	2.7	2.0	10.3	24.0	< 0.001

Table 5-9: Speech recognition error rates (in percentage) on the test data. "WER" is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. "UER" is the utterance or sentence error rate. The significance level between the baseline performance and the improved performance is also listed.

that of /*ey*/ (as in Mond*ay*), regardless of the stress properties. By grouping all vowels into a few stress categories, the intrinsic values contribute to large variances in the stress models, causing extensive overlap among the distributions. There are two approaches to improving the models: (1) normalizing the prosodic measurements by vowel intrinsic values, and (2) building separate models for different vowels. We experimented with the second approach, because there are plenty of training data in our corpus. One extreme is to build prosodic models for the complete inventory of vowels with different stress properties. However, the speech recognition performance with the new set of models (of significantly larger size) was virtually unchanged. We also tried less refined categories, by grouping vowels with similar intrinsic durations into classes. However, the changes to speech recognition results were also negligible.

Puzzled by these results, we performed an experiment in which all vowels were mapped into one class to form a single model for "vowel-like". The recognition performance was virtually the same as using the four-class models. This seems to suggest that the gain by using prosodic models is mainly achieved by eliminating implausible hypotheses, e.g., preventing vowel/non-vowel or vowel/non-phone confusions, rather than by distinguishing the fine differences among various stress and vowel classes.

We applied the basic prosodic model on the test data and obtained similar improvements. The detailed recognition results are summarized in Table 5-9. The significance level of the *matched pairs segment word error test* between the baseline performance and the improved performance is less than 0.001. This implies that the improvements by using prosodic models, although small, are statistically significant.

The recognition experiments described so far did not use any prosodic models to score

for segments other than vowels. This is due to the consideration that some prosodic features are not well-defined for all phones. For example, the durations of various silence models, such as those used for modeling the pauses before and after an utterance or between words, are somewhat arbitrary; the durations of fricatives and affricates are also likely to have large variances. Furthermore, F_0 is simply not defined for unvoiced sounds. We believe that it is more advantageous to score for only a *subset* of phones for which the prosodic measurements are "meaningful" and more informative. If we blindly build prosodic models for all phones, the scores for non-vowel segments are likely to be very noisy (due to noisy prosodic measurements), which will dilute the more informative scores for vowels. We conducted a recognition experiment to examine this hypothesis, in which prosodic models for all phones were trained and applied in the recognition search. We found that the recognition performance was worse than that in the baseline system. This suggests that the incorporation of noisy prosodic scores for non-vowel segments actually increased confusions during the recognition search.

5.5 Summary

In this chapter, we tested the approach of scoring the lexical stress patterns of recognizer hypotheses to improve automatic speech recognition performance. The motivation is that substitution, insertion and deletion errors sometimes result in mismatched stress characteristics between the hypothesized syllable nucleus and its acoustics; thus, the additional constraints provided by the stress models can improve speech recognition performance by reducing such errors.

Towards this goal, we first examined the correlation of various pitch, energy, and duration measurements with lexical stress on a large corpus of spontaneous utterances in the JUPITER domain. We found that the distributions of most prosodic features differed for different lexical stress classes; however, the extent of overlap among classes was also significant. We then performed classification experiments to identify the most informative features for lexical stress. The best single feature for stress classification was the integral of energy over the nucleus vowel, while the best set of prosodic features consists of the integral of energy,

raw duration, pitch slope, and the average probability of voicing. We observed that the best set of prosodic features were completely computable from information extracted from the segmental region alone. It is also convenient that F_0 difference performed better than F_0 average; thus, the sentence-level normalization is not required. Higher stress classification accuracy was achieved by using spectral features (MFCCs) in addition to the prosodic features. In the recognition experiments, however, it was found that the gain using only prosodic features was greater than when MFCC features were used.

We integrated the stress model into the recognizer first-pass Viterbi search. We found that using a simple four-class stress model achieved small but statistically significant gain over the state-of-the-art baseline performance on JUPITER. However, more refined models taking into account the intrinsic prosodic differences failed to improve the performance further. Our recognition results of a one-class (including all vowels) prosodic model suggest that the gain obtained by using prosodic models was mainly due to the elimination of implausible hypotheses, rather than by distinguishing different stress and segmental classes. This also explains the discrepancy between the recognition and stress classification experiments regarding the choice of optimal features, i.e., the improved recognition performance was not due to better classification of stress subcategories.

Although using one-class prosodic model performed as well as using more refined stress models in improving speech recognition performance in our experiments, the ability to reliably classify the underlying stress of vowels has many other potential uses in constraining a recognition system. It is known that the acoustic realization of a stressed vowel tends to be more distinctive than that of its unstressed counterpart, and consonants in pre-stressed position are also known to be more clearly enunciated (Lehiste 1970). In addition, phonological variations for stressed syllables are more limited compared to unstressed ones (Lea 1980). For example, phoneme /t/ is less likely to be flapped at pre-stressed position. We can build more accurate acoustic models for stressed vowels and pre-stressed consonants, and improve phonological rules by include stress as a factor in the conditional probability of alternative realizations. These more sophisticated acoustic models and phonological rules can be applied conditioned on the outcome of stress classification for the syllable nucleus vowels for improved speech recognition performance.

Previous experiments have found that adding phonetic segment models in addition to boundary models did not improve recognition performance, unless tri-phone models were used (Ström et al. 1999). However, in our experiments, the addition of very simple prosodic segment models was able to reduce the baseline word error rate by about 5.5%. This seems to suggest that prosodic features, which are not accessible to the boundary models, offer greater hope for independent information to improve recognition performance. We have also found that it is more advantageous to apply prosodic constraints selectively, i.e., only on phones for which the prosodic measurements are "meaningful" and more informative. We have developed a mechanism which is able to score for a subset of phones, and to incorporate these scores, normalized appropriately, into the first-pass recognition search.

Chapter 6

Recognition Confidence Scoring Enhanced with Prosodic Features

Recognition confidence scoring concerns the problem of evaluating the quality of the recognizer outputs, so that proper actions can be taken by a dialogue system depending on the reliability of the recognition hypotheses. Confidence scoring can naturally be formulated as a two-way classification or detection problem, i.e., whether a recognizer hypothesis is correct or not. The classification can be done on the utterance level, to determine whether to accept or reject a sentence hypothesis. It can also be done on the more refined word level, so that the dialogue system can be more specific about which words in the sentence hypothesis might be problematic. Instead of accepting or rejecting the whole utterance, the system can either request confirmation explicitly about the less reliable parts, or ignore those parts if a coherent meaning can be derived from the rest of the utterance hypothesis. A probabilistic implementation of such a framework has been developed for the JUPITER weather system by Hazen *et al.* (2000a, 2000b) at the Spoken Language Systems group.

In this chapter, we examine the approach of using prosodic cues to improve recognition confidence scoring on both the utterance level and the word level. Prosodic information can potentially assist the detection of speech recognition errors for several reasons. First, speech recognition performance depends on speaker characteristics and speaking style, or more generally, the amount of *similar* training data; and such properties are often correlated

with distinctive prosodic features. For example, the recognition performance for female and child speech is significantly worse than for male speech in the JUPITER system (Glass et al. 1999); and hyperarticulated speech is also likely to cause inferior speech recognition performance (Soltau and Waibel 1998). Female and child speakers typically have higher F_0 than male speakers; and hyperarticulated speech is characterized prosodically by slow speaking rate, increased pitch and loudness, etc. Second, there are prosodic cues to speech artifacts, which are also a significant source of recognition errors. For example, when a user calls JUPITER from a noisy environment, the background can also be picked up by the telephone and recorded with the speaker utterance. It is usually difficult for the recognizer to distinguish the background speech from the user speech. However, the energy of the background speech should be much weaker compared to that of the user utterance. Thus, the word confidence model can utilize the energy cues to improve accept/reject decisions in this case. Third, we expect that recognition errors sometimes result in mismatches between the hypotheses and the prosodic measurements. For example, insertion and deletion errors are likely to result in very short or long hypothesized phones, which are revealed by durational cues. In the previous chapter, our recognition experiments have suggested that prosodic models can decrease the likelihood of implausible hypotheses. We hope that "unusual" prosodic measurements are sometimes indicative of speech recognition errors, thus, prosodic cues can be used to improve the detection of erroneous word hypotheses in the confidence experiments.

In the following sections, we first introduce previous work done by Hirschberg and colleagues on using prosodic cues in utterance-level confidence scoring. Then we describe the basics of the confidence scoring framework used in our experiments. After that, we report the utterance-level and word-level confidence scoring experiments in detail. Finally, we conclude with a short discussion and summary.

6.1 Related Research

Using prosodic information to predict speech recognition performance has been explored by Hirschberg *et al.* (1999, 2000) on a couple of recognition systems and application domains.

The main motivation is that there exist prosodic cues to utterances that are typically not well modeled by a speech recognizer's training corpus, such as high pitch, too loud, too long, etc. In addition, a speaker often hyperarticulates when trying to correct system errors (Oviatt et al. 1996), and hyperarticulated speech is subsequently more likely to cause recognition errors by the system (Soltau and Waibel 1998). Thus, the prosodic characteristics of an utterance are likely to correlate with the performance of the speech recognition system on this utterance. Eight prosodic features were examined by the authors as potential cues to predict system errors in recognizing or understanding each user *turn* or utterance. These features include maximum and mean F_0 values, maximum and mean energy values, total duration, length of the pause preceding the turn, number of syllables per second in the turn (tempo), and percentage of silence within the turn. It has been found that there exist statistically significant differences in the *mean* values of a subset of these prosodic features between correctly recognized vs. misrecognized user turns, although the features in the subset vary for the two systems used in the studies. These prosodic cues were used by a *rule-based* classifier to perform accept/reject decisions on recognition outputs, in conjunction with other information such as acoustic confidence score, language model, recognized string, likelihood score, and system prompt. The results seem to suggest that the efficacy of prosodic features depends highly on the quality of the recognition system. In the system which used "older" recognition technology and "poorer performing" acoustic and language models, the prosodic features achieved a large improvement over using acoustic confidence alone (over 50% reduction in classification errors), and the best-performing rule set included prosodic features. However, in the system which was better trained for the recognition task, adding prosodic features improved only modestly over acoustic confidence features alone (less than 7% error reduction).

In our investigation, we first test if the approach of using prosodic cues in utterance-level confidence scoring can be generalized to the JUPITER system, which has been well-trained on a large corpus of speech data. We found that there are differences in both the means and the variances of some prosodic measurements between correctly and incorrectly recognized utterances, with the variances generally larger for misrecognized utterances. This is consistent with the intuition that "outliers" are more likely to be incorrectly recognized.

Since the confidence scoring module in our experiments is based on a Bayesian classification framework, we use an information theoretic measure (namely, the *mutual information* between a feature and the correct/incorrect labeling) to evaluate the effectiveness of prosodic features. We will also perform experiments comparing the recognition error detection performance using only features derived from the recognizer outputs, such as acoustic scores, with that obtained when the feature set is augmented with additional prosodic features. We also examine if prosodic features can be used to better distinguish correctly and incorrectly recognized *words*. Although the methodology is quite similar to that used in the utterance-level confidence scoring, the underlying assumptions are somewhat different, as described previously. Mutual information measures and word-level confidence experiments will be used to test this approach.

6.2 Experimental Background

In this section, we provide some background knowledge for the experiments described in this chapter, including the basics of the confidence scoring module, the speech data, and the labeling of the data. The details of the confidence scoring framework can be found in (Kamppari and Hazen 2000; Hazen et al. 2000a; Hazen et al. 2000b).

6.2.1 Experimental Framework

The confidence scoring module developed by Hazen *et al.* is based on a Bayesian formulation. For each recognition hypothesis, a set of confidence measures are computed to form a confidence feature vector. The feature vector is reduced to a single dimension using a simple linear discrimination projection. Distributions of this raw confidence score for correct and incorrect hypotheses are obtained from the training data. A probabilistic confidence score is then obtained using maximum *a posteriori* probability (MAP) classification, with the raw confidence score as the input. The threshold of the MAP log likelihood ratio can be varied to set the operating point of the system to a desired location on the *receiver-operator characteristic* (ROC) curve, to balance between high detection rate and low false alarm rate. In our experiments, the minimum classification error rate is chosen as the desired

operating point for both the utterance-level and word-level recognition error detection.

Hazen's confidence models used 15 features for detecting utterance-level recognition errors, and 10 features for detecting word-level recognition errors. These features try to measure whether the input speech is a good fit to the underlying models used by the system, as well as whether there are many competing hypotheses that have similar scores. For example, among the 15 utterance features, the total score (i.e., the sum of acoustic, language model, and pronunciation model scores) for the top sentence hypothesis is used to measure the match between the hypothesis and the models; while the drop in total score between the top hypothesis and the second hypothesis in the N-best list tries to measure the "distance" between competing hypotheses. The complete inventory of the 25 utterance and word features is listed in Appendix A for reference. These features will be referred to as ASR features in the rest of this chapter, because they come from the automatic speech recognition (ASR) system. The ASR features are used to train baseline utterance and word confidence models, to be compared with confidence models using additional prosodic cues.

The ultimate evaluation criterion for the confidence scoring module is the understanding performance. However, since we do not address the problem of incorporating the confidence scores into a dialogue system, we will only evaluate the classification performance. In particular, the *figure of merit* (FOM), which is the area under the ROC curve, and the minimum classification error rate are used for evaluation.

6.2.2 Data and Labeling

We have provided a description of various data sets in the JUPITER corpus in Table 5-2. The confidence models should be trained using a set of data that is independent of the data used for training the recognizer, because the models are used to evaluate the recognizer's performances on *unseen* data. In our experiments, the utterance and word confidence models are trained on the "all_test_1" data set. In addition, the "all_test_3" set is used as held-out data for feature selection, and the "all_test_2" set is used for testing.

The data must be labeled for training and testing the confidence models. Each utterance is first passed through a recognizer to generate a 10-best list, because some utterance-level ASR features depend on the information in the N-best outputs. The labeling of

Set	Utterance			Word		
	# Cor.	# Incor.	ER	# Cor.	# Incor.	ER
Training	1,602	866	35.1%	9,907	2,017	16.9%
Development	1,050	811	43.6%	6,010	1,629	21.3%
Test	1,415	920	39.4%	9,013	1,795	16.6%

Table 6-1: Number of correct (cor.) and incorrect (incor.) hypotheses on the training, development, and test data for the confidence experiments. The error rate (ER) is the number of incorrect hypotheses divided by the total number of hypotheses. The error rate for the utterance hypotheses corresponds to the sentence error rate. The error rate for the word hypotheses is different from the recognition word error rate, because the deletion errors are not counted.

the utterance hypotheses is somewhat arbitrary and depends on the purpose. In Hazen's experiments, the goal of the utterance confidence scoring is to reject *very poorly* recognized utterances, because the word-level confidence scoring can deal with hypotheses with partial errors. Thus, fairly complex criteria are used to label the utterance hypotheses. We simply mark an utterance as incorrectly recognized if there are any errors in the best sentence hypothesis. This follows the example in Hirschberg's experiments so that the results of the utterance-level confidence experiments can be compared. The labeling for the word-level hypotheses is very straightforward. Correctly recognized words in the hypothesis are labeled as correct, and incorrectly recognized words are labeled as incorrect. Only words in the top sentence hypothesis are used for training. Table 6-1 summarizes the number of correct and incorrect utterance and word hypotheses on the training, development, and test data sets. Notice that word deletion errors do not contribute to incorrect word hypotheses. The overall error rates on these data sets are much higher than on clean in-vocabulary speech, due to the inclusion of utterances with out-of-vocabulary words and artifacts.

6.3 Utterance-Level Confidence Scoring

6.3.1 Utterance-Level Prosodic Features

We have examined twelve utterance-level prosodic features as potential candidates for predicting speech recognition errors. Three features are related to F_0, two features are related

to energy, and the remaining seven features capture various kinds of timing information of a user utterance. The details of these features are listed as follows:

- **utterance_mean_$\mathbf{F_0}$**: the average F_0 of all vowels in an utterance.
- **utterance_max_$\mathbf{F_0}$**: the maximum F_0 of all vowels in an utterance.
- **utterance_mean_pv**: the average probability of voicing of all vowels in an utterance.
- **utterance_mean_energy**: the mean RMS energy of all vowels an utterance.
- **utterance_max_energy**: the maximum RMS energy of all vowels an utterance.
- **pause1_duration**: the duration of silence before the utterance in a recording. This is to indicate if an utterance is truncated at the beginning because the speaker started speaking before the recording started.
- **pause2_duration**: the duration of silence after the utterance in a recording. This is to indicate if an utterance is truncated at the end due to end-point detection errors.
- **utterance_duration**: the duration of an utterance (excluding anterior and posterior pauses).
- **utterance_percent_silence**: the percentage of silence (as indicated by sum of inter-word pause durations) within an utterance.
- **utterance_speaking_rate**: the utterance speaking rate as defined in Equation 5.1, which is computed as the sum of expected vowel durations divided by the sum of measured vowel durations in an utterance.
- **utterance_num_syllables**: the number of syllables in an utterance.
- **utterance_tempo**: the number of syllables in an utterance divided by the utterance duration.

6.3.2 Feature Selection

We are not sure which of the utterance-level features described in the previous section contribute substantial information to the detection of utterance hypothesis errors, and how they compare to the ASR features. In (Hirschberg et al. 1999), a T-test was used to determine if the means of prosodic features differed statistically for correctly and incorrectly recognized user turns. Our preliminary analysis revealed that there were differences in both the means and the variances of the prosodic measurements between the two classes, with the variances generally larger for misrecognized utterances. Given that the confidence scoring module uses a probabilistic framework, we believe that the mutual information measure will be a good indication of the effectiveness of each confidence feature. In the following, we first give a brief introduction of the concept of mutual information. We then compute the mutual information between each utterance feature and the utterance "correctness" labels. Finally we describe the feature selection procedure based on the mutual information ranking of the features.

Mutual Information

For two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$, the mutual information $I(X;Y)$ is defined as:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \; log \frac{p(x,y)}{p(x) \; p(y)} \tag{6.1}$$

It can be proven that

$$I(X;Y) = H(X) - H(X|Y) \tag{6.2}$$

where $H(X)$ is the entropy of the random variable X, and $H(X|Y)$ is the conditional entropy of X given Y. Thus, the mutual information can be interpreted as the relative reduction of uncertainty of X due to the knowledge of Y, and vise versa. The mutual information will have a value of 0 if X and Y are independent of each other.

Utterance-Level Feature Ranking

Assume that X is the correct/incorrect labeling of a hypothesis, and Y is an utterance feature. To compute the mutual information between X and Y, we need to obtain the joint and marginal distributions of X and Y.

Although most of the utterance confidence features used in our study are continuously distributed, we can use the histograms to approximate the probability mass functions. The histogram distributions of a feature for correct and incorrect utterance hypotheses, i.e., $p(y|x = correct)$ and $p(y|x = incorrect)$, can easily be obtained from the training data. The joint distribution can then be computed from the conditional distributions as follows:

$$p(x, y) = p(y|x = correct)p(x = correct) + p(y|x = incorrect)p(x = incorrect) \qquad (6.3)$$

The *priors* of X, $p(x)$, and the histogram distribution of the feature Y, $p(y)$, can be directly obtained from the training data. Thus, we can compute the mutual information according to Equation 6.1.

Table 6-2 summarizes the mutual information between each utterance feature and the utterance correctness label. The features in the table are ordered by the mutual information measure, and the prosodic features are indicated in **bold** fonts. As shown in the table, the features with high mutual information are mostly from the ASR system. This is not surprising, because the ASR features are directly linked to the performance of a recognition system. Nevertheless, some prosodic features also provide significant information about the labels. In particular, the percentage of silence within an utterance, average and maximum F_0 values, utterance duration and tempo are among the "best" prosodic features. Figure 6-1 shows the histogram distributions of the percentage of silence feature and the mean F_0 feature for the correct and incorrect classes. It seems that utterances with a high percentage of internal silence are more likely to be incorrectly recognized. The internal pauses are usually associated with hesitation, emphasis, or hyperarticulation, which are not typical in the JUPITER training data. Utterances with high mean F_0 are also more likely to be incorrectly recognized. This is consistent with the recognition results that female

Feature	**Mutual Information**
total_drop	.409
average_purity	.296
frac_high_purity	.247
acoustic_score_per_bound	.243
total_score_per_word	.195
lexical_score_per_word	.177
nbest_frac_high_purity	.171
lexical_drop	.161
nbest_average_purity	.157
total_score	.145
utterance_percentage_silence	**.144**
acoustic_score	.117
lexical_score	.114
utterance_mean_F_0	**.099**
utterance_max_F_0	**.092**
utterance_total_duration	**.092**
utterance_tempo	**.090**
utterance_max_energy	**.082**
utterance_mean_pv	**.075**
acoustic_drop	.075
pause1_duration	**.070**
utterance_speaking_rate	**.068**
num_words	.056
utterance_mean_energy	**.054**
pause2_duration	**.049**
utterance_num_syllables	**.043**
nbest	.006

Table 6-2: Ranking of utterance-level confidence features by mutual information.

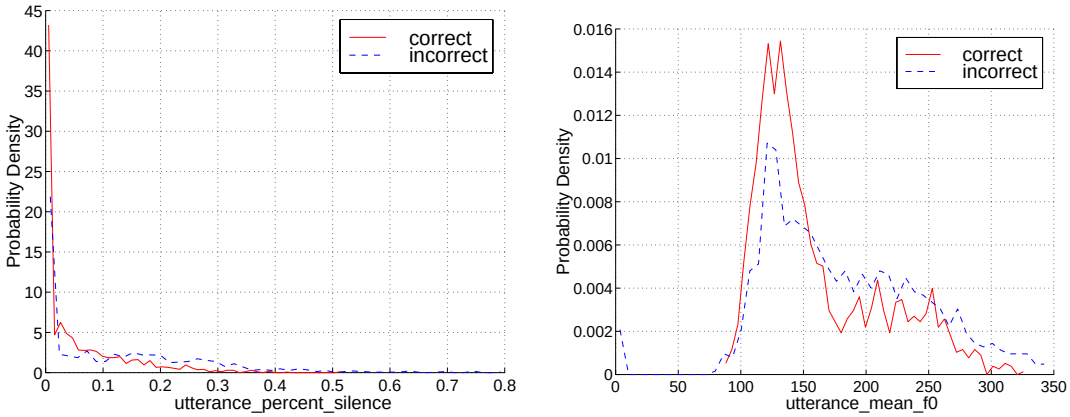

Figure 6-1: Histogram distributions of two example prosodic features for the correctly and incorrectly recognized utterances.

and child speech have considerably higher error rates. There are also a few incorrectly recognized utterances with measured mean F_0 of 0. This is because there are no vowels in the hypothesis, which is a clear indication that the hypothesis is likely to be wrong.

Feature Selection Procedure

It is desirable to obtain the performances of all feature combinations on development data to find the optimal feature set. However, the computation requirement is tremendous, because N features can lead to $2^N - 1$ feature combinations. Instead, we pick only features that provide substantial information about the labels, i.e., those with high *mutual information* with the correct/incorrect labeling of the utterance hypotheses. Specifically, we rank-order the utterance features according to the mutual information measure, and add each feature incrementally into the feature set. In this way, at most N feature combinations need to be tested. The utterance confidence model using the subset of features is trained on the training data, and tested on the development data. The feature set that yields the best performance on the development data is chosen as the final feature set. The underlying assumption of this process is that, if a feature can not improve the performance, then any feature with lower mutual information can not improve the performance either. This is only an approximation, because the features are generally not independent of one another.

System	FOM	MER (%)	Significance Level
Baseline	.900	16.9	
+ Prosodic Features	.912	15.6	*.018*

Table 6-3: Figure of merit (FOM) and minimum classification error rate (MER) for the utterance-level confidence scoring with only ASR features and with ASR and prosodic features combined. The McNemar significance level between the two classification results is also listed.

6.3.3 Experimental Results

We obtained the performance of utterance-level accept/rejection decisions with only ASR features and with ASR and prosodic features combined. The features are selected using the procedure described in the previous section. In the experiment which used only ASR features, we found that all 15 ASR features improved the performance on the development data when added incrementally into the feature set. In the experiment which used both ASR and prosodic features, we found that the top 25 features (out of 27 total features) in Table 6-2 improved the performance on the development data when added incrementally.

Figure 6-2 plots the ROC curves of the utterance-level classification experiments on the test data. As shown in the figure, the addition of prosodic features pushed the ROC curve towards the upper-left corner slightly. This means that the correct acceptance rate is improved if the false acceptance rate is kept the same, or, the false acceptance rate is reduced if the correct acceptance rate is maintained. The figure of merit and the minimum classification error rate are summarized in Table 6-3 for the two system configurations. The McNemar significance level between the two classification results is 0.018, thus, the improvement is statistically significant given a 0.05 threshold.

6.4 Word-Level Confidence Scoring

6.4.1 Word-Level Prosodic Features

We have examined 9 word-level prosodic features as potential candidates for predicting word-level speech recognition errors. Three features are related to F_0, two features are

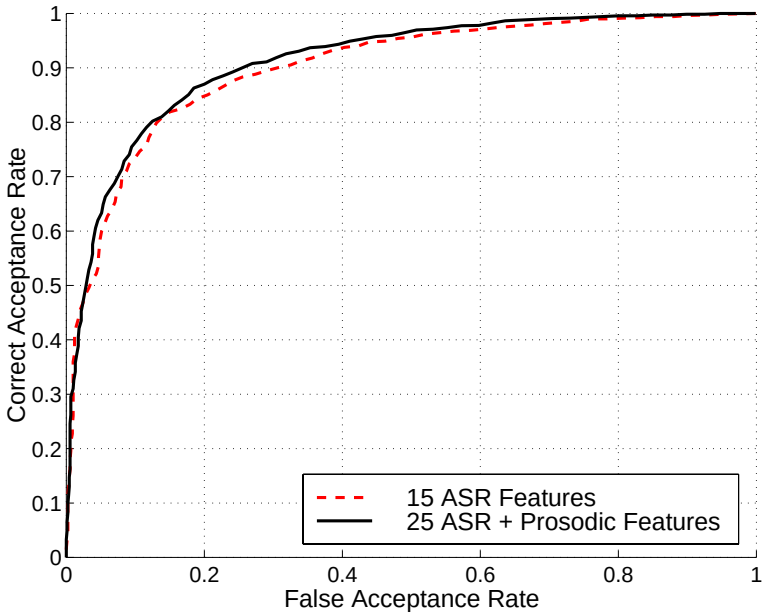

Figure 6-2: ROC curves of utterance-level speech recognition error detection using only ASR features and using both ASR and prosodic features.

related to energy, and the remaining four features capture various timing information of a hypothesized word.

- **word_mean_$\mathbf{F_0}$**: the average F_0 of all the vowels in a word.
- **word_max_$\mathbf{F_0}$**: the maximum F_0 of all the vowels in a word.
- **word_average_pv**: the average probability of voicing of all the vowels in a word.
- **word_mean_energy**: the difference between the utterance maximum energy and the mean RMS energy of all vowels in a word.
- **word_max_energy**: the difference between the utterance maximum energy and the maximum RMS energy of all vowels in a word.
- **word_speaking_rate**: the word speaking rate, computed as the sum of the expected vowel durations in a word divided by the sum of the measured vowel durations in the word.

- **word_num_syllables**: the number of syllables in a word.
- **word_duration**: the duration of a word.
- **after_word_pause_duration**: duration of pause after a word.

6.4.2 Word-Level Feature Ranking

Table 6-4 summarizes the mutual information between each word feature and the word correctness label. The features in the table are ordered by the mutual information, and the prosodic features are indicated in **bold** fonts. The word energy features, which have been normalized by the maximum utterance energy, are among the "best" prosodic features. This is possibly because they are good indications of background speech, as discussed previously. Similar to the utterance features, the top word features are all ASR features. However, prosodic features compare favorably to some ASR features; and more importantly, they provide independent information, and hence, are more likely to bring additional gain.

6.4.3 Experimental Results

We obtained the performance of word hypothesis error detection with only ASR features and with ASR and prosodic features combined. The features are selected using the same procedure as for utterance features. In the experiment which used only ASR features, we found that all 10 ASR features improved the detection performance on the development data when added to the feature set. In the experiment which used both ASR and prosodic features, only the top 13 features in Table 6-4 improved detection performance on the development data when added incrementally to the feature set.

Figure 6-3 plots the ROC curves of word-level classification experiments on the test data. As shown in the figure, the addition of prosodic features also pushed the ROC curve towards the upper-left corner slightly. Table 6-5 summarizes the FOM and MER for the two system configurations. The McNemar significance level between the two classification results is 0.0005, which implies that the difference is statistically significant.

Feature	**Mutual Information**
utt_score	.264
frac_nbest	.164
bound_diff_from_max_mean	.161
bound_score_mean	.152
bound_score_min	.113
word_max_energy	**.043**
word_mean_energy	**.043**
bound_norm_score_mean	.043
word_speaking_rate	**.040**
word_mean_F_0	**.040**
word_mean_pv	**.038**
word_max_F_0	**.033**
after_word_pause_duration	**.032**
bound_score_std_dev	.031
word_duration	**.025**
num_bounds	.012
word_num_syllables	**.006**
num_nbest	.001
bound_likelihood_mean	.0003

Table 6-4: Ranking of word-level confidence features by mutual information.

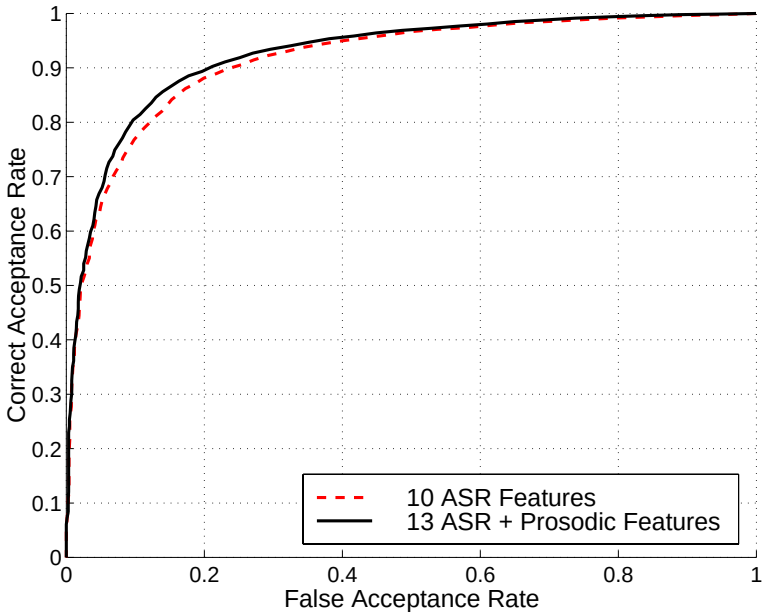

Figure 6-3: ROC curves of word-level speech recognition error detection using only ASR features and using both ASR and prosodic features.

6.5 Summary

In this chapter, we explored the use of prosodic cues in recognition confidence scoring. Our utterance-level confidence experiment results are generally consistent with previous work by Hirschberg and colleagues. We observed that prosodic cues achieved small but statistically significant improvement in the detection of utterance hypothesis errors. There are two potential explanations for the relatively smaller gain. First, the user behavior in the two cases seems to differ. The data used in the experiments in (Hirschberg et al. 1999) contain a high percentage of hyperarticulated speech due to high recognition error rates in the dialogue system used for data collection. These hyperarticulated user turns are likely to be misrecognized; thus, prosodic cues are effective for detecting such problems due to the distinctive prosodic characteristics of hyperarticulation. On the other hand, the JUPITER recognizer has been trained on large amounts of speech data and performs relatively well on in-domain queries. Most user queries are spoken with "normal" prosody; thus, prosodic cues are less indicative of the recognizer performance. However, prosodic cues can help

System	FOM	MER (%)	Significance Level
Baseline	.913	10.9	
+ Prosodic Features	.925	10.2	*0.0005*

Table 6-5: Figure of merit (FOM) and minimum classification error rate (MER) for the word-level confidence scoring with only ASR features and with ASR and prosodic features combined. The McNemar significance level between the two classification results is also listed.

identify female or child speech, which usually have inferior performance to male speech under the JUPITER recognizer. Second, the confidence classifier model used in our experiments is relatively simple. In particular, the classifier is only able to deal with numerical features; in addition, all features are reduced to a one-dimensional feature by a linear discrimination projection, which implies a linear decision boundary in the feature space. We expect the prosodic cues to be more effective if the classifier can pay more attention to "outliers", because many prosodic features have a "wider" distribution for incorrectly recognized utterances than for correctly recognized utterances. The same analysis also applies to the word-level experimental results.

The best utterance-level prosodic feature (as indicated by the mutual information measure) is the percentage of silence in the utterance; and utterances with a high percentage of silence are more likely to be incorrectly recognized. This is in direct contrast to results obtained in (Hirschberg et al. 2000), where the *correctly* recognized turns have a significantly higher percentage of internal silence. This seems to suggest that the prosodic characteristics of correctly and incorrectly recognized utterances could have different characteristics in different systems. However, we believe that the principle of looking for prosodic cues to recognition errors is generally applicable.

The framework for word-level confidence scoring can be generalized to perform phrase boundary detection. We will explore this possibility in detail in Section 8.2 when discussing future work.

Chapter 7

Characterization of English Intonation Contours

Research on using intonation in the linguistic analysis of spoken utterances has been sparse. Among the few inquiries reported in the literature, most methods rely on an intermediate prosodic transcription to serve as a bridge between the acoustic realization of the intonation and the syntactic/semantic structure of the utterance (Ostendorf et al. 1993; Kompe et al. 1997). These methods need to address several difficult issues. First, prosodic transcription, e.g., using ToBI convention for English (Silverman et al. 1992), is a challenging and time-consuming task, which makes it impractical to transcribe large speech corpora manually. Second, automatic recognition of intonational events (especially pitch accents, phrase tones, etc.) from the acoustic signal is difficult and error-prone (Ostendorf and Ross 1995). This also hampers the development of reliable methods to perform automatic prosodic transcription. Third, the mapping between prosodic events and the syntax/semantics of an utterance is still very poorly understood, except for a general correspondence between prosodic phrase boundaries and syntactic boundaries. For this reason, most studies have focused on using prosodic phrase boundary locations to resolve syntactic ambiguities (Ostendorf et al. 1993; Hunt 1995) or to improve parsing efficiency (Kompe et al. 1997). Although there have been efforts towards automatically describing and classifying intonation contours (Grigoriu et al. 1994; Jensen et al. 1994; Ostendorf and Ross 1995; ten Bosch 1995), their use in

linguistic analysis or speech recognition has been limited, largely due to the missing link with linguistic identities.

We are interested in developing a framework to model the intonational aspects of certain syntactic/semantic structures in English utterances, without using intermediate prosodic transcriptions. There are two important characteristics in our proposed approach. First, we want to build acoustic models directly for certain linguistic components in an utterance, without using prosodic labels as an intermediate layer. In this way, we can avoid the labor-intensive prosodic labeling process as well as the necessity of predicting prosodic labels from linguistic analyses. We can use data-driven methods to derive distinct F_0 patterns/categories for the linguistic components in our modeling framework, which can be regarded as analogous to prosodic labels. Second, we only model part of the prosodic space of an utterance, in particular, phrases that bear important communicative functions. We believe that such an approach is more robust than trying to characterize the intonation of an entire utterance, especially for spontaneous speech. This is based on the observation that spontaneous speech consists of "islands of acoustic reliability" surrounded by casually enunciated "fillers". The "fillers" are likely to contribute "noise" if they are included in the modeling framework. We envision that the phrase models can potentially be applied to score the intonation patterns of recognizer hypotheses, which can in turn be used to resort the N-best outputs for improved recognition accuracy or to support the rejection of erroneous hypotheses.

In this chapter, we examine the feasibility of such a framework by performing a pilot study on characterizing the pitch contours of some selected English phrases in the JUPITER domain. As a starting point, we select five common types of phrases in the JUPITER corpus, such as "what is", "tell me", city names, etc., to carry out our study. These phrases also carry important information in the JUPITER domain, so that they are likely to have a significant impact on the system performance. We will develop acoustic models to characterize the intonation contours of these phrases, and seek to answer the following questions in our experiments:

- Can we identify phrases based on F_0 contour alone?

- Does phrase F_0 pattern generalize across similar but not identical utterances?
- Does each phrase have some set of canonical patterns?
- Are there interdependencies among phrases in an utterance?
- Will this information be useful to speech recognition or understanding?

In the following sections, we first give an overview of our experimental methodology, including selection of data for training and testing, acoustic characterization of the selected phrases, and the objectives of the experiments. After that, we present and discuss the results of each experiment in detail. Aspects related to prosodic theories and potential applications of this framework in speech recognition and understanding will also be addressed in the discussions. Finally, we conclude with a brief summary.

7.1 Experimental Methodology

7.1.1 Data Selection

One of the key issues in intonation modeling is to find an inventory of model units. In our experimental framework, we want to explore the feasibility of directly modeling certain linguistic structures in English utterances. Thus, we decide to begin with a number of common phrases in JUPITER utterances in our initial investigation. In this way, the unit set covers some "typical" basic linguistic patterns in the JUPITER utterances, and there will be sufficient data for acoustic model training.

Five classes of two-syllable words/phrases are selected from the JUPITER corpus, including "<what_is>", "<tell_me>", "<weather>", "<SU>", and "<US>". Each "phrase" class consists of a list of words/phrases with the same stress pattern, which have also been chosen to have similar semantic properties, so that they are likely to serve similar syntactic functions. In particular, each phrase class consists of words that can be substituted into the following sentence template to produce a well-formed sentence:

<what_is> | <tell_me> the <weather> in|for|on <SU> | <US>

<**what_is**>:	what is, how is, ...
<**tell_me**>:	tell me, give me, show me, ...
<**weather**>:	weather, forecast, dew point, wind speed, ...
<**SU**>:	Boston, Paris, Monday, ...
<**US**>:	Japan, Detroit, tonight, ...

Table 7-1: Five common phrase classes and examples for each class in the JUPITER weather domain.

For example, the "<weather>" class contains words or compound words like "weather", "forecast", "wind speed", "dew point", etc., all of which have "stressed unstressed" stress pattern and refer to some kind of weather information (general or specific); the "<US>" class consists of "unstressed(**U**) stressed(**S**)" two-syllable words for place names or dates; while the "<SU>" class consists of "stressed(**S**) unstressed(**U**)" two-syllable words for place names or dates, etc. Example words/phrases in each class are listed in Table 7-1. The complete listing of distinct entries in each phrase class is given in Appendix B.

Utterances that match exactly the above sentence template in the JUPITER corpus are chosen to form a test set. We will conduct experiments to classify the intonation contours of the five phrase classes on this set, and to study the correlation of the intonation contour patterns among the phrases in these utterances. However, we must train acoustic models for the phrase contours using different data. To ensure similarity between the training and test data for the five phrases, an instance of a phrase is used for training only if it occurs at particular positions in an utterance. Specifically, the "<what_is>" and "<tell_me>" phrases are constrained to be from the beginning of an utterance; the "<weather>" phrase is limited to be from an intermediate position in an utterance; and the "<SU>" or "<US>" phrases are selected only from the end of an utterance. Thus, the training set consists of utterances which contain the five phrases at positions described above, excluding those that match exactly the test sentence template. In this way, we can ensure the independence of training and test data and examine if the phrase F_0 patterns can generalize across similar but not identical utterances. The data selection criteria are summarized and illustrated by some example training and test utterances in Table 7-2.

Test	**<what_is>** | **<tell_me> the <weather> in|for|on <SU>** | **<US>** *What is* the *weather* in *Detroit*? *Give me* the *wind speed* for *Friday*.
Training	**<what_is>** | **<tell_me> ...** *What is* the humidity in Honolulu Hawaii? *Give me* the *weather* for Chicago for *tomorrow*.
	... <weather> ... Yes, I would like to know the *weather* in *New York*. Can you tell me the *sun rise* for anchorage?
	... <SU> | **<US>** *Tell me* the *wind speed* for concord new Hampshire *today*. And what is the time in *Frankfurt*?

Table 7-2: Criteria for selecting training and test utterances ("|" means "or", and "..." means "any words"). Test utterances are selected to match the "test" template. Training utterances are selected to match any of the three "train" templates but not the "test" template. Two example utterances for each template are shown below the template, with the *effective phrases* for training and testing highlighted in *italics*.

7.1.2 Feature Extraction

The principal acoustic correlates of intonation include all three prosodic features: fundamental frequency, energy, and duration (Ladd 1996). However, F_0 is the feature most closely related to intonation. Hence, to limit the scope of our initial investigation, we use only F_0-based measurements to characterize the phrase intonation pattern. Specifically, we describe the F_0 contour of a phrase using its constituent syllable F_0 contours, with each syllable F_0 contour characterized by the F_0 average and slope. Thus, for the two-syllable phrases chosen for our study, each token will be represented by a four-dimensional vector, consisting of the F_0 averages and slopes of the two syllables. In general, a phrase of N syllables can be characterized by a $2N$-dimensional vector using the same paradigm. The phrases forming the model inventory do not have to be of equal number of syllables, because the phrase models are intended to be applied in a post-processing framework. We have chosen only two-syllable phrases in our study, mainly to obtain classification performance to evaluate the feasibility of our proposed approach.

Partial Syllabification

We only have phonetic and word alignments for the utterances in the JUPITER domain, which were obtained automatically by running the JUPITER recognizer in forced recognition mode. To extract F_0 features from a syllable, or more precisely, the *sonorant region*[1] of a syllable, we need to determine the boundaries of the sonorant regions from phonetic and word transcriptions. This can be formulated as the problem of deciding the association of sonorant consonants between two adjacent syllable nuclei (vowels or syllabic consonants such as "*el*", "*em*"and "*en*"). The task is somewhat easier than completely syllabifying a phone sequence, because the non-sonorant phones do not need to be processed, and in fact, provide clear separating boundaries. We designed a multi-pass procedure to perform the task, which looks for cues (e.g., fricatives, stops, word boundaries, etc.) that can unambiguously divide sonorant phones between two adjacent syllable nuclei, and uses heuristics to handle ambiguities. Although this procedure is currently implemented to process phonetic and word transcription files, the basic algorithm can also easily be implemented within a recognizer to process N-best outputs or a phone graph.

The following rules (in the order listed) are used by the algorithm to decide the attachment of sonorant consonants between two adjacent syllable nuclei:

1. If there are one or more non-sonorant phones (i.e., stops, fricatives, silences, etc.) between two syllable nuclei, then the sonorants before the first non-sonorant phone are attached to the first syllable nucleus, and the sonorants after the last non-sonorant phone are attached to the second syllable nucleus.

2. If there is an /*ng*/ or post-vocalic /*l*/ or /*r*/[2] between two syllable nuclei, then the sonorant boundary is at the end of this phone.

3. If there is only one sonorant consonant separating two vowels, then the sonorant consonant is split into two halves, with each half merged with the adjacent vowel.

[1]The sonorant region of a syllable consists of the nucleus vowel and any preceding and following sonorant consonants (nasals and semivowels).

[2]Post-vocalic /*l*/ and /*r*/ are distinguished by specific labels in the JUPITER recognizer.

4. If there is a word boundary between two adjacent vowels, then the sonorant boundary is at the word boundary.

5. For any remaining un-processed two consecutive sonorant consonants, the /*m y*/ and /*n y*/ phones are merged with the second syllable nucleus, others are separately attached to the adjacent vowels.

Rule 3 is very important for dealing with ambiguities in determining syllable boundaries. We believe that splitting an inter-vocalic sonorant consonant is a more robust solution than trying to associate it with one of the syllable nuclei, because segmentation involving a "vowel nasal/semivowel vowel" sequence is already error-prone in the automatically derived phonetic transcriptions. Thus, we prefer this rule to rule 4, which relies on word boundary information to infer syllable boundaries. However, this will not affect the case where two words are separated by a short pause, because rule 1 should have been applied to make the correct association.

Rule 5, although simplistic, is an adequate solution for our data. After the first four rules are applied, the remaining un-processed sonorant sequences are those with two sonorant consonants between two syllable nuclei *within a word* (we have not observed three or more consecutive sonorant consonants within a word). There is no English syllable with two consecutive nasals, which implies that there must be a syllable boundary between two nasals. Syllables containing two consecutive semivowels are rare in English, with the exception of the /*r l*/ sequence in a few words such as "snarl". Hence, it is safe to assume that there is a syllable boundary between two semivowels in our data. Syllables with a "semivowel nasal" sequence seem to only exist in mono-syllabic words or word-final syllables (e.g., in words like "warm", "snowstorm", etc.), which should have been taken care of by rule 4. Thus, we can assume that there is a syllable boundary between a semivowel and a nasal after rule 4 is applied (e.g., in words like "warmer", "morning", etc.). For "nasal semivowel" combinations, we have observed that /*m y*/ and /*n y*/ seem to always occur within one syllable (e.g., in "accu*mu*lation", "Califor*nia*", etc.), while the others seem to be separated by a syllable boundary (e.g., in "someone", "only", etc.). Thus, after merging /*m y*/ and /*n y*/ with the following vowel, we can assume that there is a syllable boundary between

any remaining two sonorant consonants.

7.1.3 Experiments and Objectives

Three experiments are conducted to address the questions proposed at the beginning of this chapter. We only give a brief description of the experiments here. The details are covered in the next section.

Acoustic models for the five phrase classes are trained on the training data, and phrases in both the training and the test utterances are classified using these models. The classification accuracy reflects the ability to identify these phrases based on F_0 information only, while the classification performance differences between the training and test sets are indicative of how well the phrase F_0 patterns generalize from the training utterances to the test utterances.

Intonation theories have described the intonation contour as a sequence of discrete events with distinctive categories (Ladd 1996). We are interested in knowing if we can identify canonical F_0 contour patterns, which are analogues to distinctive categories in acoustic representation, in a "data-driven" manner. We are also interested in knowing if the acoustic realizations of various phrases within the same utterance are dependent on one another. For example, is the F_0 pattern of the "<what_is>" phrase (at the beginning of an utterance) correlated with that of the "<SU>" phrase (at the end of an utterance); or are these two phrases realized independently of each other, due to the "distance" between them in the utterance?

Data clustering is performed on the phrase F_0 contours in the training utterances, to identify typical F_0 contour patterns for each of the five phrase classes. Initial clusters are first obtained by unsupervised clustering. These clusters are then filtered by a "self-selection" process, to ensure that the clusters are indeed "typical" patterns for the data. We then use a *mutual information* measure to quantify the correlation of various F_0 contour patterns among the phrases within an utterance. Each phrase in the test data set is classified into one of the F_0 patterns obtained in the clustering experiment. Mutual information is computed for each pair of phrase F_0 patterns to determine if there exist interdependencies among phrases in the utterance.

7.2 Experiments and Discussions

7.2.1 Phrase Classification

We first perform classification experiments to examine how well phrases can be identified by their F_0 contours. As described in the previous section, each phrase token is represented by a four-dimensional vector, consisting of the F_0 averages and slopes of the two constituent syllables. A principal component analysis is first applied on the collection of training vectors to "whiten" the observation space. Mixtures of diagonal Gaussian models are then trained to characterize the distributions of the rotated feature vectors for the five phrase classes. Maximum likelihood (ML) classification is used, because our purpose is to evaluate the ability to identify phrases based on F_0 information alone, without the assistance/interference of *priors*. The F_0 contour of each utterance has been normalized by its mean value, to reduce variances due to speaker pitch differences.

To examine how well the phrase models generalize from training data to test data, we applied the phrase models to classify the phrases in both the training and the test utterances. The five-class classification accuracy is 60.4% on the training data, and 56.4% on the test data. The detailed classification confusions among phrases in the training data and in the test data are summarized in Table 7-3 and Table 7-4, respectively. The performance on the *unseen* test data is only slightly worse than that on the training data, and the confusion matrices for both sets clearly show that the results are significantly better than chance. We are inclined to conclude that there exists information in the F_0 contours of the five phrases that can be used to distinguish these phrases.

As shown in the tables, the confusions are high between phrases at the same utterance positions, and significantly lower between phrases at opposite positions (i.e., the beginning vs. the end of an utterance). This is possibly due to the general declination of F_0 contours, which causes the F_0 levels of the utterance-initial phrases to be higher than those of the utterance-final phrases. Thus, the "position" differences among phrases might have helped the classification performance. However, we believe that the ability to distinguish the phrases is not entirely due to F_0 level differences caused by F_0 declination, as indicated by the confusions between phrases at the same utterance positions. To further verify that, we

	<what_is>	<tell_me>	<weather>	<SU>	<US>	# Tokens
<what_is>	**55.87%**	23.74%	13.43%	5.80%	1.16%	6193
<tell_me>	11.29%	**68.11%**	16.40%	2.49%	1.71%	762
<weather>	6.94%	10.45%	**68.74%**	7.20%	6.67%	3013
<SU>	6.46%	3.90%	10.82%	**60.29%**	18.53%	6130
<US>	2.87%	4.05%	11.33%	24.32%	**57.43%**	592

Table 7-3: Classification confusions among phrases in the training utterances. The reference labels are shown in the first column, the hypothesized labels for the phrases are shown in the first row, and the number of tokens for each phrase class is summarized in the last column.

	<what_is>	<tell_me>	<weather>	<SU>	<US>	# Tokens
<what_is>	**45.82%**	19.22%	25.63%	7.94%	1.39%	718
<tell_me>	16.88%	**53.25%**	15.58%	9.09%	5.20%	77
<weather>	4.91%	0.63%	**68.18%**	12.45%	13.83%	795
<SU>	4.46%	3.42%	15.33%	**52.53%**	24.26%	672
<US>	3.25%	1.63%	13.01%	16.26%	**65.85%**	123

Table 7-4: Classification confusions among phrases in the test utterances. The reference labels are shown in the first column, the hypothesized labels for the phrases are shown in the first row, and the number of tokens for each phrase class is summarized in the last column.

performed two-class classification experiments for phrases at the same utterance positions, i.e., "<what_is>" vs. "<tell_me>", and "<SU>" vs. "<US>". The classification accuracy on the test utterances is 68.9% for "<what_is>" vs. "<tell_me>" (at utterance start), and 73.2% for "<SU>" vs. "<US>" (at utterance end), both well above chance on *unseen* data. This strongly suggests that the F_0 patterns of these phrases differ, in the absence of significant F_0 declination effects. We will analyze the F_0 contours of these five phrases in detail in the data clustering experiment.

7.2.2 Data Clustering

We performed K-means clustering on the training tokens of each phrase class to identify if there exist canonical F_0 patterns. As in the classification experiments, a principle component analysis is also applied on the four-dimensional feature vector prior to the clustering, mainly to normalize the variance on each dimension. A Euclidean distance metric is used

in the clustering algorithm.

The K-means clustering algorithm can always find an arbitrary number of data clusters, especially when the input data are noisy and the specified number of clusters is small compared to data size. In order to ensure that the output clusters are indeed "typical" patterns for the data, we experimented with the number of clusters used by the algorithm, and applied a self-selection process to filter out "bad" clusters after the initial clusters were obtained. The selection procedure works as follows. We train a diagonal Gaussian model for each data cluster obtained by clustering, and re-assign each token into one of the clusters through classification. A threshold is applied on the classifier probability score of each token, to discard tokens that are not well-represented by any of the Gaussian models. The remaining tokens in each data cluster are then counted, and clusters with a significant number of tokens are retained as "typical" patterns for the phrase.

We settled on a maximum number of 8 clusters in the K-means clustering algorithm. After the "self-selection" procedure was applied, 3 to 4 F_0 contour pattern clusters emerged for each type of phrase. We represented each cluster by the mean F_0 contour of the constituent tokens in the cluster, as displayed in Figure 7-1. These clusters will be referred to as subclasses in our discussions hereafter. The number of tokens in each cluster/subclass is shown in the legends of the figure, along with the name of each subclass, which has been chosen arbitrarily to uniquely identify the clusters.

As shown in Figure 7-1, most phrase subclasses differ significantly in shape, except for subclasses of the "<tell_me>" phrase and two subclasses of the "<what_is>" phrase (i.e., "what_is-C4" and "what_is-C6"). This suggests that the subclasses are not simply due to variations caused by speaker F_0 range differences. Several interesting observations can be made from the figure. For example, the "what_is-C4" and "what_is-C6" patterns seem to demonstrate the "peak delay" phenomenon (Silverman and Pierrehumbert 1990; van Santen and Hirschberg 1994), i.e., the accented syllable has a rising F_0 contour, while the following unaccented syllable has a significantly higher F_0 level than the previous accented syllable. The subclasses of the "<SU>" and "<US>" phrases are particularly "expressive", possibly due to the fact that these phrases are likely to be accented (because they convey important information such as a place or a date), and they carry a phrase tone as well (because they are

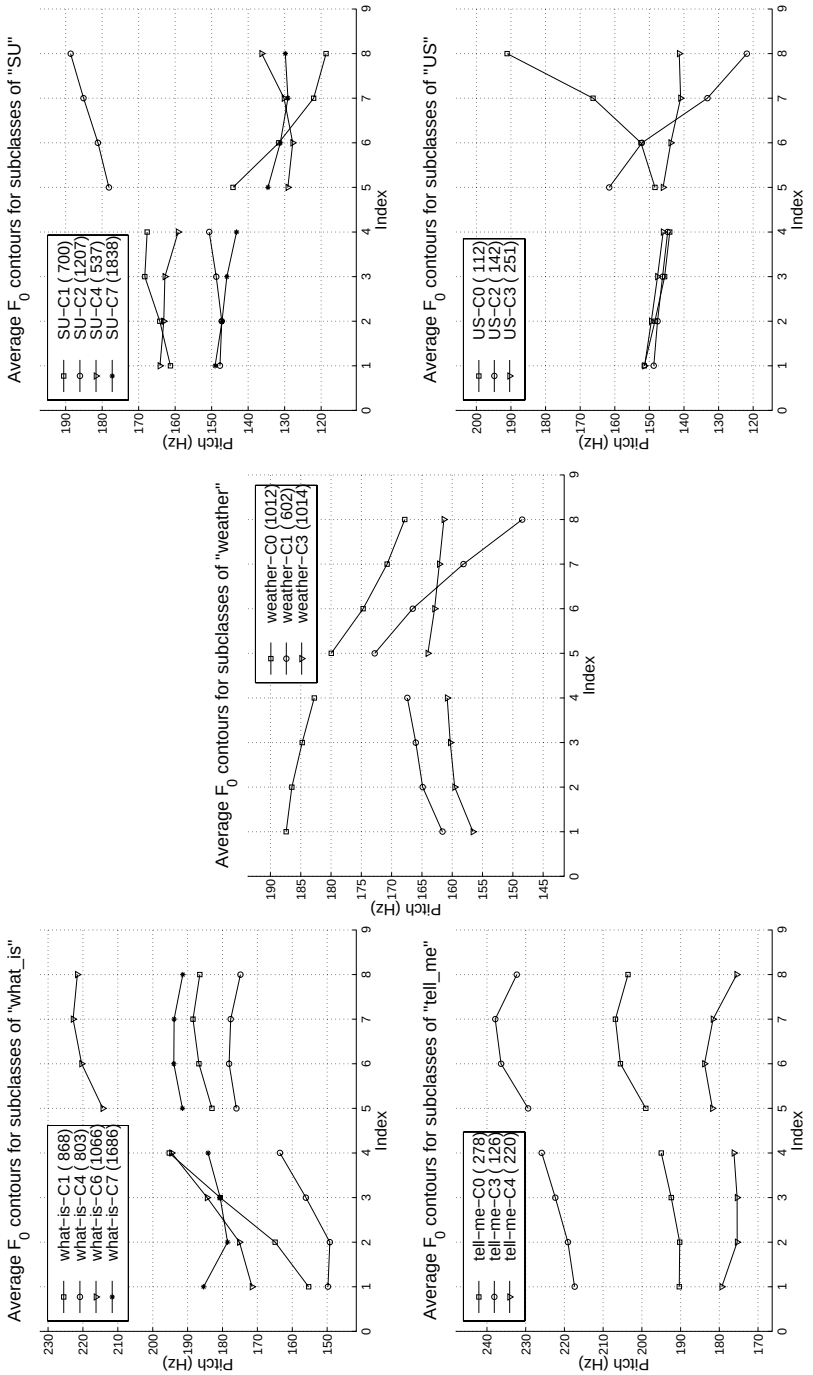

Figure 7-1: Mean F_0 contours for phrase clusters obtained by unsupervised clustering. Each phrase F_0 contour is shown as a sequence of two syllable F_0 contours, each represented by 4 connected samples. The indices of these samples are shown on the X-axis in the plots. The number of tokens in each cluster is shown in parentheses after the cluster name, which has been chosen arbitrarily to uniquely identify the clusters.

at the end of an utterance). There are three patterns for the "<US>" phrase: rising ("US-C0"), falling ("US-C2"), and flat[3] ("US-C3"). It is interesting to note that the average F_0 contours of the first syllable (unstressed) for the three subclasses are virtually the same, while the large differences among subclasses are only on the second syllable (stressed). This seems to be consistent with the intonation theory that only stressed syllables are likely to be accented. The first syllable lacks variation because it does not carry any intonational events, while the second syllable is responsible for signaling both the accents (if any) and the phrase boundary tone. The "<SU>" phrase also has the basic rise, fall, and flat patterns. However, the first syllable in the "<SU>" phrase also demonstrates variations, possibly due to its role in carrying pitch accents. In particular, the "SU-C1" and "SU-C4" patterns have higher F_0 levels for the first syllable. We suspect that the first syllable in these two subclasses is more accented. The "SU-C7" pattern is fairly "plain", and its mean F_0 contour is very similar to that of the "US-C3" pattern.

We listened to some utterances labeled with the "SU-C2" or "US-C0" patterns, and generally perceived a rising (question) intonation. These subclasses possibly correspond to the $L^*\ H^-H\%$ and $L^*\ L^-H\%$ patterns described in the ToBI labeling convention (Silverman et al. 1992). However, we are unable to systematically relate these acoustically derived classes to categories defined in prosodic labeling conventions. It will be interesting to perform the data clustering experiment on prosodically labeled data to facilitate such comparisons.

7.2.3 Correlations of Phrase Patterns

We have identified a set of canonical F_0 patterns for each phrase using a "data-driven" approach. We now use these subclasses to examine if there exist correlations among the acoustic realizations of the phrases within an utterance, e.g., if certain F_0 contour patterns of the "<what_is>" phrase are more likely to occur with certain F_0 contour patterns of the "<SU>" phrase in the same utterance. The test set is used to carry out this study, because the utterances in the test set are more homogeneous and each contains exactly three phrases.

[3]The slight falling slope of this subclass is likely due to an overall F_0 declination.

	weather-C0	weather-C1	weather-C3	SU-C1	SU-C2	SU-C4	SU-C7
what_is-C1	-0.16	0.42	-0.29	0.25	-0.35	-0.21	0.13
what_is-C4	**-0.58**	0.06	0.06	0.10	**0.58**	-0.35	**-1.01**
what_is-C6	**0.67**	-0.15	-0.10	**-0.68**	-0.27	**0.52**	0.47
what_is-C7	0.12	-0.28	0.12	-0.03	-0.28	0.10	0.23
weather-C0	-	-	-	**-0.70**	**-1.59**	**0.71**	**0.87**
weather-C1	-	-	-	0.28	-0.08	-0.11	-0.21
weather-C3	-	-	-	-0.07	0.23	-0.13	-0.14

Table 7-5: Mutual information between each pair of subclass patterns calculated for phrases in utterances matching "<what_is> the <weather> in|for|on <SU>." The subclass patterns have been described graphically in Figure 7-1. Mutual information measures larger than 0.5 or smaller than -0.5 are highlighted in **boldface**. A total number of 610 utterances is used in the computation.

We use the *mutual information* measure to quantify the correlation, which is based on the frequency counts of the phrase subclasses in the test utterances. Analogous to the "self-selection" process used in data clustering, we trained a diagonal Gaussian model using the training tokens in each phrase subclass[4], and classified the phrases in the test utterances into one of the subclasses. We then counted the number of each individual subclass and the number of each subclass *pair* in the test utterances. The mutual information between a pair of subclasses, A and B, is computed as follows:

$$M(A,B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(AB)}{P(A)P(B)} \tag{7.1}$$

The mutual information between A and B is zero if A and B are statistically independent of each other, positive if A is more likely to occur with B, and negative if A is less likely to occur with B.

Table 7-5 shows the mutual information between each pair of subclasses for three phrases, i.e., "<what_is>", "<weather>", and "<SU>", computed using 610 test utterances. The number of tokens in each phrase subclass on the test data is summarized in Table 7-6. The "<tell_me>" and "<US>" phrases are ignored, due to sparse data for these two phrases in

[4]The five-class classification accuracy on the test data can be improved to 58.8% if the phrase models are trained using only training tokens in the subclasses, i.e., data that have "survived" the selection procedure in data clustering, instead of using the entire training data.

<what_is>	# Tokens	<weather>	# Tokens	<SU>	# Tokens
what_is-C1	135	weather-C0	71	SU-C1	181
what_is-C4	167	weather-C1	200	SU-C2	181
what_is-C6	81	weather-C3	339	SU-C4	79
what_is-C7	137	-	-	SU-C7	169

Table 7-6: Number of tokens in each subclass of "<what_is>", "<weather>", and "<SU>" phrases on the test data.

the test utterances. Most of the mutual information measures between subclasses are close to zero, which implies that the correlation between those subclasses are small. However, we have observed a few subclasses which have large positive or negative values for their mutual information with some other subclasses. We plotted the *compatible* subclass pairs (those with large positive mutual information measures) in Figure 7-2, and the *incompatible* subclass pairs (those with large negative mutual information measures) in Figure 7-3, for easy reference.

The "weather-C0" subclass (with a high, falling mean F_0 contour) seems to have strong "preferences" with regard to other phrase subclasses. For example, the mutual information between "what_is-C6" (with a very high, rising mean F_0 contour) and "weather-C0" is 0.67. From the mean F_0 contours of "what_is-C6" and "weather-C0" shown in Figure 7-2, it seems that these two F_0 patterns may form one intonation phrase. On the other hand, the mutual information between "what_is-C4" (with the lowest F_0 among all subclasses of "<what_is>") and "weather-C0" is -0.58, which suggests that these two patterns are highly incompatible. The mean F_0 contours of "what_is-C4" and "weather-C0" are shown in Figure 7-3 along with other incompatible pairs. Intuitively, we think that it is difficult (and unnatural) to start a "<weather>" word from an F_0 onset that is higher than the F_0 offset of the preceding syllable. However, the number of "weather-C0" tokens in the test data is relatively small; thus, the mutual information might not be very robustly estimated. There also seem to be some "long-distance" correlations between the "<what_is>" phrase and the "<SU>" phrase. For example, the "what_is-C4" pattern (very low F_0) seems to be more compatible with the "SU-C2" pattern ("question intonation"), and least compatible with "SU-C7" (the "plain" pattern).

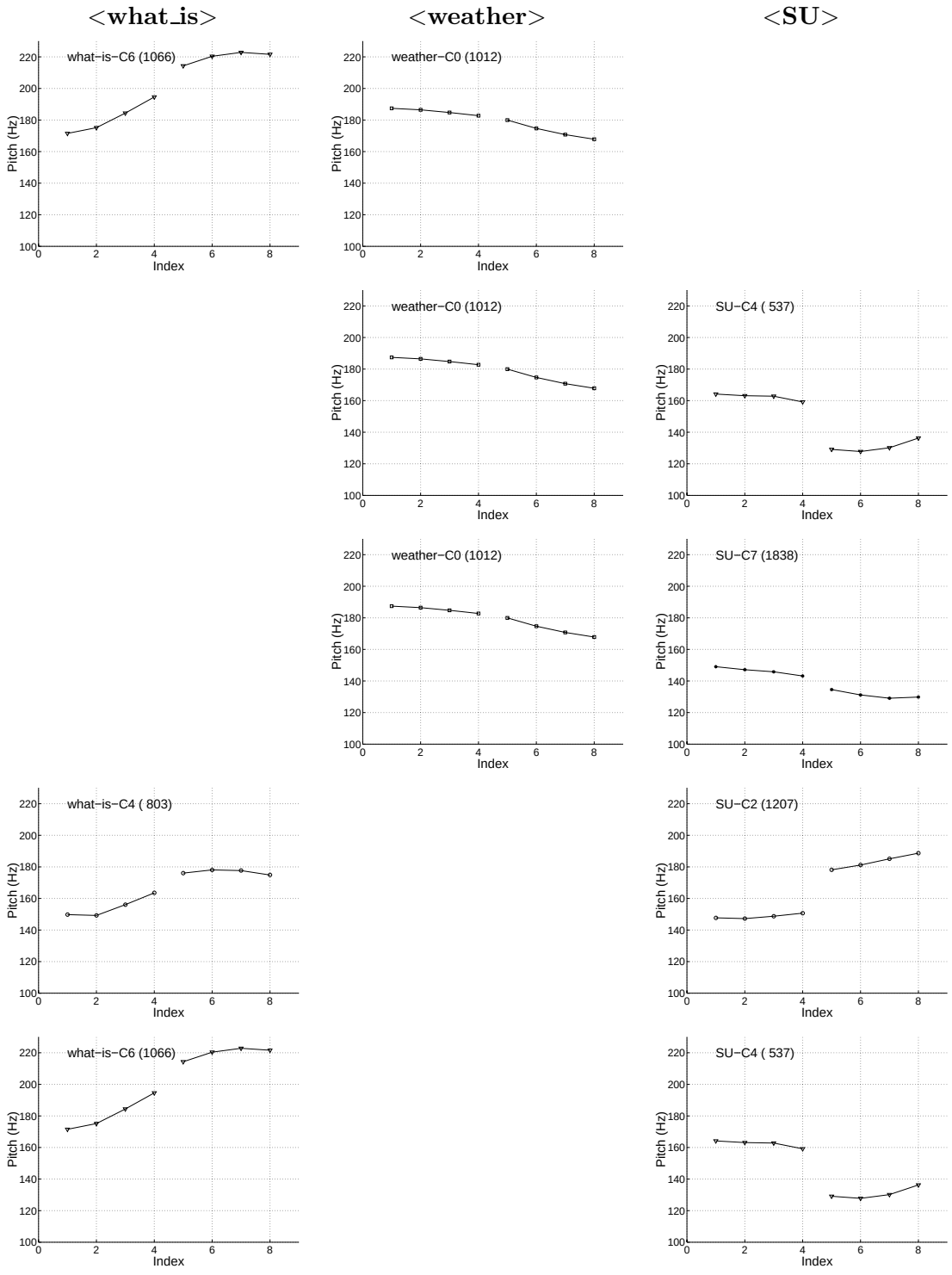

Figure 7-2: Compatible subclass pairs as indicated by mutual information measures.

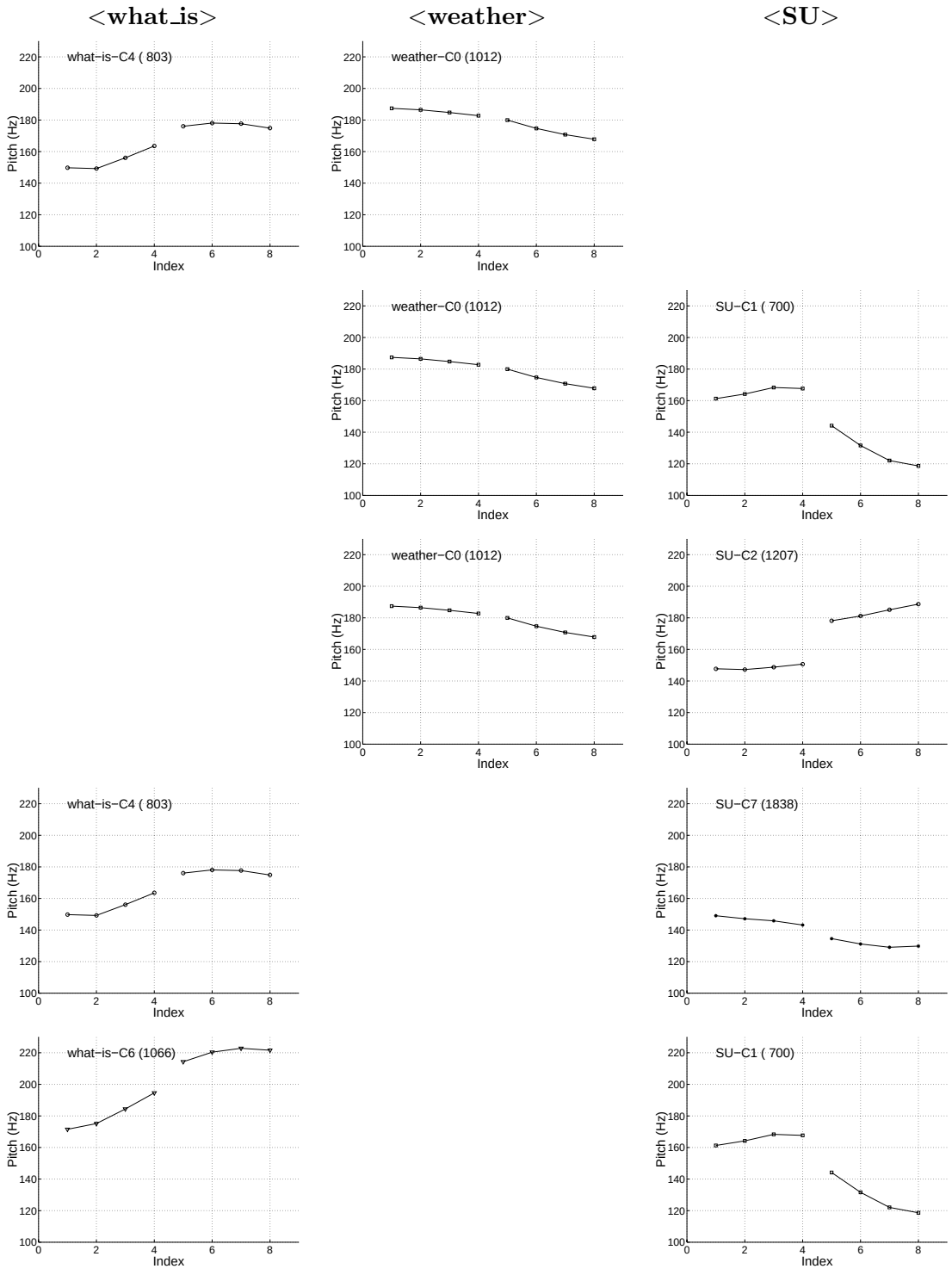

Figure 7-3: Incompatible subclass pairs as indicated by mutual information measures.

Although we are generally unable to derive linguistic explanations for these observations, it is interesting to know that there exist certain correlations among phrases in an utterance. We developed a framework which is able to quantify these correlations using statistical methods. Such information can potentially be utilized to provide additional constraints in scoring phrase F_0 patterns, in a way similar to using language models in word recognition.

7.3 Summary

In this chapter, we presented preliminary experiments towards developing a framework to model the intonational aspects of certain syntactic/semantic structures in English utterances, without using intermediate prosodic transcriptions.

We selected five common two-syllable "phrases" from the JUPITER corpus to form our initial model inventory. We characterized these phrase contours using a concatenation of F_0 features extracted from its constituent syllables, and trained a diagonal Gaussian mixture model for each phrase. We obtained a five-class classification accuracy of 60.4% on the training set, and 56.4% on unseen test data. Our various classification results clearly indicate that there exists information in the F_0 contours of the five phrases that can be used to distinguish these phrases. We can expand this framework to include more phrases, even of variable lengths. In general, a phrase of N syllables can be characterized by a concatenation of its constituent syllable features. These models can be applied in a post-processing framework, to score the intonation patterns of recognizer hypotheses. We hypothesize that these scores can be used to resort the N-best outputs for improved recognition accuracy, or to support the rejection of erroneous hypotheses.

We also performed an unsupervised data clustering experiment to identify typical F_0 contour patterns for each of the five phrases. This is of both practical and theoretical interest, because intonation theories have described the intonation contour as a sequence of categorical events (Ladd 1996). We want to know if we can identify canonical F_0 contour patterns, which are analogous to distinctive categories in acoustic representation, in a "data-driven" manner. We found some interesting F_0 patterns from the clustering process. However, we are unable to systematically relate these acoustically derived classes to cate-

gories defined in prosodic labeling conventions. It will be interesting to perform the data clustering experiment on prosodically labeled data to facilitate such comparisons.

Using the automatically derived phrase subclasses, we are able to describe a phrase contour symbolically and use statistical methods to quantify the correlations between each pair of phrase patterns. We found that there were compatible and incompatible phrase pattern pairs, and some of these observations correspond well with our intuitions. Although we are generally unable to derive linguistic explanations for these observations, we can nevertheless describe them in a quantitative way. Such information can potentially be utilized to provide additional constraints in scoring phrase F_0 patterns, in a way similar to using language models in word recognition.

We consider our initial experimental results as promising. As possible directions for future work, we can expand this framework to include more phrase patterns, incorporate more prosodic features into acoustic modeling, and test the models in recognition and confidence modeling tasks.

Chapter 8

Summary and Future Work

8.1 Summary

As we have pointed out in the introduction (Chapter 1), prosodic cues, namely, F0, duration, and energy, play an important role in human speech communication. At the lexical level, prosody helps define words and shapes the segmental property of sounds. Above the lexical level, prosody structures a message into sentences and smaller phrases, determines the sentence mood, and marks the focus of a sentence. Prosody also conveys extra-linguistic information such as gender, emotion and attitude of a speaker. Prosodic models can potentially be used in many aspects of a human-computer dialogue system, including speech recognition, syntactic/semantic analysis, topic segmentation, dialogue act determination, dialogue control, speech synthesis, etc.

In this thesis, we have explored prosodic models for Mandarin Chinese and English telephone speech along various dimensions, within the context of improving speech recognition and understanding performance in dialogue systems. In the following, we briefly recapitulate the methodologies and main results of our explorations, followed by a summary of the main contributions of this thesis.

Robust Pitch Tracking

Pitch detection is a critical first step in the analysis and modeling of speech prosody. The fundamental frequency is an important feature for many prosodic components, such as

lexical stress, tone, and intonation. However, pitch estimation errors and the discontinuity of the F_0 space make F_0 related measurements noisy and undependable. Pitch detection algorithms also have inferior performance on *telephone speech*, due to signal degradation caused by the noisy and band-limited telephone channel. To address these problems, we have implemented a novel *continuous* pitch detection algorithm (CPDA), which has been designed explicitly to promote robustness for telephone speech and prosodic modeling (Chapter 2). The algorithm derives reliable estimations of pitch and the temporal change of pitch from the entire harmonic structure. The estimations are obtained easily with a logarithmically sampled spectral representation (i.e., DLFT spectrum), because signals with different F_0 can be aligned by simple *linear shifting*. The correlation of the DLFT spectrum with an ideal harmonic template provides a robust estimation of F_0. The correlation of two DLFT spectra from adjacent frames gives a very reliable estimation of F_0 change. The constraints for both $\log F_0$ and $\Delta \log F_0$ are then combined in a dynamic programming search to find a very smooth pitch track. The DP search is able to track F_0 continuously regardless of the voicing status, while a separate voicing decision module computes a probability of voicing per frame. We have demonstrated that the CPDA is robust to signal degradation inherent in telephone speech. In fact, the overall gross error rate for studio and telephone speech is nearly the same (4.25% vs. 4.34%). We have also demonstrated that the CPDA has superior performance for both voiced pitch accuracy and tone classification accuracy compared with an optimized algorithm in XWAVES.

Lexical Tone Modeling for Mandarin Chinese Speech Recognition

We first performed empirical studies of Mandarin tone and intonation, focusing on analyzing sources of tonal variations (Chapter 3). We demonstrated an F_0 downtrend for Mandarin Chinese using both position dependent tone statistics and the average F_0 contour of a set of aligned utterances. The data show that F_0 decreases consistently within a phrase; while there is a jump of F_0 level after each phrase boundary. However, the F_0 hike is relatively small compared to the declination, and the overall change of F_0 level is predominantly decreasing. We then characterized the effects of phrase boundary, tone coarticulation, and tone sandhi using a similar method, by comparing average tone contours in different

immediate contexts. The most obvious effects of a phrase boundary seem to be on the tone excursion range. Tone 2, tone 3 and tone 4 at internal phrase-final positions reach a lower F_0 target than at other positions; tone 2 at phrase-initial positions also seems to rise to a higher F_0 target than at other positions. Tone coarticulation is manifested as both carry-over and anticipatory effects, with the carry-over effects appearing to be more significant. The carry-over effects mainly change the F_0 onset of the following tone, and the change is assimilatory in nature. The anticipatory effects are more complex, with both assimilatory and dissimilatory effects present in the data. The sandhi-changed tone 3 is similar to tone 2. It seems that a context dependent model using both left and right tone context should be able to capture the tone sandhi variation.

We incorporated tone models into speech recognition and tried to account for the tonal variation factors in tone modeling for improved tone classification and speech recognition performances (Chapter 4). We first developed a segment-based tone classification framework, which used discrete Legendre decomposition to parameterize tone F_0 contours and Gaussian classifiers to estimate tone probability scores. Using this basic framework, we demonstrated that tone recognition performance for *continuous* Mandarin speech can be significantly improved by taking into account sentence declination, phrase boundary, and tone context influences. We then implemented two mechanisms in the SUMMIT speech recognition system to incorporate tone model constraints: first-pass and post-processing. Integration of a simple four-tone model into the first-pass Viterbi search reduced the baseline speech recognition error rate by 30.2% for the digit domain and by 15.9% for the spontaneous utterances in the YINHE domain. Using the simple four-tone model to resort the recognizer 10-best outputs yielded similar improvements for both domains. However, further improvements by using more refined tone models were small and not statistically significant. This suggests that a simple and efficient strategy to utilize tone information can be achieved by integrating a simple four-tone model into the first-pass Viterbi search.

Lexical Stress Modeling for Spontaneous English Speech Recognition

Lexical stress in English is the analogy of tone in Mandarin Chinese. Leveraging the same mechanisms developed for Mandarin tone modeling, we tested the approach of scoring

the lexical stress patterns of recognizer hypotheses to improve speech recognition performance (Chapter 5). The motivation is also similar to that for tone modeling, i.e., erroneous hypotheses will have worse "stress scores" than the correct hypothesis. However, unlike Mandarin tones, the acoustic manifestations of lexical stress are quite obscure. To address this issue, we first examined the correlation of various pitch, energy, and duration measurements with lexical stress on a large corpus of spontaneous utterances in the JUPITER domain. We found that the distributions of most prosodic features differed for different lexical stress classes; however, the extent of overlap among classes was also significant. We then performed classification experiments to identify the most informative features for lexical stress. The best single feature for stress classification was the integral of energy over the nucleus vowel, while the best set of prosodic features consisted of the integral of energy, raw duration, pitch slope, and the average probability of voicing. Higher stress classification accuracy was achieved by using spectral features (MFCCs) in addition to the prosodic features. In the recognition experiments, however, it was found that the gain using only prosodic features was greater than when MFCC features were also used. We observed that the best set of prosodic features were completely computable from information extracted from the segmental region alone. It is also convenient that F_0 difference performed better than F_0 average; thus, the sentence-level normalization is not required.

We integrated the stress model into the recognizer first-pass Viterbi search. We found that using a simple four-class stress model achieved small but statistically significant gain over the state-of-the-art baseline performance on JUPITER. However, more refined models taking into account the intrinsic prosodic differences among vowels failed to improve the performance further. Our recognition results of a one-class (including all vowels) prosodic model seemed to suggest that the gain of using prosodic models was mainly due to the elimination of implausible hypotheses, e.g., preventing vowel/non-vowel or vowel/non-phone confusions, rather than by distinguishing different stress and segmental classes. We have also found that it is more advantageous to apply prosodic constraints selectively, i.e., only on phones for which the prosodic measurements are "meaningful" and more informative.

Recognition Confidence Scoring Enhanced with Prosodic Features

Moving beyond improving speech recognition, we examined the use of prosodic cues in recognition confidence scoring for improved accept/reject decisions (Chapter 6). Hirschberg and colleagues (1999, 2000) have found that there exist statistically significant differences in the mean values of certain prosodic features between correctly and incorrectly recognized user turns, and these prosodic cues can be used to improve accept/reject decisions on recognition outputs. However, it was also found that the efficacy of the prosodic information was dependent on the quality of the recognition system. We first tested if the approach of using prosodic cues in utterance-level confidence scoring can be generalized to the JUPITER system, which has been well-trained on a large corpus of speech data. We found that there were differences in both the means and the variances of some prosodic measurements between correctly and incorrectly recognized utterances, with the variances generally larger for misrecognized utterances. This is consistent with the intuition that "outliers" are more likely to be incorrectly recognized. We observed that prosodic cues achieved small but statistically significant improvement in the detection of utterance hypothesis errors. We also examined if prosodic features can be used to better distinguish correctly and incorrectly recognized *words*. Although the methodology is quite similar to that used in the utterance-level confidence scoring, the underlying assumptions are somewhat different. We expect that there exist prosodic cues to speech artifacts (such as background speech), which are a significant source of recognition errors. Furthermore, "unusual" prosodic measurements are sometimes indicative of speech recognition errors. We found that prosodic cues achieved small but statistically significant improvement in the detection of word errors as well. The receiver-operator characteristic (ROC) curves were also improved overall in both cases.

Characterization of English Intonation Contours

We presented preliminary experiments towards developing a framework to model the intonational aspects of certain syntactic/semantic structures in English utterances, without using intermediate prosodic transcriptions (Chapter 7). We selected five common two-syllable "phrases" from the JUPITER corpus to form our initial model inventory. We characterized these phrase contours using a concatenation of F_0 features extracted from their constituent

syllables, and trained diagonal Gaussian mixture models for these phrases. We obtained a five-class classification accuracy of 60.4% on the training set, and 56.4% on unseen test data (which can be improved to 58.8% if using only "clean" training examples). Our various classification results clearly indicate that there exists information in the F_0 contours of the five phrases that can be used to distinguish these phrases. These models can be applied in a post-processing framework, to score the intonation patterns of recognizer hypotheses. We hypothesize that these scores can be used to resort the N-best outputs for improved recognition accuracy, or to support the rejection of erroneous hypotheses.

We also performed an unsupervised data clustering experiment to identify typical F_0 contour patterns for each of the five phrases. This is of both practical and theoretic interest, because intonation theories have described the intonation contour as a sequence of categorical events (Ladd 1996). We want to know if we can identify canonical F_0 contour patterns, which are analogues to distinctive categories in acoustic representation, in a "data-driven" manner. We found some interesting F_0 patterns from the clustering process. However, we are unable to systematically relate these acoustically derived classes to categories defined in prosodic labeling conventions. It will be interesting to perform the data clustering experiment on prosodically labeled data to facilitate such comparisons.

Using the automatically derived phrase subclasses, we were able to describe a phrase contour symbolically and use statistical methods to quantify the correlations between each pair of phrase patterns. We found that there were compatible and incompatible phrase pattern pairs, and some of these observations correspond well with our intuition. Although we are generally unable to derive linguistic explanations for these observations, we can nevertheless describe them in a quantitative way. Such information can potentially be utilized to provide additional constraints in scoring phrase F_0 patterns, in a way similar to using language models in word recognition.

Thesis Contributions

In summary, we have made the following contributions to research in the area of prosodic modeling in this thesis:

- The development of a *continuous* pitch tracking algorithm that is designed specially

for telephone speech and prosodic modeling applications.

- An empirical study of Mandarin tone and tonal variations, which analyzes the effects of tone coarticulation, tone sandhi, and some intonation components, on the acoustic realizations of tone.
- The development of a mechanism which is able to combine multiple classifiers and to selectively score for a subset of phones in the recognition first-pass search.
- The development and analysis of a preliminary framework for characterizing pitch contours of spoken English utterances without intermediate prosodic transcription.
- Improvements in speech recognition and confidence scoring performance using prosodic information.

8.2 Future Directions

This thesis has explored a wide range of topics in the area of prosodic modeling. Many aspects of the work presented in this thesis can be improved or extended. Some methodologies and empirical results are also potentially useful for other applications. In this section, we mention some of these directions for future work.

Several aspects of the pitch tracking algorithm can be improved. First, the dynamic programming search back traces the optimum pitch contour upon arriving at the last frame in the utterance, which causes a significant delay in the overall pipe-lined recognition process. This can be improved by allowing the DP to back track periodically, e.g., whenever the best node score is much higher than the scores of its competitors, or upon transition from voiced regions to unvoiced regions. In case of conflicting paths, back tracking from later frames should have higher priority than any back tracking from previous frames. Although this is still not a completely pipe-lined design, the delay could be significantly reduced. Second, it is usually necessary to normalize a pitch contour by its average value in prosodic modeling. This function can be implemented within the pitch tracking algorithm, e.g., the average F_0 value can be estimated during back tracking, utilizing voicing probabilities to discount F_0 values from unvoiced frames.

We have designed various ways to separate tonal and intonational aspects in Mandarin Chinese utterances, which are both manifested mainly as F_0 movements. We demonstrated the F_0 downtrend for Mandarin Chinese using position-dependent tone statistics and the average F_0 contour of a set of aligned utterances. We also characterized the effects of phrase boundary, tone coarticulation, and tone sandhi using a similar method, by comparing average tone contours in different immediate context. This methodology can be extended to studying the effects of pitch accent or lexical stress, by comparing the average tone contours between accented/unaccented or stressed/unstressed syllables. However, such a study is likely to rely on the availability of a corpus with pitch accent and lexical stress labels.

The empirical study on Mandarin tone and intonation is not only useful for improving tone recognition, but also useful for Mandarin speech synthesis. The F_0 downtrend and the context-dependent tone models (including dependencies on tone context, phrase boundary, pitch accent, lexical stress, etc.) can be utilized to construct a target F_0 contour from a prosodically tagged Chinese sentence. The F_0 downtrend modeling can be improved for synthesis applications. We have found that F_0 downtrend can be better characterized on the phrase-level than on the utterance-level; however, we have chosen a sentence-level modeling due to the unavailability of phrase boundary information during recognition. A phrase-level characterization can easily be incorporated into a synthesis system, because the phrase boundaries are usually provided in the text input.

The dependency of tone expression on phrase boundary and pitch accent can also potentially be utilized to identify phrase boundaries and pitch accents, which can in turn be used in higher-level linguistic processing. For examples, tone 2, tone 3 and tone 4 before a internal phrase boundary reach a lower F_0 target than at other positions, while tone 2 at a phrase-initial position seems to rise to a higher F_0 target than at other positions. Other studies have found that pitch accent (focus) enlarges the F_0 range of words at a non-final focus, and the F_0 range after the focus is both lowered and reduced (Xu 1999). Such information can be used in addition to durational and pause-related cues in detecting phrase boundaries and pitch accents.

Although using prosodic features improved accept/reject decisions in both utterance-

level and word-level confidence scoring, the gain by using prosodic features is relatively small. We suspect that the simple confidence classifier model used in our experiments might not be optimal for incorporating prosodic features. In particular, the classifier is only able to deal with numerical features; in addition, all features are reduced to a one-dimensional feature by a linear discrimination projection, which implies a linear decision boundary. We expect the prosodic cues to be more effective if the classifier can pay more attention to "outliers", because many prosodic features have a "wider" distribution for incorrectly recognized utterances than for correctly recognized utterances. It is desirable to design a probabilistic classifier which is able to (1) use both numerical and symbolic features, (2) exploit dependencies among features, and (3) handle complex decision boundaries.

The confidence scoring framework can easily be adapted to perform phrase boundary detection. Prosodic features on the utterance-level and word-level have been extracted to support accept/reject decisions of recognition hypotheses. These features include not only simple prosodic measurements, such as mean and maximum F_0 for a word or an utterance, pause duration after a word, etc., but also complex measurements such as utterance and word speaking rates. In general, we can obtain raw or normalized prosodic measurements from utterance-, word-, syllable-, and segment- levels, given the recognizer N-best outputs. With a modest amount of labeled training data, we can identify the best features for phrase boundary classification, using the mutual information based feature selection procedure described in Section 6.3.2. Hypothesized phrase boundaries with probability scores can be inserted into the recognizer hypotheses, which can be input to the TINA probabilistic parsing system (Seneff 1992). The parsing grammar can be augmented with optional phrase boundaries at appropriate locations, similar to the approach described in (Kompe 1997).

We have also begun to develop a framework to model the intonational aspects of certain syntactic/semantic structures in JUPITER utterances, without using intermediate prosodic transcriptions. Our experiments on classifying five common two-syllable phrases on unseen data, based on F_0 information only, have shown promising results. We can expand this framework to include more phrases, potentially of variable lengths. These phrases can be chosen from common structures in the parsing grammar, with considerations for stress patterns. These phrase models can be applied in a post-processing framework, to score the

intonation patterns of recognizer hypotheses. We can use these scores, appropriately normalized, to resort the recognizer N-best outputs. This is similar to the approach of scoring for the lexical stress patterns of recognizer hypotheses, except that the phrase models capture higher-level constraints and more contextual information. We can also build intonation models for the key concepts in a domain, such as place names in the JUPITER system. We can use these models to support the rejection of erroneous key word hypotheses, which are particularly important to the understanding performance.

Appendix A

ASR Confidence Features

The following 15 utterance-level confidence features are taken from (Hazen et al. 2000b) to train the baseline utterance confidence model:

1. **total_score:** the total score from all models (i.e., the acoustic, language, and pronunciation models) for the top-choice hypothesis.
2. **total_score_per_word:** the average score per word from all models for the top-choice hypothesis.
3. **lexical_score:** the total score of the N-gram model for the top-choice hypothesis.
4. **lexical_score_per_word:** the average score per word of the N-gram model for the top-choice hypothesis.
5. **acoustic_score:** the total acoustic score summed over all acoustic observations for the top-choice hypothesis.
6. **acoustic_score_per_bound:** the average acoustic score per acoustic observation for the top-choice hypothesis.
7. **total_drop:** the drop in the total score between the top hypothesis and the second hypothesis in the N-best list.
8. **acoustic_drop:** the drop in the total acoustic score between the top hypothesis and the second hypothesis in the N-best list.

9. **lexical_drop:** the drop in the total N-gram score between the top hypothesis and the second hypothesis in the N-best list.

10. **average_purity:** the average N-best purity of all words in the top-choice hypothesis. The N-best purity for a hypothesized word is the fraction of N-best hypotheses in which that particular hypothesized word appears in the same location in the sentence.

11. **frac_high_purity:** the fraction of words in the top-choice hypothesis which have an N-best purity of greater than one half.

12. **nbest_average_purity:** the average N-best purity of all words in all of the N-best list hypothesis.

13. **nbest_frac_high_purity:** The percentage of words across all N-best list hypotheses which have an N-best purity of greater than one half.

14. **nbest:** the number of sentence hypotheses in the N-best list. This number is usually its maximum value of ten but can be less if fewer than ten hypotheses are left after the search prunes away highly unlikely hypotheses.

15. **num_words:** the number of hypothesized words in the top-choice hypothesis.

The following 10 word-level confidence features are taken from (Hazen et al. 2000b) to train the baseline word confidence model:

1. **bound_score_mean:** the mean log likelihood acoustic score across all acoustic observations in the word hypothesis.

2. **bound_norm_score_mean:** the mean of the acoustic likelihood scores (not the *log* scores) across all acoustic observations in the word hypothesis.

3. **bound_score_min:** the minimum log likelihood score across all acoustic observations in the word hypothesis.

4. **bound_score_std_dev:** the standard deviation of the log likelihood acoustic scores across all acoustic observations in the word hypothesis.

5. **bound_diff_from_max_mean:** the average difference, across all acoustic observations in the word hypothesis, between the acoustic score of a hypothesized phonetic unit and the acoustic score of highest scoring phonetic unit for the same observation.

6. **bound_likelihood_mean:** mean score of the catch-all model across all observations in the word hypothesis.

7. **num_bounds:** the number of acoustic observations within the word hypothesis.

8. **frac_nbest:** the fraction of the N-best hypotheses in which the hypothesized word appears in the same position in the utterance.

9. **num_nbest:** the number of sentence level N-best hypotheses generated by the recognizer.

10. **utt_score:** the utterance confidence score generated from the utterance features described above.

Appendix B

Complete Word/Phrase List

The following is a complete list of words and phrases occurred in the training data for each of the five phrase classes:

- **<what_is>**

  how is what are what is what was

- **<tell_me>**

  give me show me tell me

- **<weather>**

  dew point forecast pressure sun rise sun set
  weather wind speed wind speeds

- **<SU>**

  Aalten Akron Amherst Asheville Asia
  Athens Austin Baghdad Bangkok Bangor
  Beaumont Bedford Belgium Berkeley Bismarck
  Bombay Bosnia Boston Bridgeport Brisbane
  Brownsville Brunswick Brussels Cairo Cambridge
  Camden Charleston Charlotte Chile China
  Christchurch Cleveland Concord Dallas Dayton

Denmark	Denver	Dublin	Durham	Egypt
England	Erie	Europe	Fairbanks	Fairfax
Falmouth	Fargo	Fiji	Finland	Flagstaff
Florence	Frankfurt	Fresno	Friday	Georgia
Glasgow	Greenland	Greenville	Hamburg	Hartford
Hilo	Holland	Houston	Huntsville	Iceland
Ireland	Israel	Jackson	Jordan	Juneau
Kansas	Kenya	Lansing	Lhasa	Lincoln
Lisbon	London	Lowell	Mali	Maui
Medford	Melbourne	Memphis	Monday	Moscow
Munich	Nashville	Nassau	New York	Newark
Newport	Norfolk	Oakland	Paris	Phoenix
Pittsburgh	Pittsfield	Plymouth	Poland	Portland
Portsmouth	Princeton	Provo	Pueblo	Quito
Raleigh	Richmond	Rio	Russia	Rutland
Salem	Scotland	Scranton	Springfield	Stockholm
Strasbourg	Stuttgart	Sunday	Sydney	Tampa
Texas	Thursday	today	Tokyo	Trenton
Tucson	Tuesday	Tulsa	Tunis	Utah
Venice	Warsaw	Wednesday	weekend	Whitehorse
Worcester	Zaire	Zurich		

- **<US>**

Beirut	Belize	Berlin	Brazil	Cancun
Detroit	Eugene	Iran	Iraq	Japan
Kunming	Kuwait	Madrid	Marseille	Nepal
Peru	Pierre	Quebec	Shanghai	Spokane
Tahoe	Taipei	Taiwan	Tehran	Tibet
today	tonight	Ukraine	Vermont	Xian

Bibliography

Adda-Decker, M. and G. Adda (1992). Experiments on stress-dependent phone modeling for continuous speech recognition. In *Proc. ICASSP'92*, San Francisco, USA, pp. 561–564.

Aull, A. M. (1984). Lexical stress and its application in large vocabulary speech recognition. Master's thesis, Massachusetts Institute of Technology.

Aull, A. M. and V. W. Zue (1985). Lexical stress determination and its application to large vocabulary speech recongnition. In *Proc. ICASSP '85*, Tampa, USA, pp. 1549–1552.

Cao, Y., Y. Deng, H. Zhang, T. Huang, and B. Xu (2000). Decision tree based Mandarin tone model and its application to speech recognition. In *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1759–1762.

Carter, D. M. (1987). An information-theoretic analysis of phonetic dictionary access. *Computer Speech and Language 2*, 1–11.

Chang, P.-C., S.-W. Sun, and S.-H. Chen (1990). Mandarin tone recognition by Multi-Layer Perceptron. In *Proc. ICASSP'90*, Albuquerque, USA, pp. 517–520.

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, S.-H. and Y.-R. Wang (1990). Vector quantization of pitch information in Mandarin speech. *IEEE Transactions on Communications 38*(9), 1317–1320.

Chen, S.-H. and Y.-R. Wang (1995). Tone recognition of continuous Mandarin speech based on Neural Networks. *IEEE Transactions on Speech and Audio Processing 3*(2), 146–150.

Chung, G. (1997). Hierarchical duration modelling for a speech recognition system. Master's thesis, Massachusetts Institute of Technology.

Chung, G. and S. Seneff (1999). A hierarchical duration model for speech recognition based on the ANGIE framework. *Speech Communication 27*, 113–134.

Daly, N. A. (1994). *Acoustic-Phonetic and Linguistic Analyses of Spontaneous Speech: Implications for Speech Understanding*. Ph. D. thesis, Massachusetts Institute of Technology.

Dellaert, F., T. Polzin, and A. Waibel (1996). Recognizing emotion in speech. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 1970–1973.

Droppo, J. and A. Acero (1998). Maximum *a posteriori* pitch tracking. In *Proc. ICSLP'98*, Sydney, Australia, pp. 943–946.

Dumouchel, P. and D. O'Shaughnessy (1995). Segmental duration and HMM modeling. In *Proc. EUROSPEECH'95*, Madrid, Spain, pp. 803–806.

Flammia, G., J. Glass, M. Phillips, J. Polifroni, S. Seneff, , and V. Zue (1994). Porting the bilingual VOYAGER system to Italian. In *Proc. ICSLP '94*, Yokohama, Japan, pp. 911–914.

Freij, G. J. and F. Fallside (1990). Lexical stress estimation and phonological knowledge. *Computer Speech and Language 4*, 1–15.

Fu, S. W. K., C. H. Lee, and O. L. Clubb (1996). A survey on Chinese speech recognition. *Communications of COLIPS 6*(1), 1–17.

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

Gao, Y., H.-W. Hon, Z. Lin, G. Loudon, S. Yogananthan, and B. Yuan (1995). TANGERINE: A large vocabulary Mandarin dictation system. In *Proc. ICASSP'95*, Detroit, USA, pp. 77–80.

Gao, Y., T. Huang, Z. Lin, B. Xu, and X. D. (1991). A real-time Chinese speech recognition system with unlimited vocabulary. In *Proc. ICASSP'91*, Toronto, Canada, pp. 257–260.

Gårding, E. (1987). Speech act and tonal pattern in standard Chinese. *Phonetica 44*, 113–29.

Geoffrois, E. (1996). The multi-lag-window method for robust extended-range f_0 determination. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 2399–2402.

Gillick, L. and S. Cox (1989). Some statistical issues in the comparison of speech recognition algorithms. In *ICASSP'89*, Glasgow, Scotland, pp. 532–535.

Glass, J. (1988). *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*. Ph. D. thesis, Massachusetts Institute of Technology.

Glass, J., J. Chang, and M. McCandless (1996). A probabilistic framework for feature-based speech recognition. In *ICSLP'96*, Philadelphia, USA, pp. 2277–2280.

Glass, J., G. Flammia, D. Goodline, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue (1995). Multilingual spoken language understanding in the MIT VOYAGER system. *Speech Communication 17*, 1–18.

Glass, J., T. Hazen, and L. Hetherington (1999). Real-time telephone-based speech recognition in the JUPITER domain. In *Proc. ICASSP'99*, Phoenix, USA, pp. 61–64.

Goddeau, D., E. Brill, J. Glass, C. Pao, and M. Phillips (1994). Galaxy: A human-language interface to on-line travel information. In *ICSLP'94*, Yokohama, Japan, pp. 707–710.

Grigoriu, A., J. P. Vonwiller, and R. W. King (1994). An automatic intonation tone contour labelling and classification algorithm. In *Proc. ICASSP'94*, Adelaide, Australia, pp. 181–184.

Hakkani-Tür, D., G. Tür, A. Stolcke, and E. Shriberg (1999). Combining words and prosody for information extraction from speech. In *Proc. EUROSPEECH'99*, Budapest, Hungary, pp. 1991–1994.

Halberstadt, A. K. (1998). *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*. Ph. D. thesis, Massachusetts Institute of Technology.

Hazen, T. J., T. Burianek, J. Polifroni, and S. Seneff (2000a). Integrating recognition confidence scoring with langauge understanding and dialogue modeling. In *Proc. ICSLP'00*, Beijing, China.

Hazen, T. J., T. Burianek, J. Polifroni, and S. Seneff (2000b). Recognition confidence scoring for use in speech understanding systems. In *Proceedings 2000 IEEE Workshop on Automatic Speech Recognition and Understanding*, Paris, France.

Hermes, D. (1988). Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America 83*(1), 257–273.

Hess, W. (1983). *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag.

Hieronymus, J. L., D. McKelvie, and F. McInnes (1992). Use of acoustic sentence level and lexical stress in HSMM speech recognition. In *Proc. ICASSP'92*, San Francisco, USA, pp. 225–227.

Hirose, K. (1995). Disambiguating recognition results by prosodic features. In Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody*, pp. 327–342. Springer.

Hirose, K. and K. Iwano (2000). Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition. In *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1763–1766.

Hirschberg, J., D. Litman, and M. Swerts (1999). Prosodic cues to recognition errors. In *Proceedings 1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, USA.

Hirschberg, J., D. Litman, and M. Swerts (2000). Generalizing prosodic prediction of speech recognition errors. In *Proc. ICSLP'00*, Beijing, China.

Ho, T.-H., H.-M. Wang, L.-F. Chien, K.-J. Chen, and L.-S. Lee (1995). Fast and accurate continuous speech recognition for Chinese language with very large vocabulary. In *Proc. EUROSPEECH'95*, Madrid, Spain, pp. 211–214.

Hon, H.-W., B. Yuan, Y.-L. Chow, S. Narayan, and K.-F. Lee (1994). Towards large vocabulary Mandarin Chinese speech recognition. In *Proc. ICASSP'94*, Adelaide, Australia, pp. 545–548.

Howie, J. M. (1974). On the domain of tone in Mandarin. *Phonetica 30*, 129–148.

Hsieh, H.-Y., R.-Y. Lyu, and L.-S. Lee (1996). Use of prosodic information to integrate acoustic and linguistic knowledge in continuous Mandarin speech recognition with very large vocabulary. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 809–812.

Huang, H. C.-H. and F. Seide (2000). Pitch tracking and tone features for Mandarin speech recognition. In *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1523–1526.

Huber, R., E. Nöth, A. Batliner, J. Buckow, V. Warnke, and H. Niemann (1998). You beep machine - emotion in automatic speech understanding systems. In *Proc. First Workshop on Text, Speech, Dialogue*, pp. 223–228.

Hunt, A. (1994). A generalised model for utilising prosodic information in continuous speech recognition. In *Proc. ICASSP'94*, Adelaide, Australia, pp. 169–172.

Hunt, A. (1995). Training prosody-syntax recognition models without prosodic labels. In Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody*, pp. 309–325. Springer.

Hurley, E., J. Polifroni, and J. Glass (1996). Telephone data collection using the World Wide Web. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 1898–1901.

Huttenlocher, D. P. (1984). Acoustic-phonetic and lexical constraints in word recognition: lexical access using partial information. Master's thesis, Massachusetts Institute of Technology.

Jenkin, K. L. and M. S. Scordilis (1996). Development and comparison of three syllable stress classifiers. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 733–736.

Jensen, U., R. K. Moore, P. Dalsgaard, and B. Lindberg (1994). Modelling intonation contours at the phrase level using continuous density hidden Markov models. *Computer Speech and Language 8*, 247–260.

Jones, M. and P. C. Woodland (1994). Modelling syllable characteristics to improve a large vocabulary continuous speech recogniser. In *Proc. ICSLP'94*, Yokohama, Japan, pp. 2171–2174.

Jurafsky, D., E. Schriberg, B. Fox, and T. Curl (1998). Lexical, prosodic, and syntactic cues for dialogue acts. In *ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pp. 144–120.

Kamppari, S. O. and T. J. Hazen (2000). Word and phone level acoustic confidence scoring. In *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1799–1802.

Kannan, A. and M. Ostendorf (1997). Modeling dependency in adaptation of acoustic models using multiscale tree processes. In *Proc. EUROSPEECH'97*, Rhodes, Greece, pp. 1863–1866.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America 82*(3), 737–793.

Klatt, D. H. and K. N. Stevens (1972). Sentence recognition from visual examination of spectrograms and machine-aided lexical searching. In *Proceedings 1972 Conference on Speech Communication and Processing*, Bedford, USA, pp. 315–318.

Kompe, R. (1997). *Prosody in Speech Understanding Systems*. Springer.

Kompe, R., A. Kieβling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T. Ruland, and H. U. Block (1997). Improving parsing of spontaneous speech with the help of prosodic boundaries. In *Proc. ICASSP '97*, Munich, Germany, pp. 811–814.

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

Lea, W. A. (1973). Evidence that stressed syllables are the most readily decoded portions of continuous speech. *Journal of the Acoustical Society of America 55*, 410(A).

Lea, W. A. (1980). Prosodic aids to speech recognition. In W. A. Lea (Ed.), *Trends in Speech Recognition*, pp. 166–205. Englewood Cliffs, New Jersey: Prentice-hall, Inc.

Lee, L.-S., C.-Y. Tseng, H.-Y. Gu, F.-H. Liu, C.-H. Chang, Y.-H. Lin, Y. Lee, S.-L. Tu, S.-H. Hsieh, and C.-H. Chen (1993). Golden Mandarin (I) - a real-time Mandarin speech dictation machine for Chinese language with very large vocabulary. *IEEE Transactions on Speech and Audio Processing 1*(2), 158–179.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman, M. and J. Pierrehumbert (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle (Eds.), *Language Sound Structure*, pp. 157–233. Cambridge, MA: M.I.T. Press.

Lin, C.-H., C.-H. Wu, P.-Y. Ting, and H.-M. Wang (1996). Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units. *Speech Communication 18*, 175–190.

Liu, L.-C., W.-J. Yang, H.-C. Wang, and Y.-C. Chang (1989). Tone recognition of polysyllable words in Mandarin speech. *Computer Speech and Language 3*, 253–264.

Lyu, R.-Y., L.-F. Chien, S.-H. Hwang, H.-Y. Hsieh, R.-C. Yang, B.-R. Bai, J.-C. Weng, Y.-J. Yang, S.-W. Lin, K.-J. Chen, C.-Y. Tseng, and L.-S. Lee (1995). Golden Mandarin (III) - a user-adaptive prosodic-segment-based Mandarin dictation machine for Chinese language with very large vocabulary. In *Proc. ICASSP'95*, Detroit, USA, pp. 57–60.

Martin, P. (1982). Comparison of pitch detection by cepstrum and spectral comb analysis. In *Proc. ICASSP '82*, New York, USA, pp. 180–183.

Modern Chinese Dictionary (1978). *Xian Dai Han Yu Ci Dian* 现代汉语词典. Beijing, China: Shang wu yin shu guan 商务印书馆.

Noll, A. M. (1970). Pitch determination of human speech by the harmonic products spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In *Symposium on Computer Processing in Communication*, pp. 779–797. New York: University of Brooklyn.

Ostendorf, M. and K. Ross (1995). A multi-level model for recognition of intonation labels. In Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody*, pp. 291–308. Springer.

Ostendorf, M. and N. M. Veilleux (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics 20*(1), 27–53.

Ostendorf, M., C. W. Wightman, and N. M. Veilleux (1993). Parse scoring with prosodic information: An analysis-by-synthesis approach. *Computer Speech and Language 7*, 193–210.

Oviatt, S., A. Levow, E. Moreton, and M. MacEachern (1998). Modeling global and focal hyperarticulation during human-computer error resolution. *Journal of the Acoustical Society of America 104*(5), 3080–3098.

Oviatt, S., G.-A. Levow, M. MacEachernn, and K. Kuhn (1996). Modeling hyperarticulate speech during human-computer error resolution. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 801–804.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Ph. D. thesis, Massachusetts Institute of Technology.

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

Polzin, T. S. (2000). Verbal and non-verbal cues in the communication of emotions. In *Proc. ICASSP'00*, Istanbul, Turkey, pp. 2429–2432.

Price, P. J., S. Ostendorf, S. Shattuck-Hufnagel, and C. Fong (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America 90*(6), 2956–2970.

Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal (1976). A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Audio, Signal, and Speech Processing 24*, 399–417.

Sagisaka, Y., N. Campbell, and N. Higuchi (Eds.) (1995). *Computing Prosody*. Springer.

Schroeder, M. R. (1968). Period histogram and product spectrum: new methods for fundamental-frequency measurement. *Journal of the Acoustical Society of America 43*, 829–834.

Secrest, B. G. and G. R. Doddington (1983). An integrated pitch tracking algorithm for speech synthesis. In *Proc. ICASSP '83*, pp. 1352–1355.

Selkirk, E. O. (1984). *Phonology and Syntax: The Relationship between Sound and Structure*. Cambridge, MA, USA: MIT Press.

Seneff, S. (1978). Real-time harmonic pitch detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*(4), 358–365.

Seneff, S. (1992). TINA: A natural language system for spoken language applications. *Computational Linguistics 18*(1), 61–86.

Seneff, S., J. Glass, T. Hazen, Y. Minami, J. Polifroni, and V. Zue (2000). MOKUSEI: A Japanese spoken dialogue system in the weather domain. *NTT R&D 49*(7), 376–382.

Seneff, S., R. Lau, J. Glass, and J. Polifroni (1999). The MERCURY system for flight browsing and pricing. *MIT Spoken Language System Group Annual Progress Report*, 23–28.

Serridge, B. (1997). Context-dependent modeling in a segment-based speech recognition system. Master's thesis, Massachusetts Institute of Technology.

Shattuck-Hufnagel, S. and A. E. Turk (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research 25*(2), 193–247.

Shen, X.-N. (1989). Interplay of the four citation tones and intonation in Mandarin Chinese. *Journal of Chinese Linguistics 17*(1), 61–74.

Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics 18*, 281–295.

Shih, C. (1997). Declination in Mandarin. In G. K. A. Botinis and G. Carayannis (Eds.), *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, pp. 293–296. Athens, Greece.

Shih, C. and R. Sproat (1992). Variations of the Mandarin rising tone. In *Proceedings of the IRCS Research in Cognitive Science*, Philadelphia, USA, pp. 193–200.

Shih, C.-L. (1986). The prosodic domain of tone sandhi in Chinese. *Dissertation Abstracts International 47*, 889A.

Shinoda, K. and T. Watanabe (1996). Speaker adaptation with autonomous model complexity control by MDL principle. In *ICASSP'96*, Atlanta, USA, pp. 717–720.

Shriberg, E., R. Bates, and A. Stolcke (1997). A prosody-only decision-tree model for disfluency detection. In *Proc. EUROSPEECH'97*, Rhodes, Greece, pp. 2383–2386.

Shriberg, E., E. Wade, and P. Price (1992). Human-machine problem solving using spoken language ststems (SLS): factors affecting performance and user satisfaction. In *Proc. of the DARPA Speech and Natural Language Workshop*, pp. 49–54.

Shu, H., C. Wooters, O. Kimball, T. Colthurst, F. Richardson, S. Matsoukas, and H. Gish (2000). The BBN Byblos 2000 conversational Mandarin LVCSR system. In *Proc. 2000 Speech Transcription Workshop*.

Silverman, K., M. B. Beckman, J. Pirelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992). ToBI: A standard for labeling English prosody. In *Proc. ICSLP'92*, Banff, Canada, pp. 867–870.

Silverman, K. E. A. and J. B. Pierrehumbert (1990). The timing of prenuclear high accents in English. In J. Kingston and M. E. Beckerman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge University Press.

Sjölander, K. and J. Högberg (1997). Using expanded question sets in decision tree clustering for acoustic modeling. In *Proceedings 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 179–184.

Sluijter, A. and V. van Heuven (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America 100*(4), 2471–2485.

Soltau, H. and A. Waibel (1998). On the influence of hyperarticulated speech on recognition performance. In *Proc. ICSLP'98*, Sydney, Australia.

Stolcke, A., E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc. ICSLP'98*, Sydney, Australia, pp. 2247–2250.

Ström, N., L. Hetherington, T. Hazen, E. Sandness, and J. Glass (1999). Acoustic modeling improvements in a segment-based speech recognizer. In *Proceedings 1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, CO.

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

ten Bosch, L. F. M. (1995). Automatic classification of pitch movements via MLP-based estimation of class probabilities. In *Proc. ICASSP'95*, Detroit, USA, pp. 608–611.

Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustical Society of America 55*, 1061–1069.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research 1*, 155–182.

van Kuijk, D. and L. Boves (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication 27*(2), 95–111.

van Kuijk, D., H. van den Heuvel, and L. Boves (1996). Using lexical stress in continuous speech recognition for dutch. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 1736–1739.

van Santen, J. and J. Hirschberg (1994). Segmental effects on timing and height of pitch contours. In *Proc. ICSLP'94*, Yokohama, Japan, pp. 719–722.

Waibel, A. (1988). *Prosody and Speech Recognition*. London: Pitman.

Wang, C. (1997). Porting the GALAXY system to Mandarin Chinese. Master's thesis, Massachusetts Institute of Technology.

Wang, C., S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue (2000). MUXING: a telephone-access Mandarin conversational system. In *Proc. ICSLP'00*, Beijing, China, pp. 715–718(II).

Wang, C., J. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue (1997). YINHE: A Mandarin Chinese version of the GALAXY system. In *Proc. EUROSPEECH'97*, Rhodes, Greece, pp. 351–354.

Wang, C. and S. Seneff (1998). A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition. In *Proc. ICSLP'98*, Sydney, Australia, pp. 635–638.

Wang, C. and S. Seneff (2000a). Improved tone recognition by normalizing for coarticulation and intonation effects. In *Proc. ICSLP'00*, Beijing, China, pp. 83–86 (II).

Wang, C. and S. Seneff (2000b). Robust pitch tracking for prosodic modeling of telephone speech. In *Proc. ICASSP'00*, Istanbul, Turkey, pp. 1143–1146.

Wang, C.-F., H. Fujisaki, and K. Hirose (1990). Chinese four tone recognition based on the model for process of generating F_0 contours of sentences. In *Proc. ICSLP'90*, Kobe, Japan, pp. 221–224.

Wang, H.-M. and L.-S. Lee (1994). Tone recognition for continuous Mandarin speech with limited training data using selected context-dependent hidden Markov models. *Journal of the Chinese Institute of Engineers 17*(6), 775–784.

Wang, H.-M., J.-L. Shen, Y.-J. Yang, C.-Y. Tseng, and L.-S. Lee (1995). Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data. In *Proc. ICASSP'95*, Detroit, USA, pp. 61–64.

Wang, W. S.-Y. and K.-P. Li (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research 10*, 629–636.

Wang, Y.-R. and S.-H. Cheng (1994). Tone recognition of continuous Mandarin speech assisted with prosodic information. *Journal of the Acoustical Society of America 96*(5), 2637–2645.

Wang, Y.-R., J.-M. Shieh, and S.-H. Cheng (1994). Tone recognition of continuous Mandarin speech based on hidden Markov model. *International Journal of Pattern Recognition and Artificial Intelligence 8*(1), 233–245.

Wu, Y., K. Hemmi, and K. Inoue (1991). A tone recognition of polysyllabic Chinese words using an approximation model of four tone pitch patterns. In *Proc. IECON'91*, Kobe, Japan, pp. 2115–2119.

Wu, Z. and M. Lin (1989). *Shi Yan Yu Yin Xue Gai Yao* 实验语音学概要. Beijing, China: Higher Education Press.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America 95*(4), 2240–2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics 25*, 62–83.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica 55*, 179–203.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f_0 contours. *Journal of Phonetics 27*, 55–105.

Yang, W.-J., J.-C. Lee, Y.-C. Chang, and H.-C. Wang (1988). Hidden Markov model for Mandarin lexical tone recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 36*(7), 988–992.

Ying, G. S., L. H. Jamieson, R. Chen, C. D. Michell, and H. Liu (1996). Lexical stress detection on stress-minimal word pairs. In *Proc. ICSLP'96*, Philadelphia, USA, pp. 1612–1615.

Zue, V. and J. Glass (2000). Conversational interfaces: advances and challenges. *Proceedings of the IEEE, Special Issue on Spoken Language Processing 88*(8), 1166–1180.

Zue, V., S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington (2000). JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing 8*(1), 100–112.

Zue, V., S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill (1993). The MIT ATIS system: December 1993 progress report. In *Proc. ARPA Spoken Language Technology Workshop*, Plainsboro, USA, pp. 67–71.