JAIST Repository
https://dspace.jaist.ac.jp/

Title	Studies on Spectral Modification in Voice Transformation
Author(s)	Nguyen, Binh Phu
Citation	
Issue Date	2009-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/8002
Rights	
Description	Supervisor:instruction Prof. Masato Akagi, 情報科学研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Studies on Spectral Modification in Voice Transformation

by

Binh Phu Nguyen

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology

March 2009

Abstract

This dissertation aims to propose spectral modelings and spectral modification algorithms to improve the quality of modified speech in voice transformation.

Voice transformation is a process of changing certain perceptual properties of speech while leaving other properties unchanged. Voice transformation has many applications in our lives. For example, we employ voice transformation techniques to create various wave sounds from a limited pre-recorded database in a Text-to-Speech system. In foreign language learning, it will be much easier to listen when slowing down the speed of sounds. To enhance the hearing abilities of deaf people, we can adjust the frequency of sounds so that it is located in their hearing portion.

One of the core processes in voice transformation is spectral modification. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing, such as manipulations of the formant structures, amplitude manipulations. The challenge of spectral modification is to modify the spectral/acoustical features without degrading the speech quality. Most high-quality spectral modification methods operate in time-frequency domain. The representations of speech signals in the time-frequency domain can describe the speech signals well. The reason is that when a human produces voices, the continual motion of the articulators forms a time-varying acoustic filter which is responsible for the generation of the speech waveform.

Although many high-quality spectral modification methods have been proposed in the literature, they still have issues. In the time axis, most of them process speech signals frame by frame. They lack of a model to describe the temporal evolution of parameters. Therefore, they do not ensure the smoothness of synthesized speech after modification. In addition, they either do not guarantee to keep the natural evolution of parameters of speech signals. These limitations degrade the quality of modified speech. In the frequency axis, spectral modification can be performed by the rule-based approach or the statistical approach. The rule-based approach often needs small training data. However, this approach only stores basic rules, and the basic rules do not reflect the natural characteristics of the speech signals. While the statistical approach performs spectral modification by using machine learning techniques, and it requires large training data. Therefore, applying which approach depends on applications and specific conditions. Three main issues of spectral modification methods can be summarized as follows.

1. The first issue is lack of efficient spectral modelings for speech modification. The spectral modelings determine which attributes can be processed and how these attributes can be modified. One of requirements of these spectral modelings is that they allow to ensure smoothness of modified speech signals and to perform efficient spectral modification.

2. The second issue is insufficient smoothness of modified spectra between frames. Conventional methods of spectral modification often perform spectral modification frame by frame. When unexpected modifications happen, there are discontinuities of

speech spectra between frames. This leads to degradation of the quality of modified speech.

3. The third issue is ineffective spectral modification. The rule-based techniques of spectral modification, such as linear prediction (LP)-based methods and frequency warping methods, are limited by their inability to independently control important formant characteristics such as amplitude, bandwidth or to control the spectral shape. The statistical techniques of spectral modification take advantages of mathematical or statistical models, but they lack of acoustic constraints. It therefore requires a new method for performing efficient spectral modification.

The main goals of this dissertation are to deal with the three major issues of spectral modification mentioned above, i.e. lack of spectral modelings for speech modification, insufficient smoothness of the modified spectra between frames, and ineffective spectral modification.

To perform spectral modification, we first develop an analysis/synthesis framework. When human beings produce speech voices, the continual motion of the articulators forms a time-varying acoustic filter which is responsible for the generation of the speech waveform. To characterize this type of properties, we need a joint time-frequency representation. In the first part of this dissertation, we first introduce improvements of speech spectral envelope modeling, and then we present a new modeling of speech spectral sequence. Conventional representations of a speech spectral envelope, such as LP coefficients or non-parametric representations, meet difficulties in controlling spectral peaks and spectral shape. In our framework, we explore the Gaussian mixture model parameters proposed by Zolfaghari et al. (called spectral-GMM parameters in this dissertation) to model a speech spectral envelope. In this technique, formants are assumed to be represented by Gaussian distributions, and a speech spectral envelope could be represented by a Gaussian mixture model. Although the original method well models and flexibly controls the speech spectral envelope, it still has two main problems. The first problem is difficulty in modifying a speech spectral envelope in both axes, frequency and amplitude. The second problem is that a Gaussian distribution does not always fit to a formant very well, especially for other kinds of voices which are not reading voices, such as singing voices. To solve these drawbacks, we propose two improvements in the speech spectral envelope modeling. We employ constraints to model a spectral peak by using only one Gaussian component. We also use an asymmetric Gaussian mixture model to model a speech spectral envelope, instead of a Gaussian mixture model. We then develop a new modeling of speech spectral sequence based on temporal decomposition (TD), which originally proposed by Atal, and spectral-GMM parameters. In our modeling, the TD algorithm is utilized to model the temporal evolution (in the time domain), and spectral-GMM parameters are used to model the speech spectral envelope (in the frequency domain). Experimental results show that our modeling models speech signals very well. In addition, our analysis/synthesis method is potential to ensure the smoothness of modified speech, and to perform efficient spectral modification.

To solve the second issue, the insufficient smoothness of modified spectra between frames, one of efficient ways is to control spectral dynamics. In this dissertation, we employ the TD technique, which decomposes speech into event targets and event functions, to control spectral dynamics to improve the quality of synthesized speech. Based on the TD technique, we propose a new method to improve the quality of modified speech

signals in concatenative speech synthesis. Concatenative speech synthesis systems form utterances by concatenating pre-recorded speech units. In concatenative speech synthesis systems, output speech is limited by the contents of the pre-recorded databases, and inevitable concatenation errors can lead to audible discontinuities. To reduce the discontinuities between speech units, many methods have been presented in the literature. However, some steps in these methods need to be manually performed, due to preparation of "fusion" units, or extraction of formants. Therefore, it is necessary to have a new method which can automatically smooth the mismatch between speech segments. In our method, we automatically decompose speech units by using the TD technique. The same event functions evaluated for the spectral parameters are also used to describe the temporal patterns of the excitation parameters. The modifications of spectral parameters, F0 and gain information at the joint parts of the speech units are performed by altering the last and the first event and excitation targets of the first and the last speech units, respectively. As a result, the mismatch of spectral, F0, and gain information are reduced at the concatenation points.

To solve the third issue, the ineffective spectral modification, we develop a new efficient algorithm of spectral modification which is applied for rule-based methods, and propose an improvement of GMM-based voice conversion methods.

In our proposed algorithm of spectral modification, we utilize spectral-GMM parameters to model a speech spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. It is well-know that formant frequencies are some of the most important parameters in characterizing speech, and control of formants can effectively modify the spectral envelope. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. Therefore, we develop a new algorithm to modify spectral-GMM parameters in accordance with formant frequencies. We then apply our algorithm to two areas, emotional speech synthesis which requires modifications of both formant frequency and power, and voice gender conversion which requires a large amount of spectral modification. Experimental results show the effectiveness of our spectral modification algorithm.

In state-of-the-art voice conversion systems, GMM-based spectral voice conversion methods are regarded as some of the best systems. However, the quality of converted speech is still far from natural. This dissertation presents a new spectral voice conversion method to deal with two drawbacks of the conventional GMM-based spectral voice conversion methods, insufficient precision of GMM parameters and insufficient smoothness of the converted spectra between frames. For improvement of the estimation of GMM parameters, we utilize phoneme-based features of event targets as spectral vectors to take into account relations between spectral parameters in each phoneme, and to avoid using spectral parameters in transition parts. For enhancement of the continuity of speech spectra, we only need to convert event targets, instead of converting source features to target features frame by frame, and the smoothness of the converted speech is ensured by the shape of the event functions. Experimental results confirm the high-quality of our converted speech.

In summary, the main contributions of this dissertation are as follows.

1. Propose a new approach of spectral modeling for speech modification. Our method is potential to ensure smoothness of modified speech, and to perform efficient spectral

modification.

2. Propose an efficient framework to model and control temporal evolution, which can ensure the naturalness and the smoothness of modified speech. This feature improves the quality of modified speech.

3. Propose a new efficient algorithm of spectral modification. Our algorithm performs spectral modification directly on the speech spectral envelope, which is flexible to modify the speech spectral envelope, and does not produce artifacts.

4. Apply phoneme constraints to the GMM-based spectral voice conversion method to improve the quality of transformation functions, which leads to enhancement of converted speech quality.

Key words: spectral modification, voice transformation, temporal decomposition, spectral-GMM

Acknowledgments

This dissertation would not have been possible without the continuous and unconditional supports of many people to whom I owe a great debt of gratitude.

First and foremost, I would like to thank my supervisor, Professor Masato Akagi, for bringing me into his laboratory and providing a tremendously supportive environment for my research. He opened a door for me to the world of speech signal processing. He followed my research closely when I needed it and gave me freedom to do research when I would like. He provided much invaluable knowledge about academic life. His seriousness and dedication to work have been inspirational to me.

I would also like to thank Associate Professor Masashi Unoki for his supports and comments on my academic research. Associate Professor Masashi Unoki is one of two professors, along with Professor Masato Akagi, who have been a major influence on my life since I commenced my PhD course. Their invaluable comments and guidance helped me a great deal when I started to enter the field of speech signal processing.

I am very grateful to Associate Professor Isao Tokuda for his supervision of my sub-theme research. His enthusiasm and useful guidelines gave me a good opportunity to learn another domain of speech signal processing. This knowledge is very helpful as I pursue my research at the present and in the future.

I wish to thank Professor Jianwu Dang for serving as a member of my dissertation committee and giving me valuable comments on my work in the PhD defense as well as during the joint meetings of the AIS and IIP laboratories.

I am also grateful to Associate Professor Tatsuya Kitamura from Konan University, Japan for serving as a member of my dissertation committee and giving me many insightful comments.

Thanks also go to Professor Donna Erickson from Showa Music University, Kawasaki City, Japan and Associate Professor Ken-Ichi Sakakibara from the Department of Communication Disorders, Health Sciences University of Hokkaido, Japan for their useful discussions on my sub-theme.

I thank my lab-mates for their contributions in developing a wonderful and supportive academic environment, especially Mr. Yasuki Murakami, my "tutor", who gave much help throughout the years I lived at JAIST. Their generous assistance made my life easier and more enjoyable.

I appreciate Dr. Matt Stuttle of Cambridge University, UK for kindly providing me with his doctoral dissertation. It gave me the initial concepts of modeling the speech spectrum using the Gaussian mixture model. Thanks also go to Dr. David Suendermann, New York City, USA for providing me with his voice conversion Matlab toolbox. I also owe many people whom I cannot name here a great deal for their enthusiastic helps and useful comments on my academic topics.

I would like to acknowledge the financial supports from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), the SCOPE of the Japanese Ministry of Internal Affairs and Communications (MIC).

I would like to thank JAIST for offering me a wonderful research environment. I owe a great deal to the Student Service Department at JAIST, the JAIST Technical Communication Program, especially Professor Mary Ann Mooradian, for their valuable helps.

My special thanks go to Professor Ho Tu Bao for his thoughtful advice and supports, the Vietnamese community at JAIST for sharing ups and downs in the last three years.

Finally, I have saved the best for the last. I thank my family for everything they have done in my upbringing. Their sacrifice, love and supports are always an endless source of inspiration for me to move forwards. Most of all, I thank my beloved wife, Dam Thi Quynh Lien, and my daughter, Nguyen Chi Mai. Without their understanding, supports and love, none of this dissertation would have even begun.

Contents

Abstract **i**

Acknowledgments **v**

Acronyms **xiii**

1 Introduction **1**
1.1 Definition of voice transformation 1
1.2 The need for voice transformation 2
1.3 Analysis/synthesis methods 3
1.4 Spectral modification techniques 5
1.4.1 Rule-based approach 6
1.4.2 Statistical approach 9
1.5 Motivation and scope of the research 11
1.6 Main contribution of the dissertation 15
1.7 Outline of the dissertation 16
1.8 Summary 18

2 Research Background **21**
2.1 Speech production 21
2.2 The source-filter model for speech production 23
2.3 Linear prediction model 25
2.4 Temporal decomposition 28
2.4.1 Introduction 28
2.4.2 Atal's method of temporal decomposition 29
2.4.3 Modified restricted temporal decomposition (MRTD) 31
2.5 Speech spectrum modeling using Gaussian mixture model 36
2.5.1 Introduction 36
2.5.2 Estimation of spectral-GMM parameters 37
2.5.3 Initialization 40
2.5.4 Issues in estimating spectral-GMM parameters from a speech spectrum 41
2.5.5 Properties of the spectral-GMM parameters 43
2.6 STRAIGHT 43
2.6.1 Outline of STRAIGHT 43

2.6.2 Derivation of LSF parameters 45
2.7 Summary 45

3 Spectral Modelings for Speech Modification 46
3.1 Introduction 46
3.2 Modelings of a speech spectral envelope 47
3.2.1 Speech spectrum modeling using Gaussian mixture model 47
3.2.2 Smoothed-spectrum representation by STRAIGHT 48
3.2.3 Improvement of spectral peak estimation using Gaussian mixture model 48
3.2.4 Spectral envelope modeling using asymmetric Gaussian mixture model 50
3.2.5 Experiments and results 54
3.2.6 Conclusions 58
3.3 Modeling of the speech spectral sequence 58
3.3.1 Introduction 58
3.3.2 Temporal decomposition 60
3.3.3 Modeling of the event function using polynomial fitting 61
3.3.4 Proposed modeling of the speech spectral sequence 62
3.3.5 Merits of our modeling in speech modification 63
3.3.6 Experiments and results 65
3.3.7 Conclusions 67
3.4 Summary 67

4 Spectral Smoothing for Concatenative Speech Synthesis based on Temporal Decomposition 68
4.1 Introduction 68
4.2 Related work 70
4.3 Proposed method 71
4.3.1 Overview of our proposed method 71
4.3.2 Proposed method 72
4.4 Experiments and results 74
4.5 Conclusions 74

5 Rule-based Approach to Spectral Modification 76
5.1 Introduction 76
5.2 Proposed spectral modification algorithm 78
5.3 Experiments and results 79
5.3.1 Application to emotional speech synthesis 79
5.3.2 Application to voice gender conversion 82
5.4 Conclusions 85

6 Statistical Approach to Spectral Modification 86
6.1 Introduction 86
6.2 Problem formulation 88
6.3 Conventional GMM-based voice conversion 90

6.3.1 Training procedure 90
6.3.2 Transformation procedure 90
6.4 Temporal decomposition 91
6.5 Proposed spectral voice conversion method 92
6.5.1 Spectral parameters 92
6.5.2 Proposed method 92
6.6 Experiments and results 93
6.6.1 Experimental conditions 93
6.6.2 Objective test 94
6.6.3 Subjective tests 95
6.7 Conclusions 96

7 Summary and Future Work **97**
7.1 Summary of the dissertation 97
7.2 Further research directions 100

References **104**

Publications **117**

List of Figures

1.1 Block diagram of a spectral processing framework 5
1.2 Examples of linear warping functions over a speech spectral envelope. The warping functions cause an expansion or compression of the speech spectral envelope. 8
1.3 Schematic overview of the dissertation. 19
1.4 Correspondence between our work and the flowchart of voice transformation. 20

2.1 The human vocal organs [29] 22
2.2 Block diagram of simplified model for speech production. 24
2.3 Block diagram of the MRTD algorithm [110]. 32
2.4 Example of two adjacent event functions in the second order TD model. . . 32
2.5 The path in parameter space described by the sequence of spectral parameters $\mathbf{y}(n)$ is approximated by means of straight line segments between breakpoints. 33
2.6 Determination of the event functions in the transition interval $[n_k, n_{k+1}]$. The point of the line segment between $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ (a), between $\hat{\mathbf{y}}(n-1)$ and $\mathbf{a}_{k+1}$ (b) with minimum distance from $\mathbf{y}(n)$ is taken as the best approximation. 34
2.7 Examples of a well-shaped event function (a) and an ill-shaped event function (b). 34
2.8 Example of event targets and event functions which are decomposed from the MRTD algorithm. 36
2.9 Block diagram of estimation of spectral-GMM parameters from speech signals 38
2.10 Probability distribution $P(x_k)$. 39
2.11 Schematic structure of STRAIGHT (adapted from [12]). 44

3.1 A Gaussian mixture model (M=8) fits to a STRAIGHT spectral envelope. 48
3.2 Illustration of the Empirical rule of Gaussian distribution. 49
3.3 Diagram of our proposed method for modeling the speech spectral envelope. 50
3.4 Example of the spectral envelope restored by our proposed method: the thin lines indicate Gaussian components, the bold line indicates the restored spectral envelope (top), and STRAIGHT spectral envelope (bottom). 51
3.5 Example of Gaussian and AG distribution: Gaussian (left), and asymmetric Gaussian (right). 52
3.6 Illustration of the Empirical rule of AG. 53
3.7 Comparison of listening test results for the compared methods. 56

3.8 Comparison of fitting in the STRAIGHT smoothed spectrum by using GMM (10 components), and AGMM (10 components). The solid line is a STRAIGHT spectrum of a frame, dash-dot line is AGMM fitting in the STRAIGHT spectrum, and the dashed line is GMM fitting in the STRAIGHT spectrum 58
3.9 Comparison of mean opinion scores (MOSs) over varying number of Gaussians in the mixture using spectral-AGMM method and spectral-GMM method. 59
3.10 A fitted curve (the solid line) using the non-linear least square method for an event function (the dashed line) (top), and a time-scale modification of this event function (bottom). 62
3.11 Diagram of proposed modeling of speech spectral sequence based on temporal decomposition and Gaussian mixture model. 63
3.12 Diagram of our proposed method for modifying amplitudes of spectral peaks 65
3.13 Example of the amplitude modification of the first three formants by 2, 3, and 2.5 times, respectively. 24 Gaussian components (M=24) are used to model this speech spectral envelope. 66

4.1 Diagram of our proposed method. 73
4.2 Results of subjective tests of concatenative speech synthesis. 74
4.3 Parts of the LSF contours before and after modification at the concatenation points by replacing the vowel *"u"* in the word *"takumi"* by the vowel *"e"* in the word *"jiten"*. The dot line indicates the LSF contours of the two speech units before modification. The solid line indicates the LSF contours of the two speech units after modification by using our proposed method. . 75

5.1 Block diagram of our spectral modification algorithm. 78
5.2 Example of our spectral modification algorithm applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$. 80
5.3 Example of formant modification algorithm of the LSF-based method [99] applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$. 81
5.4 Evaluation measure of Scheffe's paired comparison (five grades: -2, -1, 0, 1 and 2). 82
5.5 Subjective listening results for emotional speech synthesis. 82

6.1 General architecture of a voice conversion system. 87
6.2 Diagram of our proposed voice conversion method training procedure (top), and transformation procedure (bottom). 93

List of Tables

3.1 Analysis conditions for experiments of testing methods 55
3.2 Average LSD, and percentage number of outlier frames obtained from (1) the original GMM method and (2) our proposed method. 56
3.3 Analysis conditions for experiments of AGMM performance 57
3.4 Average LSD, and percentage number of outlier frames obtained from the spectral-AGMM and spectral-GMM methods. The first line in each row is the result of the spectral-GMM method, and the second line is the result of the spectral-AGMM method. 57
3.5 Average PESQ for analysis/synthesis methods. 67

5.1 Analysis conditions for experiments of emotional speech synthesis. 81
5.2 Analysis conditions for experiments of VGC system. 84
5.3 Subjective listening results for VGC system (1) STRAIGHT + LSF (2) STRAIGHT + framewise-GMM (3) our proposed system (STRAIGHT + TD-GMM). 85

6.1 Analysis conditions for experiments on the voice conversion methods. . . . 94
6.2 Objective results for the voice conversion methods (1) Conventional method (GMM method) (2) TD+GMM method (3) our proposed method (Phoneme-based TD+GMM method). 95
6.3 MOS results for voice conversion methods (1) Conventional method (GMM method) (2) TD+GMM method (3) our proposed method (Phoneme-based TD+GMM method). 95
6.4 ABX results for voice conversion methods (1) Conventional method (GMM method) (2) TD+GMM method (3) our proposed method (Phoneme-based TD+GMM method). 96

Acronyms

ABS	Analysis by synthesis
AG	Asymmetric Gaussian
AGMM	Asymmetric Gaussian mixture model
DFT	Discrete Fourier transform
DTW	Dynamic time warping
EM	Expectation maximization
FFT	Fast Fourier transform
GMM	Gaussian mixture model
HMM	Hidden Markov model
LP	Linear prediction
LPC	Linear predictive coding
LSD	Log spectral distortion
LSF	Line spectral frequency
MFCC	Mel frequency cepstral coefficient
ML	Maximum likelihood
MOS	Mean opinion score
MRTD	Modified restricted temporal decomposition
OLA	Overlap-add
PESQ	Perceptual evaluation of speech quality
PSOLA	Pitch-synchronous overlap-add
SEEVOC	Spectral envelope estimation vocoder
SFTR	Spectral feature transition rate
spectral-AGMM	spectral asymmetric Gaussian mixture model
spectral-GMM	spectral Gaussian mixture model
STFT	Short-time Fourier transform
STRAIGHT	Speech transformation and representation using adaptive interpolation of weighted spectrum
TD	Temporal decomposition
TTS	Text-to-Speech
VQ	Vector quantization
VTLN	Vocal tract length normalization

Chapter 1

Introduction

In this chapter, we briefly introduce the research context, the research motivations, as well as the major contributions of the dissertation. We begin by introducing the need for voice transformation. In the next parts, we present an overview of analysis/synthesis methods, and discuss typical spectral modification methods. We then state the motivation and scope of this dissertation. The main contributions of this dissertation are also shortly mentioned. Finally, the structure of this dissertation is outlined.

1.1 Definition of voice transformation

This dissertation aims to study on spectral modification of speech to improve the quality of modified speech in the field of voice transformation.

The capability of speaking a language is one of the most wonderful skills of human beings. Speech is one of the communication ways that human beings most widely use. It serves as a very effective way of communication, such as sharing experiences, feelings, concepts and ideas among people. As the most natural way of communication between human beings, speech is a subject which has attracted much interest and attention. The structure of speech, speech production, and speech perception mechanisms have been researched by linguists, psychologists and physiologists.

Nowadays, advances in electronic and computer technology are causing an explosive growth in the use of machines for processing information. Scientists and engineers try to improve the communication not only among human beings but also between human beings and machines via human beings' voices in different environments. In general, two major technologies for the communication between human beings and machines are speech recognition and speech synthesis. Speech recognition is a technique to recognize a set of words from voices of any speakers with high accuracy rate. It can be seen as a technique for processing information input. On the other hand, speech synthesis is a technique for processing information output. It can be defined as the artificial production of human speech. In this context, speech synthesis is the reverse of speech recognition.

The central topic of this dissertation, spectral modification in voice transformation, can be considered as a core part of the speech synthesis area. The goals of voice transformation systems are to generate wave sounds from a pre-recorded speech database, or to alter styles

of speech utterances/segments without losing the utterance/segment content, etc. The styles which can be changed include the speaker's gender, the speaker's identity, or the speaker's emotion, and so on. A voice transformation system often has to be capable of accomplishing two main tasks:

1. *Training phase:* The system determines the optimal transformation for converting speech utterances/segments into the other ones from training data.

2. *Transformation phase:* The system applies this optimal transformation to convert new input utterances/segments.

1.2 The need for voice transformation

Voice transformation is an important technique in Text-to-Speech. In addition, voice transformation also has many applications in our lives, such as in education, aid-to-the handicapped, and entertainment.

Text-to-Speech: The main goal of Text-to-Speech is to produce natural speech sounds from texts. The most successful TTS approach to-date is called concatenative synthesis. In this approach, natural speech utterances of speakers first are recorded and stored in an acoustic inventory. During synthesis, individual portions of a speech utterance are retrieved from the inventory, optionally modified, and then concatenated in the desired sequence. Therefore, it is necessary to smooth the mismatch between these speech units. In addition, to create various kinds of databases, we employ voice conversion techniques to modify utterances of a speaker to those of another speaker, or we convert neutral utterances to emotional utterances.

Education: When learning foreign languages, proper intonation of sentences and pronunciation of words is one of the most difficult tasks for learners. Slowing down conversions helps the learners to properly pronounce foreign words/sentences, as well as to practice their listening skill.

Aid-to-the handicapped: Voice transformation techniques can be employed to enhance the speech quality for both speakers and listeners. For a handicapped speaker, we can improve the speech intelligibility, since the intelligibility of his/her voices may be affected by abnormal controls over phoneme duration and pitch variations. For a handicapped listener, e.g. a person with hearing disabilities, we can employ voice transformation techniques to put the frequency space to the hearing portion of his/her ears.

Entertainment: In the game industry, voice transformation can create virtual voices to achieve interesting atmospheres. In the film industry, voice transformation can be used to dub the voice of actors/actresses in different languages.

Others: Voice transformation can be integrated into car navigation systems, voice-enabled e-mail systems, etc.

Moreover, voice transformation can be also applied to the most classical areas of the speech technology, such as speech coding, speech recognition, speaker verification and identification, speech enhancement. In speech coding, we perform time-scale modification for data reduction to save storage space and transmission bandwidth. In speech recognition, there is variety of speakers. Each speaker has own characteristics, such as gender, accent, identity; even the same speaker also has different styles, due to changes of emotions. Therefore, to improve the accuracy rate of speech recognition, we need to perform speaker adaptation or vocal tract length normalization to reduce the variety of speakers. In speaker verification and identification, we apply speech modification methods to remove the linguistic contents, and keep the speaker's identity. In speech enhancement, voice conversion techniques are useful tools to improve the speech quality and/or intelligibility.

1.3 Analysis/synthesis methods

There are various attributes of speech signals which can be modified, such as speed, formants, fundamental frequency. Speech modification can be classified into three main groups as follows.

- **Time-scale modification:** It is a process of modifying the duration of a speech signal, while maintaining other quality, such as the pitch and the timbre, unchanged.
- **Pitch-scale modification:** Aim of pitch-scale modification is to change fundamental frequency information of a speech signal, while maintaining the original time and spectral properties.
- **Spectral modification:** Purpose of spectral modification is to change spectral attributes of a speech signal, while keeping the duration and fundamental frequency.

In voice transformation, analysis/synthesis methods are very important, since they determine which features and algorithms for modifications. Many analysis/synthesis methods are available in the literature. They process speech signals in the time-domain [51, 95, 96], in the frequency-domain [33, 40, 98], and in the time-frequency domain [42, 43, 44, 91, 92]. We now present some typical analysis/synthesis methods.

Many analysis/synthesis methods are waveform-based systems, such as pitch-synchronous overlap-add (PSOLA) [51, 95, 96]. PSOLA processes directly on the waveform to incorporate the desired prosody information. It operates by sampling windowed portions of the original signal and then re-synthesizing them with a basic overlap-add procedure. Time-scale modification is performed by deleting or repeating windowed sections prior to the overlap-add procedure. Pitch-scale modification is also possible by adjusting the spacing between overlapped windows during re-synthesis. These methods have been a popular choice mainly because of their simplicity and capability of high fidelity playback. However, they offer only crude modifications that often result in objectionable artifacts.

The source-filter model for speech production [39] is based on a simple model of speech production. According to this model, the simplified human voice production system is

decomposed into three components: glottal source, vocal tract, and radiation impedance. The main challenge of the source-filter model is the estimation of the glottal excitation and vocal tract filter parameters from the speech signal. In this model, it is assumed that there is no interaction between the source and the filter. Hence, the individual acoustic properties of the source and the filter can be separately simulated. The vocal tract filter can be modeled as an acoustic tube with a varying cross-sectional area formed by the pharynx, the oral cavity, the nasal cavity, and the lips. Depending on the shape of the acoustic tube, a sound wave traveling through it is reflected in a certain way so that interferences generate resonances at certain frequencies. These resonances are called formants. The ability to manipulate the characteristics of the vocal tract largely depends on the formant structure of the vocal tract spectrum. Formant characteristics have long been known to be important in the area of speech signal processing. Modification of the formant structure can be performed in a number of ways. All-pole models, such as linear prediction (LP) models, offer formant modification through the shifting and scaling of pole locations. Other methods modify the spectral envelope with functions that warp the envelope along the frequency and/or amplitude axes. These methods, however, are capable of performing only limited modifications and offer little control over important formant characteristics. For example, pole modification does not allow a particular formant's bandwidth and amplitude to be independently controlled.

Sinusoidal model was initially explored by McAulay and Quatieri [91, 92]. It is based on modeling the time-varying spectral characteristics of a sound as sums of time-varying sinusoids. An input sound is modeled by

$$s(t) = \sum_{i=1}^{R} A_i(t) cos(\theta_i(t)) \tag{1.1}$$

where $A_i(t)$ and $\theta_i(t)$ are the instantaneous amplitude and phase of the ith sinusoid, respectively.

By representing a voiced waveform as a sum of sinusoidal components, the sinusoidal model has been shown to be capable of producing high-quality speech, even after pitch and time-scale transformations. Later work with this model showed the potential for time-scale modification and pitch alteration [122, 123]. An extension to McAulay and Quatieri's work, the Analysis-by-Synthesis/Overlap-Add (ABS/OLA) model, was developed by George and Smith [42, 43, 44]. This model is based on the combination of a block overlap-add sinusoidal representation and an analysis-by-synthesis parameter estimation technique. ABS/OLA performs synthesis by employing an efficient FFT implementation. The sinusoidal model is an attractive representation of speech. However, because of a high number of sinusoidal amplitudes, frequencies and phases involved, the sinusoidal modeling is less flexible to modify spectral features than the source-filter model. In addition, this model does not yield control over the speech in terms of formant frequencies and bandwidths [160].

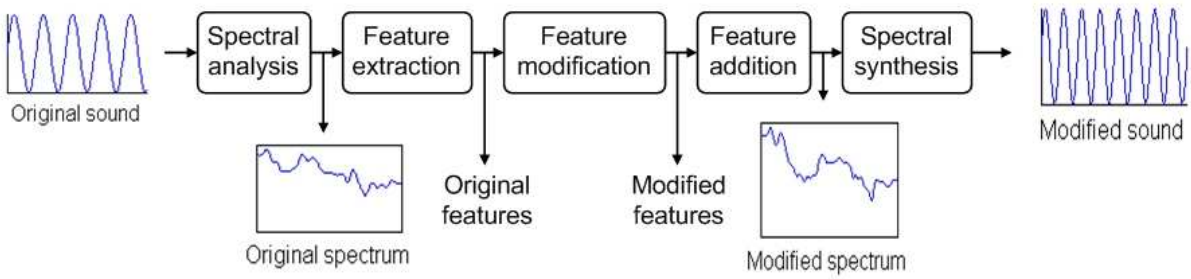

Figure 1.1: Block diagram of a spectral processing framework

1.4 Spectral modification techniques

Spectral modification is one of the core processes in voice transformation. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing, such as manipulations of the formant structures, amplitude manipulations. The diagram of spectral processing is shown in Figure 1.1. The basic idea of spectral processing is to convert a time-domain digital signal into its representation in a time-frequency domain. Most approaches start by developing an analysis/synthesis technique from which the speech signal is reconstructed with minimum loss of sound quality. Then, the main issues have to be resolved: what kind of representation and which parameters are chosen for the application of the desired speech processing. The question of which speech representation to be used is tightly related to the question of which features and algorithms to be employed for modifications. The challenge of spectral modification is to modify the spectral/acoustical features without degrading the speech quality.

Among other sophisticated representations, the short-term Fourier transform (STFT) and its magnitude are good mathematical tools [31]. To further process the spectral information, we often obtain the spectral envelope from a Fourier magnitude spectrum by successively smoothing its curve to get rid of the rapid fluctuations. In general, the spectral envelope is a smoothed version of the frequency spectrum of a sound, and is often independent of the fundamental frequency. In the literature, there are two popular approaches to spectral modification:

- **Rule-based approach:** This approach works based on a set of rules that have been established by analyzing training data. In this approach, the rules are described by specific clauses, such as IF THEN. However, the manual development of an understanding component is time-consuming, since each application requires an own adaptation.

- **Statistical approach:** This approach works based on mathematical and statistical algorithms. The statistical models are derived from the automatic analyses of large speech databases.

1.4.1 Rule-based approach

The main idea of the rule-based approach is modifying speech signals by applying rules which can be clearly described by using specific clauses, such as IF THEN. This approach includes two main following techniques.

- All-pole-based methods which scale poles by a complex factor to alter formant characteristics.
- Frequency warping methods which modify the spectral envelope directly.

Pole modification

When linear prediction (LP) analysis is used to estimate the spectral envelope, formants are assigned to poles, and can then be modified to correspond with desired formant locations. The algorithm for changing formant frequencies and bandwidths is described as follows.

It has been previously mentioned that the prediction error filter or the LP analysis filter $A(z)$ can be expressed in terms of the LP coefficients a_i in the following form:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}. \tag{1.2}$$

The roots obtained from Eq. (1.2) are:

$$z_i = r_i e^{\pm j\omega_i} \quad (i = 1, 2, ..., p/2) \tag{1.3}$$

Each formant is related to a complex conjugate pair of poles $z_i = r_i e^{\pm j\omega_i}$. Formant frequency, F_i, and formant bandwidth, B_i, can be calculated as follows.

$$F_i = \frac{\omega_i}{2\pi} \; Fs \tag{1.4}$$

$$B_i = -\frac{logr_i}{\pi} \; Fs \tag{1.5}$$

where Fs is the sampling frequency.

Formant modifications can be performed by scaling the angle, ω_i, and magnitude, r_i, of each pair of poles. However, this kind of algorithms has two main problems, i.e. pole interaction, and difficulty in controlling the spectral shape. When two poles are shifted too closely to one another in frequency, only one peak appears in the spectrum. This is a symptom of pole interaction. Based on the fact that the formant energy is more important than the formant bandwidth in speech perception [75], a number of iterative algorithms has been developed to compensate for pole interaction, such as [56, 94]. While these methods can produce spectral envelopes with desired formant amplitudes at the formant frequencies, one drawback of them is that the bandwidth and amplitude of each formant can not be controlled independently. As shown in Eq. (1.5), each formant's bandwidth is dependent on the magnitude of the corresponding pole. Therefore, the amplitude and bandwidth of each formant can not be independently modified with these

procedures. Recently, a method for directly modifying formant locations and bandwidths in the line spectral frequency (LSF) domain has been developed [99]. From now, we refer to the method in [99] as the LSF-based method. By taking advantage of the nearly linear relationship between the LSF coefficients and formants, modifications are performed based on desired shifts in formant frequencies and bandwidths. However, the main drawback, i.e. the lack of control over the spectral shape, has not been solved.

Frequency warping

Frequency warping is a simple method for shifting formants by applying a frequency warping function directly to the spectral envelope. In frequency warping methods, we need to define three parameters, i.e. the lower and upper frequencies, f_L and f_U, and a warping function. The lower and upper frequencies, f_L and f_U, determine the range of the spectral envelope to be affected. The warping function determines how to modify frequencies in the defined range. The warping function can be either a piecewise linear function or a non-linear function.

Shifting the formants by a constant factor may be performed via linear expanding or compressing the speech spectral envelope $X(f,t)$ along the frequency axis. It can be called a linear frequency scale mapping of the speech spectral envelope. It corresponds to:

$$Y(f,t) = X(W(f,t)) \tag{1.6}$$

where $X(f,t)$ and $Y(f,t)$ represent the source spectrum envelope and the transformed spectrum envelope at time t, respectively. The linear frequency scale mapping function $W(f,t)$ is given by:

$$W(f,t) = \frac{f}{\gamma} \tag{1.7}$$

where γ is the constant formant modification factor corresponding to the ratio of the target speaker frequency f_{target} and the source speaker frequency f_{source}:

$$\gamma = \frac{f_{target}}{f_{source}} \tag{1.8}$$

An example of a linear warping function is illustrated in Figure 1.2. In [126], Rentzos et al. proposed a new frequency warping method to modify frequencies, bandwidths and amplitudes of formants. They divided the frequency of aperiodic component into N bands, which is accordance with N formant ranges. The equation for spectrum mapping is expressed as follows.

$$Y(f,t) = \eta(f,t)X[W(f,t), f, t] \tag{1.9}$$

where X, Y and f denote the source spectrum, the transformed spectrum and the frequency variables, respectively. The frequency warping function $W(f,t)$ is the mapping functions for the peak frequency, and $\eta(f,t)$ is the magnitude shaping function used to map the

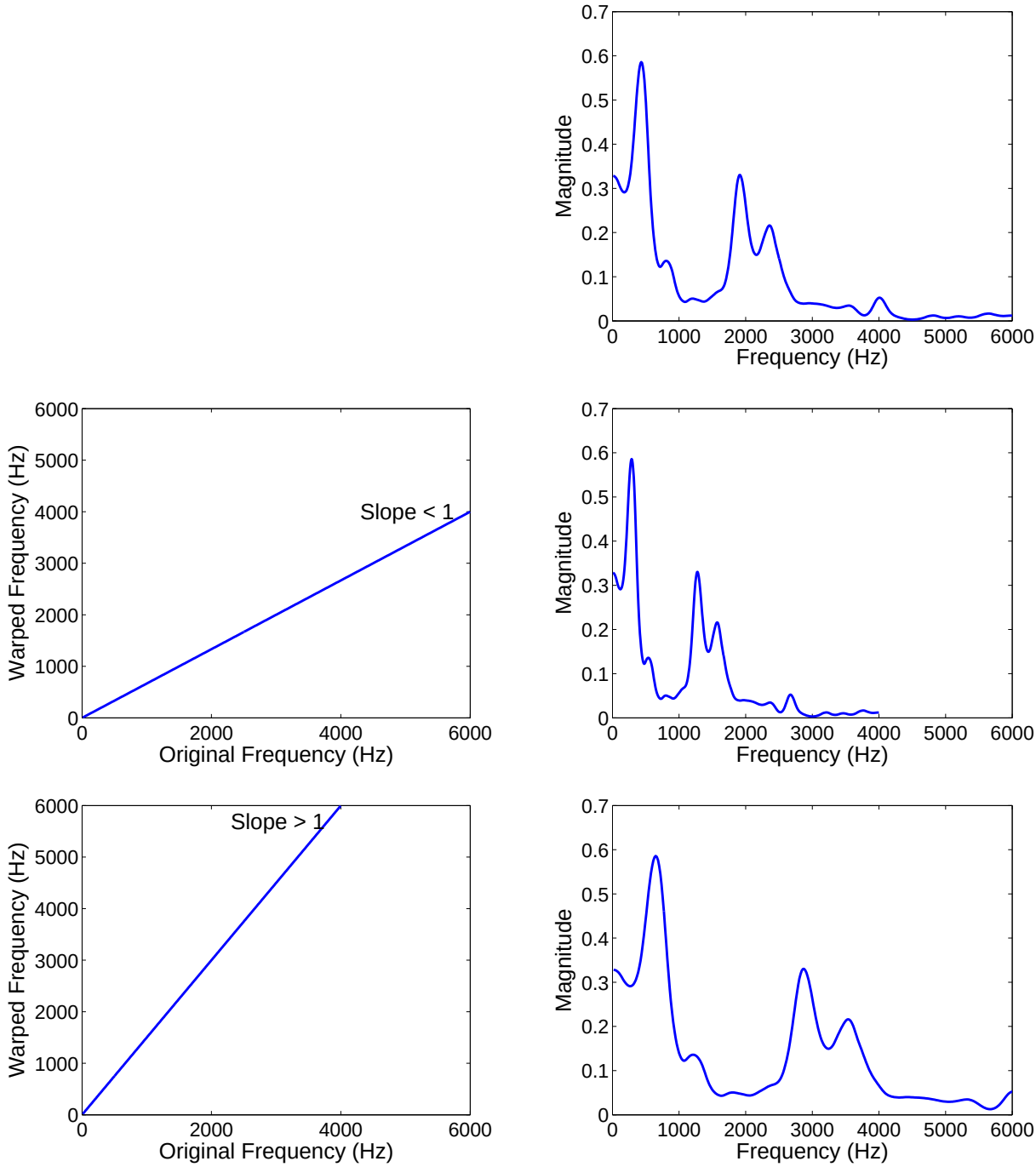

Figure 1.2: Examples of linear warping functions over a speech spectral envelope. The warping functions cause an expansion or compression of the speech spectral envelope.

spectral magnitude between the source and the transformed peaks. However, in this study, formants first have to been estimated, and this process is a difficult task.

The disadvantage of the linear warping method is that for $\gamma > 1$ the frequencies higher than $Fs/2\gamma$ are missed, and for $\gamma < 1$ the frequencies higher than $Fs/2\gamma$ must be padded by arbitrary values (see Figure 1.2). Furthermore, if NF-point fast Fourier transform (FFT) is used for the original spectrum envelope computation, the number of points of the modified speech spectral envelope is lower than NF for $\gamma > 1$, and higher than NF for $\gamma < 1$. Therefore, the transformed spectrum envelope would have to be re-sampled so that its length would be NF.

To overcome this drawback, some non-linear warping methods have been proposed, such as [120]. For example, in [120], one warping function was chosen as a circle arc going through the points $[f_L,\ f_L]$ and $[f_U,\ f_U]$ as follows.

$$W(f) = y \pm \sqrt{r^2 - (f - x)^2} \tag{1.10}$$

where

$$y = \frac{Fs}{8\Gamma}\frac{3\Gamma^2 - 1}{R - 1}$$

$$x = \frac{Fs}{2} - y$$

$$r^2 = x^2 + y^2$$

and Γ is the formant modification factor corresponding to the middle of the modified frequency range, $(f_L + f_U)/2$.

Although frequency warping methods allow a high level of control over formant characteristics, this type of modification still has some limitations.

- Frequency warping methods only perform spectral modification when the original and modified formants are spaced far enough apart so as to be nearly independent of one another. When formants are too close to one another, it is difficult to modify their bandwidths to desirable specifications. This is similar to the pole interaction problem suffered by pole modification techniques.

- The conventional frequency warping methods do not estimate and model spectral peaks. Therefore, they meet difficulties in controlling spectral peaks, such as preserving shapes of peaks, and emphasizing spectral peaks around 3 kHz in transformation of speaking voice into singing voice.

- Frequency warping methods do not allow formants to be merged or split, which is often desired in formant modification processes [78].

1.4.2 Statistical approach

The original purpose of this approach is for speaker adaptation [134] and spectral voice conversion [23, 34, 37, 61, 74, 81, 107, 143, 148, 163]. After that, researchers apply this

approach in other areas of speech technology, such as in noise reduction [101]. To convert the spectral envelopes of the source speaker to the spectral envelopes of the target one, a training (or learning) step is necessary. The training data includes two parts. One part is the analyzed data set of source parameters, the other part is the analyzed data set of target parameters. The source and target training data sets are aligned. Mathematical or statistical algorithms are applied to formulate mapping functions. We utilize the mapping functions to convert features of a speech signal, and synthesize the converted and other features to output the converted speech. This subsection reviews some typical techniques and functions used for spectral transformations.

One of the earliest approaches to the spectral conversion problem is the mapping codebook method proposed by Abe et al. [1]. The basic idea of this technique is to make mapping codebooks which represent the correspondence between the two speakers. A conversion of acoustic features from one speaker to another is therefore reduced to the problem of mapping the codebooks of the two speakers [1]. The procedure for the mapping codebooks construction is given below [1]:

1. The source and target speakers pronounce a learning word set. Then all words are vector quantified frame by frame.

2. Using a time warping technique (DTW), the correspondence between vectors of the same words for the two speakers is determined.

3. The vector correspondences between the speakers are accumulated as histograms. Using each histogram as a weighting function, the mapping codebook is defined as a linear combination of the target speaker's vectors.

4. Steps 2 and 3 are repeated to refine the mapping codebook.

The acoustic space of a speaker is typically modeled by a vector quantization codebook of 32-256 vectors derived using the LBG algorithm [50]. In the conversion phase, the source speaker's input utterance is LPC-analyzed and the spectrum parameters are vector quantized using the source speaker's own codebook. Then, they are decoded using the source-target mapping codebooks. The main shortcoming of the systems based on vector-quantization is that there are discontinuities in the transformation function near the transitions between classes. This limitation leads to degradation of the transformed speech quality.

The appearance of discontinuities in the transformation function near the transitions between classes is solved by dividing the acoustic space into overlapping classes, so that all the input vectors have a certain probability of belonging to each of the acoustic classes. Stylianou et al. [143] propose a Gaussian mixture model (GMM)-based method for voice conversion to solve the discontinuity problem. In [143], a Gaussian mixture model (GMM) is fitted to the training acoustic vectors of the source speaker as follows.

$$p(x) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m, \Sigma_m) \tag{1.11}$$

where M is the number of Gaussian components, $\mathcal{N}(x; \mu_m, \Sigma_m)$ is a Gaussian vector distribution defined by the mean vector μ_m and the covariance matrix Σ_m, and α_m is the weight assigned to the mth Gaussian component. The transformation function is given as follows.

$$F(x) = \sum_{m=1}^{M} p_m(x)[\mathbf{v}_m + \mathbf{\Gamma}_m \Sigma_m^{-1}(x - \mu_m)] \tag{1.12}$$

where $p_m(x)$ denotes the probability that x belongs to the mth Gaussian component. The vectors $\mathbf{v}_m$ and matrices $\mathbf{\Gamma}_m$ are calculated during the training phase. The soft acoustic classification based on GMM avoids the appearance of typical artifacts caused by the discontinuities in the transformation function. This technique has become one of the best statistical methods of spectral voice conversion. However, it still has some drawbacks, such as over-smoothing effect of converted spectrum that degraded the transformed speech quality. Although many studies have improved the GMM-based method, such as [23, 34, 37, 61, 74, 81, 107, 147, 148, 163], the quality of the transformed speech is still far from natural.

1.5 Motivation and scope of the research

Speech is one of the most convenient ways for people to communicate. The advances in technology lead to the rapid growth of communication environments between human beings and machines. The areas of speech processing technology related to human-machine communication are speech recognition and speech synthesis. The former is a technique to extract linguistic content in a spoken utterance. The aim of the latter is to generate an artificial voice from a machine. However, in many cases, the communications between human beings and machines are not friendly, since voice processing techniques related to the problems of man-machine communication do not provide high quality communication.

Our work is restricted to the voice transformation area. We focus on how to improve the quality of transformed speech. Most speech processing applications utilize certain properties of speech signals in accomplishing their tasks. Voice transformation algorithms aim to change some attributes of speech while leaving other attributes unchanged. Voice transformation generally consists of the process of analyzing a speech signal into a number of parameters, modifying these parameter values in accordance with a desired goal and synthesizing the correspondingly modified speech signal. Applications of voice transformation can be founds in many areas. In Text-to-Speech, voice transformation is used to create new speech signals from a pre-recorded speech database. We can combine and smooth speech units to generate a long utterance. Since recording a speech database is time-consuming and expensive, to create other types of speech, e.g. emotional or expressive speech or speech of various speakers, we modify available speech databases. In education, slowing down voices helps learners to discover proper intonation of sentences and pronunciation of words, as well as to practice listening skill. Voice transformation is used to aid for the handicapped people, such as placing speech signals to other frequency ranges to improve the recognition rate for the deaf people. In addition, voice

transformation can be applied in car navigation, voice-enabled e-mail, etc.

In the voice transformation area, a number of key factors determines the success and usefulness of a voice transformation system:

- the quality (including intelligibility and naturalness) of the transformed speech
- the flexibility to generate the speaker's styles (e.g. speaker's emotion, speaker's gender, and speaker's identity) to achieve more flexible speech database

To achieve these goals, we need to solve many questions. For example, how to analyze and synthesize speech signals, which properties of speech should be manipulated, how to modify these properties. Speech modification can be classified into three main groups: time-scale modification, pitch-scale modification, and spectral modification.

In this dissertation, we concentrate on spectral modification, which is one of the core processes in voice transformation. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing. Spectral modification is used to perform a variety of modifications to speech spectra, such as modifications of formant structures, amplitude. The challenge of spectral modification is to modify the spectral/acoustical features without degrading the speech quality.

From a mathematical point of view, we can represent a speech signal in various ways. For example, we can describe the speech signal as a function of time in the time-domain, which shows how signal magnitude changes over time. We can also describe the signal as a functions of frequency in the frequency-domain, which shows how quickly signal frequencies changes. Representation and processing of speech signals in the time-domain and the frequency-domain are important for some aspects of speech applications. However, these representations in these domains do not reflect the most essential property of speech. The reason is when a human produces voices, the continual motion of the articulators forms a time-varying acoustic filter responsible for the generation of the speech waveform. To characterize this type of properties, we need a joint time-frequency representation.

Most methods of spectral modification process speech signals frame by frame, and they rarely consider the relations between neighboring frames. When there are unexpected modifications in some frames, the modified speech may be not smooth. As a result, there are some clicks in the modified speech, which lead to degradation of the speech quality. In addition, they often do not model the temporal evolution of parameters. Time-scale modification is performed by replication/omission of some of the windowed segments [156] or using interpolation functions [65]. Therefore, the natural evolution of speech signals is not guaranteed, which also leads to degradation of the modified speech quality.

To perform spectral modification, among other sophisticated representations, the short-term Fourier transform (STFT) and its magnitude are good mathematical tools [31]. To further process the spectral information, we often obtain the spectral envelope from a Fourier magnitude spectrum by successively smoothing its curve to get rid of the rapid fluctuations, and modify the spectral envelope. The spectral envelope can be represented by non-parametric or parametric methods. Modification of spectral envelope can be carried out by the rule-based approach or the statistical approach.

The rule-based approach works based on a set of rules that have been established by analyzing training data. In this approach, the rules are described by specific clauses, such as IF THEN. For the non-parametric methods, frequency warping techniques, such as [151], are often employed to modify the spectral envelope. Although this kind of methods provides high-quality, the modification is still not successful, because frequency warping methods meet difficulties in modifying spectral peaks, such as preserving shapes of peaks, and emphasizing spectral peaks around 3 kHz in transformation of speaking voice into singing voice, since they do not estimate spectral peaks. Moreover, the frequency warping methods do not allow formants to be merged or split, which is often desired in formant modification processes [78]. High-quality, flexible modifications can be achieved by parametric methods when processing speech. A number of parameterization methods exist for speech spectral modeling, such as LP-based methods [94, 99]. Although these spectral modelings have been successfully used in various applications, such as in speech coding and speech recognition, the main drawback to these types of models, the lack of control over the spectral shape, has not been solved.

The statistical approach works based on mathematical and statistical algorithms. In this approach, a training (or learning) step is necessary to find rules for mapping. The mapping rules are described by mathematical or statistical models. We should add acoustic constraints to improve the quality of modified speech.

Although many methods of spectral modification have been proposed, there are still three main issues:

1. lack of efficient spectral modelings for speech modification,
2. insufficient smoothness of modified spectra between frames, and
3. ineffective spectral modification

This dissertation focus on solving three issues mentioned above. To perform spectral modification, we first develop an analysis/synthesis framework. The analysis/synthesis method is very important, since it decides that which and how features can be modified. Many high-quality analysis/synthesis methods have been proposed. For example, the sinusoidal model [91, 92] and its extensions [42, 43, 44] produce the high quality of synthesized speech. However, a number of parameters is high, and their parameters are not directly related to formants. These limitations prevent the sinusoidal model from modifying the spectral envelope in accordance with scaling factors of formants. Although the STRAIGHT method [65] produces very high quality of synthesized speech and modified speech, it still processes speech signals frame by frame and its spectral information is non-parametric representation. Therefore it is necessary to have a spectral modeling which is effective and flexible to ensure the smoothness of modified speech and perform the efficient spectral modification.

In the second part of this dissertation, we focus on solving the second issue, insufficient smoothness of modified spectra between frames. The discontinuities between frames exist in most applications of speech modification when manipulating speech signals if we do not consider the relation between neighboring frames. It is necessary to have a method for

modeling the temporal relations. In the literature, a hidden Markov model (HMM) is well-known for being a typical model for modeling temporal trajectories of spectral parameters. A HMM can be used to represent a given speech segment in a stochastic manner. It models each of the various quasi-stationary spectral zones characteristic of particular word as a state in a Markov chain with an associated observation probability density function. The HMM model is widely employed in automatic speech recognition [80], speech coding [55], in formant tracking [3], speech synthesis systems [149], etc. However, the quality of the HMM model much depends on training data. The HMM model is not suitable for applications of speech modification when limited training data is available. Therefore, it requires a new method for modeling temporal evolution of speech parameters. We aim to present an efficient model of temporal evolution. After that, we verify this model in concatenative speech synthesis.

Concatenative speech synthesis systems form utterances by concatenating pre-recorded speech units. In a concatenative speech synthesis systems, output speech is limited by the contents of the acoustic inventory (not just the linguistic content, but also the emotional state of the speaker, degree of articulation, etc.), and inevitable concatenation errors can lead to audible discontinuities. To solve the discontinuities between speech units, many methods for reducing the mismatch between speech units have been presented, such as [53, 161, 62]. However, some steps in these methods need to be manually performed, due to preparation of "fusion" units [161], or extraction of formants [62]. Therefore, it is required to have a new method which can automatically smooth the mismatch between these units.

In the third part of this dissertation, we concentrate on solving the third issues, the ineffective spectral modification. We deal with both kinds of spectral modification approaches: rule-base and statistical approaches.

First, our purpose is to develop a new efficient algorithm of spectral modification. One of the most important requirements of spectral modification is that it is flexible enough to perform a variety of modifications within the spectral envelope. Formant frequency is one of the most important parameters in characterizing speech, and it also plays an important role in specifying speaker characteristics. Therefore, alteration of formant frequencies can control other features that are directly connected to the speech production process. Conventional spectral modification methods, such as [94, 99, 151], often control formants to modify the speech spectral envelope. However, these methods are limited by their inability to independently control important formant characteristics such as amplitude and bandwidth, or to control the spectral shape. In our solution, we use spectral-GMM parameters to model the speech spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. We propose a new algorithm for modifying spectral-GMM parameters in accordance with formant frequencies. We then apply our algorithm to two areas, emotional speech synthesis which requires modification of both formant frequency and power, and voice gender conversion which requires a large amount of spectral modification.

Second, we aim to improve the quality of spectral modification in voice conversion

systems. The purpose of the voice conversion is to transform a voice of a speaker (source speaker) so that it is perceived by listeners as if it were uttered by a different specific speaker (target speaker). The limitation in current voice conversion systems is that manipulating the speech signal for converting the source voice into the target voice still degrades its quality. Although there are various speaker-dependent voice characteristics, the conversion of spectral characteristics is a major process in a voice conversion system. Many spectral voice conversion methods have been proposed in the literature, and GMM-based spectral voice conversion methods are regarded as some of the best systems. However, the quality of converted speech is still far from natural. To improve the quality of converted speech, we focus on solving two main problems of the degradation of the quality of converted speech: (i) modeling the distribution of acoustic features in voice conversion often uses unstable frames, which degrades the precision of GMM parameters (ii) the transformation function may generate discontinuous features if frames are processed independently.

In summary, this dissertation focuses on improving the quality of modified speech in the area of voice transformation. We aim to propose new spectral modelings and efficient spectral modification algorithms to solve three main issues of spectral modification in voice transformation, i.e. the lack of efficient spectral modelings for speech modification, the insufficient smoothness of the modified spectra between frames, and the ineffective spectral modification. Four main applications discussed in this dissertation include concatenative speech synthesis, emotional speech synthesis, voice gender conversion, and spectral voice conversion.

1.6 Main contribution of the dissertation

As mentioned earlier, this dissertation focuses on spectral modification of speech in voice transformation. Being motivated by these research objectives, this dissertation has been conducted, and some research results have been revealed. Our methods can solve the three main issues as mentioned above, i.e. the lack of efficient spectral modelings for speech modification, the insufficient smoothness of modified spectra between frames, and the ineffective spectral modification. The major contributions presented in this dissertation can be summarized as follows.

1. Developed two improvements of modeling of the speech spectral envelope using statistical methods (Chapter 3). This kind of representation meets most requirements of a spectrum representation for speech modification. Those are preciseness and stability, which ensure to model and reconstruct the speech spectral envelope precisely, locality (without affecting the intensity of frequencies further away from the point of manipulation), and flexibility and ease of manipulation. Our spectral modelings not only model the speech spectral envelope well but also flexibly modify the speech spectral envelope in both dimensions, frequency and amplitude.

2. Developed a new modeling of speech spectral sequence for speech modification (Chapter 3). Our modeling is based on temporal decomposition and spectral-GMM.

Our modeling is potential to ensure the smoothness of modified speech and perform efficient spectral modification. In addition, our modeling is also potential to control temporal evolution of speech signals. This characteristic helps to improve the speech quality of synthesized speech when performing time-scale modification.

3. Proposed a new method for spectral smoothing in concatenative speech synthesis (Chapter 4). Our method reduces the mismatch of spectral, gain, and F0 information at concatenation points.

4. Developed a new efficient spectral modification algorithm (Chapter 5). Our algorithm performs spectral modification directly on the speech spectral envelope, which is flexible to modify the speech spectral envelope, and does not produce artifacts. We then apply our algorithm to two areas, emotional speech synthesis, which requires modification of both formant frequency and power, and voice gender conversion, which requires a large amount of spectral modification.

5. Proposed a new method for spectral voice conversion using temporal decomposition and Gaussian mixture model (Chapter 6). Our method solved two drawbacks of GMM-based voice conversion methods: insufficient precision of GMM parameters, and insufficient smoothness of the converted spectra between frames.

Although this dissertation only focuses on improving the quality of modified speech in the area of in voice transformation, our work can find other applications in other fields of speech signal processing, such as speech recognition, speech perception, speaker verification and identification. In our work, we model the temporal evolution of spectral parameters. We also directly model the speech spectral envelope. These modelings allow us to effectively and flexibly perform time-scale modification and spectral modification. These two operations are employed in most areas of speech technology.

1.7 Outline of the dissertation

The rest of the dissertation is organized as follows. A schematic overview of this dissertation and the correspondence between our work and the flowchart of voice transformation are shown in Figure 1.3, and Figure 1.4, respectively.

Chapter 2 - Research Background

In this chapter, we introduce the background knowledge for our dissertation. It includes speech production and its modelings, source-filter model, spectral modification algorithms, temporal decomposition, speech spectrum modeling using Gaussian mixture model, and STRAIGHT.

Chapter 3 - Spectral Modelings for Speech Modification

The purpose of this chapter is to deal with the first issue of spectral modification, the lack of efficient spectral modelings for speech modification. This chapter first introduces two

improvements in modeling speech spectral envelope using spectral-GMM parameters [166, 167, 170]. In the first improvement, we not only model the speech spectral envelope well but also ensure a correspondence between spectral peaks and Gaussian components. In the second improvement, we use asymmetric Gaussian mixture model to model the speech spectral envelope, instead of using Gaussian mixture model. This chapter also presents a new method for speech spectral sequence modeling in the time-frequency domain.

Publications related to this chapter are [103, 104, 105, 106, 109].

Chapter 4 - Spectral Smoothing for Concatenative Speech Synthesis based on Temporal Decomposition

The aim of this chapter is to solve the second issue of spectral modification, the insufficient smoothness of modified spectra between frames. To improve the quality of modified speech, one of requirements of spectral modification methods is to ensure the smoothness of modified speech. To overcome the discontinuities of speech after modification, one of efficient ways is to control spectral dynamics. Knagenhjelm and Kleijn [73] point out that spectral dynamics is more important than spectral distortion in human perception. This chapter employs the temporal decomposition technique [5, 113] to control spectral dynamics in concatenative speech synthesis. We address discontinuities (spectral, fundamental frequency, and gain information) in concatenation points in a concatenative speech synthesis system.

Publications related to this chapter are [107].

Chapter 5 - Rule-based Approach to Spectral Modification

The aim of this chapter is to solve the third issue of spectral modification, the ineffective spectral modification. We develop a spectral modification algorithm. Formant frequency is one of the most important parameters in characterizing speech, and control of formants can effectively modify the spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. We solve this problem in this chapter. We evaluate the effectiveness of our proposed method in two areas, emotional speech synthesis which requires modification of both formant frequency and power, and voice gender conversion which requires a large amount of spectral modification.

Publications related to this chapter are [103, 104, 105, 106].

Chapter 6 - Statistical Approach to Spectral Modification

This chapter continues to solve the third issue of spectral modification, ineffective spectral modification. In Chapter 5, we develop a spectral modification algorithm which is applied

for the rule-based approach. In this chapter, we focus on other kind of spectral modification approach, i.e. how to efficiently perform statistical approach to spectral modification. We improve the GMM-based method, one of most successful statistical techniques. In our work, we deal with the two following drawbacks in a GMM-based spectral voice conversion system, insufficient precision of GMM parameters and insufficient smoothness of the converted spectra between frames.

Publications related to this chapter are [107, 108].

Chapter 7 - Summary and Future Work

In this chapter, we summarize the contribution of this dissertation, and gives some future research directions.

1.8 Summary

In this chapter, we have presented the overview of our work. Our motivation and scope have been given. The contributions to the knowledge in the field of spectral analysis and spectral processing have been summarized. The structure of this dissertation has been outlined.

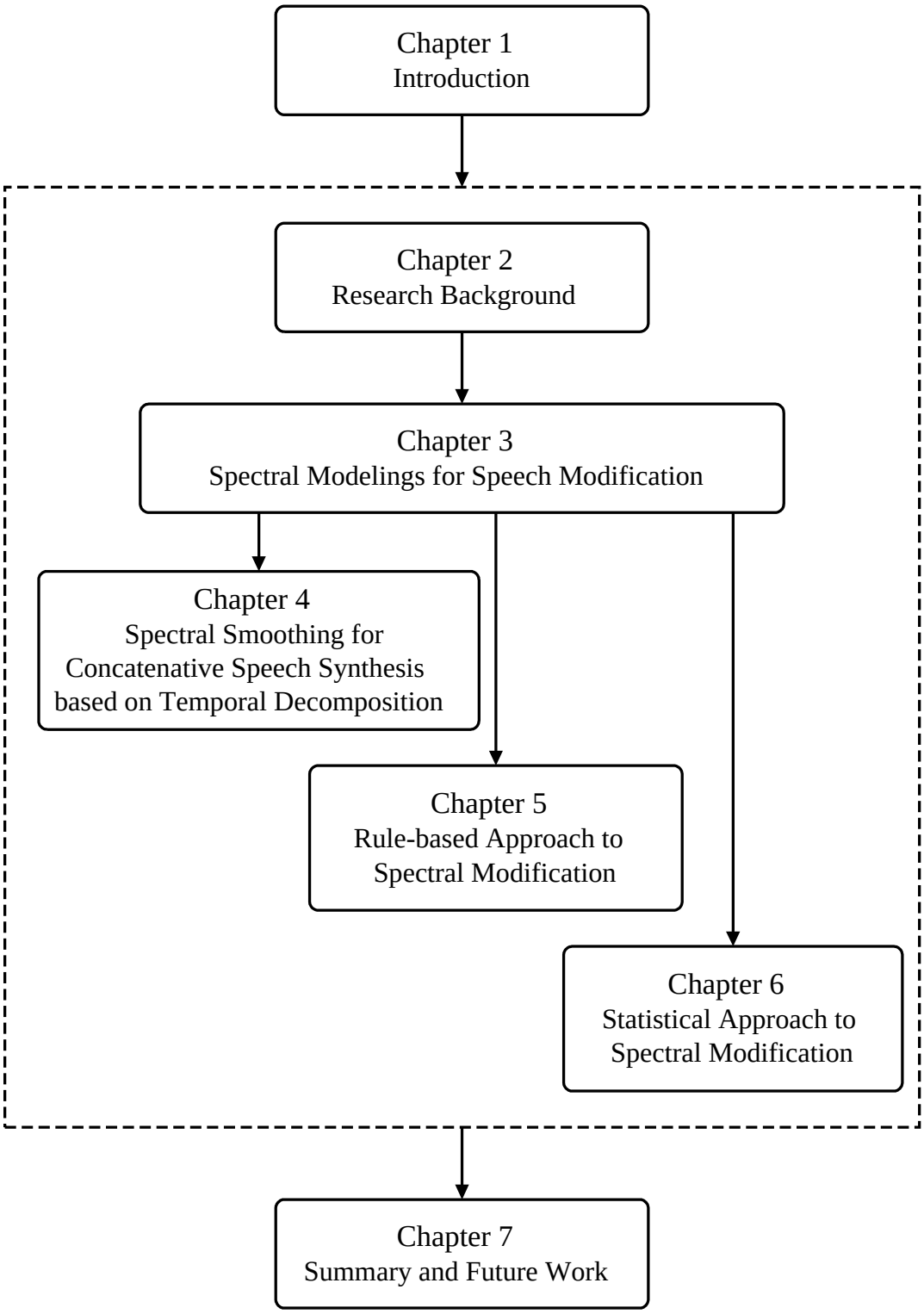

Figure 1.3: Schematic overview of the dissertation.

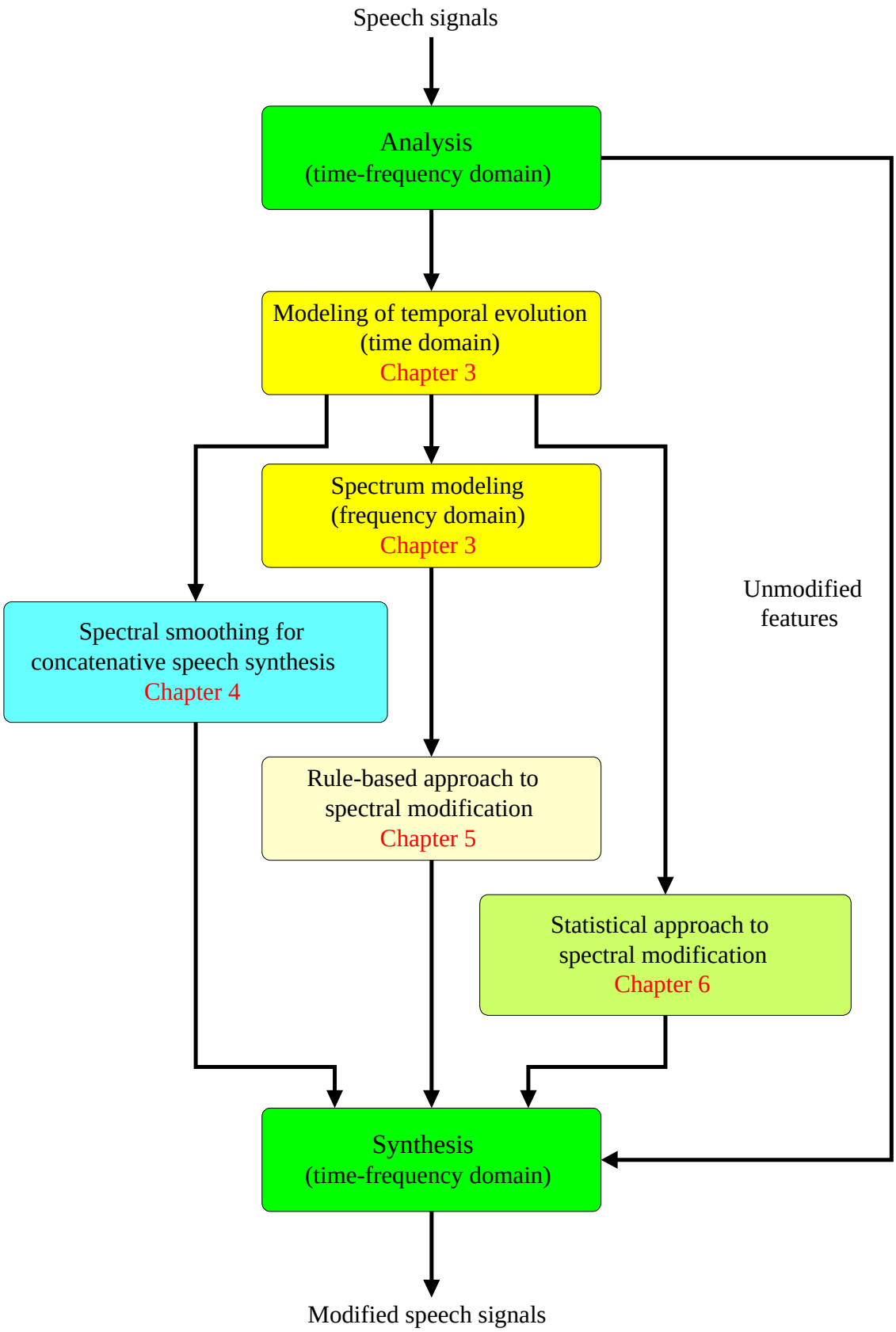

Figure 1.4: Correspondence between our work and the flowchart of voice transformation.

Chapter 2

Research Background

In this chapter, we present the background for our work. In the first part, speech production is outlined. In the next part, source-filter model for speech production is introduced. The overview of the linear predictive analysis and LSF coefficients are presented. Finally, three major methods which are used extensively in the remaining chapters of this dissertation are presented. They are temporal decomposition [5, 113], modeling of speech spectrum using Gaussian mixture model [166, 167, 170], and a speech analysis/modification/synthesis system, STRAIGHT [65].

2.1 Speech production

When processing speech signals for many possible applications, the task is made much easier if one understands how human beings generate speech, and how various voice processes of human beings can be modeled in computer. This section briefly introduces the human speech production.

The main organs of the human body which are responsible for producing speech are the lungs, larynx, pharynx, nose, and mouth. These organs are shown in Figure 2.1. The production of a sound is described as follows.

1. The human speech production starts with the downward movement of the diaphragm to let air flow up to the lungs.

2. Air pressure from the lungs creates a steady flow of air through the trachea, larynx, and pharynx.

3. The vocal folds in the larynx vibrate to create fluctuations in air pressure that are known as sound waves.

4. Resonances in the vocal tract add characteristics to these sound waves according to the positions and the shapes of the lips, jaw, tongue, soft palate, and other speech organs.

5. Openings of mouth and nose radiate the sound waves into the environment.

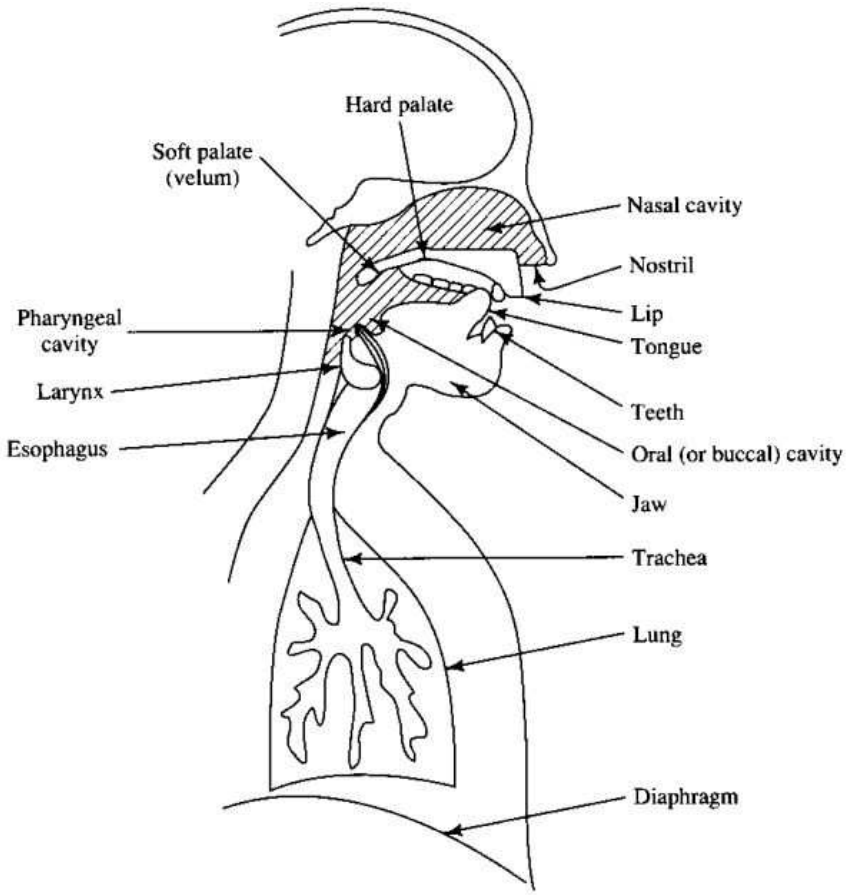

Figure 2.1: The human vocal organs [29]

When we hear somebody making a voiced sound, we hear the fundamental frequency and a series of harmonics. The fundamental frequency is determined by the amount of vibrations of the vocal folds in one second which is measured in Hertz (cycles per second). This fundamental frequency is also known as the lowest resonant frequency of a vibrating object. Speech signals are generally considered as voiced, unvoiced, or in some cases it can manifest itself between these two. Voiced sounds consist of F0's whereas (purely) unvoiced sounds do not consist of any F0's. Therefore, an unvoiced sound can be considered as a white noise. When the vocal folds vibrate, the harmonic components of the F0 are produced. Harmonics or vibrations of the vocal folds at different frequencies give the sound with their particular character. The relation between the harmonic and the F0 is that the harmonic is an integer multiple of the F0. Because the vocal tract modifies the wave signal, formant (pole) and sometimes antiformant (zero) frequencies occur. The most important properties of a formant are its frequency, amplitude and bandwidth. The fundamental frequencies and formant frequencies are considered to be the most important concepts in speech synthesis and speech processing.

In our daily lives, we mainly communicate with other people through voice. We can use our voice perception to understand the information conveyed in the human voice. The human voice does not only contain speech information but also help us in identifying

individuals and perceiving their mental and emotional state. The human voice contains the following information.

Linguistic information

Fujisaki defined the linguistic information as "symbolic information that is represented by a set of discrete symbols and rules for their combination" [41]. The human voice is a medium on which spoken words are carried. It is an acoustic signal which is used by a language to communicate with other people (who at least understand the language).

Non-linguistic information

Speech is a natural way of communication between people. Besides the words that are said, it contains much other information. When listening to a voice, we perceive not only what is said but also how it is said. The way of speaking conveys a lot of information that is automatically processed in our brains to give an overall impression of the message we hear. The information hidden in the words is called non-linguistic information. The non-linguistic information "concerns such factors as the age, gender, idiosyncrasy, and physical and emotional states of the speaker, etc." [41]. The non-linguistic information can be further classified as follows.

- **Identity information**. This information help us to identify who speaking. Each person has own characteristics of articulation, such as the size of the vocal tract and larynx. When people produce speech signals, identity information is contained in speech signals. Listeners can determine the gender, and the age of the speaker from voice. Apart from that, listeners may realize where the speaker comes from, due to the specific pattern of pronunciation which is related to some regional factors (accents).

- **Affective information** This information in the voice is directly related to the feeling, mood and tone of the speaker. Usually it is paired with observable external manifestations. Affective information in the voice occurs when there are changes in the acoustic parameters induced by the autonomic influence and specific patterns of muscular contraction corresponding to various affective states. Affective information is probably the most important study field in emotional speech analysis. Not just because it influences the way we speak but it also dictates the way our speech organs produces the sound. Therefore, it can be implied that there is a strong relation between the human voice and human emotion.

2.2 The source-filter model for speech production

Obviously, there is no single representation and processing system which are optimal for everything. In this section, we give a brief introduction to the source-filter model for speech production, which forms the basis for many speech production models, including the methods used in this research.

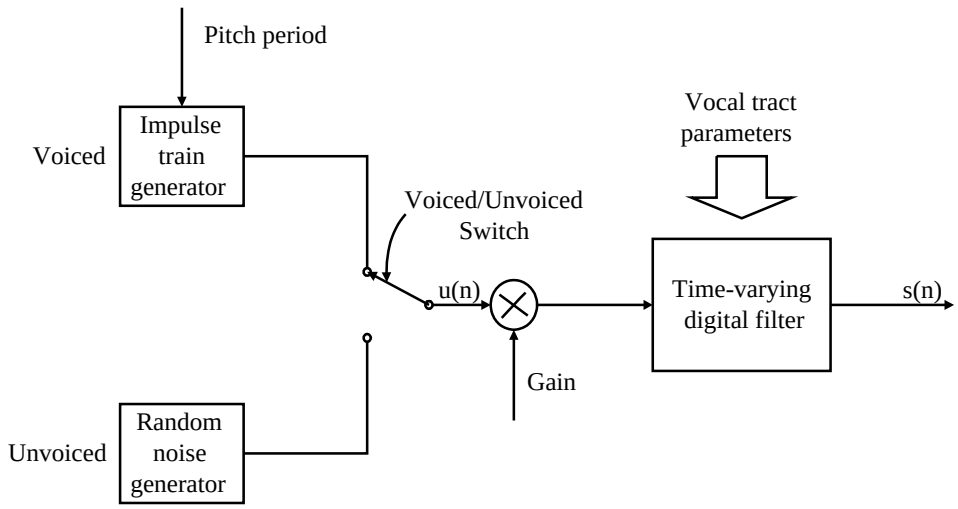

Figure 2.2: Block diagram of simplified model for speech production.

In 1960, Fant [39] introduced the source-filter model for speech production. This model has become a very useful tool for speech analysis and applications such as vocoders, speech synthesis, speech coding, speech modification, speech enhancement. The source-filter theory hypothesizes that an acoustic speech signal can be seen as a source signal (the glottal source, or noise generated at a constriction in the vocal tract), filtered with the resonances in the cavities of the vocal tract downstream from the glottis or the constriction. Therefore, a speech signal is represented as follows.

$$S(z) = E(z)G(z)V(z)R(z) \tag{2.1}$$

where S(z) is the acoustic speech waveform, E(z) is partial realization of a white noise process for an unvoiced sound or a discrete-time impulse train of period for a voiced sound, G(z) is the glottal waveform, V(z) is the vocal tract filter, and R(z) is the radiation impedance.

Fant demonstrated that by modeling the vocal tract as a series of concatenated, lossless acoustic tubes, a linear model for speech production can be derived. As shown in the block diagram in Figure 2.2, the excitation of this filter is either an impulse train for producing voiced speech and zero mean, unit variance, Gaussian noise for unvoiced speech. For voiced speech, the period of the impulse train corresponds to the pitch, T0, of the speaker.

In most of the speech modification algorithms, the radiation part and the vocal tract part are interchanged so the input to the vocal tract transfer function is a differentiated voice source waveform. In some cases, such as in linear prediction analysis, it is convenient to combine the glottal pulse, radiation, and vocal tract components all together and represent them as a single transfer function, $H(z)$, as follows.

$$H(z) = G(z)V(z)R(z) \tag{2.2}$$

It should be noted that the source-filter model has got some limitations, and does not account for all types of speech production. In reality, the vocal tract is not lossless.

Additionally, sounds produced by the nasal tract are not provided in this model. Also, there are certain forms of speech, such as voiced fricatives, which require dual excitation modes. However, this model forms the foundation for later models which have attempted to account for additional factors.

2.3 Linear prediction model

One of the most powerful speech analysis techniques is the linear prediction (LP) model. This model is an early method originally developed for speech coding and compression. Because of the special properties of this method, it can also be used for spectral envelope estimation. LP represents the spectral envelope as an all-pole filter. This representation is based on the concatenated lossless acoustic tube model. In this model, the composite spectral effects of glottal excitation, vocal tract, and lip radiation are represented by a time-varying all-pole filter with the transfer function $H(z)$ of the form.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2.3}$$

where $S(z)$ and $U(z)$ are the z-transforms of output and input signals, respectively. G denotes the gain of the filter. If the order p of the LP filter is high enough to capture the spectral envelope of speech, this all-pole model performs a good reconstruction of speech for all speech sounds when it is excited by an accurate enough input signal (excitation). In the simplest synthesis structure, the filter is excited by an impulse train for voiced speech and by random noise for unvoiced speech. The main advantage of this model is that the filter, a_i, and gain parameter, G, can be estimated in a computationally efficient manner using linear predictive analysis.

The basic idea behind the LP model is that a given speech sample at time n, $s(n)$, can be approximated as a linear combination of the past p speech samples, as follows.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p) = \sum_{i=1}^{p} a_i s(n-i) \tag{2.4}$$

where $a_i (1 \leq i \leq p)$ are assumed constant over the speech analysis frame.

In source-filter model, the speech samples $s(n)$ are related to the excitation $u(n)$ by the following equation.

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{2.5}$$

A linear predictor with prediction coefficients, a_i, is defined as a system whose output is

$$\widehat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{2.6}$$

The prediction error, $e(n)$, is defined as

$$e(n) = s(n) - \widehat{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{2.7}$$

Eq. (2.7) implies that the prediction error is the output of a system with the transfer function given by

$$A(z) = 1 - \sum_{i=1}^{p} a_i s(n-i) \tag{2.8}$$

The prediction error filter, $A(z)$, is an inverse filter for the system, $H(z)$, as follows.

$$H(z) = \frac{G}{A(z)} \tag{2.9}$$

The basic problem of linear prediction analysis is to determine the optimum set of prediction coefficients, a_i, from speech signal, $s(n)$. Because of the time varying characteristics of speech, prediction coefficients are estimated over short-time frames of speech of duration of approximately 20-30 ms. The evaluation of a_i involves the minimization of the prediction error $E(n)$ for a window of speech around the sample of index n.

$$E_n = \sum_m e_n^2(m) \tag{2.10}$$

$$= \sum_m (s_n(m) - \widehat{s}_n(m))^2 \tag{2.11}$$

$$= \sum_m \left(s_n(m) - \sum_{i=1}^{p} a_i s_n(m-i) \right)^2 \tag{2.12}$$

We can find the values of a_i that minimize E_n in Eq. (2.12) by setting $\frac{\partial E_n}{\partial a_j} = 0$ for $j = 1, 2, ..., p$, and we obtain the following equations.

$$\sum_m s_n(m-j)s_n(m) = \sum_{i=1}^{p} a_i \sum_m s_n(m-j)s_n(m-i) \quad 1 \leq j \leq p \tag{2.13}$$

If we define

$$\phi_n(j,i) = \sum_m s_n(m-j)s_n(m-i) \tag{2.14}$$

Eq. (2.13) can be written as follows.

$$\sum_{i=1}^{p} a_i \phi_n(j,i) = \phi_n(j,0) \quad j = 1, 2, ..., p \tag{2.15}$$

The linear system of equations given in Eq. (2.15) can be solved to determine the LP coefficients, a_i. The quantities $\phi(j,i)$ can be calculated either using the autocorrelation method [84, 85, 88] or the covariance method [6].

Line spectral frequency (LSF)

In this subsection, we briefly introduce the line spectral frequency (LSF) coefficients which are discussed more often in this dissertation.

LP coefficients have other representations which are directly derived from the LP coefficients: line spectral frequencies (LSF), reflection coefficients (RC), autocorrelations (AC), log area ratios (LAR), impulse responses of LP synthesis filter (IR), etc. They effectively have an one-to-one relationship with the LP coefficients, and they preserve all the information from the LP coefficients. Among them, some are computationally efficient. Some of them have special features which make them attractive for different purposes. For the purpose of spectral modification, among LP representations, LSF coefficients have some advantages as follows.

- The LSF coefficients have locality property. The locality requirement states that it is possible to achieve a local change of the spectral envelope, i.e. without affecting the intensity of frequencies further away from the point of manipulation. Ideally, the representation would fulfill the requirement of orthogonality, where one component of the spectral envelope can be changed without affecting the others at all. An adverse alteration of one LSF coefficient results in a spectral change only around that frequency [117].

- The LSF coefficients correspond to the bandwidths and approximate locations of the formant frequencies. Hence, we can directly obtain the information about formant locations and bandwidths from the LSF coefficients.

- The LSF's movements in time are more predictable and gradual compared to LP coefficients. Therefore, the LSF coefficients have been found to be suitable as an input of the temporal decomposition algorithm [5], which is a technique we employ in this dissertation.

- Stability check is easy. If the LSF coefficients are in ascending order in the range [0,π], the resulting filter is guaranteed to be stable.

We briefly describe the procedure to calculate LSF coefficients. The more details of this procedure can be referred to [60, 138]. It has been previously mentioned that the prediction error filter or the LP analysis filter $A(z)$ can be expressed in terms of the LP coefficients (direct form predictor coefficients) a_i in the following form:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}. \tag{2.16}$$

The LSF coefficients are calculated using a symmetric and an anti-symmetric polynomial obtained from $A(z)$. The symmetric polynomial, $P(z)$, and the anti-symmetric polynomial, $Q(z)$, are obtained from $A(z)$ as follows.

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{2.17}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{2.18}$$

$P(z)$ and $Q(z)$ satisfy three conditions below [137].

- All the roots of $P(z)$ and $Q(z)$ polynomials lie on the unit circle.
- Roots of $P(z)$ and $Q(z)$ are interlaced.
- The minimum phase property of $A(z)$ can be preserved, if the first two properties are intact after quantization or interpolation.

From the first property, we see that the roots of $P(z)$ and $Q(z)$ can be expressed in terms of w_i (as e^{jw_i}). These w_i are called the LSFs. The LSF parameters may be calculated from Eqs. (2.17) and (2.18) using several methods. Soong and Juang [138] computed the LSF parameters by applying a discrete cosine transformation, and Kabal and Ramachandran [60] used Chebyshev polynomials.

2.4 Temporal decomposition

As mentioned above, one of issues of spectral modification methods is that there are discontinuities when unexpected modifications happen. To solve this issues, in our dissertation, we employ the temporal decomposition (TD) [5] as an analysis/synthesis framework. This section introduces the temporal decomposition technique, and the modified restricted temporal decomposition (MRTD), which is one of improvements of the TD algorithm, and is employed in this dissertation.

2.4.1 Introduction

In 1983, a new technique for efficient coding of linear predictive coding (LPC) parameters, i.e. spectral parameters, was introduced by Atal [5]. Considering that speech events do not occur at uniformly spaced time intervals and that articulatory movements are sometimes fast, sometimes slow, he concluded that uniform time sampling of speech parameters is not efficient. Thus, he proposed the temporal decomposition (TD) to represent the continuous variation of these parameters as a linear-weighted sum of a number of discrete elementary components. In other words, the observed spectral parameter vectors are approximated by a linear combination of a number of vectors of the same dimension called event targets. The interpolation functions used in this approximation are later referred to as event functions. The computational procedure of Atal's method includes three main steps as follows.

1. Detection of event locations by expressing the event functions as a linear combination of orthogonal functions using singular value decomposition (SVD) of the spectral parameter matrix.
2. Computation of event functions by basing on the minimization of a compactness measure of event functions.
3. Calculation of event targets by minimizing the mean squared error between the original and reconstructed spectral parameters.

Although efficient speech coding was his primary objective, the concept of temporal decomposition of speech has attracted many researchers in different application areas, such as speech coding or voice storage [5, 7, 8, 9, 10, 24, 25, 45, 47, 48, 69, 70, 79, 102, 112, 113, 127], speech recognition [68, 89, 115, 153, 154, 155], speech segmentation [11], and speech synthesis [4, 18, 19], speaker identification [111], voice morphing [135], speech modification [103, 105, 106, 107, 108, 109].

2.4.2 Atal's method of temporal decomposition

Temporal decomposition of speech [5] was first used as a method for efficient coding of LPC parameters. Suppose that a given speech utterance has been produced by a sequence of K movements aimed at realizing K acoustic targets. Let us denote the speech parameters corresponding to the kth target by $\mathbf{a}_k$, and the temporal evolution of this event by a function, $\phi_k(n)$. The frame number n varies between 1 and N, the number of frames in the speech segment. In temporal decomposition of speech, the observed speech parameters, $\mathbf{y}(n)$, are approximated by $\hat{\mathbf{y}}(n)$, a linear combination of event targets as follows:

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.19}$$

where

$$\begin{aligned} \mathbf{a}_k &= [a_{1k}\ a_{2k}\ \cdots\ a_{Pk}]^T \\ \mathbf{y}(n) &= [y_1(n)\ y_2(n)\ \cdots\ y_P(n)]^T \\ \hat{\mathbf{y}}(n) &= [\hat{y}_1(n)\ \hat{y}_2(n)\ \cdots\ \hat{y}_P(n)]^T \end{aligned}$$

The superscript T on a vector or matrix means its transpose. In matrix notation, Eq. (2.19) can be written as

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi} \tag{2.20}$$

where P is the dimension of the spectral parameters, $\hat{\mathbf{Y}}$ is a $P \times N$ matrix whose n^{th} column is $\hat{\mathbf{y}}(n)$, $\mathbf{A}$ is a $P \times K$ matrix whose k^{th} column is $\mathbf{a}_k$, and $\mathbf{\Phi}$ is a $K \times N$ matrix whose k^{th} row is $\boldsymbol{\phi}_k$.

In Eqs. (2.19) and (2.20), both the event targets $\mathbf{A}$ and event functions $\mathbf{\Phi}$ are unknown, only the speech parameter sequence of a given utterance $\mathbf{Y}$ is known. To find $\mathbf{A}$ and $\mathbf{\Phi}$, $\mathbf{Y}$ is to be decomposed through orthogonalization [5, 155]. The analysis procedure in the original TD method [5] is described as follows. The more details of this procedure can be referred to [5].

First, the spectral parameter matrix of a windowed speech segment of about 200-300 ms is decomposed into two orthogonal matrices and a diagonal matrix of eigenvalues, using the so-called singular value decomposition.

$$Y^T = UDV^T$$

where Y^T is the $N \times P$ spectral parameter matrix, U is a $N \times P$ orthogonal matrix, V is a $P \times P$ orthogonal matrix, and D is a diagonal matrix of eigenvalues. N is the number of frames in the windowed speech segment, and P is the order of the spectral parameters. This allows the event functions to be represented as a linear combination of a set of orthogonal functions, and also allows the number of events, M, to be fixed in the windowed speech segment under analysis, by taking into account only the number of significant eigenvalues. A window of about 200-300 ms often gives $M = 5$.

$$\phi_k(n) = \sum_{i=1}^{M} b_{ki} u_i(n)$$

where $u_i(n)$ is the element (n, i) of the matrix U and b_{ki} are a set of coefficients.

Next, the nearest event function, $\phi(n)$, to the center of the windowed speech segment, $n = n_c$, is evaluated by considering the minimization of a distance measure, $\theta(n_c)$.

$$\theta(n_c) = \sqrt{\frac{\sum_{n=1}^{N}(n - n_c)^2 \phi^2(n)}{\sum_{n=1}^{N} \phi^2(n)}}$$

Minimization of $\ln(\theta(n_c))$, with respect to the coefficients b_i leads to an eigenvector problem of a matrix $R \in R^{K \times K}$.

$$R\mathbf{b} = \lambda \mathbf{b}$$

where the element (i, r) of the matrix R is given by

$$R_{ir} = \sum_{n=1}^{N} (n - n_c)^2 u_i(n) u_r(n),$$

and $\mathbf{b}$ is the vector of coefficients b_i. The solution corresponding to the smallest eigenvalue λ provides the optimum $\mathbf{b}$.

To analyze a complete utterance the above procedure should be repeated with windows located at intervals through out the utterance. In Atal's method, to ensure that no event function is missed, the window is required to be shifted by a small interval, i.e. by a frame interval. Therefore, if the total number of windows is L, SVD and eigenvector solving should be performed L times. SVD is a highly involved computational procedure and this is known to be the major reason for the high computational complexity of the Atal's method.

Since the window is shifted by a small interval at each time, the same event function is generally found for several adjacent windows. To find the locations of the event functions, and to reduce the total set of event functions, a reduction algorithm based on a zero crossing criterion of a timing function, $\nu(l)$, is incorporated.

$$\nu(l) = \frac{\sum_{n=1}^{N}(n - l)\phi^2(n)}{\sum_{n=1}^{N} \phi^2(n)}$$

The function $\nu(l)$ crosses the $\nu(l) = 0$ axis from positive to negative at each location

l which equals the location of one of the $\phi_k(n)$ for some k.

By considering the minimization of the squared error between reconstructed and original spectral parameters, E_i, with respect to a_{ik}'s, the spectral targets, $\mathbf{a}_k$, are determined in the following expression.

$$E_i = \sum_{n=1}^{N} \left(y_i(n) - \sum_{k=1}^{K} a_{ik}\phi_k(n) \right)^2, \quad 1 \leq i \leq P$$

where N and K are the total number of frames and events in the entire utterance, respectively. Finally, an iterative refinement procedure is used to improve the event function shapes and to reduce the reconstruction error. The refined set of event functions are evaluated by minimizing the reconstruction error, E_n, of spectral vectors.

$$E_n = \sum_{i=1}^{P} \left(y_i(n) - \sum_{k=1}^{K} a_{ik}\phi_k(n) \right)^2, \quad 1 \leq n \leq N$$

The resultant $\phi_k(n)$'s are used to obtain an even better estimates of the targets, $\mathbf{a}_k$'s. The procedure is repeated until both $\phi_k(n)$'s and $\mathbf{a}_k$'s converge to a set of stable values.

2.4.3 Modified restricted temporal decomposition (MRTD)

Although the original implementation of temporal decomposition of speech [5] was mathematically solid, it is known to have the following two major drawbacks: (i) The method is computationally costly, making it impractical; and (ii) High parameter sensitivity of the number and locations of the events. In other words, they are very sensitive to some trivial changes in the analysis parameters [87, 155].

In the literature, the TD algorithm has been attractive to many researchers, and numerous improvements of TD have been proposed. Studies focus on solving two main drawbacks, high computational costs, and high parameter sensitivity to the number and locations of events. To alleviate the complexity problem with the original method, a number of modifications has been proposed in the literature [4, 9, 11, 24, 32, 45, 46, 69, 102, 112, 113, 115, 136, 155]. To alleviate the sensitivity problem with Atal's method, some solutions have also been proposed, such as [102, 112, 113, 155].

In this dissertation, we employ the modified restricted temporal decomposition (MRTD) algorithm [113]. The reasons for using the MRTD algorithm in this work are twofold: (i) the MRTD algorithm enforces a new property on event functions, named the "well-shapedness" property, to model the temporal structure of speech more effectively [113]; (ii) event targets can convey the speaker's identity [111].

In this subsection, we briefly describe the MRTD. The details of MRTD algorithm can be referred to [110]. The block diagram of the MRTD algorithm is given in Figure 2.3. In the MRTD algorithm, some following constraints are applied.

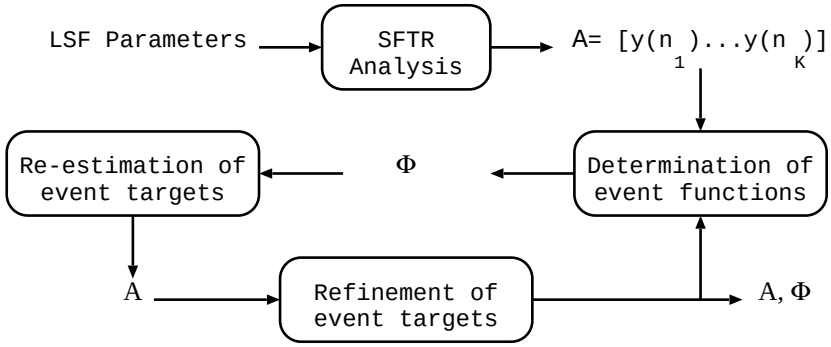

Figure 2.3: Block diagram of the MRTD algorithm [110].

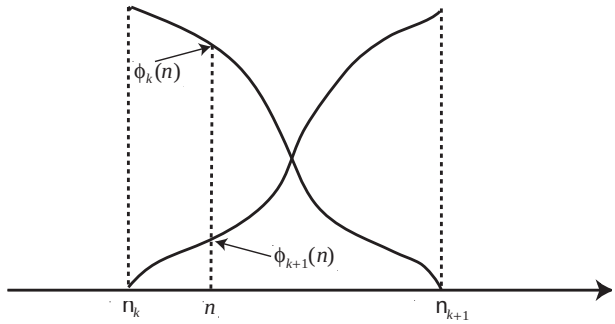

Figure 2.4: Example of two adjacent event functions in the second order TD model.

Restricted second order TD model

Assume that the co-articulation in speech production described by the TD model in terms of overlapping event functions is limited to adjacent events, the second order TD model [9, 115, 136], where only two adjacent event functions can overlap as depicted in Figure 2.4, is given by Eq. (2).

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k \phi_k(n) + \mathbf{a}_{k+1} \phi_{k+1}(n), \;\; n_k \leq n < n_{k+1} \tag{2.21}$$

where n_k and n_{k+1} are the locations of event k and event $(k+1)$, respectively.

The so-called restricted second order TD model was utilized in [32, 69] and this work with an additional restriction to the event functions in the second order TD model that all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in [32]. Eq. (2) can be rewritten as follows.

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k \phi_k(n) + \mathbf{a}_{k+1} \left(1 - \phi_k(n)\right), \;\; n_k \leq n < n_{k+1} \tag{2.22}$$

Determination of event functions

Assume that the locations n_k and n_{k+1} of two consecutive events are known. Then, the right half of the kth event function and the left half of the $(k+1)^{th}$ event function can be optimally evaluated by using $\mathbf{a}_k = \mathbf{y}(n_k)$ and $\mathbf{a}_{k+1} = \mathbf{y}(n_{k+1})$. The reconstruction error,

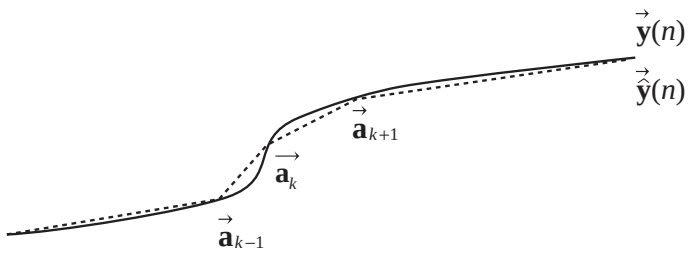

Figure 2.5: The path in parameter space described by the sequence of spectral parameters $\mathbf{y}(n)$ is approximated by means of straight line segments between breakpoints.

$E(n)$, for the nth spectral parameter vector is

$$\begin{aligned} E(n) &= \| \mathbf{y}(n) - \hat{\mathbf{y}}(n) \|^2 \\ &= \| (\mathbf{y}(n) - \mathbf{a}_{k+1}) - (\mathbf{a}_k - \mathbf{a}_{k+1})\phi_k(n) \|^2 \end{aligned} \tag{2.23}$$

where, $n_k \leq n < n_{k+1}$. Therefore, $\phi_k(n)$ should be determined so that $E(n)$ is minimized.

Geometric interpretation of TD

TD yields an approximation of a sequence of spectral parameters by a linear combination of event vectors. Since TD's underlying distance metric is Euclidean, a natural requirement is to have this approximation be invariant with respect to a translation or rotation of the spectral parameters. Dix and Bloothooft [32] considered the geometric interpretation of TD results and found that TD is rotation and scale invariant, but it is not translation invariant.

To overcome this shortcoming and describe TD as a breakpoint analysis procedure in a multidimensional vector space, where breakpoints are connected by straight line segments, Dix and Bloothooft [32] enforced two constraints on the event functions: (i) at any moment of time only two event functions, which are adjacent in time, are non-zero; and (ii) all event functions at any time sum up to one. In other words, the restricted second order TD model was utilized in [32]. These constraints are needed to approximate the path in parameter space by means of straight line segments between breakpoints (see Figure 2.5).

Geometrically speaking, the two event vectors $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ define a plane in P-dimensional vector space. The determination of event functions $\phi_k(n)$ and $\phi_{k+1}(n)$ in the interval $[n_k, n_{k+1}]$ is now depicted in Figure 2.6(a) as the projection of vector $y(n)$ onto this plane. Clearly the following holds: $\phi_k(n_k) = 1$, $\phi_k(n_{k+1}) = 0$, and $0 \leq \phi_k(n) \leq 1$ for $n_k \leq n \leq n_{k+1}$.

While n ranges from n_k to n_{k+1}, the movement of vector $\mathbf{y}(\mathbf{n})$ is described by the transition of $\hat{\mathbf{y}}(n)$ along the straight line segment connecting two breakpoints $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$. As time is moving forward, the transition of $\hat{\mathbf{y}}(n)$ should be monotonic.

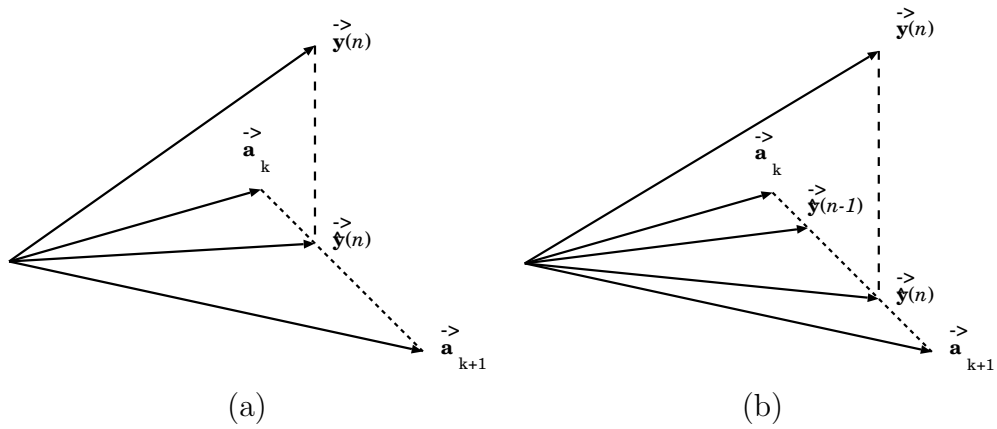

Figure 2.6: Determination of the event functions in the transition interval $[n_k, n_{k+1}]$. The point of the line segment between $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ (a), between $\hat{\mathbf{y}}(n-1)$ and $\mathbf{a}_{k+1}$ (b) with minimum distance from $\mathbf{y}(n)$ is taken as the best approximation.

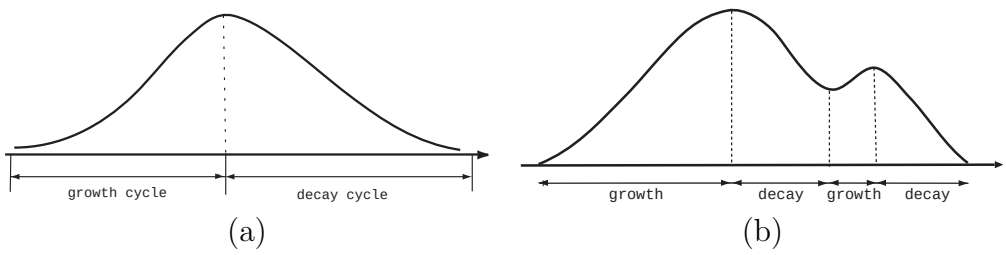

Figure 2.7: Examples of a well-shaped event function (a) and an ill-shaped event function (b).

New determination of event functions

The TD model is based on the hypothesis of articulatory movements towards and away from targets. An appealing result of the above properties of event functions is that one can interpret the values $\phi_k(n)$ as a kind of activation values of the corresponding event. During the transition from one event towards the next the activation value of the left event decreases from one to zero, whilst the right event increases its activation value from zero to the value of one. As mentioned earlier, to model the temporal structure of speech more effectively no backwards transitions are allowed. Therefore, each event function should have a growth cycle; during which the event function grows from zero to one and a decay cycle; during which the event function decays from one to zero. In other words, each event function should have only one peak, which is called the well-shapedness property. On the contrary, an ill-shaped event function can be viewed as an event function which has several growth and decay cycles, i.e. having more than one peak.

Figure 2.7 shows examples of well-shaped and ill-shaped event functions. It can be seen that well-shaped event functions are desirable from speech coding point of view because the well-shapedness property helps reduce the quantization error of event functions when vector quantized.

However, the determination of event functions in [32] has not guaranteed the well-

shapedness property for them since their changes during the transition from one event towards the next may not be monotonic, which results in ill-shaped event functions. In particular, one may wonder that if an event function has some values of one interlaced by other values, causing the next event function to have more than one lobe, which is not acceptable in the conventional TD method. Ill-shaped event functions are also undesirable from speech coding point of view. They increase the quantization error when vector quantized because the uncharacteristic valleys and secondary peaks are not normally captured by the codebook functions. This is because an event function is quantized by its length and shape in the interval between its and the next event function's locations. In that interval, a well-shaped event function is always a decreasing function while an ill-shaped event function is always non-monotonic.

Taking into account the above considerations, we have modified the determination of event functions corresponding to the point of the line segment between $\hat{\mathbf{y}}(n-1)$ and $\mathbf{a}_{k+1}$ (see Figure 2.6(b)) instead of $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ as considered in [32], with minimum distance from $\mathbf{y}(n)$. In mathematical form, the above determination of event functions can be written as

$$\phi_k(n) = \begin{cases} 1 - \phi_{k-1}(n), & \text{if } n_{k-1} < n < n_k \\ 1, & \text{if } n = n_k \\ \min(\phi_k(n-1), \max(0, \hat{\phi}_k(n))), & \text{if } n_k < n < n_{k+1} \\ 0, & \text{otherwise} \end{cases} \tag{2.24}$$

where

$$\hat{\phi}_k(n) = \frac{\langle (\mathbf{y}(n) - \mathbf{a}_{k+1}), (\mathbf{a}_k - \mathbf{a}_{k+1}) \rangle}{\| \mathbf{a}_k - \mathbf{a}_{k+1} \|^2} \tag{2.25}$$

Here, $< ., . >$ and $\| . \|$ denote the inner product of two vectors and the norm of a vector, respectively.

This modification ensures that the value of event function ϕ_k at n is always not greater than the value of event function ϕ_k at $n-1$ in the interval $[n_k; n_{k+1}]$ (see the third line of Eq. (5)), and thereby the well-shapedness property is guaranteed. It should be noted that in [32], $\phi_k(n)$ is determined as $\min(1, \max(0, \hat{\phi}_k(n))$, if $n_k < n < n_{k+1}$.

Identification of event location

In the MRTD algorithm, the local minima of the spectral feature transition rate (SFTR) based on LSF parameters is used as the initial locations of events [69, 102]. SFTR is calculated as follows.

$$\text{SFTR}: \quad s(n) = \sum_{i=1}^{P} c_i(n)^2, \quad 1 \le n \le N \tag{2.26}$$

where

$$c_i(n) = \frac{\sum_{m=-M}^{M} m \mathbf{y}_i(n+m)}{\sum_{m=-M}^{M} m^2}, \quad 1 \le i \le P \tag{2.27}$$

The window size, $2M$, of SFTR analysis is the only parameter that effects the initial

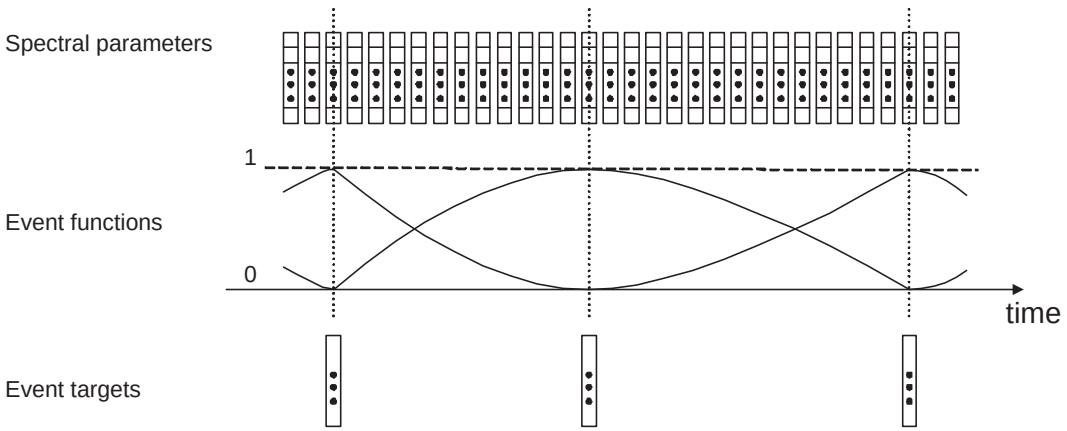

Figure 2.8: Example of event targets and event functions which are decomposed from the MRTD algorithm.

number and locations of events.

An example of event targets and event functions which are decomposed from the MRTD algorithm is shown in Figure 2.8.

MRTD of excitation parameters

In the MRTD algorithm, the same event functions evaluated for the spectral parameters are also used to describe the temporal pattern of the excitation parameters. Let $b(n)$ be an excitation parameter, i.e. F0, gain, and aperiodic component (AP), in the nth frame. $b(n)$ can be approximated by using event target of excitation $\mathbf{b}_k$ and its event function $\phi_k(n)$, as follows.

$$\hat{\mathbf{b}}(n) = \sum_{k=1}^{K} \mathbf{b}_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.28}$$

where $\phi_k(n)$ is estimated from Eq. (2.19).

2.5 Speech spectrum modeling using Gaussian mixture model

2.5.1 Introduction

To perform spectral modification, we first represent the speech spectrum by using parametric or non-parametric methods. For the parametric representations, such as LP coefficients, spectral modification is done by changing LP coefficients [56, 94, 99]. For the non-parametric representations, spectral modification is performed by applying directly frequency warping techniques to the spectral envelope [120, 151].

Parametric methods for spectral modification, such as representing the spectral envelope by using LP coefficients, is flexible. They change properties of formants to perform

spectral modification. However, there is a number of problems associated with the use of formants as features. For example, formants are often poorly defined in certain types of phonemes, such as fricatives or nasalized consonants. LP-based methods do not extract any amplitude information of formants from the speech signal, since it is difficult to control formant amplitudes when manipulating the formant frequencies. LP-based methods also meet a pole interaction. Moreover, the accuracy of the LP-based methods can hardly be high because it is not always clear to determine whether a root obtained forms a formant or just shapes the spectrum [29]. These drawbacks prevent performing high-quality spectral modification from using representations of LP coefficients.

Although frequency warping functions, which perform spectral modification directly on the speech spectral envelope [120, 151], give high quality, they are non-parametric methods. Therefore, it is difficult to emphasize on specific peaks, such as converting a speaking voice into a singing voice, or merging or splitting spectral peaks.

In 1996, Zolfaghari and Robinson [167] proposed a new statistical method for estimating parameters of Gaussian mixture model (called spectral-GMM parameters in this dissertation) from the speech spectrum. Originally proposed as a parametric method for representing the speech spectrum, it was later developed as a low-bit speech codec [167, 168]. The technique assumes that a set of Gaussian components can represent a distribution based on the spectral envelope. The GMM parameters are iteratively estimated using the expectation maximization (EM) algorithm [30]. The spectral-GMM parameters have flexibility to model and control the speech spectrum.

In following subsections, we introduce the technique for modeling a speech spectrum using Gaussian mixture model [140, 164, 167] and related issues. The more details of this technique can be found in [140, 164].

2.5.2 Estimation of spectral-GMM parameters

The main purpose of this subsection is to find an optimal parameter set of Gaussian mixture model to fit to a speech spectrum. The parameters of a probability density function are the number of Gaussian components M, the mean μ_m, the standard deviation σ_m, and the weighting factors α_m. The flowchart to estimate spectral-GMM parameters from a speech signal is shown in Figure 2.9, and the procedure of estimation of spectral-GMM parameters is described as follows.

In a single frame, the spectrum $X(e^{j\omega_n})$ is viewed as a probability distribution $P(x_k)$, where x_k are the bin numbers ($k = \{1, ..., N\}$) and 2N is the FFT size. $P(x_k)$ is simply a normalized spectral density which is to be approximated by a Gaussian mixture model. Figure 2.10 illustrates $P(x_k)$ for a single frame.

Let x_k be the observed incomplete data and (x_k, y_k) be the complete data, where y_k is an unobservable integer between 1 and M indicating the number of Gaussian components. We optimize the expected complete data log likelihood instead of the actual complete data log likelihood, since we can not observe the values of variables y_k. Auxiliary function Q is the expected complete data log likelihood for multiple observed data $X = \{x_1, ..., x_N\}$ and multiple unobserved data $Y = \{y_1, ..., y_N\}$. We assume that a parametric family of mixture probability density functions is given, and a particular $\overline{\Phi}$ is the parameter value

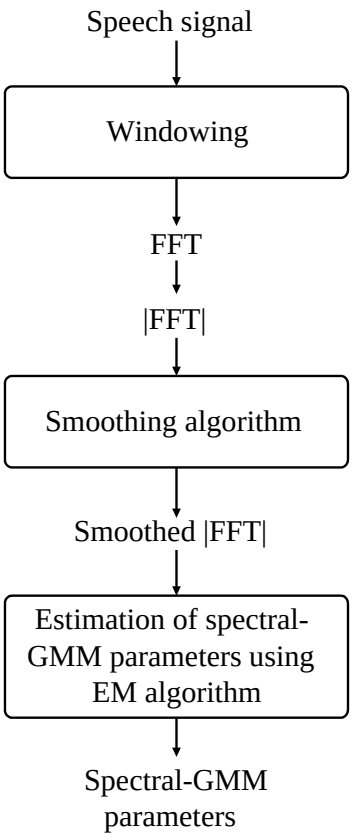

Figure 2.9: Block diagram of estimation of spectral-GMM parameters from speech signals

to be estimated.

In this model, each bin x_k is weighted by its corresponding intensity $P(x_k)$, and we can write the incomplete data log-likelihood as

$$\begin{aligned} \mathcal{L}(X,\Phi) &= log\left[\prod_{k=1}^{N} f(x_k|\Phi)^{P(x_k)}\right] \\ &= \sum_{k=1}^{N} P(x_k) log f(x_k|\Phi) \end{aligned} \tag{2.29}$$

The log-likelihood of one complete data point (x_k, y_k) is obtained as

$$f(x_k, y_k|\Phi) = \alpha_{y_k} f(x_k|y_k, \phi_{y_k}) \tag{2.30}$$

where α_{y_k} is the a priori probability (the mixture weight). A log-likelihood of one incomplete data point x_k is

$$f(x_k, \Phi) = \sum_{y_k} f(x_k, y_k|\Phi) \tag{2.31}$$

and the posterior probability is formalized by

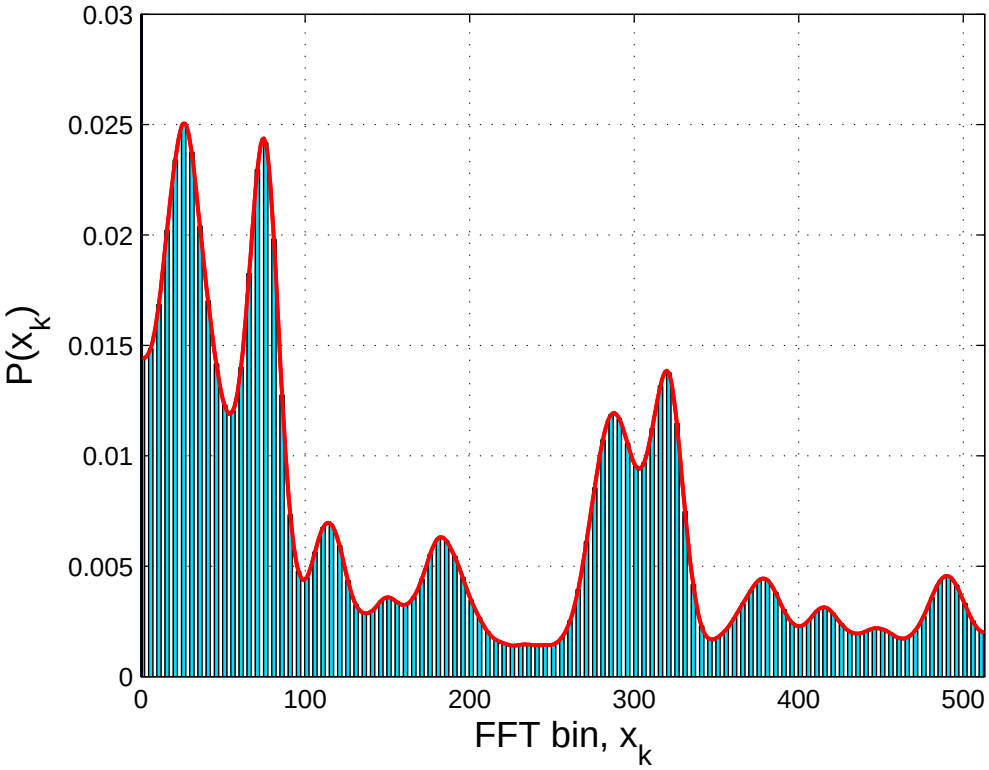

Figure 2.10: Probability distribution $P(x_k)$.

$$P(y_k|x_k,\Phi) = \frac{\alpha_{y_k} f(x_k|y_k,\phi_{y_k})}{\sum_{y_k} \alpha_{y_k} f(x_k|y_k,\phi_{y_k})} \tag{2.32}$$

Hence the Q-function can be represented as

$$Q(\Phi,\overline{\Phi}) = \sum_{k=1}^{N} P(x_k)\Big\{\sum_{y_k} P(y_k|x_k,\phi_{y_k}) log[\overline{\alpha}_{y_k} f(x_k|y_k,\overline{\phi}_{y_k}]\Big\} \tag{2.33}$$

Since the inner summation is over all y_k and $y_k \in 1...M$ for each k, we can denote y_k by i, that is y_k = i if the kth sample was generated by the ith mixture. Splitting the two log-terms, the Q-function can be written as.

$$\begin{aligned} Q(\Phi,\overline{\Phi}) = \sum_{i=1}^{M}\Big\{\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)\Big\} log\overline{\alpha}_i + \\ + \sum_{i=1}^{M}\Big\{\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i) log f(x_k|i,\overline{\phi}_i)\Big\} \end{aligned} \tag{2.34}$$

The Gaussian distribution is used as the density, and is described as follows.

$$f(x_k|i,\phi_i) = \mathcal{N}(x_k,\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\ exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\}$$

where μ_i is the mean, and σ_i is the standard deviation. Note also that the mixture weights of the mixture model satisfy.

$$\sum_{i=1}^{M}\alpha_i = 1; \qquad \alpha_i \geq 0 \tag{2.35}$$

Maximization of the Q-function is achieved by maximizing each term in Eq. (2.34) with respect to $\overline{\alpha}_i$ and $\overline{\phi}_i$. The following equation is obtained from the second term of Eq. (2.34)

$$\begin{aligned}\frac{\partial Q_1(\Phi,\overline{\Phi})}{\partial\overline{\phi}_i} &= \sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)\frac{\partial}{\partial\overline{\phi}_i}[logf(x_k|i,\overline{\phi}_i] \\ &= 0 \end{aligned} \tag{2.36}$$

Differentiating for $\overline{\mu}_i$, and $\overline{\sigma}_i^2$ in Eq. (2.36), and setting it equal to zero, we obtain new parameter estimates $\overline{\mu}_i$, and $\overline{\sigma}_i^2$.

$$\overline{\mu}_i = \frac{\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)x_k}{\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)} \tag{2.37}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)(x_k-\overline{\mu}_i)^2}{\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)} \tag{2.38}$$

From the first term of Eq. (2.34), the mixture weights are formulated as follows

$$\overline{\alpha}_i = \frac{1}{N}\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i) \tag{2.39}$$

Using the EM algorithm, we can estimate the spectral-GMM parameters to model the shape and the peaks of the speech spectrum.

2.5.3 Initialization

The EM algorithm [30] is an iterative process. Initialization of EM is a critical issue because EM converges to a local maximum of the likelihood function: the final estimate depends on the initialization. Therefore, an important consideration is that the parameter estimates need to be initialized for the first iteration of the algorithm. The choice of initial parameters constrains the solution found by the EM algorithm and hence is very important. There are several options for the choice of initial parameters. In [164, 166], Zolfaghari et al. presented two methods of initialization of EM algorithm. They include:

- Initializing the means uniformly over the interval. The variances are made significant

with respect to the interval and the number of Gaussian components in the mixture. The mixture weights are set equal values.

- The final iteration of the previous frame is used for initializing the current frame.

The second method for choosing the initial values for the EM algorithm has a drawback. The estimation of the current frame is poor, if the estimation of the previous frame is poor. Therefore, in [164, 166], Zolfaghari et al. suggested the combination of two methods. In [140], via small initial experiment, Stuttle suggested that the good initialization values for each Gaussian component i in the GMM would be as follows.

$$\mu_i = \frac{N(i+0.5)}{M} + 0.5 \tag{2.40}$$

$$\sigma_i^2 = \frac{N^2}{M^2} \tag{2.41}$$

$$\alpha_i = \frac{1}{M} \tag{2.42}$$

This initial values are similar to the first method in [164, 166]. In this dissertation, we employ the initial values in Eqs. (2.40), (2.41), and (2.42).

2.5.4 Issues in estimating spectral-GMM parameters from a speech spectrum

In the previous subsection, the theory for estimating spectral-GMM parameters from a speech spectrum was presented. In this subsection, a number of issues with the implementation of this algorithm is examined.

Spectral smoothing

The characteristic shape of the speech spectrum can give problems for estimating spectral-GMM parameters. The voiced speech spectrum is characterized by a number of pitch peaks separated by the fundamental frequency. The choice of initial parameters can determine the maxima found by the EM algorithm. There exist many local maxima the EM algorithm could find at each of these pitch peaks. If the pitch peaks are separated by a high fundamental frequency, a maximum could be found estimating a Gaussian component to a single pitch peak, and ignoring the adjacent harmonics. The Gaussian component which models the harmonic has a very small variance, but does not represent the general spectral envelope. To represent the phonetic class of the speech spectrum, it is desirable that the GMM models the spectral envelope, and avoids the problem of components converging upon the pitch peaks.

There are three possible solutions to the problem of the Gaussian components representing the spectral harmonics:

- **Variance flooring:** applying a variance floor to the Gaussian components prevents any component becoming too narrow and representing only a single pitch peak.

- **Spectral smoothing:** using a smoothing algorithm, such as the SEEVOC algorithm [118], to remove the pitch or voicing from the spectrum and estimate the vocal tract function.

- **Overlapping bin functions:** rather than the FFT bins being represented by a rectangular histogram function, the bins could be allowed to overlap each other. This technique can be related to a spectral smoothing approach.

Prior distributions

The technique for estimating spectral-GMM parameters from a spectral histogram in Section 2.5.2 places no prior constraint on the parameters extracted. Small changes in the spectral histogram could lead to large changes in the solutions found by the EM algorithm. It is possible that the estimated parameters could vary greatly between two frames that appear similar. Therefore, some constraints, or priors, should been applied on the locations of the peaks modeled.

Issues from the EM algorithm

When estimating the spectral-GMM parameters, we employ the EM algorithm [30]. The configuration of the EM algorithm, i.e. the number of Gaussian components, the initialization, and number of iterations, affects the estimated spectral-GMM parameters. These issues are described as follows.

- *Number of components.* One variable to consider in estimating spectral-GMM parameters from a spectrum is the number of Gaussian components. Note that increasing the number of components leads to better modeling of the speech spectrum. However, it increases the size of the feature vector, and gives difficulties in controlling Gaussian components to alter the characteristic of the spectral envelope.

- *Initialization of the EM algorithm.* The EM algorithm is sensitive to the values used to initialize the estimated parameters. The choice of initial parameters affect the local maxima, and it is an important parameters. In [164], the use of the previous frame values for the initialization of the spectral-GMM parameters was mentioned. However, initializing the EM algorithm with the spectral-GMM parameters from the previous frame causes problems when the speech changes suddenly, such as a plosive sound. The estimated features respond poorly to rapid changes in the speech. The mixture weights of some components weights can approach zero and the variances become very large. When the values are passed onto the next frame, the small weights and large variances of Gaussian components used to initialize will lead to the EM algorithm finding a local maximum with the variances approaching infinity and the priors approaching zero.

- *Maximum number of iterations.* When estimating spectral-GMM parameters, another important parameter of the EM algorithm is the maximum number of iterations. In general, in terms of the objective function, decreasing the maximum

number of iterations will yield a less precise estimation, but will take less time, and increasing the maximum number of iterations will yield a more precise estimation, but take more time. It should be noted that in some cases, the large maximum number of iterations only increases precise estimation in term of the objective function. If too many iterations are used, the model will find local maxima, and ignore estimation of some important features of the speech spectrum (e.g. important spectral peaks).

2.5.5 Properties of the spectral-GMM parameters

In the previous subsections, we have outlined the technique for extracting spectral-GMM parameters from a speech spectrum, and some issues of this technique. In this subsection, we discuss the properties of the spectral-GMM parameters.

Gaussian parameters as formant-like features

The spectral-GMM parameters can be considered to be analogous to a set of formant-like features [167]. The means, standard deviations, and mixture weights of Gaussian components correspond to the formant locations, the formant bandwidths, and the formant amplitudes, respectively. Once the spectral-GMM parameters have been extracted from a speech spectrum, they are ordered according to their frequency values. These formant-like features have been employed in speech recognition [140, 141, 142], and adaptation of children's speech [28]. In addition, Qin et al. [121] found out the correspondence between the mean values of the first Gaussian component and fundamental frequencies.

Flexible control over the speech spectrum

Modeling of speech spectrum using spectral-GMM parameters gives flexibilities to control the speech spectrum. We present a method for modifying amplitudes of spectral peaks in Chapter 3, and another method for modifying frequencies of spectral peaks in Chapter 5.

2.6 STRAIGHT

2.6.1 Outline of STRAIGHT

The source-filter model, which separates speech information into mutually independent filter parameters and source parameters, has been found to produce synthetic speech of very high quality. However, for such systems to attain this high quality the parameters need careful attention and hand-editing [165]. The STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) system [65] based on the source-filter model allows flexible control of speech parameters. Successive refinements on extraction procedure of source and spectral parameters enable the total system to re-synthesize high-quality speech. The STRAIGHT independently allows for over 600 % manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, without introducing further degradation due to parameter manipulation.

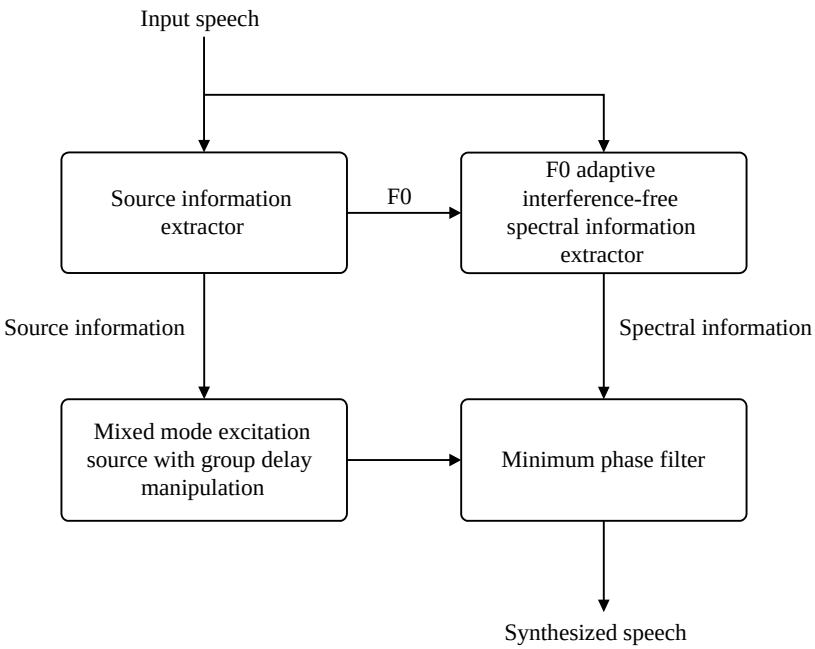

Figure 2.11: Schematic structure of STRAIGHT (adapted from [12]).

Although STRAIGHT is basically a channel VOCODER [65], its design objective greatly differs from its predecessors. We briefly describe the main parts of STRAIGHT. The detailed mathematical treatment of STRAIGHT can be found in [64, 65]. Figure 2.11 shows the schematic diagram of STRAIGHT. It consists of three key subsystems:

- **Source information extractor:** Dedicated $F0$ extractors for STRAIGHT are based on instantaneous frequency. The fundamental frequency is accurately estimated to smooth out the periodic bouncing in the short-term spectrum using an $F0$-adaptive filter.
- **Smoothed time-frequency representation extractor:** The central feature of STRAIGHT is the extended pitch synchronous spectral analysis that provides a smooth artifact-free time-frequency representation of the spectral envelope of the speech signal. The STRAIGHT spectrum is basically an $F0$-independent representation.
- **Synthesis engine:** It consists of an excitation source and a time varying filter. The time varying filter is implemented as the minimum phase impulse response calculated from the smoothed time-frequency representation through several stages of FFTs.

In this dissertation, we employ STRAIGHT as an analysis/synthesis method for two main reasons. First, STRAIGHT is an extremely high-quality analysis/modification/synthesis method [65]. This method has been successfully applied in many speech application areas, especially in speech modification [66, 81, 104, 107, 108, 121, 146, 147, 148]. Second,

STRAIGHT uses a pitch-adaptive spectral analysis scheme combined with a surface reconstruction method in the time-frequency plane to remove signal periodicity. This results in a smooth spectral representation both in the time domain and in the frequency domain [65]. It follows that the LSF parameters extracted from the STRAIGHT spectra are correlated among frames, and thus the corresponding LSF contours are smooth also. It is not the case of a normal LP analysis method, where LSF parameters are extracted independently on a frame-by-frame basis.

2.6.2 Derivation of LSF parameters

STRAIGHT is based on a simple channel VOCODER [65]. It decomposes input speech signals into source information (F0, aperiodic indexes) and spectral information (STRAIGHT spectral envelope). LSF parameters are extracted from a STRAIGHT spectrum as follows.

The amplitude spectrum $X[n]$, where $0 \leq n \leq \frac{NF}{2}$ (NF is the number of samples in the frequency domain), obtained from STRAIGHT analysis is transformed into the power spectrum using Eq. (2.43).

$$S[n] = \mid X[n] \mid^2, \quad 0 \leq n \leq \frac{NF}{2} \tag{2.43}$$

The ith autocorrelation coefficient, R[i], is then calculated using the inverse Fourier transform of the power spectrum as follows.

$$R[i] = \frac{1}{NF} \sum_{n=0}^{NF-1} S[n] \exp\{j\frac{2\pi ni}{NF}\}, \quad 0 \leq i \leq NF-1 \tag{2.44}$$

where $S[n] = S[NF-n]$. Assuming that the speech samples can be estimated by a Pth order all-pole model, where $0 < P < NF$, the reconstruction error is calculated as given in Eq. (2.45).

$$G_L = R[0] - \sum_{l=1}^{P} a_l^P R[l] \tag{2.45}$$

where $\{a_l^P\}$, $l = 1, 2 \cdots P$, are the corresponding linear predictive coding (LPC) coefficients. G_L hereafter is referred to as gain. By minimizing G_L with respect to a_l^P, a_l^P could be estimated. They are then transformed into LSF parameters.

2.7 Summary

In this chapter, we have presented the background for our dissertation. First, we have briefly described the speech production. We then have given introduction of the source-filter model for speech production. We also have presented the representation of the vocal tract filter in the source-filter model, i.e. LP coefficients, and LSF parameters. Last, we have presented the key details of the temporal decomposition technique, speech spectrum modeling using spectral-GMM parameters, as well as brief introduction of STRAIGHT.

Chapter 3

Spectral Modelings for Speech Modification

The aim of this chapter is to solve the first issue of spectral modification, the lack of efficient spectral modelings for speech modification. In general, speech modification consists of the process of analyzing a speech signal into a number of parameters, modifying these parameter values in accordance with a desired goal, and synthesizing the correspondingly modified speech signal. Therefore, to obtain the high-quality of modified speech, the first task is to develop a spectral modeling which can allow to perform efficient modification. This chapter first presents two improvements of the speech spectral envelope modeling using spectral-GMM parameters. We then introduce a new modeling of the speech spectral sequence.

3.1 Introduction

Spectral modification techniques are used to perform a variety of modifications to speech spectra, such as manipulations of the formant structures, amplitude manipulations. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing. It can be applied in many areas. Spectral modification methods are a powerful technology for customizing Text-to-Speech (TTS) systems, such as by converting source features to target features [1, 143], changing a male voice into a female voice and vice versa [77], and applying emotional speech synthesis [66]. Spectral modification techniques are often applicable to automatic speech recognition tasks [49], and speech enhancement [15].

The basic idea of spectral processing is to convert a time-domain digital signal into its frequency-domain representation. Most of the approaches start by developing an analysis/synthesis technique from which the speech signal is reconstructed with minimum loss of sound quality. Then, the main issues have to be resolved: what kind of representation and which parameters are chosen for the application of the desired speech processing. The challenge of spectral modification is to modify the spectral/acoustic features without degrading the speech quality.

This chapter presents spectral modelings for speech modification. In the first part, we

present two improvements of modeling of a speech spectral envelope. In the second part, we propose a new framework for modeling speech spectral sequence.

3.2 Modelings of a speech spectral envelope

3.2.1 Speech spectrum modeling using Gaussian mixture model

One of the most important requirements of spectral modification is that it be flexible enough to perform a variety of modifications within the spectral envelope. Formant frequency is one of the most important parameters in characterizing speech, and it also plays an important role in specifying speaker characteristics. Therefore, using formant frequency as a parameter can control other features that are directly connected to the speech production process. Conventional spectral modification methods, such as [94, 99, 151], often control formants to modify the speech spectrum. However, these methods are limited by their inability to independently control important formant characteristics such as amplitude and bandwidth, or to control the spectral shape.

Zolfaghari et al. proposed a technique to fit a Gaussian mixture model to a smoothed magnitude spectrum of a speech signal [166, 167, 170]. This technique is briefly described as follows.

In a single frame, the normalized spectrum $X(e^{j\omega_n})$ is viewed as a probability distribution $P(X)$, where $X = \{x_1, ..., x_L\}$, x_l $(1 \leq l \leq L)$ is the frequency bin number, and $2L$ is the FFT size. $P(x_l)$ is simply a spectral density. The overall density of a Gaussian mixture model is written by

$$u(x) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m, \sigma_m^2) \tag{3.1}$$

where M is the number of mixture components, $\mathcal{N}(x; \mu_m, \sigma_m^2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(x-\mu_m)^2}{2\sigma_m^2}}$ is the mth local Gaussian component, μ_m, σ_m are called mean and standard deviations of Gaussian component m respectively, and $\{\alpha_m\}_{m=1}^{M}$ are mixture weights satisfying $0 \leq \alpha_m \leq 1$ and $\sum_{m=1}^{M} \alpha_m = 1$.

Zolfaghari et al. assumed that formants could be represented by Gaussian distributions, and a speech spectrum could be represented by a Gaussian mixture model. The EM algorithm [30] is often used to optimize the log likelihood of the histogram of the speech spectrum at time t with respect to the model parameters $u(x)$ in Eq. (3.1). The estimated means, standard deviations, and mixture weights of the Gaussian components can be related to the locations, bandwidths, and amplitudes of the formants, respectively [167]. The ability to independently control the parameters of each Gaussian component enables precise estimation of the spectral envelope, enables a wide variety of modifications, and enables independent control of the formants.

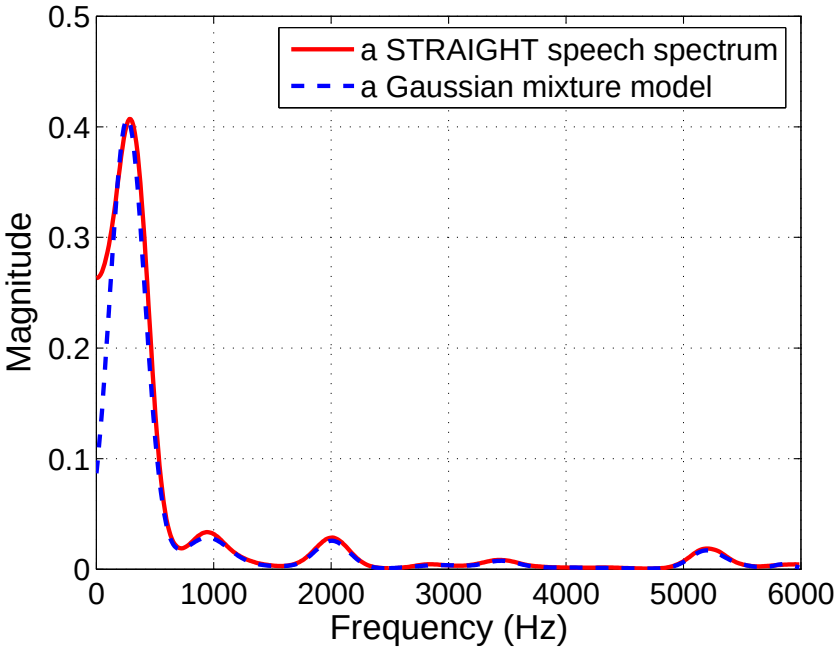

Figure 3.1: A Gaussian mixture model (M=8) fits to a STRAIGHT spectral envelope.

3.2.2 Smoothed-spectrum representation by STRAIGHT

The characteristic shape of the speech spectrum can present problems for estimating a set of Gaussian components. The voiced speech spectrum is characterized by a number of pitch peaks separated by the fundamental frequency. If the pitch peaks are separated by a high fundamental frequency, a maximum can be found by estimating a Gaussian component for a single-pitch peak, and ignoring the adjacent harmonics. This results in a very small variance for that Gaussian component. Therefore, the high-quefrency effects of the excitation from the spectrum are removed to improve the representation of the spectral envelope by the GMM fitting method. In this dissertation, we model a STRAIGHT spectral envelope using spectral-GMM parameters. STRAIGHT [65] uses a pitch-adaptive spectral analysis scheme combined with a surface reconstruction method in the time-frequency plane to remove signal periodicity. This results in a smooth spectral representation free of glottal excitation information. Figure 3.1 shows that a Gaussian mixture model of eight Gaussian components can fit to a STRAIGHT spectral envelope (at 12 kHz sampling frequency) well.

3.2.3 Improvement of spectral peak estimation using Gaussian mixture model

The ability to independently control the parameters of each Gaussian component enables precise estimate of the spectral envelope, enables a wide variety of modifications, and enables independent control of the formants. However, the original methods [166, 167, 170] do not ensure an one-to-one correspondence between spectral peaks and Gaussian compo-

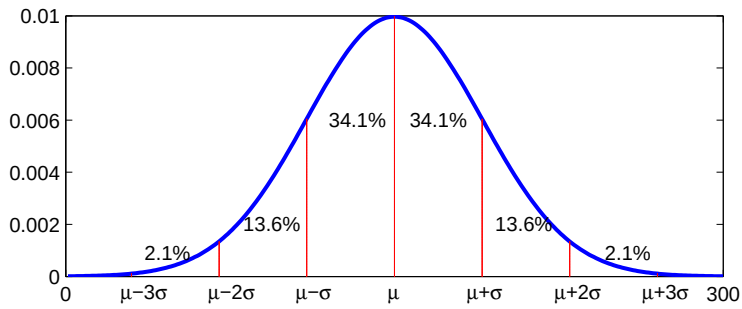

Figure 3.2: Illustration of the Empirical rule of Gaussian distribution.

nents. This creates difficulties in modifying the speech spectral envelope in both dimensions, frequency and amplitude. To overcome this drawback, we propose an improvement of modeling speech spectral envelope for speech modification. The aim of our proposed method is not only to model the speech spectral envelope well, but also to ensure an one-to-one correspondence between spectral peaks and Gaussian components. We describe one important characteristic Gaussian distribution , called "68-95-99.7 rule" or the "empirical rule", which we used to solve these problems. The rule is described as follows. About 68% of values drawn from a Gaussian distribution $\mathcal{N}(z;\mu,\sigma^2)$ fall in the interval $\mu-\sigma$ to $\mu+\sigma$, about 95% of the values fall in the interval $\mu-2\sigma$ to $\mu+2\sigma$, and about 99.7% fall in the interval $\mu-3\sigma$ to $\mu+3\sigma$. This characteristic is illustrated in Figure 3.2. Based on this geometric characteristic, we develop an improvements to model a speech spectral envelope as follows, and the diagram of our method is shown in Figure 3.3.

First, from a speech spectral envelope, we start to estimate spectral-GMM parameters with the initial 18 Gaussian components. The requirement of the initial number of Gaussian components is that the number of components be high enough to model the speech spectral envelope well. Note that the initial Gaussian components depend on the sampling frequency. We assume that if the linear sum of two Gaussian components has only one peak, these two Gaussian components are dependent, and also that, if the linear sum of two Gaussian components has two peaks, these two Gaussian components are independent. After we get the spectral-GMM parameters in the first iteration, we check whether or not all Gaussian components are independent components. If not, we divide the spectral-GMM parameters into two groups. The first group models the spectral shape of the speech spectral envelope, and the other group models the spectral peaks of the speech spectral envelope. On the basis of the geometric characteristics of normal distribution, i.e. the empirical rule, we assume that a Gaussian component m is a spectral shape factor if there are at least two other Gaussian components located between $[-3\mu_m, 3\mu_m]$, where μ_m is the mean of this Gaussian component m. If two Gaussian components i, j are dependent spectral peaks, we merge these two Gaussian components by the following equations.

$$\mu_{ij} = \frac{\alpha_i\mu_i + \alpha_j\mu_j}{\alpha_i + \alpha_j} \tag{3.2}$$

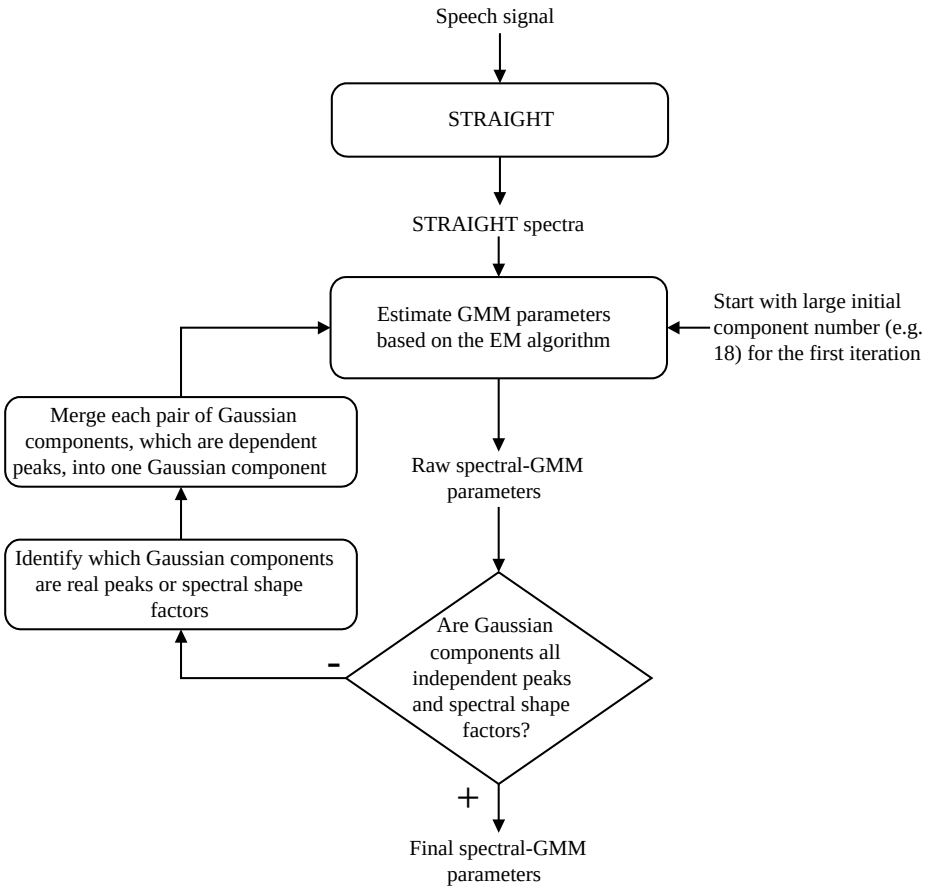

Figure 3.3: Diagram of our proposed method for modeling the speech spectral envelope.

$$\sigma_{ij}^2 = \frac{\alpha_i(\sigma_i^2 + (\mu_{ij} - \mu_i)^2) + \alpha_j(\sigma_j^2 + (\mu_{ij} - \mu_j)^2)}{\alpha_i + \alpha_j} \tag{3.3}$$

where $\mu_i, \sigma_i, \alpha_i$ and $\mu_j, \sigma_j, \alpha_j$ are the means, standard deviations, and mixture weights of Gaussian components i and j, respectively. After merging Gaussian components, a new process of estimating spectral-GMM parameters is executed, with the condition that the initial parameters for the new process are current Gaussian components. The process of spectral-GMM estimation continues to iterate, and terminates when all Gaussian components are independent components. Obviously, this algorithm always converges, since after each iteration, the number of Gaussian components decreases by at least 1. An example of the spectral envelope restored by our proposed method is illustrated in Figure 3.4.

3.2.4 Spectral envelope modeling using asymmetric Gaussian mixture model

Zolfaghari et al. [166, 167, 170] proposed a technique to fit a Gaussian mixture model to a smoothed magnitude spectrum of a speech signal. As mentioned above, the ability

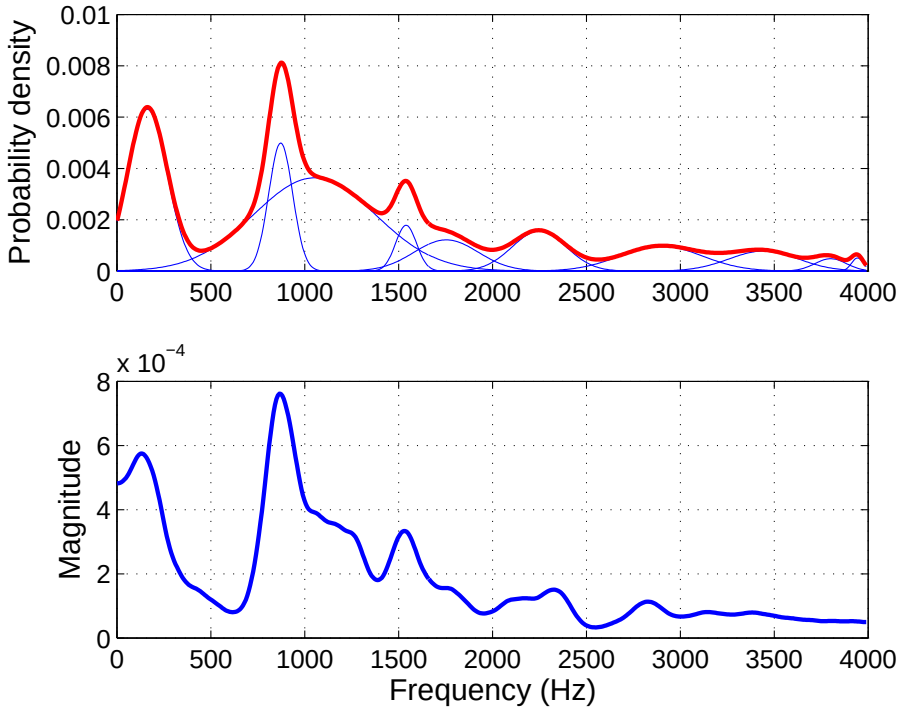

Figure 3.4: Example of the spectral envelope restored by our proposed method: the thin lines indicate Gaussian components, the bold line indicates the restored spectral envelope (top), and STRAIGHT spectral envelope (bottom).

to independently control parameters of each Gaussian component enables a precise estimate of the spectral envelope, a wide variety of modifications, as well as independently controlling formants. In this model, Zolfaghari et al. assumed that formants could be represented by Gaussian distributions. However, mixture of Gaussians does not always fit any distribution of patterns.

To overcome this drawback, we utilize asymmetric Gaussian mixture model (AGMM) to model a speech spectral envelope, instead of using Gaussian mixture model. The main advantage of AGMM is that we can independently modify both sides of each asymmetric Gaussian component.

Asymmetric Gaussian mixture model

The normal distribution, also called the Gaussian distribution, is an important family of continuous probability distributions, and has been applicable in many fields. Many measurements, ranging from psychological to physical phenomena (in particular, thermal noise) can be approximated, to varying degrees, by the normal distribution. The normal distribution is the most widely used family of distributions in statistics, and many statistical tests are based on the assumption of normality. In probability theory, normal distributions arise as the limiting distributions of several continuous and discrete families of distributions. In addition, the normal distribution maximizes information entropy

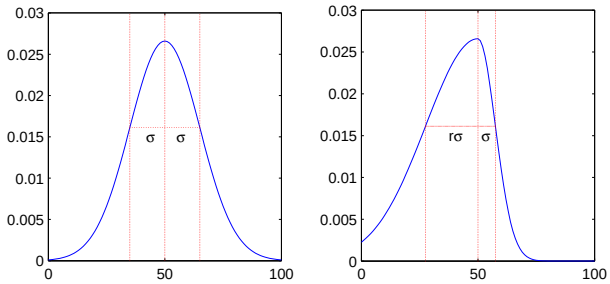

Figure 3.5: Example of Gaussian and AG distribution: Gaussian (left), and asymmetric Gaussian (right).

among all distributions with known mean and variance, which makes it be the natural choice of underlying distribution for data summarized in terms of sample mean and variance.

However, the normal distribution and the Gaussian mixture model do not always fit any distribution of patterns. In the literature, to increase the accuracy rate, some methods [13, 63] use asymmetric Gaussian (AG), instead of using Gaussian distribution, and they get some promising results. The asymmetric Gaussian can be defined as follows.

$$\mathcal{N}(z;\mu,\sigma_1^2,\sigma_2^2) = \frac{1}{\sqrt{2\pi}}\frac{1}{\frac{1}{2}(\sigma_1+\sigma_2)}\begin{cases} exp\big(-\frac{(z-\mu)^2}{2\sigma_2^2}\big) & if \quad z<\mu \\ exp\big(-\frac{(z-\mu)^2}{2\sigma_1^2}\big) & if \quad \mu \le z \end{cases} \tag{3.4}$$

Kato et al. [63] represented the asymmetric Gaussian distribution in another form as shown in Eq. (3.5).

$$\mathcal{N}(z;\mu,\sigma^2,r) = \frac{2}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma^2}(r+1)}\begin{cases} exp\big(-\frac{(z-\mu)^2}{2\sigma^2}\big) & if z>\mu \\ exp\big(-\frac{(z-\mu)^2}{2\sigma^2 r^2}\big) & otherwise \end{cases} \tag{3.5}$$

where μ, σ^2, r are parameters of $\mathcal{N}(z;\mu,\sigma^2,r)$, and μ, σ, r are called mean, standard deviation, and shape adjustment factor of AG component, respectively. In the same manner as Gaussian, the d-dimensional AG has a latent variable $z \in R^d$, and the observation variable x is modeled using z and an orthonormal matrix $\Phi \in R^{d\times d} : x = \Phi z$. The difference between the AG and the Gaussian is the distribution of the latent variable z. Examples of Gaussian and AG are shown in Figure 3.5.

It is easy to realize that AG is an extension of Gaussian, since AG with $r = 1$ is equivalent to Gaussian. Moreover, AG also has the "68-95-99.7 rule" or the "empirical rule". The rule is described as follows. About 68% of values drawn from an AG distribution $\mathcal{N}(z;\mu,\sigma^2,r)$ fall in the interval $\mu - r\sigma$ to $\mu + \sigma$, about 95% of the values fall in the interval $\mu - 2r\sigma$ to $\mu + 2\sigma$, and about 99.7% fall in the interval $\mu - 3r\sigma$ to $\mu + 3\sigma$. This characteristic is illustrated in Figure 3.6.

Like the Gaussian distribution, the AG can not represent densities with multiple

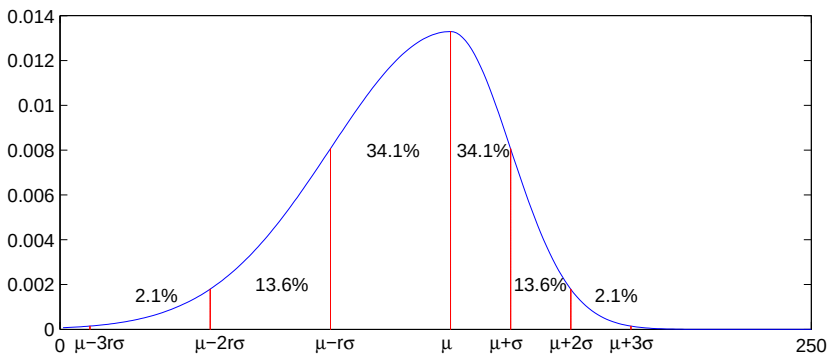

Figure 3.6: Illustration of the Empirical rule of AG.

modes. It is therefore straightforward to consider a mixture of AG, which is able to model complex data structures with a linear combination of local AGs. The overall density of the M-component mixture model is written by

$$p(x) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m, \sigma_m^2, r_m) \tag{3.6}$$

where M is number of mixture components, $\mathcal{N}(x; \mu_m, \sigma_m^2, r_m)$ is the mth local AG, and $\{\alpha_m\}_{m=1}^{M}$ are mixture weights satisfying $0 \leq \alpha_m \leq 1$ and $\sum_{m=1}^{M} \alpha_m = 1$.

EM-based spectral envelope modeling using asymmetric Gaussian mixture model

The main purpose of this subsection is to find an optimal parameter set of asymmetric Gaussian mixture model to fit to the speech spectral envelope. The parameters of a probability density function are the number of AGs M, the mean μ_m, the standard deviation σ_m, the weighting factors α_m, and shape adjustment factor r_m of each AG function. To find these parameters to optimally fit a certain probability density function for a set of data, an iterative algorithm, the EM algorithm [30], is often used.

The technique to estimate spectral-AGMM parameters is similar to the technique to estimate spectral-GMM parameters, which is described in Subsection 2.5.2. Except that we use the AG distribution to present the density, instead of using Gaussian distribution. Asymmetric Gaussian distribution is described as follows.

$$\begin{aligned} f(x_k|i, \phi_i) &= \mathcal{N}(x_k, \mu_i, \sigma_i^2, r_i) \\ &= \frac{2}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_i^2}(r_i+1)} \begin{cases} exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) & if x > \mu_i \\ exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2 r_i^2}\right) & otherwise \end{cases} \end{aligned} \tag{3.7}$$

where μ_i is the mean, σ_i is the standard deviation and r_i is the shape adjustment factor. Maximization of the Q-function is achieved by maximizing each term in Eq. (2.34) with

respect to $\overline{\alpha}_i$ and $\overline{\phi}_i$. The following equation is obtained from the second term of Eq. (2.34).

$$\begin{aligned}\frac{\partial Q_1(\Phi,\overline{\Phi})}{\partial\overline{\phi}_i} &= \sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i)\frac{\partial}{\partial\overline{\phi}_i}[logf(x_k|i,\overline{\phi}_i] \\ &= 0 \end{aligned} \tag{3.8}$$

Differentiating for $\overline{\mu}_i$, $\overline{\sigma}_i^2$ and $\overline{r}_i$ in Eq. (3.8), and setting it equal to zero, we obtain new parameter estimates $\overline{\mu}_i$, $\overline{\sigma}_i^2$ and $\overline{r}_i$

$$\overline{\mu}_i = \frac{\sum_P P(x_k)P(i|x_k,\phi_i)x_k + \sum_Q P(x_k)P(i|x_k,\phi_i)x_k r_i^2}{\sum_P P(x_k)P(i|x_k,\phi_i) + \sum_Q P(x_k)P(i|x_k,\phi_i)r_i^2} \tag{3.9}$$

$$\begin{aligned}\overline{\sigma}_i^2 = &\frac{\sum_P P(x_k)P(i|x_k,\phi_i)\frac{(x_k-\overline{\mu}_i)^2}{r_i^2}}{\sum_{i=1}^{N} P(x_k)P(i|x_k,\phi_i)} \\ &+\frac{\sum_Q P(x_k)P(i|x_k,\phi_i)(x_k-\overline{\mu}_i)^2}{\sum_{i=1}^{N} P(x_k)P(i|x_k,\phi_i)}\end{aligned} \tag{3.10}$$

$$\begin{aligned}&\sum_{i=1}^{N} P(x_k)P(i|x_k,\phi_i)\overline{v}_i\overline{r}_i^3 \\ &-\sum_P P(x_k)P(i|x_k.\phi_i)(x_k-\overline{\mu}_i)^2\overline{r}_i \\ &-\sum_P P(x_k)P(i|x_k.\phi_i)(x_k-\overline{\mu}_i)^2 = 0\end{aligned} \tag{3.11}$$

where $P = \{\forall x_k | x_k \leq \mu_i\}$ and $Q = \{\forall x_k | x_k > \mu_i\}$. From the first term of Eq. (2.34), the mixture weights are formulated as follows

$$\overline{\alpha}_i = \frac{1}{N}\sum_{k=1}^{N} P(x_k)P(i|x_k,\phi_i) \tag{3.12}$$

Using the EM algorithm, we can estimate the spectral-AGMM parameters to model the shape and the peaks of the speech spectral envelope. Like spectral-GMM parameters, spectral-AGMM parameters are also related to formant information.

3.2.5 Experiments and results

In this subsection, we conduct experiments to evaluate our proposed modelings of the speech spectral envelope. We evaluate the improvement of spectral peak estimation using Gaussian mixture model in the first part, and the speech spectral envelope modeling using asymmetric Gaussian mixture model in the second part. In both experiments, we compare

Table 3.1: Analysis conditions for experiments of testing methods

STRAIGHT	Sampling frequency	8 kHz
	Window length	40 ms
	Window shift	2 ms
	FFT points	1024

our methods with the original method for estimating spectral-GMM parameters from a speech spectral envelope [167].

Evaluation of the improvement of spectral peak estimation using Gaussian mixture model

To evaluate the effectiveness of our proposed method, we conducted some experiments. We used an objective measure and a subjective test to evaluate the quality of synthesized speech which was restored by our proposed method. We compared our proposed method with the original GMM method. The objective measure used here is the average log spectral distortion (LSD) between STRAIGHT spectra and the spectral envelopes reconstructed from the original GMM method, and from our proposed method, respectively. This criterion is a function of the distortion introduced in the spectral density of speech in each particular frame. Log spectral distortion between source and target for the nth frame, D_n, is defined (in dB) as follows.

$$D_n = \sqrt{\frac{1}{F_s}\int_0^{F_s}[10log_{10}(P_n(f)) - 10log_{10}(\hat{P}_n(f))]^2 df} \tag{3.13}$$

where F_s is the sampling frequency; $P_n(f)$ and $\hat{P}_n(f)$ are the power spectrum corresponding to the nth frame of the source and the nth frame of target, respectively. In this subsection, $P_n(f)$ is the nth STRAIGHT power spectrum, and $\hat{P}_n(f)$ is the nth power spectrum which is reconstructed from the original GMM method, and our proposed method. The results are provided in terms of average log spectral distortion and percentage outliers.

For the objective test, both methods used an approximate number of Gaussian components. A set of 150 sentence utterances of the ATR Japanese speech database [2] was selected as the speech data. This dataset spoken by 6 speakers (3 male & 3 female) is resampled at 8 kHz sampling frequency. The analysis conditions for these experiments are shown in Table 3.1. The number of Gaussian components in the original GMM method is 9, which is approximately the average number of Gaussian components in our proposed method (9.1). Table 3.2 gives a comparison of the log spectral distortion results the original GMM method and our proposed method. Experimental results indicate that the performance of our proposed method is better than that of the original GMM method.

For the subjective test, we randomly presented each of four utterances restored from both the original GMM method and our proposed method to eight Japanese graduate students with normal hearing ability, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Figure 3.7

Table 3.2: Average LSD, and percentage number of outlier frames obtained from (1) the original GMM method and (2) our proposed method.

Method	Avg. LSD (dB)	0-2 dB (%)	2-4 dB (%)	>4 dB (%)
Original GMM method	2.6878	18.820	73.082	8.098
Our proposed method	1.9004	66.996	30.262	2.742

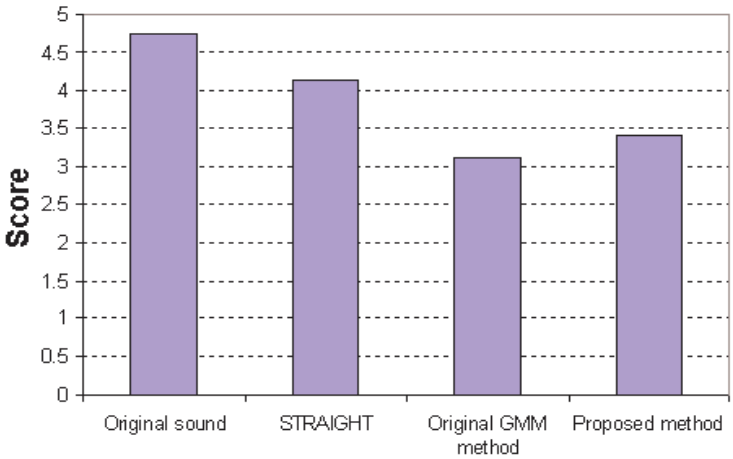

Figure 3.7: Comparison of listening test results for the compared methods.

shows the average scores, which indicate that the speech quality of our proposed method is better than that of the original GMM method.

Evaluation of speech spectral envelope modeling using asymmetric Gaussian mixture model

Experiments were performed to determine the effectiveness of modeling speech spectral envelope using spectral-AGMM method. In this part, we compare the quality of the spectral envelopes restored by using our proposed method with the spectral envelopes restored from spectral-GMM parameters [166, 167, 170] by using an objective test and a subjective test.

The objective measure used here is also log spectral distortion (LSD), as shown in Eq. (3.13). In this part, $P_n(f)$ is the nth STRAIGHT power spectrum, and $\hat{P}_n(f)$ is the nth power spectrum which is reconstructed from one of two kinds of spectral parameters, spectral-GMM and spectral-AGMM. The results are also provided in terms of average log spectral distortion and percentage outliers.

A set of 150 sentence utterances of the ATR Japanese speech database [2] was selected as the speech data. This dataset spoken by 6 speakers (3 male & 3 female) is re-sampled at 8 kHz sampling frequency. The analysis conditions for these experiments are shown in Table 3.3. Table 3.4 gives a comparison of the log spectral distortion results between our proposed method and the spectral-GMM method. Results indicate that the performance

Table 3.3: Analysis conditions for experiments of AGMM performance

STRAIGHT	Sampling frequency Window length Window shift FFT points	8 kHz 40 ms 2 ms 1024
GMM method	Iteration of EM algorithm	30 times
AGMM method	Iteration of EM algorithm	30 times

Table 3.4: Average LSD, and percentage number of outlier frames obtained from the spectral-AGMM and spectral-GMM methods. The first line in each row is the result of the spectral-GMM method, and the second line is the result of the spectral-AGMM method.

Number of components	Avg. LSD (dB)	0-2 dB (%)	2-4 dB (%)	> 4 dB (%)
4	4.94 4.92	0.24 0.34	34.13 34.27	65.63 65.39
6	3.87 3.85	1.90 2.93	62.45 61.48	35.65 35.59
8	3.24 3.20	6.17 9.99	75.82 71.49	19.01 18.52
10	2.83 2.72	13.82 23.29	76.32 67.05	9.86 9.66
12	2.49 2.34	26.54 41.80	68.06 52.90	5.40 5.30
14	2.21 2.03	43.47 60.26	53.06 36.54	3.47 3.10

of our proposed method is better than that of the method using GMM parameters. Note that results of the average LSD in Table 3.4 were calculated for whole frames of all utterances which include both voiceless and voiced frames. One example of restored spectral envelopes from the two methods is illustrated in Figure 3.8.

For the subjective test, we randomly presented a set of four kinds (2 male & 2 female) of synthesized sounds including utterances synthesized from both AGMM and GMM methods with different components (i. e. 4, 6, 8, 10, 12, 14) to eight Japanese graduate students with normal hearing ability, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Fig. 3.9 shows the average scores, which indicate that the speech quality of our proposed method is slightly better than that of GMM method.

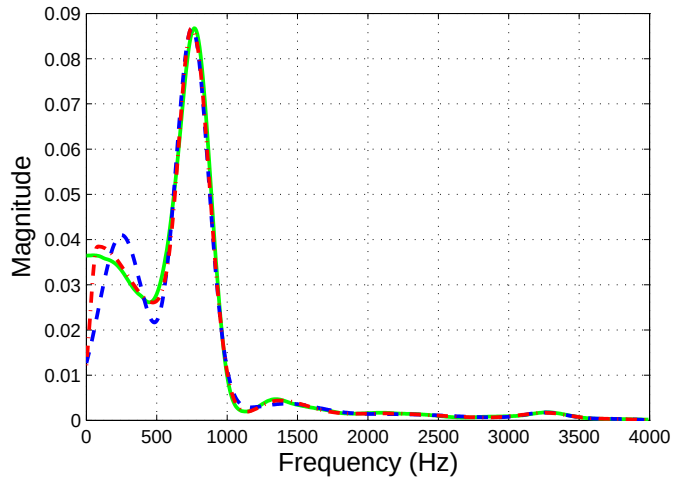

Figure 3.8: Comparison of fitting in the STRAIGHT smoothed spectrum by using GMM (10 components), and AGMM (10 components). The solid line is a STRAIGHT spectrum of a frame, dash-dot line is AGMM fitting in the STRAIGHT spectrum, and the dashed line is GMM fitting in the STRAIGHT spectrum

3.2.6 Conclusions

In this section, we have presented two improvements of modeling of a speech spectral envelope using Gaussian mixture model. In the first improvement, our proposed method not only models the speech spectral envelope well, but also ensures an one-to-one correspondence between spectral peaks and Gaussian components, which provides advantages when modifying the spectral envelope in both dimensions, frequency and amplitude. In the second part, we use spectral-AGMM parameters to model a speech spectral envelope, instead of using spectral-GMM parameters. Since both sides of asymmetric Gaussian component can be independently modified, spectral-AGMM parameters have been found to be fitted to a speech spectral envelope better than spectral-GMM parameters. In addition, this modeling gives more flexibilities to modify a speech spectral envelope.

3.3 Modeling of the speech spectral sequence

3.3.1 Introduction

Modeling of the speech spectral sequence plays an important role in spectral modification. We should add acoustic constraints to this modeling for two reasons.

The first reason is that we have to control natural evolution of spectral parameters when altering duration of speech. Several approaches are available in the literature for time-scale modification. Altering the time-scale of a speech signal can be achieved in the time domain [156], or frequency domain [33]. Time-domain techniques are based on overlap-add (OLA) methods [156]. These techniques first segment the waveform into a series of overlapping frames by windowing the speech signal with a suitable window

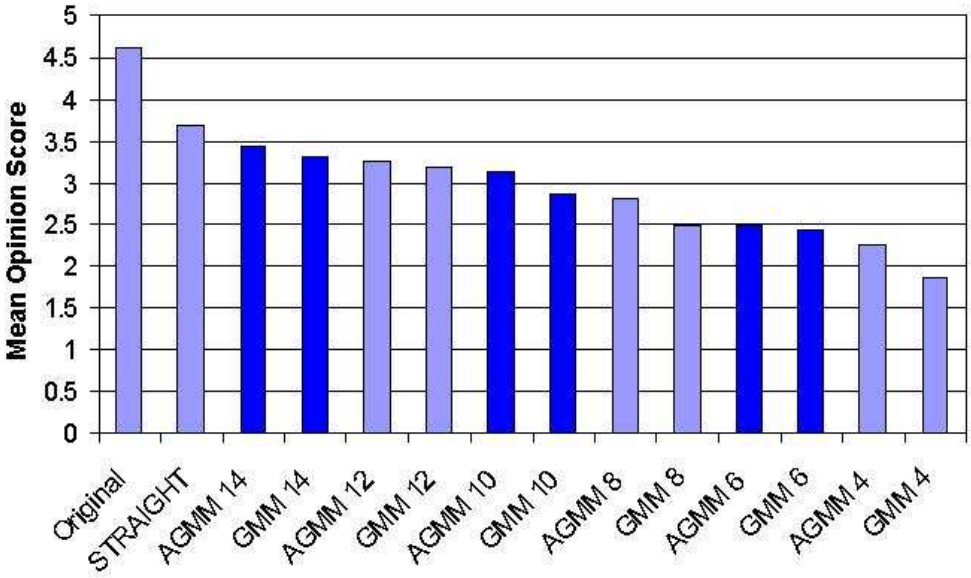

Figure 3.9: Comparison of mean opinion scores (MOSs) over varying number of Gaussians in the mixture using spectral-AGMM method and spectral-GMM method.

function. To perform time-scale modification, some of the windowed segments are either replicated or omitted. In these cases, the information about the pitch markers is not used for splitting the speech signal into short segments. As a result, the periodicity due to pitch is not preserved well after time-scale modification [125]. Therefore, these techniques tend to perform poorly when large modification factors must be used (e.g., factors greater than $\pm$ 20% to $\pm$ 30%) [76]. Frequency-domain techniques are based on short-time Fourier transform (STFT) or phase vocoder methods [33]. These algorithms require high computation costs, but are capable of providing high-quality output. However, they still suffer from some distortion, mainly due to the effects of "phase dispersion" [114]. That is, while the scaled signal has the same frequency, the phases between the components change, resulting in a different wave shape. In the STRAIGHT method [65], the analysis algorithm does not extract phase information. Its reconstruction algorithm adopts the minimum phase assumption for the spectral envelope, and further applies all-pass filters to reduce the buzz timbre of the reconstructed signal. This method offers high-quality modified speech signals without introducing the artificial timbre. However, this approach still processes speech signals frame by frame, and speech manipulation is performed by using interpolation functions. This method does not consider the temporal evolution of parameters when modifying speech signals.

The second reason is that, when performing spectral modification, if unexpected modifications happen, there are discontinuities, which leads to degradation of modified speech. Some methods have been proposed to solve the discontinuities of modified speech. For example, to maintain a continuous transformation in consecutive frames, Chen et al. [23] employ a median filter and a low pass filter to smooth the converted features along the time axis. However, applying these filters could lead to a loss of temporal resolution, and it was a relatively crude implementation.

This section proposes a framework to model the speech spectral sequence in the time-frequency domain. In the following subsection, we describe the overview of the temporal decomposition technique [5, 113], which is utilized to model the speech spectral sequence.

3.3.2 Temporal decomposition

A shortcoming of conventional spectral modification methods is that they do not take into account the correlation between frames after modification. There are some clicks in the modified speech because of discontinuous spectral contours. Therefore, we employ TD to deal with the problem.

In articulatory phonetics, speech can be described as a sequence of distinct articulatory gestures. Each gesture produces an acoustic event that should approximate a phonetic target. Adjacent gestures overlap in time, which results in overlap of these phonetic targets.

Atal proposed a method based on the temporal decomposition of speech into a sequence of overlapping target functions and corresponding event targets [5], as given in Eq. (3.14).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{3.14}$$

where $\mathbf{a}_k$ is the spectral parameter vector corresponding to the kth event target. The temporal evolution of this target is described by the kth event function, $\phi_k(n)$. $\hat{\mathbf{y}}(n)$ is the approximation of the nth spectral parameter vector $\mathbf{y}(n)$, and is produced by the TD model. N and K are the number of frames in the speech segment, and the number of event functions, respectively ($N \gg K$).

To modify the speech spectra, we only need to modify the event targets $\mathbf{a}_k$ and the corresponding event functions $\phi_k(n)$, instead of modifying the speech spectra frame by frame. The smoothness of modified speech is ensured by the shape of the event functions $\phi_k(n)$. This feature leads to easy modification of the speech spectra, as well as ensuring the smoothness of the speech spectra between frames, and thereby enhances the quality of modified speech.

The original method of TD is known to have two major drawbacks, high computational cost and high parameter sensitivity to the number and locations of events. A number of modifications have been explored to overcome these drawbacks. In this study, we employ the MRTD algorithm [113]. The reasons for using the MRTD algorithm in this work are twofold: (i) the MRTD algorithm enforces a new property on event functions, named the "well-shapedness" property, to model the temporal structure of speech more effectively [113]; (ii) event targets can convey the speaker's identity [111]. In the MRTD algorithm, LSF parameters are chosen for the input of TD, because of their spectral sensitivity (an adverse alteration of one coefficient results in a spectral change only around that frequency [117]) and their stability and interpolation advantages (LSFs result in low spectral distortion when being interpolated and/or quantized [116]). In this section, LSF parameters are extracted from spectral envelope information of STRAIGHT [65]. The STRAIGHT spectra are suitable for TD, because they are smooth in the time-frequency domain.

3.3.3 Modeling of the event function using polynomial fitting

New method for identifying the event locations

The MRTD algorithm uses a spectral stability criterion to determine the initial event locations [113]. It is assumed that each acoustic event that exists in speech gives rise to a spectrally stable point in its neighborhood. Therefore, the locations of the spectrally stable points and the corresponding spectral parameter sets can be used as good approximations of event locations and event targets, respectively. This algorithm is automatically performed, and the subsequent computation of refined event targets and event functions is much less demanding than the traditional TD method. This algorithm is useful for applications in speech coding [113], and speaker identification [111]. However, this algorithm does not ensure one-to-one correspondence between events and phonemic units. This correspondence can give some advantages in voice transformation (e.g. alignment between two utterances), speech perception (e.g. sharing the event functions, event targets), etc. Therefore, it requires a new method for identifying the event locations.

In [109, 133], a new method for determination of event locations based on phonemes has been proposed. In our method, we use labeled data of utterances to segment speech signals into phonemes. Each phoneme is divided into K equal segments, and the $K+1$ points marking these segments are used for identifying the event locations. For requirements of each application, we adjust the number of event functions in each phoneme.

New method for modeling of the event function

To control the event functions, the event functions should be modeled. The MRTD algorithm enforces the "well-shapedness" property of event functions. That is, the event functions in the MRTD are monotonic during the transition from one event towards the next. In addition, the MRTD method employs the restricted second order TD model, in which only two event functions at any moment of time can overlap and all event functions sum up to one [113]. Therefore, to model the event function, polynomial fitting for the event function is performed by using the nonlinear least square method as follows.

$$Z = -\left(\frac{X}{c}\right)^{M} + e \tag{3.15}$$

where e is the maximum value of ϕ, and e is equal to 1. Z is equal to 0 when

$$X = c \tag{3.16}$$

where c is the duration of two consecutive events. The polynomial fitting was done in $0 \leq \phi \leq 1$. The value of M indicates slope of event function. Shape of the event function can be changed according to the values of c and M. As a result, it is possible to control the event function. Figure 3.10 shows an example of modeled event function by the proposed method for an event function extracted by MRTD.

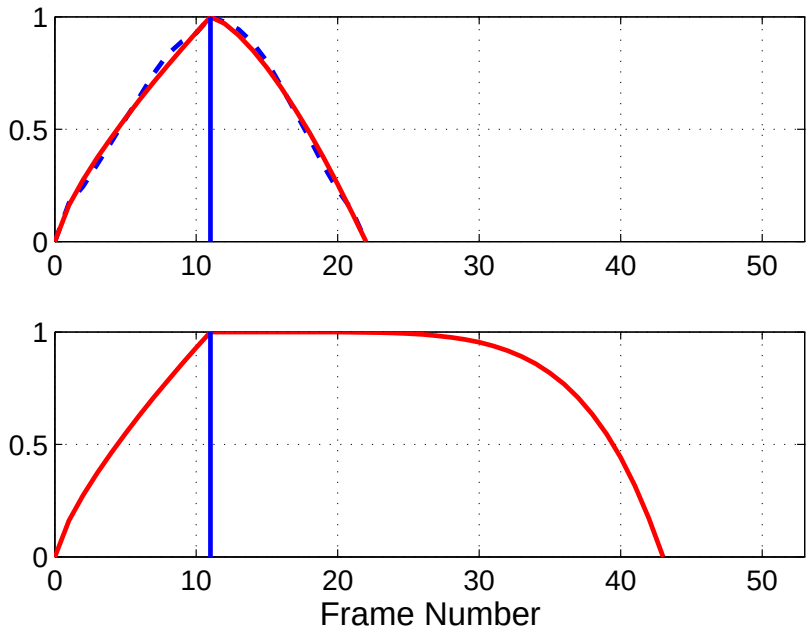

Figure 3.10: A fitted curve (the solid line) using the non-linear least square method for an event function (the dashed line) (top), and a time-scale modification of this event function (bottom).

3.3.4 Proposed modeling of the speech spectral sequence

One of the advantages of TD is that it ensures the smoothness of modified speech signals. However, if event targets are represented by linear prediction (LP) coefficients, such as LSF coefficients, we meet difficulties in performing spectral modification. To overcome these drawbacks, we use spectral-GMM parameters [109, 166, 167, 170] to model each event target. Based on TD and spectral-GMM, we propose a new modeling of speech spectral sequence for speech modification which can deal with these two drawbacks of conventional spectral modification methods, the insufficient smoothness of the modified spectra between frames, and the ineffective spectral modification. In addition, since glottal and vocal tract information are not independent, modifying them separately often degrades the quality of modified speech signals. Therefore, a high-quality analysis/modification/synthesis framework, STRAIGHT, is utilized in this study.

Proposed modeling

The processing flow of our proposed method is as follows, and is shown in Figure 3.11.

First, STRAIGHT [65] decomposes input speech signals into spectral envelopes, F0, and AP. Since the spectral envelopes can be further analyzed into LSF parameters, MRTD [113] is employed in the next step to decompose the LSF parameters into event targets and event functions. To identify the event locations, we have presented a new method based on the phoneme as described in 3.3.3. The event functions are modeled by using Eq. (3.15). Since the event targets are valid LSF parameters [113], the spectral envelope of each event

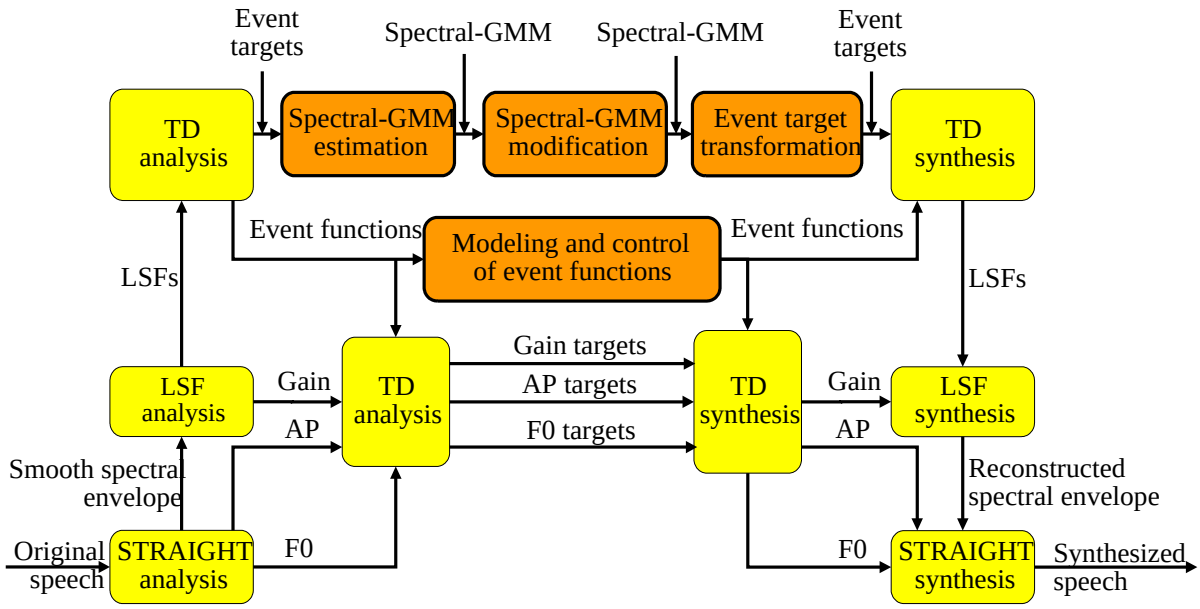

Figure 3.11: Diagram of proposed modeling of speech spectral sequence based on temporal decomposition and Gaussian mixture model.

target can be restored, and then the spectral envelopes are converted to spectral-GMM parameters. By using spectral-GMM parameters to model the event targets, we can flexibly perform modifications of the event targets. The same event functions evaluated for the spectral parameters are also used to describe the temporal pattern of the F0, gain, and AP. The modified event targets are then re-synthesized as modified LSF by TD synthesis. In the following step, the modified LSF parameters are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the synthesized speech. Note that our proposed method is integrated with the STRAIGHT method, it therefore can use the merits of the STRAIGHT to modify the F0.

3.3.5 Merits of our modeling in speech modification

Spectral modeling is very important, since it determines that which and how features of speech signals can be modified. In the previous subsection, we propose the modeling of speech spectral sequence for spectral modification. In this subsection, we present how to perform time-scale modification and spectral modification, two of the three most popular kinds of speech modification.

Time-scale modification

In our method, spectral dynamics is represented by using event functions of the TD algorithm. Moreover, event functions are modeled by using the polynomial fitting technique as in Eq. (3.15). Since event functions can be seen as temporal evolution of spectral parameters, to perform time-scale modification, we need to change length of each event function. From Eq. (3.15), we can modify the duration of the speech segment by changing

the value of c. We modify the values of M to alter the slope of the event function. By changing the values of c and M, we can control the evolution of all important parameters of speech signals (i.e. F0, gain, AP, and spectral parameters). Therefore, it enhances the quality of modified speech. An example of time-scale modification of an event function is also shown in Figure 3.10. In this example, to show the flexibility of our proposed method, we only altered the shape of the right side of the event function by the modification factors of c and M 3 and 3.9 times, respectively. In addition, we can ensure the smoothness of modified speech between frames by controlling the shape of event functions.

Spectral modification

Formants are some of the most important of speech. Controlling formants can change characteristics of speech. In this subsection, we discuss how to modify amplitudes of spectral peaks. Modification of spectral frequencies is discussed in Chapter 5. Spectral-GMM parameters meet most requirements of a spectrum representation for speech modification. Those are preciseness and stability, which ensure to model and reconstruct the speech spectral envelope precisely, locality (without affecting the intensity of frequencies further away from the point of manipulation), and flexibility and ease of manipulation. We now present a new method to control amplitude of spectral peaks. The algorithm for modifying amplitudes of spectral peaks is described as follows, and is shown in Figure 3.12.

First, we determine the spectral peak which need to be modified. When estimating spectral-GMM parameters, we do not apply any constraints to ensure the one-to-one correspondence between spectral peaks and Gaussian components. That is why there are some Gaussian components contributing to the same spectral peak. In our algorithm, we merge these Gaussian components to one Gaussian component by using the following conditions.

- If linear sum of two Gaussian component, i and j, have more than one peak, we consider that these Gaussian components are independent.
- If linear sum of two Gaussian components, i and j, have only one peak, we consider that these Gaussian components are dependent, and we merge these Gaussian components to one Gaussian component by using following equations.

$$\mu_{ij} = \frac{\alpha_i \mu_i + \alpha_j \mu_j}{\alpha_i + \alpha_j} \tag{3.17}$$

$$\sigma_{ij}^2 = \frac{\alpha_i(\sigma_i^2 + (\mu_{ij} - \mu_i)^2) + \alpha_j(\sigma_j^2 + (\mu_{ij} - \mu_j)^2)}{\alpha_i + \alpha_j} \tag{3.18}$$

  where $\mu_i, \sigma_i, \alpha_i$ and $\mu_j, \sigma_j, \alpha_j$ are the means, standard deviations, and mixture weights of the Gaussian components i and j, respectively.

The above iterative process continues until there is only one Gaussian component used to model the desired spectral peak. We easily can control properties of this spectral peak,

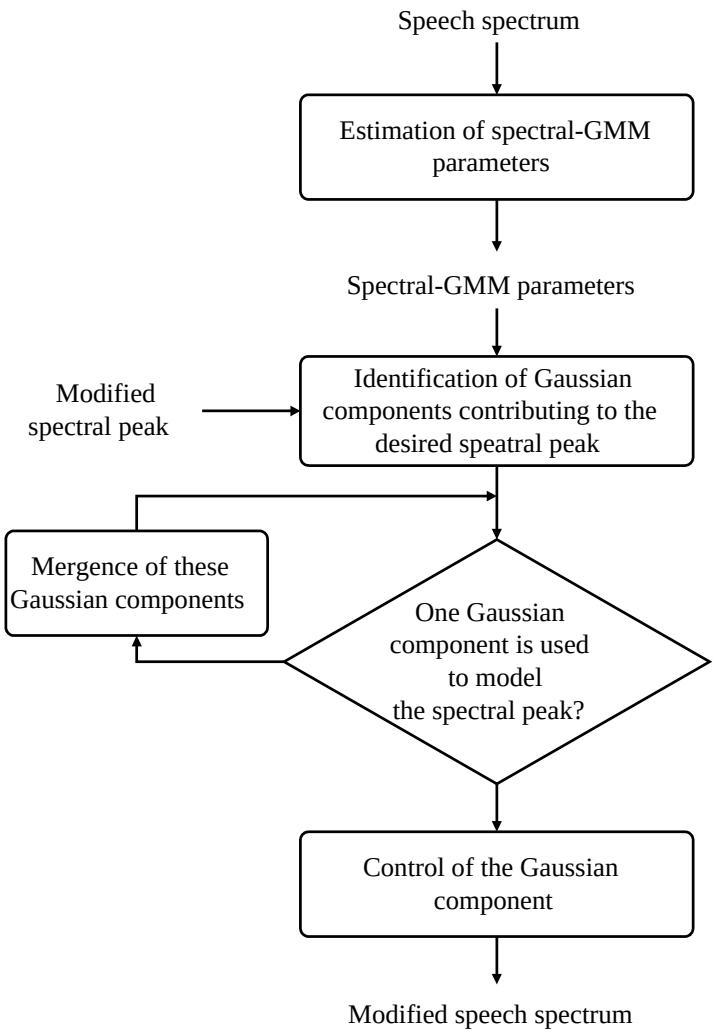

Figure 3.12: Diagram of our proposed method for modifying amplitudes of spectral peaks

such as modifications of amplitude or bandwidth. An example of amplitude modification of the first three formants is shown in Figure 3.13. In this example, we increase the amplitudes of the first three formants by 2, 3, and 2.5 times, respectively.

In summary, one of advantages of modeling of the speech spectral envelope using spectral-GMM parameters is that we can flexibly control the speech spectral envelope. This is a major advantage of spectral-GMM parameters if comparing with conventional representations of speech spectral envelope, such as LP coefficients or non-parametric representations.

3.3.6 Experiments and results

In our proposed modeling of the speech spectral sequence, since we use spectral-GMM parameters to model each event target, the order of LSFs has to be high enough to precisely restore the spectral envelope. Via a small experiment, by calculating the average log spectral distortion (LSD) between STRAIGHT spectra and the spectral envelopes

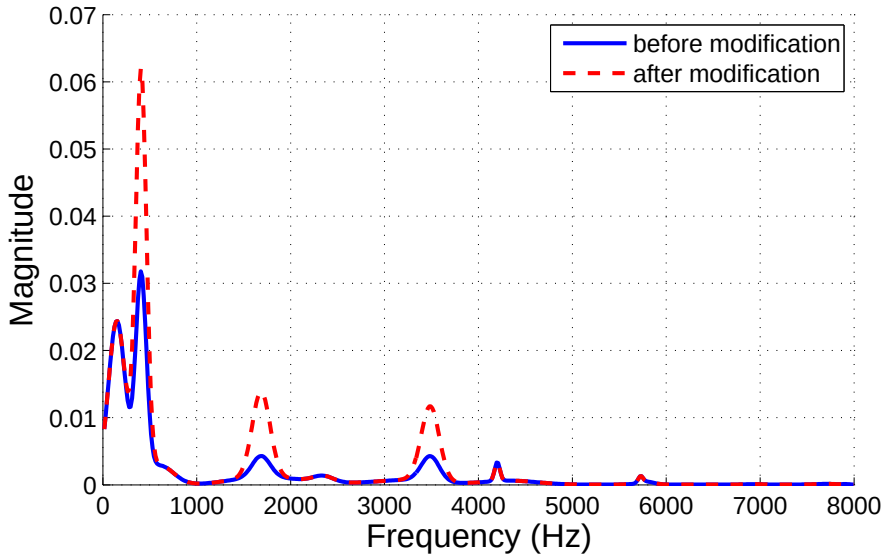

Figure 3.13: Example of the amplitude modification of the first three formants by 2, 3, and 2.5 times, respectively. 24 Gaussian components (M=24) are used to model this speech spectral envelope.

restored from LSFs with different orders in a set of 250 sentence utterances of the ATR Japanese speech database [2] at sampling frequency of 16 kHz, we chose the LSF order of 40 in this framework. With this order, the average LSD is smaller than 1 dB.

A set of 100 sentence utterances of the ATR Japanese speech database [2] was selected as the speech data. This speech dataset is spoken by 4 speakers (2 male & 2 female) resampled at 16 kHz sampling frequency.

To evaluate the performance of our proposed analysis/synthesis method, we compare the quality of synthesized speech restored by our method with that of the framewise-GMM method [167]. In the framewise-GMM method [167], we estimated the spectral-GMM parameters from the STRAIGHT spectrum frame by frame. In this section, we used the perceptual evaluation of speech quality (PESQ) score (ITU-T P.862) to evaluate the quality of synthesized speech. The PESQ uses a sensory model to compare the original, unprocessed signal (reference signal) with the degraded signal from a network or an analysis/synthesis system. The PESQ scores are calibrated using a large database of subjective tests. Having high correlation ($\rho > 0.92$) with subjective listening tests, the PESQ can be used reliably to predict the subjective speech quality of codec in a very wide range of conditions, including those with background noise, analogue filtering, and/or variable delay [128]. The score of PESQ ranges from -0.5 to 4.5. The higher the score, the better the perceptual quality. We used the original sounds as the reference signals, and the synthesized utterances restored by STRAIGHT, the framewise-GMM method, and our proposed method as the degraded signals. Since the average number of Gaussian components in our proposed method is 9.2, we chose 9 Gaussian components to model each speech spectral envelope in the framewise-GMM method. The average PESQ results are shown in Table 3.5. These results indicate that the quality of synthesized speech of our

Table 3.5: Average PESQ for analysis/synthesis methods.

STRAIGHT method	3.56
Framewise-GMM method	3.03
Proposed method	3.32

proposed method is better than that of the framewise-GMM method. Moreover, in our proposed method, both the speech spectra (i.e. the spectral evolution and the speech spectral envelope) and the temporal evolution of the excitation parameters (i.e. F0, gain, and AP) can be modeled, which gives the flexibility to control these parameters.

3.3.7 Conclusions

In this section, we have presented high-quality analysis/synthesis methods based on temporal decomposition for speech modification. In our proposed methods, we utilize TD to model the spectral evolution, and spectral-GMM parameters to model the event targets. Our proposed methods not only effectively describe the temporal trajectories between frames, but also flexibly model the event targets. Moreover, processing rules are more effectively applied, since we only need to process the event targets, instead of processing frame by frame, and the event targets may be associated with ideal articulatory positions. Moreover, the same event functions evaluated for the spectral parameters are also used to describe the temporal pattern of the F0, gain, and AP. We then model the event functions by using polynomial fitting. These models give the flexibility to control the speech signals in both time and frequency domains.

3.4 Summary

In this chapter, we have presented our solutions to the first issue of spectral modification, the lack of efficient spectral modelings for speech modification.

First, we have presented the two improvements of modeling of speech spectral envelope. In the first improvement, we not only model the speech spectral envelope well but also ensure a correspondence between spectral peaks and Gaussian components. In the second improvement, we use asymmetric Gaussian mixture model to model the speech spectral envelope.

Second, we have presented a new method for modeling speech spectral sequence, and have discussed the advantages of our method. In our proposed modeling of the speech spectral sequence, apart from well modeling of speech signals, it is potential to alter duration and perform efficient spectral modification.

Chapter 4

Spectral Smoothing for Concatenative Speech Synthesis based on Temporal Decomposition

The purpose of this chapter is to solve the second issue of spectral modification, i.e. the insufficient smoothness of modified spectra between frames. When producing speech sounds, the articulators of human beings move slowly, which makes speech signals change gradually. If unexpected modifications happen, there are discontinuities of modified speech, and the quality of speech degrades. This chapter presents the TD algorithm as an efficient model to control the smoothness of synthesized speech, and confirms that the reduction of discontinuities in concatenation points improves the quality of synthesized speech in concatenative speech synthesis.

4.1 Introduction

In the process of generating audible speech from a textual representation, a Text-to-Speech (TTS) system first converts text into a linguistic representation, which is then used to generate an appropriate acoustic waveform. This second step is achieved by using a speech synthesis model that describes the relationship between linguistic units and acoustic features. These speech synthesis models vary in their complexity. The first intelligible synthesizer used an approach called formant synthesis, which utilizes relatively simple models of the glottal source and vocal tract. Model parameters can be generated either by rule [72] or from a database [86]. Most aspects of speech are controllable, including the degree of articulation and characteristics of the speaker. The resulting speech is highly intelligible, but is often judged as not very natural. In an effort to increase naturalness without decreasing flexibility, researchers have increased the complexity of the speech synthesis model to take into account more physiological and physical details about the speech production process; this approach is called articulatory synthesis [131]. Unfortunately, it is difficult, in practice, to generate the high-dimensional parameter trajectories which are necessary to drive articulatory synthesis models, because the relationships between linguistic units and parameter trajectories are complicated and cannot be learned

easily. Both formant and articulatory synthesis are examples of parametric synthesis.

The most successful TTS approach to-date is called concatenative synthesis. In this approach, natural speech utterances of a single speaker must first be recorded and stored in an acoustic inventory. During synthesis, individual portions of speech are retrieved from the inventory, optionally modified, and then concatenated in the desired sequence. Intelligibility and naturalness of speech are very high in the concatenative synthesis approach [16]. However, output speech is limited by the contents of the acoustic inventory (not just the linguistic content, but also the emotional state of the speaker, degree of articulation, etc.), and inevitable concatenation errors can lead to audible discontinuities. To overcome the problems of limited content and discontinuities, researchers either significantly increase the size of the database to include more variability, or introduce additional modeling to modify and thus control the natural speech signal. In the former case, we need to spend a big part of the effort on preparing the database. It has to be correctly designed to cover all the variability of the language, and well recorded for the system to have high voice quality. These sometimes are difficult to perform, since it is time-consuming and expensive. In the latter case, models that include prosodic control of pitch and duration are common [144]. In addition to prosodic modifications, researchers have also proposed spectral modifications, for example smoothing spectral balance discontinuities at concatenation points, expressed as energies in four bands [93], smoothing formant discontinuities [22, 82, 94], and controlling the degree of articulation [159].

To improve the quality of concatenative speech synthesis, spectral smoothing methods are proposed in the literature. One important issue in spectral smoothing is to determine which circumstances smoothing should be performed. If two segments have a sufficiently close spectral match, distortion caused by smoothing may negate the performance gain. Moreover, many smoothing techniques are inappropriate for use with unvoiced speech. Another issue is to determine the best time span over which to interpolate. The fundamental frequency remains continuous if inserted data is equal to an integer number of pitch periods.

In this section, we review some basic smoothing techniques which show encouraging results.

Conkie and Isard [27] introduce the optimal coupling technique which allows the segment boundaries to move to improve the spectral match between adjacent segments. In this algorithm, we do not modify the existing speech, we only scan the boundaries of two speech units to find out which points of these units are most suitable. Spectral discontinuity measures are often used to determine the amount of mismatch. Although this technique is simple and easy to implement, it may cut an important part of a sound.

Another method also does not perform audio signal interpolation but instead mask discontinuities. The continuity effect is a psychoacoustic phenomenon that is suggested here as a possible method for spectral smoothing. When two sounds are alternated, a less intense masked sound may be heard as continuous despite being interrupted by a more intense masking sound. The sensory evidence presented to the auditory system does not make it clear whether or not the obscured sound has continued. Psychologists call this effect "closure" [20, 97]. However, this technique does not provide great spectral smoothing in all situations, and sometimes it still possesses noisy quality.

Waveform interpolation (WI) is a speech-coding technique which takes advantage of the gradual evolution of the shape of pitch-period waveforms. The WI coder operates on a frame-by-frame basis. In each segment, the pitch track is calculated and characteristic waveforms are extracted. Each characteristic waveform is typically one pitch period long, but the length may be an integer number of periods. In coding, characteristic waveforms are extracted from the original signal at regular time intervals. This method is simple, but it does not consider formant locations. However, it is useful on LP residual.

LP interpolation techniques are often used to smooth LP-filter coefficients in LP coding (LPC) and sometimes also for speech synthesis. The basic strategy is to model the speech signal as separate spectral and residual (filter and source) components and to adjust each component separately. Here, we perform LP spectral parameter interpolation in one of several domains, while the residual is interpolated using WI. This technique is regarded as one of the most effective spectral smoothing methods.

4.2 Related work

In the literature, there are many spectral smoothing methods which interpolate LP coefficients of LP-related features. In [119], Plumpe et al. propose a HMM-based smoothing technique. This technique represents a speech signal as being composed of a number of frames, where each frame can be synthesized from LSF coefficients. Each frame is represented by a state in an HMM, where the output distribution of each state is a Gaussian random vector consisting of x and Δx. The set of LSF vectors that maximizes the HMM probability is the representation of the smoothed speech output. However, a large training database is required to estimate the HMM parameters. In [119], the training database used for the experiments contains over 7,000 sentences from a single speaker. This is the limitation of this technique. Heng and Wenju [53] proposed a new kind of LPC parameters by divide the frequency into two parts: low and high frequency parts. They then use high order of LPC to model the low frequency part and low order of LPC to model the high frequency part. After representing the spectral envelope by this kind of LPC parameters, they apply this presentation to spectral smoothing. However, this method is only effective when the total of LPC order is constant. Moreover, there is a gap between two parts of frequencies, which degrade the quality of synthesized speech. To reduce concatenation mismatch, Wouters and Macon [161] propose a method which control spectral dynamics. In this approach, synthesis is performed by combining information from two tiers of speech units, denoted concatenation units and fusion units. The concatenation units specify initial estimates of the spectral trajectories for an utterance, while the fusion units characterize the spectral dynamics at the join points between concatenation units. These two unit tiers are fused during synthesis to obtain natural spectral transitions throughout the synthesized speech. However, this method needs to prepare a fusion unit for each concatenation point. Kain et al. [62] also proposed a new method of controlling spectral in concatenative speech synthesis which has same idea with the work of Wouters and Macon [161]. They smooth the trajectory of formant frequencies. In [62], it is not necessary to prepare the fusion units. Apart from that, this method considers the smoothness of energy. However, since this method use formant frequencies

as parameters to interpolate between two segments, some steps in this method need to be manually performed. Therefore, a new method for concatenative speech synthesis which is automatically performed is needed.

4.3 Proposed method

4.3.1 Overview of our proposed method

Spectral dynamics

One of issues of spectral modification is discontinuities of spectral parameters if unexpected modifications happen. One of efficient ways to solve the discontinuities of spectral parameters is control of spectral dynamics. Spectral dynamics is widely applied in many areas in speech signal processing, such as in speech recognition [124], speaker verification [14, 21, 139], speech coding [100], text-to-speech [22, 62, 161], voice conversion [34, 146, 147].

In the viewpoint of engineering, speech is a signal continuously changing in a time. Spectral dynamics, which refers to the temporal characteristics in spectral information, is a very important feature of speech. Spectral dynamics provides most information about phonetic properties of speech sounds (i.e. formant transitions). These correlations can be captured to some extent by augmenting the original set of acoustic features (static features) which dynamic features.

The dynamic features are often referred to as time derivatives or deltas. The simplest way to calculate spectral dynamics is computing the difference between the feature values of two consecutive frames.

$$\Delta y_t = y_{t+1} - y_t \tag{4.1}$$

where y_t is the spectral feature of frame t, such as MFCC, LSF, LPC. We also can calculate the spectral dynamics within several frames as follows.

$$\Delta y_t = y_{t+U} - y_{t-U} \tag{4.2}$$

where U typically takes a value of 1 or 2 (look forward and backward one or two frames).

Although time difference features have been used successfully in many systems, they are sensitive to random fluctuations in the original static features, and therefore tend to be "noisy". In [124], a more robust measure of local change is obtained by applying linear regression over a sequence of frames as follows.

$$\Delta y_t = \frac{\sum_{i=1}^{D} i(y_{t+i} - y_{t-i})}{2\sum_{i=1}^{D} i^2} \tag{4.3}$$

The delta features described in Eqs. (4.1), (4.2), and (4.3) are the first-order time derivatives. We can in turn calculate the second-order time derivatives $\Delta\Delta y_t$ (referred to as delta-deltas) from the first-order time derivatives.

TD as a framework for modeling spectral dynamics

Studying on modeling of spectral dynamics has been attractive sciences and engineers in the area of speech signal processing. There are two compelling reasons for carrying out dynamic speech modeling. First, mathematical modeling of speech dynamics provides an effective tool in the scientific methods of studying the speech chain. Second, advancement of human language technology, especially which in automatic recognition of natural-style human speech is also expected to benefit from comprehensive computational modeling of speech dynamics.

In the literature, a hidden Markov model (HMM) is well-known for being a typical model for modeling spectral dynamics. A HMM can be used to represent a given speech segment in a stochastic manner. However, the HMM model requires a large database for modeling the spectral dynamics. This requirement is not suitable for this application, since we have a limited data.

In 1983, Atal proposed a method based on the temporal decomposition of speech into a sequence of overlapping event functions and corresponding event targets [5]. Temporal decomposition (TD) can be seen as a model of speech spectral evolution where a sequence of spectral parameters is described as a linear combination of a limited set of vectors (event targets). Event function between two event targets can be seen as interpolation of these event targets, and is a way to model transitions between successive sounds. In a TD process, we estimate event targets and event functions based on only this speech segment, and we do not need training data. Therefore, in this dissertation, we employ TD [5, 113] to control the spectral dynamics. In the following sections, we describe our proposed method based on TD to reduce the mismatch at concatenation points.

4.3.2 Proposed method

Since controlling spectral dynamics can improve the quality of concatenation speech, we propose a new method for concatenative speech synthesis based on temporal decomposition [5, 113]. Our algorithm is described as follows, and is shown in Figure 4.1.

First, STRAIGHT [65] decomposes each speech segment into spectral envelopes, F0 (fundamental frequency) information, and aperiodic indices. Since the spectral envelopes can be further analyzed into LSF parameters, MRTD is employed in the next step to decompose the LSF parameters of each speech segment into event targets and event functions. The same event function evaluated for LSF parameters are used to decompose the fundamental frequency and gain to get fundamental frequency targets and gain targets. In the ideal case, the last target of the first speech segment and the first target of second speech segment are identical. However, in concatenative speech synthesis, two event targets are often different. We need to modify these targets to smooth the transition between two speech segments. Since each event target is valid LSFs, we should modify event targets so that they become valid LSFs. In our algorithm, the modified event target is calculated by applying following equation.

$$LSF_i^{modified} = \beta LSF_i^{last\ \ ET} + (1-\beta) LSF_i^{first\ \ ET} \tag{4.4}$$

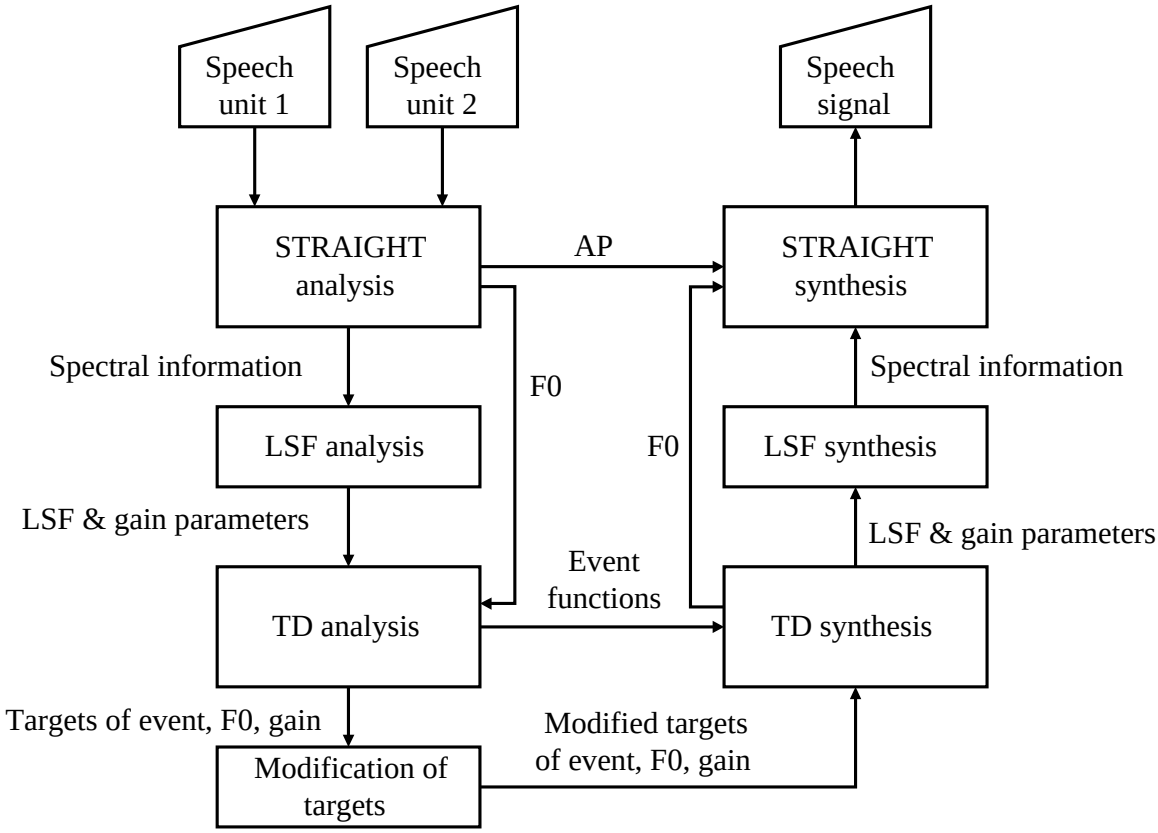

Figure 4.1: Diagram of our proposed method.

where $i = 1 \ldots P$, P is the order of LSF. The $LSF^{last\ ET}$ and $LSF^{first\ ET}$ are the LSF parameters of the last event target of the first speech segment and the first event target of the second speech segment, respectively. β is the weight factor, and satisfies $0 \leq \beta \leq 1$. We can adjust the value of β to control the degree of modification of each concatenation part in accordance with their importance. After combination of the last event target of the first speech segment and the first event target of the second speech segment, we also modify the fundamental frequency targets, and gain targets to smooth all of the most important parameters in the concatenation point. The modified event targets, modified fundamental frequency targets, and modified gain targets are then re-synthesized as modified LSFs, modified fundamental frequency information, and modified gain information by TD synthesis, respectively. In the next step, the modified LSF parameters and modified gain information are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the synthesized speech. Note that when we modify these targets, the spectral and source information of adjacent frames around on the concatenation point are also modified, and the smoothness is ensured by the shape of the event functions.

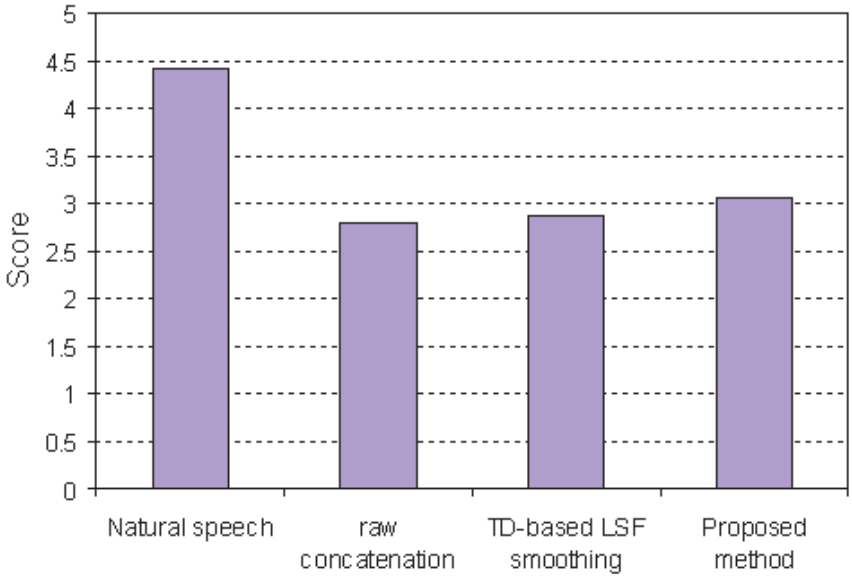

Figure 4.2: Results of subjective tests of concatenative speech synthesis.

4.4 Experiments and results

Stimuli consisted of the five Japanese vowels (/a/, /e/, /i/, /o/, and /u/) in a consonant-vowel-consonant (CVC) context. We selected a dataset consisting of five words containing the five Japanese vowels from the ATR Japanese speech database [2]. We exchanged the vowels in these words, and smoothed the borders by using different methods. Some synthesized words were meaningless. The main analysis conditions for these experiments are as follows. Sampling frequency is 16 kHz, and the order of LSF is 32.

To evaluate the performance of our proposed method, we performed subjective experiments regarding speech quality. We compared our proposed method with two other methods. In the first method, we only concatenated speech segments together (the raw concatenation method), and in the second method, we only smoothed spectral parameters by using TD, but we did not smooth F0 and energy (TD-based LSF smoothing method). We presented the synthesized sounds to eight Japanese graduate students with normal hearing ability, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Results of the subjective tests are shown in Figure 4.2. These results indicate that the quality of words modified by using our proposed method is the best in all three methods. Figure 4.3 shows the parts of the LSF contours before and after modification at the concatenation points by replacing the vowel *"u"* in the word *"takumi"* by the vowel *"e"* in the word *"jiten"*.

4.5 Conclusions

In this chapter, we have presented a framework for controlling the smoothness of modified speech. In our solution, we employ the TD algorithm to control the spectral dynamics, one of the main sources of discontinuities in synthesized speech. On the application side,

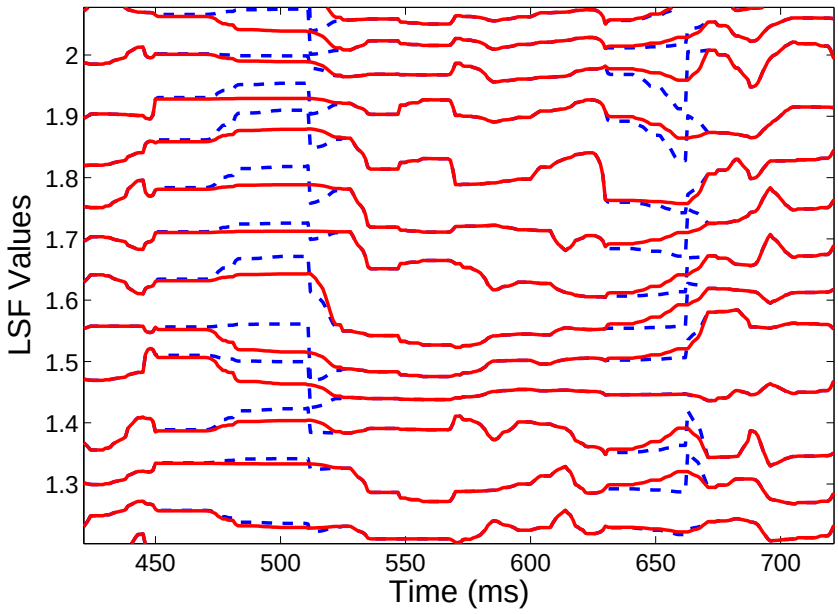

Figure 4.3: Parts of the LSF contours before and after modification at the concatenation points by replacing the vowel *"u"* in the word *"takumi"* by the vowel *"e"* in the word *"jiten"*. The dot line indicates the LSF contours of the two speech units before modification. The solid line indicates the LSF contours of the two speech units after modification by using our proposed method.

we propose a new interpolation method to smooth the discontinuities between speech segments at concatenation points. Our proposed method not only deals with the spectral discontinuities but also solves the discontinuities of fundamental frequency and energy. Experimental results prove that the TD algorithm is utilized as an efficient model to ensure the smoothness of synthesized speech, and also confirm the effectiveness of our proposed method in concatenative speech synthesis.

Chapter 5

Rule-based Approach to Spectral Modification

The purpose of this chapter is to solve the third issue of spectral modification, i.e. the ineffective spectral modification. In this chapter, we propose a new efficient rule-based algorithm of spectral modification, which performs directly on a spectral envelope. In Chapter 3, spectral-GMM parameters are used to model the spectral envelope. Spectral-GMM parameters are related to formant information. The formant features are some of the most important parameters in characterizing speech, and control of formants can effectively modify the spectral envelope. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. This chapter first presents a new algorithm to modify spectral-GMM parameters in accordance with formant frequencies. We then apply our algorithm to two areas, emotional speech synthesis which requires modification of both formant frequency and power, and voice gender conversion which requires a large amount of spectral modification.

5.1 Introduction

Spectral modification techniques are used to perform a variety of modifications to speech spectra, such as manipulations of the formant structures, amplitude manipulations. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing. It can be applied in many areas. Spectral modification methods are a powerful technology for customizing Text-to-Speech (TTS) systems, such as by converting source features to target features [1, 143], changing a male voice into a female voice and vice versa [77], and applying to emotional speech synthesis [66]. Spectral modification techniques are often applicable to automatic speech recognition tasks [49], and speech enhancement [15].

The basic idea of spectral processing is to convert a time-domain digital signal into its representation in time-frequency domain. Most of the approaches start by developing an analysis/synthesis technique from which the speech signal is reconstructed with minimum loss of sound quality. Then, the main issues have to be resolved: what kind of

representation and which parameters are chosen for the application of the desired speech processing. The challenge of spectral modification is to modify the spectral/acoustical features without degrading the speech quality.

A variety of spectral modification methods have been discussed in the literature. They can be classified into two popular approaches: linear prediction (LP)-based methods [94, 99] and frequency warping methods [151]. LP-based methods are often affected by the pole interaction problem suffered by pole modification techniques. An iterative algorithm for overcoming pole interaction during formant modification was developed by Mizuno et al. [94]. This method produces spectral envelopes with desired formant amplitudes at the formant frequencies. However, the amplitude and bandwidth of each formant cannot be independently modified, since each formant's bandwidth is dependent on the magnitude of the corresponding pole. Recently, a method for directly modifying formant locations and bandwidths in the line spectral frequency (LSF) domain has been developed [99]. We refer to the method in [99] as the LSF-based method. By taking advantage of the nearly linear relationship between the LSF coefficients and formants, modifications are performed based on desired shifts in formant frequencies and bandwidths. However, the main drawback, i.e. the lack of control over the spectral shape, has not been solved. Frequency warping methods, such as by Turajlic et al. [151], give high quality of modified speech. However, frequency warping methods meet difficulties in modifying spectral peaks, such as preserving shapes of peaks, and emphasizing spectral peaks around 3 kHz in transformation of speaking voice into singing voice, since they do not estimate spectral peaks. Moreover, frequency warping methods do not allow formants to merge or split, which is often desired in formant modification processes [78].

In addition, some methods mentioned above [94, 99] only mention how to modify the speech spectral envelope in a frame, and they [94, 99, 151] rarely deal with constraints between frames after modification. This limitation may cause a discontinuity problem between adjacent frames. As a result, there are some clicks in the modified speech when unexpected modifications happen in some frames. Moreover, Knagenhjelm and Kleijn [73] point out that spectral discontinuities between adjacent frames are one of the major sources of quality degradation in speech coding systems. Therefore, this problem should be solved to enhance the quality of modified speech.

In this chapter, we employ a new modeling of speech spectral sequence based on temporal decomposition (TD) [5, 113] and spectral-GMM [166, 167, 170] in Section 3.3. In this speech spectra model, we employ the modified restricted temporal decomposition (MRTD) algorithm [113] to model the spectral evolution. We then use spectral-GMM parameters [167] to model each event target. Note that the spectral-GMM parameters used here are to approximate a spectral envelope, which are different from those often used to model the distribution of acoustic features in state-of-the-art methods for voice conversion. The question is now how to modify the event targets. In the next section, we present a new algorithm to answer this question.

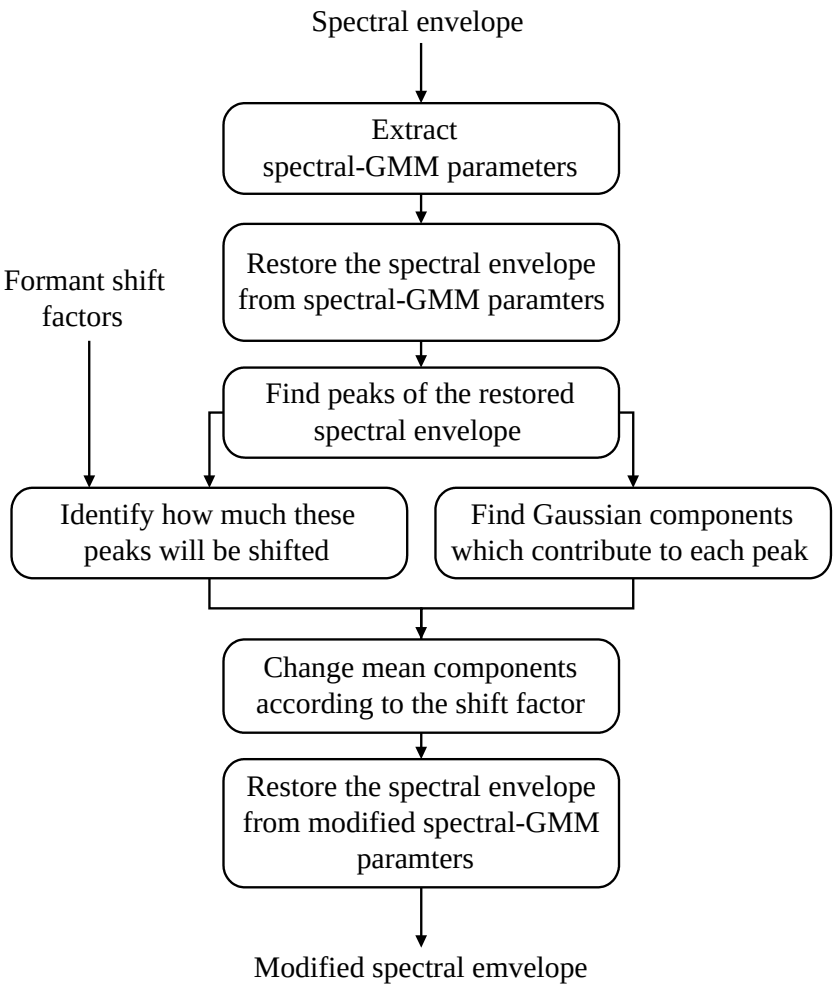

Figure 5.1: Block diagram of our spectral modification algorithm.

5.2 Proposed spectral modification algorithm

Control of formants is an effective way to perform modification of a speech spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. When estimating spectral-GMM parameters from a spectral envelope, we just try to minimize the distance between the histogram of the spectral envelope and the Gaussian mixture model. As a result, there may be some components which contribute to one peak of the spectral envelope restored from the spectral-GMM parameters, which make it difficult to modify the spectral-GMM parameters.

In this section, based on the geometric characteristic of the Gaussian distribution, we propose a new algorithm for modifying GMM parameters in accordance with formant frequencies. The spectral modification algorithm is described as follows, and is corresponding to Figure 5.1.

We first extract spectral-GMM parameters from the smooth spectral envelope. In the next step, we find the peaks of the spectral envelope reconstructed from the spectral-GMM parameters. Since not all these peaks are formants, we have to decide how much these peaks are shifted. For spectral modification, the first formants are most important, and often considered for modification. In this study, we also focus on modifying factors related to the first four formants. We isolate spectral regions of the input signal by dividing it into four non-overlapping bands (0 - 800 Hz, 800 - 2500 Hz, 2500 - 3500 Hz, 3500 - sampling frequency/2 Hz) which cover the first four formant frequency ranges [62]. The scaling factor of each peak is the scaling factor of the formant to which the peak belongs. Based on the geometric characteristics of normal distribution, i.e. the empirical rule, we find which Gaussian components contribute to this peak. If this peak is located between $[\mu_m - 3\sigma_m; \mu_m + 3\sigma_m]$, where μ_m is the mean and σ_m is the standard deviation of Gaussian component m, we regard Gaussian component m as contributing to this peak. We shift the mean parameter of this Gaussian component by the scaling factor of this peak. In this algorithm, we only modify the mean parameters of Gaussian components, and we do not modify the other parameters of Gaussian components (i.e. standard deviations and mixture weights). Note that mean parameters are sorted in ascending order, and every mean parameter is shifted only once. After shifting the mean parameters of the Gaussian components, we reconstruct the modified spectral envelope. Consequently, we can independently modify each spectral peak. An example of our proposed algorithm applied to a spectrum is shown in Figure 5.2. For comparison with our method, an example of formant modification of the LSF-based method [99] is shown in Figure 5.3. In the LSF-based method, since attributes of a formant depend on properties of a conjugate pole pair, when we change formant frequencies, amplitudes of a speech spectral envelope also change. On the contrary, we can control the spectral shape using our method.

5.3 Experiments and results

In this section, to evaluate the effectiveness of our proposed algorithm spectral modification, we investigate it in two areas, emotional speech synthesis which requires modification of both formant frequency and power in Subsection 5.3.1, and voice gender conversion which requires a large amount of spectral modification in Subsection 5.3.2. In the both applications, we employ our modeling of speech spectral sequence in Subsection 3.3. Note that we employ the local minima of the spectral feature transition rate (SFTR) based on LSF parameters [69, 102] to automatically identify the event locations, and the spectral-GMM parameters [164, 167] to model each event target. We then apply our algorithm of spectral modification to modify event targets.

5.3.1 Application to emotional speech synthesis

In this subsection, we investigate our spectral modification method for emotional speech synthesis, where formant frequencies are shifted by small scaling factors (below 8 percent), and power envelopes need to be modified.

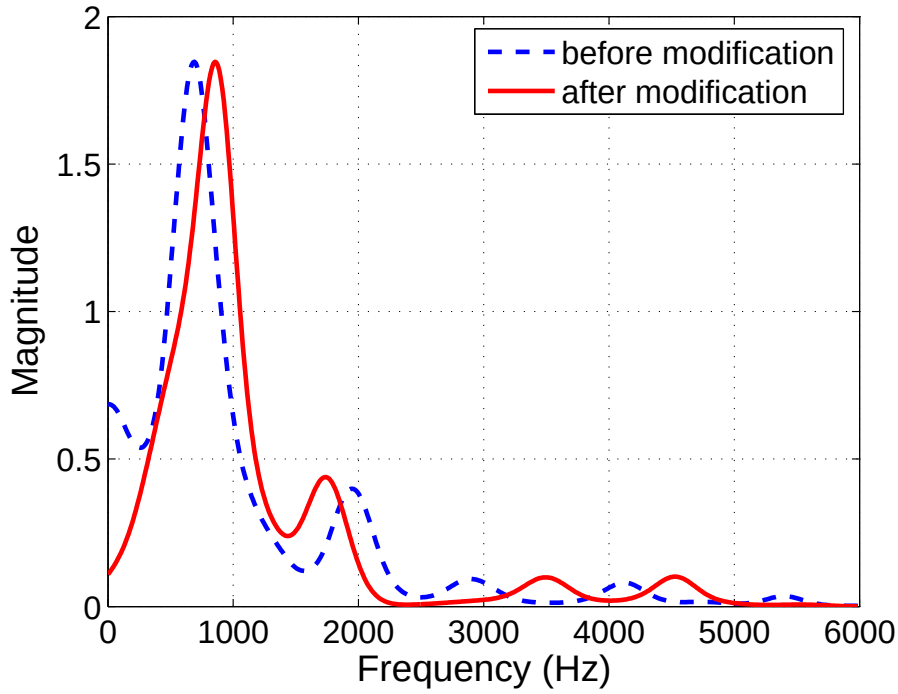

Figure 5.2: Example of our spectral modification algorithm applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$.

Speech is one of the most natural methods for human beings to convey linguistic contents. In addition, it is also one of the useful methods for conveying emotion. Speech is probably the only method that can convey linguistic contents and the speaker's emotion simultaneously. In synthesized speech databases, there are two important requirements, i.e. intelligibility and naturalness. Therefore, the need for increasing naturalness becomes more palpable. One of the aspects of naturalness most obviously missing in synthetic speech is appropriate emotional expressivity. Attempts to add emotion effects to synthesized speech have been attractive to many researchers.

Many earlier studies of expressive speech only focused on statistical correlations between expressive speech and acoustic features without taking into account the fact that human perception is vague rather than precise [58]. Huang and Akagi [57, 58] propose a multi-layer approach to modeling perception of expressive speech. Unlike most other studies that deal with the direct relationship between emotional speech and acoustic features, this model consists of three layers, emotional speech, semantic primitives, and acoustic features. This model is a rule-based conversion system, and therefore it is necessary to control each parameter independently.

In [57, 58], it was necessary to modify both power envelopes and formants. In the standard spectral modification techniques, such as [99], when formant frequencies are shifted, the magnitude of the speech spectrum is also changed accordingly. It is difficult to independently modify both power and formant frequencies with the defined scaling factors. To overcome this drawback, we employ our spectral modification method. Since our

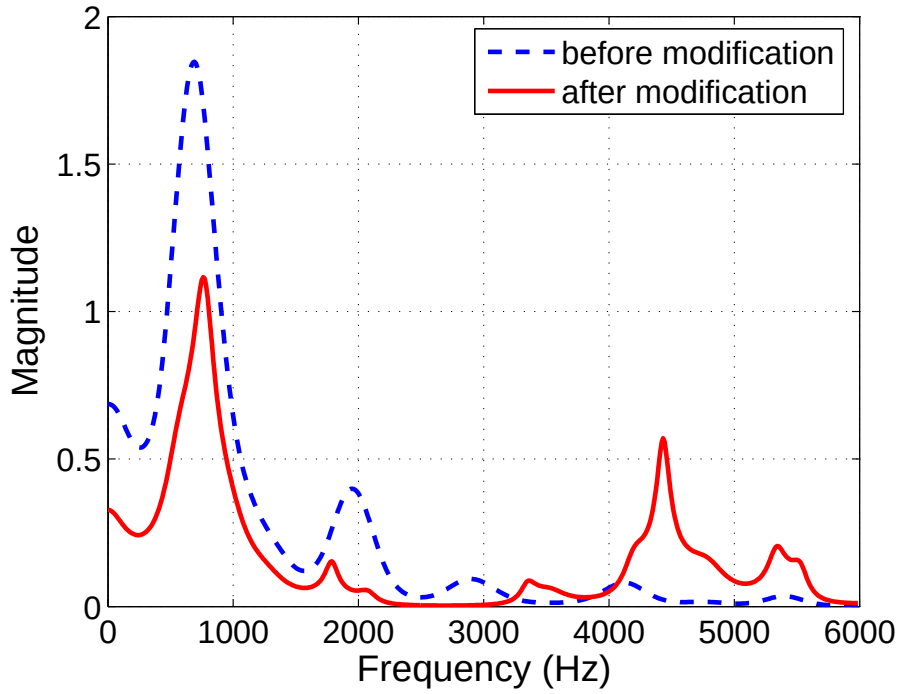

Figure 5.3: Example of formant modification algorithm of the LSF-based method [99] applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$.

method uses spectral-GMM parameters to directly model and modify the spectral envelope, the magnitude of the spectral envelope is almost the same when formant frequencies are shifted, and each parameter's value can be modified independently. In addition, the smoothness of synthesized speech is ensured by using TD.

To verify the effectiveness of our spectral modification method, we conducted a listening experiment to compare it with the LSF-based method [99], which enabled a high level of control of formant characteristics. Both the LSF-based method and our spectral modification method had been applied in [57, 58], while other processes and morphing rules were kept the same. A neutral utterance was used to morph emotional utterances, i.e. cold anger, hot anger, sadness, and joy. The analysis conditions are listed in Table 5.1.

Table 5.1: Analysis conditions for experiments of emotional speech synthesis.

STRAIGHT	Sampling frequency Window length Window shift FFT points	22.05 kHz 40 ms 1 ms 1024
LSF-based method	LSF order	24
Proposed method	Gaussian components	24

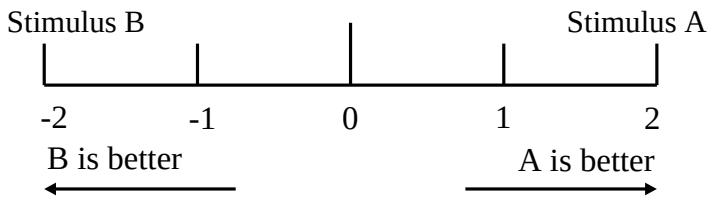

Figure 5.4: Evaluation measure of Scheffe's paired comparison (five grades: -2, -1, 0, 1 and 2).

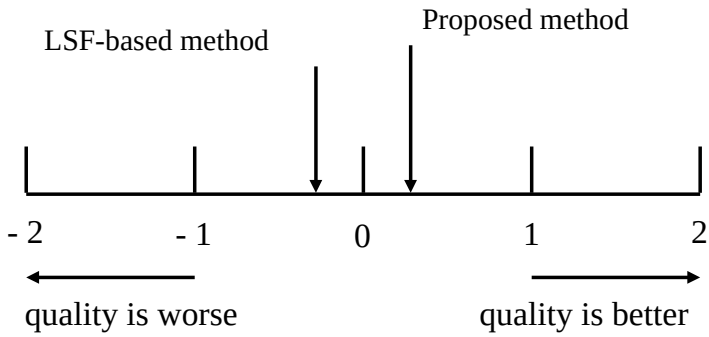

Figure 5.5: Subjective listening results for emotional speech synthesis.

The subjective test was carried out using Scheffe's method of paired comparison [130]. In this subsection, five grades from -2 to 2 were used, as shown in Figure 5.4. Eight Japanese graduate students known to have normal hearing ability were recruited for the listening experiment. Paired stimuli A and B were presented to each listener, and listeners were asked to grade stimuli according to his/her perception of speech quality. Experimental results are shown in Figure 5.5. According to a two-tailed t-test, these results are statically significant at a 95% confidence level (p-$value$= $9.2 \cdot 10^{-3}$). These experimental results also indicate that the speech quality of our proposed method is better than that of the LSF-based method [99]. In this application, since scaling factor is small (less than 8 percent), the difference of the results between the LSF-based and TD-GMM methods is small.

5.3.2 Application to voice gender conversion

In Subsection 5.3.1, our proposed spectral modification method was effectively applied to shift the small formant frequencies, below 8 percent. In this subsection, we explore the effectiveness of our spectral modification method in a voice gender conversion (VGC) system which requires much spectral modification, about 20 percent.

The aim of voice gender conversion is to modify female (male) speech so that it sounds as if it was spoken by a male (female). Voice gender conversion has applications in voice output systems such as text to speech synthesis, multimedia voice applications, or in voice gender normalization for improved speech compression or recognition. The voice

gender conversion challenge is to convert the gender-related parameters of the speech signal without affecting smoothness and naturalness.

A typical male vocal tract is about 17.5 cm in length (i.e. from vocal cords to lips) while that of a female is about 15.2 cm. The adult male larynx is about 1.2 times the size of that of the female. During puberty the male larynx undergoes a change in shape (Adam's apple protrusion) such that the adult male vocal cord membrane length reaches about 1.6 times that of the female. Analysis of male and female voiced utterances shows that the female formant frequencies are about 15 % higher than male. This difference is in close agreement with the male-to-female vocal tract length ratio. Female pitch is generally about 1.7 times that of male. This difference is mainly attributed to the difference in vocal cord membrane length although other factors such as male/female differences in the way in which the cords open and close are believed to be relevant also [145]. A detailed explanation of the speech production mechanism can be found in [132].

For a long time, it was believed that pitch was the dominant cue in voice gender perception. However, Childers and Wu [26] showed that grouped formant information gave a higher automatic gender distinction success rate than pitch information. Therefore, both the glottal and vocal tract related features of the source speech signal need to be modified in voice gender conversion systems.

A variety of approaches to voice gender conversion have been discussed in the literature. Most voice gender conversion methods are based on a parametric source-filter model of speech production [6, 59, 77]. In [6] and [77], formant modification is done by linear frequency scale mapping applied to the spectral envelope. This does not reflect accurately the frequency difference between male and female, and the quality of converted speech is not very natural. To overcome the disadvantages of linear frequency scale mapping, Jung et al. [59] refined the method proposed in [77] by splitting the speech signal into two complementary frequency bands to separate F4 from the other formants, and modifying each sub-band with different formant scaling factors. However, this method still uses LP (linear prediction) coefficients to represent and modify the spectral envelope. Due to the limitation of standard LP-based techniques in independently modifying important formant characteristics such as amplitude and bandwidth, the quality of speech is not enhanced.

In addition, all methods mentioned above modify spectral envelope and fundamental frequency frame by frame, and rarely apply any constraints between frames. When there are unexpected modifications in some frames, the modified speech may be not smooth. As a result, there are some clicks in the converted speech, which leads to degradation of speech quality.

Our perception of spoken-voice gender relies heavily on the phonation or voicing process, which is associated mainly with vowel sounds. We first extracted the fundamental frequencies, and the first four formant frequencies from the five Japanese vowels spoken by two speakers (one male & one female) in the ATR Japanese speech database [2]. We then used these values to formulate the scaling factors for our VGC system. In this subsection, we used labeled data of each utterance to identify the distance between an event location and vowels. The scaling factors for an event target is the scaling factors of the vowel which was nearest to this event target.

Table 5.2: Analysis conditions for experiments of VGC system.

STRAIGHT	Sampling frequency Window length Window shift FFT points	12 kHz 40 ms 1 ms 1024
LSF-based method	LSF order	14
Framewise-GMM method	Gaussian components	14
Proposed method	Gaussian components	14

To evaluate the performance of our proposed system, we conducted a listening experiment. We compared the performance of our system with the performance of two other systems. All three systems used STRAIGHT to modify fundamental frequencies. In the first system, the LSF-based method [99] was employed to modify formant frequencies (STRAIGHT+LSF). In the second system, speech spectra were modified frame by frame using only the framewise-GMM method, without using TD (STRAIGHT + framewise-GMM method).

A set of 50 sentence utterances of the ATR Japanese speech database was selected as the speech data. This dataset spoken by 2 speakers (one male & one female) was re-sampled at 12 kHz sampling frequency. The analysis conditions are listed in Table 5.2.

We randomly presented the synthesized sounds of each of six utterances which had been spoken by two speakers (one male & one female), to eight Japanese graduate students with normal hearing ability, and asked them to identify the gender of the person who was speaking, and to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). The average scores are shown in Table 5.3. When comparing our proposed method (STRAIGHT + TD-GMM) with the first method (STRAIGHT + LSF), a two-tailed t-test at a 95% confidence level shows that the speech quality of our proposed method is superior to that of the first method for both kinds of conversions ($p\text{-}value = 1.0 \cdot 10^{-4}$ for male to female conversion, and $p\text{-}value = 1.7 \cdot 10^{-5}$ for female to male conversion). In this application, since scaling factor is large (more than 20 percent), the difference of the results between the LSF-based and TD-GMM methods is large. The LSF-based method cannot give acceptable voice quality, since the quality of most converted speech signals is diminished by a discernible buzzy sound. Our proposed method produces better voice quality than the other methods. The speech quality of the second method (STRAIGHT + framewise-GMM) and our proposed method are almost equivalent (*p-value* of a two-tailed t-test at a 95% confidence level for male to female and female to male conversions are 0.7529 and 0.6802, respectively). According to the experimental results in Subsection 4.1, our proposed method is better than the framewise-GMM method in terms of spectral modeling. In this application, there are no reference speakers, and the modified speech of the framewise-GMM and our proposed methods are hardly perceptually distinguishable. The reason is that we used the same scaling factor for every frame in a vowel in this application. Therefore, the smoothness of spectra is preserved in voiced frames when modifying, and the effectiveness of TD is not shown

Table 5.3: Subjective listening results for VGC system (1) STRAIGHT + LSF (2) STRAIGHT + framewise-GMM (3) our proposed system (STRAIGHT + TD-GMM).

Type of conversion	Correct gender identification (%)			Quality evaluation score		
	(1)	(2)	(3)	(1)	(2)	(3)
Male to Female	83.3	93.8	93.8	2.73	3.15	3.19
Female to Male	100	100	100	3.10	3.58	3.63

clearly. It should be noted that both the second method and our proposed method used the algorithm in Section 5.2 to perform spectral modification.

5.4 Conclusions

In this chapter, we have presented an efficient algorithm for modifying spectral-GMM parameters in accordance with formant scaling factors. Our proposed algorithm performs spectral modification directly on the speech spectral envelope, and it produces the high quality of spectral modification. Our proposed algorithm is especially useful when we change the speech spectral envelope by large factors, while conventional methods can not make great changes. The experimental results prove the effectiveness of our proposed algorithm.

There is however an issue which still remains to be solved. In this chapter, we only change mean parameters of Gaussian components to perform spectral modification. It is well-known that amplitudes and bandwidth of spectral peaks are also important. The next stage of this research is how to change other Gaussian components (i.e. standard deviations and mixture weights) to modify amplitudes and bandwidth of spectral peaks.

Chapter 6

Statistical Approach to Spectral Modification

This chapter continues to solve the third issue of spectral modification, i.e. the ineffective spectral modification. In Chapter 5, we propose a new efficient algorithm of spectral modification which belongs to the rule-based approach. In this chapter, we improve a statistical approach. The statistical approach to spectral modification has been applied in many areas of speech technology, such as speaker adaptation [134], spectral voice conversion [23, 34, 37, 61, 74, 81, 107, 143, 148, 163], noise reduction [101]. This chapter improve the GMM-based spectral voice conversion method. In our proposed method, we improve the parameters of the converting function by adding constraints to the estimation of GMM parameters, and enhance the transformation stage by employing the model of temporal evolution of spectral parameters.

6.1 Introduction

The aim of voice conversion is to convert a speaker voice (source speaker) to sound as if it were the voice of a defined speaker (target speaker). Applications of voice conversion systems can be found in several fields, such as Text-to-Speech customization, automatic translation, education, medical aids and entertainment.

In voice conversion, we have focused on improving both speech intelligibility and speaker's identity. Until now, the quality of speech intelligibility can be acceptable, but speaker's identity of converted utterances is far from natural. Researchers study on the relationship between voice individuality and certain acoustic features. Technology of voice conversion nowadays tries to convert spectral information, fundamental frequency, energy, duration. The diagram of a typical voice conversion system is shown in Figure 6.1.

The voice conversion process can be decomposed into two stages: the training procedure and the transformation procedure. In the training procedure, we first prepare training data. Depending on the conversion technology, the training data can be parallel (text-dependent) or non-parallel data (text-independent). The parallel data is the data of the same sentences which are uttered by both source and target speakers, whereas in the non-parallel data, the spoken utterances by the source and target speakers need not

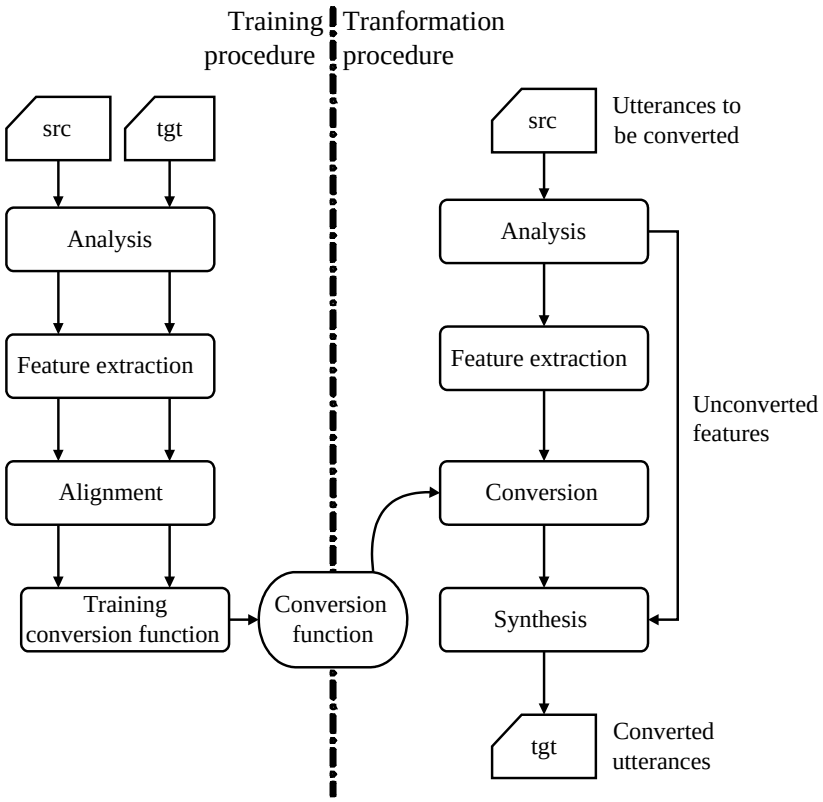

Figure 6.1: General architecture of a voice conversion system.

be the same. We then align utterances of the source and target speakers. For the parallel data, the utterances of the source and target speakers can be automatically aligned by some algorithms, such as the dynamic time warping (DTW) algorithm, the hidden Markov models (HMM) technique. For the non-parallel data, the biggest challenge is how to align the corresponding training data. After aligning the data, we apply mathematic/statistical models to obtain conversion functions. The text-dependent voice conversion gives higher quality than text-independent voice conversion. However, the parallel data is not always available in real-life applications. In the transformation procedure, the system applies the conversion function to transform new input utterances of the source speaker.

The core process in a voice conversion system is the transformation of the spectral envelope of the source speaker to match that of the target speaker. There are many statistical methods which have been proposed to implement the transform function for converting source features to target features, such as codebook-based conversion [1], neural network-based conversion [157], hidden Markov model (HMM)-based conversion [67], and Gaussian mixture model (GMM)-based conversion [23, 34, 37, 61, 74, 81, 107, 143, 148, 163]. Among those techniques, the vast majority of the current voice conversion systems focus on data-driven GMM-based transformation on the spectral aspects of conversion. Research results found in the literature have shown that the GMM-based methods can be successfully used in voice conversion. These methods are still regarded as robust and capable of producing high speech quality [163].

For the conversion of fundamental frequency, many studies [37, 38, 147] apply linear conversion techniques. They assume that fundamental frequency is characterized by a log-normal distribution. In the training procedure, they calculate the average value μ and variance σ of $logf_0$ for both source and target speakers. The conversion of fundamental frequency is done by following equation.

$$logf_0^{converted} = \mu^{target} + \frac{\sigma^{target}}{\sigma^{source}}(logf_0^{source} - \mu^{source}) \tag{6.1}$$

Some researchers also apply GMM-based methods to convert fundamental frequency, such as [52]. Other studies [36, 52] investigate the correlation between fundamental frequency and spectral information, and then apply the GMM-based methods to estimate the fundamental frequency from the converted spectral information.

Along with conversions of spectral information and fundamental frequency, energy and duration information are also investigated to convert. However, because of lack of modeling of these features, we only perform simple conversion. For example, in [37, 38], power conversion is done by constant multiplicative factors for sub-bands of speech spectrum. Durations are kept the same to the utterances of source speaker, or duration conversion is done by multiplying by the average factors of each phoneme.

In this dissertation, we focus on spectral voice conversion, one of core process in a voice conversion system. In the next parts, we formulate the problems of spectral voice conversion, and then present our solution.

6.2 Problem formulation

Among spectral voice conversion methods, GMM-based methods are widely used. Although the GMM-based voice conversion methods can give reasonably acceptable speech, the quality of converted speech is still far from natural. Three major problems remain to be solved, i.e. insufficient precision of GMM parameters, insufficient smoothness of the converted spectra between frames, and over-smooth effect in each converted frame. This section deals with the first two of the three drawbacks in a GMM-based voice conversion system, the insufficient precision of GMM parameters, and the insufficient smoothness of the converted spectra between frames.

A GMM-based voice conversion method normally includes two parts, a training procedure and a transformation procedure. In the training procedure, the methods are often based on parallel training data, where both the source and target speakers utter the same sentences. In this case, the dynamic time warping (DTW) algorithm or the hidden Markov models (HMM) technique are often used to align the two signals to extract matching source and target training vectors. Both unstable frames, which often come from transition parts between phonemes, and stable frames are used to model the distribution of acoustic features. This leads to addition of noise to the GMM parameters. To overcome this drawback, some solutions have been proposed. Kumar and Verma [74] explicitly partition acoustic space of a speaker into phones by using the phonetic alignments. After that, GMM parameters are used for finer modeling of each phone. This approach can prevent the interference of frames between phones. However, it still uses unstable

frames in each phone. Liu et al. [81] segment frames according to each phoneme, and eliminate unstable frames in each phoneme by proposing a method for identifying stable frames based on limitation of maximal variation range for the first three formant frequencies. After getting the stable frames, Liu et al. also use GMM parameters to model the distribution of acoustic features. Nguyen and Akagi [107] use event targets as spectral vectors to estimate GMM parameters, instead of using spectral parameters of aligned frames. However, all methods in [74, 81, 107] do not take into account the relations between frames when estimating the GMM parameters. The GMM parameters therefore are more precisely estimated when being considered the relations between frames.

In the transformation procedure, there are two main drawbacks, i.e. insufficient smoothness of the converted spectra between frames, and over-smooth effect in each converted frame. Until now, most voice conversion methods perform voice transformation function frame by frame. This means that to convert one frame, the information about past and future frames is not relevant. This may cause a discontinuity problem between adjacent frames when unexpected modifications happen in some frames. As a result, there are some clicks in the converted speech. Moreover, Knagenhjelm and Kleijn [73] pointed out that spectral discontinuities between adjacent frames were one of the major sources of quality degradation in speech coding systems. Some approaches to deal with this problem were discussed. To maintain a continuous transformation in consecutive frames, Chen et al. [23] smooth the converted features along the time axis by employing a median filter and a low pass filter. However, applying these filters can lead to a loss of temporal resolution, and it is a relatively crude implementation. Duxans et al. [34] include dynamic information in their GMM-based voice conversion system to take into the relations between frames. However, according to Duxans et al. [34], this method does not improve the performance of a GMM-based voice conversion system. Therefore, the discontinuity problem between adjacent frames should be solved to enhance the quality of converted speech. The problem of over-smooth effect happens in each converted frame, because of the statistical averaging operation [148]. Some works attempted to solve it [23, 148, 163], but defining solutions for this problem is beyond the scope of this section.

This section addresses two of the three main issues mentioned above, the insufficient precision of GMM parameters, and the insufficient smoothness of the converted spectra between frames. We propose a new spectral voice conversion method based on temporal decomposition (TD) [5, 113] and GMM [61, 143]. In our proposed method, we employ the modified restricted temporal decomposition (MRTD) algorithm [113] in both training and transformation procedures. We extract a set of phoneme-based features of event targets. We then use them as spectral vectors for training to take into the relations between spectral parameters in each phoneme, and to avoid using spectral parameters in transition parts. In the transformation procedure, we only need to convert event targets, instead of converting spectral parameters frame by frame, and the smoothness of converted speech is ensured by the shape of the event functions. In addition, since the fundamental frequency and vocal tract information are not independent, modifying them separately often degrades the quality of converted speech. Therefore, a high quality analysis-synthesis framework, STRAIGHT [65] is utilized in this section.

6.3 Conventional GMM-based voice conversion

As previously mentioned, the GMM-based voice conversion methods are found to be superior to other methods. In this section, we describe the basic GMM-based voice conversion method which is employed as our baseline system. A GMM-based voice conversion method often includes two parts, the training procedure and the transformation procedure.

6.3.1 Training procedure

The source speech is represented by a time sequence $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, where $\mathbf{x}_i$ is a D dimensional feature vector for the ith frame, i.e. $\mathbf{x}_i = [x_1, x_2, ..., x_D]^T$. The target speech is represented by a time sequence $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m]$, where $\mathbf{y}_j = [y_1, y_2, ..., y_D]^T$. The DTW algorithm is then adopted to align source features with their counterparts in target series to obtain feature pair series $Z = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_q]$ where $\mathbf{z}_q = [\mathbf{x}_i^T, \mathbf{y}_j^T]^T$.

The distribution of Z is modeled by Gaussian mixture model, as in Eq. (6.2).

$$p(z) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(z; \mu_m, \Sigma_m) = p(x, y) \tag{6.2}$$

where M is the number of Gaussian components. $\mathcal{N}(z; \mu_m, \Sigma_m)$ denotes the 2D dimension normal distribution with the mean μ_m and the covariance matrix Σ_m. α_m is the prior probability of $\mathbf{z}$ having been generated by component m, and it satisfies $0 \leq \alpha_m \leq 1$, $\sum_{m=1}^{M} \alpha_m = 1$. The parameters $(\alpha_m, \mu_m, \Sigma_m)$ for the joint density $p(x, y)$ can be estimated using the expectation maximization (EM) algorithm [30].

6.3.2 Transformation procedure

The transformation function that converts source feature $\mathbf{x}$ to target feature $\mathbf{y}$ is given by Eq. (6.3).

$$\begin{aligned} F(x) = E(y|x) &= \int y p(y|x) dy \\ &= \sum_{m=1}^{M} p_m(x) \left(\mu_m^y + \Sigma_m^{yx} (\Sigma_m^{xx})^{-1} (x - \mu_m^x) \right) \end{aligned} \tag{6.3}$$

$$p_m(x) = \frac{\alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx})}{\sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx})} \tag{6.4}$$

where $\mu_m = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix}$, $\Sigma_m = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix}$, and $p_m(x)$ is the probability of $\mathbf{x}$ belonging to the mth Gaussian component.

6.4 Temporal decomposition

A shortcoming of the conventional GMM-based voice conversion methods is that they do not take into account the correlation between frames in both training and transformation procedures. As a result, the precision of estimated GMM parameters is degraded, and there are some clicks in the converted speech because of discontinuous spectral contours. Therefore, we employ TD to deal with the problem.

In articulatory phonetics, speech is described as a sequence of distinct articulatory gestures, each of which produces an acoustic event that should approximate a phonetic target. Due to the overlap of the gestures, these phonetic targets are often only partly realized.

Atal [5] proposed a method based on the temporal decomposition of speech into a sequence of overlapping target functions and corresponding event targets, as given in Eq. (6.5).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{6.5}$$

where $\mathbf{a}_k$ is the speech parameter corresponding to the kth event target. The temporal evolution of this target is described by the kth event function, $\phi_k(n)$. $\hat{\mathbf{y}}(n)$ is the approximation of the nth spectral parameter vector $\mathbf{y}(n)$, and is produced by the TD model. N and K are the number of frames in the speech segment, and the number of event functions, respectively ($N \gg K$).

This chapter also employs the MRTD algorithm [113], one of improvements of the TD algorithm. In the MRTD algorithm, LSF parameters are chosen for the input of TD because of their sensitivity (an adverse alteration of one coefficient results in a spectral change only around that frequency) and efficiency (LSFs result in low spectral distortion when being interpolated and/or quantized). In this chapter, LSF parameters are extracted from spectral envelopes of STRAIGHT [65]. The STRAIGHT spectra are suitable for TD, because they are smooth in the time-frequency domain.

In a voice conversion system, we need to align utterances of source and target speakers. The original MRTD algorithm does not ensure an one-to-one correspondence between event locations and phonemic units. This makes it difficult to align parallel training data in voice conversion systems. Therefore, we use the algorithm which is described in Subsection 3.3.3. This algorithm for identifying the event locations is based on phonemes. To increase the accuracy of phoneme segmentation, this algorithm is effectively used when labeled data of utterances are available. Each phoneme is divided into four equal segments, and the five points marking these segments are used for identifying the event locations. Specially, since we can represent each phoneme by five event targets, these five event targets of each phoneme can be regarded as a "voice font". It should be noted that we can easily increase the quality of synthesized speech by increasing the number of event locations in each phoneme.

6.5 Proposed spectral voice conversion method

6.5.1 Spectral parameters

The overall shape of the spectral envelope provides an effective representation of the vocal tract characteristics of the speaker. However, the dimension of the spectral envelope is rather high, and it is not effective for direct use in a voice conversion system. We therefore often use another representation of the spectral envelope. MFCC coefficients are used to represent the spectral envelope in [74, 143, 148], while line spectral frequency (LSF) coefficients are used in [23, 34, 37, 61, 163] for the reason that LSFs have better linear interpolation attributes. In our voice conversion system, we choose LSFs for the representation of the spectral envelope. The reason for selecting LSFs is that these parameters closely relate to formant frequencies, but in contrast to formant frequencies they can be estimated quite reliably. Also, they have good interpolation characteristics, and a badly predicted component adversely affects only a portion of the frequency spectrum. Moreover, they are easily integrated with the MRTD algorithm, which uses LSFs as its input.

6.5.2 Proposed method

As previously mentioned, our proposed method focuses on spectral voice conversion, and is based on the GMM method [61, 143]. The processing flow of our spectral voice conversion system, which includes training and transformation procedures, is described as follows, and is shown in Figure 6.2.

In the training procedure, STRAIGHT [65] decomposes input speech signals into spectral envelopes, F0 (fundamental frequency) information, and aperiodic components (AP). Since the spectral envelopes can be further analyzed into LSF parameters, MRTD [113] is employed in the next step to decompose the LSF parameters into event targets and event functions. Note that the method for determination of event locations is from [133]. Each phoneme is represented by five event targets, and a vector of phoneme-based features of event targets $\mathbf{EV} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \mathbf{a}_3^T, \mathbf{a}_4^T, \mathbf{a}_5^T]$, where $\mathbf{a}_k (1 \leq k \leq 5)$ is the kth event target in each speech segment (a phoneme), can be a good vector to present the relations between event targets in a phoneme. Moreover, each event target $\mathbf{a}_k$ in the MRTD algorithm [113] is a valid LSF coefficient. An important property of LSFs $\{LSF_i\}$ is that they are ordered $(0, \pi)$, as follows.

$$0 < LSF_1 < LSF_2 < \ldots < LSF_P < \pi \tag{6.6}$$

where P is the order of LSF. To prevent a bad initialization in estimation of GMM parameters, we normalize the vectors of phoneme-based features of event targets extracted from each phoneme in utterances of source and target speakers, $\mathbf{EV_x}$ and $\mathbf{EV_y}$, as follows.

$$\mathbf{EV_x} = [\mathbf{a}_{x1}^T, \mathbf{a}_{x2}^T + \pi, \mathbf{a}_{x3}^T + 2\pi, \mathbf{a}_{x4}^T + 3\pi, \mathbf{a}_{x5}^T + 4\pi]^T \tag{6.7}$$

$$\mathbf{EV_y} = [\mathbf{a}_{y1}^T, \mathbf{a}_{y2}^T + \pi, \mathbf{a}_{y3}^T + 2\pi, \mathbf{a}_{y4}^T + 3\pi, \mathbf{a}_{y5}^T + 4\pi]^T \tag{6.8}$$

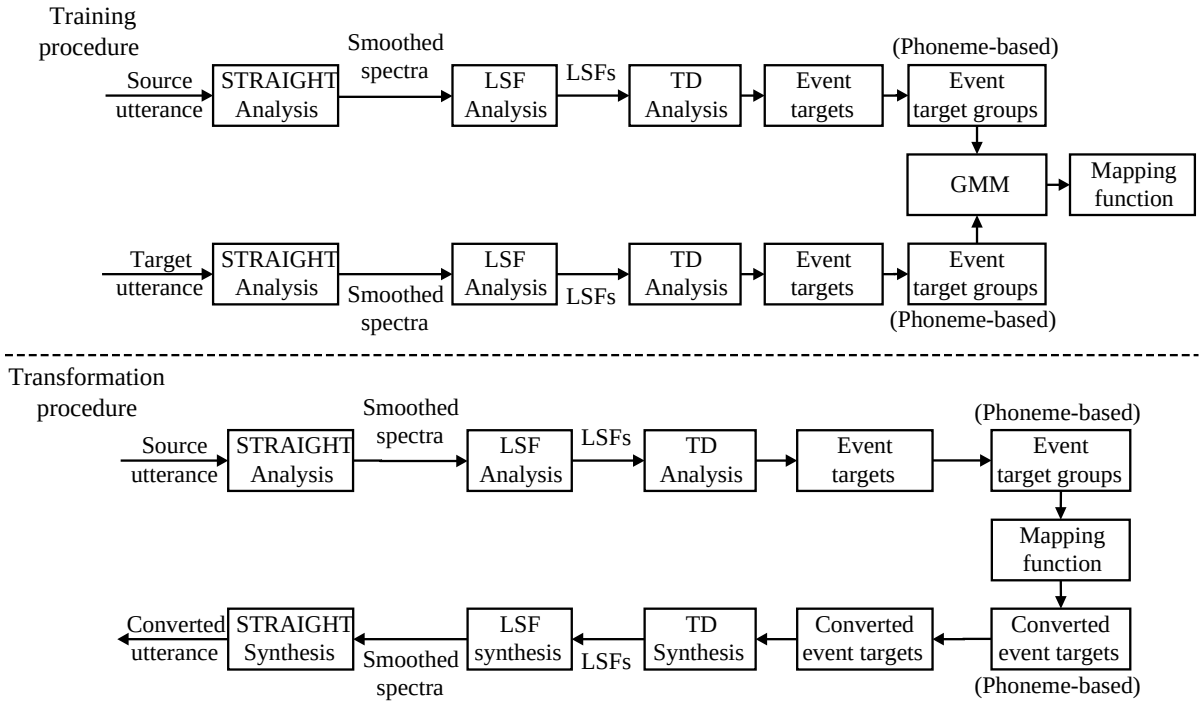

Figure 6.2: Diagram of our proposed voice conversion method training procedure (top), and transformation procedure (bottom).

where $\mathbf{a}_{xk}$, $\mathbf{a}_{yk}$ are the kth event targets in each phoneme of the source and target speakers, respectively. As a result, the vectors $\mathbf{EV_x}$ and $\mathbf{EV_y}$ are ordered $(0, 5\pi)$. All the phoneme-based features are then aligned according to each phoneme, and modeled by GMM parameters in Eq. (6.2).

In the transformation procedure, normalized phoneme-based features are also extracted from each utterance of the source speaker by using STRAIGHT and MRTD. We then convert each of the normalized phoneme-based features by using Eq. (6.3), and convert back to event targets. The converted event targets are re-synthesized as converted LSF by MRTD synthesis. In the following step, the converted LSF parameters are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the converted speech. Note that this section does not deal with prosodic, energy conversion. Therefore, to implement a complete voice conversion system, our proposed method should be integrated with some methods for prosodic, energy conversion, such as in [37, 148].

6.6 Experiments and results

6.6.1 Experimental conditions

The corpus used for the experiments is a dataset consisting of 460 sentences spoken once each by two speakers (one male & one female) in the MOCHA-TIMIT English speech database [162]. The speech data was recorded at 16KHz sampling rate. In our

Table 6.1: Analysis conditions for experiments on the voice conversion methods.

Sampling frequency	16 kHz
Window length	40 ms
Window shift	1 ms
FFT points	1024
LSF order	18
Gaussian components	20

experiments, two different voice conversion tasks were investigated: male-to-female, and female-to-male conversion. For each kind of conversion, we used 300 pair utterances for training, and 30 other pair utterances for evaluation.

To evaluate the performance of our proposed method, we performed an objective test, and also subjective evaluation experiments regarding speech quality and speaker individuality. We compared our proposed method (the phoneme-based TD+GMM method) with two other methods. The first method used for comparison is the conventional method (the GMM method) [61, 143]. The second method used for comparison also used event targets for training, and the transformation procedure was performed for each event target (the TD+GMM method). The difference between the second method and our proposed method is that the second method does not take into account the relations between event targets in training and transformation procedures. Since we only focus on spectral voice conversion, we automatically copy the prosody information and energy from the utterances of the target speaker to converted utterances. In addition, because the problem of the over-smooth effect in each converted frame is outside the scope of this section, without loss of generality, all three methods utilize the same transformation mapping function of the conventional method [61, 143] (see Eq. (6.3)). The analysis conditions for these experiments are shown in Table 6.1.

6.6.2 Objective test

We use LSF performance index PI_{LSF} for the objective test. This measure is defined as follows.

$$PI_{LSF} = 1 - \frac{E_{LSF}(t(n), \widehat{t}(n))}{E_{LSF}(t(n), s(n))} \tag{6.9}$$

where $t(n)$ represents the utterance of the target speaker, $s(n)$ represents the utterance of the source speaker, and $\widehat{t}(n)$ represents the converted utterance. $E_{LSF}(t(n), \widehat{t}(n))$ is the mean transform LSF error, and $E_{LSF}(t(n), s(n))$ is the mean inter-speaker LSF error, defined as follows.

$$E_{LSF}(A, B) = \frac{1}{L} \sum_{l=1}^{L} \sqrt{\frac{1}{P} \sum_{i=1}^{P} \left(LSF_A^{l,i} - LSF_B^{l,i} \right)^2} \tag{6.10}$$

Table 6.2: Objective results for the voice conversion methods (1) Conventional method (GMM method) (2) TD+GMM method (3) our proposed method (Phoneme-based TD+GMM method).

Type of conversion	LSF performance index		
	(1)	(2)	(3)
Male to Female	0.3692	0.3819	0.4013
Female to Male	0.3517	0.3745	0.3829

Table 6.3: MOS results for voice conversion methods (1) Conventional method (GMM method) (2) TD+GMM method (3) our proposed method (Phoneme-based TD+GMM method).

Type of conversion	Mean opinion score		
	(1)	(2)	(3)
Male to Female	3.17	3.50	3.89
Female to Male	2.67	3.13	3.67

where L is the number of frames, P is the order of LSF, and $LSF^{l,i}$ is the LSF component i in the frame l.

$PI_{LSF} = 0$ indicates that the output of the system is no more similar to the target than the source is, whereas $PI_{LSF} = 1$ indicates that the output of the system is identical to the target. In general, a higher value for PI_{LSF} suggests a better system.

The results of this objective test are shown in Table 6.2. These results indicate that the performance of our proposed method is significantly better than that of the conventional method, and also better than that of the second method (the TD+GMM method).

6.6.3 Subjective tests

Subjective tests concerning speech quality and speaker individuality were carried out. Six graduate students known to have normal hearing ability were recruited for the listening experiments.

In the test of speech quality, we randomly presented each of ten converted utterances from both kinds of conversion (male-to-female and female-to-male) to listeners, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Table 6.3 shows the average scores, which indicate that the speech quality of our proposed method (the phoneme-based TD+GMM method) is superior to that of the conventional method (the GMM method), and also better than that of the second method (the TD+GMM method).

In the test of speaker individuality, an ABX test was conducted. A represents the source speaker, B represents the target speaker, and X represents the converted speech, which supplied from each one of the three test systems. The listeners were asked to select if

Table 6.4: ABX results for voice conversion methods (1) Conventional method (GMM method) (2) TD+GMM method (3) our proposed method (Phoneme-based TD+GMM method).

Type of conversion	ABX score		
	(1)	(2)	(3)
Male to Female	4.06	4.17	4.50
Female to Male	3.44	3.56	4.00

X was closer to A or B, and adjusted the score from 1 to 5 according to his/her perception of speaker individuality when comparing. The score of 1 means that the converted speech is very similar to the source speaker, and the score of 5 means that the converted speech is very similar to the target speaker. Results of the ABX test are shown in Table 6.4. These results also indicate that the speech individuality of converted utterances of our proposed method is the most similar to the target speaker among the three methods. It should be noted that the score of the test of speaker individuality is rather high because in this section, we only focus on spectral conversion, and we therefore copied prosodic information and energy from utterances of the target speaker for all three methods.

6.7 Conclusions

In this chapter, we have continued to present another solution for the third issue of spectral modification, the ineffective spectral modification. We improve the GMM-based spectral voice conversion method, one kind of the statistical approach to spectral modification. We deal with the two of three main drawbacks of conventional GMM-based voice spectral conversion methods, the insufficient precision of GMM parameters, and the insufficient smoothness of the converted spectra between frames. Our proposed method considers the relations between frames when estimating GMM parameters by using a set of phoneme-based features of event targets as spectral vectors for training. Therefore, our approach can improve the precision of GMM parameters. Our proposed method also ensures the smoothness of converted speech by performing the conversion procedure for event targets, instead of converting the spectral parameters frame by frame. The experimental results prove the effectiveness of our proposed method.

There are however issues which still remain to be solved. Although prosodic conversion, and duration conversion are outside of the scope of this section, they are important features for the realization of speaker personality, and to improve the natural quality of converted speech. Prosodic conversion, duration conversion, and the problem of over-smooth effect in each converted frame is considered in our future work.

Chapter 7

Summary and Future Work

In this chapter, we conclude the dissertation with summaries and discussions of future work.

7.1 Summary of the dissertation

This dissertation aims to propose spectral modelings and spectral modification algorithms to improve the quality of modified speech in the area of voice transformation.

Most high-quality methods of speech modification process speech signals in the time-frequency domain. In conventional spectral modelings, frames are modeled independently, and there is no models of temporal evolution of parameters. In the time axis, most methods of speech modification process speech signals frame by frame, and do not consider the temporal evolution. Therefore, they do not ensure the smoothness of synthesized speech after modification. In addition, they either do not guarantee to keep the natural evolution of parameters of speech signals. These limitations degrade the quality of modified speech. In the frequency axis, speech modification methods often process the spectral envelope which is estimated from DFT. They then apply the rule-based approach or the statistical approach to modify the spectral envelope. The rule-based approach needs small training data, but this approach does not reflect natural characteristics of speech signals. Moreover, it meets difficulties in modifying the formants or the spectral shape. Although the statistical approach produces high quality of modified speech, it requires large training data. In addition, it works based on mathematical and statistical algorithms, some acoustic constraints need to be added to discover all most interesting and understandable rules in the analyzed data set. Therefore, spectral modification methods including both the statistical approach and the rule-based approach do not offer great flexibilities and effectiveness in modifying speech signals. Three main issues which need to be solved consist of the lack of efficient spectral modelings for speech modification, the insufficient smoothness of modified spectra between frames, and the ineffective spectral modification.

This dissertation consists of seven chapters in which the four main chapters 3, 4, 5, 6 have described and solved three main problems of spectral modification which are mentioned above. Chapter 3 has solved the problem of lack of efficient spectral modelings for speech modification. Chapter 4 has dealt with the problem of insufficient smoothness of

modified spectra between frames. Chapters 5 and 6 have solved the problems of ineffective spectral modification. On the application side, we have improved the quality of modified speech in four main areas, concatenative speech synthesis in Chapter 4, emotional speech synthesis, voice gender conversion in Chapter 5, and spectral voice conversion in Chapter 6. The main works in this dissertation can be summarized as follows.

In Chapter 3, we have focused on solving the first issue, i.e. the lack of efficient spectral modelings for speech modification. To modify the speech spectra, we first have to develop an analysis/synthesis technique. Speech analysis extracts features which are pertinent for different applications, while removing irrelevant aspect of the speech. The analysis/synthesis method is very important, since it decides what kind of representation and which parameters we can choose to modify. The most essential property of speech is the spectral contents (including resonance peaks) which change continuously over time. The physical reason underlying such an essential property is the continual motion of the articulators that forms a time-varying acoustic filter responsible for the generation of the speech waveform. To characterize this type of property, we need a time-frequency representation.

We first introduces two improvements in modeling speech spectral envelope. Although spectral-GMM parameters [166, 167, 170] are flexible to control the speech spectrum, the original method does not ensure an one-to-one correspondence between spectral peaks and Gaussian components. In addition, GMM does not always fit any distribution of patterns, so it is meaningful to provide another probability model which can be chosen instead of GMM. In the first improvement, we not only model the speech spectral envelope well but also ensure a correspondence between spectral peaks and Gaussian components. In the second improvement, we use asymmetric Gaussian mixture model to model the speech spectral envelope. This chapter also presents a new method for speech spectral sequence modeling in the time-frequency domain. Our proposed modeling of speech spectral sequence is potential to allow to ensure the smoothness of modified speech, and to perform efficient spectral modification.

In Chapter 4, we have focused on solving the second issue, i.e. the insufficient smoothness of modified spectra between frames. To ensure the smoothness of modified speech signals, one of efficient ways is to control spectral dynamics. Knagenhjelm and Kleijn [73] point out that spectral dynamics is more important than spectral distortion in human perception. Therefore, control of spectral dynamics improves the quality of synthesized speech. In this dissertation, we employ the temporal decomposition technique [5, 113] to control spectral dynamics in concatenative speech synthesis to addresses the mismatch between concatenation units in a concatenative speech synthesis system.

The most successful TTS approach to-date is called concatenative synthesis. Output speech is limited by the contents of the acoustic inventory, and inevitable concatenation errors can lead to audible discontinuities. Since preparing a large acoustic inventory is expensive and time-consuming, one research direction is smoothing mismatches between concatenation units when having a limited database. For discontinuities at boundary points, spectral mismatch is one of the main problems should be solved. Therefore, in this part, we solve the mismatch of spectral, gain, F0 information at concatenation points.

In Chapters 5 and 6, we have focused on solving the third issue, i.e. ineffective spectral

modification. In Chapter 5, we develop a new efficient rule-based spectral modification algorithm. In Chapter 6, we improve the GMM-based spectral voice conversion method, one kind of the statistical approach.

In Chapter 5, we present a new efficient algorithm for spectral modification and its applications. In Chapter 3, we propose a new model of the speech spectral sequence for speech modification. This chapter solves the next step in stages of speech modification, i.e. how to modify the spectral-GMM features. Formant frequency is one of the most important parameters in characterizing speech, and control of formants can effectively modify the spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. After proposing a new algorithm for spectral modification, we evaluate the effectiveness of our proposed algorithm in two areas, emotional speech synthesis which requires modification of both formant frequency and power, and voice gender conversion which requires a large amount of spectral modification.

Chapter 6 focuses on dealing with drawbacks in a Gaussian mixture model (GMM)-based voice conversion system. In state-of-the-art voice conversion systems, GMM-based spectral voice conversion methods [143] [61] are regarded as some of the best systems. However, the quality of converted speech is still far from natural. There are three main reasons for the degradation of the quality of converted speech: (i) modeling the distribution of acoustic features in voice conversion often uses unstable frames, which degrades the precision of GMM parameters (ii) the transformation function may generate discontinuous features if frames are processed independently (iii) over-smooth effect occurs in each converted frame. Chapter 6 solves the first two of these three drawbacks, insufficient precision of GMM parameters and insufficient smoothness of the converted spectra between frames.

In summary, this dissertation focuses on spectral modification of speech in the area of voice transformation. We propose a framework for speech modification to solve three main problems of spectral modification, i.e. lack of efficient spectral modelings for speech modification, insufficient smoothness of the modified spectra between frames, and ineffective spectral modification. The main contributions of this dissertation are as follows. First, we can flexibly control spectral dynamics of speech signals, which is one important property for ensuring the smoothness of speech signals, and performing time-scale modification. Second, our algorithm performs spectral modification directly on the speech spectral envelope, which is easy to convert, and does not produce artifacts. Third, we add a phoneme constraint to the GMM-based spectral voice conversion method, which improves the quality of converted speech.

Although we only focus on improving the quality of modified speech in the area of voice transformation, we can apply our work in other fields of speech signal processing, such as speech recognition, speech perception, speaker verification and identification. In our work, we model temporal evolution of spectral parameters. We also model directly a speech spectral envelope. These modelings allow us to effectively and flexibly perform time-scale modification and spectral modification. These two operations are employed in

most areas of speech technology.

7.2 Further research directions

In this dissertation, we have discussed the advantages of temporal decomposition [5, 113] and spectral-GMM parameters [104, 106, 107, 108, 109, 166, 167, 170] in spectral modeling and speech modification. Some research topics can be further investigated from our research. We give some suggestions as follows.

Spectral voice conversion

Until now, many methods have been proposed to improve the quality of converted speech signals. However, the quality is far from natural. One of solutions to improve the quality of converted speech signals is the combination of the GMM-based spectral voice conversion and frequency warping conversion, such as [37, 38, 81]. It is well-known that frequency warping methods produce a high-quality speech signals. However, frequency warping methods often decrease the speaker identity, since it is difficult to chose warping factors for each frame. When converting speech signals, they often chose the same warping factors for a number of frames. This implementation is not good, since speech signals change continuously when human beings produce speech sounds. On the contrary, the GMM-based statistical conversion methods increase the speaker identity, but decrease the quality of converted speech signals. In [37, 38, 81], although the combination of GMM-based statistical conversion and frequency warping conversion is employed, the quality still need improving, since they only warp the speech spectral envelope in frequency dimension. The relation of amplitudes between frequency bins is not discussed. Therefore, the quality of converted speech is not high enough. In our work, we use spectral-GMM parameters to model the speech spectral envelope. This modeling allows us to flexibly modify the speech spectral envelope in both dimensions, frequency and amplitude. It is interesting to investigate spectral-GMM parameters as parameters for frequency warping.

In addition, most spectral voice conversion methods use LSFs [23, 34, 37, 61, 107, 108, 163] or MFCCs [74, 143, 148] as features for training and converting. Although these studies get promising results, one of drawbacks of these coefficients is that relations between these coefficients in a speech spectral envelope are difficult to control. This characteristic effects the control over the spectral shape of the speech spectral envelope. That is one of reasons for degradation of converted speech signals. As we mentioned above, spectral-GMM parameters are rather independent. Changing values of each Gaussian component only effects to formants and the part of the spectral shape located around the mean value. Therefore, it is also interesting to investigate spectral-GMM parameters as spectral parameters, instead of using LSFs or MFCCs, in spectral voice conversion.

Speech recognition and speaker recognition

Some studies [19, 68, 111] already focused on the application of temporal decomposition in speech recognition. However, most studies only used event targets as good features

for speech recognition [19, 68] or speaker recognition [111]. Although event targets are found as superior features than conventional features (e.g. MFCCs), these studies have not investigated the combination of static features and dynamic features. Many studied [14, 21, 124, 139] have shown that only the joint use of static and dynamic information can lead to performance improvement in areas of speech and speaker recognition. Moreover, because spectral dynamics calculated from two consecutive frames is sensitive to random fluctuations in the original static features, some studies [17, 21, 150] use spectral dynamic information which is obtained by linking spectral dynamics across multiple frames, and they got promising results. In our analysis/synthesis method [109], M indicates a slope of an event function. Therefore, M can be seen as dynamic information between multiple frames. It is interesting to investigate the combination of event targets and the values of M in speech and speaker recognitions.

Application of spectral-GMM parameters to vocal tract length normalization in speech recognition

One of the major factors affecting the performance of speaker independent speech recognition is the variability in speech signal arising due to the physiological differences of vocal tract of speakers. To increase the accuracy rates of speech recognition, we need to perform acoustic feature speaker normalization. The most popular speaker normalization technique is vocal-tract length normalization (VTLN). A natural issue is the choice of the warping function, which specifies the relationship between the frequency axis used to represent spectra produced by the "standard" speaker and the frequency axis describing the spectral productions of a new speaker. When examining normalization techniques, the important issues are the choice of warping function and how it is selected. In the literature, many studies [35, 49, 158] proposed vocal tract length normalization to improve accuracy rate in speech recognition. However, until now, most methods only deal with frequency normalization. They rarely consider amplitude normalization. Amplitude parameters are so important [54]. As mentioned above, spectral-GMM parameters are flexible to control both frequency and amplitude axes. Spectral-GMM parameters were investigated in speech recognition area [28, 141, 142]. In [141, 142], mean features of spectral-GMM parameters were combined with MFCC to make good features in speech recognition. While in [28], mean parameters of Gaussian components were used to normalize frequency axis. However, these approaches [28, 141, 142] did not take advantages of spectral-GMM parameters to normalize amplitude axis. Therefore, it is of interest to investigate spectral-GMM parameters in speaker normalization, both frequency and amplitude axes.

Speech enhancement

In this dissertation, we dealt with spectral-GMM parameters in speech modification. Moreover, spectral-GMM parameters are also investigated in speech coding [168] [169], speech recognition [141] [142]. In modeling speech spectral envelope using spectral-GMM parameters, we based on the geometric characteristics of formants that can be represent by Gaussian distribution. Applications of geometric features have been investigated in speech

signal processing. For example, in [32, 112, 113], geometric interpretation is employed to ensure the monotonic characteristic of event functions of TD. In [83], Lu and Loizou proposed a new geometric approach to spectral subtraction for eliminating the audible musical noise. In [90], Master presented a model using spectral-GMM parameters to model a DFT magnitude spectrum combining multiple harmonic sources. This model was then used to perform signal separation, recovering the independent component sounds from the original combined sounds. Therefore, it is interesting to investigate spectral-GMM parameters in speech enhancement, due to the geometric characteristics of spectral-GMM parameters.

Applications of temporal decomposition and spectral-GMM parameters to speech perception

We already discussed the temporal decomposition and spectral-GMM parameters in speech modification. In [109, 133], we proposed a new method to identify event locations based on phoneme, and a new method to model event functions. This leads to the merits that we can easily exchange event targets and event functions between two speakers or two emotional styles of a speaker, or changing the temporal evolution. In addition, in [104, 106, 109], we used spectral-GMM parameters to model an event target. Our proposed methods provide flexible tools for investigation of speech perceptions. For example, we can apply our methods in the following experiments.

- Investigate the transitions of vowels and consonants to speech perception by sharing or changing event functions.
- Investigate the change of amplitude of spectral peaks to speech perception.
- Explore the temporal cues on speech perception by sharing or changing the temporal evolution.

Improving the estimation of spectral-GMM parameters

In this dissertation, we discussed the advantages of spectral-GMM parameters in speech modification. When estimating the spectral-GMM parameters, the EM algorithm [30] is used. The EM algorithm for Gaussian mixture models often gets caught in local maxima of the likelihood which involve having too many Gaussians in one part of the space and too few in another, widely separated part of the space. Therefore, we have not ensured the one-to-one correlation between Gaussian components and formants. Some techniques, such as [152], in the area of machine learning can be investigated for improving estimation of spectral-GMM parameters.

Others

In [109, 133], we presented some improvements of the MRTD algorithm, such as identification of event locations based on phoneme, and modeling of event functions. These improvements bring advantages in modifying spectral information and duration. Moreover, speech spectrum modeling using spectral-GMM parameters [104, 106, 107, 108, 109,

166, 167, 170] gives flexibilities to control the speech spectrum. It is interesting to investigate TD and spectral-GMM parameters in emotional speech synthesis and conversion of speaking voices to singing voices in the following directions.

- In [57, 58], Huang and Akagi proposed a three-layer model for expressive speech perception. In this model, Huang and Akagi investigated acoustic features in the third layer. Although they already investigated most acoustic features which have been found to be strongly correlated with emotion, such as F0, energy, three first formant frequencies, there is still another important kind of features which need to be explored for improving the quality of the synthesized speech. Those are amplitudes of the first two harmonics (H1, H2). These features have been also found strongly correlated with emotion, and they are investigated in many studies in the area of emotion, such as [71]. In addition, the amplitudes of the first three formants (A1, A2, and A3) are related to voice quality. We can estimate H1, H2, A1, A2, and A3, but it is difficult to modify the spectrum by their desired factors if using conventional representations of speech spectrum (i.e. LP coefficients or non-parametric representation). Investigation of spectral-GMM parameters in [57, 58] is our future work.

- For the conversion of speaking voices to singing voices, in [129], Saitou et al. pointed out that emphasizing spectral peaks around 3 kHz should be done in the stage of spectral modification. However, they employed the STRAIGHT spectrum to represent the spectrum information. Since the STRAIGHT spectrum is a non-parametric representation, this work is manually done by using weighting functions. In addition, we also need to modify duration of vowels in their transition parts. In [129], using STRAIGHT spectra makes it less flexible. In Chapter 3 of this dissertation, we can flexibly modify duration by using TD. Therefore, it is also interesting to explore TD and spectral-GMM parameters to automatically carry out some steps of the conversion of speaking voice to singing voice in [129].

Bibliography

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, pp. 655–658, 1998.

[2] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database user's manual," *ATR Technical Report, TR-I-0166*, 1990.

[3] A. Acero, "Formant analysis and synthesis using hidden Markov models," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'99)*, pp. 1047–1050, 1999.

[4] G. Ahlbom, F. Bimbot, and G. Chollet, "Modeling spectral speech transitions using temporal decomposition techniques," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'87)*, pp. 13–16, 1987.

[5] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'83)*, pp. 81–84, 1983.

[6] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, 1971.

[7] C. Athaudage, "Speech compression using optimized temporal decomposition for voice storage applications," *Ph.D. Thesis, Royal Melbourne Institute of Technology*, 2001.

[8] C. N. Athaudage, A. Brabley, and M. Lech, "On performance evaluation of a temporal decomposition based speech coder," *Proceedings of the International Conference on Information, Communications, and Signal Processing (ICICS'01)*, 2001.

[9] C. N. Athaudage, A. B. Brabley, and M. Lech, "Optimization of a temporal decomposition model of speech," *Proceedings of the International Symposium on Signal Processing and Its Applications (ISSPA'99)*, pp. 471–474, 1999.

[10] C. N. Athaudage, A. B. Brabley, and M. Lech, "Efficient compression of MELP spectral parameters using optimized temporal decomposition," *Proceedings of the Australian International Conference on Speech Science and Technology (SST-2000)*, pp. 386–391, 2000.

[11] G. Bailly, P. F. Marteau, and C. Aubry, "A new algorithm for temporal decomposition of speech. Application to a numerical model of coarticulation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'89)*, pp. 508–511, 1989.

[12] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation," *Acoustical Science and Technology*, vol. 28, no. 3, pp. 140–146, 2007.

[13] P. N. Bennett, "Using asymmetric distributions to improve text classifier probability estimates," *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, pp. 111–118, 2003.

[14] C. Berasconi, "On instantaneous and transitional spectral information for text-dependent speaker verification," *Speech communication*, vol. 9, no. 4, pp. 129–139, 1990.

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by additive noise," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'79)*, pp. 208–211, 1979.

[16] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," *Proceedings of the Joint Meeting of ASA*, 1999.

[17] B. Bielefeld, "Language identification using shifted delta cepstrum," *In Fourteenth Annual Speech Research Symposium*, 1994.

[18] F. Bimbot, G. Ahlbom, and G. Chollet, "From segmental synthesis to acoustic rules using temporal decomposition," *Proceedings of the 11th, ICPhS*, pp. 31–34, 1987.

[19] F. Bimbot, G. Chollet, and P. Deleglise, "Speech synthesis by structured segments, using temporal decomposition and a glottal excitation," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'89)*, pp. 2183–2186, 1989.

[20] A. S. Bregman, "Auditory scene analysis: The perceptual organization of sound," *MIT Press, Cambridge, MA*, 1990.

[21] J. R. Calvo, R. Fernańdez, and G. Hernńdez, "Application of shifted delta cepstral features in speaker verification," *Proceedings of the Conference of the International Speech Communication Association (Interspeech'07)*, pp. 734–737, 2007.

[22] D. T. Chappell and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, pp. 343–373, 2002.

[23] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'03)*, pp. 2413–2416, 2003.

[24] Y. M. Cheng and D. OShaughnessy, "Short-term temporal decomposition and its properties for speech compression," *IEEE Transactions on Signal Processing*, vol. 39, no. 6, pp. 1282–1290, 1991.

[25] Y. M. Cheng and D. OShaughnessy, "On 450-600 b/s natural sounding speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 207–219, 1993.

[26] D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis," *Journal of the Acoustical Society of America*, vol. 90, pp. 1841–1856, 1991.

[27] A. D. Conkie and S. Isard, "Optimal coupling of dipphones," *Progress in speech synthesis*, pp. 293–304, 1997.

[28] X. Cui and A.Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment," *Computer Speech & Language*, vol. 20, no. 4, pp. 400–419, 2006.

[29] J. R. J. Deller, J. H. L. Hansen, and J. G. Proakis, "Discrete-time processing of speech signals," *The Institute of Electrical and Electronics Engineers, Inc., New York, U. S. A*, 2000.

[30] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, pp. 1–38, 1977.

[31] L. Deng and D. O'Shaughnessy, "Speech processing: A dynamic and optimization-oriented approach," *Marcel Dekker, Inc., New York, U. S. A*, 2003.

[32] P. J. Dix and G. Bloothooft, "A breakpoint analysis procedure based on temporal decomposition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 9–17, 1994.

[33] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, pp. 14–27, 1986.

[34] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including dynamic and phonetic information in voice conversion systems," *Proceedings of the International Conference on Spoken Language Processing (Interspeech'04)*, pp. 1193–1196, 2004.

[35] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, pp. 346–348, 1996.

[36] T. En-Najjary, O. Rosec, and T. Chonavel, "A voice conversion method based on joint pitch and spectral envelope transformation," *Proceedings of the International Conference on Spoken Language Processing (Interspeech'04)*, pp. 1225–1228, 2004.

[37] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech'07)*, pp. 1965–1968, 2007.

[38] D. Erro, T. Polyakova, and A. Moreno, "On combining statistical methods and frequency warping for high-quality voice conversion," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pp. 4665–4668, 2008.

[39] G. Fant, "Acoustic theory of speech production," *The Netherlands: Mouton-The Hague*, 1960.

[40] J. L. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[41] H. Fujisaki, "Information, prosody, and modeling with emphasis on tonal features of speech," *Proceedings of Speech Prosody*, pp. 1–10, 2004.

[42] E. B. George, "An analysis-by-synthesis approach to sinusoidal modeling applied to speech and music processing," *Ph. D. thesis, Georgia Institute of Technology*, 1991.

[43] E. B. George and M. J. T. Smith, "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, 1992.

[44] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 5, pp. 389–406, 1997.

[45] S. Ghaemmaghami and M. Deriche, "A new approach to very low-rate speech coding using temporal decomposition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, pp. 224–227, 1996.

[46] S. Ghaemmaghami, M. Deriche, and S. Sridharan, "Hierarchical temporal decomposition: A novel approach to efficient compression of spectral characteristics of speech," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, pp. 2567–2570, 1998.

[47] S. Ghaemmaghami and S. Sridharan, "Very low rate speech coding using temporal decomposition," *Electronics Letters*, vol. 35, no. 6, pp. 456–457, 1999.

[48] S. Ghaemmaghami, S. Sridharan, and V. Chandran, "Coding speech at very low rates using temporal decomposition based spectral interpolation and mixed excitation in the LPC model," *Applied Signal Processing*, vol. 6, no. 4, pp. 203–223, 1999.

[49] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Journal of Computer Speech & Language*, vol. 20, pp. 107–123, 2006.

[50] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, pp. 4–29, 1984.

[51] C. Hamon, E. Mouline, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'89)*, pp. 238–241, 1989.

[52] Z. Hanzlicek and J. Matousek, "F0 transformation within the voice conversion framework," *Proceedings of the International Conference on Spoken Language Processing (Interspeech'07)*, pp. 1961–1964, 2007.

[53] K. Heng and L. Wenju, "Selective-LPC based representation of STRAIGHT spectrum and its applications in spectral smoothing," *Proceedings of the International Conference on Spoken Language Processing (Interspeech'06)*, pp. 2050–2053, 2006.

[54] J. M. Hillenbrand, R. A. Houde, and R. T. Gayvert, "Speech perception based on spectral peaks versus spectral shape," *J. Acoust. Soc. Am.*, vol. 119, pp. 4041–4054, 2006.

[55] T. Hoshiya, S. Sako, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Improving the performance of HMM-based very low bitrate speech coding," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, p. 800803, 2003.

[56] Y. S. Hsiao and D. G. Childers, "A new approach to formant estimation and modification based on pole interaction," *Proceedings of the Conference Record of the Thirtieth Asilomar Conference on Signals, Systems and Computers*, pp. 783–787, 1996.

[57] C. F. Huang and M. Akagi, "A rule-based speech morphing for verifying an expressive speech perception model," *Proceedings of the Conference of the International Speech Communication Association (Interspeech'07)*, pp. 2661–2664, 2007.

[58] C. F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[59] E. Jung, A. T. Schwarzbacher, K. Humphreys, and R. Lawler, "Application of real-time AMDF pitch-detection in a voice gender normalisation system," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02)*, pp. 2521–2524, 2002.

[60] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 6, pp. 1419–1426, 1986.

[61] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, pp. 285–288, 1998.

[62] A. Kain, Q. Miao, and J. van Santen, "Spectral control in concatenative speech synthesis," *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, 2007.

[63] T. Kato, S. Omachi, and H. Aso, "Asymmetric Gaussian and its application to pattern recognition," *Lecture Notes in Computer Science*, vol. 2396, pp. 405–413, 2002.

[64] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proceedings of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA'01)*, 2001.

[65] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Journal of Speech Communication*, vol. 27, pp. 187–207, 1999.

[66] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *Proceedings of the European Conference on Speech Communication and Technology (Interspeech'03)*, pp. 2401–2404, 2003.

[67] E. K. Kim, S. Lee, and Y. H. Oh, "Hidden Markov model based voice conversion using dynamic characteristics of speaker," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'97)*, pp. 2519–2522, 1997.

[68] S. Kim, "Very low bit rate speech coding based on temporal decomposition of line spectral frequencies," *Ph.D. Thesis, Korea Advanced Institute of Science and Technology (KAIST), Korea*, 2000.

[69] S. Kim and Y. Oh, "Efficient quantisation method for LSF parameters based on restricted temporal decomposition," *Electronics Letters*, vol. 35, no. 12, pp. 962–964, 1999.

[70] S. J. Kim, S. H. Lee, W. J. Han, and Y. H. Oh, "Efficient quantization of LSF parameters based on temporal decomposition," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, pp. 2575–2578, 1998.

[71] T. Kitamura, "Acoustic analysis of imitated voice produced by a professional impersonator," *Proceedings of the Conference of the International Speech Communication Association (Interspeech'08)*, pp. 813–816, 2008.

[72] D. Klatt, "Review of text-to-speech conversion for English," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[73] H. P. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, pp. 732–735, 1995.

[74] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, pp. 720–723, 2003.

[75] H. Kuwabara and K. Ohgushi, "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech," *Acustica*, vol. 63, no. 2, pp. 120–128, 1987.

[76] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *Speech and Audio Processing, IEEE Transactions on*, pp. 323–332, 1999.

[77] R. Lawlor and A. D. Fagan, "A novel efficient algorithm for voice gender conversion," *XIVth International Congress of Phonetic Sciences, University of California, Berkeley, USA*, pp. 77–80, 1999.

[78] M. E. Lee, "Acoustic models for the analysis and synthesis of the singing voice," *Ph.D. Dissertation, Georgia Institute of Technology*, 2005.

[79] A. N. Lemma, W. B. Kleijin, and E. F. Deprettere, "LPC quantization using wavelet based temporal decomposition of the LSF," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'97)*, pp. 1259–1262, 1997.

[80] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[81] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through combining modified GMM and formant mapping for Mandarin," *Proceedings of the International Conference on Digital Telecommunications (ICDT'07)*, p. 10, 2007.

[82] P. H. Low, C. H. Ho, and S. Yaseghi, "Using estimated formant tracks for formant smoothing in text to speech synthesis," *Proceedings of the Automatic Speech Recognition and Understanding (ASRU'03)*, pp. 688–693, 2003.

[83] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, 2008.

[84] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE Publication*, pp. 561–580, 1975.

[85] J. Makhoul and J. Wolf, "Linear prediction and the spectral analysis of speech," *BBN Report No. 2304*, 1972.

[86] R. H. Manell, "Formant diphone parameter extraction utilising a labelled single-speaker database," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'88)*, 1998.

[87] S. M. Marcus and R. A. J. M. Van Lieshout, "Temporal decomposition of speech," *IPO Annual Progress Report 19*, pp. 26–31, 1984.

[88] J. D. Markel and A. H. Gray, "Linear prediction of speech," *Springer-Verlag, New York*, 1976.

[89] P. F. Marteau, G. Bailly, and M. T. Janot-Giorgetti, "Stochastic model of diphone-like segments based on trajectory concepts," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'88)*, pp. 615–618, 1988.

[90] A. Master, "Speech spectrum modeling from multiple sources," *Master's thesis, Engineering Dep, Cambridge, England*, 2000.

[91] R. McAulay and T. Quatieri, "Magnitude-only reconstruction using a sinusoidal speech model," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'84)*, pp. 441–444, 1984.

[92] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 744–754, 1986.

[93] Q. Miao, X. Niu, E. Klabbers, and J. van Santen, "Effects of prosodic factors on spectral balance: analysis and synthesis," *Speech prosody*, 2006.

[94] H. Mizuno, M. Abe, and T. Hirokawa, "Waveform-based speech synthesis approach with a formant frequency modification," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'93)*, pp. 195–198, 1993.

[95] E. Molines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[96] E. Molines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175–205, 1995.

[97] B. C. J. Moore, "An introduction to the psychology of hearing," *Fourth ed. Academic Press, New York*, 1997.

[98] J. A. Moore, "The use of the phase vocoder in computer music applications," *Journal of the Audio Engineering Society*, vol. 24, no. 9, pp. 717–727, 1978.

[99] R. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *IEEE Signal Processing Letters*, vol. 9, pp. 19–21, 2002.

[100] P. Motlicek, H. Hermansky, H. Garudadri, and N. Srinivasamurthy, "Speech coding based on spectral dynamics," *Lecture Notes in Computer Science, Springer*, vol. 4188, pp. 471–478, 2006.

[101] A. Mouchtaris, J. V. der Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to single channel speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1180–1193, 2007.

[102] A. Nandasena, P. Nguyen, and M. Akagi, "Spectral stability based event localizing temporal decomposition," *Computer Speech and Language*, vol. 15, no. 4, pp. 381–401, 2001.

[103] B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," *Journal of Acoustical Science and Technology (in press).*

[104] B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," *Proceedings of the International Speech Communication Association (Interspeech'07)*, pp. 538–541, 2007.

[105] B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *Proceedings of the RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'07)*, pp. 481–484, 2007.

[106] B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *J. Signal Processing*, vol. 11, pp. 333–336, 2007.

[107] B. P. Nguyen and M. Akagi, "Control of spectral dynamics using temporal decomposition in voice conversion and concatenative speech synthesis," *Proceedings of the RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08)*, pp. 279–282, 2008.

[108] B. P. Nguyen and M. Akagi, "Phoneme-based spectral voice conversion using temporal decomposition and Gaussian mixture model," *Proceedings of the International Conference on Communications and Electronics (ICCE'08)*, pp. 224–229, 2008.

[109] B. P. Nguyen, T. Shibata, and M. Akagi, "High-quality analysis/synthesis method based on temporal decomposition for speech modification," *Proceedings of the International Speech Communication Association (Interspeech'08)*, pp. 662–665, 2008.

[110] P. C. Nguyen, "A study on efficient algorithms for temporal decomposition of speech," *Ph.D. Thesis, Japan Advanced Institute of Science and Technology (JAIST), Japan*, 2003.

[111] P. C. Nguyen, M. Akagi, and T. B. Ho, "Temporal decomposition: A promising approach to VQ-based speaker identification," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, pp. 184–187, 2003.

[112] P. C. Nguyen, M. Akagi, and B. P. Nguyen, "Limited error based event localizing temporal decomposition and its application to variable-rate speech coding," *Speech Communication*, vol. 49, no. 4, pp. 292–304, 2008.

[113] P. C. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low bit rate speech coding," *IEICE Transactions on Information and Systems*, vol. E86-D, pp. 397–405, 2003.

[114] B. Ninness and S. Henriksen, "Time and frequency scale modification of speech signals," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, pp. 1295–1298, 2000.

[115] M. Niranjan and F. Fallside, "Temporal decomposition: A framework for enhanced speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'89)*, pp. 655–658, 1989.

[116] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'95)*, pp. 1029–1032, 1995.

[117] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *Speech and Audio Processing, IEEE Transactions on*, pp. 3–14, 1993.

[118] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-29, no. 4, pp. 786–794, 1981.

[119] M. Plumpe, A. Acero, H. W. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, 1998.

[120] A. Pribilova and J. Pribil, "Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description," *Speech Communication*, vol. 48, pp. 1691–1703, 2006.

[121] L. Qin, G. Chen, Z. Ling, and L. Dai, "An improved spectral and prosodic transformation method in STRAIGHT-based voice conversion," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, pp. 21–24, 2005.

[122] T. Quatieri and R. McAulay, "Speech transformations based on a sinusoidal representation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'85)*, pp. 489–492, 1985.

[123] T. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *Signal Processing, IEEE Transactions on*, vol. 40, no. 3, pp. 497–510, 1992.

[124] L. Rabiner and B. H. Juang, "Fundamental of speech recognition," *Prentice Hall*, 1993.

[125] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE transactions on audio, speech and language processing*, vol. 14, pp. 972–980, 2006.

[126] D. Rentzos, S. Vaseghi, Q. Yan, and C. H. Ho, "Parametric formant modelling and transformation in voice conversion," *International Journal of Speech Technology*, vol. 8, pp. 227–245, 2005.

[127] C. H. Ritz and I. S. Burnett, "Temporal decomposition: A promising approach to low rate wideband speech compression," *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech'01)*, pp. 2315–2318, 2001.

[128] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'01)*, pp. 749–752, 2001.

[129] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using STRAIGHT," *Proceedings of the Conference of the International Speech Communication Association (Interspeech'07)*, pp. 4005–4006, 2007.

[130] H. Scheffe, "An analysis of variance for paired comparisons," *Journal of the American Statistical Association*, vol. 37, pp. 381–400, 1952.

[131] C. Shadle and R. Damper, "Prospects for articulatory synthesis: A position paper," *Proceedings of the ISCA Tutorial and Research Workshop*, 2001.

[132] D. O. Shaughnessy, "Speech Communication: Human and Machine," *2nd Edition. IEEE Press, New York*, 2000.

[133] T. Shibata and M. Akagi, "A study on voice conversion method for synthesizing stimuli to perform gender perception experiments of speech," *Proceedings of the RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08)*, pp. 180–183, 2008.

[134] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," *Circuits and Systems, IEEE International Sympoisum on*, vol. 1, pp. 594–597, 1991.

[135] Y. Shiraki, "Optimal temporal decomposition for voice morphing preserving Δ cepstrum," *IEICE Trans Fundam. Electron. Commun. Comput. Sci.*, vol. E87-A, no. 3, pp. 577–583, 2004.

[136] Y. Shiraki and M. Honda, "Extraction of temporal pattern of spectral sequence based on minimum distortion criterion," *Proceedings of the Autumn Meeting of ASJ*, pp. 233–234, 1991, (in Japanese).

[137] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'84)*, pp. 1.10.1–1.10.4, 1984.

[138] F. K. Soong and B. H. Juang, "Optimal quantization of LSP parameters," *Speech and Audio Processing, IEEE Transactions on*, vol. 1, pp. 15–24, 1993.

[139] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 6, pp. 871–879, 1988.

[140] M. N. Stuttle, "A Gaussian mixture model spectral representation for speech recognition," *Ph.D. thesis, Cambridge University*, 2003.

[141] M. N. Stuttle and M. J. F. Gales, "A mixture of Gaussians front end for speech recognition," *Proceedings of the European Conference on Speech Communication and Technology*, pp. 675–678, 2001.

[142] M. N. Stuttle and M. J. F. Gales, "Combining a Gaussian mixture model front end with MFCC parameters," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02)*, pp. 1565–1568, 2002.

[143] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[144] P. Taylor, A. Black, and R. Caley, "The architecture of the Festival speech synthesis system," *Proceedings of the third ESCA workshop on speech synthesis*, 1998.

[145] I. R. Titze, "Physiologic and acoustic differences between male and female voices," *Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1699–1707, 1989.

[146] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, pp. 9–12, 2005.

[147] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[148] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'01)*, pp. 841–844, 2001.

[149] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'93)*, pp. 1315–1318, 2000.

[150] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. D. Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02)*, pp. 89–92, 2002.

[151] E. Turajlic, D. Rentzos, S. Vaseghi, and C. H. Ho, "Evaluation of methods for parameteric formant transformation in voice conversion," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, pp. 724–727, 2003.

[152] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "Split and merge EM algorithm for improving Gaussian mixture density estimates," *J. VLSI Signal Process. Syst.*, vol. 26, no. 1-2, pp. 133–140, 2000.

[153] A. Van Dijk-Kappers, "Comparison of parameter sets for temporal decomposition," *Speech Communication*, vol. 8, no. 3, pp. 203–220, 1989.

[154] A. Van Dijk-Kappers, "Temporal decomposition of speech and its relation to phonetic information," *Ph.D. Thesis, Eindhoven University of Technology, The Netherlands*, 1989.

[155] A. Van Dijk-Kappers and S. Marcus, "Temporal decomposition of speech," *Speech Communication*, vol. 8, no. 2, pp. 125–135, 1989.

[156] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'93)*, pp. 554–557, 1993.

[157] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," *Proceedings of the International Conference on Spoken Language Processing (Interspeech'02)*, pp. 285–288, 2002.

[158] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, pp. 761–764, 1999.

[159] J. Wouters, "Analysis and synthesis of degree of articulation," *Ph.D. thesis*, 2001.

[160] J. Wouters and M. Macon, "Spectral modification for concatenative speech synthesis," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, pp. 941–944, 2000.

[161] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, pp. 30–38, 2001.

[162] A. Wrench, "The MOCHA-TIMIT articulatory database," *Queen Margaret University College, http://www.cstr.ed.ac.uk/artic/mocha.html*, 1999.

[163] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Transactions on Speech and Audio Processing*, pp. 1301–1312, 2006.

[164] P. Zolfaghari, "Sinusoidal model based segmental speech coding," *Ph.D. thesis, Cambridge University*, 1998.

[165] P. Zolfaghari, Y. Atake, K. Shikano, and H. Kawahara, "Investigation of analysis and synthesis parameters of straight by subjective evaluation," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*, pp. 498–501, 2001.

[166] P. Zolfaghari, H. Kato, Y. Minami, A. Nakamura, S. Katagiri, and R. Patterson, "Dynamic assignment of Gaussian components in modelling speech spectra," *Journal of VLSI Signal Processing Systems*, vol. 45, pp. 7–19, 2006.

[167] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," *Proc. ICSLP*, pp. 1229–1232, 1996.

[168] P. Zolfaghari and T. Robinson, "A formant vocoder based on mixtures of Gaussians," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, pp. 1575–1578, 1997.

[169] P. Zolfaghari and T. Robinson, "Speech coding using mixture of Gaussians polynomial model," *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1495–1498, 1999.

[170] P. Zolfaghari, S. Watanabe, A. Nakamura, and S. Katagiri, "Bayesian modelling of the speech spectrum using mixture of Gaussians," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, pp. 553–556, 2004.

Publications

Awards

1. Student paper award of the 2007 RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'07).

Journals

1. B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," Journal of Acoustical Science and Technology (in press).
2. B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," Journal of Signal Processing, vol. 11, pp. 333-336, 2007.
3. P. C. Nguyen, M. Akagi, and B. P. Nguyen, "Limited error based event localizing temporal decomposition and its application to variable-rate speech coding," Speech Communication, vol. 49, pp. 292-304, 2007.

International Conferences

1. B. P. Nguyen, Takeshi Shibata, and M. Akagi, "High-quality analysis/synthesis method based on temporal decomposition for speech modification," Proceedings of the Conference of the International Speech Communication Association (Interspeech'08), pp. 662-665, 2008.
2. B. P. Nguyen and M. Akagi, "Phoneme-based spectral voice conversion using temporal decomposition and Gaussian mixture model," Proceedings of the International Conference on Communications and Electronics (ICCE'08), pp. 224-229, 2008.
3. B. P. Nguyen and M. Akagi, "Control of spectral dynamics using temporal decomposition in voice conversion and concatenative speech synthesis," Proceedings of the RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08), pp. 279-282, 2008.
4. B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," Proceedings of the Conference of the International Speech Communication Association (Interspeech'07), pp. 538-541, 2007.
5. B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," Proceedings of the RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'07), pp. 481-484, 2007.

Domestic Conferences and Others

1. B. P. Nguyen, I. Tokuda, and D. Erickson, "Analysis of the roles of glottal features for emotion classification in spontaneous and acted emotional speech," Proceedings of the Autumn Meeting of the ASJ, pp. 367-370, 2008.
2. B. P. Nguyen and M. Akagi, "Improvement of spectral peak estimation using Gaussian mixture model for speech modification," Proceedings of the Spring Meeting of the ASJ, pp. 303-304, 2008.
3. B. P. Nguyen and M. Akagi, "Temporal decomposition-based speech spectra modeling using asymmetric Gaussian mixture model," Proceedings of the Autumn Meeting of the ASJ, pp. 359-362, 2007.
4. B. P. Nguyen and M. Akagi, "A flexible temporal decomposition-based spectral modification method using asymmetric Gaussian mixture model," Technical Report of IEICE, pp. 389-394, 2007.
5. B. P. Nguyen, C-F. Huang, and M. Akagi, "Temporal decomposition-based spectral modification and its application to emotional speech synthesis," Proceedings of the Spring Meeting of the ASJ, pp. 239-240, 2007.
6. C-F. Huang, M. Akagi, and B. P. Nguyen, "Rule-based speech morphing for evaluating expressive speech perception," Proceedings of the Spring Meeting of the ASJ, pp. 241-242, 2007.