Processing of speech utterances for computer aided training of speaking skills

Zhao, Sixuan

2014

Zhao, S. (2014). Processing of speech utterances for computer aided training of speaking skills. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/61736>

<https://doi.org/10.32657/10356/61736>

PROCESSING OF SPEECH UTTERANCES FOR COMPUTER AIDED TRAINING OF SPEAKING SKILLS

ZHAO SIXUAN

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2014

Acknowledgement

I would like to thank my first supervisor Professor Koh Soo Ngee and current supervisor Professor Soon Ing Yann. I owe them a lot because they provided numerous and invaluable guidance to me in the past several years. They are excellent advisors and very nice friends who supported me consistently and patiently in many areas. I always receive precious suggestions and advice from them throughout my PhD years. Therefore, I would like to express my utmost gratitude to them for all their help.

Special thanks must also go to my co-supervisor Professor Luke Kang Kwong. His profound knowledge in linguistic areas contributes a lot to our research topics. All the precious comments and suggestions from him have helped me progress in my project. His patience and meticulous proofreading is indispensable to all of our publications.

I wish to gratefully acknowledge the Institute for Media Innovation as well as the School of Electrical and Electronic Engineering in Nanyang Technological University for the PhD grant provided. Otherwise, I can never conduct my PhD study in Singapore.

I would also like to express my gratitude to Mrs. How Seok Lai at the Media Technology Laboratory for all the technical support given to me. She made great efforts to maintain a safe and pleasant working environment for me and the others in the laboratory.

I am indebted and grateful to many of my good friends, Mr. Deng Jun, Mr. Yu Gang, Mr. Liu Benxu, Mr. Wang Li, Mr. Wang Hongxing, Mr. Ren Zhou, Mr. Wang Qiang, Ms. Li Yusha, Mr. Ye Wenbin, Mr. Chen Binbin, Ms. Zhang Xiaoyan, Dr. Zhang Lining, Mr. Ye Wei, Dr. Lei Baiying, Dr. Dai Peng, Dr. Liu Xiaobei, Mr. Xie Xiaohua, Mr. Yang Zai, Mr. Fang Yuming, Mr. Zhang He, Mr. Zhou Yuan, Ms. Liu Yisi, Mr. Hou Xiyuan, Mr. Liang Hui, Mr. Yi Xiang, Ms. Wang Xiaoyan, Ms. An Jianing, Ms. Wei Caiyi, and all the others if I have not included their names. Particularly, I would like to thank my girlfriend Liu Jia for

all of her understanding and encouragements through my PhD study. I would also like to thank all the other good friends at the Institute for Media Innovation, Media Technology Laboratory and other institutes in Nanyang Technological University for their help and encouragements.

Last and most importantly, I wish to express my gratitude to my family. I am grateful to all my family members for their generous understanding and long-term support for my PhD study. I would like to thank my parents for their everlasting encouragements and unconditional support. Their love is my pillar of strength which propels me to advance when I was frustrated. Without their support, I can never finish my pursuit to PhD and could have burned out so easily by the tremendous pressure.

Table of Contents

Acknowledgement	i
Table of Contents	iii
Abstract	vii
List of Figures	ix
List of Tables	xii
List of Abbreviations	xiii
Chapter 1 Introduction	1
1.1 History of Computer-aided Language Learning	2
1.2 Structure of CALL System for Speaking Skills Training	3
1.3 Research Topics	6
1.3.1 Prosody Evaluation for Speaking Skills Assessment	6
1.3.2 Feedback Utterances for Language Learning	8
1.3.3 Phonetic Segmentation	11
1.4 Major Contributions	13
1.5 Organization of the Thesis	15
Chapter 2 Literature Review	17
2.1 Prosody Evaluation.....	17
2.1.1 Reference-dependent Evaluation.....	17
2.1.2 Reference-independent Evaluation.....	22
2.2 Feedback Utterances Generation	27

2.2.1	Feedback Utterances in CALL Systems.....	27
2.2.2	Differences between Voice Conversion and Accent Reduction.....	30
2.2.3	Rule-Based Accent Reduction.....	31
2.2.4	Reference-Based Accent Reduction	34
2.3	Phonetic Segmentation	37
2.3.1	Text-independent Segmentation	38
2.3.2	Text-dependent Segmentation	40
2.4	Summary	49
Chapter 3	Prosody Evaluation Based on Prosodic Unit Segmentation	51
3.1	Motivations & Proposed Work.....	51
3.2	Prosodic Unit.....	52
3.3	Automatic Quasi-foot Segmentation	54
3.4	Reference-dependent Prosody Evaluation Based on Prosodic Unit.....	59
3.4.1	Framework of Reference-dependent Evaluation System	59
3.4.2	Distance Measurement	60
3.4.3	Evaluation Experiment	63
3.5	Reference-independent Prosody Evaluation Based on Prosodic Unit.....	65
3.5.1	Support Vector Machine.....	66
3.5.2	Segmentation Method & Features	69
3.5.3	Experimental Results.....	72
3.5.4	Feature Selection Test	76
3.5.5	Relationship between Accented Word Ratio and Human Score	79
3.6	Summary	82
Chapter 4	Generation of Feedback Utterances	84
4.1	Motivations & Proposed Work.....	84
4.2	Multi-corpora Accent Reduction.....	86
4.2.1	Accent Reduction Scheme.....	86

4.2.2	Prosodic Modification	89
4.2.3	Segmental Modification	90
4.2.4	Experimental Conditions	94
4.2.5	Assessments of Accent Reduced Utterances	95
4.2.6	Effects of the Nationality of Learners and the Corpus	103
4.2.7	Effects of Different Accent Reduction Setups	104
4.3	Feedback Utterances Generated by Different Synthesis Methods and Voice Conversion Method.....	105
4.3.1	Accent Reduction Methods	106
4.3.2	Voice Conversion Method.....	110
4.3.3	Comparison of Different Feedback Utterances	113
4.3.4	Accented Word Comparison	117
4.3.5	Discussion.....	119
4.4	Summary	121
Chapter 5	A Hybrid Scheme for Automatic Phonetic Segmentation	123
5.1	Motivations & Proposed Work.....	123
5.2	Baseline Phonetic Segmentation based on Forced Alignment	127
5.3	Refinement Scheme for Phonetic Segmentation	130
5.3.1	Statistical Correction	130
5.3.2	Multi-resolution Fusion	135
5.3.3	Refinement Using Predictive Models.....	137
5.4	Experimental Results on a Single Corpus	141
5.5	Cross-corpora Segmentation	149
5.6	Summary	154
Chapter 6	Conclusion & Recommendations for Future Research	156
6.1	Prosody Evaluation Based on Quasi-foot Segmentation.....	156
6.2	Generation of Feedback Utterances	157

6.3 Refinement of Phonetic Segmentations	159
6.4 Recommendations for Future Research	160
List of Publications	163
References	165

Abstract

Computer-aided language learning (CALL) involves the studies and applications of speech and language processing technologies to improve the process of language acquisition. Ideally, an effective computer-aided language learning system should be able to accurately assess the performance of a language learner and generate meaningful feedback during his learning process. This thesis addresses several issues which are relevant to computer-aided language learning systems, particularly for learning of English as a second language (L2).

The first issue is about the evaluation of prosody of the learner's speech utterances. Prosody evaluation plays an important role in automatic assessment of English proficiency of L2 learners. It requires segmentation of a speech utterance into appropriate units to achieve effective modeling of prosodic features. A segmentation scheme is proposed to improve the prosody evaluation results by taking into account prosodic units. Unlike lexical units such as word or phoneme, prosodic units correspond to the phrasing and rhythm information and are more appropriate for the purpose of prosody evaluation. An algorithm is designed to segment the speech signal into prosodic units automatically, and it is shown that the algorithm can detect the proposed prosodic unit with reasonable accuracy.

The production of audio feedback which is an important component of CALL is studied in this thesis. The learner's vocal features and the teacher's linguistic gestures are combined to produce effective feedback utterances which can facilitate the acquisition of English speaking skills. An accent reduction scheme which reduces the perceived accents in the learner's utterances is studied. A multi-corpora experiment designed to examine effects of external factors on the accent reduction results resolves some ambiguities in the literature. In addition, different speech synthesis methods are described and implemented to perform accent reduction. Voice conversion is also applied as a new method to generate

feedback utterances which possess the learner's vocal features and the teacher's linguistic gestures. The feedback utterances generated by various accent reduction methods are compared with that produced by voice conversion in order to identify an optimal way to produce feedback utterances with high nativeness and acoustic quality. Consequently, a multi-stage feedback scheme is proposed.

Finally, the phonetic segmentation process is studied and its performance is improved to produce more accurate phone boundary information. Such kind of information can contribute to the development of speech technology areas which can be applied to the design of computer-aided language learning systems. Three different refinement methods, i.e., statistical correction, multi-resolution fusion, and predictive model based refinement, are presented. These methods are combined appropriately to improve the accuracy of the baseline phonetic segmentation system using forced alignment. The proposed refinement scheme is also extended to a cross-corpora scenario, which enables the analysis of a new corpus with limited labeled data and thus facilitates the application of the new corpus for various purposes such as speech recognition and linguistic research.

List of Figures

Figure 1.1 Structure of Computer-aided Language Learning System	5
Figure 1.2 Illustration of Prosodic Features.....	6
Figure 1.3 Feedback Utterances to Non-native English Learners	10
Figure 1.4 Phonetic Segmentation	12
Figure 2.1 Basic Flowchart of Reference-dependent Evaluation	18
Figure 2.2 DTW Alignment between Two Parallel Speech Utterances	19
Figure 2.3 MFCC Based Alignment	21
Figure 2.4 Intonation Contour Types.....	22
Figure 2.5 Basic Flowchart of Reference-independent Evaluation	23
Figure 2.6 Intonation Recognition Grammar.....	25
Figure 2.7 Comparison between Voice Conversion and Accent Reduction.....	30
Figure 2.8 Rule-based Accent Reduction	31
Figure 2.9 Reference-based Accent Reduction.....	34
Figure 2.10 Text-independent Segmentation Purely Based on Acoustic Features	39
Figure 2.11 Regression Fusion Methods for Phonetic Segmentation Refinement	43
Figure 2.12 Fusion Based Phonetic Segmentation Using Candidate Selection...	44
Figure 2.13 Predictive Model Based Refinement – Searching around the Preliminary Boundary	46
Figure 2.14 Singing Voice Segmentation by DTW and HMM	49
Figure 3.1 Waveform, Pitch and Intensity of the Sentence “ <i>I/ felt that I/ might/ never stop the ma/chine from/ running</i> ”	53
Figure 3.2 Logistic Regression	55
Figure 3.3 Example of Quasi-foot Segmentation	58

Figure 3.4 Framework of Reference-dependent Evaluation System	60
Figure 3.5 Example of Non-linear SVM.....	69
Figure 3.6 Cross-validation Results of Different Segmentation Methods	74
Figure 3.7 Relationship between Feature Dimension and Evaluation Accuracy.....	77
Figure 3.8 Relationship between Accented Word Ratio and Human Score	80
Figure 3.9 Accented Word Ratio Distribution across Scores Given by Three Evaluators	81
Figure 4.1 Modification of Pitch and Duration Using PSOLA	88
Figure 4.2 Accent Reduction Flowchart Using LP-PSOLA.....	89
Figure 4.3 Interpolation of DTW Alignment.....	91
Figure 4.4 Interpolation of Spectral Envelopes	92
Figure 4.5 VTLN Warping to Preserve Speaker Identity	93
Figure 4.6 Segmental Modification by Replacing LPCs	94
Figure 4.7 Accentedness Score of BURNC Based Corpus	97
Figure 4.8 Accentedness Score of ARCTIC Corpus	98
Figure 4.9 Flowchart of ITU P.563.....	100
Figure 4.10 Acoustic Quality MOS by P.563.....	102
Figure 4.11 Reduction of Perceived Accents for Different Students.....	103
Figure 4.12 Scheme of Prosodic Modification with Different Synthesis Methods	107
Figure 4.13 HSM based Prosodic Modification	108
Figure 4.14 Synthesis Scheme of STRAIGHT Model.....	110
Figure 4.15 Voice Conversion Scheme	113
Figure 4.16 Accentedness Scores of Different Stimuli.....	114
Figure 4.17 Acoustic Quality of Different Stimuli.	116
Figure 4.18 Pitch Accent Similarity Ratio – Percentage that the Learner Have the Same Pitch Accent Label as the Teacher	118

Figure 4.19 Spectrogram Comparison: (a) <i>Original Speech of the Learner</i> (b) <i>Feedback Utterances Generated by STRAIGHT Prosodic Modification</i> (c) <i>Feedback Utterances Generated by Voice Conversion</i>	120
Figure 4.20 Proposed Feedback Scheme	121
Figure 5.1 Proposed Refinement Scheme	126
Figure 5.2 Illustration of Forced Alignment Process	129
Figure 5.3 Training & Testing of the Baseline Automatic Phonetic Segmentation System Using Forced Alignment	130
Figure 5.4 Decision Tree based Clustering of Bi-phone Classes	131
Figure 5.5 Segmentation Error Distribution	132
Figure 5.6 Illustration of Relative Correction Method	134
Figure 5.7 State Selection for Relative Statistical Correction	135
Figure 5.8 Separation of Two Groups of Data Using LDA Projection	140
Figure 5.9 Segmentation Results of HMMs with Different Resolutions and Multi-resolution Fusion	145
Figure 5.10 Improvements of the Whole Test Set and Individual Phones on TIMIT	147
Figure 5.11 Overall Improvements of Segmentation Results	148
Figure 5.12 Cross-corpora Improvements Using the Hybrid Refinement Scheme	153

List of Tables

Table 3.1 Pitch Accent and Foot Boundary Detection Results.....	59
Table 3.2 Conditions of the Reference-dependent Evaluation Experiment.....	63
Table 3.3 Prosody Evaluation Results with Different Segmentation Methods.	65
Table 3.4 Experimental Conditions Based on BURNC.....	73
Table 3.5 Experimental Results Based on Different Segmentation Methods...	75
Table 3.6 Subjective Score Distribution from Three Evaluators (Number of Sentences for Each Score).....	80
Table 4.1 Experimental Conditions for Accent Reduction	95
Table 5.1 HMMs Performance with Different Stepsizes	136
Table 5.2 Segmentation Results Using Mono-/Bi- Phone Models with / without Statistical Correction.....	142
Table 5.3 Relative Correction with Different Search Ranges.....	143
Table 5.4 Refinements Using Predictive Models.....	146
Table 5.5 Cross-corpora Segmentation on WSJCAM0 Using AMA	151

List of Abbreviations

AAM	American Acoustic Model
ANN	Artificial Neural Network
ASM	Automatic Segmentation Machine
BAM	British Acoustic model
BSM	Baseline Segmentation Machine
BURNC	Boston University Radio New Corpus
CALL	Computer Aided Language Learning
DTW	Dynamic Time Warping
f0	Fundamental Frequency
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HSM	Harmonics Stochastic Model
KNN	K-Nearest Neighborhood
LDA	Linear Discriminative Analysis
LPC	Linear Predictive Coefficient
LSF	Line Spectral Frequency
LVCSR	Large Vocabulary Continuous Speech Recognition
MAE	Mean Absolute Error
MAP	Maximum A Posterior
MFCC	Mel Frequency Cepstral Coefficient
MLP	Multiple-Layer Perceptron
MOS	Mean Opinion Score
mRMR	minimum-Redundancy-Maximum- Relevance

PESQ	Perceptual Evaluation of Speech Quality
PSOLA	Pitch Synchronous Overlap and Add
RMSE	Root Mean Square Error
SNR	Signal-to-Noise Ratio
STRAIGHT	Speech Transformation and Representation by Adaptive Interpolation of Weighted Spectrogram
SVM	Support Vector Machine
SVR	Support Vector Regression
TOBI	Tone and Break Index
TSM	Tone Sequence Model
TTS	Text-to-Speech Synthesis
VC	Voice Conversion
VTLN	Vocal Tract Length Normalization

Chapter 1 Introduction

Accompanying the rapid internationalization of the world, global interactions and cooperation are increasing at an unprecedented speed. As a result, foreign language skills, especially English, are becoming more and more important in communication, trade, industry, culture and other aspects in the development of open societies. Traditional way of language learning in a classroom setting, even though widely used by most people, still has a lot of limitations which include inflexibility of learning materials and restrictions on time and location. To address the increasing demands of English language learners, computer-aided language learning (CALL) systems have been experimented and adopted to complement teaching of English in recent decades.

Compared to traditional learning in a classroom environment, CALL systems for English training possess the following benefits:

1. Self-adaptive Learning: it enables language learners to schedule and pace their learning according to their time constraint and ability.
2. Anxiety-free Practice: it allows learners to practice the language without direct interaction with other people and thus reduces the anxiety and/or shyness of learners.
3. Personalized Learning: learners can choose the most suitable learning materials or style offered by various CALL systems.

Therefore, a CALL system can be an effective tool to assist the English learning process especially of non-native learners.

In recent years, with the ever increasing computer processing speed and memory capacity at decreasing costs and the availability of advanced speech processing technology, CALL systems are becoming more and more sophisticated and effective. Although a decade ago, CALL techniques were specialized only for limited use in classroom teaching environments, it has now become a popular and convenient choice for students and teachers all over the

world for either self-learning or interactive-learning. This trend is also accelerated by the ubiquitous internet connectivity in many countries.

1.1 History of Computer-aided Language Learning

The history of CALL can be traced to 1960s. According to [1], the early developmental history of CALL possesses three stages: behavioristic CALL, communicative CALL, and integrative CALL.

From 1960s to 1970s, most behavioristic CALL systems mainly worked as a machine tutor to provide students with information and studying materials for their own practices, such as grammar learning, through computer software. In the communicative CALL era from 1970s to 1980s, with the advent of PC, the developed systems can not only work as a machine tutor, but also offer students a platform for communicative and interactive learning in groups, creating opportunity for language learning in a more realistic environment. From 1990s, the prosperous internet technology enables the development of integrative CALL systems, which integrate the features of both behavioristic CALL and communicative CALL to provide students with a more enriched learning environment.

From the history of CALL in early stages, it is clear that CALL progressed rapidly by incorporating novel information and computer technologies. Numerous reviews and summaries on CALL have been reported in [1, 2], which demonstrate the demands for and the value of CALL. However, limited by the less than mature technology in speech processing, CALL systems before 1990s mainly focused on text-based learning such as grammar and vocabulary. The very important part of language learning, i.e., speaking skills, was ignored for a long time. After 1990s, with the development of modern speech recognition and speech processing techniques, there was a revival in R & D interests for computer-aided speaking skills training and more systems have been proposed and developed based on new speech processing technologies as presented in [3-5]. In [3], the authors focus on adoption of speech recognition and speech processing method, and discuss the task design, model construction and teaching of linguistic structure in a modern CALL system. In [4], the online technique and

CD-Rom based teaching system on Chinese are discussed. In [5], a group of popular techniques and correlated software for CALL are listed and discussed and a summary of the developments in this field is given.

1.2 Structure of CALL System for Speaking Skills Training

The literatures listed in the previous section show that CALL systems targeting at speaking skills training receive more and more attention recently. Specifically, speaking skills training systems for second language (L2) English learners are in high demand. The reason is that L2 learners suffer from a number of problems such as entrenched effects and habits arising from the use of their own native language and low frequency of using L2 [6]. All of these problems hinder them from producing native utterances and lead to a perceived foreign accent in their speech utterances.

Foreign accent perceived in an utterance can be defined as the deviation from the standard segmental (formants, spectral envelope) and prosodic (pitch, duration and energy) norms of a spoken language from native speakers [7]. The English speech from non-native speakers may possess a certain degree of understandability and intelligibility to a listener [8], but the word recognition of those accented utterances always requires more efforts, especially in a noisy environment [9]. Some studies have shown that foreign accent in an utterance can actually affect the performance of speech recognition by both humans and computers [10-13]. As demonstrated in [10], speech utterances with foreign accent reduce the intelligibility by 40% compared with speech pronounced by native American English speakers. The increased speaking rate in some scenarios also reduces the intelligibility of accented speech [11]. The experiments performed by [12, 13] using speech recognition algorithms for different American accents also show significantly different recognition accuracies across American accents, which means even accents of native speakers may affect the recognition process. In addition, speakers with foreign accents may suffer from other disadvantages even when their speech utterances are understandable and intelligible to the listeners. The study in [14] demonstrates that the required processing time (the time taken to understand and

react to the speech) for listeners of non-native speech is generally longer than that of native speech, handicapping non-native speakers in scenarios in which response time is very important, e.g., business meeting or lifesaving. Besides, foreign accent can lead to discriminations or adverse treatments on the speakers as shown in [15-18] . Studies in [8] also indicate that accented speech is especially obvious and even annoying to native speakers. Such a disadvantaged situation is even more pronounced if the foreign accented speaker works with native speakers because the perceived foreign accent level may increase due to the comparison with the surrounding native speakers [19].

Factors influencing L2 learning have been discussed in a number of papers [6, 20-22]. Age of learning is proposed as a factor in determining the foreign accent of a non-native speaker in some studies [6, 21]. Other factors such as length of residence in an L2-speaking country, gender, formal instruction, motivation, language learning aptitude and amounts of native language (L1) used in daily life are also highlighted by [6]. However, the uncertainty about these factors limits their applications on the development of effective feedback for the acquisition of English by L2 learners. Although a “critical learning period” is proposed by some studies (before age of 15 as proposed in [22]), there are also counter-arguments to such a hypothesis as presented in [23, 24]. It is argued in [23] that adults may learn a foreign language more efficiently and accurately (especially in terms of phonology) according to certain neurolinguistic studies. Native-like foreign pronunciation can also be achieved by adult after puberty as shown in [24] . According to these conclusions, it is possible for an L2 learner, even starting from post-pubertal period, to obtain proficient foreign language skills.

Therefore, long-term training of speaking skills is feasible and necessary, making CALL a convenient and effective tool for this purpose due to its properties of self-adaptive and personalized learning. In addition, it is also shown that the acquisition of English speaking skills and the reduction of perceived accents are critical issues for L2 English learners. Because of the negative effects of foreign accents as well as the importance of speaking skills in international communications, this thesis mainly addresses several aspects of the CALL system for speaking skills training of L2 learners.

Figure 1.1 Structure of Computer-aided Language Learning System

An effective CALL system for speaking skills training should take into account both the assessment of the learner's performance as well as the feedback information provided to the learner. The automatic assessment of the learner's speaking skills enables the system to "understand" the learner's proficiency in English so that appropriate feedback can be provided. In particular, the feedback can give the learner valuable cues about his overall performance. Moreover, the feedback provided by a CALL system should also include the information which can help the learner identify the problems of his utterances so that his speaking skills can be improved accordingly. The proposed system diagram of CALL with feedback to learners is shown in Figure 1.1.

According to the diagram, the learner utters an English sentence as an input to the CALL system. The system first assesses the proficiency of the input utterance, which is defined as the deviation in terms of prosody and pronunciation of the utterance from a reference norm, such as native British, Australian or American English, and generate a machine score, e.g., from 1 to 5, to indicate the overall performance of the learner. In addition to this machine score, the system will also generate feedback information which identifies the

differences between the learner's non-native utterance and the reference norm, providing cues for the learner to practice and improve their speaking skills.

1.3 Research Topics

1.3.1 Prosody Evaluation for Speaking Skills Assessment

Speech signals can be represented by both segmental features, i.e., pronunciation, and supra-segmental features, i.e., prosody. Unlike segmental features associated with spectral information, prosody is defined as the combination of intonation and rhythm which reflect phrasing and prominence (or stress) information in the speech signal [25]. Particularly, intonation refers to the contour of pitch and intensity not used to distinguish words, because pitch is used to reflect lexical meanings in some tonal languages like Mandarin. Rhythm is correlated to the appropriate manipulation of duration and stress, which divides an utterance into equal portions based on stress (e.g., English), syllable (e.g., Spanish), or mora (e.g., Japanese). This definition of prosody will be used in the scope of this thesis. English is a stress-timed language and prosody in English can be generally modeled by some acoustic features of speech

Figure 1.2 Illustration of Prosodic Features

utterances, e.g., pitch, duration and intensity. A simple illustration of these prosodic features is shown in Figure 1.2, with the top plot showing speech waveform, the middle plot showing the corresponding pitch contour, and the bottom plot showing the corresponding intensity contour.

An important part of a CALL system is automatic scoring, which can indicate the proficiency of the learner by evaluating his speaking skills automatically. Considering the composition of speech signals, two kinds of automatic scoring scheme exist, i.e., pronunciation evaluation which focuses on the accuracy of the pronunciation as well as the phoneme errors of a language learner and prosody evaluation which assesses the prosody of the utterances produced by language learners.

Although pronunciation conveys lexical meanings and requires efforts in the English learning process, prosody is also very important for second language acquisition. As prosody mainly consists of supra-segmental features such as stress, rhythm and intonation of speech, it can reflect various non-lexical features of utterances which include the emotion of speakers, the type of sentences (statement, question or command), the irony, emphasis or contrast of sentences and other language elements that cannot be conveyed by grammar or vocabulary. In addition, prosody can help the listener to identify the lexical contents of an utterance even with reduced intelligibility [26]. Therefore, the evaluation of the prosody of speech utterances can give feedback about the learner's manipulation of stress, phrasing, and other non-lexical elements of English, thus contributing to the language acquisition process.

Besides, from the viewpoint of language acquisition, prosody is the hardest and the last to be acquired [27]. This conclusion given by linguists further argues for the importance of developing a CALL system for prosody training. Prosody acquisition process requires the learner to understand the use of prosody appropriately and form a speaking style similar to that of native speakers. It is quite a demanding and long-term task which particularly requires help from fields beyond education and linguistics, such as CALL. Basically, there are three main prosody problems of non-native English learners:

1. Prosody influenced by the accents and knowledge of their own native language.
2. Failure to manipulate stress and phrasing in English.
3. Inappropriate usage of prosody in different situations.

Although pronunciation training systems have been in existence for years with the advent of speech recognition techniques [28-30], prosody evaluation demands further investigations and receives increased attention. A number of papers concerning prosody evaluation are published in recent years [25, 31-34]. In an attempt to achieve more accurate evaluation results, however, it is desirable to develop an improved evaluation method based on the combination of prosodic concepts and signal processing/machine learning methods so as to obtain more meaningful and accurate assessment results. Therefore, one aspect to be addressed in this thesis is prosody evaluation which can assess the proficiency of an English learner in terms of the nativeness of prosody by incorporating linguistic concepts.

1.3.2 Feedback Utterances for Language Learning

Feedback information functions as an important part of a CALL system for speaking skills training as well. The machine score can only provide the learner with a general understanding of his overall performance, giving no information to the learner on how to improve his speaking skills. Although some systems like [35] also highlight the mispronounced components in a learner's utterance, more information is still required to guide the learner in the practice stage.

One way is to show the diagram of the human pronunciation system as well as the movements of tongue, teeth and other related organs [36]. This visual feedback can work in certain situations, but it may be difficult for some learners to actively control the movements of their organs appropriately according to the provided demonstration. The alternative method, which is quite popular not only for CALL but also in the real world tutoring process, is to ask the learner to listen to the sentence produced by a native speaker and reproduce it. This kind of feedback is mainly based on the auditory system and it enables the learner to listen and practice repeatedly to improve his speaking skills.

The conventional approach for speaking skills training requires the learner to repeat a sentence after listening to the utterance from a native speaker. However, it is proposed in [37] that a native speaker who possesses the voice with the highest similarity to the learner would be the best in offering the most appropriate feedback to the L2 learner, enabling him to concentrate on the pronunciation and prosody issues. Unfortunately, even if voices from a number of different speakers are recorded in the system, it is difficult to guarantee that an appropriate pattern for every user can be included. Therefore, a number of papers [38-41] have argued that it is beneficial for language learners to listen to their own accent-corrected voices rather than follow the teacher's utterances. Those methods to obtain the learner's own accent-reduced speech as feedback for language learning purposes are called "accent reduction" or "accent conversion". For consistency, we will use the term "accent reduction" in this thesis.

As the proposed feedback is the combination of the correct prosody or pronunciation of the teacher and the vocal features (in the scope of this thesis, vocal feature refers to features related to speaker identity) of the learner, an alternative method to achieve the desired feedback is to convert the vocal features of the teacher's speech to those of the learner by using a voice conversion (VC) [42, 43] method. Although voice conversion techniques have been in existence for years, such an application has not been proposed in the past. Therefore, it is desirable to study the generation of feedback utterances using voice conversion techniques.

An illustration of different feedback utterances is given in Figure 1.3(a) shows the traditional feedback utterances which are just utterances pronounced by a native speaker directly. Figure 1.3(b) shows the feedback utterances generated by an accent reduction method. This kind of feedback modifies the linguistic gestures, i.e., features related to linguistic contents such as prosody and pronunciation, of the learner's utterances, so that the re-synthesized feedback utterances are more native-like while preserving the learner's identity, simulating the utterances given by a native speaker whose vocal features are highly similar to that of the learner. The target of Figure 1.3(c) is similar to that of Figure 1.3(b), i.e., generating utterances with the learner's identity and the teacher's linguistic

Figure 1.3 Feedback Utterances to Non-native English Learners

gestures. However, it starts from the teacher's utterances and uses voice conversion techniques to change the vocal features rather than linguistic gestures.

Accent reduction refers to the modification of accented speech so that the processed utterances have less deviation from a reference norm in terms of prosodic and/or segmental features and sound more native-like. In our case, the reference norm is standard American English utterances produced by native speakers. Generation of feedback utterances based on voice conversion is a new application in CALL. Therefore, a series of experiments are needed to evaluate the performance and applicability of this kind of methods for feedback generation. Comparison between feedback utterances given by voice conversion and accent reduction methods should also be conducted to decide situations in

which application of a specific method is appropriate as well as the effectiveness of the synthesized feedback utterances.

1.3.3 Phonetic Segmentation

The labeling of speech signals plays an important role in many speech processing areas. Labeling at the phone level is one of the most popular ways of segmentation, as phone is the smallest unit of speech utterance with which other units (e.g., syllable, word, etc.) can be constructed. The time labeling of the basic unit of speech utterances, i.e., phonetic segmentation, can provide valuable cues for a number of speech processing applications which can potentially be used for advanced CALL systems. In order to perform accurate speech recognition, speech synthesis, or other applications, an appropriate and informative speech corpus is indispensable. To facilitate the development of speech tools, the corpus must provide information about the content (labeling) as well as the time alignment (segmentation) of speech signals [44, 45].

One of the well-known topics related to phonetic segmentation is the concatenative text-to-speech (TTS) synthesis [46, 47]. It requires accurate time labeling to generate segmented speech units for synthesis. The TTS system may be used for the purpose of CALL to provide desired stimuli conveniently. In addition, it has been shown that segmentation results are linked to speech recognition processes as highlighted in [48, 49]. These studies indicate that poor segmentation can lead to increased errors in a context-dependent phone recognition system. Hence, refining the segmentation scheme appropriately can generate more accurate identification of phone boundaries and improve the recognition results. The enhanced phone recognition systems can enhance pronunciation evaluation, which is one of the important topics of CALL.

Accurate phonetic segmentation can also facilitate the accent reduction process by providing phone boundaries for the mapping between non-native and native prosodic features. Different from voice conversion which can be performed without transcriptions [42], accent reduction requires accurate time alignment to enable the modification of prosodic features such as pitch and duration of the non-native speech at the phone-level. Considering the wide

applications of phonetic segmentation, an appropriate scheme must be developed to obtain reliable phone boundaries.

One example of phonetic segmentation, which identifies the onset and end of each phone of an utterance, is shown in Figure 1.4, with both the spectrogram and its transcription displayed. Phonetic segmentation can be done either manually or automatically. Manual segmentation refers to the process in which the trained linguistic experts identify phone boundaries using audio and visual information about the speech signal. Human labelers can first observe the waveform or spectrogram to get the “coarse segmentation” in their mind. Then, they can zoom into the region around that coarse segmentation and listen to the speech signal in detail to find the accurate phone boundary along with visual information. For instance, the onset of voicing can be identified from the appearance of periodic glottal pulses in the waveform or spectrogram as well as the increase of energy level due to airflow pushed from the oral tract by the raising of the velum. Although the accuracy of manual segmentation is highest, several drawbacks prevent it from applications in a CALL system and result in the demand for a reliable automatic phonetic segmentation system.

First, manual segmentations can introduce subjectivities from the labeler and

Figure 1.4 Phonetic Segmentation

may not be able to maintain the consistency [50, 51]. It means that the results generated by different labelers can vary somewhat, making it difficult to compare manually obtained results. Furthermore, the segmentation results given by the same human labeler for the same samples may also vary due to the change of concentration, perspective or even mood of the labeler across time.

Second, manual segmentation is very labour-intensive and time-consuming as discussed in [52, 53]. As a result, automatic segmentation is needed to facilitate the training process of a concatenation-based TTS system. This is because one utterance with a length of several seconds usually contains a lot of phones and each of them demands efforts from the labeler to delimit its boundaries. Normally, it could take up to 400 times real time [54] or 30 seconds per phone [55]. The compensation required to employ human experts for segmentations is also very high.

Finally, the most important problem which prevents manual segmentation to be applied in a CALL system is that manual segmentation cannot be done in a real-time manner. As CALL applications using phonetic segmentations always require a real-time or online response, the manual approach cannot be used under such a scenario.

Automatic segmentation systems can overcome all of the disadvantages mentioned above. The trained segmentation models can maintain the consistency of generated phone boundaries. Further, it possesses a much higher efficiency and can be implemented in real time. As phone is the lowest level from the linguistic perspective, segmentations based on the other phonetic units such as syllable or word can also be obtained from the phone level segmentations and the transcriptions of the speech signals. Due to all the benefits and potential applications, an advanced phonetic segmentation system is necessary and desirable for the enhancement of current CALL systems.

1.4 Major Contributions

This thesis addresses three aspects of a CALL system for speaking skills training, i.e., prosody evaluation, feedback utterances generation, and phonetic segmentation. The main contributions of the thesis are listed below. More

elaborate discussions on the contributions are presented at the end of each of the following chapters.

1. Prosody Evaluation:

- A new ‘quasi-foot’ segmentation scheme is proposed to give an improved modeling of prosodic features. An algorithm based on word-level pitch accent detection is designed to identify prosodic units with reasonable accuracy.
- Text-independent prosody evaluation is implemented based on this new segmentation scheme to provide learners with increased flexibility.

2. Feedback Utterance Generation:

- A multi-corpora accent reduction is experimented and the effects of various factors, such as nationality and corpus, in relation to the produced feedback utterances are examined.
- Three different synthesis methods are involved for the purpose of accent reduction. Experiments are performed to identify the most appropriate method in terms of various criteria like acoustic quality and nativeness.
- Voice conversion is applied for the generation of feedback utterances. A multi-stage learning scheme combining accent reduction and voice conversion methods is proposed.

3. Phonetic Segmentation:

- A statistical correction based on relative ratio is used to refine phone boundaries, with state-selection performed to identify the most appropriate search range of each phone boundary class. Multi-resolution fusion is proposed by taking into account of HMMs with different time resolutions.
- A comprehensive refinement scheme which appropriately combines statistical correction, multi-resolution fusion, and predictive model is proposed to improve the phonetic segmentation results.

- The refinement scheme is experimented under a new cross-corpora scenario which conducts training and testing on two different corpora to investigate the possibilities of segmenting a new corpus with limited labeled data.

1.5 Organization of the Thesis

The organization of this thesis is as follows:

In Chapter 2, a literature review of the three topics mentioned above is given. Previous studies are presented to provide an overview in these areas.

Chapter 3 discusses prosody evaluation based on prosodic unit segmentation. The incorporated linguistic concepts and reasons for using such concepts are introduced and discussed. A scheme is proposed to detect the prosodic unit automatically. The proposed segmentation is employed in both reference-dependent and reference-independent prosody evaluation systems for experiments. Several different segmentation methods are applied and compared. Different measurements based on accented word ratio and a feature selection experiment are incorporated to assess the accuracy and robustness of the proposed scheme.

In Chapter 4, a detailed discussion on the generation of feedback utterances is provided. Both accent reduction and voice conversion methods are applied to provide English learners with audio feedback by maintaining their own vocal features and quality. To study the effects of external factors on the overall performance of an accent reduction system, multi-corpora experiments which involve a number of different non-native English learners are carried out to assess the quality of feedback utterances produced by students with different nationalities using different corpora. In addition, Feedback utterances generated by accent reduction using three different speech synthesis methods and voice conversion are discussed and compared in terms of a number of measures.

Automatic phonetic segmentation is studied in Chapter 5. A multi-stage refinement process is proposed to improve the accuracy of phonetic segmentation. With the preliminary phone boundaries generated by forced alignment, the statistical correction method is first applied to correct the

systematic biases of the acoustic models of phone boundaries. After that, a multi-resolution fusion step is performed to further improve the segmentation accuracy by considering the effects of time resolutions. Finally, predictive models taking into account a number of acoustic features around the preliminary phone boundaries are used to refine the phone boundaries. Assessments in terms of various criteria are performed and analyzed to test this segmentation scheme. In addition to experimenting on a single corpus, a cross-corpora segmentation scenario is also studied for evaluating the effectiveness of the refinement scheme.

Chapter 6 summarizes all the previous chapters to fulfill the requirement of a comprehensive CALL system. The main contents of each chapter are reviewed and summarized in detail. Several prospective research directions are also proposed and discussed to facilitate future studies.

Chapter 2 Literature Review

In this chapter, a literature review of the three topics relevant to CALL as discussed in the previous chapter is presented. The basic structures and relevant techniques of recent prosody evaluation systems are described. The generation of feedback utterances for non-native English learners is discussed and the past endeavors on accent reduction methods are studied. Different refinement schemes for automatic phonetic segmentation are also reviewed.

2.1 Prosody Evaluation

Although there is no standard taxonomy of prosody evaluation systems in the literature, a general classification can be made according to the system's structure and reliance on reference speech utterances in the evaluation process. Therefore, prosody evaluation schemes can be categorized into two groups: reference-dependent scheme and reference-independent scheme.

2.1.1 Reference-dependent Evaluation

The reference-dependent prosody evaluation assesses a student's prosody based on reference utterances. It provides the machine score of a learner's prosody using pre-recorded reference speech utterances. A reference-dependent evaluation system mainly consists of a reference corpus pronounced by native speakers, a segmentation method to segment the speech utterance into smaller units, a feature extraction algorithm to obtain feature vectors, and a distance measurement scheme to calculate the distance between the learner's prosody feature vector and that of the reference speaker. Detailed descriptions of such kind of systems can be found in [25, 31, 56].

The basic flowchart of the reference-dependent prosody evaluation scheme is given in Figure 2.1. It is clear that both the learner's utterance and the reference utterance are processed in the same way in the above scheme. Therefore, reference-dependent evaluation is a kind of "symmetrical" system as the processing procedures are the same for both the reference and the learner's utterance. Since the distance measurement mainly relies on the reference

Figure 2.1 Basic Flowchart of Reference-dependent Evaluation

utterance pronounced by native speakers, the construction of a reference corpus is an important aspect of the system design. The reference utterances should be pronounced by native speakers without regional accents, e.g., by radio news broadcasters. The distance measurement is performed by dynamic time warping (DTW) [57], so that two feature vectors with different lengths and shapes can be compared appropriately. This process is demonstrated in Figure 2.2. The similarity matrix between feature vectors from two parallel speech files with different time durations is shown, with the darkness indicating the value of cost, i.e., difference between each pair of frames of the two vectors. The two vectors are aligned appropriately by the dash line which denotes the correspondence between them. It can be seen that the dash line goes through the light parts of the similarity matrix and thus obtains the lowest-cost path to align the two vectors appropriately. More details of DTW will be given in Chapter 3.

With DTW alignment, the baseline reference-dependent evaluation system first segments the speech signal into individual words and then measures the distances between two feature vectors to obtain the evaluation score:

$$W_j = \sum_{n=1}^{N_j} D(s_j(n), t_j(n)) \quad (2.1)$$

$$S = \sum_{j=1}^J W_j \quad (2.2)$$

where W_j is the j -th word, $s_j(n)$ and $t_j(n)$ indicate the n -th feature vector of j -th word from the student and the teacher respectively, D is the distance calculated via DTW, and S indicates the sentence level score. A number of

Figure 2.2 DTW Alignment between Two Parallel Speech Utterances

studies relevant to reference-dependent prosody evaluation have been performed in [25, 31, 32, 34, 58, 59], which are discussed below.

In [31], a prosody evaluation system used for evaluating the intonation score based on multiple reference sentences is proposed. It focuses on how to deal with the existence of multiple scores in the evaluation process. Since reference-dependent method evaluates the machine score by calculating the distance between two prosodic contours, it is possible to include multiple reference utterances for each sentence. In such a case, the distance measurements obtained by comparing the learner's utterance with each of the reference utterances should be combined appropriately to yield a reliable evaluation result. A total of four methods are experimented to combine the obtained measurements. These methods include: Minimum, choosing the best score as the final score; Maximum, choosing the worst score as the final score; Median, choosing the median of the all scores; Regression, linearly combining all the scores by some weighting coefficients. It shows that the regression method can achieve the highest human-machine correlation.

Studies presented in [34, 58] pertain to a prosody evaluation system with a structure similar to that in [31] and propose a parameter called “word important

factor”. A number of different factors obtained by a decision tree are used to weigh different words. The detailed algorithm for designing the decision tree can be found in [34]. Questions related to the characteristics of the word such as its position in the sentence and part of speech (POS) information are used to determine to which class the word belongs, and the word important factor is then calculated for each class. The evaluation scores are calculated at the word level, with the sentence level score obtained by the weighted summation of word level scores.

In [59], Mahalanobis rather than Euclidean distance is applied for distance measurement. Unlike Euclidean distance which measures the difference between a pair of points with the same weight, Mahalanobis distance takes into account the correlation of two vectors. The Mahalanobis distance between two sequences s and t with covariance matrix Σ is calculated as follows:

$$D(s, t, \Sigma) = \sqrt{(s - t)^T \Sigma^{-1} (s - t)} \quad (2.3)$$

where D is the Mahalanobis distance. Using the covariance matrix in distance calculation, Mahalanobis distance introduces a weight to each dimension of features by considering the correlation information. Therefore, the distance measurement process can be more reliable. In addition, this paper also presents the idea which uses synthesized rather than recorded natural speech as the reference for prosody evaluation. The human-machine correlation of the proposed reference pattern is comparable to that obtained via natural speech. Although experiments are only performed on Japanese accented English speech, this work shows the potential of using a speech synthesis system to evaluate prosody under the reference-dependent scheme.

In [25], a prosody evaluation system using Mel frequency cepstral coefficients (MFCCs) for alignment is proposed as shown in Figure 2.3. In this system, the alignment process is done by using MFCCs rather than prosodic features. DTW is used to align two sequences of MFCCs and the similarity estimation between two utterances is then obtained by comparing the two sets of intonation feature vectors based on the obtained alignments. Therefore, the distance measurement process combines both MFCCs and intonation features,

Figure 2.3 MFCC Based Alignment

with the former for alignment and the latter for distance measurement. After obtaining the alignment information using MFCCs, the pitch score is calculated by measuring the correlation between the learner's pitch contour and that of the reference speech, while the stress score is obtained by comparing the learner's energy contour with that of the reference speech. Subsequently, the pitch scores and stress scores are merged together by a linear function to provide the assessment result.

Tone and Break Index (TOBI) [60] is used to model the intonation contour in [32] and some other papers which will be mentioned in the next section. It is a set of conventions for transcribing and annotating prosody. In TOBI, "H" and "L" indicate the rise or fall of intonation contour, "*" indicates the pitch accent which marks the stressed syllable of specific words for certain semantic effects, "%" indicates boundary tone which is the end of the phrase, and "-" indicates phrase tone which is the interval between the last pitch accent and the final boundary tone. The combination of these symbols is then used to denote the prosody of a specific utterance.

In [32], a different intonation measurement scheme is used. Four types of intonation contours are used: H*L, L*H, H*LH and L*HL, which can be considered as "falling", "rising", "falling followed by rising" and "rising followed by falling", as showing in Figure 2.4. If the intonation types of the

H*L	L*H	L*HL	H*LH

Figure 2.4 Intonation Contour Types

learner's and the reference speech utterances are the same, a distance of "0" is assigned to the syllable. If one of the intonation types is H*LH and the other is L*HL, a distance of "2" is assigned. In all the other cases, a distance of "1" is assigned. The evaluation is done at the syllable level and the syllable level scores are combined to obtain the sentence score. However, rather than measuring distances directly, the intonation contour of each syllable is divided into four types mentioned before and the score is generated according to the discrete score of "0", "1" and "2" obtained from the comparison of intonation types of the learner's and the reference speech utterances in each syllable.

The study in [56] also assesses the prosody of a student by comparing the intonation contour of the learner with that of the reference utterances. In addition, it designs a number of tasks to test the prosody evaluation system from different aspects, including "lexical stress task" (asking the student to follow a non-sense word), "emphatic stress task" (asking the student to repeat an utterance with a specific word stressed), "affect task" (asking the student to say a fixed phrase with different emotions), etc. Results show that all of these tasks can be accomplished by the prosody evaluation system to a certain degree, and pitch is identified as a very important factor in the evaluation process.

2.1.2 Reference-independent Evaluation

The reference-independent prosody evaluation, unlike the reference-dependent one, estimates the machine score of prosody based on a pre-trained model which can assess the nativeness of prosody given relevant features of the input utterance without referring to the pre-recorded reference utterances.

Compared to reference-dependent method, the reference-independent method consists of three steps: (1) segmentation and feature extraction which play the same roles as those in the reference-dependent method, (2) training of the

Figure 2.5 Basic Flowchart of Reference-independent Evaluation

evaluation model which is used to estimate the machine score, and (3) estimation process which generates the machine score given a set of input feature vectors. The flowchart of the reference-independent method is given in Figure 2.5. The model is first trained based on the training utterances and their corresponding subjective scores given by human evaluators. After obtaining parameters of the evaluation model, the evaluation process can be performed on the extracted features of input speech pronounced by a learner. Since it is based on machine learning and pattern recognition techniques rather than distance measurement, the learner can pronounce any sentences and obtain the machine score as feedback. In contrast, in the reference-dependent method, the pronunciations of learners are limited to a pre-designed transcription set. Inclusion of any new transcriptions requires the addition of new reference utterances in the reference corpus so that the distance measurement can be performed. Therefore, the use of reference-independent evaluation method allows the learners to have increased flexibility which results in a more stimulating experience.

Unlike reference-dependent methods which basically apply distance measurement to evaluate the prosody, the evaluation models used in reference-independent method may vary. Some studies related to reference-independent methods have been reported in [61-66]. They use either hidden Markov model (HMM) or regression methods for prosody evaluation.

In [61-63], HMM is used to evaluate the prosody of the utterances from English learners. This idea originates from tone recognition in Mandarin speech recognition. Unlike other stress-timed languages such as English, Mandarin always uses tones for differentiating lexical meanings of words. For example, the

pronunciation of “ma” in Mandarin may mean “mother” (flat tone), “dull” (rising tone), and “horse” (falling-rising tone). Therefore, tone recognition is used in Mandarin speech recognition. In HMM based speech recognition for Mandarin, a series of tone models are constructed to recognize the tone of each syllable. In HMM based prosody evaluation system, these schemes are applied to model intonational units in English and evaluate the learner’s prosody.

The HMM based prosody evaluation system in [61-63] includes three steps. First, the HMMs are trained to model a series of intonational units. TOBI symbols as given in the previous section are used as different models to recognize intonational units. Again, the difference between tone and intonation is that intonation conveys no lexical meanings of the utterance. Second, for each utterance, the recognition process is performed to recognize the sequence of intonational units present in the utterance. The output can be a sequence of labels like HL*HL%. Finally, the confidence score is calculated based on the posterior probability of each label to indicate the nativeness of the learner’s prosody. Overall, this kind of method follows the similar steps used in pronunciation evaluation systems, with MFCCs replaced by prosodic features and acoustic models replaced by intonation models. The intonation recognition model is similar to that for speech recognition, i.e., :

$$P(M_t|O) = \frac{P(O_t|M_t)P(M_t)}{\sum_n P(O_t|M_n)P(M_n)} \quad (2.4)$$

$$S = \frac{1}{T} \sum_{t=1}^T P(M_t|O) \quad (2.5)$$

where O is the speech observation in prosodic features, M_t is the recognized model and corresponds to one of the TOBI intonation label, $P(M_t)$ denotes the prior probability, T is the number of intonation labels, and S is the sentence level score.

The intonation recognition grammar is shown in Figure 2.6. The upper one is the simpler intonational grammar which only recognizes the pitch accent “*” and boundary tone “%”. It starts with a silence, and after one pitch accent, can go back to the previous state to handle the situation that multiple pitch accents exist in a sentence. The first “%” of this sentence represents the initial boundary,

Figure 2.6 Intonation Recognition Grammar

which is also modeled by an HMM. The lower diagram shows the more detailed grammar which includes the low and high pitch accents or boundary tones.

In [64], a similar but more complex evaluation scheme is proposed. To evaluate the prosody, the HMM based model is used, with additional models included for intonation recognition. In addition to units like H^* , L^* , $H\%$, and $L\%$ mentioned in [63], HMMs are also trained for H- and L- to model phrase tones which indicate the intervals between the last pitch accent and the final boundary tone. Besides, since prosodic features are suprasegmental and affected by surrounding intonations, the posterior probability of observations should also relies on the previous state. As a result, a bigram model is proposed:

$$P(M_t|O, M_{t-1}) = \frac{P(O_t|M_t)P(M_t|M_{t-1})}{\sum_n P(O_t|M_n)P(M_n|M_{t-1})} \quad (2.6)$$

where O is the speech observation in terms of prosodic features, M_t is the recognized model or phone intonation label, $P(M_t|M_{t-1})$ is the bigram probability given the previous model, and n is the number of HMMs. Experimental results in [64] demonstrate the superiority of this bigram scheme in terms of the correlation coefficient with subjective scores given by human evaluators.

In contrast to the approach discussed above, a different regression based evaluation model is proposed recently as in [65, 66]. In [65], prosody evaluation is achieved using a regression technique. Since human evaluation of prosody is to give a score according to the input feature vectors extracted from the learner's

utterance, the machine score can be obtained by building a mapping from the prosodic features to human scores while minimizing the mapping errors. In this paper, support vector regression (SVR) is used to generate the machine score based on the input prosodic features. A number of features which describe various properties about the pitch and energy contours like their positions, ranges and slopes are involved. The evaluation model is then constructed by the input features and the corresponding human scores to assess the prosody.

A similar scheme is also applied in [66]. It involves a subjective test which asks a group of experienced labelers to identify several sentences which are most prone to prosody errors for L2 speakers (from the ISLE corpus). These sentences are then used as the transcriptions of the prosody evaluation experiment. Features similar to [65] are involved and cross-validated forward feature selection (CV-FS) is used to add features sequentially based on the performance of the prosody evaluation system. Different from the approach adopted in [65], linear multiple regression rather than SVR is trained in this scheme to assess the prosody.

Another study [67] uses the same evaluation scheme but experiments on three different groups of labelers, i.e., experts, phoneticians, and naïve labelers who do not receive formal training of the labeling process. It is demonstrated that subjective scores from naïve labelers can be used to train the prosody evaluation system if a sufficient number of labelers are involved. This observation shows that the efforts and compensations for finding qualified labelers can be saved if a large group of naïve labelers are available.

Compared to the HMM based evaluation, the regression method presented in [65, 66] does not require TOBI labeling for training purposes, which can be difficult or at least time-consuming to obtain. Although prosody evaluation using regression methods is an effective and convenient approach, current studies only take into account lexical units for the segmentation of sentences. It is also noted that experiments reported in [65] only involve German and Japanese, which are very different from English in terms of prosodic structures. Therefore, it is worthy to study the potential of reference-independent evaluation along this direction.

2.2 Feedback Utterances Generation

In a CALL system, feedback utterances can be generated in various ways to provide the learner with informative cues for the process of speaking skills training. To generate effective feedback utterances, the characteristics of the traditional and alternative feedback are first discussed and recent studies on the generation of feedback utterances are then reviewed.

2.2.1 Feedback Utterances in CALL Systems

Traditional speaking skills training in classroom, though widely used by most teachers and learners, still has some drawbacks which include inflexibility of learning materials and constraints on time and location [1, 2]. To address the increasing demands from English language learners, CALL systems play an important role in the acquisition of good speaking skills. Considerable efforts have been made in this area and a number of software has been developed [35, 68-70]. The traditional speaking skills training systems still focus on the evaluation of the learner's speech and provide the learner with assessment score for their reference and practice. Both pronunciation evaluation [30, 71] and prosody evaluation [34, 58, 65] are studied in the development of computer-aided speech evaluation systems. However, one problem of these CALL systems using only evaluation scheme is the limited accuracy as all the systems presented in the literature can only achieve a human-machine correlation from 0.4 to 0.8, which is bounded at the upper end by the correlation across human scores. To address this issue, some systems alternatively detect the problematic phone or phrase and highlight those part for the learner to repeat and practice [35]. The limitation of such a system is that the learner may fail to make improvement upon achieving certain benchmarks without further feedback [29]. Besides, studies in [72] compare the influence of different feedback on foreign accent reduction, showing that aural feedback is the most important to learners in their learning process. Therefore, it is essential to create appropriate aural feedback, in addition to numerical scores, for the learners to practice and reduce their foreign accents.

A popular way to give feedback for speaking skills training in current CALL software is to ask the learner to repeat the sentence after a native speaker's utterance is played to him (e.g., ISLE [68], CASTLE [70] and The CUHK Experience [35]). However, there are two issues with such a scheme. First, the dissimilarity between the vocal features of the learner and that of the native speaker may reduce the learning efficiency as illustrated in [37]. As the vocal features between the two speakers can be quite different, they may distract the information to be conveyed by the feedback utterances. It means that the learner may be distracted by the vocal features of the teacher when listening to the feedback utterances, and hence fails to focus on the differences in his prosody and pronunciation from those of the teacher.

Secondly, as illustrated in a linguistic study [73], a system which provides the learner with reference utterances by considering his English proficiency is more effective. This study proposes that an adaptive feedback which takes into account the changes of the learner's performance can be more effective, because different feedback utterances are desirable in different learning stages. When the learners' speaking skills are poor, following a very native speech may sometimes frustrate them, and thus lowering their interest in the learning process. Further, using only the teacher's utterances as the feedback may not meet the preferences of different English learners. For example, some English learners may prefer to focus on prosodic issues while other students may prefer to concentrate on pronunciation problems, depending on their backgrounds and learning experiences. Therefore, there should be some intermediary feedback utterances to provide the English learners with more choices in the learning process.

Some research also proposes assistant conversions of speech utterances to obtain more informative audio feedback for the learner, such as CASTLE [70] which extends the duration and enlarges the pitch variation of native speech to help English learners or WinPitch LTL [74] which allows manual manipulations of prosody. Exaggeration of the f_0 (fundamental frequency of the speech signal) differences between the teacher and the learner as well as synthesis of a third-person speech as audio feedback are also applied in [75]. These investigations,

however, do not correctly reduce the perceived accents in the learner's speech utterances.

Compared to the traditional way of feedback which is simply the utterance from a native speaker, a different reference utterance converted from the learner's own voice seems to be a better choice. According to [37, 73], a "golden speaker" who can consider the learner's English proficiency and speaking characteristics can offer the most appropriate feedback to L2 learners. The "golden speaker", as proposed in [37], possesses the voice with the highest similarity with the learner, thus enabling the learner to focus on the pronunciation and prosody issues. However, the inclusion of a "golden speaker" cannot be achieved easily, because a large number of native speakers must be recorded in the corpus so that there can be some utterances with vocal features similar to those of the learner. Moreover, even if a group of speakers are included in the system, it can hardly be guaranteed that a "golden speaker" for any user can be picked from the group. For example, even when the recorded utterances include those from native speakers of different ages and genders, it is likely that some learners may have special vocal features and therefore no perfect "golden speaker" can be found using pre-recorded utterances.

To overcome this problem, studies in [38-41, 76] suggest that language learners can listen to their own accent-corrected voices in order to achieve better training results. It means that the learner's own utterances are processed to become more native-like without changing the speaker identity of the learner. As a person is the most appropriate "golden speaker" for himself, the accent-corrected utterance of a learner whose vocal features are preserved in the correction process can be used as an effective feedback for the learner to imitate and practice. Specifically, subjective evaluation results presented in [39, 41] demonstrate the reduction of perceived accent after appropriate modifications of non-native speech. Further, pedagogical study in [40] suggests that prosody-corrected speech of the learner is a more effective stimuli for L2 learners compared with pre-recorded native speech. Therefore, it is desirable to correct the learner's utterances and use the corrected utterances as feedback.

2.2.2 Differences between Voice Conversion and Accent Reduction

Accent reduction shares some properties with voice conversion because both of them involve transformations of speech signals. However, there are two main differences between them.

The first difference lies in their objective: unlike voice conversion which transforms the vocal features of the source speaker (the one who pronounces the original utterance to be converted) to that of the target speaker (the one with the desired vocal features), accent reduction only reduces the accent of the non-native speaker to a more native one, while leaving the vocal features unchanged. As shown in Figure 2.7, voice conversion changes the vocal features of utterances. The objective is to modify the speech signal so that the output is perceived as if it were pronounced by the target speaker rather than the source speaker. In contrast, accent-reduced speech signals have the same vocal features as the input, but the perceived accents are reduced. Therefore, the objectives of the two schemes are related to vocal features and perceived accents, respectively.

The second difference stems from the dependence on linguistic information. As voice conversion only concerns the conversion of speaker identity features, the linguistic information can be separately treated to a certain extent and therefore it is possible to implement cross-lingual voice conversion under some

(a) Voice Conversion Process – Vocal features Changed

(b) Accent Reduction Process – Perceived Accent Changed

Figure 2.7 Comparison between Voice Conversion and Accent Reduction

scenarios [77]. For accent reduction, however, the linguistic information is very important and alignment information is always required, because the transformation is performed on the linguistic gestures rather than the vocal features to reduce perceived foreign accents.

These unique features of accent reduction make it stand as an independent component of speech processing techniques and it certainly deserve further studies. Recent literature about accent reduction can mainly be grouped into two categories: rule-based reduction and reference-based reduction.

2.2.3 Rule-Based Accent Reduction

Rule-based accent reduction relies on a thorough study of a group of accents and involves modifications of those features which make different accents distinguishable, to convert or reduce the perceived accents. With this kind of accent reduction methods, the linguistic and phonetic differences between two different accents are studied by analyzing two large groups of utterances belonging to each accent. As the differences across accents characterize the features used to identify or distinguish them, some specific rules can be developed to change the perceived accents by adjusting parameters correlated to these differences. A general flowchart of this kind of system is shown in Figure 2.8.

Figure 2.8 Rule-based Accent Reduction

Some studies related to rule-based accent reduction have been done in [39, 78-81]. In [78], a group of stimuli from Japanese L2 learners of French are recorded and manipulated using phonetic analysis software *PRAAT* [82] for f0 contour and duration analysis. Pitch-synchronous overlaps and adds (PSOLA) [83] method is used to modify prosodic features. Experimental subjects are required to listen to short French phrases synthesized with French and Japanese segments in combination with duration and f0 given by Japanese learners and French native speakers. Different kinds of short phrases are played to the subjects: 1) phrases synthesized using standard and Canadian French utterances in combination with duration and f0 produced by Japanese and French native speakers, 2) phrases read by Japanese learners and French native speakers which are re-synthesized by manipulating local duration and f0. The final lengthening and continuing rise of f0 lacking in Japanese L2 Learners are added, and subjective tests performed on this modification indicate reduced perceptions of accent.

The study in [79] is based on formant space comparison using a formant tracker developed based on HMMs and linear predictive coefficients (LPCs). The first step, which is the rule-developing phase, of this study is to compare the formant spaces of three different regional accents of English, i.e., American, British and Australian English. The formant spaces of each vowel under different accents are plotted and the characteristics of each kind of accent are summarized. A 2-D HMM scheme is applied to track the formants of all the three accents. According to the analysis, the formants of vowels are distinct among those different English accents. For example, American vowels display higher values of F2, i.e., the second formant frequency (similar symbols applied for other formant frequencies), compared to Australian English, whereas the F3 and F5 are consistently higher in the Australian English in comparison with British English. Further, lower F3 and F4 are detected in the utterances of American males, accounting for the relatively rare case of the vowels followed by “r” in American English.

All of these observations provide the guidelines for rule-based accent conversion. The second step is the conversion stage which modifies the formant

spaces of different accents based on the developed rules to achieve the transformation of perceived English accents. Formants are modeled by LPC analysis and modified by frequency warping and spectral shaping of the spectrum. The ABX tests, i.e., asking the subject to determine whether the transformed speech X is closer to target accent B or source accent A, performed in this study demonstrate that more than 75% of converted speech utterances are perceived as the desired accent. Related studies are also performed in [80] with the same formant tracker, showing that these different English regional accents have distinguished spectral features which can be potentially used for accent reduction purposes. In this paper, cross-entropy measurements are used to quantify the differences among formant spaces as well as cepstra of regional dialects of English. A clustering using a phonetic tree based on cross-entropy also demonstrates dissimilarities of different English accents in terms of formant spaces.

An accent reduction scheme is also employed among the same three regional accents of English in [81]. However, it is based on the modification of prosodic features. Rather than transforming spectral information, the duration and intonation contours of the three dialects are studied and transformation rules for prosody-related accent reduction are developed. It is shown that British speakers demonstrate the largest pitch range among all the three accents. British speakers prefer to use low-rise tone in non-final intonation groups whereas American and Australian speakers prefer to use high-rise tone. British speakers also incline to have a steeper pitch contour at the end of the sentence. According to all of those observations, pitch and duration modifications are performed using PSOLA synthesis methods to transform the perceived accents of the input speech. The ABX tests demonstrate that more than 70% of the converted stimuli are perceived as the target accents. Reviewing what have been done in [79, 81], it is noted that both segmental features, e.g., formant and spectrum, and supra-segmental features, e.g., duration and intonation, affect the perceived accents.

In [39], the tone sequence model (TSM) is used for intonation modeling of American L2 learners of German. Similar to TOBI as mentioned before, it models an f_0 contour as a series of discrete labels. These labels can be used to

define particular f_0 configurations. The speech utterances pronounced by non-native learners are re-generated by modifying the intonation according to a series of linguistic rules specified for both American English and German (e.g., the comparatively lower f_0 value at the beginning of German speech and the less steep falling of the f_0 contour at the stressed syllables of American English). Experiments show that the reconstructed German speech produced by American L2 learners follows the German intonational rules and thus has less perceived accents.

2.2.4 Reference-Based Accent Reduction

Different from rule-based accent reduction, the reference-based accent reduction performs the transformation according to parallel speech utterances from the learner and the teacher, as demonstrated in [38, 40, 41, 76, 84]. The flowchart of reference-based accent reduction is shown in Figure 2.9. Parallel speech signals from the source (accented) and target (native) corpus are first parameterized by a synthesis model. Then, the parameters of the source speech signal are modified according to that of the target one to address the accent without changing the vocal features. Finally, the accent reduced speech signal is reconstructed by the modified parameters.

The study in [38] presents an accent reduction for prosody modification at the

Figure 2.9 Reference-based Accent Reduction

word level. The learner is asked to pronounce a number of phrases and these phrases are then modified according to the duration and pitch produced by native speakers. Improved nativeness can be observed from the accent-reduced phrases. However, this work only focuses on the word level without considering more general cases such as prosodic and segmental modifications at the sentence level.

Research in [84] generates feedback utterances to correct stress issues for the learning of Japanese. The Japanese L2 learner is first required to pronounce a number of homonym pairs (Japanese words with the same pronunciation which can only be distinguished by their written communication or pitch patterns). For instance, “hashi” can mean “chopsticks”, “bridge” or “edge” in Japanese, depending on the pitch pattern. These pairs are to be segmented into mora (basic unit of Japanese) by a forced alignment system. Then, the accent types of these pairs are assessed according to f_0 contours. In comparison to that of the teacher, it can be determined whether the f_0 contour of each homonym is correct or not. For those homonyms with stress issues, the learner’s utterances should be modified appropriately to produce the audio feedback. The pitch and duration of each phone from the learner are modified and resynthesized by PSOLA method using new pitch marks according to the parallel speech signals from the teacher. Recordings from eight non-native Japanese speakers are involved in experiments and subjective evaluations are conducted by eight native speakers. It is shown that the proposed audio feedback is effective for the language acquisition process. Similar f_0 modifications can also be found in [75], in which f_0 contours of the learner are resynthesized according to that from the teacher to provide audio feedback for English and Japanese learners.

In [40], the accent reduction process is performed to reduce the perceived accents in German utterances produced by non-native Italian speakers using the PSOLA method to change prosodic features. Different feedback utterances are provided in this paper. They include the native speech given by native German speakers, the accent reduced speech which is generated by correcting and resynthesizing prosodic features of Italian L2 learners according to parallel native German speech, and the accent reduced speech putting emphasis on stressed syllables. Over 30 subjects with various ages are involved in the

evaluation tests and each of them is asked to assess the generated stimuli. The experimental results show that feedback utterances generated by accent reduction preserve the learner's own voice and are more effective for the purpose of stress pronunciation training in comparison with pre-recorded reference utterances spoken by German native speakers.

Another study [76] performs accent reduction in a spoken language conversion system, i.e., resynthesizing a L2 language from the L1 corpus of a particular speaker, which may also be used in a speech-to-speech translation system. The perceived accents in the synthesized Japanese utterance using an English TTS system are reduced by this scheme. The input of the accent reduction system is the synthesized Japanese utterance generated by an English TTS system. Due to the different characteristics and phone sets between Japanese and English, the generated Japanese speech is not native-like. Therefore, the synthesized speech is considered as accented speech which is modified using parallel reference speech. Both segmental (spectral envelope) and prosodic features (pitch and duration) are modified to reduce the perceived accents. The prosodic modifications are performed by adjusting pitch and durations using PSOLA and the segmental modifications are performed by replacing the spectral envelope of the source speech signal by that of the corresponding frame in the target speech signal. The combination of segmental and prosodic modifications improves the intelligibility of synthesized speech from 56% to 84% based on subjective listening tests.

More thorough studies with intensive subjective tests are presented in [41, 85]. Modification schemes similar to that in [76] are performed in these work, with a pool of subjective tests performed on voice quality, accentedness and speaker identity. The speech utterances are separated into excitations, which are related to speaker identities, and spectral envelopes, which are mainly responsible for linguistic gestures. Then, the prosodic features of the learner's speech are modified according to the corresponding features of the teacher at the phone level, leading to the prosodic modification. The spectral envelopes of the non-native utterance are also replaced appropriately by that of the native one to modify segmental features. In the experiments, the non-native utterances are

modified to different extents, i.e., prosodic modification only, segmental modification only, and combined modification. Subjective assessments are performed by a group of subjects to indicate the perceived accents, acoustic quality, as well as speaker identity of the speech utterances before and after accent reduction. It is shown that segmental and prosodic features can affect the perceived accents and acoustic quality of the accent-reduced utterances differently. In addition, the speaker identity of the non-native speaker is also preserved after accent reduction. As the subjective evaluation process of accent reduced utterances is time consuming and inconsistent, studies presented in [85] propose several different objective measurements to assess the acoustic quality and perceived accents of the modified speech. Outputs given by those objective measurements are compared with subjective evaluation results, demonstrating high human-machine correlations.

Considering all the discussions above, accent reduction can produce effective feedback utterances and provide more options of stimuli to help L2 learners in the acquisition of speaking skills. However, the effects of some external factors like nationalities of learners and the corpus in use are still unclear. Further, as most current manipulations of accents are based on PSOLA, implementation of accent reduction using different synthesis methods is also of interest. The objective is to find the most appropriate way and explore more possibilities for accent reduction.

2.3 Phonetic Segmentation

Phonetic segmentation can affect a number of speech technologies, such as concatenation based TTS, speech recognition, and accent reduction. As all of these applications have potentials to be applied in a modern CALL system, it is worthwhile to study how to improve the accuracy of phonetic segmentation systems.

There are mainly two kinds of phonetic segmentation methods: text-independent segmentation, which only uses acoustic features, and text-dependent segmentation which uses both acoustic features and transcriptions of the speech signal.

2.3.1 Text-independent Segmentation

Some studies on text-independent phonetic segmentation are done in [86-88]. The proposed methods detect phone boundaries using only acoustic features. This kind of segmentation systems does not use information from the linguistic content, i.e., the transcription of the speech utterances. Rather, it uses probability models to examine the existence of phone boundaries according to specific acoustic features.

Segmentation systems in [86, 87] apply the same scheme for text-independent segmentation. They first identify a number of frames which are detected as candidate boundaries according to spectral features. Next, probability modeling of each candidate frame and its neighboring frames is conducted to examine each candidate boundary. This segmentation process is shown in Figure 2.10. Supposing frame 1, frame m , and frame n are candidate frames, all the frames, i.e., frame 1 to n , can be segmented into two segments with the boundary located between frame m and $m + 1$. The two segments are either modeled by two separate statistical models ($M1$) or one single statistical model ($M0$). Acoustic features are assumed to be normally distributed. For $M0$, it is assumed that all the n frames of data follows a normal distribution $(x_1, x_2, \dots, x_n) \sim \mathcal{N}(\mu_Z, \Sigma_Z)$, i.e., the same distribution source Z (all the frames originating from the same phone). In contrast, $M1$ assumes that frame 1 to m follows one normal distribution $(x_1, x_2, \dots, x_m) \sim \mathcal{N}(\mu_X, \Sigma_X)$, and frames $m + 1$ to n follows the other normal distribution $(x_{m+1}, x_{m+2}, \dots, x_n) \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, indicating a phone transition between the two segments. As a result, the possible phone boundary can be determined by calculating the log likelihood ratio given by the two models $M1$ and $M0$. If the likelihood given by the single statistical model $P(X|M0)$, where X is the feature vector, is greater, the two frames are more likely to fall into the same statistical distribution and thus are not likely to be separated by a phone boundary. Otherwise if $P(X|M1)$ is greater, a phone boundary is likely to exist between two frames as they fall into different statistical distributions. In this detection scheme, three candidate frames (e.g., frame 1, frame m and frame n in Figure 2.10) are examined and the process moves forward continuously so as to go through all the candidate frames. Once a

Figure 2.10 Text-independent Segmentation Purely Based on Acoustic

candidate frame is rejected as a phone boundary, it will not be considered in the next examination.

To measure $P(X|M0)$ and $P(X|M1)$, Bayesian information criterion (BIC) is used. The μ and Σ of assumed normal distributions can be calculated from the feature vectors belonging to each segment, so that BIC of $M1$ and $M0$ in Equation (2.7) can be produced to determine whether a phone boundary should exist between two segments:

$$\begin{aligned}
 \text{BIC}(M1, M0) &= \text{BIC}(M1) - \text{BIC}(M0) \\
 &= -2 \left\{ \sum_{i=1}^{N_X} \ln P(x_i | \mu_X, \Sigma_X) + \sum_{i=N_X+1}^{N_Z} \ln P(x_i | \mu_Y, \Sigma_Y) \right\} \\
 &\quad + 2 \left\{ \sum_{i=1}^{N_Z} \ln P(x_i | \mu_Z, \Sigma_Z) \right\} + \ln K
 \end{aligned} \tag{2.7}$$

where K indicates the constant term which is not relevant to the features, N_Z indicates the total number of frames considered, and N_X indicates the frame number belonging to the left segment. If this BIC result is negative, $P(X|M1)$ is greater than $P(X|M0)$ which indicate that there should be a phone transition between two segments.

Particularly, the approach proposed in [86] first calculates MFCCs of each frame and measures the spectral changes so as to identify a number of candidate

frames with peak spectral changes. Then, the process described above is used to examine each of these candidate frames to decide whether the candidate frame should be a phone boundary or not. To overcome the small size problem and outliers which may exist in the segmentation scheme, this paper also applies a selection scheme to choose the most appropriate method from a pool of models.

In [87], the microcanonical-multiscale formalism (MMF) resulted from physics field is applied to detect the candidate phonetic boundary by measuring the singularity exponential (SE) function over time. MMF function of a given signal $s(t)$ is defined by a scale-dependent function τ_r :

$$\tau_r(s(t)) = \alpha(t)r^{h(t)} + o(r^{h(t)}) \quad r \rightarrow 0 \quad (2.8)$$

where $h(t)$ is the SE function, $\alpha(t)$ is a function of t , and $o()$ indicates convergence faster than what in the bracket. It can then be estimated by choosing τ_r to be the gradient-modulus measure as discussed in [89]. The SE function relates to the scaling of local power-law behaviors in the signal domain and is almost linear inside each phone, leading to a clear change in its slope at the phone boundary. Therefore, a number of candidate phone boundaries are identified according to SE functions. Subsequently, the two statistical models $M0$ and $M1$ as mentioned above are applied on each candidate frame to eliminate those incorrect candidates and thus obtain the phonetic boundaries.

Studies in [88] proposes a biometric model of the human auditory processing scheme for phone boundary detection. To calculate the neural features of frequency synchrony as well as average signal levels, a speech signal is filtered by a head-related transfer function and divided into different components using gammatone filters. The local maxima are located by a peak-picker scheme and the synchrony information is generated between peaks. With that information, the rate of synchrony and the average signal level are calculated accordingly. After obtaining all of these auditory based features, a two-layered support vector machine (SVM) system is used to identify phone boundaries.

2.3.2 Text-dependent Segmentation

Although endeavors on text-independent phonetic segmentation may contribute to certain speech processing areas, many target applications of phonetic

segmentation are those in which linguistic information is available, such as TTS system or the training of acoustic models in speech recognition. Furthermore, text-independent phonetic segmentation methods generally underperform the text-dependent methods which utilize linguistic information. Therefore, more focuses should be given to the text-dependent segmentation method.

The current mainstream text-dependent segmentation methods are based on forced alignment [90-94]. This process first trains acoustic models using HMMs as in speech recognition systems. The acoustic models consists of a group of HMMs, with each HMM representing one phone or phone boundary and modeled by three to four states. The acoustic features, normally MFCCs plus first and second order delta, are then input to the trained acoustic models along with the phone level transcription to detect the phone boundaries. The phone level transcription can be obtained by the word level transcription and a pronunciation dictionary. Normally, the speech utterances to be segmented are assumed to be pronounced correctly, i.e., it exactly matches the given transcription. With the obtained phone-level transcription, all the HMMs of each individual phones are connected to form the state sequence which passes through the whole sentence. The posterior probability of each input frame is calculated to decide whether to transfer to the next state or stay at the current state, so that the number of frames belonging to each state is estimated. The phone duration can then be obtained by summing the frame numbers of each state. More detailed descriptions on the forced alignment process will be given in Chapter 5.

Such a baseline segmentation system using forced alignment results in reasonable segmentation accuracy in the range of 30 ms – 50 ms as shown in [91-93, 95]. However, further studies on phonetic segmentation are still necessary to improve the detection process for improving other applications, e.g., avoiding audible errors from TTS systems as in [96]. To overcome these segmentation errors, a number of techniques have been proposed to refine the phone boundaries given by the baseline system.

- **Modification of Modeling Scheme**

There are a number of studies that propose to modify the modeling scheme of segmentation systems. A large margin algorithm which uses a framework similar

to support vector machine (SVM) is proposed in [97] to do phonetic segmentation. This study casts the task of speech segmentation to a large margin problem and proposes effective solutions to this problem. A set of base functions is learned to measure the confidence for an alignment in a segmentation system, leading to performances better than the baseline HMMs method in [90]. In addition, it not only experiments on speech database, but also performs music-to-score alignment which aligns music signal to its corresponding transcriptions.

In [98], a phone boundary based segmentation model is proposed. In addition to phone HMMs, phone boundaries are also modeled with special 1-state HMMs. Each special 1-state HMM has only one emitting state, modeling the transitional property of phone boundaries. These phone boundary models are then combined with phone HMMs to perform forced alignment. State numbers used to train phone HMMs are also not identical, with 1-state, 3-state, and 5-state HMMs applied to different phones. It also demonstrates that using only isolated-unit training, i.e., training HMMs based on individual phones without embedded training, can lead to higher segmentation accuracy.

A modified HMM scheme incorporating artificial neural network (ANN) is also proposed in [99] to obtain phone boundaries. Instead of Gaussian mixture model (GMM), six ANN classifiers are used to generate phone and phone-transition probabilities. For each ANN, cepstral features plus energy-based features are taken as input, while the output indicates six kinds of phonetic features including manner, place, height and their transitions. These phonetic features are then combined by Bayes' rule in the Viterbi decoding process to generate accurate phone alignments. All of these methods can contribute to the segmentation by providing more flexible structures and improved segmentation results. However, applying such kind of methods requires a re-structuring and re-training of all the acoustic models and thus it cannot refine segmentations given by existing acoustic models conveniently. Furthermore, some acoustic modeling techniques may sacrifice the segmentation performance for a certain range of resolution to improve the overall performance. For example, the segmentation accuracy of 5 ms is reduced as reported in [99] due to the inclusion of phone transition models.

- **Statistical Correction and Fusion Methods**

Statistical correction and fusion methods have also been proposed to improve phonetic segmentation results, as in [91, 100-102]. In [91], the concept of statistical corrections on phonetic segmentation is proposed and the experimental results demonstrate a significant improvement in segmentation accuracy. It uses a scheme to calculate statistics, i.e., two ratios, for each phone boundary class to refine the preliminary phone boundaries. The correction ratios are calculated from the difference between the manual and automatic segmentations from the training corpus. This scheme takes into account the relative duration of each phone as well as the phone transcription information to correct the phone boundary dynamically. However, it neither studies the performance of this method on context independent (i.e., monophone) models nor examines the effects of the change of search ranges on statistical correction.

In [101], a fusion method is used to combine a number of different HMMs to refine the phonetic segmentation results, as demonstrated in Figure 2.11. The system first trains 112 baseline segmentation engines (BSEs), which are HMMs with various parameter settings, including no. of mixtures, no. of states, context-

Figure 2.11 Regression Fusion Methods for Phonetic Segmentation Refinement

dependent vs. context-independent, etc. In the training process, forced alignments are performed on each input speech signal to generate segmentation results using these different BSEs. These automatic segmentation results are then combined with manual segmentations to calculate the parameter P for the fusion purpose. With obtained parameters and BSEs, segmentation results for the input speech signals are first generated by each individual BSE. A fusion is then performed to combine all the time alignments from different BSEs using the obtained parameters P . The author experiments on a number of fusion methods, such as linear regression, decision tree, and support vector regression. The experimental results show that the fusion based on support vector regression yields the best results.

Fusion methods are also adopted in [100]. The difference is that they do not include all the segmentation results for each individual phone boundary. Rather, a selection process is performed to find the most appropriate segmentation results from different automatic segmentation machines (ASMs). The flowchart of this system is given in Figure 2.12. Similar to [101], a number (55 in this case) of ASMs, which are also HMM based acoustic models with different parameters, are trained. Then, segmentations are performed on the training corpus in which

Figure 2.12 Fusion Based Phonetic Segmentation Using Candidate Selection

the manual segmentations are available. Automatic segmentation results for those training speech utterances are obtained from each individual ASM. In the candidate selection process, all the observations from the same phone boundary are collected to calculate the difference between manual segmentation and automatic segmentation given by each of the ASM:

$$M(i) = \operatorname{argmin}_k \sum |S^R(i) - S_k^A(i)| \quad (2.9)$$

where S^R is the real boundary of the i -th boundary class, S_k^A is the automatic boundary generated by model k , and $M(i)$ is the model selected for i -th boundary class. The phone boundary here is defined as a combination of two neighboring phones. The difference between manual and automatic segmentation can be used to select the most appropriate ASM for each phone boundary. With the selected ASMs, the time alignment of a specific phone boundary will be given by the selected ASM, thus yielding more relevant segmentation results. This work is extended in [102], which uses the appropriate combination of a subset of all the ASMs to detect phone boundaries for each boundary class.

A prominent drawback of fusion methods is that the massive processing time required for phonetic segmentations, because segmentation results must be generated by many different HMMs (from 55 to 112) and then merged together. Also, one drawback of all the work on statistical correction and fusion method mentioned above is that most of them are not tested on a standard corpus, except for [101] which uses TIMIT corpus to test the performance.

- **Predictive Model Based Refinement**

Text-dependent phonetic segmentation results have also been refined using various predictive models as in [93, 95, 103-106]. Such kind of methods searches frames around the preliminary boundary, identifying the frame which maximizes the posterior probability obtained from feature vectors and the refinement model as the corrected final boundary:

$$t^{fin} = \operatorname{argmax}_{t_n \in (t^i - r^i, t^i + r^i)} \{P(t_n | o_n)\} \quad (2.10)$$

where t^{fin} is the corrected final boundary, t^i is the i -th preliminary phone boundary given by forced alignment, either with or without statistical correction

as stated before, r^i is the search range of i -th phone boundary, and o_n is the observed feature vector. A demonstration of this process is shown in Figure 2.13. Either classification model, which classifies frames far from or close to the real boundaries into different classes, or regression model, which estimates a continuous value for each frame, can be trained to calculate the probability score $P(t_n|o_n)$. The data required for the training process mainly involves acoustic features, manual segmentations and sometimes automatic segmentations of the training corpus. Various models, like k-nearest neighborhood (KNN), multi-layer perception (MLP), and SVM can be implemented to refine phone boundaries.

A hybrid method is used in [103] to perform feature selection using sequential forward selection and to refine the boundary using KNN according to selected features. A number of features related to the phone boundary information like pitch or entropy are sequentially added to form the feature vector. Next, each KNN is trained for one phone boundary given the manual segmentation results and the feature vectors. Speech frames in the training set are grouped into two categories, frames far from the manual boundary and frames close to the manual boundary, to train the KNN model. In the refinement process, the frame

Figure 2.13 Predictive Model Based Refinement – Searching around the Preliminary Boundary

belonging to the class close to the manual boundary with the highest probability is identified as final boundary. The most suitable feature set is then decided by the segmentation accuracy given by the KNN model. This method improves the segmentation results given by the baseline forced alignment system, but the feature selection process is time consuming.

Work presented in [107] designs a different topology of HMMs to refine the segmentation process. During this scheme, a 3-state HMM topology is used to train the boundary classifier. Those frames at the left side of the manual boundary are used to model the first state, while frames at the right side of the manual boundary are used to model the last state. The frame corresponding to the manually labeled boundary is used to model the middle state. These boundary classifiers are then applied to examine a set of candidate frames around the preliminary boundaries given by a forced alignment system. However, the presented experiments are only based on a limited number of phone pairs.

A multi-layer perceptron (MLP) based refinement process is studied in [104]. It trains a MLP for each kind of phone transition group and using the obtained models for correction. In this method, the MLP model is trained for each phone boundary class to perform a post-refinement on the results given by a forced alignment system. The frame corresponding to the manual boundary is labeled as 1 and the frames neighboring to that frame are labeled as 0.5, with the rest of frames labeled as 0. An iterative training process is performed to partition the phone boundary space so that phone transitions with similar properties are refined by the same MLP. This process is performed as follows. First, an initial partition of the phone boundary space is given and MLPs are trained to refine the phone boundaries. Second, the obtained phone boundaries are compared to the manual segmentations to find the most appropriate MLP for each phone boundary which generates the smallest distance. Third, the phone boundary space is re-partitioned to regroup each phone boundary class into a new MLP by minimizing the distance between the real segmentations and the automatic segmentations given by the MLP to which the phone boundary class belongs. Finally, the MLPs are retrained with the modified partitions. This process is repeated a number of times to segment the phone boundary space into four

categories and then one MLP model is trained for each category of phone boundaries to refine the segmentation results. It is proven in [104] that this iterative scheme can finally converge and yield improved segmentation accuracy.

Several other studies [93, 95, 105, 106] refine the phone boundary using SVM models. The feature vectors combining both spectral and prosodic information (MFCCs, entropy, f0, etc.) are used to represent the frames around the preliminary boundary, and then SVM based classification or regression is used to identify the most probable boundary. Binary SVMs are trained for different phone boundary classes. Frames in the training corpus are grouped into different categories in a way similar to that presented in [103] to obtain these SVMs. Particularly, the study presented in [106] uses the SVM model to refine the segmentation on Chinese singing voices given by DTW and HMM models. In this paper, the note boundaries are generated by two systems based on both the DTW method, which compares the pitch information of the singing voice to the corresponding note transcription, and the HMM method, which performs segmentation based on forced alignment process. The SVM trained from acoustic features and manual segmentations are then applied to estimate the confidence scores of two boundaries given by DTW and HMM. The boundary with the higher confidence score is then selected and refined by SVM as in Equation (2.10) to produce the final boundary. This process is shown in Figure 2.14.

Statistical corrections and fusion methods mainly address the systematic bias – they either compensate for the drift of phone boundaries produced by acoustic models using a shift or combine results from multiple acoustic models so that those biases from different models are compensated by each other. However, these methods fail to compensate for all the possible errors inside each phone boundary class. One reason is that different phone boundaries in the same class can still be affected by its context information. Further, different segmentation machines may not be able to compensate for all the errors from each other. In contrast to these two methods, predictive models examine the relevant acoustic features of frames around each preliminary boundary to refine the results, but it

Figure 2.14 Singing Voice Segmentation by DTW and HMM

may not be able to incorporate the global statistics effectively and may suffer from low implementation speed.

2.4 Summary

This chapter covers a literature review on three topics relevant to a CALL system, namely prosody evaluation, feedback utterances generation, and phonetic segmentation.

In the review on prosody evaluation, reference-dependent and reference-independent methods are studied. The basic schemes and investigations related to the two evaluation methods are discussed. For reference-dependent evaluation, recent proposals about word important factor, combination of multiple scores, MFCC based alignment, Mahalanobis distance and intonation type comparison are elaborated. Different structures of the reference-independent evaluation methods are discussed, including HMM and regression based models.

The feedback utterance is an important component of an advanced CALL system. A detailed discussion on different types of feedback is given, showing that accent reduction is a reasonable choice to produce stimuli with the learner's own vocal features. Two different kinds of schemes, namely rule-based and the reference-based schemes are discussed. Recent studies relevant to these two schemes are reviewed, which reduce perceived accents in the learner's non-native speech using both prosodic and segmental modifications. These studies

show that accent-reduced feedback utterances can provide learners with effective stimuli to practice their speaking skills.

Phonetic segmentation can contribute to various speech technologies potentially beneficial to CALL systems, such as text-to-speech, speech recognition, and accent reduction systems. Recent work on phonetic segmentation includes text-independent system, which purely rely on acoustic features and spectral information without prior knowledge about the phonetic sequences, and text-dependent systems, which incorporate both phonetic transcriptions and acoustic features to produce phone boundaries. More focuses are given to the latter scheme. Recent studies on text-dependent systems are reviewed, including statistical correction, fusion, and predictive models. Implementation issues as well as the pros and cons of these methods are also discussed.

In this thesis, four corpora are used for various studies described in Chapter 1. These corpora include Boston university radio news corpus (BURNc), ARCTIC corpus, TIMIT corpus, and a non-native corpus collected by the author. BURNc is the corpus produced by Boston University for the purpose of speech synthesis and prosody modeling. ARCTIC is a phonetically balanced corpus developed by Carnegie Mellon University for the purpose of unit selection speech synthesis. TIMIT is developed by MIT, SRI, and Texas Instrument for the purpose of acoustic modeling and speech segmentation. The author collected corpus consists of a number of non-native English utterances produced by local and international students in Singapore. The detail of this corpus will be given in Chapter 3. All of the four corpora use a sampling rate of 16k Hz.

Chapter 3 Prosody Evaluation Based on Prosodic Unit Segmentation

3.1 Motivations & Proposed Work

In prosody evaluation, one important issue is the segmentation of input sentences. Before the evaluation, all the pronounced sentences should be segmented into small units appropriately, as shown in Chapter 2.

Currently, segmentation for prosody evaluation is mainly based on lexical units as studied in [31, 32, 34]. It means that sentences are segmented into units in terms of words and syllables. Although such a segmentation method is suitable and often used for pronunciation evaluation, it is not very appropriate for prosody evaluation. Unlike lexical units which include phoneme, syllable and word, prosody is basically the supra-segmental features of a sentence and may not be related to the lexical boundaries strictly. Hence, if we segment a sentence according to lexical or syllabic units, the evaluation results may fail to reflect the learner's mastery of verbal articulatory accurately.

One logical solution is to perform the segmentation in the prosodic domain rather than in the lexical domain, i.e., the basic unit for segmentation should be a prosodic rather than a lexical unit. Unlike lexical units which mainly affect the lexical and syntactic meaning of an utterance, prosodic units can reflect the emotions, intentions and rhythm of an utterance. Considering these information given by prosodic units, it seems more reasonable to segment speech utterances in this way for prosody evaluation. In this chapter, segmentation based on a specific prosodic unit, i.e., foot, is applied for the evaluation of prosody. As manual labeling is time-consuming, an algorithm is also designed to perform the segmentation automatically.

To test the performance of the proposed segmentation in prosody evaluation, two different kinds of methods, reference-dependent evaluation and reference-independent evaluation, are experimented. Experiments using different segmentation methods for the reference-dependent and the reference-independent evaluation systems are performed, with the results given by the proposed segmentation method comparable to subjective scores from human experts.

3.2 Prosodic Unit

In order to segment a speech signal in the prosodic domain, it is pertinent to understand the hierarchy of prosodic units. Prosody is a supra-segmental layer of speech which consists of pitch, duration and intensity. It is generally used by a speaker to organize phonetic segments (vowels and consonants) into systematic units of various sizes. Since prosodic features are supra-segmental, many different levels exist, e.g., foot, phonological word, clitic, intonation unit, declination unit and utterance (from the low level to the high level) [108]. Hence, when applying prosodic units in the segmentation of utterances, the first problem is to select the appropriate unit.

In English, foot is defined as a phonological unit consisting of an accented syllable followed by any number of unaccented syllables. Specifically, a foot starts from the beginning of a stressed syllable and end at the beginning of the next stressed syllable. A phenomenon which should be noticed is that feet are mainly delimited by the stresses in a sentence. As a result, foot boundaries may not correspond to word boundaries. Thus, one word may cross two feet or one foot may consist of multiple words. For instance, the sentence “*I felt that I might never stop the machine from running*” and its pitch and intensity contours (extracted by *Praat*) are shown in Figure 3.1, with the discontinuous solid line indicating the pitch contour and the continuous dotted line indicating the intensity contour. The foot segmentation is “*I/ felt that I /might/ never /stop the ma/chine from/ running*” (delimited by vertical lines). Clearly, in this sentence, the boundary of a foot may locate inside a lexical word.

Figure 3.1 Waveform, Pitch and Intensity of the Sentence “*I/ felt that I/ might/ never stop the ma/chine from/ running*”

One exception to the foot level definition is anacrusis. Since each foot is combined by a stress and the following unstressed parts, it is necessary to define the component before the first stress. Anacrusis is defined as one or more unstressed syllables preceding the first stressed syllable in an utterance. For example, the initial unaccented phrase “*I*” (the first accented syllable is “*felt*”) is called anacrusis and considered as an incomplete unit from linguistics point of view. In this thesis, anacrusis is taken as a foot in the proposed segmentation scheme to model a complete prosody contour.

Foot is selected due to its appropriate position in the prosodic domain as well as its correlation with stresses. First, to segment a pronounced sentence appropriately, it is obvious that the segmentation units should not be too long; otherwise, it may give rise to problems in the segmentation process of short sentences and reduce the accuracy of evaluation results. The position of foot in the hierarchy of the prosodic domain generally coincides with that of word in the lexical domain. It implies that foot is a suitable unit to measure prosody. Second, the definition of foot pertains to stresses which contain a lot of rhythm information. When the same sentence is spoken by the same native speaker, there

is a tendency to keep the speaking rate of each foot not far away from the norm (relative to the tempo at the moment of the utterance). Accordingly, feet in a sentence can express significant rhythm information, which may contribute to the accuracy of evaluation results. Finally, when automatic segmentation is done at the foot level, existing techniques of stress detection can be used.

3.3 Automatic Quasi-foot Segmentation

The manual segmentation of foot is time-consuming and only feasible during the experimental stage and thus cannot be used in a real-time CALL system. Therefore, an automatic segmentation method is needed for the design of a CALL system.

For linguists who perform the segmentation manually, many detailed information including lexical meanings may be used to achieve accurate foot segmentation results. However, the basic and simplest definition of foot is unambiguous, i.e., a phonological unit consisting of an accented syllable followed by a number of unaccented syllables. Therefore, it is obvious that foot correlates to the stress or accentuation in English closely. Considering that a lot of work on stress and pitch accent detection has been done [109-113], segmenting sentences into feet according to stress or pitch accent detection technique is a reasonable and feasible approach. In English, pitch accent refers to the use of variation of pitch to express prominences and stresses. Therefore, it delimits the perceived stress as well as timing of English utterances.

The work in [110] conducts pitch accent detection at phone, syllable, and word level by determining whether there is a pitch accent in a phone, syllable, or word. The experiments are done by using the Naïve-Bayesian, Conditional Random Field and Logistic Regression methods. Boston University Radio News Corpus (BURNC) [114] is used for model training and testing. Experimental results demonstrate that pitch accent detection on word-level can reach a high level of accuracy. Besides, most of the foot boundaries coincide exactly with word boundaries.

Therefore it is reasonable to consider using word-level pitch accent detection method to obtain “quasi-foot” segmentations. Even though other methods such

Figure 3.2 Logistic Regression

as HMM based detection are also proposed by [109, 113], logistic regression as proposed in [110] is employed in our study because of its superior performance and simplicity.

Logistic regression is a kind of generalized linear regression. It is well known that linear regression approximates the real value by multiplying a feature vector with a weight vector. In logistic regression, the multiplication of the feature vector with a weight vector is first performed, and the result is further processed to obtain the probability of accentuation. The basic function of logistic regression is:

$$f(z) = \frac{e^z}{e^z + 1} \quad (3.1)$$

$$z = w \cdot x \quad (3.2)$$

where x is the feature vector, w is the weight vector, z is the dot product of weight and feature vectors, $f(z)$ is the output with a value between 0 and 1, regardless of the value of z . Therefore, $f(z)$ is especially useful to describe probability. In this case, the output is taken as a probability to indicate whether the word is accentuated. Normally, the threshold for accentuation judgement is set at 0.5.

The quasi-foot segmentation involves the following steps: first, word level segmentation is obtained by forced alignment which is implemented by HTK [115]. HMM based acoustic models are trained by the Wall Street Journal (WSJ) corpus involving context-dependent models [116]. A discussion about forced alignment will be given in Chapter 5. Second, logistic regression is performed to detect pitch accent at the word level. The parameters of logistic regression, i.e., the w in Equation (3.2), are trained by machine learning software WEKA [117] through feature vectors and manual labels. As proposed in [110], normalized pitch, energy and duration are used as elements of input feature vector. Pitch is extracted by the subharmonic-to-harmonic ratio method (SHR) [118], with an average estimation error of 5 Hz based on CSTR database as reported in [119]. Energy is defined in the root mean square (RMS) form. Pitch slope, calculated as the first order pitch regression line across two neighboring frames, is also added as an extra feature. The normalization is performed by dividing mean pitch and energy at the word-level by those at the sentence level. In prosody evaluation, the boundaries of detected accented words are taken as the quasi-foot boundaries.

In Figure 3.3, an example of automatic quasi-foot segmentation is shown. The solid lines are the word boundaries obtained by forced alignment, whereas the dash lines are quasi-foot boundaries detected by logistic regression. Almost all the manually detected foot boundaries are correctly identified in this example. From Figure 3.3(a), those words with POS tag as noun, verb, adjective, or adverb like “that’s”, “supporters”, “Safe”, “Road”, “Act”, “hopping” and “work” are detected as accented words. Similarly, in Figure 3.3(b), Figure 3.3(c) and Figure 3.3(d), words containing main information of the sentence like “choice”, “association”, “grilled”, “initially”, “involve”, “inmates”, “consequences” and “driving” are correctly identified. The obtained segmentations are reasonable since human speakers also usually put stress or emphasis on those kinds of words. Also, there is a series of unstressed syllables following those words just as what has been stated in the definition of foot.

(a)

(b)

(c)

Figure 3.3 Example of Quasi-foot Segmentation

To analyze the performance of the quasi-foot segmentation method, experiments are performed on the BURNC corpus. In pitch accent detection, the detected pitch accents are compared with human labeled accentuation provided by the corpus. The estimated boundaries are compared with human labeled boundaries and the F-Measures are shown in Table 3.1. F-measure is defined as the harmonic mean of “precision” and “recall”, as given below:

$$\text{Precision} = \frac{\text{number of correct results}}{\text{number of all the returned results}} \quad (3.3)$$

$$\text{Recall} = \frac{\text{number of correct results}}{\text{number of results that should have been returned}} \quad (3.4)$$

$$\text{F-Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.5)$$

The obtained results show that F-Measure for pitch accent detection accuracy is 0.81. The F-Measure of quasi-foot segmentation is 0.76 because some feet boundaries do not correspond to word boundaries. Although the detection is not perfect, the F-measure is comparable to those achieved in [120], which are considered as accurate for pitch detection and thus is good enough for prosody evaluation.

Table 3.1 Pitch Accent and Foot Boundary Detection Results

	Detection Method	Evaluation Method	Detection Accuracy
Pitch Accent Detection	Logistic Regression	Compared with real pitch accent	0.81
Quasi-foot Segmentation	Forced Alignment +Logistic Regression	Compared with manually detected foot boundaries	0.76

3.4 Reference-dependent Prosody Evaluation Based on Prosodic Unit

3.4.1 Framework of Reference-dependent Evaluation System

The system configuration used for reference-dependent prosody evaluation is shown in Figure 3.4. For each of the learner's utterance, the corresponding teacher's utterance must be available and used in the evaluation process. Segmentation is first performed on both of the teacher's and the learner's utterances. The manual segmentations were provided by human experts who are linguistic experts from School of Humanities and Social Sciences in Nanyang Technological University (NTU).

Secondly, feature extraction is then implemented on each unit which could be word or foot. In our study, each vector has four features. They are normalized log pitch, normalized log intensity and their first derivatives. As the human auditory system perceives tones in a way similar to the logarithmic function (instead of in a linear function), the logarithmic scale is a more appropriate measure of the perceptual effects of the differences in intonation. In addition, average values are subtracted from log pitch and log intensity for normalization.

Figure 3.4 Framework of Reference-dependent Evaluation System

After obtaining the sequences of prosodic features of the teacher and the learner, alignments between the two sequences are performed. Then, the accumulative distance between the teacher’s and the learner’s prosodic features is calculated unit by unit to obtain the score at unit level. Finally, the overall sentence score can be obtained by averaging all the unit scores.

3.4.2 Distance Measurement

Distance measurement is the most important part of the reference-dependent evaluation. It compares the feature vector of an input speech utterance with that of the reference utterance so as to obtain a “measure of the differences” between the two vectors as the evaluation results. However, the direct comparison between two feature vectors is not feasible due to two reasons.

Firstly, in most of cases, the lengths of the two feature vectors are different. Since there is no way to calculate the Euclidean distances between two vectors with different lengths, direct comparison does not work. Secondly, even in the very rare case that two feature vectors possess the same length, the intonation contour of the learner’s utterance and the reference utterance can be quite different. Therefore, a direct comparison between two feature vectors can lead to inaccurate evaluation results. One possible solution is to compare the

corresponding salient components of two feature vectors, e.g., the peak of one feature with those of the other. In such a comparison, the distance measurement concentrates on the prosodic contours represented by the feature vectors, which provides most of the information about the nativeness of the speech utterances in terms of prosody.

Considering the two problems mentioned above, dynamic time warping (DTW) is used. It is a scheme which can solve the line-up or alignment problem between two sequences with the same or different lengths. An example of DTW has been shown in Section 1 of Chapter 2. In very early years, DTW was used as a tool for isolated word recognition. With the advent of statistics based speech recognition techniques such as HMM and artificial neural network (ANN), its adoption in speech recognition has ceased. However, it is still an important tool for alignment of time series and sequences.

A lot of previous work on DTW can be found in [57, 121, 122]. From [122], the basic procedure of DTW can be summarized as follows:

Given two sequences, T with index n and R with index m :

$$T = \{T(1), T(2), \dots, T(N)\} \quad (3.6)$$

$$R = \{R(1), R(2), \dots, R(M)\} \quad (3.7)$$

To measure the distance between two feature sequences, the accumulative distance D_A is calculated to measure the difference between the two sequences:

$$D_A(n, m) = d(T(n), R(m)) + \min[D_A(n-1, m), \quad (3.8)$$

$$D_A(n, m-1), D_A(n-1, m-1)]$$

The D_A function defines the basic DTW iteration which allows each point to move to its 3 neighboring points. The movement direction should be the one leading to the smallest increase in D_A . This process iterates continuously until the terminal point is achieved by going through the two sequences. Then, the final distance between the two sequences is:

$$\bar{D} = D_A(N, M) \quad (3.9)$$

where N and M are the numbers of frames of two sequences, respectively.

Therefore, distance measurement in reference-dependent method can be achieved appropriately by the DTW algorithm. The approach of DTW resolves the alignment problem and hence allows the distance between two feature vectors to be calculated more appropriately. The evaluation score is calculated based on the distance between the learner's and the teacher's feature sequences.

In previous papers [31, 34, 123], Euclidean distance is used to measure the accumulative distance. However, it is noticed that the covariances of the features may also affect the evaluation results, as correlations and interactions among different features do exist. To overcome this problem, Mahalanobis distance is used as proposed in [59]. The Mahalanobis distance between two feature vectors s and t with covariance matrix Σ is calculated by considering their covariance matrix:

$$d_M(s, t, \Sigma) = \sqrt{(s - t)^T \Sigma^{-1} (s - t)} \quad (3.10)$$

Let $s_j(n)$ and $t_j(n)$ be the n -th feature vector of the j -th prosodic unit (e.g. foot) of the learner's and the teacher's utterance, respectively. Assuming $M_j(n)$ is the Mahalanobis distance of the n -th frame of the j -th prosodic unit, the machine score can be calculated as follows given the covariance matrix:

$$M_j(n) = d_M(s_j(n), t_j(n), \Sigma) \quad (3.11)$$

$$U_j = \sum_{n=1}^{N_j} M_j(n) \quad (3.12)$$

$$S^W = \sum_{j=1}^J U_j(n) \quad (3.13)$$

where N_j and J are the number of frames of the j -th prosodic unit and the number of units in a sentence, respectively. U_j and S^W represent the accumulative distance of the j -th prosodic unit and that of the whole sentence. The covariance matrix can be calculated as:

$$\Sigma = \frac{1}{N-1} \sum_{j=1}^J \sum_{n=1}^{N_j} (t_j(n) - \tilde{t})(t_j(n) - \tilde{t})^T \quad (3.14)$$

where $\tilde{t}$ is the average of the teacher's feature vector.

In this way, the distance between each pair of the teacher's and the learner's prosodic features can be calculated. The distance can be used to indicate the similarity between the learner's and the reference utterances in terms of their

prosodic contours, thus giving the assessment of the learner's prosody. To test the accuracy of the evaluation results, subjective scores of each utterance done by human evaluators are first obtained. The correlation coefficient between human evaluation scores and machine measured distances can then be calculated to assess the performance of the prosody evaluation system.

3.4.3 Evaluation Experiment

To test the validity of the proposed segmentation method, an experiment based on both word and foot level segmentation is performed. In the experiment, sentences pronounced by a native speaker are used as the teacher's utterances. The experimental conditions are described in Table 3.2.

Six unique sentences are used. They include two long sentences and four short sentences; five of which are statements and one is a question. Pronounced sentences from NTU students whose native language is not English are recorded for evaluation. Eight students were invited to record their utterances which are taken as the learners' utterances.

With automatically detected word or foot boundaries, feature vectors of each prosodic unit are extracted for distance measurement. The pitch contours are obtained using harmonics-to-subharmonics ratio techniques as in [118]. Subjective scores are given by linguistic experts from School of Humanities and Social Sciences in NTU. The subjective evaluators were asked to evaluate each

Table 3.2 Conditions of the Reference-dependent Evaluation Experiment

Learners	8 students from School of Humanities and Social Science and School of Electrical and Electronic Engineering in NTU
No. of Sentences	6 different sentences with 2 long sentences and 4 short sentences
Feature Vector	normalized log pitch, normalized log intensity, and their first derivatives
Subjective Evaluation Score	5 point evaluation score (from 1 to 5)

of the utterances to give a score ranging from 1 to 5 (worst to best). The correlation coefficient between scores given by human evaluators or between human and machine scores used in this section is defined as follows:

$$r = \frac{\sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (Y_i - \mu_Y)^2}} \quad (3.15)$$

where X_i and Y_i are two sequences of scores, μ_X and μ_Y are the mean of the two sequences, and n is the number of scores.

With the automatic quasi-foot detection, prosody evaluation results are obtained based on the corpus as mentioned in Table 3.3. The inter-human correlation is 0.43. In Table 3.3 the human-machine correlation coefficients are given with the negative sign changed to positive for convenience. It can be observed that the human-machine correlation based on foot level segmentation is around 0.39. In contrast, experiments with the same conditions performed at word level give the human-machine correlation of around 0.33. Also, compared to other prosody evaluation systems based on word level segmentation [34, 59], the overall correlation coefficient obtained by the proposed foot level segmentation is higher.

The word-level segmentation using forced alignment can yield the same accuracy as that produced by manually segmented words. This is because forced alignment has evolved and been refined for years, and it has become a reasonably reliable segmentation scheme based on word boundary detection. The human-machine correlation obtained by quasi-foot segmentation is slightly lower than that given by the human labeled foot boundaries, but it still outperforms word level segmentation. The lower accuracy is mainly due to the uncertainty in pitch accent detection. T-tests are applied to test the significance of correlation coefficients. The statistics is defined as:

$$t = r\sqrt{(n-2)/(1-r^2)} \quad (3.16)$$

where r is the correlation coefficient, and n is the number of observations.

Table 3.3 Prosody Evaluation Results with Different Segmentation Methods

Segmentation Unit	Techniques	Human-machine Correlation	T-test Results (p value)
Word (manually or automatically)	Forced Alignment done by a speech recognizer	0.33	0.0164
Foot	Human-labeling	0.39	0.0053
Quasi-foot (automatically)	Forced Alignment + Pitch Accent Detection	0.38	0.0065

3.5 Reference-independent Prosody Evaluation Based on Prosodic Unit

Even though the reference-dependent method can evaluate the prosody automatically, it is still not ideal for CALL application. One obvious problem is that the learner can only practice on those pre-recorded sentences pronounced by native speakers; otherwise, the evaluation process which measures the distance between prosodic features cannot be conducted. To allow for more flexibility in the language learning process, a number of reference-independent systems as described in Chapter 2 are also proposed. As the proposed prosodic unit based segmentation does not rely on a specific model, it can also be applied to the reference-independent evaluation process. To test the performance of reference-independent evaluation using the proposed quasi-foot segmentation method, the SVM based prosody evaluation as mentioned in [65] is used in our investigations.

A detail discussion of SVM is given to facilitate the understanding of the reference-independent evaluation process. The following sections first review the

theories of SVM and then discuss how it is used in a the prosody evaluation scheme.

3.5.1 Support Vector Machine

Support vector machine is a classical pattern recognition algorithm whose history can be traced to 1990s by Vladimir N. Vapnik [124]. With increasing demands for efficient and effective algorithms for pattern recognition, more and more researchers recognize the importance of SVM, and thus a lot of work related to SVM has been carried out. Recent reviews and discussions on SVM can be found in [125-128].

The objective of SVM is to solve classification or regression problems based on an existing database with corresponding labels. Label here means the correct class of each sample in a classification problem or the real value corresponding to each sample in a regression problem. Therefore, two main steps are involved in a typical SVM problem:

1. Training: The model of support vector machine is trained according to the samples in a training set and their corresponding labels. In such a process, the parameters for the SVM model are estimated to minimize misclassification errors or the difference between the real and estimated values. The selection of training samples should be appropriate to ensure the performance of SVM.

2. Testing: The parameters obtained in the training process can be used to estimate the values of the testing samples. The estimated values are then used to compare the difference between the actual and the estimated labels to assess the performance of the trained model.

The fundamental form of SVM addresses the classification problem of linearly separable data. It is essentially a binary classification model. However, multiple-class models can be easily generated by using either one-versus-one or one-versus-all modeling of binary classifiers. To represent the binary classification problem using SVM, the labeled feature vector set is defined as $\mathbf{x}_i, y_i \in R^n$ ($i = 1, 2, \dots, l$) and $y_i \in (-1, +1)$. It is supposed that the two classes can be separated by a hyperplane satisfying $\mathbf{w}^T \mathbf{x} + b = 0$, where $\mathbf{w}$ is the weight and b is the intercept.

In SVM, the margin is defined as the distance from the hyperplane to the nearest sample. Obviously, larger margin indicates a better separation. Therefore, the objective of SVM is to separate two kinds of data while maximizing the margin. To solve the linearly inseparable case, a slack variable ξ_i is added so that the soft margin scheme is introduced. Without loss of generality, the classification problem can be described as:

$$\mathbf{w}^T \mathbf{x}_i + b \geq 1 \quad \text{for } y_i = +1 + \xi_i \quad (3.17)$$

$$\mathbf{w}^T \mathbf{x}_i + b \leq -1 + \xi_i \quad \text{for } y_i = -1 + \xi_i \quad (3.18)$$

which is equivalent to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad (3.19)$$

The margin can be defined as ρ according to the following derivations:

$$\begin{aligned} \rho(\mathbf{w}, b) &= \min_{y=-1} d(\mathbf{w}, b; \mathbf{x}) + \min_{y=+1} d(\mathbf{w}, b; \mathbf{x}) \\ &= \min_{y=-1} \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|} + \min_{y=+1} \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|} \\ &= \frac{1}{\|\mathbf{w}\|} \left(\min_{y=-1} |\mathbf{w}^T \mathbf{x}_i + b| + \min_{y=+1} |\mathbf{w}^T \mathbf{x}_i + b| \right) = \frac{2}{\|\mathbf{w}\|} \end{aligned} \quad (3.20)$$

Therefore, maximization of the margin is equivalent to the minimization of $\|\mathbf{w}\|$. The corresponding Lagrange equation is:

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \\ &\quad - \sum_{i=1}^l \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^l \beta_i \xi_i \end{aligned} \quad (3.21)$$

where α_i represents the Lagrange multipliers, which is the parameter of SVM, and β_i is also the Lagrange multipliers but can be eliminated by taking the dual problem of L_p . C is the penalty parameter to be chosen by the user. It controls the tradeoff between the flatness of hyperplane and the upper bound of errors. By

taking the derivatives of L_p and substituting the results to the original Equation (3.21), L_p can be changed to a dual problem, i.e.,

$$\begin{aligned} L_D &= \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ &\text{subject to } 0 \leq \alpha_i \leq C \\ &\text{and } \sum_i \alpha_i y_i = 0 \end{aligned} \quad (3.22)$$

where α_i is calculated as the SVM parameter. Those vectors $\mathbf{x}_i$ corresponding to $\alpha_i > 0$ are called “support vector”, since if all the other points are removed and only support vectors are retained, the obtained SVM model will be the same. After obtaining α_i , the separation problem can be solved by considering the sign of:

$$f(x) = \sum_{i=1}^N \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle \quad (3.23)$$

where N is the number of training data, $\mathbf{x}_i$ is the support vector and $\mathbf{x}$ is the testing feature vector. If the sign of $f(x)$ is “+”, then the input vector should be categorized as class “+1”; otherwise, it should be categorized as class “-1”.

Although slack variables are introduced, linear SVM cannot work well in many complex and non-separable situations. Therefore, SVM should be generalized to be more than a linear function. From previous equations of SVM, it is obvious that the training data set $\mathbf{x}_i$ only exists in the inner product with the input test feature vector. Therefore, to generalize SVM into a non-linear function, a mapping from the original Euclidean space to a higher dimension can be proposed as:

$$\Phi: R^d \rightarrow H \quad (3.24)$$

Therefore, the inner product of training samples becomes $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. For the convenience of description as well as calculation, a kernel function is defined as $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. After mapping to the higher dimension, the data can become linearly separable and a linear function in the higher dimension is able to separate the nonlinear data in the original low dimension. Therefore, the nonlinear classification problem in SVM can be solved by introducing the kernel function, i.e., :

Figure 3.5 Example of Non-linear SVM

$$f(x) = \sum_{i=1}^N \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \quad (3.25)$$

One example of non-linear SVM is shown in Figure 3.5. The two dotted curves indicate the margins of SVM and the circles located on the dotted curves refer to support vectors. It can be observed that the two data sets of circles are separated by a non-linear hyperplane with a small set of support vectors.

By using a kernel function, the complex calculation of mapping to $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$ can be reduced to the simple calculation of $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Since $K(\mathbf{x}_i, \mathbf{x}_j)$ is determined only by the product of $\mathbf{x}_i$ and $\mathbf{x}_j$ in the low dimensional space, the complexity is much reduced. For instance, the polynomial kernel function is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^n \quad (3.26)$$

Clearly, the computation complexity is decided by $\mathbf{x}_i \cdot \mathbf{x}_j$ with any exponential variable n . As a result, the kernel function can reduce the computation complexity greatly.

3.5.2 Segmentation Method & Features

To construct the evaluation model, both speech utterances and their corresponding subjective scores are required to train the evaluation model. In the

evaluation process, feature vectors should be constructed in the sentence level as the input of the evaluation model so that a score for each sentence can be generated. Sentences are first segmented into smaller units and the feature vectors for each segment are then extracted. The segment-level feature vectors are combined to form the sentence level feature vector as the input of the evaluation model.

To test the performance of prosody evaluation using quasi-foot segmentation, SVM based models are applied to assess the prosody of an input utterance without the use of reference utterances. Three segmentation methods based on (1) foot, (2) word and (3) voiced/unvoiced (v/uv) are used and the corresponding prosody evaluation results are compared. Segmentations based on smaller units such as syllable or pseudo-syllable are not included as they are too short to carry sufficient prosodic information. The v/uv segmentation is performed based on the method proposed in [65]. If the unvoiced part is longer than 50ms, the two neighboring voiced parts are separated into two segments; otherwise, they will be merged and interpolated as one segment.

Prosody evaluation is performed by SVM based regression. To be consistent with previous work in [58, 59, 65], a 5-level score (with 1 as the worst and 5 as the best) is assigned to each sentence by human evaluators. After training the SVM based on mean subjective scores from human evaluators, the machine score can be estimated according to the SVM regression model and the input feature vector. Although the SVM classification model is not so appropriate for prosody evaluation due to the loss of order information, it is also included as an alternative for comparison purposes. The loss of order information refers to the fact that the relationship between SVM class corresponding to score 5 and that corresponding to score 1, i.e., the former is superior to the latter, is not reflected by the classification models. In such case, each score given by the human evaluator is taken as one class. For example, if a 5 point scoring scheme is used in subjective assessment, then there are 5 classes in the SVM based classification. The SVM classification model is trained on a series of input utterances and their subjective scores, and then a class corresponding to one of the five scores is assigned during the evaluation process. The performances using

these two methods, i.e., regression and classification models, are evaluated in terms of human-machine correlation and classification accuracy, respectively.

A 20-dimensional feature vector is extracted from each segment given by various segmentation methods. The extracted features are as follows:

Pitch Related Features: maximum of pitch value, position of maximum pitch value, minimum of pitch value, position of minimum of pitch value, start pitch value, end pitch value, position of start pitch value, position of end pitch value, regression slope, regression intercept, and regression error.

Energy Related Features: maximum of energy value, position of maximum energy value, minimum of energy value, position of minimum of energy value, start energy value, end energy value, regression slope, regression intercept, and regression error.

All the pitch and energy values are converted to the log scale by taking the log of the original values. A total of 20 prosodic features are extracted for each segment. These features in all the segments are then combined to form the sentence level feature vector. The sentence level feature vector is composed by extracting the maximum, minimum and mean of each of the 20 features across all the segments. Therefore, a total of 60 features are obtained for each utterance. The 60-dimensional feature vector is then used for training and testing of the SVM based evaluation models.

To train SVM models appropriately, the kernel function must also be decided. As mentioned before, kernel function determines the performance of SVM model in the nonlinear case. Since most prosodic features are not linearly separable, an appropriate kernel function is necessary to improve the evaluation results. There are several kinds of popular kernel functions: linear kernel function, polynomial kernel function, radial basis function (RBF), sigmoid function and other pre-computed functions. It is shown in [129] that “if RBF is used with model selection, then there is no need to consider the linear kernel”. Since the distribution of prosodic features is always non-linear and complex, linear kernel function is first excluded. Studies in [130] also demonstrates that the positive definite property of kernel matrix may not be guaranteed by using sigmoid kernel and it generally underperforms RBF

kernel. Polynomial kernels can generate reasonable results with lower dimensions, but numerical difficulties tend to rise when higher dimensions are used. Besides, functions like Fourier kernel function are always too complicated to guarantee the generality. Therefore, RBF kernel seems a reasonable choice. In addition, from recent papers on kernel selection of SVM [131-133], RBF kernel is most commonly used. Since RBF kernel maps a feature vector into infinite dimensions, the empirical risk can be controlled to a certain extent [125]. Besides, even though its Vapnik–Chervonenkis (VC) dimension is high, experiments show that the expected risk is still small and acceptable as discussed in [131-133]. It means that RBF kernel can provide better regression or classification results while maintaining reasonable generality. Due to all the considerations above, RBF kernel is used in the prosody evaluation system. The RBF kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-g|\mathbf{x}_i - \mathbf{x}_j|^2\} \quad (3.27)$$

where parameter g is a pre-defined parameter to control the performance of RBF kernel.

3.5.3 Experimental Results

To test the proposed prosody evaluation method, a larger database is also collected based on BURNC corpus as listed in Table 3.4. In this corpus, 20 unique sentences from BURNC are selected, including 15 normal sentences and 5 long sentences. Ten students in NTU with various nationalities including Chinese, Indians, Vietnamese and Singaporean were invited to read each of the 20 sentences, generating 200 utterances. Besides, utterances from BURNC by 3 speakers for each of the 20 sentences are also selected to form the 60 teachers' utterances. Therefore, a total of 260 utterances are included in the corpus. The subjective scores of the students' utterances in terms of the nativeness of prosody are given by linguists, while all of the teachers' utterances are assigned a full score of 5. The overall correlation among human evaluators is 0.64. Both the training and testing of SVM are based on the corpus for prosody evaluation with 10-folds cross validation, i.e., dividing all the samples into 10 equal parts and use 9 of them for training and the rest of them for testing. This process is repeated

Table 3.4 Experimental Conditions Based on BURNC

Database	Boston University Radio News Corpus (BURNC)
No. of Transcriptions	20 unique sentences selected from BURNC
No. of Students' utterances	A total of 200 students' utterances from 10 students including Chinese, Indian, Vietnamese and Singaporean
No. of Teachers' utterances	A total of 60 teachers' sentences from BURNC
Subjective Evaluation	3 evaluators with 5 point evaluation (inter-human correlation 0.64)

after each of the 10 parts is used for testing, and the mean evaluation results are reported. Particularly, each of the 10 sets has different sentences, so that the testing set consists of sentences not existing in the training set. This design can verify the text-independent property of the prosody evaluation process.

The SVM algorithm is implemented by LIBSVM [134], a public SVM software developed by National Taiwan University. The parameters involved in the SVM training process such as penalty variable C and kernel parameter g are determined by grid search as proposed in [127]. This process first sets a number of values for C and g separately, and then uses those value to form a 2-D “grid” to experimentally decide the most appropriate parameters. As recommended by [127], the value of C is in the range of $[2^{-5}, 2^{-3}, \dots, 2^{15}]$ while g is in the range of $[2^{-15}, 2^{-13}, \dots, 2^3]$. The SVM model is to be trained across all the points in the grid formed by the combination of two parameters with different values. After that, a similar but finer grid search is performed around the previously detected optimal point to refine the obtained parameters. The obtained values for C and g which can yield the best results are then used for the SVM. The 10-set

cross validation experimental results of three different segmentation methods mentioned before are shown in Figure 3.6.

Figure 3.6 Cross-validation Results of Different Segmentation Methods

Table 3.5 Experimental Results Based on Different Segmentation Methods

Segmentation Methods	Human-machine correlation (regression)	Classification Accuracy (classification)
quasi-foot	0.61	51.7%
word	0.54	46.4%
v/uv	0.45	42.3%

It can be observed that the human-machine correlation using quasi-foot segmentation outperforms the other two methods across different subsets. Also, the influences of these differences can be averaged to assess the overall performance of each segmentation method. The overall performance in terms of classification accuracy and human-machine correlation are shown in Table 3.5.

Classification accuracy refers to the percentage that the machine scores obtained by the classification model are equal to the corresponding human scores. T-tests are performed to test the significance level of the difference between classification accuracies and correlation coefficients obtained by the three different segmentation methods, showing $p < 0.01$ for each pair of segmentations. From the obtained results, the quasi-foot segmentation achieves the best results in terms of both human-machine correlation and classification accuracy. The quasi-foot segmentation shows the highest human-machine correlation (0.61) which is close to the inter-evaluators correlation (0.64).

From Table 3.5, the voiced/unvoiced segmentation underperforms the other two methods because neither lexical nor prosodic information is modeled in such a segmentation scheme. It is also shown that quasi-foot segmentation can achieve better evaluation results in terms of human-machine correlation and classification accuracy. The word segmentation approach can obtain acceptable results, but fails to match the performance of quasi-foot segmentation. The proposed segmentation is therefore a good choice for prosody evaluation.

According to the experimental design, the transcriptions of testing sentences do not exist in the training set. It means that the SVM based prosody evaluation can assess input utterances with new transcriptions which are not presented in the training corpus. Hence, this text-independent evaluation scheme can raise the interest of learners by enabling them to practice any sentences rather than being restricted by pre-designed practice sentences as is usually the case in reference-dependent systems like [58, 59].

3.5.4 Feature Selection Test

Although good results can be obtained by the proposed method with a 60-dimensional feature vector, feature selection is still desirable to determine the most reasonable and effective feature subset for prosody evaluation. An appropriate feature subset not only can improve the efficiency of the feature extraction and prosody evaluation process, but also can improve the evaluation results by eliminating redundant features. Feature selection methods in [135-137] have been reviewed and the minimum-Redundancy-Maximum-Relevance (mRMR) method proposed by [137] is used for its reliability and theoretical basis.

In this method, the correlation between labels and features is maximized whereas the inter-feature redundancy is minimized to select the best feature subset. Max-relevance refers to searching of features that maximize the relevance which is measured by mutual information between feature set S and class c . The optimal feature subset S is calculated according to the mutual information values between the individual feature x_i and the class c :

$$\max D(S, c), \quad \text{where } D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \quad (3.28)$$

where $I(x_i; c)$ is defined as the mutual information between x_i and c :

$$I(x_i; c) = \iint p(x_i, c) \log \frac{p(x_i, c)}{p(x_i)p(c)} dx_i dc \quad (3.29)$$

In contrast, Min-redundancy refers to searching of features that minimize the dependence which is also measured by mutual information between different

features in the obtained subset. With two features highly dependent on each other, the classification result will not change much even if one of the two is removed. Therefore, Min-redundancy can reduce the feature dimension and therefore simplify the feature extraction process. This method is implemented by minimizing the mutual information between each pair of features, i.e., :

$$\min R(S), \text{ where } R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \quad (3.30)$$

Using the mRMR algorithm, the optimal feature subset for prosody evaluation can be obtained by maximizing the relevance between the feature vector and the class, while minimizing the dependence between each pair of features.

In the feature selection experiment, the dimension of a feature vector is restricted from 10 to 60 as the target of this study is reducing the dimension of feature vector while achieving acceptable evaluation performance. The experimental results concerning the relationship between the feature dimension and the evaluation accuracy are presented in Figure 3.7.

Figure 3.7 Relationship between Feature Dimension and Evaluation Accuracy

It is shown that the quasi-foot based segmentation method results in the highest human-machine correlation compared with the two other methods. This segmentation method achieves the best correlation coefficient of 0.66 with a feature dimension of 18, which is smaller than that for word segmentation (44) and v/uv segmentation (57). It means that quasi-foot segmentation can obtain a reasonable accuracy with a comparatively smaller feature subset, leading to less computational complexity.

Furthermore, the human-machine correlation achieved by quasi-foot segmentation generally outperforms that of the word based segmentation across different feature subset, demonstrating better robustness to the change of features. One explanation for such a phenomenon is that quasi-foot segmentation has implicitly involves some rhythmic information, thus its performance can outweigh that of word segmentation which contains few prosodic information for the cases where low dimensional feature vectors are used.

To study the obtained feature subset, the first 15 preferred features selected by mRMR algorithm for the quasi-foot segmentation are listed as follows: (1) maximum of energy regression error, (2) minimum of start pitch position, (3) minimum of maximum pitch position, (4) maximum of minimum energy, (5) minimum of minimum energy position, (6) minimum of maximum energy, (7) maximum of pitch regression slope, (8) minimum of minimum pitch, (9) mean of energy regression intercept, (10) mean of maximum pitch position, (11) maximum of end energy, (12) mean of pitch regression error, (13) minimum of energy regression slope, (14) maximum of minimum pitch position, and (15) minimum of pitch regression slope.

From this feature list, it is obvious that for the most important 15 features, the number of pitch correlated features and that of energy correlated features are nearly the same, demonstrating that both energy and pitch are important to produce natural prosody. Besides, the fact that a number of features related to the regression line are included in the 15 most important features shows the importance of pitch and energy contours in prosody evaluation, as those features are highly correlated to the contour information. This observation is in line with

the common idea that prosody is tightly linked to the contours of pitch and energy.

The feature selection process reduces the feature dimensions while retaining acceptable evaluation accuracy. In addition, the experimental results also show that the quasi-foot approach can outperform the other two methods in terms of evaluation accuracy as well as robustness to the change of features used for evaluation.

3.5.5 Relationship between Accented Word Ratio and Human Score

To further examine the contribution of the proposed segmentation approach, the relationship between the accented word ratios and subjective scores is studied. Accented word is defined as the word that has a pitch accent. It can be calculated as the number of accented words divided by the total number of words in an utterance. The distribution of subjective scores given by three evaluators (T1, T2 and T3) is shown in Table 3.6. It can be observed that very few sentences are assigned a score of 1, whereas the distributions of other scores are relatively even. Therefore, score 1 is excluded in the analysis of the relationship between accented word ratio, and the ratios are calculated for these sentences with a prosody score from 2 to 5. The mean value of accented word ratios and the distributions of the accented word ratios across different scores given by the three evaluators are shown in Figure 3.8 and Figure 3.9, respectively. The lines in Figure 3.9 connect the mean values of ratios in different groups.

It can be seen that the accented word ratio decreases as the subjective scores given by human evaluators increase. This phenomenon is explainable. As non-native speakers with poorer speaking skills tend to accentuate most of the words in their attempt to make correct pronunciations, more accented words will be detected in their utterances. In contrast, speakers with proficient speaking skills can manipulate stresses well and accentuate the words appropriately according to the rhythm of the sentence, thus leading to fewer accented words. Hence, accented words reflect the learners' manipulation of English rhythm and affect the nativeness of their prosody.

Table 3.6 Subjective Score Distribution from Three Evaluators (Number of Sentences for Each Score)

	Score 1	Score 2	Score 3	Score 4	Score 5
T1	1	35	72	36	56
T2	3	28	63	57	49
T3	12	48	54	37	49

Word or voiced/unvoiced segmentation does not differentiate accented and unaccented word, thus the obtained feature vector will not be affected by the rhythm information. In contrast, quasi-foot segmentation is based on accented word detection and each segment correlates to one accented word. As a result, the rhythm information correlated to accented words is thus modeled by the quasi-foot segmentation and contributes to the prosody evaluation. This observation further shows that the proposed prosodic unit segmentation is an appropriate approach for prosody evaluation.

Figure 3.8 Relationship between Accented Word Ratio and Human Score

(a) Accented Word Ratio Distribution According to Score Given by T1

(b) Accented Word Ratio Distribution According to Score Given by T2

(c) Accented Word Ratio Distribution According to Score Given by T3

Figure 3.9 Accented Word Ratio Distribution across Scores Given by Three Evaluators

3.6 Summary

In this chapter, a segmentation approach based on a prosodic unit called foot is proposed for prosody evaluation. Since each prosodic unit corresponds to the phrasing information, such kind of segmentation is more similar to how human verbally articulate the prosodic contour.

To reduce the burden of human labeling, a segmentation method is proposed to obtain foot boundaries automatically based on pitch accent detection. As an important component in prosody study, the pitch accents of an utterance can be estimated from pitch, duration and intensity of a speech utterance. Since the pitch accent correlates to stresses in an utterance closely and most of foot boundaries still lie on word boundaries, it is reasonable to use accented words detection to achieve automatic quasi-foot segmentation. The proposed quasi-foot segmentation based on forced alignment and logistic regression can obtain reasonable segmentation results in comparison with human labeled data. In the experiment using reference-dependent evaluation, the human-machine correlation obtained by segmentation based on foot outperforms that obtained by word level segmentation, showing the effectiveness of the proposed method.

The proposed segmentation method is also tested on reference-independent evaluation system to provide more flexibility for the learners. The reference-independent evaluation is based on SVM classification and regression models. The adoption of kernel function in SVM enables the formation of complicated separation hyperplane while reducing computation complexity. The reference-independent evaluation is tested on a corpus based on BURNC using three different segmentation methods, i.e., word segmentation, quasi-foot segmentation, and voiced/unvoiced segmentation. The experimental results demonstrate that the proposed quasi-foot segmentation can outperform the other two segmentation methods in terms of both classification accuracy and human-machine correlation.

It can be observed that the results obtained from reference-independent evaluation outperform those given by reference-dependent evaluation. There are two main reasons that explain this observation. Firstly, the modeling scheme of

reference-independent method is more flexible and effective. This is because machine learning techniques like SVM can be used under such a scheme, and the evaluation models are optimized for the given features and thus it can link the correlated features to the labels (scores) in a precise manner. Secondly, the reference-dependent method uses DTW which only involves a direct comparison between the feature vectors from two speakers, without optimizations to increase the robustness of the system. DTW based measurement has already been proved to be ineffective in comparison with other methods like HMM or SVM in the topic of speech recognition. It can also be observed in the literature that the DTW based reference-dependent method underperforms reference-independent methods [65, 138].

A feature selection based on mRMR is performed to obtain the optimal feature subset. Experimental results show that the quasi-foot segmentation can maintain a reasonable human-machine correlation even with the reduced feature subset, demonstrating higher robustness as well as efficiency. The accented word ratio is also analyzed and discussed. The comparison between human scores and accented word ratios demonstrates that there is a negative relationship between them, i.e., the higher the accented word ratio the lower the human score. This observation shows that accented words are correlated to the human evaluation of prosody. The quasi-foot segmentation incorporates such kind of rhythm information appropriately and thus improves the prosody evaluation results.

Chapter 4 Generation of Feedback Utterances

4.1 Motivations & Proposed Work

This chapter studies the generation of feedback utterances for non-native English learners. Different methods are used to produce feedback utterances, with experiments performed to obtain stimuli for evaluation.

Although a number of studies have been done in the literature, current accent reduction methods still have a number of problems. One of the major disadvantages of rule-based accent reduction lies in the required intensive study of linguistic rules. Although statistical studies can be easily performed on regional accents of English as in [79, 81], it is difficult to conduct the same study on other accent pairs in which the English corpus from L2 learners is not available, e.g., Vietnamese English, Thai English or Cambodian English. As a result, the rule-based accent reduction can only work for English learners whose regional English corpus is available, compromising its potential for commercialization and application.

Compared to rule-based accent reduction, reference-based system is more flexible as it is not constrained to a specific accent. Rather, the accented speech of the learner is modified according to the reference speech to yield the desired feedback utterances. Hence, the reference-dependent accent reduction is applicable to English learners with various accents. Unfortunately, current studies on reference-dependent accent reduction still leave some issues open to questions. In [76], the accent reduction process is performed on synthesized speech rather than natural speech, which is not the best for language learning purpose. Although [40] and [41] are based on natural speech, the experimental corpus is very limited (only 20 pairs of sentences from the same pair of learner and teacher). Some ambiguities concerning the effects of prosodic and segmental modifications still exist in the literature: while [76] and [40] indicate a higher

importance of prosodic modification, [41] shows an outperformance of segmental modification over prosodic modification. This difference also requires further study. In addition, previous work mainly uses the same synthesis method to perform accent reduction although other synthesis methods are also applicable for this purpose. Hence, an interesting issue is the effects of different synthesis methods for the accent reduction process.

Furthermore, the use of voice conversion techniques for the generation of feedback stimuli has not been studied before. Although the intent of voice conversion is different from that of accent reduction, the scheme can be applied to fulfill the target of accent reduction, i.e., combining the learner's voice and the teacher's linguistic gestures by converting the teacher's voice to that of the learner. As voice conversion only transforms the speaker identity, the native speaking styles of the teacher's speech utterances can be preserved (at least in most of the popular voice conversion methods, such as [77, 139-142]). Therefore, voice conversion technique may be used as a kind of rule-based accent reduction method (as the conversion from the teacher to the learner requires a training process). However, compared with rule-based methods in [79, 81], voice conversion methods just require a small training data set to achieve satisfactory conversion result as shown in [43]. Therefore, it is possible to provide learners with appropriate feedback utterances by using a voice conversion system. As there is a gap concerning this kind of application of voice conversion systems in the literature, it is desirable to study the feasibility of generating auditory feedback using this approach.

Therefore, three main aspects concerning feedback utterance generation are studied in this chapter:

1. Multi-corpora studies of accent reduction: the effects of different accent reduction components, e.g., prosodic or segment modification, across different corpora are studied. As stated before, there is a difference between the effects of prosodic and segmental modification. Because all the previous reference-dependent accent reduction studies are based on a single corpus, a multi-corpora study may give clues to understand the reason that leads to such a difference. In

addition, other influential factors such as nationality of English learners are also analyzed.

2. Accent reduction based on different speech synthesis methods: three popular speech synthesis methods are applied to the accent reduction process and the results are compared in terms of different criteria. This study can provide guidance for the selection of an appropriate synthesis method to generate high quality feedback utterances.

3. Application of voice conversion method to produce feedback stimuli: voice conversion is performed by using parallel speech from native and non-native English speakers to generate feedback utterances. Acoustic quality and accentedness of the converted speech are assessed and studied. Comparison between reference-dependent accent reduction and voice conversion based method is conducted, with pros and cons of each method discussed.

4.2 Multi-corpora Accent Reduction

As stated above, the influences of corpora as well as other factors from the English learners have not been studied in the literature. Therefore, this section performs a multi-corpora study by applying accent reduction on two corpora with different characteristics.

4.2.1 Accent Reduction Scheme

Current implementation of reference-dependent accent reduction process involves the well-known source-filter model, i.e., pitch-synchronous overlaps and adds (PSOLA), for its simplicity and effectiveness [41, 76]. As proposed by this model, a speech utterance can be decomposed into excitation, which is mainly responsible for the prosodic features and speaker identity, and vocal tract filter resonance, which represents the linguistic gestures of speech and partial speaker identity. As a result, accent reduction can be implemented by decomposing speech signals and processing each component (excitation and vocal tract resonance) separately to reduce the perceived foreign accent while retaining the speaker identity.

PSOLA is a kind of non-parametric method, which modifies pitch and duration by directly manipulating speech waveforms, as stated in [83, 143]. This

method decomposes speech signals into pitch-synchronous frames, and the pitch and duration modifications can be performed by overlapping and adding these frames. Hence, the first step of PSOLA is to detect the pitch marks of the speech signal, with each frame containing several pitch periods around each pitch mark.

$$x(s, n) = h(n)x(n - kt_a(s)) \quad (4.1)$$

where $x(s, n)$ is the n -th sample inside the s -th frame of the original signal, $h(n)$ is the analysis window, k is the number of pitch period involved, and $t_a(s)$ is the analysis pitch mark. Normally, a range of two local periods is used for each frame to allow around 50% overlapping across frames. To reconstruct the new signal with desired pitch and duration modifications, those pitch-synchronous frames should be overlapped and added according to new pitch marks:

$$y(u, n) = (1 - a_u)x(s, n) + a_u x(s + 1, n) \quad (4.2)$$

$$a_u = \frac{t_s(u) - t_a(s)}{t_a(s + 1) - t_a(s)} \quad (4.3)$$

Here, $y(u, n)$ is the n -th sample inside the u -th frame of the synthesized signal, a_u is the factor to add analysis frames to generate the synthesis frame which is calculated according to the original and synthesis pitch marks, and $t_s(u)$ is the synthesis pitch mark corresponding to the u -th frame of the synthesis signal. The synthesis pitch mark is decided according to the desired pitch and duration modification factors.

Overall, if the pitch is to be increased, the distance between neighboring frames should be reduced. If the pitch is to be decreased, the reverse manipulation applies. For the duration modification, the pitch-synchronous frames should be repeatedly added, i.e., frames are inserted, to increase the duration, or deleted appropriately, according to the duration modification ratio, to decrease the duration. An illustration of PSOLA based modification is shown in Figure 4.1.

There are mainly three kinds of PSOLA, namely time-domain PSOLA (TD-PSOLA), frequency domain PSOLA (FD-PSOLA), and linear predictive PSOLA (LP-PSOLA), as discussed in [143]. Theoretically, they are essentially doing the same thing – resampling in the frequency domain to change the pitch and duration. However, the implementations are different. TD-PSOLA directly overlaps and adds the pitch-synchronous frames to change the prosodic features, leading to very fast implementation speed but lower quality. In contrast, FD-PSOLA performs the overlapping and adding implementation in the frequency domain. Therefore, interpolation is required to obtain the spectrum of each frame, resulting in lower implementation speed but higher quality. LP-PSOLA first separates the excitation signal from the vocal tract filter using LPC filter. This overlapping and adding is then performed on the excitation to change the pitch and duration. The modified excitations are then combined with vocal tract filters to re-synthesize the speech. This implementation can achieve fast implementation and reasonable quality. Considering those properties described above, LP-PSOLA [83] rather than FD-PSOLA as in [41] is used, as LP-PSOLA possesses a higher efficiency (more than 10 times faster), which is important to a real-time CALL system.

In the modification process, parallel speech signals with the same transcription from two speakers (the teacher and the learner) are needed. Speech signals from both speakers are first decomposed into a series of pitch-

Figure 4.1 Modification of Pitch and Duration Using PSOLA

synchronous frames. Each frame is multiplied by a Hanning window with 50 percent overlap. Unlike [81] and [41] which only use formant features or phase vocoder to obtain the vocal tract features, the 20-order linear predictive coefficients (LPCs) of each frame are used to obtain an accurate vocal tract filter in the time domain. The speech signal is filtered by the inverse LPC filter to obtain the excitation signal. Subsequently, appropriate modifications are performed on excitations to change the prosodic features. Also, the LPCs are substituted to obtain the desired vocal tract features. Finally, the modified excitations and vocal tract filters (i.e., LPCs) are combined to synthesize the accent-reduced speech. The system diagram for accent reduction based on LP-

Figure 4.2 Accent Reduction Flowchart Using LP-PSOLA

PSOLA is demonstrated in Figure 4.2.

4.2.2 Prosodic Modification

In prosodic modification, the duration and pitch contours of the learner's speech signal are modified according to that of the teacher. The energy is left intact as the global energy level is more correlated to the speaker identity while the relative energy levels of different frequency bands are to be modified by

segmental modification. Both the duration and pitch modifications are based on the manipulation of pitch-synchronous frames using LP-PSOLA.

The first step is to obtain the phonetic time alignment. This process is done by forced alignment using HTK [115] and acoustic models trained on TIMIT [144] database. The acoustic models can generate segmentation accuracy of over 96% with 30 ms tolerance for the TIMIT database. Since the accent reduction process modifies pitch and duration of continuous utterances, it is not as sensitive to small alignment errors as a concatenative TTS system which concatenates individual phone units. Hence, the forced alignment results can generally fulfill the requirement of accent reduction.

Following that, the time-scale modification ratio α of each phone is calculated by dividing the teacher's phone length by that of the learner. This ratio α can then be used in PSOLA process to modify the learner's phone durations. To prevent unnaturalness caused by time-scale modification, the duration ratio, α , is constrained to the range of [0.25 4].

Subsequently, pitch modification is performed. The log pitch contour of each of the learner's phone is linearly interpolated to have the same length as that of the teacher, so that the modified phone will have the same number of frames as that of the teacher. Thus, a frame level mapping can be found from the learner's pitch contour to that of the teacher. Suppose the learner's log pitch value to be replaced at time t is denoted as $P^L(t)$ and the aligned teacher's one in the same phone is $\psi(P^T(t))$, with $\overline{P^T(t)}$ and $\overline{P^L(t)}$ as the mean log pitch values of the teacher and the learner, then the pitch-scale modification factor is:

$$\beta(t) = \frac{\psi(P^T(t)) - \overline{P^T(t)} + \overline{P^L(t)}}{P^L(t)} \quad (4.4)$$

It should be noted that the pitch and duration modification factors here are time-varying and calculated on a frame-by-frame basis.

4.2.3 Segmental Modification

The general idea of segmental modification is to replace the vocal tract filter of each frame of the learner by that of the teacher. Two steps are involved in this process.

The first step is the alignment between two speakers' vocal tract features. As the frame numbers of the two speakers' utterances can be quite different, it is necessary to align frame pairs between the teacher and the learner to enable spectral envelope substitution. With the phone boundaries obtained from forced alignment, the alignment process is constrained to each phone of the learner's and the teacher's speech. Two different alignment methods, linear-piecewise alignment and DTW based alignment are proposed by [41] and [76], respectively. However, each method does have some disadvantages. The DTW alignment used in [76] aligns the learner's frames and the teacher's frames inside the same phone using acoustic features of the speech segment. One problem associated with this purely DTW based alignment is the discontinuity of the aligned frames. As shown in Figure 4.3, the dotted line which indicates the purely DTW based alignment aligns one of the teacher's frame to a number of the learner's frames. Therefore, after the substitution of spectral envelope, the difference between frame 2 and frame 3 will be large and thus reducing the smoothness of the synthesized speech. The linear-piecewise alignment proposed in [41] can achieve consecutive alignment by only considering the time-axis of frames inside each phone, but the accuracy of alignment can be lower as acoustic features are not utilized.

Figure 4.3 Interpolation of DTW Alignment.

A possible solution is to combine the linear-piecewise and DTW based alignment. With the combination, DTW based alignment is first performed to obtain the gross alignment results. Then, the teacher's frames which correspond to a number of the student's frames are identified and linear interpolations are performed to achieve smoother alignment results, as shown by the solid line in Figure 4.3. Different kinds of features, i.e., short time spectrum, MFCCs and line spectral frequencies (LSF), are considered and LSF based alignment is finally adopted due to its accuracy and high acoustic quality.

The second step of segmental modification is the vocal tract (or spectral envelope) substitution. After the alignment, the LPC filter of each frame of the learner's speech should be replaced by the new LPCs. As a result of alignment, however, the index of corresponding teacher's frame may not be continuous at phone boundaries and thus distortions may be introduced. Therefore, unlike [41, 76] which combine the modified frames directly, spectral interpolation based on LSF [47] is used to smooth the phone boundaries. Suppose the new LPC coefficient of n -th frame of the learner is $V^L(n)$, and the corresponding teacher's LPC is $V^T(\psi(n))$. Then, the new LPC for n -th frame is:

Figure 4.4 Interpolation of Spectral Envelopes

$$LSF(V^L(n)) = \frac{LSF(V^T(\psi(n))) + LSF(V^T(\psi(n-1)))}{2} \quad (4.5)$$

It means that the new LPC is the interpolation of two neighboring LPCs in the form of LSFs. Therefore, the spectral envelope can become smoother and the distortion can be reduced. This interpolation is shown in Figure 4.4. The spectral envelopes of two neighboring frames at the phone boundary are closer to each other after interpolation, thus smoothing the speech and reducing unnecessary distortions introduced by accent reduction.

Before combining the modified excitation and the new vocal tract filter to complete the segmental modification, vocal tract normalization (VTLN) is performed as in [145] to preserve the learner's formant frequencies. Although previous studies in [146] show that the spectral envelope contains some speaker identity information, it is demonstrated in [77, 145] that VTLN can address this issue effectively. VTLN warps the frequency axis by mapping the formants from the teacher to that of the learner through spectrum interpolations, thus retaining the information related to the learner's speaker identity. It is an effective step to maintain the original vocal features of the learner. The VTLN mapping is demonstrated in Figure 4.5.

Finally, all of these modified pitch-synchronous frames are combined using

Figure 4.5 VTLN Warping to Preserve Speaker Identity

PSOLA algorithm as proposed by [143]. The overall reconstruction process of spectral envelope is shown in Figure 4.6.

Figure 4.6 Segmental Modification by Replacing LPCs

4.2.4 Experimental Conditions

To test the accent reduction scheme, two different corpora are applied, and the generated utterances are assessed in terms of accentedness and acoustic quality. The first corpus is based on BURNC, a prosody abundant database. It consists of 200 students' utterances and 40 teachers' utterances (as references for accent reduction). Students' utterances are recorded in a quiet lab which is not a sound-proof room to simulate the real usage environment of a CALL system. The conditions are listed in Table 4.1.

To study the effects of reference corpus on accent reduction, the other corpus based on CMU_ARCTIC is also used in the experiment. A total of 120 sentences (6 times that of the experiments in [41]) of the speaker pair including Indian speaker WSP and US speaker RMS are selected as the student's speech and the teacher's speech, respectively.

Table 4.1 Experimental Conditions for Accent Reduction

Database	BURNC
Recording Conditions	16k Hz, 16 bit, in a quiet lab
Transcriptions	20 unique sentences selected from BURNC
Students' Utterances	A total of 200 students' utterances from 10 NTU students including Chinese, Indian, Vietnamese and Singaporean
Teachers' Utterances	Utterances from speaker M1B and F2B in BURNC with selected transcriptions

Objective measurements of accentedness and acoustic quality are performed according to [85]. These objective measurements methods are used to generate machine scores for accent-reduced utterances, while comparing the results to human scores given by a number of subjects. The experiments in [85] show human-machine correlations over 0.8 for both the methods, demonstrating their effectiveness and reliability.

4.2.5 Assessments of Accent Reduced Utterances

1. Accentedness Measurement

Accentedness, or perceived accent, can be defined as the deviation of prosodic and segmental characteristics of speech utterances from a specific standard norm, which can be American, British or Australian English. Therefore, a measurement of accentedness can be used to assess the extent of the deviation of the input utterances from the native one, in terms of prosodic and segmental features. The automatic speech recognition technique provides a scheme to conduct this measurement process. The speech recognition process involves the posterior probability score calculated from the observed features and the trained acoustic model, and the score is used to select the most probable output label. The output posterior probability score indicates the probability that a speech segment corresponds to the correct acoustic model trained by native speech. Hence, it measures the deviation of the input speech from native speech (or the

standard norm, which is American English in this case) used to train acoustic models. As pitch, duration and spectral envelope are correlated to the stress, duration and pronunciation of phones which model the acoustic model, the output probability will be affected by modifications on these parameters.

However, the output probability score from speech recognizer is given at the phone level, while the accentedness for the whole sentence is required to assess the performance of the accent reduction methods. Hence, the sentence level score is proposed to be defined as:

$$S_{accent} = -\text{mean} \left\{ \log \frac{p(o_j | \lambda_j)}{p(o_j | \lambda_{max})}; j = 1, 2, \dots, n \right\} \quad (4.6)$$

where S_{accent} is the sentence level accentedness score, o_j is the observed features of j -th phone, λ_j is the correct label of j -th phone, λ_{max} is the phone label which generates o_j with the highest probability, and n is the total number of phones in the sentence. A lower score indicates lower accentedness, i.e., higher nativeness, of an utterance. The mean score rather than median score as in [85] is used to take into account the overall effects of all the scores. Further, the posterior score as suggested in [147] replaces the likelihood score in [85] to give a more accurate evaluation. The reason is that posterior score normalizes the probability score according to the specific features of the sentence, making the assessment results less susceptible to external factors such as the loudness, noise or other distortions in the utterance. This evaluation scheme is similar to the one for pronunciation evaluation as in [28, 30, 71], but here it is used to measure the similarity of the input speech signal to that of the acoustic models trained by native English to reflect the perceived accent. Experiments in [85] show a reasonable human-machine correlation for feedback utterances generated by prosodic, segmental, and combined modifications.

The speech recognition software HTK [115] is used to build acoustic models and perform the forced alignment process. The forced alignment rather than recognition is performed here, because the transcription is already known and the required output is the posterior probability rather than the phone label. The mean accentedness scores for all kinds of stimuli over 200 sentences are shown in

Figure 4.7 Accentedness Score of BURNC Based Corpus

Figure 4.7, with vertical axis indicating accentedness score and horizontal axis indicating different stimuli groups. The mean accentedness scores are connected by straight lines for comparison.

It is obvious that the accentedness score of the speech with prosodic modification and combined modification are lower than that of the students' utterances, showing a reduction of perceived accent. In contrast, speech with segmental modification possesses accentedness score close to that of the students' speech. Therefore, prosodic modification outperforms segmental modification, reducing perceived accent to a further extent. T-tests performed on each pair of stimuli show $p < 0.01$ between segmental modification and students' speech and $p < 0.001$ for all the other stimuli. The p -value is the probability that the null hypothesis, i.e., the difference between scores from two groups of stimuli is not statistically significant, is true.

It can be observed from Figure 4.7 that some of the students' accentedness scores are lower than those of the "teacher" and the generated stimuli. One explanation is that the accentedness scores are generated according to the probability score measured as the distance between individual phones and the corresponding phones of the native speakers used to train the acoustic models.

As a result, the sentence-level accentedness scores are affected by the particular transcript of the utterance. In other word, the accentedness scores for different native utterances can vary, making some of the students' scores lower than those of the "teacher" with different transcriptions.

The experiment based on ARCTIC corpus, however, shows a different pattern of segmental and prosodic modification in Figure 4.8. In the experiment with ARCTIC corpus, t-tests show significant differences for all the pairs of stimuli ($p < 0.01$), except for the pair between original students' utterances and utterances with prosodic modification ($p = 0.4$). Although the difference between the mean accentedness scores of students' and teachers' utterances is statistically significant, it is observed that the difference in ARCTIC is smaller than that in the BURNC based corpus. The main reason of this difference may stem from the flat prosody of this corpus. Figure 4.8 shows a much more significant accent reduction from segmental modification compared with that of prosodic modification, giving the same observation as in [41, 85]. Nonetheless, the reversed observation is true in the experiment with the BURNC based corpus as shown in Figure 4.7. Such conflicting observations can also be found in previous papers ([39] versus [40, 41, 85]), which are performed on different corpora

Figure 4.8 Accentedness Score of ARCTIC Corpus

respectively. It should be noted that multi-corpora experiments with the same accent reduction method are carried out in our study, which is different from the previous studies confined to a single corpus. The different performances of prosodic and segmental modifications on the two corpora shows that the corpus on which accent reduction is based can affect the performance of accent reduction methods.

2. Acoustic Quality Measurement

The assessment of acoustic quality of the modified utterances requires a suitable scheme. For acoustic quality measurement, a number of different standards are available. They include perceptual evaluation of speech quality (PESQ) implemented by ITU P.862 and single-ended evaluation as in ITU P.563 [148]. However, unlike the usual assessment in speech enhancement or communication, the reconstructed speech from accent reduction should not have the same acoustic features as the original one, due to the desired modification of prosodic and segmental features. Therefore, those double-ended assessment methods such as P.862 which compare the output speech with the original speech are not suitable for measuring accent reduction. ITU P.563 is a single-ended method, which reconstructs the input speech to reduce distortions and then assesses the input speech by comparing it with the reconstructed one. Experiments on P.563 show an average human-machine correlation of over 0.85 in terms of acoustic quality of speech coding systems as stated in [148].

Although originally designed for evaluating telephone speech in terms of naturalness of vocal tract and noises, measurements provided by P.563 are also valuable cues for assessing the accent reduced speech in terms of acoustic quality. The basic assessment flowchart of P.563 is given in Figure 4.9. A series of distortion measures are applied in P.563 to identify different types of distortions in the speech signal.

The input speech signal is pre-processed by a voice activity detector (VAD) to identify speech components from noise by setting a dynamic threshold which is iteratively estimated as the mean plus two standard deviations of the power of the noise sections over the whole signal. Normalization is performed to set the energy of the signal to a specific level.

Figure 4.9 Flowchart of ITU P.563

After the preprocessing, pitch period and voiced/unvoiced decision are also estimated. The vocal tract is then modeled by calculating pitch-synchronous linear predictive coefficients and cepstral coefficients. Those parameters are then transformed into a vocal tract area function with eight acoustic tubes using reflection coefficients [148]. The obtained acoustic tubes are further divided into three groups, i.e., rear (1-3), middle (4-6), and front (7-8) cavities. Parameters are obtained from these cavities to describe their sizes and rates of changes. Such kinds of information is used to measure the distortion and unnaturalness of the vocal tract features. In addition, each frame denoted as speech by VAD is also modeled by 21-order LPCs and the kurtosis and skewness are calculated. The average value of skewness and kurtosis across speech frames are then checked to see if they are within the expected range for natural speech.

The second component is a simulation of the intrusive (double-ended) method, providing a reference signal which is generated from a speech reconstruction module. The speech is reconstructed by first calculating the LPCs of each frame.

The LPCs are converted to LSFs and fitted to the typical vocal tract model of a human speaker. A moving average prediction based on two codebooks is used to quantize the obtained LSFs and the better of the two LSF approximations is applied to reconstruct the speech signal. A psychoacoustic model is used to compare the reconstructed reference speech with the input speech to measure the distortions removed by the reconstruction module.

Finally, various distortion-specific parameters are calculated to give a more comprehensive assessment of acoustic quality of the speech signal. Firstly, the signal to noise ratio (SNR) is calculated to identify very noisy sections in the speech signal. Then, robotization is estimated to assess the redundant periodicity in the speech signal by calculating the energy and cross-correlations among frames. Next, two types of temporal clipping are measured, including muting or interruption of speech and clipping of the front or back end of the speech signal. They are detected by analyzing the abrupt changes in the envelope of the speech signal. The last measurement is the signal correlated noise which measures multiplicative noise. This analysis process involves the evaluation of spectral statistics and uses spectral level range/deviation as the indicator of noise.

After that, a perceptual mapping is applied to include all of the above measures in calculating final mean opinion score (MOS) with a classification and regression model. The estimated quality is normalized to a standard MOS scale in the range of 1 to 5. Application of this standard on accent reduction shows an human-machine correlation over 0.8, as in [85].

To compare the performance of the proposed accent reduction scheme for different corpora, the MOS of both the BURNC and ACTIC based corpus are calculated with different modifications. The segmental modifications with and without interpolations are also compared in order to evaluate the performance of the proposed interpolation procedure. The MOS of different stimuli groups generated by the two corpora are shown in Figure 4.10.

The MOS of teachers' utterances are not calculated as accent reduction is only performed on students' utterances. Accent reduction leads to a degradation of acoustic quality of modified speech and MOS is used to reflect the quality of the modified speech. Results show that segmental modification introduces more

Figure 4.10 Acoustic Quality MOS by P.563

distortions than prosodic modification, and combined modification results in the highest degradation of acoustic quality. Such results are also consistent with those reported in previous studies [41, 85] which use the ARCTIC corpus. In addition, it can be observed that the accent-reduced speech without interpolations has a lower MOS compared to the one with interpolations, showing that the alignment and spectral interpolations are a desirable step to enhance the acoustic quality. The differences between each pair of MOS shown in Figure 4.10 are statistically significant ($p < 0.001$).

4.2.6 Effects of the Nationality of Learners and the Corpus

The evaluation results demonstrate that the perceived accent of modified speech utterances is reduced. The performance of segmental modification for BURNC based corpus, however, is worse than prosodic modification, which is different from the results obtained from the ARCTIC corpus as reported in [41].

The first possible explanation is the difference in the nationality and characteristics of English learners. Most of the foreign students in the BURNC based corpus are Chinese, Singaporean, and Vietnamese (8 out of 10), whose pronunciations are generally acceptable whereas the prosody of their utterances is poor. In contrast, the non-native speaker in the ARCTIC corpus is an Indian whose utterances have prosody close to that of a native speaker but with more issues in pronunciation. In Figure 4.11, the reduction of accentedness scores using prosodic modification are plotted for all the 10 students in the BURNC based corpus. Vertical dash lines show the maximum and minimum of reduced accentedness and boxes show the 80 percent ranges. Mean values are shown by the central dotted line. It is clear that the reduction in accentedness of student 6 and student 7 (the two Indian students) are not obvious compared to all the other 8 students. The horizontal line across Figure 4.11 indicates the mean level of

Figure 4.11 Reduction of Perceived Accents for Different Students

accentedness reduction of students 6 and 7. It can be observed that the mean reductions of accentedness for all the other students are much higher than the level specified by this line, i.e., the reduction of accentedness introduced by prosodic modification in the utterances of other students are more significant in comparison with that of students 6 and 7.

The other explanation for the difference could be the reference corpora used. Unlike ARCTIC, BURNC was collected primarily to support the generation of prosodic patterns for text-to-speech synthesis systems with abundant prosodic features. Since the prosody is comparatively flat in the ARCTIC corpus, the impact of prosodic modification is lower than that in BURNC. The flatness of prosody in ARCTIC has also been cited to explain the subjective evaluation results in [41].

4.2.7 Effects of Different Accent Reduction Setups

According to the experimental results, there is a tradeoff between accentedness reduction and speech quality: combined modification introduces more distortions but reduces accentedness to a further extent than segmental or prosodic modification alone. Although prosodic or segmental modification alone reduces accentedness to a lesser extent, it maintains a higher acoustic quality. Hence, the preference of language learners (i.e., quality vs. nativeness) should be considered when performing accent reduction.

Secondly, the nationality and characteristics of students are another issue. For students whose prosody is close to that of native speech but with pronunciation issues (e.g., the Indian speaker in ARCTIC corpus), segmental modification is more desirable. In contrast, students with unnatural prosody like Singaporean or Chinese speakers in the BURNC based corpus would prefer prosodic modification. The most effective accent reduction method can thus be selected by addressing the specific weakness of a English learner.

Finally, the reference corpus can also affect accent reduction significantly. A prosody abundant reference corpus requires more efforts in producing the desired prosody, thus prosodic modification is important. In contrast, a corpus with less

prosody dynamics reduces the difficulties in imitating the native prosody, thus enabling students to concentrate more on the practice of pronunciations.

Therefore, during the generation of feedback stimuli using accent reduction, the learner's preference, nationality, as well as the available reference corpus should all be taken into account.

4.3 Feedback Utterances Generated by Different Synthesis Methods and Voice Conversion Method

Although a number of studies about reference-based accent reduction have been presented [41, 76, 81], all these papers are based on the same synthesis method (PSOLA) which may have some limitations in producing feedback utterances with desired quality and effectiveness. To reduce the perceived accents in the learner's speech signals effectively, other speech synthesis methods can also be considered in addition to PSOLA. They include harmonic stochastic model (HSM) and speech transformation and representation by adaptive interpolation of weighted spectrogram (STRAIGHT).

However, unlike PSOLA which separates the original excitations and the vocal tract filters of the learner, HSM and STRAIGHT do not separate the information responsible for speaker identity from that responsible for linguistic contents. In HSM, the speech signal is decomposed into a series of harmonic signals whose frequency is integer times of the fundamental frequencies plus a stochastic term. Due to such kind of property, HSM can be easily applied to prosodic modification by interpolating the spectral envelope cross frames, but it is difficult to be used for segmental modification without changing the original speaker identity. Although STRAIGHT also works as a kind of source-filter model, it only separates the original spectrogram from fundamental frequencies of the learner while reconstructs excitations using random noise and values of the fundamental frequency based on zero-phase filters. Because the excitations are reconstructed in the synthesis process, the original excitations are not preserved as in the PSOLA method. As a result, all the speaker identity information is included in the spectrogram in STRAIGHT and thus linguistic contents cannot be modified separately. Actually, PSOLA is a non-parametric model which

manipulates the original speech signals, while HSM and STRAIGHT are parametric models which use a set of parameters to represent the speech signal. The substitution of parameters in HSM and STRAIGHT can lead to the change of perceived speaker identity which inconsistent with the objectives of accent reduction, i.e., reducing accentedness while maintaining the original vocal features.

Those properties of HSM and STRAIGHT prevent them from separating the segmental features and vocal features as in PSOLA method. Thus, it is not appropriate to apply these two methods to modify the segmental features of the learner's utterances. As stated in [39, 40, 149] as well as the results shown in Figure 4.7, however, prosodic features play the main role of the perceived accents in the non-native speech under the scenario of abundant prosody. Therefore, it is still worthy to study the possibility of improving feedback utterances with different speech synthesis methods. In addition, an alternative method to achieve the desired feedback is to convert the vocal features of the teacher's speech to those of the learner by using a voice conversion method [43]. Although voice conversion techniques have been in existence for years, such an application has not been proposed and studied in the past.

4.3.1 Accent Reduction Methods

This section is primarily focused on the modification of prosodic features. The overall scheme is similar to that in Figure 4.2, but only prosodic modification is performed and different synthesis methods are applied, as shown in Figure 4.12. In this modification scheme, a learner's utterance and the corresponding teacher's utterance are forced aligned to obtain the corresponding time alignment for each phone. Then, different synthesis methods are used to separate the prosodic features from the segmental features, i.e., spectrogram, so that the prosody-based accent reduction can be performed.

The phone-level time alignments, pitch modification factors, as well as duration modification factors are obtained in the same way as described in the previous section. Each phone of the learner's utterance is modified by the corresponding phone in the teacher's utterance, in terms of pitch and duration.

Figure 4.12 Scheme of Prosodic Modification with Different Synthesis Methods

The modified prosodic features are combined with the original spectral information to generate the accent-reduced feedback utterance. The same scaling factors are applied to all the three synthesis methods. Brief introductions of the three synthesis methods used are given as follows:

1. PSOLA. The first method is PSOLA which is discussed in the last section. Here, the same implementation of LP-PSOLA is performed in our study to modify the prosodic features.

2. HSM. The second method is HSM [43], which is based on harmonics and stochastic components of the speech signal. HSM decomposes the speech into different frequency components, in a way similar to the classical sinusoidal model. However, the difference is that the sinusoidal model simply describes the speech signal as a summation of multiple sine or cosine signals plus noise terms, while HSM describes the speech signal as a summation of harmonics of the fundamental frequencies plus a stochastic term. The formula of HSM is shown below:

$$x^{(k)}[n] = \sum_j A_j^{(k)} \cos(j2\pi f_0^{(k)}n + \varphi_j^{(k)}) + \sigma[n] \cdot e_{LPC}^{(k)}[n] \quad (4.7)$$

where j is the index of harmonics, k is the frame index, A_j is the amplitude, f_0 is the fundamental frequency, φ_j is the phase, e_{LPC} is the LPC filter of residuals, and σ is the white noise. Therefore, a HSM frame mainly has two parts, the

harmonics part and the stochastic part. The harmonic part is a number of cosine signals whose frequencies are harmonics of the fundamental frequencies. The amplitude and phase terms of each individual harmonic are calculated in a way to minimize the difference between the original speech signal and the reconstructed signal using those harmonic terms. In addition to the harmonic component, the stochastic component represents the information which cannot be described by harmonics, because speech signal is not perfectly periodic and not always has a stable structure as in voiced phones. Those stochastic terms are described by the LPC filters of the residuals, which come from the difference between original signals and the harmonic components.

Supposing the frame stepsize is N samples, the reconstructed signal is denoted as the interpolation of neighboring HSM frames with modified amplitude, frequency, and phase:

$$x_o^{(k)}[n] = \sum_j A_j^{*(k)} \cos(j2\pi f_0^{*(k)} n + \varphi_j^{*(k)}) + \sigma[n] \cdot e_{LPC}^{*(k)}[n] \quad (4.8)$$

$$x(kN + m) = \frac{N - m}{N} x^{(k)}[m] + x^{(k)}[m - N] \quad 0 \leq m \leq N \quad (4.9)$$

In the reconstruction stage, a white noise signal is to be filtered by the LPC

Figure 4.13 HSM based Prosodic Modification

filter, generating the stochastic component. This stochastic component is combined with the harmonics calculated earlier to obtain the reconstructed speech. The flowchart of HSM based modification is show in Figure 4.13.

The modification of pitch and duration in HSM is implemented by interpolations among HSM coefficients. Because fundamental frequency is given by the f_0 coefficient in HSM, a desirable change of f_0 can lead to the modification of pitch value. However, because the spectral envelope should remain the same, the amplitude coefficients A_j of different harmonics must be resampled to maintain the original spectral components so that the spectral envelope of the original speech signal is not subject to change. The duration modification can be done by the insertion or deletion of certain frames as in PSOLA method.

3. STRAIGHT. The third method is STRAIGHT as proposed in [150]. It is a high-quality speech synthesis method which can be viewed as an advanced version of a phase vocoder. It removes the periodicity from the spectrogram to enable flexible manipulations or reconstructions of prosodic features. The general formula of STRAIGHT is:

$$x(t) = \sum_k e_k(t) \cdot s_k(t) \quad (4.10)$$

where e_k denotes the excitation generated from the new pitch contour which is calculated from the scaling factors; s_k denotes the spectrogram of the learner's utterances after the pitch periodicity is removed; and k is the frame index. Different from PSOLA, the s_k excitation in STRAIGHT is totally reconstructed from the given pitch value by filtering white noise signal using minimum phase filters. The produced excitation is then combined with the periodicity-free spectrogram obtained from the analysis stage to generate the modified speech signal. A more detailed formulation of STRAIGHT is shown below:

$$y(t) = \sum_{t_i \in Q} \frac{1}{\sqrt{G(f_0(t_i))}} s_{t_i}(t - T(t_i)) \quad (4.11)$$

$$v_{t_i}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} S(w, t_i) \Phi(w) e^{jw(t)} dw \quad (4.12)$$

Figure 4.14 Synthesis Scheme of STRAIGHT Model.

$$T(t_i) = \sum_{t_k \in Q, k < i} \frac{1}{G(f_0(t_i))} \quad (4.13)$$

where Q represent the positions of excitations in the synthesis, $G(\cdot)$ is the pitch modification factor, $S(w, t)$ is the spectrogram, $\Phi(w)$ is the all-pass filter for fine tuning of pitch and temporal information.

The whole synthesis scheme of STRAIGHT is presented in Figure 4.14, which involves the modeling of the original speech signal and the reconstruction of the new speech signal with modified prosodic features. The high quality of STRAIGHT mainly results from two aspects -- the elimination of pitch mark detection in the analysis process and the use of group delay all-pass filters in the synthesis process for the fine control of pitch and excitations, as described in [150, 151].

4.3.2 Voice Conversion Method

An alternative way to generate desirable feedback utterances stems from voice conversion techniques. Voice conversion deals with vocal features which are correlated to speaker identities, leaving prosody and pronunciation unchanged. By performing voice conversion on the teacher's utterances to transform the voice of the teacher to that of the learner, the output utterances can possess both the teacher's linguistic gestures and the learner's voice.

The voice conversion method used in this study is mainly based on [152], which proposes high quality voice conversion using GMMs. This method models the transformation between the source and target speakers using a number of mixtures of multi-variant Gaussian models (in our case 16 mixtures are used). Each individual multi-variant Gaussian model has its own covariance matrix and mean value calculated from the training data. Line spectral frequencies (LSFs) converted from HSM parameters as in [43] are used as the feature vector. It is assumed that there are n frames of speech from the source speaker $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, with $\mathbf{x}_k = [x_1, x_2, \dots, x_d]^T$, and m frames of speech from the target speaker $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m]$, with $\mathbf{y}_k = [y_1, y_2, \dots, y_d]^T$. Here, d is the number of feature dimensions and speech signals from two speakers have the same transcription. Base on the trained GMMs and the feature vectors of the source speaker, the feature vector of converted speech can be obtained as:

$$\tilde{\mathbf{y}}_k = f(\mathbf{x}_k) = \sum_{i=1}^M P_i(\mathbf{x}_k) [\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx} (\boldsymbol{\Sigma}_i^{xx})^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_i^x)] \quad (4.14)$$

where $\mathbf{x}_k$ is the feature vector of k -th input frame, $P_i(\mathbf{x}_k)$ is the probability of $\mathbf{x}_k$ belonging to i -th mixture, $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ are mean vectors of the i -th Gaussian mixture, $\boldsymbol{\Sigma}_i^{xx}$ and $\boldsymbol{\Sigma}_i^{yx}$ are sub-matrices of the covariance matrix of the i -th Gaussian mixture estimated as in Equation (4.18), M is the number of mixtures, and $\tilde{\mathbf{y}}_k$ is the converted feature vector. For each frame of the input speech signal, the posterior probability that this frame belongs to each mixture is calculated according to the GMM and the feature vector of this frame. Then, a linear transformation proposed in Equation (4.14) is applied to map the feature vector from the source speaker to that of the target speaker. After that, all the transformed feature vectors for each mixture are combined by a weighted summation, with the weight set as the previously calculated posterior probability, i.e., $P_i(\mathbf{x}_k)$. The output feature vector is the transformed one which represents the vocal features of the target speaker.

To obtain $P_i(\mathbf{x}_k)$, the GMM can be obtained by estimating the distribution of $\mathbf{Z}$, which is the combination of $\mathbf{X}$ and $\mathbf{Y}$, i.e.,

$$P_{GMM}(\mathbf{z}_k) = \sum_{i=1}^M c_i \mathcal{N}(\mathbf{z}_k, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (4.15)$$

$$\mathbf{z}_k = [\mathbf{x}_k^T \mathbf{y}_k^T]^T \quad (4.16)$$

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \quad (4.17)$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \quad (4.18)$$

where c_i is the weight of i -th mixture which satisfies $\sum_{i=1}^M c_i = 1$. The symbol $\mathcal{N}(\mathbf{z}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes a normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. All the parameters, i.e., $(c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ representing the joint density of $\mathbf{X}$ and $\mathbf{Y}$, can be estimated from all the observation of $\mathbf{z}_k$ by using the Expectation-Maximization (EM) algorithm. After calculating these parameters, mean and covariance matrix used in Equation (4.14) can then be obtained from Equations (4.17) and (4.18). The probability of the input feature vector belonging to i -th mixture is calculated as:

$$P_i(\mathbf{x}_k) = \frac{c_i \mathcal{N}(\mathbf{x}_k, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^M c_l \mathcal{N}(\mathbf{x}_k, \boldsymbol{\mu}_l^x, \boldsymbol{\Sigma}_l)} \quad (4.19)$$

Substituting parameters obtained from Equations (4.17) and (4.18) along with $P_i(\mathbf{x}_k)$ calculated from Equation (4.19) into Equation (4.14), the output feature vector can be generated from the input feature vector $\mathbf{x}_k$. These output feature vectors can be used to synthesize the speech signal of the target speaker. Various synthesis schemes can be applied to decompose the speech signal into frames as well as to resynthesize the speech signal from the converted feature vectors.

The overall voice conversion scheme is shown in Figure 4.15. In the training stage, a number of parallel speech samples from both the learner (target speaker) and the teacher (source speaker) should be gathered first. Those speech samples are analyzed and decomposed into frames. Each frame is then represented by LSFs which are common features used for voice conversion, to train the GMMs. Each LSFs vector represents one frame of speech from either speaker. All the parameters which represent the conversion GMM, including weights, means and covariance matrices, are estimated from the stacked feature vectors from two speakers using the EM algorithm.

Training Scheme:

Conversion Scheme:

Figure 4.15 Voice Conversion Scheme

In the conversion stage, the teacher's speech signal will be analyzed and parameterized into LSFs frame by frame. Next, the LSFs are converted using the conversion formula as in Equation (4.14) to generate the new feature vectors representing the learner's vocal features. Those LSFs are then applied to synthesize the output speech signal whose vocal features are more similar to that of the learner's.

4.3.3 Comparison of Different Feedback Utterances

Experiments are performed to produce feedback utterances, by either using accent reduction based on different speech synthesis methods or the voice conversion approach, as well as to assess the generated stimuli according to several criteria. The experimental corpus contains 40 teachers' utterances (as the reference speech for accent reduction or the source speakers for voice conversion) and 200 non-native speakers' utterances as listed in Table 4.1. The same objective measurements of accentedness and acoustic quality based on [85] are used, as stated in the previous section. The HSM and STRAIGHT methods used

in this chapter are implemented by toolkits obtained from [153] and [154], respectively.

1. Accentedness Measurement

The mean accentedness scores for all the stimuli groups are shown in Figure 4.16. These stimulus groups include original learners' utterances, original teachers' utterances, and feedback utterances generated by either modifying learners' utterances with three synthesis models or converting the vocal features of teachers' utterances to those of learners using voice conversion method. The difference in accentedness scores between each pair of stimuli is statistically significant (t-tests show $p < 0.005$), except for the pair of teachers' speech and stimuli generated by the voice conversion method. After prosodic modifications using any of the three different speech synthesis models, the mean accentedness score is reduced, showing an improved nativeness. Compared to the other two synthesis methods, STRAIGHT achieves the lowest score, i.e., the highest nativeness. This should be due to the different working schemes of the three methods: STRAIGHT uses a new pitch contour to reconstruct the excitation for synthesis, whereas the other two models either overlaps and adds (PSOLA) or interpolates (HSM) original speech signal to change the intonation contour.

Figure 4.16 Accentedness Scores of Different Stimuli.

Because the excitations in the STRAIGHT method are generated from the new pitch contour using minimum phase filters and not influenced by the original (learner's) pitch contour, the synthesized speech has an intonation contour which is closest to the native one, leading to a higher nativeness. In contrast, PSOLA and HMS modify the original excitations by overlapping and interpolating, inevitably making the synthesized speech to contain prosodic features of the learner which affect the perceived accents.

The feedback utterances generated by voice conversion method demonstrate a low mean accentedness score which is similar to that of the original teachers' speech. This is expected as only the vocal features are converted while the prosody and pronunciation remain unchanged. Compared to accent reduction which improves prosodic features, stimuli generated by voice conversion also possess correct pronunciations, resulting in higher nativeness.

In addition, the mean accentedness score generated by combined modification based on PSOLA, i.e., both prosodic and segmental features are modified accordingly as described in the previous section, is also presented here for comparison. Although the inclusion of spectral envelope substitution in PSOLA based modification can reduce accentedness to a further extent, the prosodic modification performed by STRAIGHT is more effective. Therefore, the gap between PSOLA based combined modification and STRAIGHT based prosodic modification is not big even though the former still outperforms the latter in terms of nativeness. Compared to the voice conversion method, the accentedness score given by the combined modification is still significantly higher. The reason is that the spectral envelope substitution cannot retain all the linguistic gestures of the teacher, which are nearly fully preserved by the voice conversion method.

It can be noted that the teacher's accentedness score are also positive and far from 0. The reason is that native speech utterances used in the experiment cannot have exactly the same acoustic features as the samples used to train acoustic models. Because the acoustic modeling process are also affected by the gender, speaker identity, transcriptions, or other conditions in the training dataset, even the accentedness scores of the teacher cannot be very close to 0. The positive value of teachers' accentedness score can thus be considered as some systematic

biases between the acoustic models and the testing database. However, the significant and consistent difference between the teachers' and learners' accentedness score in our testing set can account for the non-native biases of the learners' speech, which results in stronger perceived accents.

2. Acoustic Quality Measurement

Acoustic quality is assessed by the MOS obtained from ITU P.563. Figure 4.17 shows the MOS of different stimuli groups, with t-tests showing a significant difference ($p < 0.01$) for each pair of stimuli. The acoustic quality of the modified speech is degraded when using accent reduction methods. However, the quality of the modified speech using STRAIGHT is closest to that of the original learners' speech. Therefore, STRAIGHT can maintain a higher quality than the other two methods. The first reason is that STRAIGHT uses new generated excitations for speech synthesis and thus avoids the distortions as introduced by overlaps and adds in PSOLA or interpolations in HSM. In addition, the group-delay manipulation used in STRAIGHT, which enables finer pitch and excitation signal control by using phase interpolation, also leads to the higher acoustic quality [150, 151].

Again, the acoustic quality of feedback utterances generated by the combined modification is also listed here for comparison. The acoustic quality from this

Figure 4.17 Acoustic Quality of Different Stimuli.

method is significantly lower than the three prosodic modification methods due to the replacement of spectral envelope. Therefore, although the combined modification using PSOLA can reduce the perceived accent more significantly compared with prosodic modification using STRAIGHT, the resulted acoustic quality is much lower. This tradeoff between the two modification methods should be considered in the selection of feedback utterances, depending on the learner's preference for nativeness or acoustic quality. However, the quality of feedback utterances generated by the PSOLA combined modification is higher than that given by the voice conversion method, as the whole spectral envelope is directly substituted without manipulations, leading to less spectral distortions compared to voice converted speech signals. The stimuli given by voice conversion have the lowest MOS, due to the distortions and over-smoothness effects, i.e., the muffled voice resulted from the statistical average of parameters, introduced by the transformation of spectral information via GMMs.

4.3.4 Accented Word Comparison

Another issue that must be considered in the evaluation of generated feedback utterances is the change of pitch accents. As an important linguistic phenomenon in English, pitch accent is an important factor to determine the nativeness of English. It has been demonstrated in Chapter 3 that the accented word plays an important role in indicating the nativeness of the learners' utterances. Therefore, an experiment is performed to test the pitch accent of original utterances from the learner and that of generated feedback utterances using various methods. This assessment is mainly related to the nativeness of prosodic features and thus suitable for comparison of prosodic modifications with different synthesis methods stated in this section.

The pitch accent detection system using logistic regression based on [110], as stated in Chapter 3, is performed to detect accented words. The accuracy of the detection system is 0.81 in terms of mean F-measure. Based on the trained pitch accent detection system, all the accented words of the original students' speech, teachers' speech and generated feedback utterances are detected. Each word is labeled by 0 or 1 to indicate whether it receives pitch accent or not. Therefore, each utterance can be represented by a binary sequence to indicate the status of

Figure 4.18 Pitch Accent Similarity Ratio – Percentage that the Learner Have the Same Pitch Accent Label as the Teacher

each word, i.e., accented or unaccented. Then, the assigned accented word labels of each of learners' utterances and feedback utterances are compared to the corresponding utterance given by the teacher with the same transcription. The results are presented as a percentage which indicates the similarity between the accented word labels of the stimuli group and the teachers' accented word labels, as shown in Figure 4.18.

It is observed that, compared to the learners' utterances, the generated stimuli exhibit pitch accents which are more similar to that of the teacher. Particularly, the accented word similarity ratio of STRAIGHT generated utterances is significantly higher than that of PSOLA and HSM generated utterances, showing that this method has better ability to imitate the teacher's rhythm. Stimuli given by voice conversion approach possess the highest similarity ratio, as the converted utterances have linguistic gestures similar to that of the teachers' utterances.

Because the teachers' pitch accents are considered to be correct and native, the pitch accents of the feedback utterances are improved. Hence, this

experiment also demonstrates that the nativeness of generated feedback stimuli is significantly higher than that of the learners' utterances. The same pattern as that in the previous accentedness measurement, i.e., voice conversion is better than STRAIGHT and STRAIGHT is better than PSOLA & HSM, can be observed from the similarity ratios in this experiment. T-tests are also performed on each pair of stimuli groups with $p < 0.001$, showing significant differences for different stimuli groups.

4.3.5 Discussion

From the experimental results on feedback utterances generated by different synthesis methods and the voice conversion technique, it can be found that the STRAIGHT method is the best for modifying the prosodic features of accented utterances because it results in the highest acoustic quality and the greatest reduction in accentedness. In addition, feedback utterances given by voice conversion approach sound more native like, resulting from the fully preserved linguistic gestures from the teacher. However, the lower quality of speech generated by voice conversion method is still an issue. Furthermore, the preservation of the learner's voice in feedback utterances should also be examined. A comparison of spectrograms, which provides useful cues about speaker identity, is shown in Figure 4.19.

The duration of the original learner's speech is longer as the teacher speaks faster than the learner. It is obvious that the spectrogram of the accent-reduced speech using STRAIGHT is similar to that of the original one. As a result, the learner's speaker identity is almost fully preserved. In contrast, the utterance generated by voice conversion method shows a smoothed spectrogram that is significantly different from the learner's one. Although some of the differences may result from the different pronunciations of two speakers, the obvious over-smoothness in the spectrogram introduced by GMMs obscures the speaker identity as well. In fact, an informal subjective listening test suggests that the voice of the converted speech is in-between that of the teacher and the learner: the converted speech still keeps some of the teacher's vocal features, even though the voice is more similar to that of the learner. What's more, the training corpus, i.e., a number of utterances produced by the learner, required by voice

Figure 4.19 Spectrogram Comparison: (a) *Original Speech of the Learner* (b) *Feedback Utterances Generated by STRAIGHT Prosodic Modification* (c) *Feedback Utterances Generated by Voice Conversion*

conversion method may be difficult for English beginners who cannot speak fluently. Those characteristics lead to some difficulties for the application of voice conversion based feedback utterances in a CALL system for English beginners.

Arising from the above observation, a feedback system which encompasses the two methods may provide learners with more choices on auditory feedback. The proposed scheme is shown in Figure 4.20. At the beginning stage, the STRAIGHT based accent reduction method can be used. With the preserved speaker identity and the improved prosody, the learner can imitate the generated stimuli to improve their speaking skills. After a period of training, it will be possible to gather the learner’s utterances to train a voice conversion system. Thus, voice conversion method can then be used to generate feedback stimuli with both prosodic features and pronunciations closer to the standard, though at the cost of some acoustic quality degradation. Although this scheme has not been

Figure 4.20 Proposed Feedback Scheme

verified through pedagogical studies due to the limit of time and resources, it proposes a promising direction for designing adaptive CALL systems and will be studied in the next stage of our research.

The accent reduction method in the first stage can also be determined according to the preference of the learner. If the learner prefers nativeness to acoustic quality, the combined modification based on PSOLA can also be applied in the accent reduction stage, as this accent reduction method yields the highest nativeness apart from the voice conversion method. However, the improvement of nativeness of this modification scheme lowers the acoustic quality in comparison with STRAIGHT method, as shown in Figure 4.16 and Figure 4.17.

4.4 Summary

This chapter studies accent reduction techniques which generate reference speech with improved nativeness while retaining the learner's vocal features for language learning purposes. To study the effects of different modifications and external factors in the accent reduction process, a multi-corpora experiment which has not been proposed before is performed. Both prosodic and segmental modifications are used to reduce the accentedness of speech. Spectral envelope interpolations are employed to enhance the quality of the generated feedback

utterances. Results from different corpora demonstrate the effects of the reference corpus and the nationality of the learner on accent reduction. A tradeoff between prosodic and combined modifications is also observed. Hence, these issues should be considered in the construction of a speech corpus for language learning purpose as well as the selection of the most appropriate accent reduction method for a specific language learner or a group of language learners sharing some common characteristics.

Different synthesis methods are also compared in order to determine the most suitable technique for accent reduction in the form of prosodic modification. Moreover, the voice conversion method is proposed to generate feedback utterances for English learners with an alternative approach. Objective measurements show that the STRAIGHT method is the best for accent reduction in terms of both nativeness and acoustic quality. Feedback stimuli generated by the voice conversion method possesses the highest nativeness, but it yields the lowest acoustic quality and a partial loss of the learner's speaker identity. In addition, the training corpus required by voice conversion creates a difficulty for English beginners. Therefore, both accent reduction and voice conversion methods have their advantages and disadvantages.

To take advantage of the two methods in the feedback generation process, a multi-stage feedback scheme is proposed to facilitate the learning process of non-native English speakers. Accent-reduced utterances are used as the feedback stimuli at the beginning stage, whereas voice conversion generated feedback utterances are used at the later training stage. As a result, the learner can receive feedback with increased nativeness and will gradually get used to the generated feedback utterances.

Chapter 5 A Hybrid Scheme for Automatic Phonetic Segmentation

5.1 Motivations & Proposed Work

Although prosodic segmentation is more effective in prosody evaluation as demonstrated in Chapter 3, accurate phonetic segmentation can facilitate computer-aided acquisition of speaking skills from different aspects. An automatic segmentation system can precisely identify the phone boundaries of learning materials assigned by teacher which normally are speech utterances without labeling. The obtained segmentations along with the segmentations of the learner's speech utterance can help learners compare the waveforms of their uttered phones with those from the learning material. Another application of phonetic segmentation is in pronunciation evaluation which focuses on the evaluation of the learner's ability to produce phones in continuous speech. This application requires the segmentation of speech signals into different phones so that the relevant features can be compared with trained assessment models. Therefore, the improvement of phonetic segmentation system is desirable for modern CALL systems. In addition, a natural text-to-speech system may provide language learners with flexible learning materials for their practices and thus is useful for CALL. The synthesis of natural speech utterances from input transcriptions always require concatenation of individual phones with different contexts, thus requiring an accurate phonetic segmentation system in the training process.

Therefore, phonetic segmentation is studied in this chapter to facilitate other aspects of CALL in addition to the evaluation and feedback components as discussed in the previous two chapters. This chapter mainly focuses on the refinement of phonetic segmentations given by a HMMs-based text-dependent segmentation system, as transcriptions are generally available for CALL purposes and can contribute to the segmentation process. As reviewed in Chapter

2, post-processing methods such as statistical correction and predictive models can be applied to any existing segmentation system directly, e.g., using a trained speech recognizer, leading to increased convenience and flexibility for the segmentation process. This kind of methods also imitates the segmentation process of human labelers: first listening to the speech signal to get the rough boundary (baseline system), then examining the waveform and spectrogram in detail to identify the accurate boundary (post-processing). Therefore, the main focus of this chapter is the development of a hybrid post-processing scheme to overcome some gaps in current studies so as to improve the segmentation results obtained from a baseline system.

Recent studies on post-processing of text-dependent segmentation results still leave some issues to be addressed. The statistical correction proposed in [91] neither studies the performance of this method on monophone models nor examines the influence of different search ranges on statistical correction. It also only experiments statistical correction on a single-speaker subset without examining the overall performance of statistical correction over a multi-speaker test set. Simple statistical correction based on fixed shifts of boundaries is also applied in some studies like [98], but this kind of correction does not take into account dynamic information of each observation and thus cannot provide very effective corrections. As reported in [100-102], fusion methods are used to combine a number of different HMMs to refine the phonetic segmentation results. A problem with this kind of methods is that the segmentation process can be time consuming because segmentation results must be generated by many different HMMs (55 to 112 in these works) and then merged together. In addition, only study reported in [101] uses a speaker-independent corpus, i.e., the TIMIT corpus [144], while work in other papers only tests on data from a single speaker. The predictive model alone also cannot incorporate the global statistics as efficiently as statistical corrections and may suffer from low implementation speed. Hence, it is desirable to combine different post-processing methods appropriately so that the refinement process can be implemented with increased effectiveness and flexibility.

A hybrid scheme combining three kinds of methods, including statistical correction which provides efficient corrections on the segmentation results, the fusion method which enables different models to compensate for biases of each other, and the predictive model which uses acoustic features to perform refinement based on a classification process, is proposed in this chapter to provide accurate and flexible refinements of phonetic segmentations. As a result, segmentation errors, especially those within small tolerances, are compensated from three aspects: (1) addressing systematic biases of acoustic models using local information from the most relevant range, (2) incorporating complementary effects from acoustic models with various stepsizes which are demonstrated to affect the results significantly, and (3) correcting non-systematic segmentation biases using predictive models.

To improve the segmentation accuracy and effectiveness, some issues are also studied and addressed in this hybrid scheme. Different from the study reported in [91], the search range of statistical correction is considered and a state selection step is proposed to exploit the benefits of different search ranges, thus improving the segmentation results. Although fusion methods have been used in previous work [100-102], none of them studies the effects of different stepsizes of HMMs on phonetic segmentation. However, this kind of information does contribute to the segmentation process. The proposed fusion method just combines limited segmentation machines to reduce the processing time in comparison with work in [100-102] which combines from 50 to over 100 HMMs. In addition to the two methods, a stepsize smaller than that of the HMMs is used in a classifier to further refine the phone boundaries. The use of a smaller stepsize in the final step aims to correct small errors ignored by the previous steps, while maintains the efficiency as only a small region around the processed boundaries is examined. Also, a joint model incorporating two kinds of classifiers is used to select the most appropriate boundaries for refinements. All of these methods will be examined on a multi-speaker corpus, i.e., TIMIT. The flowchart of this system is shown in Figure 5.1. As a post-processing method, the proposed hybrid scheme can be conveniently applied to phone boundaries generated from an existing segmentation system.

There are two main reasons for using TIMIT in this study. First, TIMIT is originally designed for acoustic modeling and speech recognition/segmentation, unlike BURNC and ARCTIC which are designed for speech synthesis. Therefore, it contains extensive manual labeling of phone boundaries and uses the standard labeling scheme. Second, using TIMIT can facilitate the comparison with other methods as most of the previous studies on phonetic segmentations are based on this corpus.

In many scenarios, it is sometimes necessary to perform phonetic segmentation on a new corpus without labeled data or only with a small set of labeled data. Automatic segmentations may be required for a new English corpus to obtain segmentation information for linguistic analysis or speech recognition purpose. One way to obtain the required phonetic segmentation is first to train acoustic models on a standard corpus and then perform segmentation on the new corpus by using the obtained acoustic models. However, there are some drawbacks of this direct application of pre-trained models. Firstly, the new corpus may have properties different from the corpus used to train the acoustic models. For instance, if the corpus for models training is a small speech recognition corpus and the new corpus is for large vocabulary continuous speech recognition (LVCSR), both the transcriptions and reading styles can be different

Figure 5.1 Proposed Refinement Scheme

in the two corpora and segmentation results will be affected. In addition, the new corpus may contain utterances with some English accents different from the corpus used to train acoustic models. Under such circumstances, the pronunciation variations in the two corpora can lead to extra biases in the segmentation process, generating poor results. Therefore, applying acoustic models from a standard corpus to a new corpus could result in much lower segmentation accuracies compared to the results given by the acoustic models trained on sufficient labeled data from the new corpus. All the recent studies of phonetic segmentations are mainly performed on a single corpus [91, 93, 95, 100-102, 105, 106], and they are not tested under a scenario of cross-corpora segmentation. To address this issue, the refinement process proposed is applied to another corpus with a different English accent to examine the effectiveness of cross-corpora segmentation.

5.2 Baseline Phonetic Segmentation based on Forced Alignment

The baseline phonetic segmentation system is based on forced alignment using a speech recognizer as stated in [90]. Forced alignment is an application of the speech recognition technique. This technique is originally applied in speech recognition to facilitate the embedded training process, but it can also be used for the purpose of phonetic segmentation. Unlike the recognition scenario in which the content of speech signal is unknown, the transcription is already given when performing phonetic segmentation. Therefore, the segmentation process only requires the acoustic models which can differentiate each phone or phone boundary using acoustic features, not relying on the language models which accounts for the grammar and linguistic structures of a specific language.

The term “forced alignment” is used because it “forces” the speech recognizer to go through specific acoustic models given corresponding transcriptions so that the phone duration information can be obtained accordingly. It uses the acoustic models developed in the training phase of the speech recognizer to identify phone boundaries. The forced alignment process is implemented by the

calculation of maximum a posterior (MAP) score of each frame of the speech signal given relevant observations (feature vectors of speech frames):

$$\begin{aligned} \max_{q_0, q_1, \dots, q_T} P(q_0, q_1, \dots, q_T | O, \lambda) \\ = \frac{\max_{q_0, q_1, \dots, q_T} P(q_0, q_1, \dots, q_T, o_1, o_2, \dots, o_T | O, \lambda)}{P(O | \lambda)} \end{aligned} \quad (5.1)$$

where q_t is the state sequence, O is the observations (feature vectors), and λ is the acoustic model corresponding to phone sequences. Therefore, forced alignment continuously selects the state sequences which can maximize the MAP given the observation of speech signals.

As the transcription is already given, the acoustic models of all the phones in the input speech utterance can be concatenated to form the “state transcription” of the sentence. The illustration of this process is given in Figure 5.2, with hollow circles as non-emitting states while filled circles as emitting states. For example, if the speech utterance has only one word “ONE (W-AH-N)” and the acoustic model of each phone has three emitting states (emitting states are those states not at the beginning or end of the HMM), the sentence level “state transcription” will be “W1-W2-W3-AX1-AX2-AX3-N1-N2-N3”, for a total of 9 states of 3 phone models. With the “state transcription”, the posterior score given each individual acoustic model and feature vectors of the input frames can be calculated. The calculated posterior probability score can be used to determine whether to stay in the current state or transit to the next state, according to the left-right structure of the standard HMMs used for speech recognition. After going through the whole utterance with the above procedure, the forced alignment system can provide the information that each phone starts and ends at which frame. As a result, the duration of each phone can be easily calculated from the number of frames involved as well as the stepsize of the frame.

It can be seen from Figure 5.2 that the entry node can go through either “one” or “two” to the exit node, depending on the given transcription. The state sequences of all the individual phones are connected to generate the sentence level “state transcription”, and the number of frames gone through by each state can be determined based on the MAP criteria. Although the utterance given in

Figure 5.2 Illustration of Forced Alignment Process

this example has only one word, it can be easily generalized to a long sentence by simply concatenating all the phones in each individual word and then applying the same scheme to obtain the state durations as well as phone durations.

The overall flowchart of a forced alignment based phonetic segmentation system is provided in Figure 5.3. Speech signals and corresponding manually labeled phone boundaries are used as input to train HMMs based acoustic models. Feature vectors extracted from the speech signals along with phone labels and corresponding time alignment information are used to train HMMs using Baum-Welch algorithm. The individual HMMs can either be context-independent, i.e., each HMM accounts for one phone, or context-dependent, i.e., each HMM accounts for one phone in combination with the previous or following phone. If context-dependent HMM are used, then there will be hundreds of acoustic models to accounts for the combination of different phones. When performing segmentation, the word transcription is first extended into phone transcription using a pronunciation dictionary. Feature extraction is also performed on the speech signal to obtain the feature vector of each frame. The time alignments of each phone can then be obtained by inputting all the feature

Figure 5.3 Training & Testing of the Baseline Automatic Phonetic Segmentation System Using Forced Alignment

vectors into the acoustic models concatenated according to phone level transcriptions as demonstrated in Figure 5.2.

5.3 Refinement Scheme for Phonetic Segmentation

5.3.1 Statistical Correction

The baseline segmentation system based on forced alignment can suffer from systematic biases. The reason is that the forced alignment system seeks to maximize the probability of frame sequences according to the built acoustic models which are optimized to minimize word recognition error rates rather than segmentation errors. It has been shown that “HMM-based models have proved to be quite ‘inventive’ in the solutions they discover” [155].

To demonstrate this phenomenon, a forced alignment based segmentation system is trained using HTK [115] on TIMIT database, which is a multi-speaker database containing a total of 6300 utterances pronounced by 630 speakers.. The general experimental conditions and setups in this paper are listed here for the convenience of discussion. The 61 phone set in TIMIT are merged into the classical 48 phone set as in [101, 156]. Baseline HMMs are trained with 4-state, 5 ms stepsize, and 25 ms frame length. Hamming window is used and the pre-emphasis coefficient is 0.97. Feature vectors consist of 12 MFCCs and normalized log energy plus their delta and acceleration (39-dimensional). HMMs are trained on TIMIT training set (3696 utterances excluding SA utterances). Both context-independent model (CI), i.e., monophone, and context-dependent model (CD), i.e., biphone, are studied in this paper. Biphone rather than triphone is used because biphone can model the transition between two phones, which is more relevant to the detection of phone boundaries between each pair of phones. A decision tree, which asks linguistic related questions (e.g., whether the left phone is a vowel, whether the right phone is a nasal, etc.) as in speech recognition tied-state clustering, is used to classify boundaries into 763 different classes, as demonstrated in Figure 5.4. To generate these classes, the

Figure 5.4 Decision Tree based Clustering of Bi-phone Classes

outlier threshold, i.e., R0 parameter in HTK, is set as 100 and the cluster log-likelihood threshold, i.e., TB parameter in HTK, is set as 300.

The distribution of overall segmentation errors in the TIMT testing set (1344 utterances excluding SA utterances) using HMMs trained by the conditions mentioned above with 5 ms stepsize and the CD scheme is given in the upper

(a) Overall Segmentation Error Distribution for the Whole Testing Set in TIMT

(b) Segmentation Error of Each Individual Phone

Figure 5.5 Segmentation Error Distribution

part of Figure 5.5. It can be seen that the statistical distribution of segmentation errors can be fitted with a Gaussian curve, whose mean is -5.05 ms and standard deviation is 10.81 ms. Because the mean is different from 0, a systematic bias exists for each trained HMM, degrading the segmentation accuracy. From the observations above, one possible way to reduce the segmentation error is to compensate for the systematic bias. A different systematic bias should be used to correct each class of phone boundary, because the acoustic features around different phone boundaries are different from each other. The mean segmentation errors of all the individual phones whose onset is given by the automatic phonetic segmentation systems are plotted in the lower part of Figure 5.5.

Two methods are considered to statistically correct the systematic bias. The first method is an absolute method whose correction term is a weighted sum of the error statistics from the statistical correction training dataset:

$$s(i) = \sum_k p_k \cdot bin[S^{As}(i) - S^M(i)]_k \quad (5.2)$$

where s is the correction term, i is the index of boundary class, p_k is the probability of segmentation errors falling into the bin ranging from k to $k + 1$ ms, according to the error distribution histogram of the boundary class i calculated from the training data. The $bin[S^{As}(i) - S^M(i)]_k$ indicates the differences between automatic and manual segmentations falling in the k -th bin, having a value of k ms. Since more than 99% of segmentation errors are smaller than 35 ms, the bin size is set as 1 ms, from -40 ms to 39 ms, and k ranges from -40 to 39. The refined phonetic segmentation can then be calculated from the automatic segmentation result and the correction term of the corresponding boundary class i :

$$S^{corr} = S^{As}(i) - s(i) \quad (5.3)$$

Although the absolute correction method (i.e., the correction term for each boundary class is a fixed number) can capture the systematic bias of acoustic models, the state-level alignment which is available from the forced alignment is not used. The automatically detected boundary is obtained by the state transition of neighboring HMMs and defined by the onset of the first state of the right phone. Therefore, the statistical correction term can be calculated as a ratio, i.e.,

a relative term, of the state-level segmentations around the automatically detected boundaries. As the manual segmentation may locate at either side of the automatic boundary, two ratios which account for the errors lie on either side of the automatically detected phone boundary should be calculated:

$$Ls(i) = \text{mean}_{i_k} \left\{ \max \left(0, \min \left(1, \frac{SR_1^{AS}(i_k) - S^M(i_k)}{SR_1^{AS}(i_k) - SL_{m-n+1}^{AS}(i_k)} \right) \right) \right\} \quad (5.4)$$

$$Rs(i) = \text{mean}_{i_k} \left\{ \max \left(0, \min \left(1, \frac{S^M(i_k) - SR_1^{AS}(i_k)}{SR_{1+n}^{AS}(i_k) - SR_1^{AS}(i_k)} \right) \right) \right\} \quad (5.5)$$

Where m is the total number of states, n is the search range (the number of considered states around the automatically detected phone boundary), $S^M(i_k)$, $SL_j^{AS}(i_k)$ and $SR_j^{AS}(i_k)$ are the manual segmentation, j -th state-level segmentation of the left phone, and j -th state-level segmentation of the right phone, corresponding to the k -th boundary of class i ; $Ls(i)$ and $Rs(i)$ are the error correction ratios on the left and right side of the boundary class i . The scheme is similar to the one used in [91], but it considers additionally the correction ratio in different search ranges. The illustration of this relative correction method is given in Figure 5.6.

Figure 5.6 Illustration of Relative Correction Method

Figure 5.7 State Selection for Relative Statistical Correction

To enjoy the benefits of different search ranges, a state selection method is used. It calculates the statistics with different search ranges during the training phase and selects the appropriate range for each phone boundary class i by choosing the one minimizing the mean distance between corrected and manual segmentations. Hence, different classes can use different search ranges to refine phone boundaries. This selection process is shown in Figure 5.7.

Once the relative correction ratios for each boundary class are obtained, the phonetic boundaries can be refined according to the class index i of the current phone and the correction terms:

$$\begin{aligned}
 S^{\text{corr}} = & S^{AS}(i) + Rs(i) \times \left(SR_{1+n}^{AS}(i) - SR_1^{AS}(i) \right) \\
 & - Ls(i) \times \left(SR_1^{AS}(i) - SL_{m-n+1}^{AS}(i) \right)
 \end{aligned} \tag{5.6}$$

5.3.2 Multi-resolution Fusion

It has been shown in [101, 157] that using regression to combine output from multiple HMMs with different parameter setups can improve the results. However, combining too many HMMs can be very time consuming. Even in systems using fusion methods like [101, 102], the differences among HMMs are only related to the number of states or mixtures, without considering the stepsize

of HMMs. However, different stepsizes may affect the segmentation results to a certain extent.

To study the effects of stepsize, segmentations are performed on TIMIT using both CI and CD HMMs with stepsize ranging from 5 ms to 12.5 ms. The same parameter settings are applied to HMMs except for their stepsizes. The training and testing set mentioned before are used for experiments. Results are shown in Table 5.1, evaluated as either percentages of the difference between manual and automatic segmentations smaller than a specific number or MAE and RMSE between manual and automatic segmentations, which are commonly used figures of merit to evaluate automatic segmentation results.

Table 5.1 HMMs Performance with Different Stepsizes

	< 5ms	< 10ms	< 20ms	< 30ms	< 40ms	< 50ms	MAE (ms)	RMSE (ms)
CI-5ms	45.10	70.60	88.83	94.18	96.36	97.77	9.57	16.60
CI-7.5ms	41.83	68.28	89.10	94.91	96.83	98.15	9.41	15.56
CI-10ms	36.57	62.23	87.36	94.76	97.05	98.37	10.38	16.09
CI -12.5ms	30.57	51.44	78.80	90.09	94.72	96.98	14.01	22.96
CD-5ms	38.36	64.98	86.23	92.79	95.88	97.52	11.31	19.13
CD-7.5ms	37.91	64.64	87.02	93.75	96.54	97.98	10.61	17.10
CD-10ms	32.33	59.44	84.93	93.72	96.80	98.23	11.50	17.47
CD-12.5ms	28.28	46.52	75.31	88.75	94.59	97.01	14.69	22.07

From Table 5.1, segmentation results given by stepsize of 12.5 ms deteriorate drastically due to the reduced resolution, showing that the stepsize should not be greater than 10 ms. In addition, it can be observed the performances of HMMs with different stepsizes vary under different tolerances. For example, the CI model with 5 ms stepsize achieves highest accuracy for 10 ms tolerance while the one with 10 ms stepsize achieves the highest accuracy for 40 ms and 50 ms tolerance. Stepsizes smaller than 5 ms, e.g., 2.5 ms, are not studied, as overly small stepsize can result in low efficiency in both training and testing phase. The effects of smaller stepsizes are to be compensated by predictive models which only search a small region around the preliminary boundary with a stepsize smaller than 5 ms. This observation shows that different stepsizes can result in different accuracies in terms of various criteria. Specifically, smaller stepsize is more preferable in addressing smaller segmentation errors whereas the reverse is true for bigger stepsize. Thus, it is worthwhile to combine the segmentation results from HMMs with various stepsizes using a regression method so that they can compensate for each other to enhance the accuracy. The regression parameters should be estimated in a manner to minimize the differences between manual and automatic segmentations from training data. In our study, support vector regression (SVR) is used to combine results generated by HMMs with different stepsizes.

5.3.3 Refinement Using Predictive Models

Although statistical correction and fusion methods can reduce the segmentation errors to a certain extent, predictive models are able to utilize flexible acoustic features to improve the segmentation accuracy and are thus worthy of consideration. A joint method using support vector machine (SVM) [125] and linear discriminant analysis (LDA) [158] is considered to further refine the phone boundaries corrected by statistical variances and multi-resolution fusion. Although other methods like artificial neural network (ANN) may also be used, SVM and LDA are more favorable as they require less parameter tuning steps and thus can be readily adaptable to various scenarios.

The first step of predictive model based refinement is feature extraction. As in [93, 95, 105, 106], 13 Mel cepstral frequency coefficients (MFCCs) are

considered. In addition, normalized energy, log pitch value, entropy, bisector frequency, and burst degree as proposed in [159] are also included to form a more informative feature vector. In addition, the dynamic features are also calculated as given in [115]:

$$DF(t) = \frac{\sum_{k=-M}^M F(t+k)k}{\sum_{k=-M}^M k^2} \quad (5.7)$$

where F is the feature vector, t is the time index, and M is the dynamic range which is set as 2 in this case. The obtained dynamic features are then combined with static features to form a 36 dimensional feature vector for each frame.

In this study, frames are labeled using a scheme different from the previous one as in [103, 104], which labels frames close to and far away from the true boundary as two different classes. To maintain the accuracy of the refinement system, only a few samples around the true boundary, e.g., within 5-10 ms, should be labeled as positive instances (frames close to true boundary) while the rest should be labeled as negative instances (frames away from true boundary). Therefore, it can lead to more negative instances than positive instances. Because classifiers are trained to minimize the number of misclassified instances, the resulted classifier will be biased to predict more instances as negative in the testing phase. To overcome this issue, binary labels -1 or 1 are assigned to each frame to represent frames locating at the left or right of the manual boundary, and the change of labels from -1 to 1 is detected as the refined phone boundary. This labeling scheme can make sure that the training instances for both classes are approximately the same, thus resolving the dataset imbalance problem. When multiple changes of labels exist, e.g., “-1 -1 1 -1 -1 -1 1 1 ...”, the change closest to the preliminary boundary is selected as the final boundary. The reason is that after statistical correction and multi-resolution fusion, the phone segmentation error is comparatively small and thus it is unlikely that the refined boundary is far from the real boundary, or at least only in very rare cases.

For each phone boundary class, a binary classifier is trained, with one class corresponding to frames on the left of the manual phone boundary while the other corresponding to frames on the right of the manual phone boundary. A 20

ms frame length and 2.5 ms stepsize are used in this process to compensate for the segmentation errors which may be neglected by the forced alignment process that has a minimum stepsize of 5 ms. The classifiers are trained by speech files and manual phone boundaries from the training set of TIMIT database. With each observation of the manual phone boundary, 20 frames left to the boundary and 20 frames right to the boundary are collected. Then, all the frames corresponding to the same phone boundary class are used to train the classification model. With the trained classification model, the refinement process can be performed by extracting 20 frames around the automatic boundary. To ensure the inclusion of enough training samples, only those boundaries with more than 10 observations in the training corpus are used to train models and refined in the testing phase using the trained models.

The SVM training process can be considered as an optimization problem which is described in section 3.5.1. The SVM used to refine phonetic boundaries is implemented by [134] using RBF kernel. A grid search method as proposed in [160] is used to decide the penalty variable C and the kernel variable g .

The LDA method estimates a projection vector which can map the two classes of data into a space by minimizing the summation of variances while maximizing the difference of means across classes:

$$J(w) = \frac{|\mu'_1 - \mu'_2|^2}{V'_1 + V'_2} \quad (5.8)$$

where w is the feature vector corresponding to two classes, μ'_i and V'_i are the mean value and variance of each class mapped to a new space by the projection vector, and $J(w)$ is the cost function to be maximized. The LDA based subspace projection can be demonstrated by Figure 5.8:

To calculate the projection vector w , the cost function can be transformed to:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (5.9)$$

$$S_W = \sum_{i=1}^2 \sum_{n \in C_i} (x_n - \mu_i)(x_n - \mu_i)^T \quad (5.10)$$

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \quad (5.11)$$

where x_n is the feature vector corresponding to two classes, μ_i is the mean value of each class of data, S_W is the within-class scatter matrix, and S_B is the between-class scatter matrix. Therefore, to maximize the function $J(w)$, we can take the derivative and finally yield:

$$\begin{aligned} \frac{d}{dw} \left[\frac{w^T S_B w}{w^T S_W w} \right] &= 0 \\ \Rightarrow [w^T S_W w] \cdot 2S_B w - [w^T S_B w] \cdot 2S_W w &= 0 \\ \Rightarrow S_W w - \left[\frac{w^T S_B w}{w^T S_W w} \right] S_W w &= 0 \\ \Rightarrow S_B w - J(w) S_W w &= 0 \\ \Rightarrow S_W^{-1} S_B - J(w) w &= 0 \end{aligned} \quad (5.12)$$

This is a generalized eigenvalue problem which can leads to the projection vector w :

Figure 5.8 Separation of Two Groups of Data Using LDA Projection

$$w = S_W^{-1}(\mu_1 - \mu_2) \quad (5.13)$$

5.4 Experimental Results on a Single Corpus

To test the proposed refinement scheme, experiments are conducted on the TIMIT corpus. The same HMMs settings, features and boundary classes as mentioned in Section 3 are used. HMMs with different stepsizes are also trained for fusion purpose which is to be discussed later. A number of experiments are performed to examine the effects of the state-level statistical correction, search range of the statistical correction, multi-resolution fusion, and predictive model. The statistical correction ratio and regression parameters for fusion are calculated from a subset (1000 sentences) of the training set of TIMIT using the difference between manual and automatic segmentations. The same subset is used to train predictive models based on manual segmentations and acoustic features from the wav files.

Firstly, the state-level statistical correction is experimented. As in many of the previous papers [90, 99, 102], we start from HMMs with 5 ms stepsize. The correction statistics are calculated for each boundary class. The results of phonetic segmentation with and without statistical correction are shown in Table 5.2, with the relative method searching in 1 neighboring states ($n=1$). It is found that the scheme based on CI outperforms that based on CD models without statistical corrections. The reasons is that CD models include the transition of two phones and thus discriminate one phone from its context accurately, while CI models trained on individual phones have the ability to discriminate the current phone (which is fixed) from its context (which varies) and generate more accurate results as mentioned in [91]. However, CD models outperform CI models after statistical corrections, which stems from the context-dependent statistics involved in CD models. As compared to the CI models with limited groups of statistics, the CD models provide more detail classifications which generate correction statistics for each phone with various contexts, providing more detailed information to detect the phone boundary.

It is also shown that the relative method outperforms the absolute method, resulting from the variation of phone durations. The variation of phone durations due to different speaking styles or sentence patterns may degrade the performance of the absolute correction, e.g., the fixed correction bias can be too small for a longer observation or too big for a shorter observation of a specific phone. On the other hand, the relative method reflects the correction term as a ratio of the neighboring state durations, making it adaptive to the phone duration. This comparison shows that phone duration information should be taken into account as in the relative method.

Since the relative correction method works better, a study of the search range (or the number of states involved for correction) is performed as shown in Table

Table 5.2 Segmentation Results Using Mono-/Bi- Phone Models with / without Statistical Correction

	< 5ms (%)	< 10ms (%)	<20ms (%)	<30ms (%)	<40ms (%)	<50ms (%)	MAE (ms)	RMSE (ms)
CI- Original	45.10	70.60	88.83	94.18	96.36	97.77	9.57	16.60
CD- Original	38.36	64.98	86.23	92.79	95.88	97.52	11.31	19.13
CI - Relative	50.39	74.29	90.55	94.84	96.75	97.99	8.82	16.72
CD- Relative	51.67	75.86	91.33	95.23	97.50	98.52	8.43	15.95
CI - Absolute	49.03	73.35	88.96	94.28	96.50	97.88	9.23	16.93
CD- Absolute	49.45	73.88	89.37	94.62	96.85	97.91	9.03	16.96

5.3. From the results, a tradeoff can be observed that the smaller search range (1-state) results in a higher accuracy for the case of a smaller tolerance, e.g., < 10 ms, whereas a larger search range (4-state) leads to a higher accuracy for the case of a higher tolerance, e.g., < 50 ms. The reasons are as follows: searching in a smaller range can provide the most accurate local information and thus generates highest accuracy for small segmentation errors, but may fail to compensate for larger errors. In contrast, searching in a larger range will be able to compensate for the large errors such as incorrect phone recognition, but may degrade the ability to correct small errors because states far from the preliminary boundary may not be able to provide the most relevant local information.

Segmentation performance using state selection is also shown in the last row of Table 5.3. It yields both lowest MAE and RMSE. This fact indicates an overall reduction of the segmentation errors by involving the state selection step.

Table 5.3 Relative Correction with Different Search Ranges

	< 5ms (%)	< 10ms (%)	<20ms (%)	<30ms (%)	<40ms (%)	<50ms (%)	MAE (ms)	RMSE (ms)
1-State ($n=1$)	51.67	75.86	91.33	95.23	97.50	98.52	8.43	15.95
2-State ($n=2$)	50.26	74.67	91.46	95.98	97.37	98.55	8.32	14.99
3-State ($n=3$)	50.46	75.18	91.05	96.25	97.70	98.72	8.11	15.03
4-State ($n=4$)	50.35	75.03	91.01	95.34	97.59	98.77	8.24	15.25
State Selection	51.83	76.51	91.72	96.01	97.97	98.85	8.01	14.89

Because state selection can choose the most appropriate search range for different phone boundary classes and generate correction statistics accordingly, the resulted correction process can be more effective.

After statistical corrections, segmentation results from HMMs with different stepsizes are fused using SVR, which is implemented by LIBSVM [134] with RBF kernel. A grid search as proposed in [160] is used to set the SVR parameters. Regression parameters are obtained from the training set by minimizing the differences between the manual boundaries and the fused boundaries given by HMMs with different resolutions. Three groups of HMMs with different resolutions (5 ms, 7.5ms, 10 ms) are used to generate phone boundaries. All the boundaries are statistically corrected with state selection and results are shown in Figure 5.9, where “Fusion (CD)” means fusion results only using CD models and “Fusion (all)” means fusion results incorporating both CD and CI models. The segmentation results given by the fusion method consistently outperform that generated by individual HMMs with various resolutions in terms of accuracy, MAE, and RMSE.

In addition, results generated by CI models are also included in the fusion process. It can be observed that the fusion of HMMs with different resolutions have the main contributions and the inclusion of CI models just achieves some incremental improvements on the results. Therefore, the fusion process enables HMMs with different resolutions to compensate each other and enhance the segmentation results.

(a) Improvements of Accuracies across Different Tolerances

(b) MAE and RMSE Given by Single Resolution Model and Fusion

Figure 5.9 Segmentation Results of HMMs with Different Resolutions and Multi-resolution Fusion

After the fusion process, predictive models are applied as the final refinement step. It attempts to provide further refinements with even smaller stepsize, i.e., 2.5ms, for resolving small segmentation errors. One classifier is trained for each phone boundary class. The same 763 phone boundary classes as used for CD models are used to train the classification models. SVM and LDA are

Table 5.4 Refinements Using Predictive Models

	< 5ms (%)	< 10ms (%)	< 20ms (%)	< 30ms (%)	< 40ms (%)	< 50ms (%)	MAE (ms)	RMSE (ms)
Original	38.36	62.98	86.23	92.79	95.88	97.52	11.31	19.13
Stats + Fusion	53.61	78.57	93.25	97.12	98.45	99.16	7.29	12.91
With LDA	55.52	79.91	93.51	97.25	98.57	99.27	7.03	12.51
With SVM	58.02	81.31	94.19	97.56	98.70	99.36	6.54	11.92

implemented by LIBSVM and Matlab, respectively. The results are shown in Table 5.4.

It can be observed that by using either LDA or SVM models, some improvements on the segmentation results can be achieved, with SVM method outperforms the LDA method. Compared to LDA, SVM is a more discriminative classifier which can maximize the margin between two classes and thus maintain the generality of the developed classification model. The RBF kernel used by SVM also maps acoustic features to a high-dimensional space which can address non-linear data more effectively, especially for the acoustic features which are not linearly separable. As compared to statistical correction which mainly addresses the systematic biases using global statistics, predictive models reduce segmentation errors based on local acoustic features around the preliminary boundary. It can be observed that segmentation accuracy with 5 ms tolerance is improved to the most extent by this step, which may result from the searching scheme of predictive model, e.g., small stepsize and preference for small shift from the preliminary boundaries. The obtained accuracy for 5 ms tolerance is better than that reported recently in [99, 101].

The error distribution after the whole refinement process as well as reduced MAE for each individual phone in TIMIT is given in Figure 5.10. The reduced MAEs are plotted for each phone whose onset is defined by the detected

boundary. It can also be noted from Figure 5.10 (a) that the error distribution has a mean value much closer to 0 and is more similar to a standard Gaussian distribution. In addition, a positive reduction in the mean segmentation errors for each individual phone can be observed in Figure 5.10 (b).

In all the experiments above, T-tests show significant differences ($t < 0.001$) of MAE, i.e., differences between manual and automatic segmentations, for

(a) Overall Error Distribution before and after Refinements

(b) Reduced MAE of Individual Phones

Figure 5.10 Improvements of the Whole Test Set and Individual Phones on TIMIT

Figure 5.11 Overall Improvements of Segmentation Results

different segmentation setups. Overall improvements by including all the proposed methods are demonstrated in Figure 5.11. The left and right vertical axis indicates the MAE & RMSE in terms of ms and the accuracies in terms of percentage, respectively. Statistical correction, state selection and multi-resolution fusion all contribute to the improvements of segmentation results in terms of accuracy, MAE, and RMSE.

The achieved segmentation results are better than those in [101], which reports MAE of 10.01 and RMSE of 17.15 based on a post-processing fusion method involving over 100 segmentation machines with the same phone sets and experimental conditions on the TIMIT corpus. Compared to results in [99, 101], the reported results are more accurate across different tolerances. The obtained results also outperform those in [98] for 10 ms to 30 ms tolerances, although there are some differences between experiments in this paper and those in [98] in terms of phone sets. Similarly, the reported segmentation accuracies in Table 5.4 for 20 ms, 30 ms, 40 ms, and 50 ms tolerances are also higher than the inter-human agreements on TIMIT which are 93.49%, 96.91%, 98.51%, and 99.06% as given in [99]. In contrast, the inter-human correlation on TIMIT for 5 ms and 10 ms tolerance are over 60 % and close to 82% respectively, desiring further improvements. Therefore, studies on phonetic segmentation systems should focus on small segmentation errors which exhibit a gap between human and

machine performances. This hybrid scheme particularly addresses this issue and leads to more improvements for small segmentation errors, i.e., those within 5 ms and 10 ms.

It should be noted that this study focuses on a post-processing scheme to refine phone boundaries given by a baseline segmentation system in comparison to the methods presented in [97-99] which change the acoustic modeling scheme in various ways. Although these methods can contribute to the segmentation accuracy significantly, the proposed scheme provides a concatenation of several different refinement methods to improve the segmentation results. The inclusion of each step, i.e., statistical correction, fusion and predictive model, can be determined according to the availability of training data and the implementation time, allowing for increased flexibility and a tradeoff between efficiency and accuracy.

5.5 Cross-corpora Segmentation

A cross-corpora segmentation study is conducted in this section. The other corpus used in this section is Cambridge Wall Street Journal (WSJCAM0) [161], which is a British version of Wall Street Journal (WSJ) corpus recorded in Cambridge University, incorporating utterances produced by 140 speakers. Different from TIMIT which consists of regional dialects in American English, WSJCAM0 contains utterances with British English which possesses characteristics different from those of American English. TIMIT and WSJCAM0 not only have different accents, but are also designed for different purposes: TIMIT is designed for small set acoustic modeling or phone recognition, while WSJCAM0 is designed for LVCSR which can be used for both acoustic and language modeling. Considering these properties, it can be expected that the segmentation results on TIMIT given by acoustic models obtained from WSJCAM0 are likely to be worse as compared to that given by acoustic models trained on TIMIT.

To study cross-corpora segmentation, the acoustic models are trained by WSJCAM0 corpus and tested on TIMIT database with/without the proposed refinement scheme, because TIMIT is a standard corpus widely used for

segmentation studies and provides manual labels which can facilitate the evaluation of segmentation results. As the WSJCAM0 corpus has much more speech data than TIMIT, only a subset of WSJCAM0 data is used for a fair comparison. All the 5013 utterances in CD1 are used as the training set to train British acoustic model which are used to obtain phone boundaries of the testing set of TIMIT database. It should also be noted that the number of phones in TIMIT is bigger than that in WSJCAM0. Closures like “cl”, “vcl” and “epi” are not labeled in WSJCAM0 and there seems to be no similar patterns in WSJCAM0 to replace those labels. As the focus of this paper is to resolve the effects of different accents or speaking styles on cross-corpora segmentations rather than dealing with transcription discrepancies, we therefore eliminate these phones in cross-corpora experiments and evaluate the results based on the rest of phones.

Under the cross-corpora segmentation task, the British acoustic models (BAM) with 5 ms stepsize and 25 ms frame length obtained from the British corpus (WSJCAM0) are used to generate segmentation results for the TIMIT testing set. The other parameter setups are the same as described in section 3. The results are then refined by 200 utterances in the training set of TIMIT. Again, TIMIT is considered as the new corpus here and only the labeling of the 200 utterances are considered available in this cross-corpora task. The refinement process including statistical correction and multi-resolution is performed with 100 utterances for statistical correction training and 100 utterances for regression training. The same 200 utterances are used to train the predictive models for refinements. This is applicable because the training process only requires the reference boundaries and speech files from the corpus, not relying on the automatic boundaries. All the refinement parameters are obtained from the training set of TIMIT while the refinement process is performed on the testing set of TIMIT. Segmentation results are compared with that generated by American acoustic models (AAMs) which are obtained from the whole TIMIT training set and also have 5 ms stepsize and 25 ms frame length. Similar as before, SA sentences in both training and testing sets are excluded.

Table 5.5 Cross-corpora Segmentation on WSJCAM0 Using AMA

	< 10ms (%)	< 20ms (%)	< 30ms (%)	< 40ms (%)	< 50ms (%)	MAE (ms)	RMSE (ms)
Baseline							
BAM	57.67	79.60	88.11	93.06	95.03	14.33	24.23
BAM + Statistical	66.84	85.74	91.91	95.26	96.52	11.50	19.51
BAM + Statistical + Fusion	68.15	87.28	93.31	95.96	97.38	10.29	17.94
BAM + Statistical + Fusion + SVM	70.04	88.36	93.82	96.35	97.56	9.82	17.17
AAM	69.36	87.80	93.59	96.27	97.64	9.97	17.36

Although the CD approach slightly outperforms the CI approach with the hybrid refinement scheme, the CD structures of the two corpora are quite distinct due to different training transcriptions and accents. In addition, the limited training corpus for statistical correction (only 200 utterances) makes the training of large groups of correction terms unreliable. Therefore, CI models are used in this cross-corpora segmentation task and the correction statistics are trained for individual phones rather than CD to reduce the required training data. The obtained results are shown in Table 5.5.

Significant improvements of segmentation accuracies can be observed if the proposed statistical correction is applied using 100 labeled utterances from the TIMIT training set. With the fusion methods applied using additional 100 labeled utterances to combine results given by HMMs with stepsizes from 5 ms to 10 ms and refinements done by predictive models, the results can be further improved. The final refined results slightly outperform that obtained by AAM in of the tolerances and MAE & RMSE. Particularly, it can be observed that the

accuracies for 10 ms and 20 ms tolerances are improved significantly. The results demonstrate that the proposed refinement process trained on 200 labeled utterances from the new corpus can result in segmentation accuracy at least comparable to that achieved by well-trained acoustic models obtained from the new corpus, i.e., TIMIT, using the whole labeled training set (3696 utterances).

The improvements in terms of MAE for the whole database and individual phones in TIMIT under this cross-corpora scenario are also plotted in Figure 5.12. The upper part of Figure 5.12 shows the error distribution before and after the proposed refinement process. Compared to the distribution generated by the baseline system, the kurtosis is corrected with the error distributions more concentrated around 0. In addition, the improvements for each individual phone whose onset is given by the automatically detected phone boundaries are plotted in the lower part. It can be observed that the MAE for most of phones is reduced significantly. In addition, vowels like “AE”, “AH”, “AY”, “EH”, and “EY” receive high reduction in error, corresponding to their distinctions between British and American as studied in [79, 162]. Such kind of distinctions can result from both pronunciation differences and context variations around these vowels across two accents. These observations show that systematic biases under this cross-corpora segmentation scenario, especially those related to the differences between the two English accents in the training and testing sets, are reduced effectively.

Based on the experiments discussed above, the proposed refinement method provides a convenient and effective way to achieve segmentations for a new corpus by using a small group of labeled data, with accuracy comparable to that generated by a forced alignment system as if it were trained by sufficient labeled data from the new corpus.

(a) Overall Error Distribution before and after Refinements – Cross-Corpora Segmentation

(b) Reduced MAE of Individual Phones – Cross-Corpora Segmentation

Figure 5.12 Cross-corpora Improvements Using the Hybrid Refinement Scheme

5.6 Summary

This section proposes a refinement process for HMMs based automatic phonetic segmentation. The preliminary phone boundaries generated by forced alignment using HMMs are refined by various refinement methods to obtain the final phone boundaries.

Statistical corrections based on both absolute and relative statistics are compared using CI and CD models. Comparisons between the two methods show that the relative method with CD models can generate the highest accuracy. The reason is that relative statistical correction method takes into account phone duration information, generating correction term variable according to the duration of each phone. A state selection method, which is not considered in previous research using statistical correction [91], is used to refine the relative statistical correction. Experimental results support that different search ranges can affect the segmentation accuracy and the proposed state selection scheme can improve the refinement process by selecting the most appropriate search range for each phone boundary class.

Studies also show that HMMs with different resolutions are able to contribute to the segmentation process differently. The proposed multi-resolution fusion step embraces the benefits of HMMs with different stepsizes, which are not studied in other fusion methods [100-102]. Experimental results demonstrate significant improvements over all criteria by incorporating the multi-resolution fusion step. As only a small number of HMMs are involved, the implementation efficiency is higher than traditional fusion methods.

In addition, refinements based on predictive models are involved to deal with the nonlinear distortions inside the segmentation results. A different predictive model is trained for each phone boundary class using the acoustic features around real boundaries. Frames around the intermediary phone boundaries are extracted and correlated feature vectors are input to the predictive model, which is a classifier, corresponding to the specific phone boundary. The binary label of each frame is then given according to the feature vector and the predictive model.

The change of output label is determined as the change from left phone to the right phone, which accounts for the most probable phone boundary.

In addition, this chapter proposes cross-corpora segmentation which has not been studied in the past. This scheme performs segmentation on a new corpus with a different accent and purpose based on the segmentation system trained on a standard corpus. For experiment purpose, TIMIT, the American corpus for acoustic modeling, and WSJCAM0, the British corpus for LVCSR, are considered as the new corpus and the standard corpus, respectively. Experiments demonstrate that the segmentation results on TIMIT given by models trained on TIMIT significantly outperform those from the models trained on WSJCAM0, showing that the differences of the characteristics and accents of corpora can affect the segmentation accuracy. Also, the proposed refinement scheme can be applied, by using a small set of labeled data from the new corpus, to improve the accuracy of cross-corpora segmentation so that the results are comparable to that as if they were generated from models trained on the new corpus with large volume of labeled data. The refinement scheme under this cross-corpora segmentation scenario can facilitate the analysis and application of a new corpus with only a small set of manual segmentations by providing a reliable starting point.

Chapter 6 Conclusion & Recommendations for Future Research

This thesis addresses several aspects of a computer-aided language learning system. Three topics including prosody evaluation, feedback generation, and phonetic segmentation are studied.

6.1 Prosody Evaluation Based on Quasi-foot Segmentation

The proposed method segments speech utterances using a prosodic unit called foot to take into account the prosodic structure. Unlike word which is a lexical unit, foot works in the prosodic domain to segment speech utterances according to relevant rhythm and stresses, making it a more appropriate unit for the purpose of prosody evaluation. Since human labeling is time-consuming and infeasible for large databases, an automatic version of foot segmentation called quasi-foot segmentation is developed. By considering the definition of foot segmentation from the viewpoint of linguistics and the correlation between foot boundary and stresses, pitch accent detection is combined with forced alignment to segment the sentence automatically. As a simple but effective method, logistic regression is applied to detect the pitch accent for automatic foot-level segmentation. Experimental results show that this segmentation scheme can achieve reasonable accuracy.

To test the effectiveness of the proposed segmentation method, experiments are performed on a reference-dependent evaluation system to assess the prosody of English learners via a comparison with reference utterances produced by the teacher. The evaluation score is calculated by measuring the distance between the learner's and the teacher's prosody. Pitch and energy related features are included and DTW is used to align the teacher's and the learner's speech utterances. Experiments are performed on the traditional word segmentation, the

manually generated foot segmentation, and the automatically generated word and quasi-foot segmentation. Results show that the automatic foot segmentation can work well in generating evaluation accuracy similar to that given by manual segmentations.

In addition to reference-dependent evaluation, a SVM based reference-independent prosody evaluation is also experimented to resolve the constraint of transcriptions used for practice. A 60-dimensional feature vector which consists of representative features of pitch and energy is extracted for each sentence. Experiments using SVM for reference-independent prosody evaluation are conducted and the corresponding results show the outperformance of foot level segmentation compared with the other two segmentation methods. Feature selection and accented word ratio comparison are also performed to verify the superiority of the proposed scheme. It is shown that the quasi-foot segmentation is more robust to the change of features and can consistently outperform other methods. In addition, the rhythm information can be modeled appropriately, leading to improved evaluation accuracy.

6.2 Generation of Feedback Utterances

To overcome those limitations of traditional feedback utterances which are recorded from the native speaker directly, a feedback scheme which combines the learner's vocal features and the teacher's linguistic gestures can be a reasonable solution. By combining the learner's vocal features and the teacher's prosody & pronunciation, the learner can listen to native-like utterances in his own voice. Therefore, the learner can just focus on prosody or pronunciation issues which discriminate his speech utterances from native ones, improving speaking skills more efficiently. Further, the proposed feedback utterances can also work as an intermediary between the learner's non-native speech and the teacher's native speech, providing the learner with more choices of different stimuli in the learning process.

A multi-corpora experiment is designed to study several issues unsolved in the literature, including the effects of corpus and nationality of learners on the resulted feedback utterances. A number of steps including alignment

interpolation, spectral interpolation, and vocal tract length normalization are added to improve the quality of accent reduced utterances. Two different corpora, the BURNC based corpus and ARCTIC corpus are used for experiments. Three different modifications, namely prosodic modification, segmental modification, and combined modification, are performed on the two corpora. Results demonstrate that prosodic modification is dominant for prosody-abundant corpus like BURNC, while segmental modification is more important for prosody-flat corpus like ARCTIC. Also, the nationality of non-native speakers can affect their choices of the desirable modification methods. Besides, acoustic quality, based on MOS, shows that the prosodic modification generates feedback utterances with the highest quality while combined modification yields the lowest quality. It means that there is a tradeoff between prosodic modification and combined modification.

To seek more possibilities of feedback generation, three different synthesis methods, namely PSOLA, HSM and STRAIGHT, are applied for the purpose of accent reduction and compared according to the produced stimuli. As HSM and STRAIGHT do not separate speech components which mainly represent speaker identity from those mainly related to linguistic contents, only prosodic modifications are performed in this study. Voice conversion method is also performed to generate feedback utterances. With this method, the teacher's utterances are converted into the learner's voice while preserving the teacher's linguistic gestures. The generated utterances can thus possess the learner's speaker identity and desired prosody and pronunciations. Experiments based on accentedness and acoustic quality are performed to assess the feedback utterances obtained by different methods. It is shown that STRAIGHT method is the most appropriate for prosody based accent reduction due to its higher acoustic quality as well as nativeness compared to the other two synthesis methods. Voice conversion method can generate the highest nativeness among all the feedback generation methods, while yielding the lowest acoustic quality due to the introduced smoothness and distortions on spectral information. Considering these pros and cons of accent reduction and voice conversion methods, a multi-stage learning scheme, which presents the learner with accent

reduced utterances at the first stage and generates feedback utterances using voice conversion method at the second stage, is proposed. This scheme enables feedback utterances adaptive to the learner's performance, offering flexible feedback stimuli in the CALL system.

6.3 Refinement of Phonetic Segmentations

A systematic bias is observed for each phone boundary class by studying the segmentation error distribution generated by forced alignment. Two different statistical correction methods, i.e., absolute and relative corrections, are proposed to compensate for this kind of bias. The first method uses an absolute or fixed value, i.e., the weighted summation of errors, as the correction term by studying the error distribution of each phone boundary class. The second method takes into account the state-level alignments and calculates two relative ratios for each phone boundary. The correction term can then be obtained by multiplying the relative ratio with the state-level duration around the preliminary boundary. For the relative method, a state selection scheme is also used to find the most appropriate search range for each phone boundary rather than applying uniform search range for all phone boundaries. Experiments performed on the TIMIT corpus show that the statistical correction can significantly improve the accuracy of segmentation results. Specifically, the relative correction outperforms the absolute correction in terms of all the measurements due to the duration information involved. A comparison across different search ranges of the relative correction method demonstrate that the proposed state selection method can achieve the highest accuracy and is thus necessary to be performed during the training of relative correction terms.

In addition to statistical correction, the resolution of HMMs is also studied. Different resolutions of HMMs can contribute differently to segmentation accuracy under different measures. Experimental results show that a smaller resolution can produce higher accuracy for smaller tolerances, while the reverse situation applies for larger resolutions. Hence, it is beneficial to use a fusion scheme to combine segmentation results generated by HMMs with different resolutions. Experimental results show a further improvement on segmentation

results after combining the corrected boundaries from HMMs with different resolutions.

Finally, a post-processing step based on predictive models is applied. For each phone boundary class, one binary classifier is trained using those speech frames around the manually labeled phone boundaries from this class. The change of labels is detected as the shift from the previous phone to the current phone, denoting a phone boundary. A further improvement is shown after involving such a predictive model based refinement scheme. In summary, the combination of statistical correction, multi-resolution fusion and predictive model based refinement can yield the most significant improvements of segmentation accuracy.

The cross-corpora segmentation, which is not explored before, is also investigated to obtain the phonetic segmentations for a new corpus with limited labeled data by using the proposed hybrid refinement scheme. Two corpora, i.e., TIMIT and WSJCAM0, which are not only in different English dialects but also designed for different purposes, are used. Experiments are performed by using WSJCAM0 to train baseline HMMs and taking TIMIT as the new corpus with limited alignment information. Segmentation results show that the proposed refinement scheme can achieve reasonable segmentation results on the new corpus as if the acoustic models were trained on the new corpus with sufficient alignment labels. Therefore, the proposed refinement scheme can facilitate the analysis as well as applications of a new English corpus in different dialects or with different purposes.

6.4 Recommendations for Future Research

Future work on prosody evaluation should focus on the improvement of automatic segmentation of the proposed prosodic unit. It is possible to merge several different detection schemes like logistic regression, HMM, and neural network for achieving improved detection accuracy. To improve the evaluation accuracy, subjective studies about prosodic features can be performed by questionnaires or linguistic analysis to examine all the relevant information such

as semantic or syntactic structures. Such kind of studies can exploit features for the identification of prosodic units as well as the evaluation of prosody.

The generation of feedback utterances in Chapter 4 may be studied in a more intensive way in cooperation with linguists or educators to consider effects of other factors, such as age and gender of language learners, on the acquisition of L2 language as well as learners' preferences of different feedback utterances. It is beneficial to develop a detail classification of L2 language learners based on their preference on feedback utterances to provide the most appropriate stimuli for facilitating the language acquisition process. A more comprehensive database containing sufficient non-native English learners with various backgrounds, nationalities, and personalities may be collected to facilitate this process. Pedagogic studies can be involved in this process by combining prosody evaluation and feedback utterance generation systems so as to evaluate the learning experience of non-native learners. The overall performance of English learners using the CALL system after a period of training can then be compared with a control group who undergoes the traditional class room training. Pedagogic studies may involve a number of issues like the selection of learners, the learning materials, the learning period, and the assessment of English proficiency before/after the training process, thus requiring cooperation across engineering, linguistic, and education researchers. Although such kind of studies is time consuming and expensive, it may provide valuable cues about the effectiveness of a CALL system.

The current cross-corpora segmentation in Chapter 5 is based on two standard corpora with different dialect accents. However, both the corpora are considered as native English corpus. Therefore, it is interesting to collect a non-native English corpus with regional dialect(s) to further test the proposed scheme of cross-corpora segmentation. To assess the results appropriately, manual labels should be obtained for the corpus with regional dialect(s). In addition, it may be possible to extend this cross-corpora segmentation scheme to a cross-language scheme. For example, acoustic models trained by English corpora may be applied to segment a French or Japanese corpus with a small set of labeled data. Although this kind of scheme may only be applied to languages with similar

overlapping phone sets, it can potentially facilitate the analysis of corpora in rare foreign languages which are difficult to find appropriate human labelers. To perform the cross-language segmentation, however, the phone differences across languages must be studied carefully so that a mapping between different labeling systems can be constructed.

List of Publications

S. Zhao, I. Y. Soon, S. N. Koh, K.K. Luke, “A Hybrid Refinement Scheme for Intra- and Cross- Corpora Phonetic Segmentation,” *Computer Speech & Language (Elsevier)*, Accepted

S. Zhao, S. N. Koh, and K. K. Luke, “Reference-independent Prosody Evaluation Based on Prosodic Unit Segmentation,” *IEICE Transactions on Information and Systems*, vol. E96-D, no. 9, 2013.

S. Zhao, I. Y. Soon, S. N. Koh, K.K. Luke, “Phonetic Segmentation Using Statistical Correction and Multi-resolution Fusion,” in *Proc. 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013 (Oral Presentation), pp. 6694-6698.

S. Zhao, S. N. Koh, I. Y. Soon, K.K. Luke., “Feedback Utterances for Computer-aided Language Learning Using Accent Reduction and Voice Conversion Method,” in *Proc. 38th International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013 (Oral Presentation), pp. 8208-8212.

S. Zhao, I. Y. Soon, S. N. Koh, K.K. Luke, “A Study on Feedback Stimuli for Non-native English Learners in Southeast Asia,” *ACM Transactions on Asian Language Information Processing*, Under Second Review

List of Publications

S. Zhao, S. N. Koh, K.K. Luke, “Accent Reduction for Computer-aided Language Learning,” in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 335-339.

S. Zhao, S.N. Koh, K.K. Luke and Y. Zhang, Computer Aided Evaluation of Intonation for Language Learning Based on Prosodic Unit Segmentation, in *Proc. Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, Singapore, 2010, pp. 788-793.

References

- [1] M. Warschauer, and D. Healey, "Computers and language learning: An overview," *Language teaching*, vol. 31, no. 2, pp. 57-71, 1998.
- [2] M. Levy, *Computer-assisted language learning: Context and conceptualization*, 1997.
- [3] F. Ehsani, and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm," *Language Learning and Technology*, vol. 2, no. 1, pp. 45-60, 1998.
- [4] D. S. Bourgerie, "Computer aided language learning for Chinese: A survey and annotated bibliography," *Journal of the Chinese Language Teachers Association*, vol. 38, no. 2, pp. 17-47, 2003.
- [5] M. Liu, Z. Moore, L. Graham *et al.*, "A Look at the Research on Computer-Based Technology Use in Second Language Learning: A Review of the Literature from 1990-2000," *Journal of Research on Technology in Education*, vol. 34, no. 3, pp. 250-274, 2002.
- [6] T. Piske, I. R. A. MacKay, and J. E. Flege, "Factors affecting degree of foreign accent in an L2: a review," *Journal of Phonetics*, vol. 29, no. 2, pp. 191-215, 2001.
- [7] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, no. 1-3, pp. 170-183, 1994.
- [8] M. J. Munro, and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73-97, 1995.
- [9] E. O. B. Wilson, and T. J. Spaulding, "Effects of Noise and Speech Intelligibility on Listener Comprehension and Processing Time of Korean-Accented English," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 6, pp. 1543, 2010.
- [10] H. Lane, "Foreign accent and speech distortion," *J. Acoust. Soc.*, vol. 35, no. 4, pp. 451-453, 1963.
- [11] J. Anderson Hsieh, and K. Koehler, "The Effect of Foreign Accent and Speaking Rate on Native Speaker Comprehension," *Language learning*, vol. 38, no. 4, pp. 561-613, 1988.
- [12] A. D. Lawson, D. M. Harris, and J. J. Grieco, "Effect of foreign accent on speech recognition in the NATO N-4 corpus," *EUROSPEECH*, 2003.
- [13] L. M. Arslan, and J. H. L. Hansen, "Language accent classification in American English," *Speech Communication*, vol. 18, no. 4, pp. 353-367, 1996.
- [14] M. J. Munro, and T. M. Derwing, "Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech," *Language and speech*, vol. 38, no. 3, pp. 289, 1995.
- [15] E. B. Ryan, and M. A. Carranza, "Evaluative reactions of adolescents toward speakers of standard English and Mexican American accented English," *Journal of Personality and Social Psychology*, vol. 31, no. 5, pp. 855, 1975.
- [16] R. Lippi-Green, *English with an accent: Language, ideology, and discrimination in the United States*: Psychology Press, 1997.
- [17] R. Kalin, and D. S. Rayko, "Discrimination in evaluative judgments against foreign-accented job candidates," *Psychological Reports*, vol. 43, no. 3, pp. 1203-1209, 1978.
- [18] A. Gluszek, and J. F. Dovidio, "Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States," *Journal of Language and Social Psychology*, vol. 29, no. 2, pp. 224, 2010.

References

- [19] J. E. Flege, and K. L. Fletcher, "Talker and listener effects on degree of perceived foreign accent," *Journal of the Acoustical Society of America*, vol. 91, no. 1, pp. 370-389, 1992.
- [20] R. McAllister, "Second language perception and the concept of foreign accent," in STiLL, Marholmen, Sweden, 1998.
- [21] J. E. Flege, M. J. Munro, and I. R. A. MacKay, "Factors affecting strength of perceived foreign accent in a second language," *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3125-3134, 1995.
- [22] S. Tahta, M. Wood, and K. Loewenthal, "Foreign accents: Factors relating to transfer of accent from the first language to a second language," *Language and speech*, vol. 24, no. 3, pp. 265, 1981.
- [23] T. M. Walsh, and K. C. Diller, "Neurolinguistic considerations on the optimum age for second language learning," in Proc. of the Fifth Annual Meeting of the Berkeley Linguistics Society, 1979, pp. 510-524.
- [24] T. Bongaerts, "Ultimate attainment in L2 pronunciation: the case of very advanced late L2 learners," in Birdsong, D. (Ed.), *Second Language Acquisition and the Critical Period Hypothesis*, Lawrence, Erlbaum, Mahway, NJ, 1999, pp. 133-159.
- [25] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, vol. 52, no. 3, pp. 254-267, 2010.
- [26] D. M. Hardison, "Generalization of computer-assisted prosody training: Quantitative and qualitative findings," *Language Learning & Technology*, vol. 8, no. 1, pp. 34-52, 2004.
- [27] M. NESPOR, "Prosody: an interview with Marina Nespor," *ReVEL*, vol. 8, no. 15, 2010.
- [28] H. Chao, Z. Feng, and F. K. Soong, "Improving Automatic Evaluation of Mandarin Pronunciation with Speaker Adaptive Training (SAT) and MLLR Speaker Adaption," in 6th International Symposium on Chinese Spoken Language Processing, 2008, pp. 1-4.
- [29] R. Hincks, "Speech technologies for pronunciation feedback and evaluation," *ReCALL*, vol. 15, no. 1, pp. 3-20, 2003.
- [30] L. Neumeyer, H. Franco, V. Digalakis *et al.*, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, no. 2-3, pp. 87-93, 2000.
- [31] A. Ito, T. Konno, M. Ito *et al.*, "Evaluation of English Intonation based on Combination of Multiple Evaluation Scores," in Interspeech 2009, 2009, pp. 596-599.
- [32] C. Kim, and W. Sung, "Implementation of an intonational quality assessment system," in Seventh International Conference on Spoken Language Processing, 2002.
- [33] G. Szaszak, D. Sztaho, and K. Vicsi, "Automatic Intonation Classification for Speech Training Systems," in Interspeech, 2009, pp. 1875-1878.
- [34] M. Suzuki, T. Konno, A. Ito *et al.*, "Automatic evaluation system of English prosody for Japanese learner's speech," in International Multi-Conference on Society, Cybernetics and Informatics, Orlando, 2007.
- [35] H. Meng, W. K. Lo, A. M. Harrison *et al.*, "Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience," in APSIPA 2010, Singapore, pp. 811-820.
- [36] I. ECL, "Pronunciation Power 1 (Beginner/Intermediate)," *Teaching English as a Second or Foreign Language*, vol. 5, no. 4, 2002.
- [37] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors-In search of the golden speaker," *Speech Communication*, vol. 37, no. 3-4, pp. 161-173, 2002.
- [38] A. Sundström, "Automatic prosody modification as a means for foreign language pronunciation training," in Proc. ISCA Workshop on Speech Technology in Language Learning (STILL 98), Marholmen, Sweden, 1998, pp. 49-52.

References

- [39] M. Jilka, and G. Möhler, "Intonational foreign accent: speech technology and foreign language teaching," in Proc. ESCA Workshop on Speech Technology in Language Learning, 1998, pp. 115-118.
- [40] M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann, "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis," in Proc. of the 11th Australian International Conference on Speech Science & Technology, University of Auckland, New Zealand, 2006.
- [41] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51(10), no. 10, pp. 920-932, 2009.
- [42] A. Kain, and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1998, pp. 285-288.
- [43] D. Erro, "Intra-lingual and Cross-lingual Voice Conversion Using Harmonic Plus Stochastic Models," *PhD Thesis*, 2008.
- [44] N. Campbell, and A. Black, "Prosody and the selection of source units for concatenative synthesis," *Progress in speech synthesis*, pp. 279-292, 1997.
- [45] S. Narayanan, and A. Alwan, *Text to speech synthesis: new paradigms and advances*: Prentice Hall Professional Technical Reference, 2005.
- [46] A. J. Hunt, and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996, pp. 373-376.
- [47] D. T. Chappell, and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, vol. 36(3), no. 3, pp. 343-373, 2002.
- [48] C. W. Wightman, and D. T. Talkin, "The aligner: Text-to-speech alignment using Markov models," *Progress in speech synthesis*, pp. 313-323, 1997.
- [49] C. D. Mitchell, M. P. Harper, and L. H. Jamieson, "Using explicit segmentation to improve HMM phone recognition," in Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Detroit, Michigan, USA, 1995, pp. 229-232.
- [50] J. P. van Hemert, "Automatic segmentation of speech," *Signal Processing, IEEE Transactions on*, vol. 39(4), no. 4, pp. 1008-1012, 1991.
- [51] B. L. Pellom, and J. H. Hansen, "Automatic segmentation of speech recorded in unknown noisy channel characteristics," *Speech Communication*, vol. 25(1), no. 1, pp. 97-116, 1998.
- [52] A. Ljolje, J. Hirschberg, and J. P. H. van Santen, "Automatic speech segmentation for concatenative inventory selection," *Progress in Speech Synthesis, Springer*, pp. 305-311, 1996.
- [53] P. Cosi, D. Falavigna, and M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," in Proc. of Eurospeech, Genova, Italy, 1991.
- [54] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1992, pp. 517-520.
- [55] H. Leung, and V. Zue, "A procedure for automatic alignment of phonetic transcriptions with continuous speech," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1984, pp. 73-76.
- [56] J. P. H. van Santen, E. T. Prud'hommeaux, and L. M. Black, "Automated assessment of prosody production," *Speech Communication*, vol. 51, no. 11, pp. 1082-1097, 2009.
- [57] M. Brown, and L. Rabiner, "An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 30, no. 4, pp. 535-544, 1982.

References

- [58] M. Suzuki, T. Konno, A. Ito *et al.*, “Automatic evaluation system of English prosody based on word importance factor,” *Journal of Systemics, Cybernetics and Informatics*, vol. 6, no. 4, pp. 83-90, 2008.
- [59] A. Ito, T. Konno, M. Ito *et al.*, “Intonation Evaluation of English Utterances Using Synthesized Speech for Computer-assisted Language learning,” *International Journal of Innovative Computing Information and Control*, vol. 6, no. 3(B), pp. 1501-1514, Mar, 2010.
- [60] K. Silverman, M. Beckman, J. Pitrelli *et al.*, “ToBI: A standard for labeling English prosody,” in Second International Conference on Spoken Language Processing, 1992, pp. 867-870.
- [61] C. Teixeira, H. Franco, E. Shriberg *et al.*, “Prosodic features for automatic text-independent evaluation of nativeness for language learners,” in Proc. of ICSLP, 2000.
- [62] C. Teixeira, H. Franco, E. Shriberg *et al.*, “Evaluation of speaker’s degree of nativeness using text-independent prosodic features,” in Proc. Of the Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark, 2001.
- [63] J. Tepperman, A. Kazemzadeh, and S. Narayanan, “A Text-free Approach to Assessing Nonnative Intonation,” in Interspeech 2007: 8th Annual Conference of the International Speech Communication Association, Vols 1-4, 2007, pp. 897-900.
- [64] J. Tepperman, and S. Narayanan, “Better Nonnative Intonation Scores through Prosodic Theory,” in Interspeech 2008: 9th Annual Conference of the International Speech Communication Association 2008, Vols 1-5, 2008, pp. 1813-1816.
- [65] A. Maier, F. Honig, V. Zeissler *et al.*, “A Language-Independent Feature Set for the Automatic Evaluation of Prosody,” in Proc. Interspeech, Brighton, UK, 2009, pp. 616-619.
- [66] F. Hönig, A. Batliner, K. Weilhammer *et al.*, “Automatic assessment of non-native prosody for english as l2,” in Speech Prosody, Chicago, USA, 2010.
- [67] F. Hönig, A. Batliner, and E. Nöth, “How Many Labellers Revisited–Naïves, Experts, and Real Experts,” in Proc. SLATE, Venice, Italy, 2011.
- [68] W. Menzel, D. Herron, P. Bonaventura *et al.*, “Automatic detection and correction of non-native English pronunciations,” *Proceedings of InSTILL 2000*, pp. 49-56, 2000.
- [69] M. Eskenazi, and S. Hansma, “The fluency pronunciation trainer,” in Proceedings of the STiLL Workshop, 1998.
- [70] J. Lu, R. Wang, L. C. D. Silva *et al.*, “CASTLE: a Computer-Assisted Stress Teaching and Learning Environment for Learners of English as a Second Language,” in InterSpeech 2010, Makuhari, Chiba, Japan, 2010, pp. 606-609.
- [71] C. Molina, N. Yoma, J. Wuth *et al.*, “ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion,” *Speech Communication*, vol. 51, no. 6, pp. 485-498, 2009.
- [72] D. J. Markham, and Y. Nagano-Madsen, “Input modality effects in foreign accent,” in Proc. of the 1996 International Conference on Spoken Language Processing, 1996, pp. 1473-1476 vol. 3.
- [73] C. S. Watson, and D. Kewley-Port, “Advances in computer-based speech training: Aids for the profoundly hearing impaired,” *Volta-Review* 91, pp. 29–45, 1989.
- [74] P. Martin, “WinPitch LTL II, a multimodal pronunciation software,” in ISCA, 2004.
- [75] M. Yoram, and K. Hirose, “Language training system utilizing speech modification,” in Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 1996, pp. 1449-1452.
- [76] M. Huckvale, and K. Yanagisawa, “Spoken language conversion with accent morphing,” in Proc. ISCA Speech Synthesis Workshop, Bonn, Germany, 2007, pp. 64-70.

References

- [77] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 676-681.
- [78] T. Kamiyama, "Perception of foreign accentedness in L2 prosody and segments: L1 Japanese speakers learning L2 French," in *International Conference on Speech Prosody* 2004.
- [79] Q. Yan, S. Vaseghi, D. Rentzos *et al.*, "Analysis and synthesis of formant spaces of British, Australian, and American accents," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15(2), no. 2, pp. 676-689, 2007.
- [80] S. Ghorshi, S. Vaseghi, and Q. Yan, "Cross-entropic comparison of formants of British, Australian and American English accents," *Speech Communication*, vol. 50, no. 7, pp. 564-579, 2008.
- [81] Q. Yan, and S. Vaseghi, "Modeling and synthesis of English regional accents with pitch and duration correlates," *Computer Speech & Language*, vol. 24(4), no. 4, pp. 711-725, 2010.
- [82] P. W. Boersma, David, "Praat: doing phonetics by computer ", 2013.
- [83] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453-467, 1990.
- [84] K. Hirose, F. Gendrin, and N. Minematsu, "A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice," in *INTERSPEECH*, 2003.
- [85] D. Felps, and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18(5), no. 5, pp. 1030-1040, 2010.
- [86] G. Almpianidis, M. Kotti, and C. Kotropoulos, "Robust detection of phone boundaries using model selection criteria with few observations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17(2), no. 2, pp. 287-298, 2009.
- [87] V. Khanagha, K. Daoudi, O. Pont *et al.*, "Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4484-4487.
- [88] S. King, and M. Hasegawa-Johnson, "Accurate Speech Segmentation by Mimicking Human Auditory Processing," in *International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8096-8100.
- [89] A. Turiel, H. Yahia, and C. J. Pérez-Vicente, "Microcanonical multifractal formalism—a geometrical approach to multifractal systems: Part I. Singularity analysis," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 1, pp. 015501, 2008.
- [90] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12(4), no. 4, pp. 357-370, 1993.
- [91] D. T. Toledano, L. A. H. Gomez, and L. V. Grande, "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on*, vol. 11(6), no. 6, pp. 617-625, 2003.
- [92] R. Yan, Y. Zu, and Y. Zhu, "Automatic Speech Segmentation Combining an HMM-Based Approach and Recurrence Trend Analysis," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 797-800.
- [93] M. Namnabat, and M. Homayounpour, "Refining Segmental Boundaries using Support Vector Machine," in *Proc. of 8th International Conference on Signal Processing*, Guilin, China, 2006.
- [94] F. Malfrère, O. Deroo, T. Dutoit *et al.*, "Phonetic alignment: speech synthesis-based vs. viterbi-based," *Speech Communication*, vol. 40(4), no. 4, pp. 503-515, 2003.

References

- [95] J. W. Kuo, H. Y. Lo, and H. M. Wang, "Improved HMM/SVM methods for automatic phoneme segmentation," *Proc. Interspeech 2007*, pp. 2057-2060, 2007.
- [96] S. Cox, R. Brady, and P. Jackson, "Techniques for accurate automatic annotation of speech waveforms," in *Proc. ICSLP*, 1998, pp. 1947-1950.
- [97] J. Keshet, S. Shalev-Shwartz, Y. Singer *et al.*, "A Large Margin Algorithm for Speech-to-Phoneme and Music-to-Score Alignment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15(8), no. 8, pp. 2373-2382, 2007.
- [98] J. Yuan, N. Ryant, M. Liberman *et al.*, "Automatic Phonetic Segmentation Using Boundary Models," in *Proc. of Interspeech*, Lyon, France, 2013.
- [99] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51(4), no. 4, pp. 352-368, 2009.
- [100] P. Seung Seop, and K. Nam Soo, "Automatic Speech Segmentation Based on Boundary-Type Candidate Selection," *Signal Processing Letters, IEEE*, vol. 13(10), no. 10, pp. 640-643, 2006.
- [101] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Computer Speech & Language*, vol. 24(2), no. 2, pp. 273-288, 2010.
- [102] S. S. Park, and N. S. Kim, "On using multiple models for automatic speech segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15(8), no. 8, pp. 2202-2212, 2007.
- [103] C. Y. Lin, K. T. Chen, and J. S. R. Jang, "A hybrid approach to automatic segmentation and labeling for Mandarin Chinese speech corpus," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 1553-1556.
- [104] K. S. Lee, "MLP-based phone boundary refining for a TTS database," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14(3), no. 3, pp. 981-989, 2006.
- [105] H. Y. Lo, and H. M. Wang, "Phone boundary refinement using ranking methods," in *Chinese Spoken Language Processing (ISCSLP)*, 2010 7th International Symposium on, 2010, pp. 488-492.
- [106] L. Cheng-Yuan, and J. Jyh-Shing, "Automatic Phonetic Segmentation by Score Predictive Model for the Corpora of Mandarin Singing Voices," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15(7), no. 7, pp. 2151-2159, 2007.
- [107] E. Akdemir, and T. Ciloglu, "HMM topology for boundary refinement in automatic speech segmentation," *Electronics Letters*, vol. 46(15), no. 15, pp. 1086-1087, 2010.
- [108] A. Fox, *Prosodic features and prosodic structure: The phonology of suprasegmentals*: Oxford University Press, 2000.
- [109] K. a. T. Imoto, Y. and Raux, A. and Kawahara, T. and Dantsuji, M., "Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system," in *Seventh International Conference on Spoken Language Processing*, 2002.
- [110] A. Rosenberg, and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009, pp. 81-84.
- [111] S. Tokuma, and Y. Xu, "The effect of F0 peak-delay on the L1/L2 perception of English lexical stress," in *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association 2009*, Vols 1-5, 2009, pp. 1619-1622.
- [112] L. Mary, and B. Yegnanarayana, "Prosodic Features For Language Identification," in *Signal Processing, Communications and Networking*, 2008. ICSCN '08. International Conference on, 2008, pp. 57-62.
- [113] J. P. Arias, N. B. Yoma, H. Vivanco *et al.*, "Word stress assessment for computer aided language learning," in *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association 2009*, Vols 1-5, 2009, pp. 1163-1166.

References

- [114] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, 1995.
- [115] S. J. Young, G. Evermann, M. Gales *et al.*, *The HTK book version 3.4*, 2006.
- [116] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: training recipes and recognition experiments," *Cavendish Laboratory, University of Cambridge*, 2006.
- [117] E. F. Mark Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.
- [118] X. J. Sun, "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," in Proc. ICASSP, Orlando, USA, 2002, pp. 333-336.
- [119] I. Luengo, I. Saratxaga, E. Navas *et al.*, "Evaluation of Pitch Detection Algorithms Under Real Conditions," in Proc. ICASSP, Hawaii, USA, 2007, pp. 1057-1060.
- [120] A. Rosenberg, and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in Proc. Human Language Technologies, 2009, pp. 81-84.
- [121] R. Yaniv, and D. Burshtein, "An enhanced dynamic time warping model for improved estimation of DTW parameters," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 216-228, 2003.
- [122] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 6, pp. 623-635, 1980.
- [123] E. J. Keogh, and M. J. Pazzani, "Derivative dynamic time warping," in First SIAM international conference on data mining, 2001.
- [124] C. Cortes, and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.
- [125] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121-167, 1998.
- [126] P. H. Chen, C. J. Lin, and B. Scholkopf, "A tutorial on v-support vector machines," *Applied Stochastic Models in Business and Industry*, vol. 21, no. 2, pp. 111-136, 2005.
- [127] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.
- [128] S. Gunn, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, 1998.
- [129] S. S. Keerthi, and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural computation*, vol. 15, no. 7, pp. 1667-1689, 2003.
- [130] H.-T. Lin, and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," *submitted to Neural Computation*, pp. 1-32, 2003.
- [131] Q. Chang, Q. C. Chen, and X. L. Wang, "Scaling Gaussian RBF kernel width to improve SVM classification," in Proceedings of the 2005 International Conference on Neural Networks and Brain, Vols 1-3, 2005, pp. 19-22.
- [132] M. G. Genton, "Classes of kernels for machine learning: a statistics perspective," *The Journal of Machine Learning Research*, vol. 2, pp. 299-312, 2002.
- [133] Y. Kamada, and S. Abe, "Support vector regression using Mahalanobis kernels," *Artificial Neural Networks in Pattern Recognition*, pp. 144-152, 2006.
- [134] C.-C. Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2(3), no. 3, pp. 27:1 - 27:27, 2011.
- [135] M. A. Hall, "Correlation-based feature selection for machine learning," *PhD Thesis*, 1999.
- [136] R. Kohavi, and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.

References

- [137] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, Aug, 2005.
- [138] T. Konno, M. Ito, M. Suzuki *et al.*, "Intonation Evaluation of English Utterances using Synthesized Speech for Computer-Assisted Language Learning." pp. 202-208.
- [139] M. Abe, S. Nakamura, K. Shikano *et al.*, "Voice conversion through vector quantization," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1988, pp. 655-658 vol. 1.
- [140] O. Turk, and L. M. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in Proc. of INTERSPEECH, 2003.
- [141] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 922-931, 2010.
- [142] H. Duxans, D. Erro, J. Pérez *et al.*, "Voice Conversion of Non-aligned Data using Unit Selection," in Spain TC-STAR Workshop on Speech-to-Speech Translation 2006.
- [143] E. Moulines, and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175-205, 1995.
- [144] e. a. John Garofolo, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.
- [145] S. Panchapagesan, and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer Speech & Language*, vol. 23, no. 1, pp. 42-64, 2009.
- [146] T. Kinnunen, and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12-40, 2010.
- [147] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832-844, 2009.
- [148] L. Malfait, J. Berger, and M. Kastner, "P. 563-the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 14, no. 6, pp. 1924-1934, 2006.
- [149] K. Nagano, and K. Ozawa, "English speech training using voice conversion," in ICSLP, Kobe, Japan, 1990.
- [150] H. Banno, H. Hata, M. Morise *et al.*, "Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation," *Acoustical science and technology*, vol. 28, no. 3, pp. 140-146, 2007.
- [151] H. Kawahara, T. Takahashi, M. Morise *et al.*, "Development of exploratory research tools based on TANDEM-STRAIGHT," in APSIPA ASC 2009, Singapore, 2009, pp. 111-120.
- [152] A. B. Kain, "High resolution voice transformation," *PhD Thesis*, 2001.
- [153] D. Erro, and A. Moreno, "UPC Voice Conversion Toolkit," <http://147.83.50.136/taip/index.php/es/recursos/herramientas/voice-conversion>.
- [154] H. Kawahara, http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html.
- [155] J. Kominek, C. L. Bennett, and A. W. Black, "Evaluating and correcting phoneme segmentation for unit selection synthesis," in Proc. of Interspeech, Geneva, Switzerland, 2003, pp. 313-316.
- [156] K. F. Lee, and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37(11), no. 11, pp. 1641-1648, 1989.
- [157] H. Kawai, and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

References

- [158] H. Yu, and J. Yang, "A direct LDA algorithm for high-dimensional data-with application to face recognition," *Pattern recognition*, vol. 34(10), no. 10, pp. 2067-2069, 2001.
- [159] C. Y. Lin, J. S. R. Jang, and K. T. Chen, "Automatic segmentation and labeling for mandarin chinese speech corpora for concatenation-based TTS," *Computational Linguistics and Chinese Language Processing*, vol. 10(2), no. 2, pp. 145-166, 2005.
- [160] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," *Technical Report, Department of Computer Science, National Taiwan University*, 2003.
- [161] T. Robinson, J. Fransen, D. Pye *et al.*, "WSJCAM0 Cambridge Read News," Linguistic Data Consortium, 1995.
- [162] Q. Yan, and S. Vaseghi, "A comparative analysis of UK and US English accents in recognition and synthesis," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, USA, 2002, pp. 413-416.