Studies on Improving the Naturalness of Synthesized Speech under Limited Data Conditions

by

Trung-Nghia Phung

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

August, 2013

Abstract

The motivation of this dissertation was to propose methods for improving the naturalness of synthesized speech under limited data conditions.

Because speech is the result of sequential linking of phonetic units, a speech synthesizer requires a database that covers all phonetic units in a specific unit set to synthesize any input text content, resulting in a requirement of significant amount of data for synthesizing. Due to the efforts of co-articulation on speech synthesis (SS), not only all context-independent phonetic units but also all context-dependent phonetic units, are necessary to synthesize natural speech. As a result, state-of-the-art speech synthesizers require large-scaled speech corpora to synthesize natural speech.

Building a large-scaled speech corpus is a costly task that takes a long time and a great deal effort by engineers, acousticians and linguists. Therefore, high-quality SS under limited data conditions is important in practice, specifically for under-resourced languages. Synthesizing speech with a limited amount of data is also critical for customizing synthesized speech with multiple voices. It is also critical in movies and games applications when we need to synthesize the voices of whom are not alive or have lost their voice characteristics. Since methods of voice transformation can be used with a limited amount of target data, a target voice can be synthesized by using a two-step procedure, which synthesizes a standard voice with a large-scaled database first and transforms the synthesized standard voice to the target voices later. However, this approach still requires a large-scaled database for the first step of synthesizing the standard voice. ***As a result, to directly build highly-natural speech synthesizers under limited data conditions is an important and interesting research topic.***

In the literature, there are a few approaches to directly improve the naturalness of synthesized speech under limited data conditions. The first approach is to maximize the use of existence contexts in the database. This approach has been shown its efficiency compared with traditional approaches. However, SS with this improvement still requires large amounts of original data for concatenation or training. The second approach is to use methods of speech modification to reduce mismatch-context errors occurred when phonetic units in matching contexts are not available due to the limitation of the database. This approach has just used in concatenative SS (CSS) and has shown a little improvement.

One core problem in synthesizing natural speech under limited data conditions is to ensure an "appropriate smoothness" in synthesized speech. Both temporal and spectral over-smoothness and over-roughness can reduce the naturalness of synthesized speech. Under limited data conditions, the over-roughness in speech synthesized by CSS and the over-smoothness in speech synthesized by Hidden-Markov-Model-based SS (HMMSS) increase. This problem significantly reduces the naturalness of synthesized speech under limited data conditions. However, ensuring an "appropriate smoothness" in synthesized speech has not taken deeply into consideration in research fields of SS.

Therefore, the unified purpose of this research is to use methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech under limited data conditions.

Temporal decomposition (TD) is an interpolation method decomposing a spectral or prosodic sequence into its sparse event targets and correspondent temporal event functions. The modified restricted TD (MRTD) is one simplified but efficient version of TD. With a determination of event functions close to the concept of co-articulation in speech, MRTD can synthesize generally smooth speech. The smoothness in synthesized speech can be adjusted by modifying event targets of MRTD. Therefore, MRTD can be used to modify / transform / and synthesize speech with an "appropriate smoothness". *As a result, MRTD was used throughout the proposed methods in this research.*

Constructing SS under limited data conditions is more practical for under-resourced languages, where large public speech corpora are missing. As the tonal-monosyllabic Vietnamese is an under-resourced language, Vietnamese datasets were used in this paper for implementations and evaluations.

Based on the general motivation and the unified purpose of this research, four specific objectives were specified and solved.

The first objective was to propose a new and efficient speech smoothness measure to control and to evaluate the smoothness in synthesized speech. Since the global variances (GV) of static features are not sufficient to measure and control the smoothness in synthesized speech, the first objective of this dissertation was to investigate the smoothness measures for speech synthesis and to propose a new and efficient measure. A speech smoothness measure based on the square sum of the variance of the delta-delta sequence, named distance of global smoothness measure (DGSM), in both time and spectral domains was proposed. The proposed DGSM showed its reliability and efficiency to measure the smoothness in different kinds of synthesized speech in both time and spectral domains.

The second objective was to solve the main problem of CSS under limited data conditions, which is reducing mismatch-context errors in CSS that cause the temporal over-roughness in synthesized speech. Although there are many methods for reducing mismatch-context error in CSS under limited data conditions, their performances are still limited. Mismatch-context errors can occur in all frames inside the co-articulated transition region between two adjacent phonemes. Mismatch-context errors cause the temporal over-roughness in synthesized speech that reduces its naturalness. Current methods are unable to localize this co-articulated transition region. In addition, mismatch-context errors are not equal for all positions inside the co-articulated transition region. Mismatch-context errors are gradually changed in the ideal case of co-articulation from the maximum value at the phoneme boundary of two neighboring phones to the minimum value at the onset/offset of the co-articulated transition region. Although some methods have attempted to weight these errors, their temporal evolutions have not been ensured due to the lack of a co-articulation model in CSS. In this research, a model of contextual effect in CSS, a method of speech modification, and a method of unit selection were proposed and combined into a non-uniform CSS to reduce mismatch-context errors in CSS under limited data conditions. MRTD was used as a core of these proposed methods to obtain synthesized speech with an "appropriate smoothness". Experimental results with Vietnamese datasets show that mismatch-context errors could be solved and the proposed CSS was natural in terms of smoothness under limited data conditions.

The third objective was to solve the main problem of HMMSS under limited data conditions, which is reducing both the temporal and spectral over-smoothness in HMMSS. When synthesized speech is over-smooth, it sounds "muffled" and far from natural. "Over-smoothness" also causes a reduction in identification emotions / expressions / styles in

speech that can also affect to the perception of the naturalness. The state-of-the-art method to reduce over-smoothness is the parameter generation algorithm that takes into consideration GV of static spectral features. The experimental results with this method revealed that the naturalness of synthetic speech was significantly improved. However, due to the inefficiency of GV of static spectral features to measure the smoothness in synthesized speech, the over-smoothness was still considerable. In this research, a hybrid SS between HMMSS and MRTD, referred to as HTD, was proposed to reduce over-smoothness in synthesized speech under limited data conditions. The event functions of MRTD decomposed from spectral sequences synthesized by HMMSS were preserved while the corresponding event targets of MRTD were rendered by a "target selection" to transform speech synthesized by HMMSS to the original speech. As a result, synthesized speech can be transformed to obtain an "appropriate smoothness". Experimental results with Vietnamese datasets show that speech synthesized by the proposed HTD had an"appropriate smoothness", resulting in an improvement on the naturalness in synthesized speech.

The fourth objective was to solve one case with ultra-limited data conditions for tonal languages when the number of tonal units is not sufficient by using a method of tone transformation. Lexical tones are usually represented by fundamental frequency (F0) contours. Therefore, a tone transformation can be considered as a F0 contour transformation applied for converting lexical tones. The state-of-the-art Gaussian-Mixture-Model-based (GMM-based) F0 transformation transforms frame-based F0 values and their temporal deltas, can be efficient for expressive speech because F0 contours are largely varied on the emotions of speech in short-term intervals. However, F0 contours of a source tonal unit and a target tonal unit are usually distinct in their long-term approximations rather than their short-term details. Therefore, transforming the short-term frame-based F0 values and their deltas may not efficiently transform the F0 contours of lexical tones but may increase the noise sensitivity. In addition, while the general GMM-based voice transformation has many advantages, it still has drawbacks, including the insufficient precision of GMM models and parameters and the over-roughness between converted frames. Therefore, a method of F0 transformation using MRTD and GMM was proposed to ensure an "appropriate smoothness" in the transformed F0 contours of the lexical tones. Experimental results with Vietnamese datasets show the effectiveness of the MRTD-GMM F0 transformation and it could be applied to improve the usability of SS of tonal languages under ultra-limited data conditions.

In summary, methods of speech modification and transformation were proposed to improve the naturalness of synthesized speech under limited data conditions based on the concept "appropriate smoothness" in synthesized speech. These methods showed their efficiencies on improving the usability under limited data conditions for both CSS and HMMSS. There results contribute to the research fields of speech processing by introducing the new concept of "appropriate smoothness" in speech. There results also contribute to the research and development fields of SS in order to make SS more convenient and more efficient for human-machine interaction systems. These results also contribute to the research fields of speech modification and transformation, in which the efficiency of MRTD to obtain an "appropriate smoothness" synthesized speech were verified and confirmed.

Acknowledgments

The author would like to dedicate this dissertation to his lovely wife Pham Thi Mai Huong for supporting him during the changes and sharing his life. Without her love, support, and inspiration, this work would be far from being complete.

The author wishes to express his sincere gratitude to his advisor Professor Masato Akagi of Japan Advanced Institute of Science and Technology (JAIST) for his constant and invaluable encouragement, supports and kind guidance during this work. Professor Masato Akagi brought the author to his laboratory, closely followed and also gave the freedom to do research. He opened a new door of academic research, specifically in the field of speech signal processing, to the author.

The author would like to express his gratitude to his co-advisor Associate Professor Masashi Unoki of JAIST for his helpful discussions and suggestions. Associate Professor Masashi Unoki made a great deal effort to train and educate the author from detail skills to the critical thinking to do research, stayed together the author when he was disappointed and encouraged him to overcome all challenges.

The author wishes to express his thanks to Professor Jianwu Dang of JAIST and Tianjin University, China for his valuable comments and suggestions, as well as for supervising the minor research and for serving as a member of the dissertation committee.

The author wishes to express his thanks to Associate Professor Luong Chi Mai of Institute of Information Technology of Vietnam (IoIT) for her continuous supports and encouragements. Associate Professor Luong Chi Mai and her colleagues in IoIT kindly supported the author Vietnamese speech corpora and documents on Vietnamese language. They also helped the author to complete the evaluations of his researches.

The author is grateful to Professor Hideki Kawahara of Wakayama University for his kind support source codes of STRAIGHT, for serving as a member of the dissertation committee and for valuable comments, suggestions, and discussions on several meetings and conferences as well as on the preliminary PhD defense of the author.

The author would like to express his thanks to Associate Professor Hirokazu Tanaka of JAIST for being a member of the dissertation committee and for helpful comments and suggestions.

The author wishes to express his gratitude to Doctor Nguyen Phu Binh of Ministry of Science and Technology of Vietnam for providing materials on the implementation of Temporal Decomposition, and Doctor Vu Tat Thang of IoIT for providing materials on the HMM-based Vietnamese TTS.

The author would like to thank Professor Ho Tu Bao and Associate Professor Huynh Van Nam of JAIST, Associate Professor Nguyen Huu Cong of Thai Nguyen University (TNU), Doctor Pham Viet Binh, Doctor Vu Vinh Quang, Doctor Nguyen Van Tao, and Doctor Vu Duc Thai of Thai Nguyen University of Information and Communication Technology (ICTU) for their valuable advices, helps and kind encouragements.

The author would like to acknowledge the financial supports from the Vietnamese Ministry of Education and Training (MOET), and would like to thank JAIST for providing a top-ranked research environment.

The author wishes to to express his special thanks to his family for their sacrifice, love and supports. Most of all, he thanks his parents Phung Thanh Binh and Pham Thi Dung, his sisters Phung Thi Hong Hanh, Phung Thi Thanh Hieu and Phung Thi Thu Hien, his wife Pham Thi Mai Huong, and his sons Phung Nam Trung and Phung Nam Bach.

The author also would like to give thanks to the Vietnamese community at JAIST for sharing ups and downs. Especially, he would like to thank to Doctor Doan Anh Vu, his great friend in JAIST, for all their common memories and experiences.

Finally, the author is grateful to all who have affected or suggested his areas of research. The author devotes his sincere thanks and appreciation to all of them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents the definition and applications of SS. Popular approaches in SS, especially the two state-of-the-art systems CSS and HMMSS, are also briefly introduced. Then, problems of SS under limited data conditions addressed in this dissertation are explained. The smoothness in the original and synthesized speech and its relations to the naturalness of speech are also analyzed and discussed. Based on these considerations, the motivation of this research, the new ideas that were proposed, and the main contributions of the dissertation are described. Finally, the outline and summarization of the dissertation are given.

1.1 Definition of speech synthesis

SS is the artificial production of human speech. A system used for this purpose is called a speech synthesizer that can be implemented in software or hardware products [1, 2].

A text-to-speech (TTS) system converts normal language text into speech that is composed of two parts: a text-processing part and a speech synthesizer [1, 2]. Therefore, SS is the core part of a TTS.

1.2 Applications of speech synthesis

SS can be used in several applications, such as applications for the blind [3, 4, 5], applications for the deafened and vocally handicapped [6, 7], educational applications [3], speech to speech translators or the universal voice translator [8, 9], and applications for telecommunications and multimedia. In principle, SS can be used in all kinds of human-machine interaction systems.

One of the most important and widely-used application field in SS is the reading and communication aids for the blind. The first commercial speech synthesizer application was probably the Kurzweil reading machine for the blind introduced by Raymond Kurzweil in the late 1970's [3]. The most serious drawback with reading machines is speech intelligibility which should be maintained with speaking rates [4]. Naturalness is also important to make these reading machines more convenient [5].

Both born-deaf people and people with hearing difficulties have speaking impossibility or difficulties. *SS gives the deafened and vocally handicapped an opportunity to communicate with the rest of the world.* Additionally, a talking head combined from both speech

and visual information can be more useful for the deaf and dumb [6, 7].

Many educational situations can be used with SS. A computer-based speech synthesizer can be programmed for specific tasks such as spelling and pronunciation teaching for different languages. It can also be used with interactive educational applications [3]. A speech synthesizer connected with word processor is also a helpful aid to proof reading.

SS is one of core parts in speech-to-speech translation system, which can be used as an universal voice translator combined from automatic speech recognition (ASR) systems, machine translators, and TTSs of several languages [8, 9]. This kind of application brings a new general and universal communication tool to connect every people in the world.

Other important applications in SS are in the area of telecommunications and multimedia. Speech synthesizer may be used to speak out some desktop messages from a computer, such as for printer activity, or used to speak out e-mails. Synthesized speech has been also used in telephone enquiry systems, in which the quality of synthesized speech has reached the acceptable level for normal customers, resulting in the increase of these systems for everyday use. Synthesized speech may also be used to speak out short text messages (SMS) in mobile phones.

Synthesized speech may also be used in several other communication and entertainment systems, such as videophones, videoconferencing, talking mobile phones, music, or healthcare industry. Synthesized speech can be also used in movies or games for regenerating the synthetic voices of people who are not alive or who have lost their voice characteristics due to the old age, or generating the synthetic voice of famous people in another language that they do not know [10, 11]. Therefore, there are many researches attempting to transform or to adapt the synthesized speech to different individual voices [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23].

An extended version of conventional reading machines is emotional speech synthesizer that aims to generate speech in different emotional modes such as excited, happy, sad, or angry by adjusting the voice characteristics of synthesized speech to suitable emotional voices [20, 24, 25, 26].

One another extended version of conventional reading machines is singing voice synthesizer, which can sing songs rather than reading the text only, as described in [24, 27, 28, 29, 30, 31, 32]. VOCALOID is a very famous singing voice synthesizer, proposed by H. Kenmochi in 2000 and was developed into a commercial product [32]. This product enables users to synthesize singing by typing in lyrics and melody.

Consequently, SS has been more and more important and popular in daily life. SS is currently used to read thousands of www-pages for the blind, is combined with visual information in 3D talking head for the deaf and dumb, is used in spelling and pronunciation teaching for different languages. SS is one of core parts in speech-to-speech translation system that brings a new universal communication tool to connect every people in the world. Synthesized speech is also widely used to speak out from SMS in mobile phones to longer messages such as email in personal computers or question/answer in automatic answering systems. Especially, current TTSs are able to not only reading the texts with neutral voices but also speak emotional or expressive voices and singing voices, and current TTSs can speak by multiple individual voices instead of using one standard voice. These advanced features make SS being applied more widely in human-machine interaction systems.

1.3 Brief review of approaches to speech synthesis

1.3.1 Formant synthesis

Formant synthesis is one of early SS [33, 34]. It is one kind of physical modeling synthesis, in which parameters such as F0, voicing, and noise levels are varied over time to create a waveform of artificial speech.

Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice.

Formant-synthesized speech can be reliably intelligible. However, formant synthesis technology generates robotic-sounding speech that would never be mistaken for human speech. Additionally, formant synthesis requires a great effort to prepare and manually correct parameters in the offline stage before synthesizing speech, resulting in the inflexibility of the synthesis. *Therefore, this kind of SS is not considered in this research.*

1.3.2 Articulatory synthesis

Articulatory synthesis is one kind of SS which is based on models of the human vocal tract and the inside articulation processes [35, 36]. The shape of the vocal tract can be controlled that involves modifying the position of the speech articulators, such as the tongue, jaw, and lips. Then, speech is created by digitally simulating the flow of air through the representation of the vocal tract.

Until recently, articulatory synthesis models have not been incorporated into commercial systems due to the complexity for implementing the models and the inconvenience to collect the articulatory data. *Therefore, this approach has not been used in this research.*

1.3.3 Concatenative speech synthesis

CSS is based on the concatenation of segments of recorded speech. State-of-the-art CSS is US system [37, 38, 39, 40] that requires a large-scaled speech corpus to select matching units for every concatenation. Therefore, all possible units of all possible contexts are required, resulting in a large-scaled database with a size up to dozens of gigabytes for concatenation, representing dozens of hours of speech [41, 42, 43].

Although differences between natural variations in speech and the nature of the segmenting techniques result in some audible glitches and in the instable synthesized trajectories in the output, US still produces the most natural-sounding synthesized speech up to now. *Therefore, this is one kind of SS studied in this research.*

More details on CSS are presented in section 2.5 of chapter 2.

1.3.4 HMM-based speech synthesis

HMMSS has been widely studied for the two last decades [15, 44, 45, 46, 47, 48, 49]. In this approach, spectral and prosodic features of speech are modeled and generated in a unified statistical framework using HMMs.

HMMSS has many advantages that have been shown in the literature, such as the high intelligibility of synthesized speech even with limited training data, the small footprint, the low computational load, and the flexibility to change the voice characteristics.

Although HMMSS has many advantages, quality of synthesized speech is still far from natural, which is mainly due to two reasons of buzziness and over-smoothness in synthesized speech. The former is a common issue with speech coding, which has recently been significantly improved, while the latter is caused by "averagely" statistical processing in HMMSS, which is still a remaining problem of HMMSS at present. *HMMSS is one kind of SS studied in this research.* More details on HMMSS are presented in section 2.6 of chapter 2.

1.3.5 Hybrid approach

Recently, hybrid approaches between CSS and HMMSS have been studied to take both their advantages.

One approach uses HMM models to smooth segment sequences obtained by US [50]. Although this approach can reduce the discontinuities at segment boundaries, it introduces some artifacts when there is mismatch between the smoothed filter coefficients and excitation signal. Another type of hybrid approach uses spectrum parameters, F0 values, and durations or part of them generated from HMMs to compute target cost for US [43, 51, 52, 53]. The newest and most successful method in this approach was HTT proposed recently by Qian et.al. [43], in which its efficiency was confirmed in Blizzard Challenge 2010. The HMM trajectory is used to guide the selection of each 5ms frame to concatenate the waveforms in HTT. The naturalness of HTT is comparable to that of unit selection and its intelligibility is comparable to that of HMMSS. Additionally, HTT is language-independent due to the use of short frames instead of phonetic-level-units. However, HTT still has drawbacks. The major one is the use of short frames, which requires a perfect selection process. If the selection process is imperfect due to a limited data corpus, it may be easy to perceive discontinuities between frames. As a result, this synthesizer requires a huge amount of data for rendering. Another drawback with HTT is its high computational load. The searching task to select the matched short frames in a huge database is not convenient in most personal hardware platforms. Additionally, HTT is not flexible for voice transformation. As a consequence, this approach can improve the quality and stability of synthesized speech but the main drawbacks of US are still remaining. *Hybrid SS was also studied in this research.*

1.4 Designing speech corpus for speech synthesis

Building speech corpora is the first step for many speech applications such as SS [54, 55, 56, 57, 58, 59]. To build a speech corpus, a text script is first selected, which contains phonetically and prosodically rich sentences. Then, this text script is usually spoken out by a native speaker and is recorded. The corresponding duration of the recorded speech files ranges from several minutes to hours. The speech files are then transcripted or labeled in different levels viz. phrases, words, syllables and phones.

One basic requirement of a speech corpus for SS is the coverage of all phonemes [57, 59]. This requirement is critical and strict to ensure speech corpora can be used to synthesize any input text content. Due to the effect of co-articulation, the use of smallest phonetic units, i.e. basic phonemes, can reduce quality of synthesized speech. Therefore, high-quality speech synthesizers usually need large-scaled speech corpora that not only cover all phonemes but also cover all larger units such as diphones or syllables.

Ensuring the phonetically balance is another basic requirement of designing a commercial speech corpus for SS [54, 58]. The ideal phonetically balance means that the occurrence frequencies of all phonetic units are equivalent. This needs a careful phonetically analysis to design the text scripts of the corpus. In practice, the ideal phonetically balance may be never reached. Therefore, people have attempted to design the text scripts with the phonetically balances as close to the ideal phonetically balance as possible by using both manual preparation and automatic algorithms [57]. As a results, *the sizes (and durations) of commercial speech corpora are usually proportion to their average occurrence frequencies of phonetic units.*

Based on the two aforementioned basic requirements, we can briefly describe the two kinds of speech corpora: small-scaled and large-scaled ones as follows.

- A small-scaled corpus has to cover all basic phonemes. The average occurrence frequency of phonemes is small.

- A large-scaled corpus has to cover not only all basic phonemes but also larger units such as diphones or syllables. The average occurrence frequency of phonetic units is large.

The number of candidates for each concatenation is small when using a small-scaled speech corpus for CSS. Besides, the number of samples for training the HMM model of each phoneme is also small when using a small-scaled speech corpus for HMMSS. Therefore, quality of both CSS and HMMSS is usually low under limited data conditions.

On the contrary, the number of candidates for each concatenation is large when using a large-scaled speech corpus for CSS and the number of samples for training the HMM model of each phonetic unit is large when using a large-scaled speech corpus for HMMSS. Therefore, quality of both CSS and HMMSS is significantly improved when increasing the size of the database.

Additionally to these two basic requirements, there are several issues involved in designing a speech corpus for SS, especially for synthesis with advanced features such as synthesizing multiple emotional or individual voices [54, 55, 56, 57, 58, 59].

1.5 The smoothness in synthesized speech

The articulators typically move smoothly during speech production [60]. Therefore, speech features of natural speech are generally smooth. However, rapid changes in speech naturally occur in some cases such as in plosives [60]. These are the "natural roughness" or "natural discontinuities" in speech features. Besides the "natural discontinuities", several kinds of "unexpected discontinuities" in SS, such as discontinuities caused by mismatch-context errors in CSS or discontinuities caused by noisy recording environments, can reduce the naturalness of synthesized speech [61, 62, 63, 64] or the naturalness of recorded speech [65].

On the contrary, the smoother speech features does not mean the more natural synthesized speech. Too smooth or over-smooth causes the "muffleness" in synthesized speech [45, 46, 47] and causes the reduction in identification emotions / expressions / styles in speech [66] that can affect to the perception of the naturalness in synthesized speech.

Instead of synthesize too smooth or too rough speech, the "optimal smoothness" that naturally exists in original speech or ***an "appropriate smoothness", referred to as an approximation of the "optimal smoothness" in this research***, have to be

reached to obtain the naturalness in synthesized speech close to that in the original speech. The "appropriate smoothness" depends on the content of speech and is different between vowels and consonants. It also depends on the observed speech features. For instance, the "appropriate smoothness" of a spectral feature differs from that of a prosodic feature such as F0 contour.

To objectively evaluate that speech is over-smooth, or appropriately smooth, or over-rough needs an objective measure of speech smoothness. In the literature, statistical variances of static spectral features have been used as a measure of smoothness in synthesized speech [47] and in noisy speech [119, 120]. By generating speech parameter with the GV close to that of original speech, the smoothness in synthesized speech has been expected to be natural [47]. Minimizing statistical variances of static spectral features has also been one of widely-used approach in noisy speech enhancement [119, 120]. However, statistical variances of static spectral features are not sufficient to measure and control the smoothness in synthesized speech, as presented in chapter 3. Therefore, a speech smoothness measure based on the square sum of the variance of the delta-delta sequence, named DGSM, in both time and spectral domains was proposed and is presented also in chapter 3.

1.6 Speech synthesis under limited data conditions

The term "limited data conditions" in this research refers to the conditions with which only small-scaled speech corpora are available. Besides, the term "ultra-limited data conditions" refers to the conditions with which the available speech corpora are even not match with basic requirements of a standard small-scaled speech corpus, such as the sufficient number of phonemes, for instance.

Speech is the result of sequential linking of phonetic units such as phonemes, which are the minimal distinctive units. Therefore, a speech synthesizer needs a database that covers all phonetic units in a specific unit set to synthesize any input text content. This database is used to analyze, or concatenate, or train parameters of all units for synthesizing. The need of covering all possible units leads to a requirement of significant amount of data to build a speech synthesizer.

The boundaries between adjacent phonetic units such as phonemes are usually blurred and smoothed, resulting in the hide of essential information in the sound transitions. This phenomenon of the mutual influence of adjacent phones, which are the acoustic realization of phonemes, is called co-articulation. Due to the efforts of co-articulation in SS, not only all context-independent phonetic units but also all possible phonetic units in all possible contexts, or all context-dependent phonetic units, are necessary to synthesize natural speech with an "appropriate smoothness". As a result, state-of-the-art speech synthesizers require large-scaled speech corpora to synthesize natural speech. For example, high-quality HMMSS requires a database with a size in gigabytes for training [41], US requires a database with a size in dozens of gigabytes for concatenation [41, 42], HTT requires a database with a size in 2-10 gigabytes for rendering [43]. On the contrary, quality of synthesized speech is drastically reduced under limited data conditions.

Unfortunately, building a large-scaled speech corpus is a costly task that takes a long time and requires a great deal of effort by engineers, acousticians, and linguists. Therefore,

to build high-quality SS with limited data is an important and practical issue, specifically for under-resourced languages with which only a few of small speech corpora are usable.

Synthesizing speech with a limited amount of data is critical for customizing synthesized speech with multiple individual and emotional voices [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. It is also critical in movies and games applications when we need to synthesize the voices of famous people who are not alive or have lost their voice characteristics [10].

As a result, to build high-quality speech synthesizers under limited data conditions is still a challenging open research question and it is the main issue of this dissertation.

To solve this problem, there are several related issues have to be considered. ***This research focuses on how to ensure synthesized speech having an "appropriate smoothness" under limited data conditions, in order to improve the naturalness of synthesized speech under these conditions.***

1.7 Motivation and scope of the research

The motivation of this dissertation was to propose methods for improving the naturalness of synthesized speech under limited data conditions.

In the literature, there are a few approaches to improve the naturalness of synthesized speech under limited data conditions. A popular indirect approach is to synthesize a standard voice with large amount of data first, then using methods of voice adaption or voice transformation to adapt / transform the standard voice to the target voice with a few target data [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. This approach is efficient for synthesizing multiple emotional or individual voices as shown in the literature [15, 16]. However, this approach still requires a large-scaled database for the first step of synthesizing the standard voice, which is usually not available under limited data conditions.

There are two current direct approaches for improving the naturalness of synthesized speech under limited data conditions. The first approach is to maximize the use of existence contexts in the database, i.e. the use non-uniform units in CSS [37] or the use of context clustering techniques in HMMSS [45]. This approach has been shown its efficiency compared with traditional approaches. However, the state-of-the-art CSS and HMMSS still require large amounts of original data. The second approach is to use methods of speech modification to reduce mismatch-context errors caused by the limitation of the database [61, 62, 63, 64]. This approach has just used in CSS and has also shown a little improvement.

The unified purpose of this research was to propose methods of speech modification and transformation to obtain synthesized speech with an "appropriate smoothness" for directly improving the naturalness of synthesized speech under limited data conditions. Under this unified purpose, four objectives were specified. The first objective was to propose a speech smoothness measure that can be applied to evaluate and to control the smoothness in synthesized speech. The second objective was to solve the main problem of CSS under limited data conditions, which is reducing mismatch-context errors in CSS to obtain an "appropriate smoothness" in synthesized speech. The third objective was to solve the main problem of HMMSS under

limited data conditions, which is reducing over-smoothness in HMMSS to obtain an "appropriate smoothness" in synthesized speech also. The fourth objective was to deal with an ultra-limited data condition in tonal languages, when the number of tonal phonetic units for synthesizing is not sufficient, by using transformed tonal speech units with an "appropriate smoothness" when the original speech is missing.

Ensuring synthesized speech with an "appropriate smoothness" is the core common concept of the proposed methods in this research. Obtaining an "appropriate smoothness" in transformed / modified / synthesized speech over modification and transformation tasks is a challenge. TD is an interpolation method decomposing a spectral or prosodic sequence into its sparse event targets and correspondent temporal event functions. MRTD is one simplified but efficient version of TD. With a determination of event functions close to the concept of co-articulation in speech, MRTD can synthesize generally smooth speech. The smoothness in synthesized speech can be adjusted by modifying event targets of MRTD [17, 18, 19, 20, 21, 22, 23]. Therefore, MRTD can be used to obtain the synthesized speech with an "appropriate smoothness". As a result, MRTD was used throughout this research to solve the four objectives of this research.

The four issues related to the four objectives of this research are presented as follows.

1. A smoothness measure for speech features is important to control and to adjust synthesized speech with an "appropriate smoothness" as well as to evaluate the performance of a speech synthesizer. Since directly using statistical variances of static spectral features is not sufficient to measure and control the smoothness in synthesized speech as shown in chapter 3, the first objective of this dissertation was to propose a new and efficient speech smoothness measure.

2. Mismatch-context errors caused by co-articulation of speech occur frequently under limited data conditions. Therefore, the performance of CSS is drastically reduced when the size of the speech corpus is reduced. In the Blizzard Challenge 2006 which provided a large-scaled speech corpus with amount of 5 h of speech, the best system was based on CSS. However, in Blizzard Challenge 2005 which provided a smaller speech corpus with amount of 1.5 h of speech, a HMMSS [44] outperformed the well-established US systems in both speech quality and intelligibility.

There are a few approaches to reducing mismatch-context errors in CSS by methods of speech modification in the literature [61, 62, 63, 64]. The simple and traditional method to reduce mismatch-context errors is by using linear spectral smoothing [61]. Some complex methods have also been proposed [62, 64]. Although mismatch-context errors occur in all frames inside the co-articulated transition region between two adjacent phonemes, current methods are unable to localize this co-articulated transition region. Additionally, the mismatch-context errors are not the same for all locations inside the co-articulated transition region. Mismatch-context errors are gradually changed from the maximum value at the phoneme boundary of two neighboring phones to the minimum value at the onset/offset of the co-articulated transition region. A few methods have attempted to weight these mismatch-context errors [64]. However, their temporal evolutions have been not ensured due to the lack of a co-articulation model in CSS. Therefore, the results obtained by D. T. Chappell [63] showed that conventional linear spectral smoothing [61] was still the most reliable and efficient method for reduce mismatch-context errors. As a consequence, to efficiently solve mismatch-context errors in CSS under limited data conditions is still a challenge and this was the second objective of this dissertation.

3. Speech synthesized by HMMSS is over-smooth. When synthesized speech is over-

smooth, it sounds muffled and far from natural [45, 46, 47, 48, 49]. Additionally, over-smoothness reduces the identification of emotions / expressions / styles in speech [66] that can affect to the perception of the naturalness in synthesized speech. Therefore, over-smoothness is the main remaining factor reducing the naturalness of HMMSS. Over-smoothness is mainly affected by the accuracy of model estimates and that of the training algorithm [48]. These factors are affected by the amount of training data [49]. The larger the amount of training data, the more accurate the model estimates and the training algorithm, and the lesser the over-smoothness in synthesized speech. As a result, the effect of over-smoothness becomes more serious in a situation with limited training data. Therefore, it is difficult to ensure the naturalness of HMMSS under limited data conditions.

There have been many studies attempting to solve over-smoothness in HMMSS. Using multiple mixtures for modeling state output probability density can reduce over-smoothness in synthesized speech [46] but it causes over-training problem due to the increased number of model parameters. A method of combining continuous HMMs with discrete HMMs, and a method of increasing the number of HMM states have also reduced the over-smoothness in HMMSS [48]. However, these methods increase the complexity of HMMs and may not convenient in practical systems. The state-of-the-art method of reducing over-smoothness in HMMSS is the method of parameter generation that take into consideration GV [47]. In this method, speech parameters are generated based on criteria of not only maximizing the HMM likelihood for static and dynamic features but also the likelihood for GV. The naturalness of synthetic speech could be significantly improved. However, to accurately estimate the GV of static spectral features of the original speech needs a sufficient amount of training data that is usually not available under limited data conditions. Moreover, since GV of static spectral features are still not sufficient to measure and control the smoothness in synthesized speech, as presented in chapter 3, over-smoothness in HMMSS is still considerable.

As a result, to efficiently reduce over-smoothness in HMMSS specifically under limited data conditions is challenging. This was the third objective of this dissertation.

4. In ultra-limited data conditions, the number of phonetic units for synthesizing may be not sufficient. Since speech modification or transformation can convert a speech unit into other units, it can be a general approach to reduce the required number of speech units to improve the usability of SS under limited data conditions. This problem is more popular and is also more serious with tonal languages in which the numbers of tonal speech units are significantly large. Therefore, tone transformations for SS of tonal languages were studied in this research to transform a original phonetic unit to other units with different combinations of tones.

A tone transformation can be considered as a F0 contour transformation applied for converting lexical tones. The simplest F0 contour transformation which can be applied for converting lexical tones is the simple exchange of the F0 contours of a source tonal unit with that of a target unit [67]. This method requires a large amount of data covering all tonal units to compute F0 contours of all tonal units in the offline stage. Therefore, this method is still not available under limited data conditions. The state-of-the-art GMM-based F0 transformation [68] transforming both frame-based F0 values and their temporal deltas is efficient for expressive speech since F0 contours are largely varied on the emotions of speech in short-term intervals. However, F0 contours of a source tonal unit and a target tonal unit are usually distinct in their long-term approximations rather than

their short-term details [69, 70, 71]. Therefore, transforming the short-term frame-based F0 values and their deltas may not efficiently transform the F0 contours of lexical tones but may increase the noise sensitivity. Additionally, GMM-based voice transformation still has drawbacks, including the insufficient precision of GMM models and parameters, the temporal over-roughness of the converted parameters between frames, and the spectral over-smoothness in each converted frame [18].

As a consequence, tone transformation is still difficult and the fourth objective of this dissertation is to propose an efficient tone transformation for SS of tonal languages that can improve the usability of these SS under limited data conditions.

Constructing SS under limited data conditions is more practical with under-resourced languages, where large public speech corpora are missing. As the tonal-monosyllabic Vietnamese is an under-resourced language, Vietnamese datasets extracted from corpus DEMEN567 [72] were used in this paper for implementations and evaluations.

1.8 Main contributions of the dissertation

As mentioned earlier, this dissertation focuses on methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech for improving the naturalness of synthesized speech under limited data conditions. The four issues were mentioned above, i.e. proposing a speech smoothness measure, reducing the mismatch-context error in CSS under limited data conditions to ensure an "appropriate smoothness" in synthesized speech, reducing the over-smoothness in HMMSS with a limited amount of data for training and rendering to ensure an "appropriate smoothness" in synthesized speech, and reducing the number of tonal speech units in SS of tonal languages by using a tone transformation that can produce an "appropriate smoothness" in transformed speech. The major contributions of this dissertation can be summarized as follows.

1. A smoothness measure for speech features was proposed to evaluate the synthesis methods and was tested different kinds of synthesized speech. The numerical analysis shows that the proposed measure is efficient and reliable.

2. Methods of reducing mismatch-context errors in CSS under limited data conditions, including a method of speech modification using MRTD, were proposed. The experimental results with Vietnamese datasets revealed that the proposed methods convincingly outperformed conventional methods in terms of both speech intelligibility and naturalness while synthesized speech could obtain an "appropriate smoothness".

3. A hybrid SS between HMMSS and MRTD, named HTD, which uses MRTD to transform speech synthesized by HMMSS to the original speech, was proposed to solve the over-smoothness problem of HMMSS under limited data conditions. The experimental results with Vietnamese datasets revealed that the proposed synthesizer articulated efficiently under limited data conditions in terms of both speech intelligibility and naturalness, the over-smoothness in synthesized speech was significantly reduced, and synthesized speech could obtain an "appropriate smoothness". Additionally, the proposed synthesizer had a small footprint, had small computational load, and could be flexible for voice transformation.

4. A method of tone transformation using GMM and MRTD to reduce the number of tonal units in the SS of tonal languages was proposed. The experimental results with Vietnamese datasets revealed that a significant number of tonal units could be reduced by

using the transformed tonal units instead of using the original ones, while the transformed tonal units were natural, high-intelligible and had an "appropriate smoothness".

Consequently, methods based on speech modification and transformation were proposed in order to improve the naturalness of synthesized speech under limited data conditions. These methods used MRTD as a tool to ensure an "appropriate smoothness" in synthesized speech. The experimental results confirmed that these proposed methods were efficient and the usability of SS under limited data conditions could be improved. Therefore, all four issues related to the four objectives of this dissertation were mostly solved based on common concept of "appropriate smoothness" in synthesized speech.

The remaining problem of this research is the lack of implementations and evaluations with several speakers and languages to confirm the speaker-independence, language-independence and the generality of the proposed solutions and methods.

The research results contribute to research fields of speech processing by introducing a new concept of "appropriate smoothness" in speech. The research results also contribute to research fields and development fields of SS in order to make SS more convenient and more efficient for human-machine interaction systems. The research results also contribute to the research fields of speech modification and voice transformation to make these systems more flexible.

1.9 Outline of the dissertation

The rest of the dissertation is organized as follows. A schematic flowchart of the research and diagram of the dissertation is combined in Fig. 1.1.

Chapter 2

This chapter presents the common backgrounds that were used throughout this research, i.e. source/filter models of speech production, the linear prediction (LP) model, the Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT), the original TD of Atal and the simplified version MRTD, the GMM-based methods for voice transformations including the conventional GMM-based method and the MRTD-GMM, basics on CSS and HMMSS, and methods for evaluating synthesized speech.

Chapter 3

This chapter aims to solve the first issue of this dissertation, which was to propose a speech smoothness measure to evaluate SS systems. First, method using GV of static spectral features is introduced and analyzed. Then, the proposed global speech smoothness measure using the square sum of the variance of the delta-delta sequence, referred to as DGSM in this dissertation, is explained and discussed.

Chapter 4

This chapter motivates to solve the second issue of this dissertation, which was to reduce mismatch-context errors in CSS to improve the naturalness of CSS under limited data conditions. The proposed model of contextual effect in CSS, the proposed method of speech modification, the proposed method of unit selection, and the proposed CSS are presented. The experiments to evaluate the proposed model, methods, and CSS, and the

results are also described and discussed.

Chapter 5

This chapter aims to solve the third issue of this dissertation, which was to improve the naturalness of HMMSS under limited data conditions by reducing the over-smoothness in synthesized speech. The proposed hybrid SS between HMMSS and MRTD, referred to as HTD in this dissertation, is described in which speech synthesized by HMMSS is transformed to the original speech to reduce its over-smoothness under limited data conditions. Both the structure of HTD and details of the proposed target selection procedure to transform speech synthesized by HMMSS into original speech are described. The differences between the proposed HTD and the HTT synthesis, which is the current state-of-the-art hybrid synthesis approach, were clearly discussed. Finally, the evaluations are described and discussed.

Chapter 6

This chapter motivates to solve the fourth issue of this dissertation, which was to reduce the number of tonal units for SS of tonal languages by using a method of tone transformation. The question why tone transformation can be used to reduce the number of tonal units for SS of tonal languages is answered. Then, the proposed MRTD-GMM method for tone transformation is described. After that, the proposed non-parallel alignment method applied for the proposed tone transformation is presented. Finally, the evaluations are described and discussed.

Chapter 7

In this chapter, the contributions of this dissertation and discussions on future researches are summarized.

APPENDIX

In this appendix, a brief introduction to Vietnamese language, Vietnamese phonology, the development of Vietnamese speech corpora and Vietnamese TTS is given.

1.10 Summary

This chapter presents the overview of the research. The motivation and the scope are given and the contributions of dissertation are summarized.

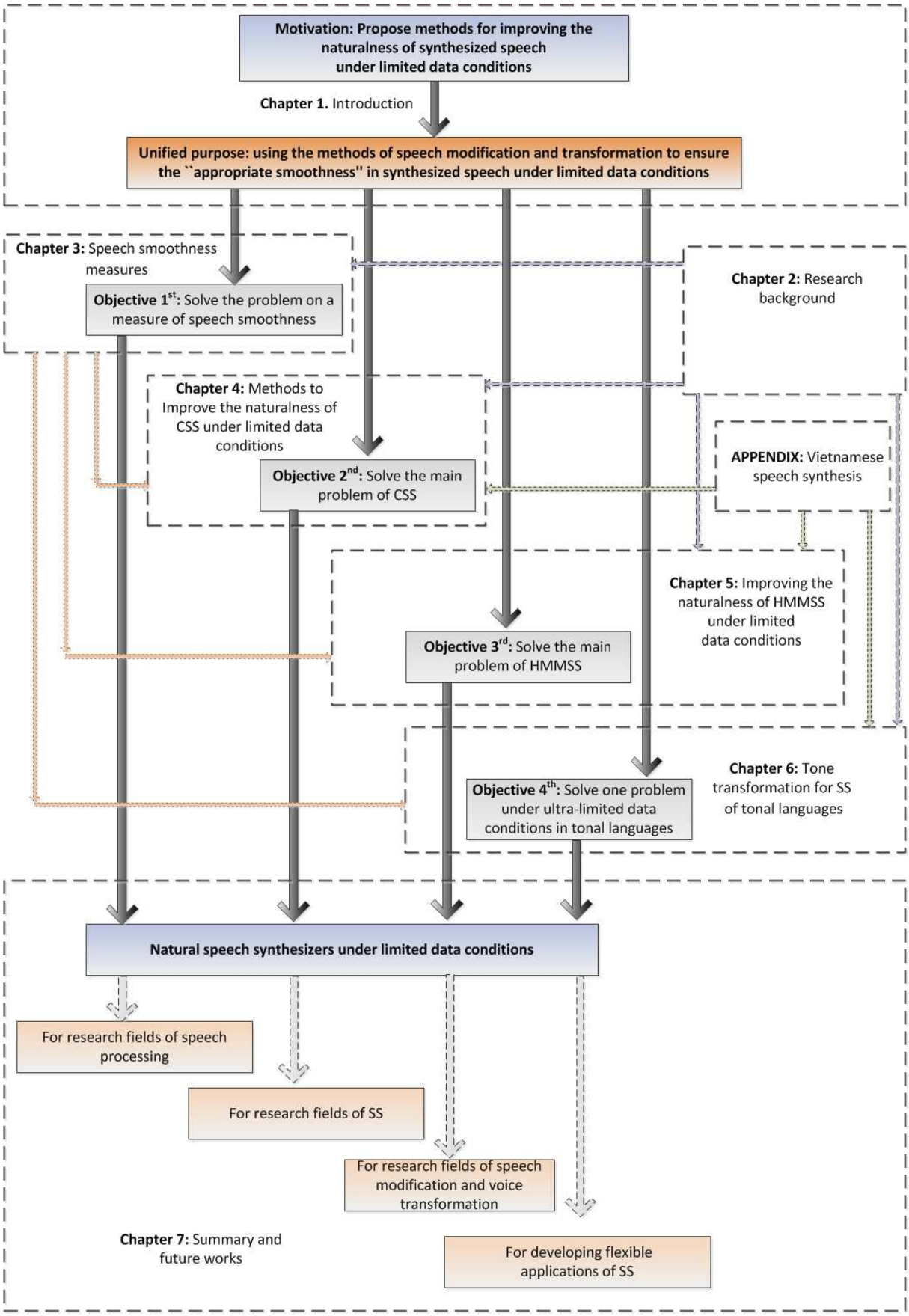

Figure 1.1: Schematic flowchart of the research and the dissertation.

Chapter 2

Research Background

This chapter presents the common backgrounds that were used throughout this research. The first section presents concepts of speech information, the source/filter model of speech production, the LP model, and the STRAIGHT. The second section describes original TD and its simplified version MRTD, which is the tool for speech modification used in chapter 4 and for speech transformation used in chapter 5. The third section presents the conventional GMM-based speech transformation and GMM-MRTD framework for spectral transformation, which is developed for the proposed tone transformation in chapter 6. The fourth section, which is also the last section of this chapter, describes basic concepts in the two state-of-the-art SS, CSS and HMMSS, and methods for evaluating synthesized speech.

2.1 Speech information

Speech is the vocalized form, which is the most popular form, of human-human communication. Speech is researched in terms of the speech production and speech perception of the sounds used in vocal languages. Several research fields involve these including acoustics, psychology, speech pathology, linguistics, cognitive science, communication studies, otolaryngology and computer science.

In speech production, manner of articulation describes how the tongue, lips, jaw, and other speech organs are involved in making a sound [60]. Normal human speech is produced with pressure provided by the lungs which creates phonation in the glottis in the larynx that is then modified by the vocal tract into different vowels and consonants. The study of speech production models is important for building several speech applications such as SS. On the theory of speech production, co-articulation is a natural phenomenon in continuous speech occurring when adjacent phonetic gestures overlap, in which phonetic gestures are linguistically significant actions of structures of the vocal tract [73, 74, 75, 76, 77]. Co-articulation causes contextual effects in speech that is the main factor affecting to the need of using large-scaled database in SS.

Speech perception refers to the processes by which humans are able to understand the sounds [78]. The study of speech perception is closely linked to the fields of phonetics, linguistics, cognitive psychology and perception in psychology. Research in speech perception seeks to understand how human listeners recognize speech sounds and use this information to understand spoken language. Speech perception is important to build several speech applications, i.e. ASR.

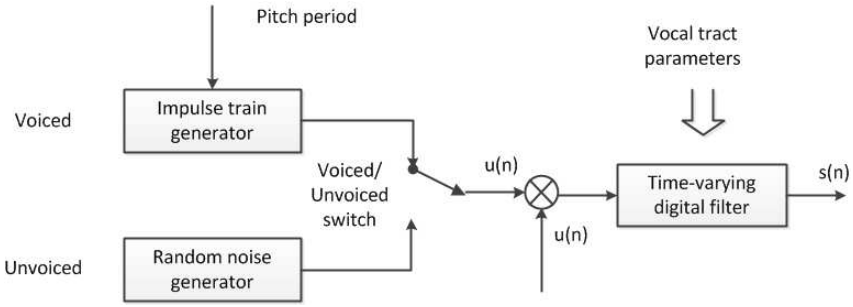

Figure 2.1: Block diagram of the source-filter model for speech production.

2.2 Source-filter model for speech production

2.2.1 Source-filter model

Since speech production models are core parts of most speech synthesizers except the waveform CSS, the source-filter model for speech production that forms the basis for many speech production models is used in this research. This sub-section presents the source-filter model for speech production.

In 1960, Fant [60] introduced the source-filter model for speech production. This model has used in vocoders that are the cores of several speech applications such as speech coding and SS. The source-filter theory hypothesizes that an acoustic speech signal can be seen as a source signal 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 given in Eq. (2.1).

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

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

Fant demonstrated that by modeling the vocal tract as a series of lossless acoustic tubes, a linear model for speech production can be derived. As shown in the block diagram in Fig. 2.1, 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 $F0$, $T0$, of the speaker.

2.2.2 Linear prediction model

The source-filter model is usually tied with the use of the LP model [79]. 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 $\mathrm{H}(z)$ of the form as shown in Eq. (2.2).

$$\mathrm{H}(z) = \frac{\mathrm{S}(z)}{\mathrm{U}(z)} = \frac{G}{1 - \sum_{i=1}^{p} b_i z^{-i}} \tag{2.2}$$

where $\mathrm{S}(z)$ and $\mathrm{U}(z)$ 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 efficient reconstruction of speech for all speech sounds when it is excited by an accurate enough input excitation signal. In the simplest 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 b_i and gain parameter G can be estimated in a computationally efficient manner.

The basic idea behind the LP model is that a given speech sample at time n, $\mathrm{s}(n)$, can be approximated as a linear combination of the past p speech samples, as described in Eq. (2.3).

$$\mathrm{s}(n) \approx b_1\mathrm{s}(n-1) + b_2\mathrm{s}(n-2) + ... + b_p\mathrm{s}(n-p) = \sum_{i=1}^{p} b_i\mathrm{s}(n-i) \tag{2.3}$$

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

In source-filter model, the speech samples $\mathrm{s}(n)$ are related to the excitation $\mathrm{u}(n)$ by Eq. (2.4).

$$\mathrm{s}(n) = \sum_{i=1}^{p} b_i\mathrm{s}(n-i) + G\mathrm{u}(n) \tag{2.4}$$

A linear predictor with prediction coefficients b_i is defined as a system whose output is given in Eq. (2.5).

$$\hat{s}(n) = \sum_{i=1}^{p} b_i\mathrm{s}(n-i) \tag{2.5}$$

The prediction error $\mathrm{e}(n)$ is defined as shown in Eq. (2.6).

$$\mathrm{e}(n) = \mathrm{s}(n) - \hat{s}(n) = \mathrm{s}(n) - \sum_{i=1}^{p} b_i\mathrm{s}(n-i) \tag{2.6}$$

Equation (2.6) implies that the prediction error is the output of a system with the transfer function given by Eq. (2.7).

$$\mathrm{B}(z) = 1 - \sum_{i=1}^{p} b_i\mathrm{s}(n-i) \tag{2.7}$$

The prediction error filter $\mathrm{B}(z)$ is an inverse filter for the system, $\mathrm{H}(z)$, as shown in Eq. (2.8).

$$\mathrm{H}(z) = \frac{G}{\mathrm{B}(z)} \tag{2.8}$$

The basic problem of LP analysis is to determine the optimum set of prediction coefficients b_i from speech signal $\mathrm{s}(n)$. 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 b_i involves the minimization of the prediction error $\mathrm{E}(n)$ for a window of speech around the sample of index n.

$$\mathrm{E}n = \sum_{m} \mathrm{e}_n^2(m) = \sum_{m} (\mathrm{s}_n(m) - \hat{s}_n(m))^2 = \sum_{m} (\mathrm{s}_n(m) - \sum_{i=1}^{p} b_i\mathrm{s}_n(m-i))^2 \tag{2.9}$$

We can find the values of b_i that minimize $\mathrm{E}(n)$ in Eq. (2.9) by setting $\frac{\partial \mathrm{E}_n}{\partial b_j}$ for $j = 1, 2, ..., p$ and we obtain Eq. (2.10).

$$\sum_m \mathrm{s}_n(m-j)\mathrm{s}_n(m) = \sum_{j=1}^{p} b_i \sum_m \mathrm{s}_n(m-i)\mathrm{s}_n(m-j), \quad 1 \leq j \leq p \tag{2.10}$$

If we define $\mathrm{r}_n(j,i) = \sum_m \mathrm{s}_n(m-j)\mathrm{s}_n(m-i)$, Eq. (2.10) can be written as Eq. (2.11).

$$\sum_{i=1}^{p} b_i r_n(j,i) = r_n(j,0), \quad j = 1, 2, ..., p \tag{2.11}$$

The linear system of equations given in Eq. (2.11) can be solved to determine the LP coefficients (LPC) b_i. The quantities $\mathrm{r}(i,j)$ can be calculated either using the covariance method [79] or the autocorrelation method [80].

2.2.3 Line spectral frequency

LSF coefficients are other representation of LPC but LSF coefficients have several advantages compared with LPC that make them superior to LPC [81, 82, 83]. The main advantages are as follows.

(1) The LSF coefficients have smaller sensitivity to quantization noise than LPC.

(2) When the roots of $\mathrm{P}(z)$ and $\mathrm{Q}(z)$ are interleaved, stability of the filter is ensured if and only if the roots are monotonically increasing. Moreover, the closer two roots are, the more resonant the filter is at the corresponding frequency. Because LSFs are not overly sensitive to quantization noise and stability is easily ensured, LSF are widely used for quantizing direct LP filters.

(3) Line spectral frequencies can be interpolated more efficiently than LPC can.

(4) 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.

(5) The LSF coefficients correspond to the bandwidths and approximate locations of the formant frequencies. Therefore, we can directly obtain the information about formant locations and bandwidths from the LSF coefficients.

We briefly describe the procedure to calculate LSF coefficients. More details of LSF can be referred to [82].

The prediction error filter or the LP analysis filter $\mathrm{B}(z)$ can be expressed in terms of the LPC b_i in Eq. (2.12).

$$\mathrm{B}(z) = 1 - \sum_{i=1}^{p} b_i z^{-i} \tag{2.12}$$

The LSF coefficients are calculated using a symmetric and an anti-symmetric polynomial obtained from $\mathrm{B}(z)$. The symmetric polynomial, $\mathrm{P}(z)$, and the anti-symmetric polynomial, $\mathrm{Q}(z)$, are obtained from $\mathrm{B}(z)$ as given in Eqs. (2.13 and 2.14).

$$\mathrm{P}(z) = \mathrm{B}(z) + z^{-(p+1)}\mathrm{B}(z^{-1}) \tag{2.13}$$

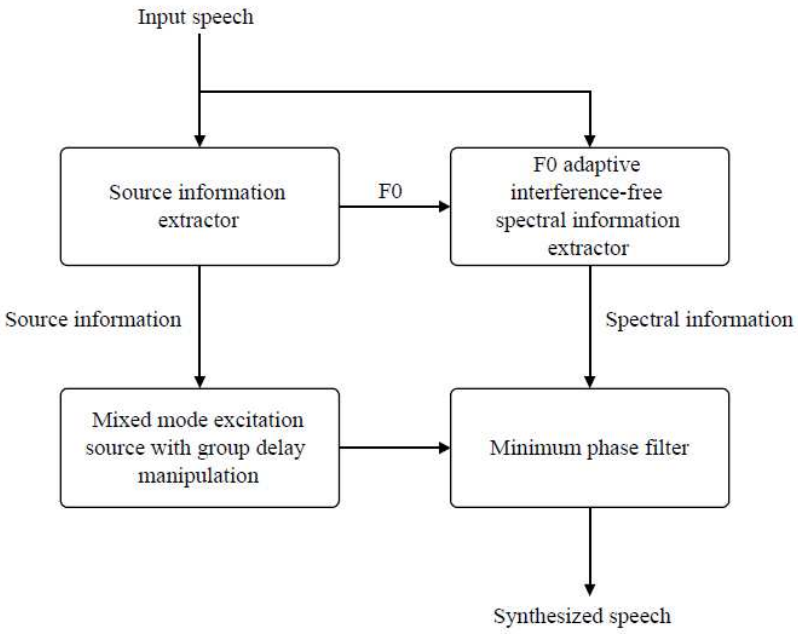

Figure 2.2: Schematic structure of STRAIGHT.

$$\mathrm{Q}(z) = \mathrm{B}(z) - z^{-(p+1)}\mathrm{B}(z^{-1}) \tag{2.14}$$

$\mathrm{P}(z)$ and $\mathrm{Q}(z)$ satisfy three conditions as shown in [83]. Those are:

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

The roots of $\mathrm{P}(z)$ and $\mathrm{Q}(z)$ can be expressed in terms of ω_i or $e^{j\omega_i}$. These ω_i are called the LSFs that can be calculated from Eqs. (2.13 and 2.14) using several methods such as using discrete cosine transformation [83] or Chebyshev polynomials [82].

2.2.4 STRAIGHT

The STRAIGHT [84, 85] is based on the source-filter model that 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. STRAIGHT independently allows for over 600% manipulation of such speech parameters as F0, vocal tract length, and speaking rate, without introducing further degradation due to parameter manipulation [84].

In this sub-section, the main parts of STRAIGHT are briefly described. More details of STRAIGHT can be referred to [84, 85]. Figure 2.2 shows the schematic diagram of STRAIGHT. It consists of three key subsystems:

(1) 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.

(2) 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.

(3) 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 Fast Fourier Transform (FFT).

STRAIGHT can be considered as the state-of-the-art vocoder for SS [44]. Therefore, it was used throughout this research.

2.2.5 Derivation of LSF parameters from STRAIGHT spectrum

In this research, both STRAIGHT and LSF are used. Therefore, the derivation of LSF parameters from STRAIGHT spectrum is required. This sub-section presents the method for this task that was mentioned in [86, 87].

The amplitude spectrum $\mathrm{X}(n)$ 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. (2.15).

$$\mathrm{S}(n) = |\mathrm{X}(n)|^2, \quad 0 \leq n \leq \frac{NF}{2} \tag{2.15}$$

The i^{th} autocorrelation coefficient $\mathrm{r}(i)$ is then calculated using the inverse Fourier transform of the power spectrum as presented in Eq. (2.16).

$$\mathrm{r}(i) = \frac{1}{NF} \sum_{n=0}^{NF-1} \mathrm{S}(n) \exp(j\frac{2\pi ni}{NF}), \quad 0 \leq i \leq NF-1 \tag{2.16}$$

where $\mathrm{S}(n) = \mathrm{S}(NF-n)$ if $\frac{NF}{2} \leq n \leq NF$. Assuming that the speech samples can be estimated by a P^{th} order all-pole model, where $0 < P < NF$, the reconstruction error is calculated as given in Eq. (2.17).

$$\mathrm{G}_L = \mathrm{r}(0) - \sum_{l=1}^{P} b_l^P \mathrm{r}(l) \tag{2.17}$$

where $\{b_l^P\}$, $l = 1, 2, ..., P$ are the corresponding LPC. G_L hereafter is referred to as gain. By minimizing G_L with respect to b_l^P, b_l^P could be estimated by Eq. (2.12). They are then transformed into LSF coefficients by Eqs. (2.12, 2.13 and 2.14).

2.3 Temporal decomposition

2.3.1 Introduction

In articulatory phonetics, speech production is considered as a sequence of overlapping articulatory gestures [73]. Each gesture produces an acoustic event that should approximate a phonetic target. Adjacent gestures overlap, resulting in the transitions between

phonemes that can be observed in almost any parametric representation of the acoustic speech signal. It has long been a difficult task to determine such targets and their temporal evolutionary patterns from the acoustic signal alone.

The so-called *TD - temporal decomposition* [88] method for analyzing speech decomposes speech into targets and their temporal evolutionary patterns. This model takes into account the above articulatory considerations and results in a description of speech in terms of event targets describing the ideal articulatory configurations of the successive acoustic events in speech, and event functions describing their temporal evolutionary patterns. Therefore, it attempts to achieve an optimal transformation from the multi-dimensional spectral parameter space to the phonetic space to be a powerful speech analysis technique.

Although original TD of Atal is mathematically solid, it suffers from its high computational costs and its high sensitivity to the number and locations of events [86, 87, 89]. To overcome drawback of original TD, there are many simplified versions that have been proposed [86, 87, 89, 90, 91]. Among them, MRTD has been shown its efficiency and compactness in many applications, especially for speech modification and voice transformation [17, 18, 19, 20, 21, 22, 23, 86, 87, 92, 93, 94, 95, 96, 97, 98, 99].

With a determination of event functions close to the concept of co-articulation in speech, MRTD can synthesize generally smooth speech. The smoothness in synthesized speech can be adjusted by modifying event targets of MRTD. Therefore, MRTD can be used to synthesize speech with the smoothness close to the "appropriate smoothness" mentioned in chapter 3. As a result, MRTD was used throughout the proposed methods in this dissertation. Specifically, MRTD was the tool for speech modification used in chapter 4 and for speech transformation used in chapters 5 and 6.

This section first presents the original concept of TD by Atal [88], then describes the core concepts and implementation of MRTD [86].

2.3.2 Original method of Atal

Original TD of Atal [88] decomposes a time sequence of speech parameters $\mathbf{y}(n)$ into K dynamic event functions ϕ_k and K static event targets $\mathbf{a}_k$ and $k = 1..K$, as given in Eq. (2.18). Here, $\hat{\mathbf{y}}(n)$ is the approximation of $\mathbf{y}(n)$. As there are K event targets in the total of N frames and $K << N$, then TD is a sparse representation of speech. The event functions are interpolation functions representing the temporal transition movements between sparse event targets.

$$\hat{\mathbf{y}}(\mathbf{n}) = \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{a_k}\phi_{\mathbf{k}}(\mathbf{n}), \quad \mathbf{1} \leq \mathbf{n} \leq \mathbf{N} \tag{2.18}$$

Equation (2.18) can be written in matrix notation as Eq. (2.19), where P is the dimension of the speech parameter. The original TD [88] was proposed for the spectral linear prediction (LP) parameter with order P. However, TD can be used for both the spectral and prosodic parameters of speech [86, 87, 89].

$$\hat{\mathbf{Y}}_{\mathbf{P\times N}} = \mathbf{A}_{\mathbf{P\times K}}\Phi_{\mathbf{K\times N}} \tag{2.19}$$

Event target $\mathbf{a_k}$ and event function ϕ_{k} are unknown and they need to be estimated with optimization tasks to minimize interpolation error. These tasks in the original TD [88] are described as follows.

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 (SVD), as given in Eq. (2.20).

$$\mathbf{Y^T} = \mathbf{UDV^T} \tag{2.20}$$

where ^{T} is the matrix transpose transform, $\mathbf{Y^T}$ is the $N \times P$ spectral parameter matrix, $\mathbf{U}$ is a $N \times P$ orthogonal matrix, $\mathbf{V}$ is a $P \times P$ orthogonal matrix, and $\mathbf{D}$ is a diagonal matrix of eigenvalues.

This allows the event functions to be represented as a linear combination of a set of orthogonal functions as given in Eq. (2.21), 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 $M = 5$.

$$\phi_k(n) = \sum_{i=1}^{M} b_{ki} \mathrm{u}_i(n) \tag{2.21}$$

where $\mathrm{u}_i(n)$ is the element (n, i) of the matrix $\mathbf{U}$ and b_{ki} are a set of coefficients.

Next, the nearest event function, ϕ(n), to the center of the windowed speech segment, $n = n_c$, is evaluated by considering the minimization of a distance measure, $\theta(n_c)$, computed as in Eq. (2.22).

$$\theta(n_c) = \sqrt{\frac{\sum_{n=1}^{N} (n - n_c)^2 \phi^2(n)}{\sum_{n=1}^{N} \phi^2(n)}} \tag{2.22}$$

Minimization of $ln(\phi(n_c))$, with respect to the coefficients c_i leads to an eigenvector problem of a matrix $\mathbf{D} \in \mathbf{D}^{K \times K}$, as described in Eq. (2.23).

$$\mathrm{Dc} = \lambda c \tag{2.23}$$

where the element (i, d) of the matrix $\mathbf{D}$ is given by Eq. (2.24).

$$\mathbf{D}_{id} = \sum_{n=1}^{N} (n - n_c)^2 \mathrm{u}_i(n) \mathrm{u}_d(n) \tag{2.24}$$

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

To analyze a complete utterance the above procedure should be repeated with windows located at intervals throughout 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, $\upsilon(l)$, is incorporated as described in Eq. (2.25).

$$\upsilon(l) = \frac{\sum_{n=1}^{N} (n - l) \phi^2(n)}{\sum_{n=1}^{N} \phi^2(n)} \tag{2.25}$$

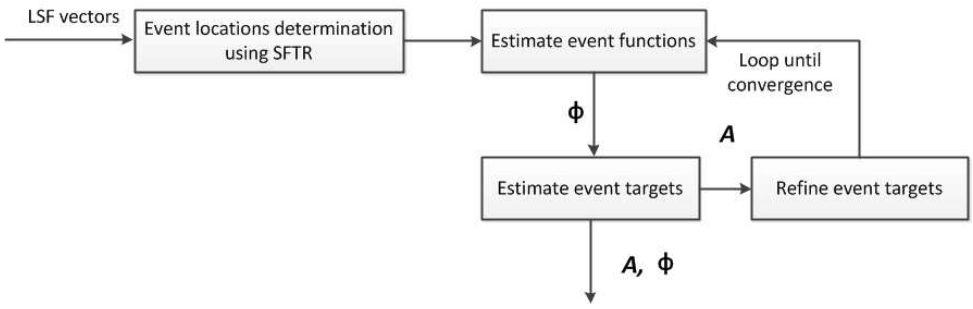

Figure 2.3: Schematic diagram of MRTD.

The function $v(l)$ crosses the $v(l) = 0$ 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 $\mathbf{a_{ik}}$'s, the spectral targets, $\mathbf{a_k}$, are determined in Eq. (2.26).

$$\mathrm{E}_i = \sum_{n=1}^{N} (\mathbf{y}_i(n) - \sum_{k=1}^{K} \mathbf{a}_{ik}\phi_k(n))^2, \quad 1 \leq i \leq P \tag{2.26}$$

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 as in Eq. (2.27).

$$\mathrm{E}_n = \sum_{n=1}^{N} (\mathbf{y}_i(n) - \sum_{k=1}^{K} \mathbf{a}_{ik}\phi_k(n))^2, \quad 1 \leq n \leq N \tag{2.27}$$

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.

After event functions are estimated, event targets are re-estimated as given in Eq. (2.28).

$$\mathbf{A} = \mathbf{Y}\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1} \tag{2.28}$$

2.3.3 Modified restricted temporal decomposition

Assume that the co-articulation in speech production described by the TD model in terms of overlapping event functions is limited to adjacent events, there are only two adjacent overlapping event functions in the second order TD model [90] as given by Eq. (2.29).

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k\phi_k(n) + \mathbf{a}_{k+1}\phi_{k+1}(n), \quad n_k \leq n \leq n_{k+1} \tag{2.29}$$

The restricted second order TD (RTD) model was utilized in [91] 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. Equation. (2.29) can be rewritten as Eq. (2.30).

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k\phi_k(n) + \mathbf{a}_{k+1}(1 - \phi_k(n)), \quad n_k \leq n \leq n_{k+1} \tag{2.30}$$

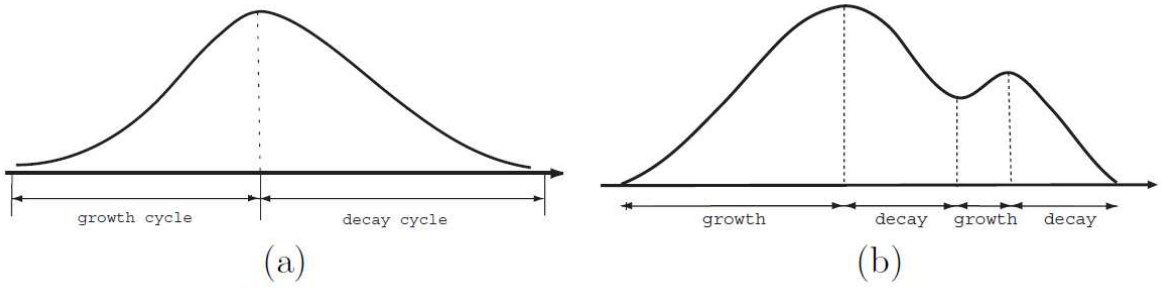

Figure 2.4: Well-shapedness property of MRTD (adopted from [87]): a well-shaped event function (a) and an ill-shaped event function (b)

P.C. Nguyen et al. [86, 87] proposed a improved version of RTD, called MRTD. The schematic diagram of MRTD is shown in Fig. 2.3. The most important improvement of MRTD is the new determination of event functions, in which each event function has only one peak, called the well-shapedness property described in Fig. 2.4. This property is desirable from speech coding point of view because it helps reduce the quantization error of event functions when vector quantized as analyzed in original work of P.C. Nguyen [86, 87]. This property also results in gradual movements of the interpolated spectral and prosodic parameters that are related to the co-articulation of speech. Therefore, MRTD can be also efficiently used to model the co-articulation of speech in the acoustical domain and speech can be analyzed / synthesized by MRTD with an "appropriate smoothness". In addition, the modification on sparse event targets directly and gradually affects to all frames inside duration in which event functions are non-zeros. Hence, speech can be modified / transformed at specific events in the time domain by independently modifying / transforming MRTD event targets and / or MRTD event functions while the smoothness in modified / transformed speech can be still ensured, as shown in [17, 18, 19, 20, 21, 22, 23, 93, 94, 95, 96, 97, 98, 99].

In mathematical form, this determination of event functions in MRTD can be derived in Eqs. (2.31) and (2.32). Here, n_k is the location of event target k and n_{k+1} is that of event target $k+1$, and $\langle ... \rangle$ and $||.||$ correspond to the inner product of two vectors and the norm of a vector.

$$\phi_k(n) = \begin{cases} 1 - \phi_{k-1}(n), \; if \; n_{k-1} < n < n_k \\ 1, \; if \; n = n_k \\ \min(\phi_k(n-1), \max(0, \hat{\phi}_k(n))), \\ if \; n_k < n < n_{k+1} \\ 0, \text{ otherwise} \end{cases} \tag{2.31}$$

$$\hat{\phi}_k(n) = \frac{\langle (\mathbf{y}(n) - \mathbf{a}_{k+1}), (\mathbf{a}_k - \mathbf{a}_{k+1}) \rangle}{||\mathbf{a}_k - \mathbf{a}_{k+1}||^2} \tag{2.32}$$

Other improvements of MRTD including the determination of event locations and the re-estimation of event targets for LSF can be referred to original works of P.C. Nguyen [86, 87].

2.4 GMM-based voice transformation

The statistical GMM-based approach has been considered as the state-of-the-art voice transformation [13, 14, 16, 68, 100, 101, 102, 103], modeling the joint distribution of the source and target spectral or prosodic sequences by a GMM. Although this approach has been popular used, it suffers from several drawbacks, including insufficiently precise GMM models and parameters, over-rough converted parameters between frames, and over-smooth converted frames [18]. Therefore, the MRTD-GMM framework for spectral transformation was proposed [18]. This section first presents the conventional GMM-based voice transformation and then describes the MRTD-GMM framework for spectral transformation.

2.4.1 Conventional GMM-based speech transformation

Speech transformation, usually known as voice transformation or voice conversion, is a task to transform some features of the source speech to those of the target speech, while other features of the source speech are kept unchanged [17].

GMM-based speech transformation was first introduced by Y. Stylianou [13] and was developed later by A. Kain [14] and T. Toda [16]. This method has been widely used until now. There are two phases: training and transformation phases in this system, as shown in Fig. 2.5.

Training GMM [14]

Assume that spectral or prosodic vector of the source and target speech are $\mathbf{x}$ and $\mathbf{y}$ respectively. Joint source-target vector $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N]$ where $\mathbf{z}_n = [\mathbf{x}_i^T, \mathbf{y}_j^T]^T$, $\mathbf{x}_i$ and $\mathbf{y}_j$ are an aligned source-target pair, $1 \leq n \leq N$. The distribution of $\mathbf{Z}$ is modeled by GMM λ, as in Eqs. (2.33) and (2.34).

$$\mathrm{p}(\mathbf{z}|\lambda) = \sum_{k=1}^{K} \alpha_k \mathrm{N}(z; \mu_k, \Sigma_k) = \mathrm{p}(\mathbf{x}, \mathbf{y}), \tag{2.33}$$

$$\mu_k = \begin{bmatrix} \mu_k^x \\ \mu_k^y \end{bmatrix}, \mathbf{\Sigma}_k = \begin{bmatrix} \sum_k^{xx} & \sum_k^{xy} \\ \sum_k^{yx} & \sum_k^{yy} \end{bmatrix} \tag{2.34}$$

where K is the number of Gaussian components.

$\mathrm{N}(\mathbf{z}; \mu_k, \mathbf{\Sigma}_k)$ denotes the 2D dimension normal distribution with the mean $\mu_\mathbf{k}$ and the covariance matrix $\mathbf{\Sigma}_k$. α_k is the prior probability of $\mathbf{z}$ having been generated by component k, and this satisfies $0 \leq \alpha_k \leq 1$, $\sum_{k=1}^{K} \alpha_k = 1$. The parameters $(\alpha_k, \mu_k, \mathbf{\Sigma}_k)$ for the joint density $\mathrm{p}(\mathbf{x}, \mathbf{y})$ can be estimated using the expectation maximization (EM) algorithm [16].

Transformation function for static frame-based values [14]

Assume that the converted spectral or prosodic sequence is $\hat{\mathbf{y}}$, the function to transform source sequence to target sequence is given by Eqs. (2.35) and (2.36).

$$\hat{\mathbf{y}} = F(x) = \sum_{k=1}^{K} \mathrm{p}_k(x). \left[\mu_k^y + \sum\nolimits_k^{yx} (\sum\nolimits_k^{xx})^{-1}(x - \mu_k^x)\right] \tag{2.35}$$

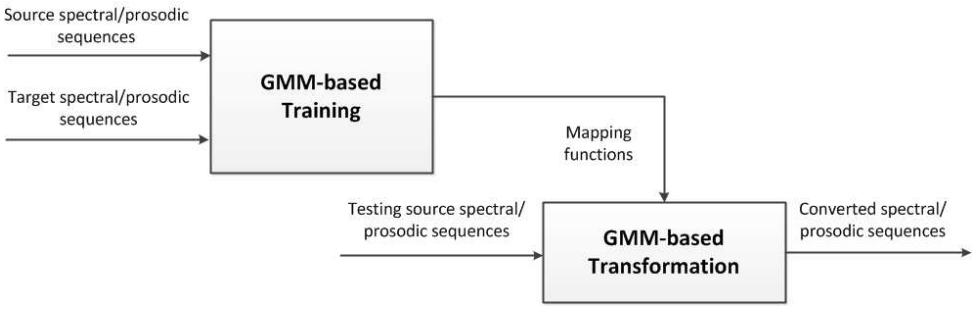

Figure 2.5: General diagram of GMM-based speech transformation.

Where

$$\mathrm{p}_k(x) = \alpha_k.\mathrm{N}(x, \mu_k^x, \sum_k^{xx}) / \sum_{l=1}^{K} \alpha_l.\mathrm{N}(x, \mu_l^x, \sum_l^{xx}) \tag{2.36}$$

is the posterior probability of vector $\mathbf{x}$ belonging to the k^{th} Gaussian.

Transformation function for features combined from static and deltas frame-based values [16]

Assume that the source spectral or prosodic sequence are appended with their deltas to form vector $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, and the joint GMM density model is λ, the suboptimum mixture component sequence $\hat{\mathbf{m}}$ is determined by maximizing the likelihood $\mathrm{p}(\mathbf{X}|m, \lambda)$ as given in Eq. (2.37).

$$\hat{\mathrm{m}} = \arg\max P(m|\mathbf{X}, \lambda) \tag{2.37}$$

Having the suboptimum mixture component sequence, the converted sequence $\hat{\mathbf{y}}$ are determined by maximizing the likelihood $\mathrm{p}(\hat{\mathbf{y}}|\mathbf{X}, m, \lambda)$ with respect to $\hat{\mathbf{y}}$. $\hat{\mathbf{y}}$ therefore is determined as given in Eqs. (2.38, 2.39, 2.40, and 2.41).

$$\hat{\mathbf{y}} = (\mathbf{W}^T\mathbf{D}_m^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{D}_m^{-1}\mathbf{U}_m \tag{2.38}$$

where $\mathbf{W}$ is the matrix for computing the static and delta features [16] and

$$\mathbf{U}m = [\mathbf{U}_1(m_{k1}), \mathbf{U}_2(m_{k2}), ..., \mathbf{U}_N(m_{kN})] \tag{2.39}$$

$$\mathbf{D}_m^{-1} = diag[\mathbf{D}(m_{k1})^{-1}, \mathbf{D}(m_{k2})^{-1}, ..., \mathbf{D}(m_{kN})^{-1}] \tag{2.40}$$

$$\mathbf{U}_n(m_k) = \mu_k^y + \sum_k^{\mathrm{yx}} (\sum_k^{xx})^{-1}(x_t - \mu_k^x) \tag{2.41}$$

$$\mathbf{D}(m_k) = \sum_k^{yy} - \sum_k^{\mathrm{yx}} (\sum_k^{xx})^{-1} \sum_k^{xy}. \tag{2.42}$$

2.4.2 MRTD-GMM spectral transformation

In MRTD-GMM framework [17, 18], spectral parameters such as LSF parameters are decomposed into TD event targets and TD event functions by MRTD. Each phoneme is represented by a specific number of equally-spaced TD event targets, which normally equals five event targets in experiments. In these five TD event targets, two edge event

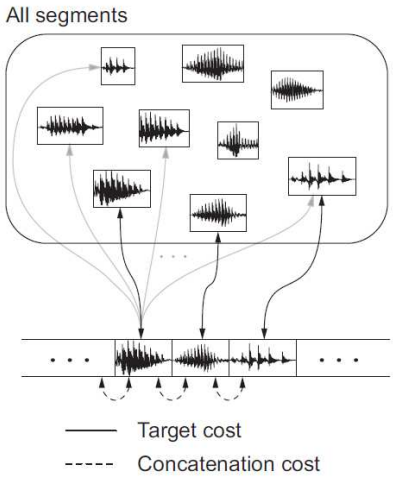

Figure 2.6: Overview of general unit-selection system, adopted from [44].

targets coincide with edge event targets of adjoining phonemes. A vector EV of phoneme-based features of event targets is formulated as given in Eq. (2.43).

$$\mathbf{EV} = \left[\mathbf{a}_1^T, \mathbf{a}_2^T, \mathbf{a}_3^T, \mathbf{a}_4^T\right] \tag{2.43}$$

where $\mathbf{a}_k \quad (1 \leq k \leq 4)$ is the k^{th} event target in each speech segment (a phoneme). To improve the estimation of GMM parameters, the vectors of phoneme-based features of event targets, $\mathbf{x}$ and $\mathbf{y}$, is normalized as in Eqs. (2.44) and (2.45).

$$\mathbf{x} = [\mathbf{a}_{s1}^T, \mathbf{a}_{s2}^T + \pi, \mathbf{a}_{s3}^T + 2\pi, \mathbf{a}_{s4}^T + 2\pi]^T \tag{2.44}$$

$$\mathbf{y} = [\mathbf{a}_{t1}^T, \mathbf{a}_{t2}^T + \pi, \mathbf{a}_{t3}^T + 2\pi, \mathbf{a}_{t4}^T + 2\pi]^T \tag{2.45}$$

where $\mathbf{a}_{sk}, \mathbf{a}_{tk}$ are the k^{th} event targets in each phoneme of the source and target speakers, respectively. The joint vector of event targets between source and target speakers $\mathbf{Z}$ is formulated as in Eq. (2.46).

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_q] \tag{2.46}$$

where $\mathbf{z}_i = [\mathbf{x}_i^T, \mathbf{y}_i^T]$ and $\mathbf{x}_i, \mathbf{y}_i$ are event target sets of i^{th} phoneme of source speaker and the corresponding event target of the target speaker, respectively.

The transformation phase is the same with that in the conventional GMM-based approach, except that the vectors for the transformation procedure are the sets on normalized phoneme-based features.

2.5 Concanenative speech synthesis

CSS is based on the concatenation of segments of recorded speech. State-of-the-art CSS produces the most natural-sounding synthesized speech up to now. However, CSS has still several drawbacks such as the requirement of huge amount of data or the inflexibility for voice transformation. CSS was studied in this research. Hence, this section presents basic common procedures in CSS. Hunt and Black [38, 39] presented a selection model for US, the state-of-the-art CSS, described in Fig. 2.6. This is the most popular and

general selection model for US. The basic notion is that of a target cost, i.e., how well a candidate unit from the database matches the required unit, and a concatenation cost, which defines how well two selected units combine. The definition of target cost between a candidate unit, $\mathbf{u}_i$, and a required unit, $\mathbf{t}_i$, is given in Eq. (2.47).

$$\mathrm{C}^{(t)}(\mathbf{t}_i, \mathbf{u}_i) = \sum_{j=1}^{p} \omega_j^{(t)} \mathrm{C}_j^{(t)}(\mathbf{t}_i, \mathbf{u}_i) \tag{2.47}$$

where j indexes over all features (phonetic and prosodic contexts are typically used). The concatenation cost is defined shown in Eq. (2.48).

$$\mathrm{C}^{(c)}(\mathbf{u}_{i-1}, \mathbf{u}_i) = \sum_{k=1}^{q} \omega_k^{(c)} \mathrm{C}_k^{(c)}(\mathbf{u}_{i-1}, \mathbf{u}_i) \tag{2.48}$$

where k may include spectral and acoustic features.

These two costs have to be optimized to find the string of units, $\mathbf{u}_{1:n} = \mathbf{u}_1, ..., \mathrm{u}_n)$, from the database that minimizes the overall cost, $\mathrm{C}(\mathbf{t}_{1:n}, \mathbf{u}_{1:n})$, as given in Eqs. (2.49) and (2.50).

$$\hat{\mathbf{u}}_{1:n} = \arg\min_{\mathbf{u}_{1:n}} \{\mathrm{C}(\mathbf{t}_{1:n}, \mathbf{u}_{1:n})\} \tag{2.49}$$

$$\mathrm{C}(\mathbf{t}_{1:n}, \mathbf{u}_{1:n}) = \sum_{i=1}^{n} \mathrm{C}^{(t)}(\mathbf{t}_i, \mathbf{u}_i) + \sum_{i=2}^{n} \mathrm{C}^{(c)}(\mathbf{u}_{i-1}, \mathbf{u}_i) \tag{2.50}$$

CSS produces the most natural-sounding synthesized speech up to now. However, differences between natural variations in speech and the nature of the segmenting techniques cause some audible glitches in the output [43]. Prosodic modeling for CSS is also a difficult task, resulting in the instability of prosodic trajectories in speech synthesized by CSS [104]. CSS requires a large-scaled speech corpus for concatenation. Under limited data conditions, the quality of CSS drastically reduced. In the Blizzard Challenge 2006 which provided a large-scaled speech corpus with amount of 5 h of speech, the best system was based on CSS. However, in Blizzard Challenge 2005 which provided a smaller speech corpus with amount of 1.5 h of speech, a HMMSS [44] outperformed the well-established CSS systems in both speech quality and intelligibility. Additionally, transforming the synthesized voices in CSS is not flexible compared with that in HMMSS.

The main reason affecting to the reduction on quality of CSS under limited data conditions is the mismatch-context errors due to the limitation of the data. Mismatch-context errors increase the temporal discontinuities in synthesized speech that reduce its quality [61, 62, 63, 64].

More details on CSS and US can be found in [38, 39].

2.6 HMM-based speech synthesis

HMMSS has been widely studied for the two last decades [15, 44, 45, 46, 47, 48, 49]. In HMMSS, spectral and prosodic features of speech are modeled and generated in a unified statistical framework using HMMs.

Figure 2.7 is a block diagram of the HMMSS.

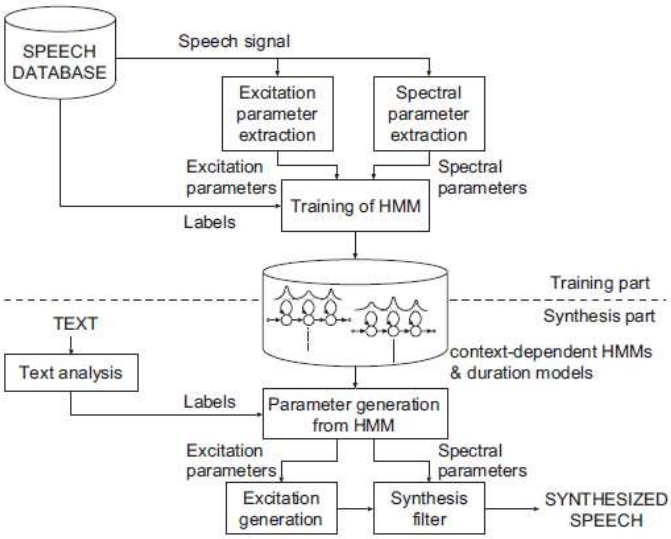

Figure 2.7: Overview of general HMM-based speech synthesis, adopted from [44].

In the training stage of a HMMSS, a maximum likelihood (ML) criterion is usually used to estimate the model parameters as shown in Eq. (2.51).

$$\hat{\lambda} = \arg\max_{\lambda}\{\mathbf{p}(\mathbf{O}|\mathbf{W},\lambda)\} \tag{2.51}$$

where λ is a set of model parameters, $\mathbf{O}$ is a set of training data, and $\mathbf{W}$ is a set of phonetic unit sequences corresponding to $\mathbf{O}$.

Speech parameters, $\mathbf{o}$, for a given phonetic unit sequence are generated in the synthesis stage, $\mathbf{w}$, from the set of estimated (trained) models, $\hat{\lambda}$, to maximize their output probabilities as given in Eq. (2.52).

$$\hat{\mathbf{o}} = \arg\max_{o}\{\mathrm{p}(\mathbf{o}|\omega,\hat{\lambda})\} \tag{2.52}$$

In the final step in HMMSS, a speech waveform is synthesized from the parametric representations of speech by using vocoders such as Mel Log Spectrum Approximation (MLSA) filter or STRAIGHT.

HMMSS has many advantages shown in the literature.

A decision-tree based context clustering technique has been used to ensure the synthesized trajectory is smooth and stable with limited amount of training data [45]. Therefore, the intelligibility of synthesized speech can be high even with limited training data. This is one major advantage of HMMSS. HMMSS can simultaneously transform the voice characteristics of synthetic speech into those of a target speaker using a small amount of target data by utilizing "average-voice-based" methods [15]. Therefore, it is flexible to adapt the synthetic voice into different target individual voices or target speaking styles with limited amounts of target data. Another convincing advantage of HMMSS is its small footprint because this technique only stores the model parameters instead of the original speech data itself. HMMSS also requires less computation than US and real-time ability can be done with such systems. Last but not least, HMMSS is able to flexibly change its

voice characteristics, speaking styles, and emotions by using the concept "average voice" [15].

Although HMMSS has many advantages as shown above, HMMSS still has problems. In HMMSS, the structure of the estimated spectrum corresponds to the average of different speech spectra in the training database since the use of the mean vector. The detail structure in the original speech is missing in this kind of approximation. This characteristic in speech synthesized by HMMSS can be considered "too medial" or "over-smooth" (or sometimes "too stable" [114]) as shown in many researches [46, 47, 48].

For speech recognition, "too medial" or "over-smooth" parameters of HMM are still enough to represent data distribution of unit such as phoneme since it is unnecessary to consider perception or quality of model parameters [48]. For speech synthesis, the synthesized speech averagely approximated by HMMSS is still highly intelligible as shown in the literature [44, 45]. However, synthesized speech is muffled and its naturalness is far from natural due to the mismatches between the smoothness in synthesized speech and the "appropriate smoothness" in the original speech. As a result, "over-smoothness" is a main remaining problem in HMMSS.

HMMSS was studied in this research, in which methods to reduce the over-smoothness of HMMSS was proposed. More details on HMMSS can be found in [15, 16, 44, 45, 46, 48, 49].

2.7 Evaluation of speech synthesis

This sections describes methods for evaluating synthesized speech.

Synthesized speech can be compared and evaluated with main respect to intelligibility and naturalness [3]. In simple applications such as reading machines for the blind, the speech intelligibility is more important than the naturalness. However, the speech naturalness is essential in multimedia applications or electronic mail readers. There are both subjective and objective tests for evaluating synthesized speech.

2.7.1 Objective evaluations

Some conventional objective methods, such as articulation index (AI) or speech transmission index (STI), have been used to evaluate quality of synthesized speech [105]. These methods can be useful to evaluate the synthesized speech transmitted through some transmission channels. However, they are not suitable for evaluating synthesized speech in general because not only the acoustic characteristics, but also high-level parts determine the final quality of synthetic speech [105].

Recently, the speech intelligibility index (SII) has also used as an objective measure for intelligibility, while the perceptual evaluation of speech quality (PESQ) has also used as an objective measure for naturalness of synthetic speech [106]. Although the experimental results show that SII and PESQ scores correlate with subjective scores, the reliability is still not ensured and *these kinds of measurement were not used in this research.*

2.7.2 Subjective evaluations

The most popular and reliable measure for evaluating synthetic speech is using subjective listening tests [107, 108, 109, 110]. The subjective evaluation can be made at both word

Table 2.1: Scales in MOS

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

level such as using diagnostic rhyme test (DRT) and modified rhyme test (MRT) [107], or sentence level such as semantically unpredictable sentences (SUS) [107, 108, 109].

When repeating the test procedure to the same listening group, the test results may increase significantly by the learning effect [111]. Therefore, considering learning effect is important in listening tests.

Currently, the most widely used to evaluate speech intelligibility is SUS, and that to evaluate speech naturalness is mean opinion score (MOS). Therefore, they were used in this research.

In SUS, the words to be tested are selected randomly from a pre-defined list of possible candidates. These are mostly mono-syllabic words with some expectations. As the sentence set is not fixed, the SUS-test is not sensitive to the learning effect.

MOS is probably the most widely used and simplest method to evaluate speech quality in general. It is also suitable for overall evaluation of synthetic speech. MOS is a five level scale from bad (1) to excellent (5). The listener's task is simply to evaluate the tested speech with scale described in Table. 2.1 below.

SUS and MOS tests have been also used in recent Blizzard Challenges [112], which is a well-known international forum for evaluating synthesized speech.

2.7.3 Evaluation challenges

The consistent evaluation of SS is difficult to be ensured because subjects for each listening test are usually different while there is a lack of universally-agreed and reliable objective evaluation criteria. In addition, different groups often use different speech data. The quality of synthesized speech also largely depends on the quality of the original corpus and on the facilities used to replay the speech.

Therefore, recently, some researchers have started to evaluate SS using a common speech dataset with common subjects by joining international forums such as the Blizzard Challenges [49, 112, 113]. Unfortunately, only speech data of high-resourced languages is provided in Blizzard Challenges.

2.8 Summary

In this chapter, the backgrounds for this dissertation were presented. The first section presented concepts of speech information, source/filter models of speech production, the LP model, and the STRAIGHT. The second section described TD and MRTD, which was the tool for speech modification used in chapter 4 and for speech transformation used in chapter 5. The third section presented the GMM-based speech transformations including conventional method and the MRTD-GMM spectral transformation, which were developed for the proposed tone transformation presented in chapter 6. The fourth section,

which was also the last section of this chapter, described basic concepts in the two state-of-the-art SS, CSS and HMMSS, and methods for evaluating synthesized speech.

Chapter 3

Speech Smoothness Measures

The unified purpose of this research is to propose methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech for improving quality of synthesized speech under limited data conditions. Under this unified purpose, four objectives are specified as mentioned in chapter Introduction. This chapter aims to solve the first issue of this dissertation, which is to propose a speech smoothness measure that can be applied to evaluate the smoothness of synthesized speech and to control synthesized speech with an "appropriate smoothness". In this chapter, the method using GV of static spectral features is introduced analyzed. Then, the proposed global speech smoothness measure using the square sum of the variance of the delta-delta sequence, named DGSM, is explained and discussed.

3.1 Introduction

Although the definition of smoothness in speech has not been clearly mentioned, it is implicit in many researches [45, 46, 47, 60, 61, 62, 63, 64, 65, 66, 114, 115, 116, 118], in which smoothness in speech can be considered as a result of transitions in speech. The slower transitions cause the smoother speech and the more rapid transitions cause the rougher speech.

The articulators typically move slowly during speech production [60]. Therefore, speech features of natural speech are generally smooth and "over-roughness" can reduce the naturalness of synthesized speech. However, rapid changes in speech features naturally occur in some cases such as in plosives [60]. These are the "natural over-roughness" or "natural discontinuities" in speech features. Besides the "natural over-roughness", several kinds of "unexpected over-roughness" in speech, such as discontinuities caused by mismatch-context errors in CSS or discontinuities caused by noisy recording environments, can reduce the naturalness of synthesized speech [61, 62, 63, 64], or of recorded speech [65].

On the contrary, over-smoothness of synthesized speech (sometime refers to over-stability in synthesized speech [114]), which is a result of too slow transitions in synthesized speech, also reduces the naturalness of synthesized speech due to several reasons presented below.

Over-smoothness causes the "muffleness" in synthesized speech [45, 46, 47] that affects its naturalness.

The degree of articulation (DoA) provides information on the style / personality [115].

DoA is characterized by modifications of the phonetic context, of the speech rate, and of the spectral dynamics (vocal tract rate of change). Over-smooth speech with too-slow transitions may affect to produce the appropriate DoA and the important information on style / personality may be loss.

Over-smooth may be acceptable for reading speech or neutral speech but not suitable for expressive speech in general. Mainly, the range and the velocity of the tongue tip movements are the primary modulation parameters associated with emotional expression [116]. Therefore, too smooth, or too stable speech with slow movements can be not efficient to represent some kinds of emotional speech with high movements of the tongue tip.

Lieberman and Michaels [66] found that smoothing the F0 contours could reduce the recognition rate of the emotion of speech. The smoother speech, the lower recognition rate of the emotion of speech.

Over-smoothness can eliminate F0 and spectral fluctuations that are important in singing voice synthesis and perception [28, 29, 30, 31, 117] and in expressive speech synthesis and perception [118].

While linguistic information in speech is critical for its intelligibility, non-linguistic information in speech, i.e. emotion, expression, individuality, and so on is important to perceive its naturalness. Therefore, over-smoothness reduces the naturalness of synthesized speech.

Consequently, both over-smoothness and over-roughness can reduce the naturalness of synthesized speech. Therefore, instead of synthesizing too smooth or too rough speech, the "optimal smoothness" that naturally exists in the original speech has to be reached to ensure the naturalness of synthesized speech.

In this research, the concept "appropriate smoothness" in speech was defined as the smoothness that approximates the "optimal smoothness" naturally existed in the original human speech. With an "appropriate smoothness", speech is supposed to be natural. This "appropriate smoothness" depends on the content of speech and is different between vowels and consonants. It also depends on the observed speech features. For instance, the "appropriate smoothness" of a spectral feature differs from that of a prosodic feature such as F0 contour.

The concept of "appropriate smoothness" was used through this research in all proposed solutions as a criteria to ensure the naturalness of synthesized speech.

3.2 Measuring speech smoothness using the global variance

In the literature, statistical variances of static spectral features have been widely used as the measures of smoothness of synthesized speech [47] or noisy speech [119, 120]. By generating speech parameter with the GV close to that of original speech, the synthesized speech has been expected to be natural [47]. This means that synthesized speech is expected to has an "appropriate smoothness" if it has GV of static spectral features close to those of the original speech. By minimizing the variance in smoothed signal PSD, smoothing methods for noisy speech enhancement were proposed [119, 120]. The enhanced speech was smoother and its quality was significantly enhanced.

GV for a static spectral feature in time and spectral domain are defined as given in

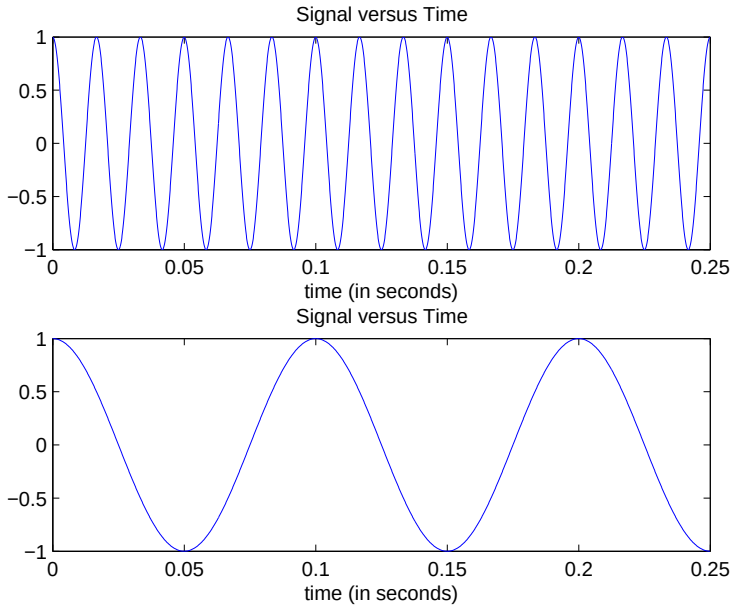

Figure 3.1: Two signals with the same mean and the same variance but different smoothness.

Eqs. 3.1 and 3.2.

$$GV_t = \frac{1}{P}\sqrt{\sum_{p=1}^{P}(\mathrm{var_t}(p))^2} \tag{3.1}$$

$$GV_s = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\mathrm{var_s}(i))^2} \tag{3.2}$$

where var_t and var_s are the variance in time domain and spectral domain of the spectral feature, P is the dimension of the feature, N is the length in time domain of the feature.

The distance between GV of synthesized speech and that of the original speech can be used to measure the closeness between the smoothness of synthesized speech and that in the original speech. Here, the distances between GV of a synthesized static feature and GV of a static feature of the original speech (GV*) are given in Eqs. 3.3 and 3.4.

$$DGV_t(x) = \frac{GV_t * -GV_t(x)}{GV_t*} \tag{3.3}$$

$$DGV_s(x) = \frac{GV_s * -GV_s(x)}{GV_s*} \tag{3.4}$$

Although variances of static spectral features can measure smoothness in speech features, they do not exactly represent the smoothness of speech features. Figure. 3.1 gives an example of two signal that have the same mean and the same variance but different smoothness. Figure. 3.2 shows an example of a LSF sequence synthesized by HMMSS with and without GV. Two drawbacks of GV of static spectral features on measuring the smoothness of the feature are shown in this figure: although increasing GV of static spectral features can make synthesized speech closer to the original speech, the synthesized

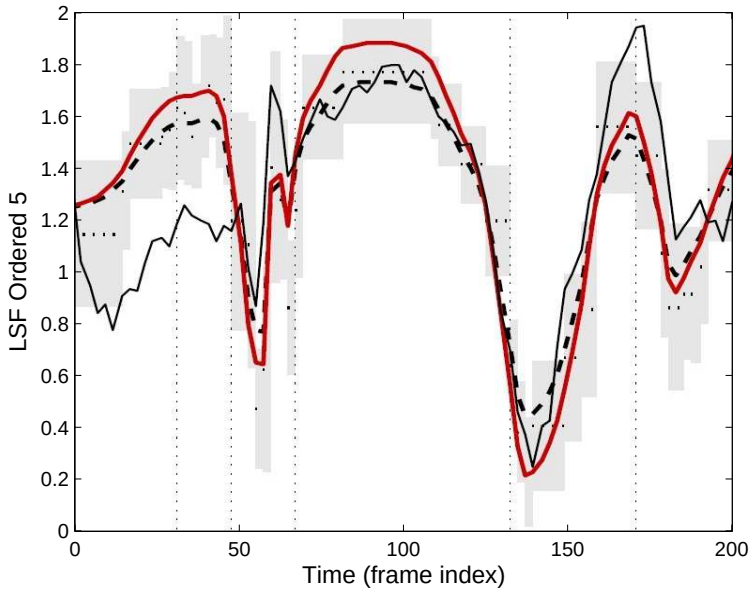

Figure 3.2: A LSF sequence synthesized by HMMSS with (red curve), without GV (dashed black curve), and the original one (solid black curve): the shades in both sides of the sequence are the standart deviation.

speech is still over-smooth compared with the original speech; the estimated GV may be inaccurate due to some reasons such as the limitation of the training data.

Due to the limitation of statistical variances of static spectral features on measuring the smoothness of speech features, to propose a new and efficient speech smoothness measure is important in the field of speech synthesis and speech processing in general.

3.3 The proposed speech smoothness measure

Smoothness in signals, time series is a result of transitions in the signals and time series. Researches on mathematics, time series analysis, and signal processing also show that the smoothness of a curve, a time series, a signal, or a feature of a signal can be measured based on the "curvature" of the envelope of the function, the time series, the signal, or the feature [121, 122, 123, 124]. The "curvature" are usually computed by using the second-order derivative of the curve, the time series or the signal [122, 125, 126]. Using second-order derivative can also represent the transition, or the changing rate in speech. Therefore, second-order derivative was used to define a measure of smoothness in speech. In time series analysis and discrete signal processing, it has been revealed that delta of delta (delta-delta) can represent a simple second-order derivative [122, 127, 128]. Therefore, the square sum of the variance of the delta-delta sequence was used to define the "global" speech smoothness measure (GSM) in this research, which is based on the formulation to measure a "global" smoothness for time series presented in [122].

Suppose that we need to measure the smoothness of a speech feature $X_{P\times N}$. $P > 1$ for a spectral feature such as LSF and $P = 1$ for a prosodic feature such as F0 contour.

$$X_{P\times N} = \{X_i^p\}, p = 1..P, i = 1..N \tag{3.5}$$

In one-dimensional feature ($P = 1$) such as F0, the smoothness can be only measured and observed in time domain named temporal smoothness of this research. However, in a P-dimensional feature (with $P > 1$) such as in LSF, the smoothness can be measured and observed in both time and spectral domain, named temporal and spectral smoothness of this research.

Measures of temporal smoothness and spectral smoothness of speech features with $P > 3$ were defined in the next Sub-sections.

3.3.1 The proposed temporal speech smoothness measure

Delta in time domain for spectral sequence $p^{th}, p = 1..P$ is given in Eq. 3.6.

$$\Delta t(X)_{P\times(N-1)} = \{X_{i+1}^p - X_i^p\}, p = 1..P, i = 1..N-1 \tag{3.6}$$

Its delta-delta in time domain is described in Eq. 3.7.

$$\Delta^2 t(X)_{P\times(N-2)} = \{\Delta t_{i+1}^p - \Delta t_{t_i}^p\}, p = 1..P, i = 1..N-2 \tag{3.7}$$

The variance of delta-delta in time domain for spectral sequence $p^{th}, p = 1..P$ is represented in Eq. 3.8.

$$Var\{\Delta^2 t(p)\} = \frac{1}{N-2}\sum_{i=1}^{N-2}(\Delta^2 t_i^p - \overline{\Delta^2 t^p})^2 \tag{3.8}$$

where

$$\overline{\Delta^2 t^p} = \frac{1}{N-2}\sum_{i=1}^{N-2}\Delta^2 t_i^p \tag{3.9}$$

Finally, the global smoothness measure (GSM) in time domain (temporal GSM) is defined in Eq. 3.10.

$$GSM_t = \frac{1}{P}\sqrt{\sum_{p=1}^{P}(Var\{\Delta^2 t(p)\})^2} \tag{3.10}$$

3.3.2 The proposed spectral speech smoothness measure

Delta in spectral domain for frame i^{th} is described in Eq. 3.11.

$$\Delta s(X)_{(P-1)\times N} = \{X_i^{p+1} - X_i^p\}, p = 1..P-1, i = 1..N \tag{3.11}$$

Its delta-delta in spectral domain is represented in Eq. 3.12.

$$\Delta^2 s(X)_{(P-2)\times N} = \{\Delta s_i^{p+1} - \Delta s_i^p\}, p = 1..P-2, i = 1..N \tag{3.12}$$

The variance of delta-delta in spectral domain for frame i^{th} is given in Eq. 3.13.:

$$Var\{\Delta^2 s(i)\} = \frac{1}{P-2}\sum_{p=1}^{P-2}(\Delta^2 s_i^p - \overline{\Delta^2 s_i})^2 \tag{3.13}$$

where

$$\overline{\Delta^2 s_i} = \frac{1}{P-2}\sum_{p=1}^{P-2}\Delta^2 s_i^p \tag{3.14}$$

Table 3.1: DGV in time domain of speech analyzed / synthesized by STRAIGHT and synthesized by HMMSS with GV

	STRAIGHT	HMM with GV
Mean	-0.007	-0.016
95% Confidence	0.016	0.038

Table 3.2: DGV in spectral domain of speech analyzed / synthesized by STRAIGHT and synthesized by HMMSS with GV

	STRAIGHT	HMM with GV
Mean	0.01	-0.019
95% Confidence	0.0007	0.007

Finally, the GSM in spectral domain (spectral GSM) is defined in Eq. 3.15.

$$GSM_s = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(Var\{\Delta^2 s(i)\})^2} \tag{3.15}$$

Based on the definitions of temporal and spectral GSM (GSM_t, GSM_s) in Eqs. 3.10 and 3.15, it reveals that the smaller GSM, the smoother speech feature, and the larger GSM, the rougher speech feature. The GSM equals zero if the feature is flat, which is the ideal most-smoothness. As a result, GSM can be used to measure the smoothness of a speech feature. However, as analysis in Section 3.1, instead of synthesizing too smooth or too rough speech, an "appropriate smoothness" has to be reached to ensure the naturalness of synthesized speech. Therefore, instead of directly using the GSM, the distance between GSM of synthesized speech and GSM^*, which is the GSM of the corresponding original speech, is proposed to measure the smoothness of the synthesized speech.

These distance of GSM (DGSM) in time and spectral domain are defined in Eqs. 3.16 and 3.17.

$$DGSM_t(x) = \frac{GSM_t^* - GSM_t(x)}{GSM_t^*} \tag{3.16}$$

$$DGSM_s(x) = \frac{GSM_s^* - GSM_s(x)}{GSM_s^*} \tag{3.17}$$

Following these definitions, it reveals that:

- If DGSM is positive, the synthesized speech is smoother than the original speech.
- If DGSM is negative, the synthesized speech is rougher than the original speech.
- If the absolute of DGSM is close to zero, the synthesized speech has an "appropriate smoothness".
- If DGSM is positive and its absolute is large, the synthesized speech is too smooth (over-smooth).
- If DGSM is negative and its absolute is large, the synthesized speech is too rough (over- rough).

In the next section, DGSM will be used to measure the smoothness of speech synthesized by some popular speech coders and synthesizers, as well as to measure the smoothness of speech mixed with noises to confirm the reliability of this proposed measure.

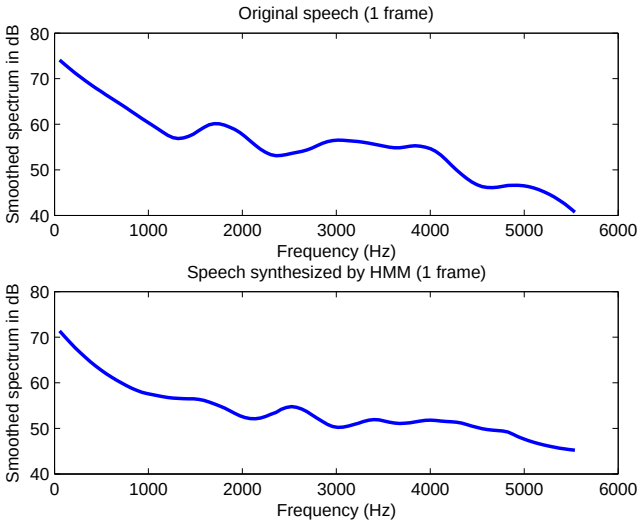

Figure 3.3: Smoothed spectrum of speech synthesized by HMMSS and the original speech.

Table 3.3: DGSM in time domain of speech analyzed / synthesized by STRAIGHT and synthesized by HMMSS with GV

	STRAIGHT	HMM with GV
Mean	-0.0809	0.5534
95% Confidence	0.01	0.0243

3.4 Examples of using the proposed speech smoothness measure

This section presents some examples of using DGSM to measure the smoothness of speech synthesized by HMMSS with speech analyzed / synthesized by STRAIGHT. A set of 100 Vietnamese utterances extracted from corpus DEMEN567 [72] was analyzed / synthesized by STRAIGHT and was compared with 100 utterances synthesized by a HMMSS with GV for Vietnamese [135] trained also with DEMEN567.

The results in Tables. 3.1, 3.2 show that DGV of speech synthesized by HMMSS in both time and spectral domains are very small and are equivalent with speech analyzed / synthesized by the high-quality STRAIGHT. However, speech synthesized by HMMSS with GV is still over-smooth in time domain as shown in Fig. 3.2. It is also over-smooth in spectral domain as shown in Fig. 3.3. Additionally, subjective evaluations in many researches show that speech synthesized by HMMSS is still muffled and over-smooth [43, 48, 49]. Therefore, it supports that GV is not efficient to measure the smoothness of synthesized speech.

The results in Tables. 3.3, 3.4 show that DGSM of speech analyzed / synthesized by STRAIGHT is very small. Therefore, speech analyzed / synthesized by STRAIGHT is very close with the original speech in term of smoothness. However, DGSM of speech

Table 3.4: DGSM in spectral domain of speech analyzed / synthesized by STRAIGHT and synthesized by HMMSS with GV

	STRAIGHT	HMM with GV
Mean	-0.098	0.2985
95% Confidence	0.018	0.0231

synthesized by HMMSS is positive and its absolute is very large. Therefore, speech synthesized by HMMSS is shown to be over-smooth in both time and spectral domains when measuring by DGSM. These results supported the reliability and the efficiency of DGSM for measuring the smoothness of synthesized speech.

3.5 Summary

This chapter showed the need of a speech smoothness measure to evaluate methods of speech synthesizing. The disadvantages of using GV to measure the speech smoothness was described. Then, the proposed DGSM using the square sum of the variance of the delta-delta sequence to measure the speech smoothness was presented and discussed. Examples of using DGSM showed that it is reliable and efficient to measure smoothness of different kinds of synthesized speech.

Chapter 4

Methods to improve quality of CSS under limited data conditions

The unified purpose of this research is to propose methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech for improving the naturalness of synthesized speech under limited data conditions. Under this unified purpose, four objectives are specified as shown in chapter Introduction. The aim of this chapter is to solve the second issue related to the second objective of the dissertation, which is to solve the main problem of CSS under limited data conditions by reducing mismatch-context errors in CSS to ensure an "appropriate smoothness" in synthesized speech.

This chapter presents the proposed model for the contextual effect in CSS, the proposed methods for speech modification and for unit selection, and the proposed CSS under limited data conditions. Finally, the experimental results are described and discussed.

4.1 Introduction

The articulators typically move smoothly during speech production [60]. Therefore, speech features of natural speech are generally smooth and "over-roughness" can reduce naturalness of synthesized speech. The rapid changes in speech features still occur in some cases such as in plosives [60]. These are the "natural over-roughness" or "natural discontinuities" in speech features. Besides the "natural over-roughness", there are several kinds of "unexpected over-roughness" in speech synthesis. One of the most popular "unexpected over-roughness" is the discontinuity caused by mismatch-context errors in CSS, which can significantly reduce the naturalness in synthesized speech [62]. Under limited data conditions, mismatch-context errors in CSS significantly increase. Therefore, solving the mismatch-context error is a main task to ensure an "appropriate smoothness" and the naturalness in speech synthesized by CSS under limited data conditions.

Approaches to reducing mismatch-context errors in CSS have been widely studied [61, 62, 63, 64]. The conventional method is by using linear spectral smoothing [61]. There are also other more sophisticated methods [62, 64]. Mismatch-context errors occur in all frames inside the co-articulated transition region between two adjacent phonemes. However, current methods are unable to localize this co-articulated transition region. In addition, the mismatch-context errors are not equal for all positions inside the co-articulated transition region. Mismatch-context errors are gradually changed in the ideal

case of co-articulation [73] from the maximum value at the phoneme boundary of two neighboring phones to the minimum value at the onset/offset of the co-articulated transition region. Although some methods have attempted to weight these mismatch-context errors [64], their temporal evolution have been not ensured due to the lack of a co-articulation model in CSS. Therefore, the results obtained by D. T. Chappell [63] showed that conventional linear spectral smoothing [61] was still the most reliable and efficient method for reduce mismatch-context errors. Consequently, it is still difficult to efficiently solve mismatch-context errors. Therefore, a model of the co-articulated transition region between two adjacent phonemes using MRTD is proposed in this chapter, where background on TD and MRTD was already presented in section 2.3 of chapter 2. A unit selection and a unit modification based on the proposed model are also proposed and are combined in the proposed CSS.

Building SS under limited data conditions is more practical with under-resourced languages, where there are a few small public speech corpora. As the tonal-monosyllabic Vietnamese is an under-resourced language as presented in appendix A, Vietnamese datasets were used in this paper for evaluations.

4.2 Modeling co-articulated transition region between phonemes in CSS

4.2.1 General model

The basic supposition in the proposed model is the assumption that each phoneme can be divided into one nuclei interval and two co-articulated transition intervals at two sides. The existence of the stationary and quasi-stationary intervals inside vowels, semi-vowels and vowel-like consonants has been confirmed [130, 131]. General Locus theory [73] suggests that there is also a nuclei interval inside a non-vowel-like (or non-stationary) consonant, referred to as the narrow region around the ideal articulatory target of the consonant. These nuclei intervals are referred to as pseudo-stationary intervals in this research due to similarities between their behaviors and those of stationary and quasi-stationary intervals under the effect of co-articulation. All of the stationary, quasi-stationary and pseudo-stationary intervals of phonemes, called nuclei intervals for short, in the proposed model are supposed to be context-less-sensitive and can be preserved for concatenation within different contexts.

The spectral transition measure (STM) [132] and MRTD [86, 87, 92] are respectively used to determine the positions and the durations of the nuclei and transition intervals of phonemes, and to interpolate speech parameters and to modify the joint transition intervals.

The context-sensitive co-articulated transition region between adjacent phonemes in the proposed model is described by the MRTD event targets and the overlapping MRTD event functions restricted by the two event targets located at the onset and offset of the co-articulated transition region shown in Fig. 4.1.

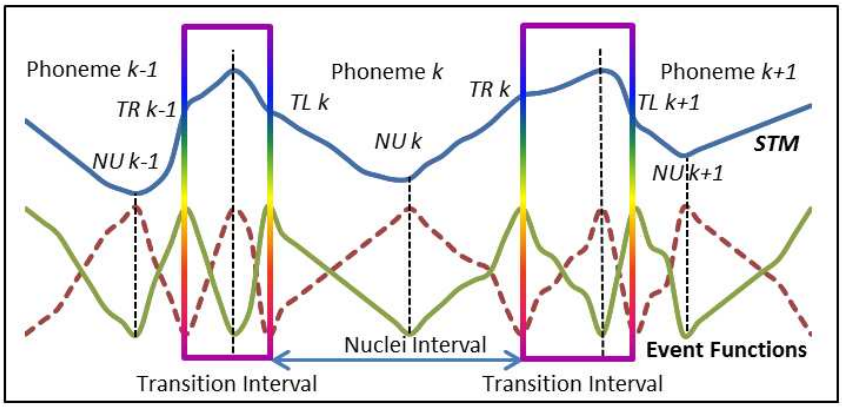

Figure 4.1: Modeling contextual effects using TD, STM and folded STM (FSTM): PBs are phoneme boundary points extracted from label data, $N_u s$ are nuclei points, $T_r s$ are onsets and $T_l s$ are offsets of joint transition regions.

4.2.2 Estimating co-articulated transition regions and TD event locations

Previous research on the co-articulation of speech has revealed that the transition movements caused by co-articulation can be observed by analyzing the transitions of formant frequencies [73]. STM [132], which is one representation of the first-order derivative of the spectral sequence, has also been used to detect the spectral transition rates of speech. STM can be used with any spectral parameter. The STM of LSF is used in this study to estimate the nuclei and co-articulated transition intervals due to the close relations between LSF and formant frequency.

The STM at time t, STM(t) is defined in Eqs. (4.1) and (4.2), where time t refers to the location of the frame in the time domain.

$$\text{STM}(t) = \frac{\partial \mathbf{LSF}}{\partial t} = (\sum_{i=1}^{P} \text{a}_i^2)/P \tag{4.1}$$

where

$$\text{a}_i = (\sum_{n=-n0}^{n0} \mathbf{LSF}_i(n).n)/(\sum_{n=-n0}^{n0} n^2) \tag{4.2}$$

Here $\mathbf{LSF}_i(n)$ is LSF coefficient $i^{th}, (1 < i < P)$, in frame n^{th} inside a window whose center is time t, and $-n_0 < n < n_0$. The regression coefficient a_i, corresponds to the linear variation in the spectral envelope pattern in a unit time. Consequently, STM(t), which is the mean-square value of $a_i, i = 1..P$, corresponds to the variation in the spectral envelope smoothed by polynomial fitting.

The nuclei interval for interpolation with TD is represented by one central event target with the location determined based on a criteria that maximize the stability of spectral transition [87], which is referred to as the location where STM is minimized. The three-step algorithm used to estimate this central event is:

Step 1. Initialize window size n0 = 1.

Step 2. Detect local minima of STM. If there is 1 minima, return, else move to step. 3

Step 3. - If there is more than 1 minima, increase n0 = n0 + 1, and return to step. 2

- If there are no minima, the location of the central event target is determined as the central location of the phoneme.

When moving from a stable nuclei interval to a dynamic transition interval (and in the inverse case), the rate at which speech is changing is maximum at the onset and offset of the dynamic interval. Therefore, the onset and offset of the co-articulated transition region are estimated based on a criteria that maximize the dynamics of spectral transition, referred to as locations where FSTM, which is one representation of the second-order derivative of spectral sequence $\partial^2\mathbf{LSF}/\partial t^2$, is maximized.

$$\text{FSTM}(t) = \frac{\partial^2 \mathbf{LSF}}{\partial t^2} = (\sum_{m=-m0}^{m0} \text{STM}(m).m) / \sum_{m=-m0}^{m0} m^2 \tag{4.3}$$

The three-step algorithm to estimate the two events located at the onset and offset of the joint co-articulated transition region is:

Step 1. Initialize window size m0 = 1.

Step 2. Detect local maxima of FSTM(t). If there is 1 maxima, return, else move to step. 3

Step. 3.

- If there is more than 1 maxima, increase m0 = m0 + 1, and return to step. 2

- If there are no maxima, the locations of outermost event targets are determined as the central locations of the left and right half-phonemes.

A total of three true event targets, including one central event target where STM is minimized and two event targets where FSTM are maximized, are used to interpolate each phoneme with TD. Two pseudo-targets at phoneme boundaries are also used to represent, select, and modify the co-articulated transition region, as explained in the next sections.

4.2.3 Representing co-articulated transition regions with pseudo-targets

A simple and unique parameter representing a co-articulated transition region is necessary to easily modify the joint co-articulated transition regions.

Following Eqs. (2.17) and (2.19) in chapter 2, the acoustical parameters for the joint transition region between two units L (left), which is restricted in locations from $\text{n}_\text{L}(K-1)$ to $\text{n}_\text{L}(K)$, and R (right), which is restricted in locations from $\text{n}_\text{R}(1)$ to $\text{n}_\text{R}(2)$, of one concatenation are represented in Eqs. (4.4) and (4.5), and are described in Fig.4.2. Here, $\text{n}_\text{L}(i)$ and $\text{n}_\text{R}(j)$ return the locations (frame indexes) of the targets i^{th} and j^{th} of the left and right units, respectively. Note that the indexes of $\mathbf{y}$ are the locations of frames, those of $\mathbf{a}$ are the locations of the sparse event targets, and the first-ordered indexes of ϕ are the locations of event targets and the second-ordered indexes of ϕ is the locations of frames.

The pseudo-event-targets are the two outermost events: event K^{th} of left unit L and event 1^{th} of right unit R. Therefore, the pseudo-target-vectors of left unit L and right unit R are $\mathbf{a_L}(K)$ and $\mathbf{a_R}(1)$.

$$\begin{aligned}\mathbf{y_L}(\text{n}_\text{L}(K-1) : \text{n}_\text{L}(K)) = \mathbf{a_L}(K-1, K) \\ \times \phi_L(K-1 : K, \text{n}_\text{L}(K-1) : \text{n}_\text{L}(K))\end{aligned} \tag{4.4}$$

$$\mathbf{y_R}(\text{n}_\text{R}(1) : \text{n}_\text{R}(2)) = \mathbf{a_R}(1, 2) \times \phi_R(1 : 2, \text{n}_\text{R}(1) : \text{n}_\text{R}(2)) \tag{4.5}$$

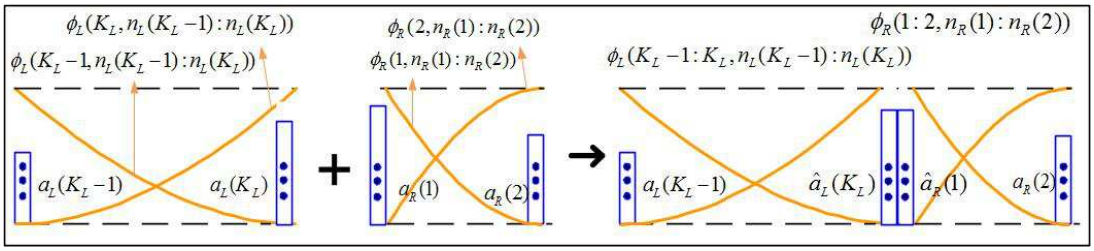

Figure 4.2: Modify joint transition regions of left unit L (left panel) and right unit R (central panel) for concatenation of unit $L+R$ (right panel): ϕ_L and ϕ_R are event functions of units L and R; $\mathbf{a_L}$ and $\mathbf{a_R}$ are event targets of units L and R; the two pseudo-targets $\mathbf{a_L}(K_L)$ and $\mathbf{a_R}(1)$ are averagely modified to $\hat{\mathbf{a}}_\mathbf{L}(K_L)$ and $\hat{\mathbf{a}}_\mathbf{R}(1)$.

Following the determination of event functions of MRTD [87], at the locations of event targets, the event function in the current interval approximates to one and other event functions approximate to zeroes. Therefore, the re-estimated target vectors, followed Eq. 2.20 in chapter 2, are just slightly different from the frame-based vectors at the same locations. However, while modifying frame-based vectors just affects to these frames, when two pseudo-targets $\mathbf{a_L}(K)$, $\mathbf{a_R}(1)$ are modified, all frames in the transition parts of left unit L and right unit R will be gradually modified respectively, derived from Eqs. (4.4) and (4.5). Therefore, while pseudo-targets and frame-based vectors are equivalent when computing concatenation costs for CSS presented in Section 4.3, only pseudo-targets can be used as unique parameters for the modification task, presented in Section 4.4.

4.3 Proposed phoneme-based selection cost with pseudo-targets

In conventional unit selection CSS, the most important part is the methods of selecting units that uses both the target cost and concatenation cost [38]. The concatenation cost is to find the best match units which join together most smoothly. The target cost is to find units in the database which best match their target predictions. It is clear that the actual target speech units are unknown in synthesis stage. Thus, the target predictions are usually determined as the centroids of the clusters of phonetic units existed in the database [39]. The conventional methods of unit clustering, such as CART method [39], need sufficient number of candidates for each unit and significant differences between candidates to ensure the accuracy of the clustering algorithm [39]. However, under limited data conditions, the number of candidates for each unit is usually small, resulting in the inaccuracy of the unit clustering and the target prediction. Therefore, determining an efficient target cost under limited data conditions is difficult and target cost is not considered in this work. Although the lack of target cost for unit selection may reduce quality of synthesized speech, especially in the timing and segmental duration, this research focuses on reducing the mismatch-context error for CSS and the use of only concatenation cost can be still sufficient to observe the efficiency of the proposed method.

Conventional concatenation cost includes both the cost of spectral features and prosody features such as F0 and power [38]. Hence, this research also uses concatenation costs that are computed by the distances between two pseudo-targets for LSF, F0, and power enve-

lope gain (PL) of two joint phonemes (or two joint boundary phonemes of non-uniform units). Equation (4.6) describes the summed concatenation cost with a set of three acoustical parameters of LSF, F0, and PL.

$$\mathrm{C} = \omega_{LSF}\mathrm{c}_{LSF} + \omega_{F0}\mathrm{c}_{F0} + \omega_{PL}\mathrm{c}_{PL} \tag{4.6}$$

The component costs $\mathrm{c}_{LSF}, \mathrm{c}_{F0}$ and c_{PL} are computed in Eqs. (4.7), (4.8) and (4.9). The $\omega_{LSF}, \omega_{F0}$ and ω_{PL} are weighted factors that can be chosen from experiments.

$$\mathrm{c}_{LSF} = \frac{|\mathbf{a}_{L\_LSF}(K) - \mathbf{a}_{R\_LSF}(1)|}{\pi} \tag{4.7}$$

$$\mathrm{c}_{F0} = \frac{|log(\mathbf{a}_{L\_F0}(K)) - log(\mathbf{a}_{R\_F0}(1)|)}{max(log(\mathbf{a}_{F0}))} \tag{4.8}$$

$$\mathrm{c}_{PL} = \frac{|\mathbf{a}_{L\_PL}(K) - \mathbf{a}_{R\_PL}(1)|}{max(\mathbf{a}_{PL})} \tag{4.9}$$

4.4 Proposed phoneme-based method of modifying co-articulated transition regions

This is the core of the proposed CSS under limited data conditions that can smooth the joint co-articulated transition region to obtain synthesized speech with an "appropriate smoothness".

Since modifying two pseudo-targets $\mathbf{a_L}(K)$ and $\mathbf{a_R}(1)$ can modify all frames in the transition regions of the two phonemes L and R as shown in Section 4.2.3, the modification of the co-articulated transition region here is the averagely modification of the two pseudo-targets for each acoustical feature given in Eqs. (4.10), (4.11).

$$\Delta_{X_i} = \frac{\mathbf{a}_{R\_X_i}(1) - \mathbf{a}_{L\_X_i}(K)}{2} \tag{4.10}$$

$$\mathbf{a}_{L\_X_i(K)} = \mathbf{a}_{L\_X_i(K)} + \frac{\Delta_{X_i}}{2}, \mathbf{a}_{R\_X_i(1)} = \mathbf{a}_{R\_X_i(1)} - \frac{\Delta_{X_i}}{2} \tag{4.11}$$

Here, $i = 1..3$, X_1 is LSF, X_2 is F0 and X_3 is PL. The smooth shapes of the two event functions, estimated by MRTD, ensure gradual changes in all frames in the modified co-articulated transition region [86, 87]. Therefore, the proposed modification can approximate the recovery of the smooth transition between adjacent phones that occurred in their original contexts. Then, synthesized speech can have an "appropriate smoothness".

Component thresholds $\delta_{F0}, \delta_{LSF}$ and δ_{PL} for the decision to modify the targets of F0, LSF, and PL are determined from experiments to avoid the modification of the joint concatenated phonemes or units that are already smooth.

4.5 Proposed TD-based CSS with non-uniform units

Signal modification can reduce the quality of original speech. Thus, natural concatenation should be preserved as much as possible. The proposed TD-based CSS uses non-uniform units, including basic phoneme and two longer units: syllables and initial/final (I/F)

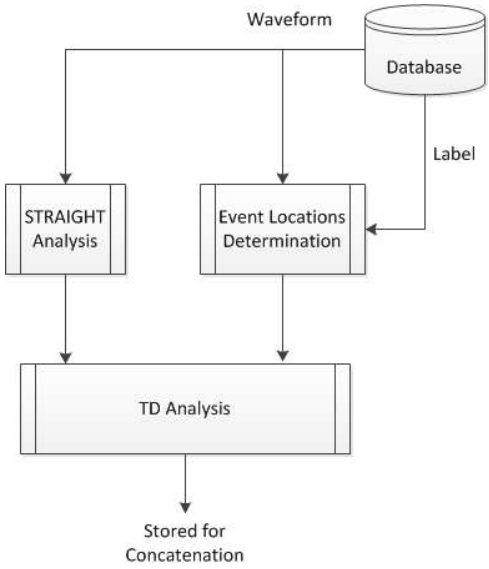

Figure 4.3: Offline stage of proposed TD-based CSS.

unit [135], which is similar to the onset / rhythm unit [133]. Syllables and the I/F unit are natural concatenations of phonemes. Therefore, the proposed phoneme-based unit selection can be directly applied to selecting non-uniform units, in which the selection process is undertaken with the pseudo-targets of the boundary phonemes of non-uniform units. The proposed phoneme-based unit modification can also be directly applied to modifying non-uniform units, in which the modification process is carried out with the co-articulated transition regions of the boundary phonemes of non-uniform units. The use of non-uniform units also improves the use of the proposed TD-based CSS under limited data conditions because the possibility of selecting units in existence contexts is maximized.

There are general diagrams of the proposed TD-based CSS in Figs. 4.3, 4.4. The phoneme-based TD event functions and targets of the original database are computed and stored in the offline pre-synthesis phase. A text searching algorithm is processed at the first stage of the synthesis phase with a parse tree based on a criteria to maximize the lengths of the units, or minimize the number of the units for concatenation [133]. After a path of units with a minimum-number of units is found from text searching, the optimal path of units with the minimum weighted sum of concatenation cost is selected based on dynamic programming. The phoneme-based TD-based modification for each concatenation pair will be run for each parameter if its concatenation cost is larger than a component threshold. Finally, the high-quality vocoder STRAIGHT [84] is used to synthesize output speech utterances.

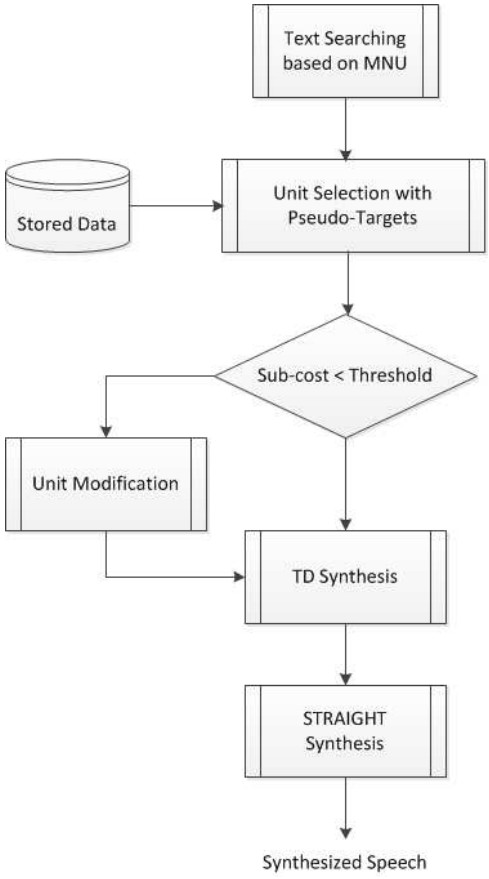

Figure 4.4: Synthesis stage of proposed TD-based CSS.

4.6 Implementations and evaluations

4.6.1 Data preparation

For implementations

A dataset in this study is considered to be under "limited data conditions" if it reaches the threshold when the phoneme coverage is approximately 100 %, as explained in section 1.6 of chapter 1. This requirement means that although all phonemes exist, their frequencies of occurrence are limited. Therefore, the possibility of selecting matching-context units for concatenation is small, and the role of modification tasks is more important.

A "limited data condition" with a dataset of 300 utterances extracted from DEMEN567, was simulated by taking this requirement into account. The tonal Vietnamese phoneme coverage is nearly 100 %. Although some monophones are still missing, most of widely used tonal phonemes appear in this dataset. The size of this dataset in WAV PCM 16 bit format is approximately 30MB and its duration is approximately 20 minutes.

This dataset was used for concatenation in the proposed TD-based CSS and it was also used to concatenate the two non-uniform unit selection CSS for Vietnamese that were used to compare with the proposed CSS. The first CSS did not have linear spectral smoothing, which is referred to as CSS A, and the second had linear spectral smoothing, which is referred to as CSS B in this study.

For evaluations

SUS has been used as a standard measure to evaluate the intelligibility of speech synthesis, but there are no designs on Vietnamese SUS lists at present. Therefore, 20 testing sentences were chosen to evaluate intelligibility with four restricted rules (rules 1–4) that prevented the subjects from easily predicting meanings, and two restricted rules (rules 5–6) that ensured the evaluation were reliable. The six rules were:

(1) The Vietnamese words in the testing sentences were all low frequency,

(2) Only sentences composed of monosyllabic words were used to prevent subjects from predicting the meanings of compound words from their constituent parts,

(3) Repeating the words between testing sentences was avoided to prevent subjects from remembering words they had heard previously,

(4) Sentences with fewer semantic relations were selected to prevent subjects from predicting the meanings of sentences,

(5) Sentences covering all Vietnamese tones that minimized the repetition of tonal phonemes were selected,

(6) Only short sentences were selected to avoid the difficulty for subjects to remember syllables that they had heard in each testing sentence.

These 20 testing sentences were chosen from a set of sentences that were not used to concatenate the proposed CSS and two conventional unit selections of CSS A and B.

Another testing dataset of 20 short sentences was used to evaluate naturalness, which was not used for concatenating the three CSSs. Since prosody trajectories were not controlled in the three CSSs, we focused on improving the quality of synthesized speech in terms of the smoothness in synthesized speech. Therefore, both testing sentences for evaluating the intelligibility and the naturalness were short sentences with a few words.

Additionally, 100 Vietnamese syllables were also extracted from the original DEMEN567 for the objective test conducted to measure the smoothness of synthesized speech. All of 100 Vietnamese syllables used in the objective evaluation were consonant - vowel (CV) syllables.

4.6.2 Experimental conditions and parameters

As analysis in section 4.3, it is difficult to determine an efficient target cost for predicting prosodic information under limited data conditions. Therefore, all the proposed CSS, CSS A, and CSS B also used only concatenation cost for selecting units. Same as in the proposed CSS, three acoustical parameters of LSF, F0, and PL were used to compute the concatenation costs. The concatenation cost for CSS A and CSS B approximate to that in the proposed CSS as explained in Section 4.2.3.

Although there are some sophisticated smoothing methods in CSS [62, 64], the linear spectral smoothing by Paliwal [61] is still one of the most popular and efficient methods for smoothing CSS [63]. Therefore, linear spectral smoothing with LSF interpolation was adopted to build CSS B in this research, in which two new frames at both sides of each concatenation were interpolated from the boundary frames (anchor frames) and were inserted as in [63].

The frame lengths were 20 ms and the frame steps were 5 ms in there CSSs. The orders of LSF in three CSSs were 24. The w_{LSF}, w_{F0} and w_{PL} corresponded to 0.8, 0.05 and 0.15 in Eq. 4.7, these weighting coefficients were also adopted to the compute concatenation costs of CSS A and CSS B. The component thresholds for the decision on modification

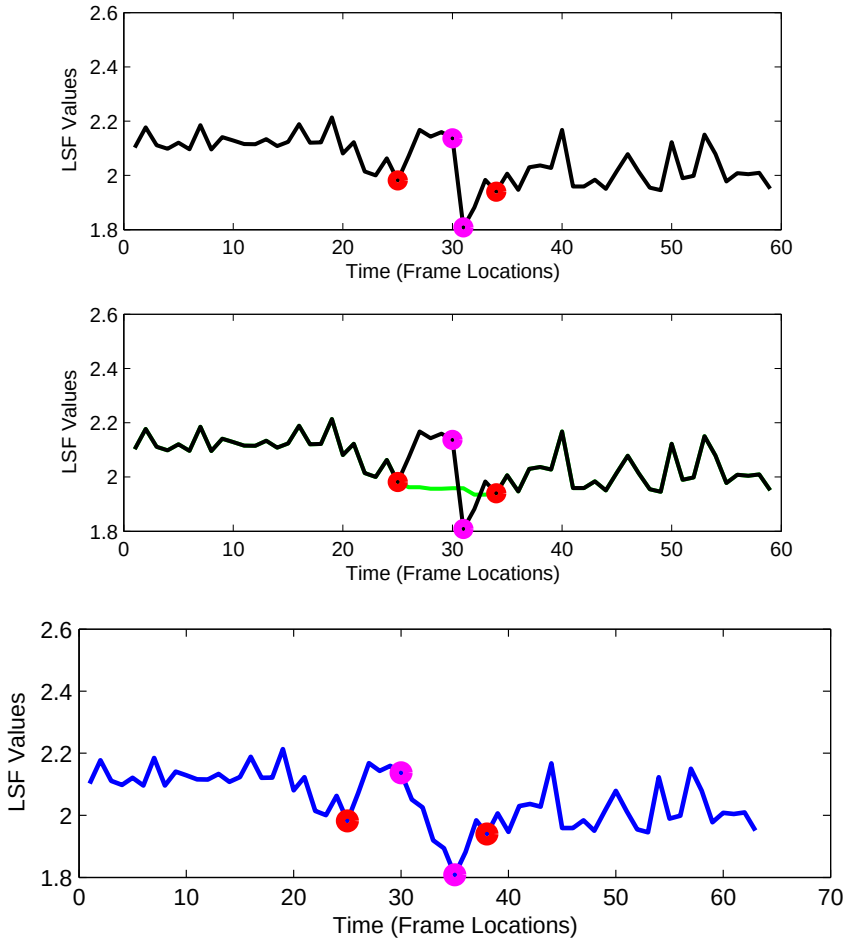

Figure 4.5: Methods of LSF smoothing: the black curve is the LSF sequence with raw concatenation; the green curve is the modified sequence inside the co-articulated transition region; the blue curve is the sequence with a interpolated region of 4 frames using linear LSF smoothing

in the proposed TD-based CSS $\delta_{LSF} = 0.01$, $\delta_{F0} = 0.1$ and $\delta_{PL} = 0.01$. $Max(\mathbf{a}_{F0}) = 800$ and $Max(\mathbf{a}_{PL}) = 0.1$ in Eqs. 4.8, 4.9.

STRAIGHT version 4 [84] was used as a vocoder to generate the output waveform of three CSSs.

Since both the text-searching algorithms are same for all three CSSs, the methods of unit selection in the three CSSs are similar, and as STRAIGHT was used for all CSSs with the same manner, the differences on the performances of the three CSSs can be mostly caused by the modification methods inside the these CSSs.

4.6.3 Objective evaluations

Since the main improvement in the proposed CSS is the MRTD-based speech modification task to reduce the mismatch-context error, this sub-section present the evaluation of the proposed modification method compared with the baseline method, which is the linear LSF smoothing [61]. DGSM, presented in chapter 3 was used to measure the smoothness

Table 4.1: Temporal DGSM of synthesized LSF sequence

	Raw Concatenation	Linear LSF Smoothing	MRTD-based Smoothing
Mean	-0.369	-0.243	0.171
95% Confidence	0.066	0.021	0.016

Table 4.2: Spectral DGSM of synthesized LSF sequence

Spectral DGSM	Raw Concatenation	Linear LSF Smoothing	MRTD-based Smoothing
Mean	0.0126	0.0225	0.0183
95% Confidence	0.001	0.0027	0.0035

of the modified speech.

100 Vietnamese syllables with four versions, including raw concatenations, speech modified with the linear LSF smoothing [61], speech modified with the proposed MRTD-based smoothing, and the original speech, were used for this evaluation. The input of DGSM is the LSF sequences were extracted from the centers of the consonants to the centers of the vowels inside syllables to emphasize the differences between your versions of speech in the concatenation regions.

The results are shown in tables. 4.1 and 4.2. It reveals that the absolute of temporal DGSM with the proposed MRTD-based smoothing method is smallest. Therefore, the proposed MRTD-based smoothing method improved the temporal smoothness compared with the baseline method.

Spectral DGSM scores were small with all versions of speech. Therefore, all speech version had "appropriated smoothness" in spectral domain. This result is reasonable because mismatch-context in CSS is a temporal phenomenon and the discontinuities caused by mismatch-context errors just occur in time domain.

An example of a raw concatenation LSF sequence, a sequence modified by the linear LSF smoothing, and a sequence modified by the proposed MRTD-based smoothing is shown in Fig. 4.5.

4.6.4 Subjective evaluations

Subjective tests on intelligibility and naturalness were conducted with five subjects who were native Vietnamese speakers with normal hearing. The intelligibility scores were measured by using WER while MOS was used to evaluate the naturalness of CSSs. The results obtained from evaluating intelligibility are summarized in Table. 4.3 and they indicate that the WERs of the proposed TD-based CSS were minimal compared with CSS A and CSS B. A F-test of the statistical analysis of variance between groups (ANOVA) was computed to evaluate the discrimination between CSS A, CSS B, and the proposed TD-based CSS. The results $F = 1302.169$ with $p < 0.001$ show that the discrimination was significant. Therefore, the proposed TD-based CSS was vastly superior to both CSS A and B.

The results from evaluating naturalness are presented in Fig. 4.6, where the proposed TD-based CSS was also superior to both CSS A and B. A F-test was computed to show the differences between MOSs of speech synthesized by the three CSSs and those of the original speech. The results are shown in Table. 4.4, which indicate that speech synthesized by the proposed TD-based CSS was closer to the original speech compared with that by CSS

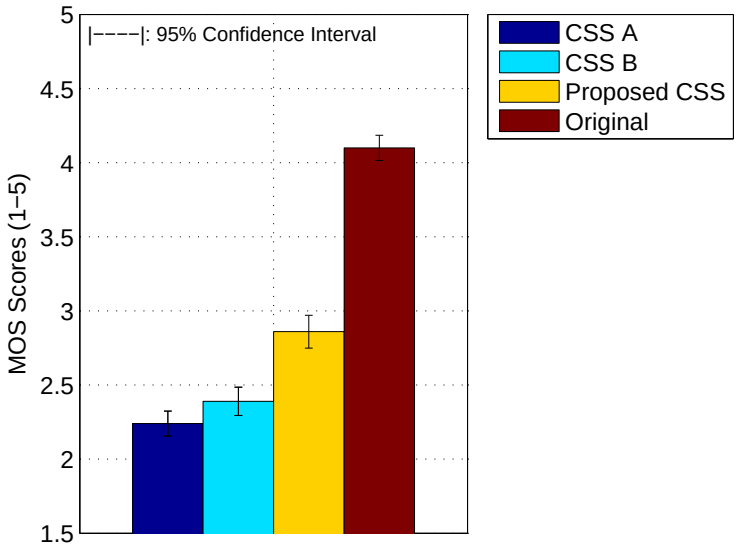

Figure 4.6: MOS scores (CSS A is without linear smoothing and CSS B is with linear smoothing)

Table 4.3: Word Error Rates for CSSs(%)

	CSS A (without linear smoothing)	CSS B (with linear smoothing)	Proposed CSS	Original
Mean	10.33	10.83	5.5	0
95% confidence	0.88	0.80	0.36	0

A and CSS B. Therefore, the proposed TD-based CSS significantly outperformed both CSS A and CSS B.

The results from both the intelligibility and naturalness evaluations confirmed that the proposed CSS efficiently reduced the mismatch-context errors in concatenations, and the proposed CSS ran efficiently under limited data conditions.

The results also show that CSS B just slightly outperformed CSS A. Thus, the linear spectral smoothing was not sufficiently efficient. On the contrary, the modification method inside the proposed CSS, which improved the smoothness in concatenated speech, showed its efficiency in terms of both intelligibility and naturalness.

4.7 Conclusions

A non-uniform CSS with methods of reducing mismatch-context errors was proposed. The experimental results with Vietnamese datasets revealed that the proposed modification method inside the proposed CSS was efficient to concatenated speech with an "appropriate smoothness". The proposed CSS convincingly outperformed two unit selection CSSs with and without speech modifications in terms of naturalness and intelligibility. As a

Table 4.4: F-test to show the difference between MOSs of speech synthesized by the three CSSs and those of the original speech

	CSS A	CSS B	Propoded TD-based CSS
F	914.029	674.222	296.548
p	< 0.001	< 0.001	< 0.001

consequence, the proposed CSS efficiently reduced the mismatch-context errors in concatenations, and the proposed CSS ran efficiently under limited data conditions. Then, the second objective of this dissertation, aforementioned in introduction chapter, was solved.

Chapter 5

Improving naturalness of HMMSS under limited data conditions

The unified purpose of this research is to propose methods of speech modification and transformation to ensure an‘ ‘appropriate smoothness" in synthesized speech for improving the naturalness of synthesized speech under limited data conditions. Under this unified purpose, four objectives are specified as shown in chapter Introduction. The aim of this chapter is to solve the third issue related to the third objective of the dissertation, which is to improve the naturalness of HMMSS under limited data conditions with an "appropriate smoothness". To achieve this objective, a hybrid SS between HMMSS and MRTD, referred to as HTD in this study, was proposed to reduce over-smoothness in speech synthesized by HMMSS under limited data conditions. Background on MRTD was already presented in section 2.3 of chapter 2. This chapter first presents overview of the proposed HTD. After that, the target selection procedure to transform speech synthesized by HMMSS into the original speech is described. The differences between the proposed HTD and the HTT, which is the current state-of-the-art hybrid synthesis approach, are also clearly specified. Finally, the experimental results are presented and discussed.

5.1 Introduction

Although HMMSS has many advantages as was previously mentioned in section 2.6 of chapter 2, HMMSS still has problems. In HMMSS, the structure of the estimated spectrum corresponds to the average of different speech spectra in the training database due to the use of the mean vector. In this case, the spectrum estimated by HMMs is an average approximation of all corresponding speech spectra in the training database. The detail structure in the original speech may be missing in this kind of approximation. This characteristic in speech synthesized by HMMSS can be considered "too medial", or "over-smooth", or sometimes "too stable" in synthesized speech.

For speech recognition, "too medial" or "over-smooth" parameters of HMM are still enough to represent data distribution of unit such as phoneme since it is unnecessary to consider perception or quality of model parameters [48]. For speech synthesis, the "over-smooth" speech synthesized by HMMSS is still highly intellligible as shown in the literature [44, 45]. In Blizzard Challenge 2010, 2011 and 2012 evaluations [49, 112, 113], the intelligibility of HMMSS is comparative to that of the original speech. However, "over-smoothness" causes the mismatches between synthesized speech and the original speech

[49, 112, 113]. Therefore, speech synthesized by HMMSS is "muffled" and far from natural [45, 46, 47, 48]. Additionally, "over-smoothness" causes a reduction in identification emotions / expressions / styles in speech [66] that can also affect to the perception of the naturalness. As a consequence, "over-smoothness" in generated parameters is still a main remaining problem in HMMSS.

Although both the spectral and prosodic trajectories generated by HMMSS are over-smooth, the effect of over-smoothness in a spectral sequence is more serious since the structures of spectral features are more complex, as shown in [48].

There have been many studies attempting to solve the problem of average parameter generation in HMMSS. Using multiple mixtures for modeling state output probability density can reduce over-smoothness in synthesized speech [46]. However, these methods cause another problem with over-training due to the increased number of model parameters. A method of combining continuous HMMs with discrete HMMs, and a method of increasing the number of HMM states has also reduced the over-smoothness in HMMSS [48]. However, these methods increase the complexity of HMMs and are not convenient in practical synthesis systems. The state-of-the-art method for reducing over-smoothness and improve the accuracy of the parameter generation in HMMSS is the method of parameter generation that take into consideration GV [47]. In this method, a GV model was trained to model the variation of parameter trajectories at utterance level. The generated parameter sequence maximizes a likelihood based on not only an HMM likelihood but also a GV likelihood, resulting in a reduction of the GV of the generated parameter trajectories. This method proved to significantly reduce the over-smoothness in synthesized speech. However, due to the disadvantages of GV as shown in section 3.2 of chapter 3, the naturalness and the individuality of recent HMMSS using GV is still not high compared with well-established USs and hybrid SSs in Blizzard Challenge 2010, 2011 and 2012 evaluations [49, 112, 113].

Parameter generation in HMMSS is mainly affected by the accuracy of model estimates and that of the training algorithm [48]. These factors are affected by the amount of training data [49]. The larger the amount of training data, the more accurate the model estimates and the training algorithm, and the more accurate parameter generation in HMMSS. As a result, the effect of average parameter generation becomes more serious in a situation with limited training data. Therefore, it is difficult to ensure the naturalness of HMMSS under limited data conditions.

Hybrid approaches between HMMSS and unit selection, such as HTT [43], have recently been studied as another solution to improving the naturalness of HMMSS and to preserving the high intelligibility of HMMSS. The HMM trajectory is used to guide the selection of each 5ms frame to concatenate the waveforms in HTT. This method was based on the concept that transforming/replacing short frames of speech synthesized by HMMSS to the physically closest frames of original speech. The naturalness of HTT is comparable to that of unit selection and its intelligibility is comparable to that of HMMSS and the original speech. In Blizzard Challenge 2010 evaluations [43, 112], the intelligibility of HTT got the first-ranked, the similarity to the original speech of HTT got also the first-ranked, the naturalness of HTT got third-ranked compared with several CSS and HMMSS that attended the evaluation. Additionally, this speech synthesis is language-independent due to the use of short frames instead of phonetic-level-units. However, HTT still has drawbacks. The major one is the use of short frames, which requires a perfect selection process. If the selection process is imperfect due to a limited data corpus, it

may be easy to perceive discontinuities between frames. As a result, this synthesizer still requires a huge amount of data for rendering. Another drawback with HTT is its high computational load. The searching task to select the matched short frames in a huge database is not convenient in most personal hardware platforms. Additionally, HTT is not able to preserve other advantages of HMMSS such as its flexibility for voice adaption or transformation.

Based on the above considerations, the objective for this chapter is to improve the naturalness of HMMSS with an "appropriate smoothness" in the original speech. As shown in section 2.3 of chapter 2, event functions of MRTD can ensure the synthesized speech smooth. The smoothness can be adjusted by modifying event targets of MRTD. Previous studies [18], [23], [132], [134] have also found that event functions of TD can represent the "languageness" or content information of speech, which is important to perceive speech intelligibility, while sparse event targets can convey the non-linguistics of style information such as speaker individuality and the emotion in speech, which is important to perceive the naturalness of speech [18], [23], [92], [134]. Therefore, these two components are independently controlled in the proposed hybrid synthesizer between HMMSS and MRTD, referred to as HTD in this dissertation, in which event functions are kept to preserve the intelligibility of HMMSS while event targets are rendered by a "target selection" to make speech synthesized by HMMSS to be transformed to the original speech. Then, synthesized speech can have an "appropriate smoothness" and the detail information related to the naturalness in synthesized speech can be recovered.

The sparse representation of MRTD itself can ensure the proposed HTD has a small footprint and low computational load. The proposed HTD can also be developed to synthesize multiple voices by utilizing the voice transformation based on "target selection" with multiple small target databases. Constructing SS under limited data conditions is more practical with under-resourced languages, where only a few small public speech corpora are available. As the tonal-monosyllabic Vietnamese is an under-resourced language as presented in appendix A, Vietnamese datasets were used in this research for evaluations.

5.2 Proposed hybrid speech synthesis combined from HMM-based speech synthesis and a TD-based rendering method

5.2.1 Outline of proposed speech synthesis

There is a diagram of the proposed HTD in Fig. 5.1. The spectral and prosodic trajectories are generated from HMMSS in the first stage. Due to the proposed HTD was implemented with Vietnamese datasets, a HMMSS with GV for Vietnamese was adopted from the original work by Vu et al. [135]. The number of HMM states in one phoneme was three. A decision-tree based context clustering technique [45] was also used to ensure the synthesized trajectory is smooth and stable with limited amount of training data. Details on this HMMSS for Vietnamese can be found in [135].

Previous results [104] show that HMMSS is significantly efficient on prosodic modeling and just needs improvements to spectral modeling. Additionally, over-smoothness in HMMSS mostly occurs in spectral sequences due to the complex structures of spectral

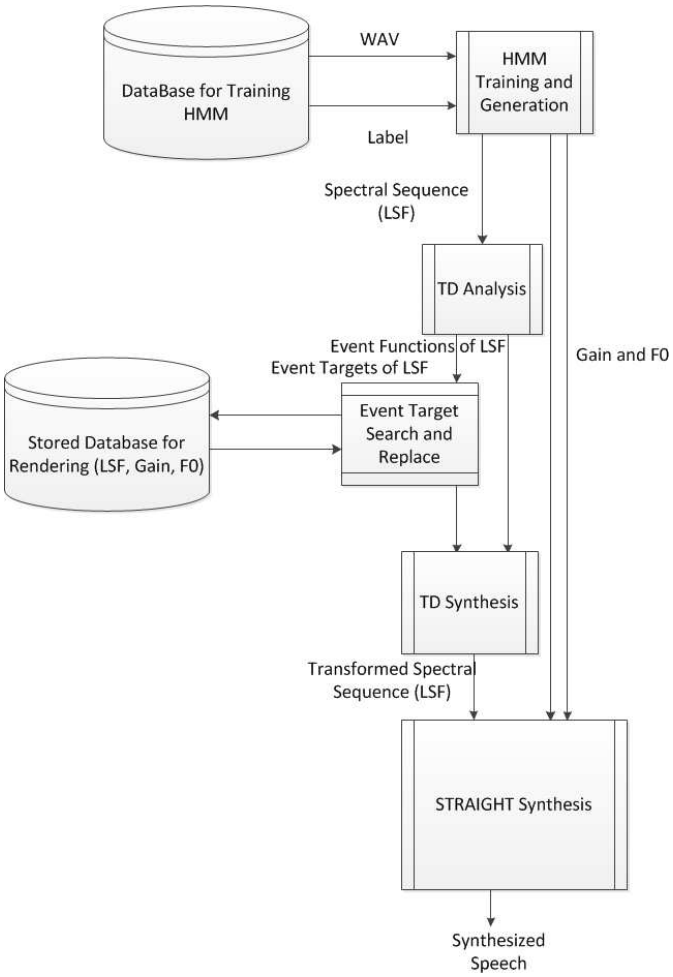

Figure 5.1: Overview of proposed HTD.

envelope features. Therefore, the prosodic trajectories of the F0 contour and gain contour of HMMSS are preserved in the proposed HTD.

A sequence of the LSF or the line spectral pairs (LSP) generated by the HMMSS is analyzed in the second stage by using MRTD [86, 87].

The event functions of the spectral sequence generated from HMMSS, which are important to perceive speech intelligibility, are preserved in the third stage of the proposed HTD, because HMMSS is already stable and highly intelligible. The target vectors are modified by selection from an original dataset to be transformed to that of the original speech. The procedure for target selection is described in more detail in the next section.

Finally, the high-quality speech vocoder STRAIGHT [84] is used to generate speech waveforms.

5.2.2 Target selection procedure inside the proposed hybrid speech synthesis

As the proposed method for selection is based on event targets, the concept behind the proposed selection procedure can be considered to be a new concept of “target selection”

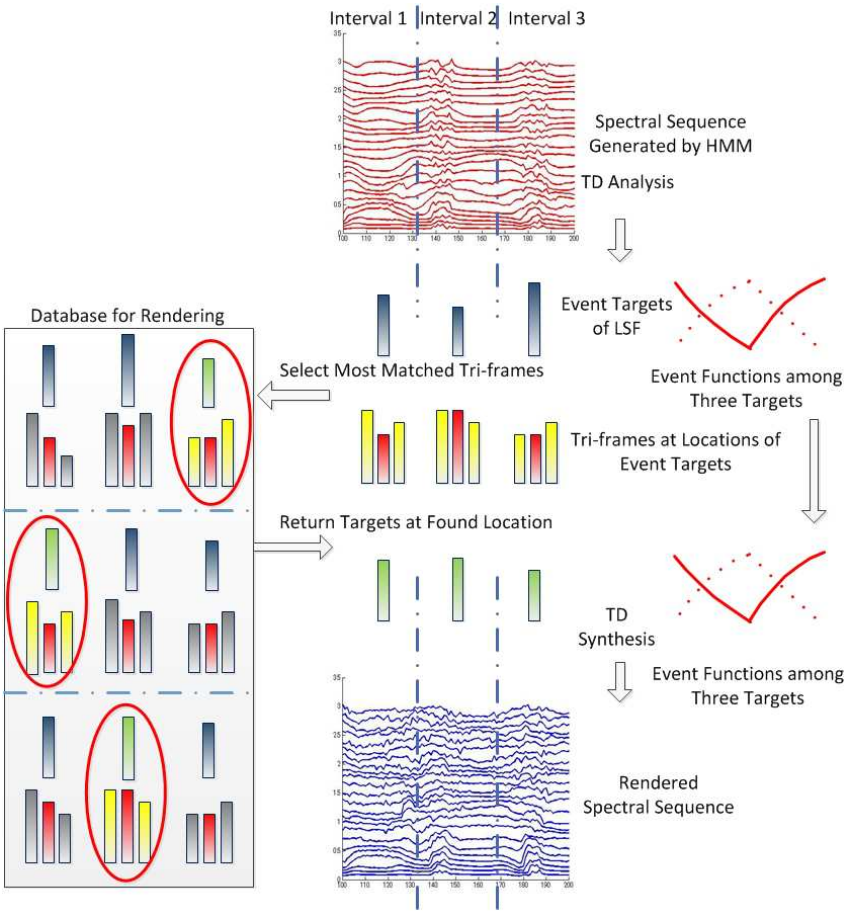

Figure 5.2: Target Selection: Single bars represent target vectors located at centers of equally-spaced portions (corresponding to HMM states); and triple bars represent three consecutive frame-based vectors (tri-frames) where their center frames are located in same positions as target vectors.

rather than the conventional concepts "unit selection" and "frame selection" [43].

The event targets of the speech trajectory generated by HMM are modified in the proposed HTD by replacing them with the most-matched event targets of original speech. Therefore, an alignment procedure in the time domain is required to accurately modify the event targets of source speech into the corresponding event targets of target speech.

Dynamic time wrapping (DTW) or nearest neighbor search (NNS) [136] can be used in the frame-based voice transformation to align the transformation in parallel form for the former and in non-parallel form for the latter. A technique of using a fixed number of equally-spaced event targets for each phoneme has been proposed when using TD-based voice transformation [18], [23]. This method involves non-parallel transformation for a syllable or an utterance but is a parallel transformation for each phoneme when each ordered event target of a source phoneme is transformed into a corresponding ordered event target of a target phoneme. Developing from this method, each phoneme is divided into three equally-spaced intervals in this work. One event target is located at the center of each of the three intervals. Therefore, there are three event targets in one phoneme. The number of event targets in one phoneme can be from one as in the original MRTD

[87], or five in [18]. There are two reasons for choosing three event targets in one phoneme in this work. The first one is that increasing the number of event targets in one phoneme larger than three does not considerably improve the quality of synthesized speech in our experiments, but increases the size of stored data for rendering. The second one is that we want to set the number of equally-spaced intervals as well as the number of event targets in one phoneme same as the number of HMM states in each phoneme with an expectation that all HMM states are rendered by the original data. Although the method of locating event targets at center frames in each HMM state in Viterbi alignment is straightforward and may increase the accuracy of the selection procedure, this method has not implemented in this research at present. This is one of our future works.

The event targets are searched and replaced as described in Fig. 5.2. Each event target of the source spectral sequence generated by HMM is replaced by an event target of original speech.

Using MRTD analysis, each event target is re-estimated by the frame-based vector at the same location, and the estimated event function at the same location as given in Eq. (2.28) of chapter 2. Therefore, event targets depend on the wide-range context, and sensitive to its locations. As a result, to directly use event targets for alignment may reduce the accuracy of the alignment procedure. Instead of that, three consecutive frames, referred to as tri-frames in this research, located at same positions of event targets, are used to align the source and target event target pairs. The matched tri-frames of the source - target pairs $s - t$ are searched by NNS with a summed cost as defined in Eqs. (5.1), (5.2), (5.3), (5.4), and (5.5) with the sub-costs of F0, LSF with order P, and PL.

$$\mathrm{d} = \mathrm{N}(\mathrm{d}_{F0}) + \mathrm{N}(\mathrm{d}_{LSF}) + \mathrm{N}(\mathrm{d}_{PL}) \tag{5.1}$$

$$\mathrm{d}_{F0} = |\log(\mathbf{F0}_t) - \log(\mathbf{F0}_s)| \tag{5.2}$$

$$\mathrm{d}_{LSF} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(\mathbf{LSF}_{i,t} - \mathbf{LSF}_{i,s})^2} \tag{5.3}$$

$$\mathrm{d}_{PL} = |\log(\mathbf{PL}_t) - \log(\mathbf{PL}_s)| \tag{5.4}$$

$$\mathrm{N}(d) = \frac{d - \mu_d}{\sigma_d} \tag{5.5}$$

The definition and concept of smoothness in speech studied in this research is physically-based rather than perception-based. Therefore, the aim of the target selection procedure is to find the spectral target of the physically closest frame in the original database for replacing to adjust the smoothness in synthesized speech physically closest to the "optimal smoothness" in the original speech, or to obtain an "appropriate smoothness" in synthesized speech.

The ideal behind the use of all F0, LSF, and PL to compute the distance cost between the trajectories generated by HMMSS and those of the original speech in HTT is to find the physically closest frames (in the waveform domain) for concatenation. This ideal was adopted to the proposed HTD to select the spectral target of the physically closest frame in the original database.

As a result, perceptual weighting the distance is not considered in this procedure. To avoid of experimental weighting the component costs, each component cost is normalized by the normal distribution similarly to that with HTT [43], as shown in Eq. (5.5), where

μ_d and σ_d correspond to the mean and standard deviation of the sample distances of all candidates.

After the matched tri-frames have been selected from the original data, the event targets of the spectral sequence generated by HMMSS in the previous stage are replaced by the outputs of the selection procedure, which are the original event targets located in the same positions as the selected tri-frames.

In our implementation, the "target selection" is supervised by labeled data to ensure its accuracy and reduce the length of searching time, in which each ordered target in a phoneme is replaced by the selected targets with the same order and in the same phoneme.

In the offline stage, the database for rendering is prepared with two steps. First, all utterances with labels are analyzed by MRTD. Then, analyzed event targets and tri-frames at the same locations are extracted from the parameters of the whole utterances by using label data, and stored for each distinct phoneme.

In the online rendering stage for each phoneme, the matched original tri-frames are selected from the original data and the event targets of the spectral sequence generated by HMM-based TTS in the previous stage are replaced by the original event targets located in the same positions as the selected original tri-frames. The "target selection" will be run with the whole database if the target phoneme for rendering is not found. Therefore, the selection procedure can still work if the number of phonemes in the database for rendering is not sufficient, such as under some limited data conditions, for instance.

5.2.3 Differences between the proposed hybrid speech synthesis and the HMM trajectory tiling synthesis

Although the proposed HTD shares some common procedures with HTT [43], their concepts are completely different. These differences are presented and discussed in this section. There are five main differences:

(1) HTT replaces all frames of the guided trajectory generated by HMMs with the most matched frames found in the original database. Therefore, HTT can be considered to be one kind of unit selection that uses HMMSS as an intermediate procedure to compute the target cost, resulting in improved stability in synthesized trajectory of speech. However, HTT shares several common disadvantages with unit selection, e.g., their requirements for huge amounts of data for selection or rendering, their huge footprints, their high computational load, and their inflexibility for voice transformations.

The proposed HTD uses HMMSS to generate spectral and prosodic trajectories. The spectral trajectory is then decomposed into its event functions and event targets. The prosodic trajectories and the event functions of the spectral trajectory are preserved to maintain the high intelligibility of HMMSS, while the sparse event targets are replaced with the most matched event targets found in the original database to reduce over-smoothness in spectral sequence. Therefore, the proposed HTD is one kind of HMMSS that uses unit selection as an intermediate procedure to limit over-smoothness, resulting in the improvement of the proposed HTD in terms of naturalness.

(2) HTT requires a huge database for rendering to ensure the smoothness of the synthesized speech because limited data may cause mismatches and discontinuities between consecutive frames. The smoothness of the synthesized trajectory in the proposed HTD is ensured by the smoothness of event functions and the stability and smoothness of the trajectory generated by HMMSS. Therefore, the matching level of the "target selection"

task does not strictly require precision as in HTT. As a result, the proposed HTD can synthesize stable and smooth speech even under limited data conditions.

(3) HTT has a high computational load because of the costly task of searching most-matched short frames (5 ms) in a huge database (with a capacity from 2 to 10 hours). "Target selection" in the proposed HTD is a sparse representation of the "frame selection" in HTT, resulting in a reduced searching space. The event rate with experimental parameters in this research is one event per eight frames. Thus, the searching space in the proposed HTD is reduced eight times compared with HTT. Additionally, the proposed "target selection" does not incur any concatenation cost as in HTT, resulting in decreased computational costs.

(4) HTT has a large footprint because it requires a huge database for "frame selection". The proposed HTD can have a small footprint because sparse "target selection" can be used with small databases for rendering. Even when the same database is used for rendering, sparse "target selection" also stores a smaller footprint compared with the "frame selection" in HTT.

(5) HTT can be combined with voice transformation by using multiple huge target databases for rendering. The requirement for huge target databases is not convenient for practical voice transformations where only a few target data are available. As was previously mentioned, "target selection" does not require a huge database. Therefore, the proposed HTD can be flexibly combined with voice transformation by using multiple small target databases for rendering.

5.3 Implementations and evaluations

5.3.1 Data preparation

A "limited data condition" was simulated with a dataset same as that in previous chapter. The size of this dataset in wav format was approximately 30MB and the duration was approximately 20 minutes. This dataset was used to train the HMMSS, which was used as input of HTT and the proposed HTD, and was used for comparisons.

Three datasets including 100, 300, and 500 utterances, extracted from DEMEN567, were used for rendering HTT and the proposed HTD to investigate the dependence of the performances of these synthesizers on the sizes of the databases used for rendering.

5.3.2 Experimental conditions and parameters

Five versions of speech were implemented for comparisons in the evaluations: speech synthesized by a HMMSS for Vietnamese [135] trained with 300 utterances, speech synthesized by HTT, speech synthesized by the proposed HTD, speech analyzed / synthesized by MRTD and STRAIGHT (MRTD-STRAIGHT) [87] and the original speech. HTT and the proposed HTD used 100, 300, and 500 utterances for rendering.

Speech analyzed / synthesized by MRTD-STRAIGHT can be considered as the ideal limitation of HTD obtained when there is sufficient data for rendering since HTD is one kind of vocoded synthesizers using both MRTD and STRAIGHT for analyzing and synthesizing. Due to reconstruction errors of MRTD and STRAIGHT, this ideal limitation of HTD is different from the original speech. The original speech can be considered as the ideal limitation of HTT when there is sufficient data for concatenation since HTT is

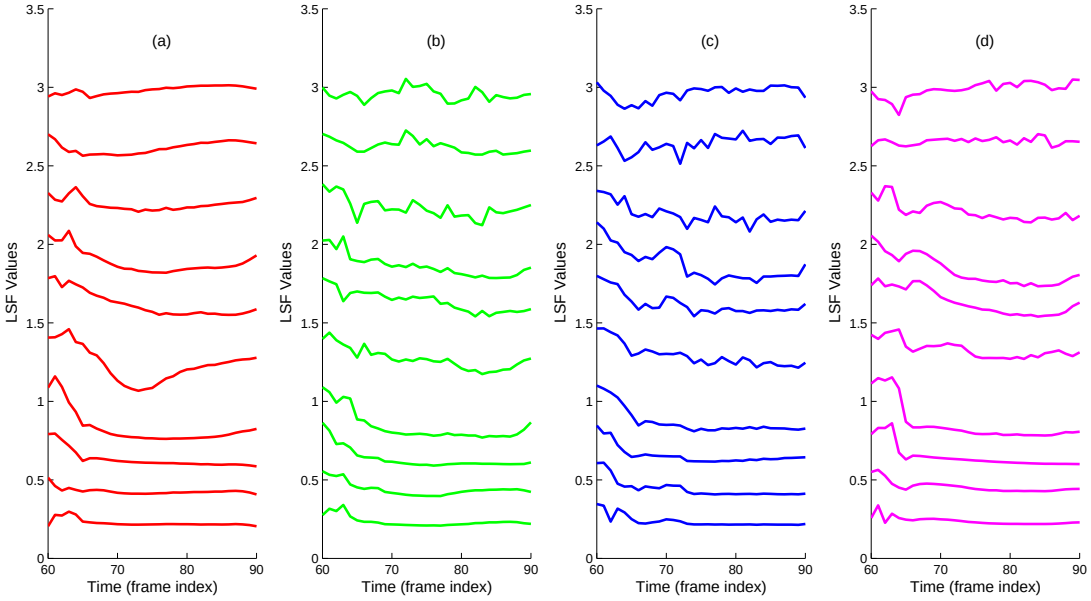

Figure 5.3: Smoothness in LSF sequences: (a) synthesized by HMMSS, (b) synthesized by HTD, (c) synthesized by HTT, (d) of the original speech.

one kind of waveform concatenation TTS used the original speech. Although these two ideal limitations of HTD and HTT can be never reached, they were used for evaluations in this paper instead of evaluating HTD and HTT with a real large-scaled speech corpus because the latter solution is not available in this research.

All experimental parameters were controlled to be equivalent for all synthesizers to enable them to be fairly evaluated: the frame length was 20 ms and the update interval was 5 ms; the spectral features for the three synthesizers were LSF with an order of 24.

The HMMSS also used the deltas of LSF. The excitation parameters for HMMSS were composed of logarithmic F0 and their corresponding delta coefficients. The context-dependent HMM used three states for one phoneme, which was the same as the number of event targets for one phoneme that was used in the proposed HTD. *The HMMSS used GV for generating parameters, in which GV was estimated from all 300 utterances in the training database.*

Other parameters of the HMMSS for Vietnamese were adopted from the original work by Vu et al. [135], while those of HTT were adopted from the original work by Qian et al.[43].

STRAIGHT version 4 [84] was used as a vocoder to generate the output waveforms. All parameters used for extracting F0, AP, and spectral envelope with STRAIGHT were default parameters except for fs, frame size and frame step.

5.3.3 Objective evaluations

100 utterances were synthesized by HMMSS, HTT and HTD in turns to evaluate the smoothness of speech synthesized by each method. The LSF sequences synthesized by the three synthesizers were used as inputs of temporal and spectral DGSM.

Table 5.1: Temporal DGSM of LSF sequences synthesized by HMMSS, HTT, and HTD

Temporal DGSM	HMM	HTT	HTD
Mean	0.5534	-0.2114	0.1871
95% confidence	0.0243	0.0222	0.0125

Table 5.2: Spectral DGSM of LSF sequences synthesized by HMMSS, HTT, and HTD

Spectral DGSM	HMM	HTT	HTD
Mean	0.2985	-0.0275	0.0731
95% confidence	0.0231	0.0127	0.0227

The results are shown in tables 5.1 and 5.2.

The results indicate that speech synthesized by HMMSS was over-smooth in both time and spectral domains, even parameter generation with GV was used. These results confirmed that HMMSS was not efficient under limited data conditions.

The results also show that speech synthesized by HTT had an "appropriate smoothness" in spectral domain. This result is reasonable because HTT was one kind of unit selection synthesizers and all selected frames were the original frames. However, speech synthesized by HTT was over-rough in time domain due to the mismatches in the frame selection under limited data conditions. This result confirmed that HTT was not efficient under limited data conditions.

The results in tables 5.1 and 5.2 show that the absolutes of DGSM scores of HTD were close to zero. Therefore, speech synthesized by HTD had an "appropriate smoothness" in both time and spectral domains. As a result, the objective evaluation results confirmed that HTD was efficient under limited data conditions with an "appropriate smoothness".

An example to show the smoothness of a syllable synthesized by HMMSS, HTT, and HTD is given in Fig. 5.3. It reveals that both the proposed synthesizer and HTT can sharpen the over-smooth LSF sequence generated by HMMSS. However, the frame-based method in HTT may have excessive sharpening and may increase the discontinuities between frames under limited data conditions, resulting in decreased intelligibility and naturalness of HTT under limited data conditions.

5.3.4 Subjective evaluations

Five versions of speech were compared in subjective evaluations, which are speech analyzed / synthesized by MRTD-STRAIGHT, speech synthesized by HMMSS, HTT, and HTD, and the original speech.

Subjective tests on intelligibility and naturalness were conducted to evaluate the synthesizers. Five subjects who were native Vietnamese with normal hearing participated in these tests. The intelligibility scores were measured by WER while MOS were used to evaluate the naturalness of the synthesizers.

20 testing sentences were chosen for evaluating the intelligibility with four restricted rules to prevent the subjects from easily predicting the meanings, and two restricted rules were chosen to ensure the evaluations were reliable. The six rules were shown in the previous chapter.

A testing dataset was used for evaluating naturalness, which contained 20 long sentences with an average length approximately 25 syllables. Evaluations of overall impres-

Table 5.3: Means of WERs (%): HMMSS was only trained with 300 utterances, speech analyzed / synthesized by MRTD-STRAIGHT and the original speech were independent with the datasets for rendering

	100 utterances	300 utterances	500 utterances
HMMSS	-	0.25	-
HTT	7.13	3.82	3.69
HTD	0.64	0.51	0.25
MRTD-STRAIGHT	-	0.25	-
Original	-	0	-

Table 5.4: F-test to show differences in the intelligibility of HMMSS and HTD in three conditions for rendering: No difference when using 500 utterances for rendering HTD

	100 Utterances	300 Utterances	500 Utterances
F	21.008	17.049	-
p	< 0.001	< 0.001	-

sions with long sentences were used to measure how close synthesized speech was to the original human speech in terms of both voice quality and segmental duration and timing.

These two testing datasets were chosen from the set of sentences that were not used for training the HMMSS and were not used for rendering with HTT and the proposed HTD. The results obtained from the intelligibility evaluations are listed in Table. 5.3. They indicate that the WERs of original speech are zeros and those of speech synthesized by HMMSS were small and were equivalent to those of speech analyzed / synthesized by MRTD-STRAIGHT. Both HTT and HTD reduced the intelligibility of speech synthesized by HMMSS. A statistical F-test was conducted to investigate how much HTT and HTD reduced the intelligibility in three conditions for rendering. The results are given in Tables. 5.4 and 5.5, which indicate that HTT significantly reduced the intelligibility of HMMSS while the reduction with HTD was not significant. With 500 utterances for rendering, the intelligibility of HTD was even equivalent with that of HMMSS and of MRTD-STRAIGHT. As a consequence, the proposed HTD was successful to preserve the intelligibility of HMMSS under limited data conditions. The results from the naturalness evaluations are presented in Fig. 5.4, in which the MOS scores of speech synthesized by HMMSS, analyzed / synthesized by MRTD-STRAIGHT, and the original speech were drawn as the same values in the three conditions, for convenience to compare all conditions for rendering. These results indicate that HTT improved the naturalness of HMMSS trained under limited data conditions with 300 sentences when using a sufficient amount of data for rendering, i.e 300 and 500 sentences. However, HTT reduced the naturalness of HMMSS when using 100 utterances for rendering. These results also indicate that HTD improved the naturalness of HMMSS trained under limited data conditions with 300 sentences when using all three datasets for rendering.

A statistical F-test was conducted to measure the significance of the improvements and reductions on naturalness of HTT and HTD compared with the HMMSS trained with 300 sentences. The results are shown in Tables. 5.6 and 5.7. They indicate that HTT significantly reduced the naturalness of the HMMSS when using 100 utterances for rendering and insignificantly improved the naturalness of the HMMSS when using 300 and 500 utterances for rendering. They also indicate that HTD significantly improved

Table 5.5: F-test to show differences in the intelligibility of HMMSS and HTT in three conditions for rendering

	100 Utterances	300 Utterances	500 Utterances
F	523.213	151.545	234.607
p	< 0.001	< 0.001	< 0.001

Table 5.6: F-test to show differences in the naturalness of HMMSS and HTT in three conditions for rendering

	100 Utterances	300 Utterances	500 Utterances
F	18.751	4.918	6.424
p	< 0.001	0.028	0.012

the naturalness of HMMSS in all three datasets used for rendering.

Consequently, the proposed HTD demonstrated its efficiency compared with HMMSS and HTT in terms of naturalness in all conditions for rendering. Especially, the propsoed HTD could improve the naturalness of HMMSS even when using an ultra-small dataset for rendering, i.e. a dataset of 100 utterances.

5.3.5 Discussions

Futher discussions of intelligibility and naturalness

The results with speech analyzed / synthesized by MRTD-STRAIGHT (ideal limitation of HTD) and the original speech (ideal limitation of HTT) show that HTT can be superior to HTD under high-resourced conditions.

The results from the intelligibility evaluation were consistent with the results from HMMSS [135] where the intelligibility scores of a Vietnamese synthesizer could reach 100%. The intelligibility of the mono-syllabic Vietnamese speech seems to be higher than that of other languages.

MOS score of 3.8 for the original speech was quite low since corpus DEMEN567 was not well recorded due to a low sampling frequency of 11025 Hz and the recording environment. The MOS scores for all HMMSS, HTT and the proposed HTD were not high compared with those of the original speech since they were implemented under a "limited data condition".

Consequently, results from evaluations demonstrated that the proposed HTD with a new rendering method preserved the intelligibility and improved the naturalness of HMMSS by reducing the problem with over-smoothness. The outperformance of the proposed HTD compared with the HTT confirmed that the proposed HTD is specifically efficient under limited data conditions.

Discussion of footprint size

The footprint size of the proposed HTD is the sum of the size of the trained HMM model's parameters and the size of the stored data for rendering. The size of the trained HMM model's parameters is up to about $1 - 2MB$ if an appropriate fixed-point representation is used.

Table 5.7: F-test to show differences in the naturalness of HMMSS and HTD in three conditions for rendering

	100 Utterances	300 Utterances	500 Utterances
F	31.749	45.213	84.644
p	< 0.001	< 0.001	< 0.001

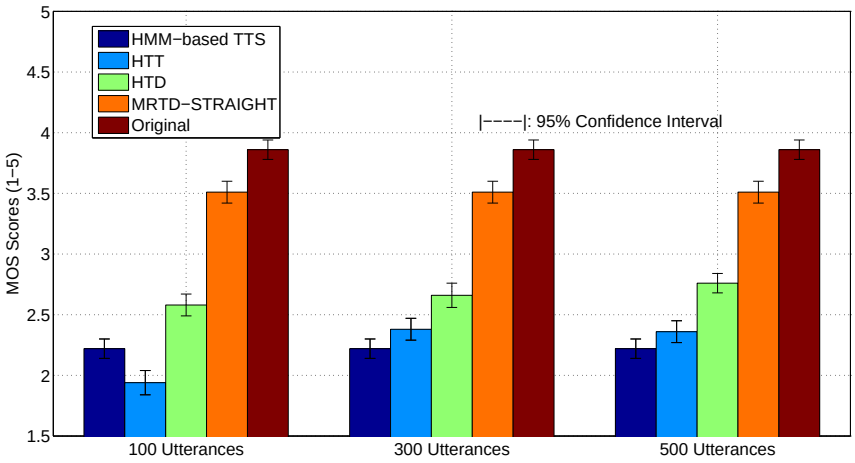

Figure 5.4: Mean MOSs for naturalness evaluations and their 95% confidence intervals: HMMSS was only trained with 300 utterances, speech analyzed / synthesized by MRTD-STRAIGHT and the original speech ware independent with the datasets for rendering

The experimental frame period was $5ms$ and $fs = 11.025$ KHz. If each sample is represented by N bytes with fixed-point representation, the size of each original frame period is $5 \times 11.025 \times N \approx 55 \times N$ bytes. The experimental event rate was approximately one target in eight frames on average. Three parameters, 24-ordered LSF, F0, and PL, were used. Each event target of each of the three parameters was stored together with their corresponding tri-frames. If each value is represented by N bytes with fixed-point representation, the size of each encoded frame is $26 \times (3+1)/8 \times N = 13 \times N$ bytes. Thus, the compression rate $rt \approx 55/13 \approx 4$. The sizes of the three original waveform databases of 100, 300, and 500 utterances for rendering were approximately 10 MBs, 30 MB, and 50 MBs, respectively. Therefore, the actual size of the smallest database for rendering was approximately $10/4 \approx 2.5$ MB, and the footprint of the proposed HTD was approximately $4-5$ MB. Although the proposed HTD increases the size of footprint compared with that of HMMSS, it was still small enough for most limited-resourced hardware platforms. If further compression techniques are used such as vector quantization to quantize LSF, F0, and PL, the size of the compressed footprint can be reduced even further.

Discussion of the possibility of the proposed HTD for voice transformations

Qian et al.s hybrid synthesizer [43] was combined with voice transformation by using multiple huge target databases for rendering. Their experimental results indicated that the quality of the transformed speech was high. However, the requirement for a huge target database is not convenient for practical voice transformations where few target data are

available. Consequently, Qian et al.s hybrid synthesizer is still not flexible enough for voice transformations.

TD-based voice transformations [18], [23] could efficiently transform speaker individuality by preserving the event functions of source speech and transforming its event targets to those of target speech. This manner is similar to the proposed HTD, when event functions of speech synthesized by HMMSS are preserved and its event targets are selected from an original database. Therefore, it is possible to develop the proposed HTD to synthesize multiple voices with a multiple-voices database.

The experimental results in this chapter revealed that the proposed HTD was efficient with a small database. Therefore, the proposed HTD can be developed for voice transformations with limited target data. Although the proposed HTD was just evaluated with a single-voice database, it will be implemented with multiple-speakers and multiple-styles databases in the future to confirm its flexibility for voice transformations.

5.4 Conclusions

A hybrid method between HMMSS and MRTD, in which speech synthesized by HMMSS was transformed to the original speech by using MRTD, was proposed to solve the over-smoothness problem with HMMSS under limited data conditions. The experimental results revealed that the proposed HTD could synthesized speech with an "appropriate smoothness" and articulated efficiently under limited data conditions in terms of both speech intelligibility and naturalness. Additionally, the proposed synthesizer had a small footprint, had small computational load. As a consequence, the third objective of this dissertation was solved. The proposed HTD will be developed in the future to other languages to confirm whether the new approach is language-independent and is flexible for voice transformation.

Chapter 6

Tone transformation for SS of tonal languages

The unified purpose of this research is to propose methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech for improving the naturalness of synthesized speech under limited data conditions. Under this unified purpose, four objectives are specified as shown in chapter Introduction. The aim of this chapter is to solve the fourth issue related to the fourth objective of the dissertation, which is to deal with an ultra-limited data condition in tonal languages, when the number of tonal phonetic units for synthesizing is not sufficient, by using transformed tonal speech units with an "appropriate smoothness" when the original ones are missing. First, the question why tone transformation can be used to solve the problem is explained. Second, the proposed MRTD-GMM method for tone transformation is described. Third, the proposed method of non-parallel alignment applied for the proposed tone transformation is presented. Finally, the experimental results are presented and discussed.

6.1 Introduction

All possible units of a specific phonetic unit set are required to build a speech synthesizer. They are needed for concatenation in CSS, or for training in HMMSS, or for extracting unit-based parameters for other unit-based synthesizers. This requirement may be easy for verbal languages, in which the number of all units of a specific phonetic unit set is limited. However, the numbers of all tonal units significantly increases in tonal languages, and it is difficult to design a small corpus that covers all possible tonal phonetic units.

Moreover, all context-dependent phonetic units are required to build a natural corpus-based SS, which leads to the need for a large database with a size of up to dozens of gigabytes for concatenation or training.

The motivation for this chapter is to improve the usability of SS under an ultra-limited data condition in which the number of tonal phonetic units is not sufficient. Therefore, a method of tone transformation was proposed to reduce the number of tonal units required for the SS of tonal languages. Lexical tones are usually represented by F0 contours. Therefore, tone transformation can be considered to be F0 contour transformation that is applied to converting lexical tones.

The simplest F0 contour transformation that can be applied to converting lexical tones is the simple exchange of the F0 contour of a source tonal unit with that of a target unit.

This method has been efficiently applied to the Thai language and reduces the size of the footprint by approximately five times [67]. However, this method requires a huge amount of data covering all tonal units to compute all their F0 contours in the offline stage. Therefore, this method is still not usable under limited data conditions.

The conventional GMM-based speech transformations have many advantages [14, 100, 101, 102, 103]. Therefore, the state-of-the-art F0 transformation is based on the GMM [68]. This method transforms both frame-based F0 values and their temporal deltas, which can be efficient for expressive speech since F0 contours are largely varied on the emotions of speech in short-term intervals. However, F0 contours of a source tonal unit and a target tonal unit are usually distinct in their long-term approximations rather than their short-term details [69, 70, 71]. Therefore, transforming the short-term frame-based F0 values and their deltas may not efficiently transform the F0 contours of lexical tones but may increase the noise sensitivity. Additionally, GMM-based voice transformations still suffer from several drawbacks, including insufficiently precise GMM models and parameters and the temporal over-roughness between frames [18].

A framework for transforming spectral sequences combined from GMM and MRTD [86, 87, 92], named MRTD-GMM [17, 18], was proposed to overcome these common drawbacks of conventional GMM-based speech transformation with significant improvements. Background on TD and MRTD was already presented in section 2.3 of chapter 2 and background on MRTD-GMM was already presented in section 2.4 of chapter 2. The results on transformation of spectral sequences obtained by B. Nguyen and Akagi [17, 18] demonstrated that converting only static event targets and preserving dynamic event functions with MRTD could efficiently improve the estimates of GMM parameters as well as efficiently eliminate the frame-to-frame discontinuities (the temporal over-roughness) compared with conventional GMM speech transformations, resulting in natural and smooth transformed speech. Due to the sparse representation of MRTD, MRTD-GMM can be more appropriate to transform the long-term features such as the tones, compared with frame-based transformations. However, MRTD-GMM still suffers from two main drawbacks when being applied to transforming prosodic features such as the F0 contour. Because dynamic features of F0 are important, both static and dynamic features of F0 need to be transformed. Normally, transforming the dynamic features with TD requires the transformation of dynamic event functions, which is not usable in original MRTD-GMM. In addition, the phoneme-based target alignment and training inside MRTD-GMM require large database covering all phonemes to train all phoneme-based GMMs. These two drawbacks were solved with the proposed MRTD-GMM-based F0 contour transformation applied to transform the lexical tones in this chapter. As Vietnamese is a tonal language as presented in appendix A, Vietnamese datasets were used in this research for evaluations.

6.2 Using tone transformations in speech synthesis of tonal languages

Changing the tone for each pronunciation in tonal languages provides a set of tonal units, referred to as a same-phonation set in this dissertation. Tone transformation can be applied to the SS of tonal languages by combining the transformed F0 contours of tonal units with the spectral envelope of a representative unit in a same-phonation set to

produce synthetic tonal units with a source/filter vocoder. A neutral unit with neutral tone, which is a tone with a flat F0 contour that is usually found in tonal languages [69], can be used as the easiest representative unit of a same-phonation set. An example of a same-phonation set for monophone /a/ in Vietnamese is *a, a‘, a′, a?, a., ã*. The spectral envelope features for all units in a same-phonation set are almost the same because they are related to similar vocal tract parameters produced by similar pronunciation behaviors. Therefore, tone transformation can be applied to the CSS of tonal languages by combining the transformed F0 contours of tonal units such as *a‘, a′, a?, a., ã* with the original spectral envelope of a representative unit in a same-phonation set such as *a* to produce synthetic sounds of these tonal units.

Assume that all tonal units are converted instead of the original ones being used and denote the theoretical percentage of data reduction as $\mathbf{r_f}$. Then, $\mathbf{r_f}$ can be approximately computed as given in Eq. (6.1),

$$\mathbf{r_f} = (1 - N_n/N_t) \times 100\% \tag{6.1}$$

where N_n is the number of neutral units and N_t is the number of tonal units.

There are a total of approximately 7000 meaningful tonal syllables and 1200 neutral syllables in Vietnamese [137, 138]. Thus, $r_t \approx 83\%$ with Vietnamese SS if the tones are transformed for all tonal syllables.

Tone transformation can be applied both for concatenating CSS and for training HMMSS. When being applied for CSS, a tonal unit is synthesized from spectral parameters of a corresponding neutral unit and the F0 contour of the transformed tone before using for concatenation. When being applied for HMMSS, a tonal unit is trained by using spectral parameters of a corresponding neutral unit and the F0 contour of the transformed tone. Therefore, tone transformation can reduce a significant amount of tonal units required for concatenating the CSS of tonal languages and for training the HMMSS of tonal languages. As a result, tone transformation can improve the usability of SS under limited data conditions.

6.3 Proposed MRTD-GMM method for tone transformation

MRTD-GMM may not be efficient for F0 transformation due to the lack of converting dynamic features. There are two options of transforming dynamic features with TD, one is transforming the dynamic event functions and the other is transforming the deltas of static event targets. As the dynamic event functions presents the relations between sparse event targets and static frames, transforming them means transforming dynamic features in all frames. This is sophisticated and may not suitable to transform the lexical tones because F0 contours of a source neutral unit and a target tonal unit are usually distinct in their approximations rather than in their details [69, 70, 71]. On the contrary, transforming the deltas of event targets is easy, suitable for statistical training, and also suitable to transform the lexical tones because only the dynamics between sparse event targets are transformed.

In addition, it has been found that low-dimensional vectors are not suitable for modeling with GMM because they might cause GMM over-fitting [139]. Therefore, using the

delta features of F0 to extend the dimensions of F0 vectors can improve the accuracy with which GMM parameters are estimated [68].

Assume that there are M F0 targets for the aligned source and target speech, where $\{\mathbf{f0}^{x_i}$ and $\mathbf{f0}^{y_i^t}\}$ correspond to the static F0 targets for the source F0 contour x and target F0 contour y^t with tone t^{th}. Here, $i = 1, 2, \cdots, M$, and $t = 2, 3, \cdots, \Im$. The $\Im$ is the number of tones and $\Im = 6$ in Vietnamese. The two-dimensional (2-D) source and target F0 target vectors $\mathbf{f0}^X$ and $\mathbf{f0}^{Y^t}$ are represented as given in Eqs. (6.2) , (6.3), and (6.4).

$$\mathbf{f0}^X = [\mathbf{f0}^{X_1{}^T}, ..., \mathbf{f0}^{X_i{}^T}, ..., \mathbf{f0}^{X_M{}^T}], \tag{6.2}$$

$$\mathbf{f0}^{Y^t} = [\mathbf{f0}^{Y_1^t{}^T}, ..., \mathbf{f0}^{Y_i^t{}^T}, ..., \mathbf{f0}^{Y_M^t{}^T}] \tag{6.3}$$

where

$$\mathbf{f0}^{X_i} = [\mathbf{f0}^{x_i}, \Delta\mathbf{f0}^{x_i}]^T, \qquad \mathbf{f0}^{Y_i^t} = [\mathbf{f0}^{y_i^t}, \Delta\mathbf{f0}^{y_i^t}]^T \tag{6.4}$$

The joint source-target vector of F0 targets $\mathbf{z}$ is computed as in Eq. (6.5).

$$\mathbf{z} = [\mathbf{f0}^{X^T}, \mathbf{f0}^{Y^t{}^T}]^T \tag{6.5}$$

The distribution of $\mathbf{z}$ is modeled by GMM λ, caculated as presented in Eq. (6.6).

$$\mathrm{p}(\mathbf{z}|\lambda) = \sum_{q=1}^{Q} \alpha_q \mathrm{N}(z; \mu_q, \Sigma_q), \tag{6.6}$$

where Q is the number of Gaussian components, $\mathrm{N}(\mathbf{z}; \mu_q, \mathbf{\Sigma}_q)$ denotes the distribution with mean μ_q and covariance matrix $\mathbf{\Sigma}_q$, and α_q is the prior probability of $\mathbf{z}$ generated by component q. The parameters $(\alpha_q, \mu_q, \mathbf{\Sigma}_q)$ are estimated using EM algorithm and the transformed F0 contour $\hat{\mathbf{y}}^t$ with target tone t^{th} is determined by maximizing the likelihood following Toda et.al. [16].

6.4 Proposed NNS-based alignment method for tone transformation

The parallel phoneme-based target alignment and training within MRTD-GMM for spectral sequence [18] is difficult to accomplish with limited amounts of training data. The non-parallel method of alignment using NNS [68] can be used with limited amounts of training data. However, Wu et al.'s method of alignment [68] searches the closest neighbors in the whole data space, which may reduce the accuracy of alignment.

Wu et al.'s NNS-based alignment [68] was modified in this research, and was integrated with the modified MRTD-GMM for F0 transformation by clustering available phonetic units based on their articulatory similarities. Each cluster produces a phonetic-dependent subspace for searching in the modified NNS-based alignment. Thus, the source and target units for each aligned source-target pair are selected from corresponding subspaces to which the source/target units belong.

When the F0 contour of lexical tones is transformed, the spectral envelope parameters for all units in each same-phonation set are almost the same because they are related to similar vocal tract parameters produced by similar pronunciation behaviors. Thus, line

spectral frequency (LSF) is used for the alignment instead of directly using F0. Then, the F0 targets in the positions of the aligned LSF target pairs are used as the inputs of the phonetic-dependent GMM models for training.

Assume that the source LSF target vector computed from neutral units is $\{\mathbf{lsf_m}\}$ and $m = 1, 2, \cdots, M$, where M is the number of event targets of these neutral units. When training for target tone t^{th}, and $t = 2, 3, \cdots, \Im$, the set of all tonal units with tone t^{th} is $\hat{ws}^t$. The $\hat{ss}^{t,m}$ is a tonal subspace of $\hat{ws}^t$ containing all units belonging to the phonetic unit cluster that $\mathbf{lsf_m}$ belongs to. The target vector for alignment is computed as given in Eq. (6.7).

$$\tilde{\mathbf{lsf}}_m = \text{NNS}(\mathbf{lsf}_m, \hat{ss}^{t,m}), \hat{ss}^{t,m} \in \hat{ws}_t. \tag{6.7}$$

The NNS function here returns the closest neighbors found in target space. The aligned LSF target pairs are therefore $\{\mathbf{lsf}_m$, and $\tilde{\mathbf{lsf}}_m\}$. The positions of the aligned LSF target pairs are needed for F0 transformation rather than their values. The positions of aligned pairs are $\{m, \text{p}(\tilde{\mathbf{lsf}}_m)\}$ in this case, where $\text{p}(\tilde{\mathbf{lsf}}_m)$ is the position of $\tilde{\mathbf{lsf}}_m$.

Target-source alignment is also used. If we assume that the target LSF target vector computed from tonal units with tone t^{th}, is $\{\tilde{\mathbf{lsf}}^t_n\}$, the source vector for alignment is computed as presented in Eq. (6.8).

$$\mathbf{lsf}^t_n = \text{NNS}(\tilde{\mathbf{lsf}}^t_n, \hat{ss}^{1,n}) \tag{6.8}$$

where $\hat{ss}^{1,n} \in \hat{ws}^1, n = 1, 2, \cdots, N$, N is the number of event targets of these tonal units, $\hat{ws}^1$ is the set of all neutral units, and $\hat{ss}^{1,n}$ is a neutral subspace of $\hat{ws}^1$ containing all neutral units belonging to the phonetic unit cluster that $\mathbf{lsf}^t_n$ belongs to. The position of aligned pairs is $\{\text{p}(\mathbf{lsf}^t_n), n\}$ where $\text{p}(\mathbf{lsf}^t_n)$ is the position of $\mathbf{lsf}^t_n$.

Combining both source-target and target-source alignments, GMM transformation function F is trained from the aligned pairs of F0 vectors: $\{\mathbf{f0}^X(m), \mathbf{f0}^{Y^t}(\text{p}(\tilde{\mathbf{lsf}}_m))\}$ and $\{\mathbf{f0}^X(p(lsf^t_n)), \mathbf{f0}^{Y^t}(n)\}$. Here, $\mathbf{f0}^X$ and $\mathbf{f0}^{Y^t}$ correspond to the F0 target vectors combined from static F0 targets and their deltas of source neutral units and target tonal units with tone t^{th} which are the same as those in Eqs. (6.2) and (6.3).

6.5 Implementations and evaluations

6.5.1 Data preparation

The original DEMEN567 corpus [72] was extracted into a syllable-based dataset of 1000 tonal syllables, which covered all six Vietnamese tones, to implement the tone transformations. A group of neutral units, which is the tonal units with Vietnamese tone Level, was used as the source while five other tonal unit groups were used as targets for the F0 contour transformations.

Ten tonal syllables of single-syllable words were evaluated for each tone. Thus, a total of 50 syllables were used for both objective and subjective evaluations.

6.5.2 Experimental parameters for implementations

The frame sizes were set to 20 ms and the update intervals were set to 1 ms for two transformations for F0 contours of lexical tones, which were the proposed method and the GMM-based method of Wu [68] used as the baseline for comparisons.

Table 6.1: Temporal DGSM in log(F0) of tonal syllables

	Frame-based GMM	MRTD-GMM
Mean	-3.18	-0.51
95% Confidence	0.02	0.02

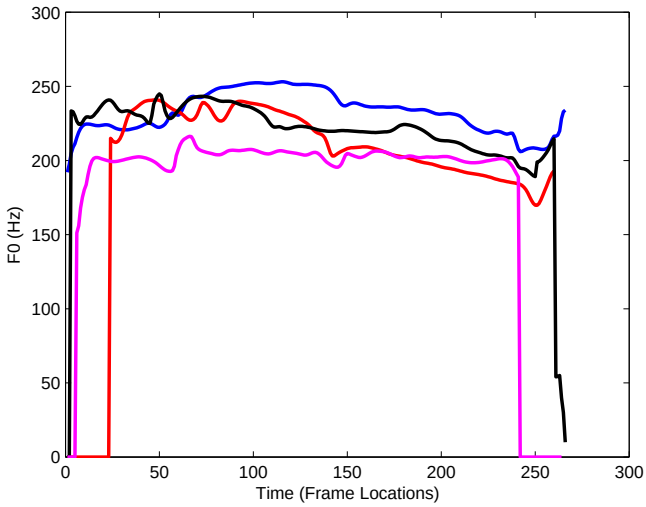

Figure 6.1: F0 contours of tonal syllables: the blue curve is of a source tonal syllable; the red curve is of the target one; the magenta curve is transformed by frame-based GMM; and the black curve is transformed by the proposed method.

When using TD analysis/synthesis, each phoneme was represented by five equally spaced F0 event targets.

For the alignments, the orders of LSF P were 32. The whole searching space was clustered into sub-spaces, in which each sub-space corresponded to each tonal phoneme. On the other words, the alignments were performed for each tonal phoneme, supervised by the text label.

For training, the numbers of GMM mixtures Q were 4.

STRAIGHT version 4 [84] was used to synthesize the transformed tonal syllables with both methods.

6.5.3 Objective evaluations

Objective tests to measure the smoothness of F0 contours of tonal syllables transformed by the proposed method, the baseline method, the the original ones were conducted, using DGSM of F0 contour presented in chapter 2.

When using DGSM to evaluate the smoothness of F0 contours, only temporal smoothness is observed and measured.

The results of DGSM of $log(F0)$, computed for all 50 testing Vietnamese tonal syllables, are shown in table. 6.1, which indicate that the tonal syllables transformed by the proposed method were smoother than those by the baseline method. Figure. 6.1 also shows that the F0 contour converted by the proposed method is closer with the target than the one converted by the frame-based GMM transformation.

Table 6.2: F-test to show the difference between WERs of tonal syllables transformed by frame-based GMM and WERs of the original tonal syllables

	tone 1	tone 2	tone 3	tone 4	tone 5
F	435.856	435.856	262.71	242.12	563.099
p	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 6.3: F-test to show the difference between WERs of tonal syllables transformed by MRTD-GMM and WERs of the original tonal syllables

	tone 1	tone 2	tone 3	tone 4	tone 5
F	5.003	35.991	43.253	70.285	72.036
p	0.038	< 0.001	< 0.001	< 0.001	< 0.001

6.5.4 Subjective evaluations

Subjective tests on intelligibility and naturalness were also conducted with five subjects who were native Vietnamese speakers with normal hearing. The intelligibility scores were measured by using word error rates (WER) while MOS was used to evaluate the naturalness of tone transformations.

The results on intelligibility in Fig. 6.2 and Tables. 6.2 and 6.3 indicate that the proposed MRTD-GMM F0 transformation significantly outperformed the state-of-the-art F0 transformation [68] in terms of speech intelligibility. The WERs of syllables transformed by the proposed MRTD-GMM F0 transformation were closer to those of the original speech compared with WERs of syllables transformed by the frame-based GMM transformation. The results on intelligibility also reveal that the proposed tone transformation was efficient for three Vietnamese tones rising, broken, and falling, while its performance was reduced with two tones curve and drop. The reason that why the proposed method was not useful for some tones may be the lack of a method of transforming the power. It has been known that although F0 contours can represent Vietnamese tones, the power contours also significantly affect to some Vietnamese tones.

The results on naturalness in Fig. 6.3 and Table. 6.4 indicate that the proposed MRTD-GMM F0 transformation also outperformed the state-of-the-art F0 transformation, frame-based GMM [68] in terms of naturalness. The MOS scores of syllables transformed by the proposed MRTD-GMM F0 transformation were closer to those of the original speech compared with MOS scores of syllables transformed by the frame-based GMM transformation.

Consequently, a significant number of tonal units for the CSS of tonal languages could be reduced by using tonal units transformed by the proposed tone transformation instead of using the original ones. However, a method of transforming the power contour needs to be developed for some tones.

Table 6.4: F-test to show the differences between MOSs of tonal syllables transformed by frame-based GMM, MRTD-GMM and MOSs of the original tonal syllables

	Frame-based GMM	MRTD-GMM
F	748.301	359.402
p	< 0.001	< 0.001

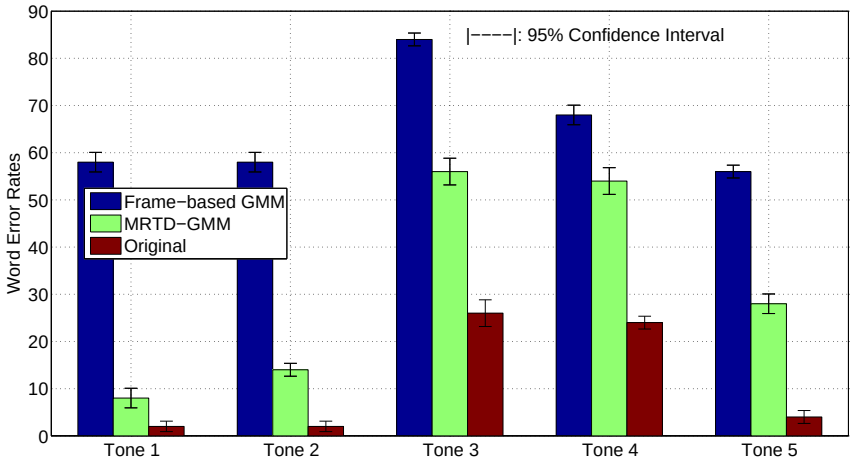

Figure 6.2: WER scores for tone transformations calculated for each of six Vietnamese tones

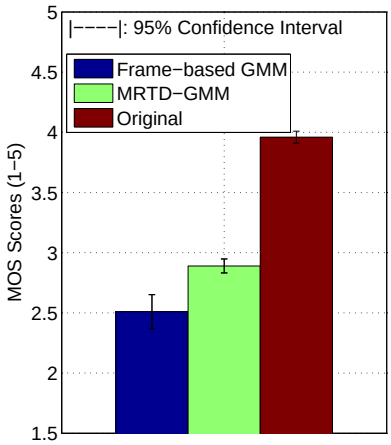

Figure 6.3: MOS scores for tone transformations calculated for all tones.

6.6 Conclusions

A method of tone transformation for reducing the number of tonal units in the SS of tonal languages was proposed. The experimental results revealed that a significant number of tonal units for the SS of tonal languages could be reduced by using transformed tonal units instead of using the original ones, while the transformed tonal units are natural, high-intelligible, with an "appropriate smoothness". As a consequence, the proposed lexical tone transformation was efficient for the SS of tonal languages and the fourth objective of this dissertation, aforementioned in chapter introduction, was solved.

Chapter 7

Summary and Future Work

In this chapter, the dissertation is concluded with summaries and discussions of future works.

7.1 Summary of the dissertation

SS can be used in several applications, covering all kinds of human-machine interaction systems. Among many kinds of SS that have been proposed, the state-of-the-art ones are CSS and HMMSS, which were studied in this research.

The need of covering all possible units leads to a requirement of significant amount of data to build a speech synthesizer. Due to the efforts of co-articulation on SS, not only all context-independent phonetic units but also all context-dependent phonetic units are necessary to synthesize natural speech. As a result, state-of-the-art speech synthesizers require large-scaled speech corpora to synthesize natural speech. For instance, high-quality HMMSS requires gigabytes of data for training and CSS requires dozens of gigabytes of data for concatenation. Unfortunately, building a large-scaled speech corpus is a costly task that takes a long time and requires a great deal of effort by engineers, acousticians, and linguists. Therefore, to build high-quality SS with limited data is an important and practical issue, specifically for under-resourced languages with which only a few of small speech corpora are usable.

Additionally, to synthesize speech with multiple individual and emotional voices requires speech corpora with multiple voices. Synthesizing speech with a limited amount of data is also critical in movies and games applications when we need to synthesize the voices of famous people who are not alive or have lost their voice characteristics. Although we can use a large-scaled corpus to synthesize speech with a standard voice first and then using methods of voice transformation / adaption to transform / adapt the standard voice to the target voices with limited target data, this kind of approaches still requires a large-scaled corpus that is not always available. As a result, to build high-quality SS under limited data conditions is still a challenge.

Under limited data conditions, one of the biggest challenges is to ensure an "appropriate smoothness" in synthesized speech. For instance, speech synthesized by CSS under limited data conditions is usually over-rough in time domain, and speech synthesized by HMMSS under limited data conditions is usually over-smooth in both time and spectral domains.

Based on these considerations, ***the motivation of this dissertation was to propose methods for improving the naturalness of synthesized speech under limited data conditions***. To achieve this motivation, this dissertation falls into the approach of using methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech. Therefore, ***the unified purpose of this research is to propose methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech for improving the naturalness of synthesized speech under limited data conditions***. Under this unified purpose, four objectives are specified:

1. The first objective is to propose a speech smoothness measure that can be applied to evaluate the smoothness in synthesized speech and to control synthesized speech with an "appropriate smoothness".

2. The second objective is to solve the main problem of CSS under limited data conditions, which is reducing mismatch-context errors in CSS to ensure an "appropriate smoothness" in synthesized speech.

3. The third objective is to solve the main problem of HMMSS under limited data conditions, which is reducing over-smoothness in HMMSS to ensure an "appropriate smoothness" in synthesized speech.

4. The fourth objective is to deal with an ultra-limited data condition in tonal languages, when the number of tonal phonetic units for synthesizing is not sufficient, by using transformed tonal speech units with an "appropriate smoothness" when the original ones are missing.

The four issues corresponding to the four objectives of the dissertation are presented in chapters 3, 4, 5, and 6.

Both over-smoothness and over-roughness can reduce the naturalness of synthesized speech. Therefore, instead of synthesizing too smooth or too rough speech, an "appropriate smoothness", approximated the "optimal smoothness" that naturally exists in original speech, has to be reached to ensure the naturalness in synthesized speech. However, to objectively classify speech smoothness into over-smoothness, or over-roughness, or appropriate smoothness has been difficult due to the lack of an efficient smoothness measure in speech.

In speech synthesis, statistical variances have been widely used to measure the smoothness in synthesized speech. However, statistical variances of static spectral features show several disadvantages as shown in chapter 3. Therefore, to propose a new and efficient speech smoothness measure is important. It was the first issue of this research, presented in chapter 3.

Mismatch-context errors caused by co-articulation of speech occur frequently under limited data conditions. Mismatch-context errors make speech synthesized by CSS over-rough in time domain that can reduce the naturalness in synthesized speech. Therefore, the performance of CSS is drastically reduced when the size of the speech corpus is reduced. In the Blizzard Challenge 2006 which provided a large-scaled speech corpus with amount of 5 h of speech, the best system was based on CSS. However, in Blizzard Challenge 2005 which provided a smaller speech corpus with amount of 1.5 h of speech, a HMMSS outperformed the well-established CSS systems in both speech naturalness and intelligibility. As a result, reducing mismatch-context errors in CSS under limited data by using methods of speech modification to modify synthesized speech to obtain an "appropriate smoothness" is critical. It was the second issue of this research, presented

in chapter 4.

Speech synthesized by HMMSS is over-smooth. When synthesized speech is over-smooth, it sounds "muffled" and far from natural. "Over-smoothness" also causes a reduction in identification emotions / expressions / styles in speech that can also affect to the perception of the naturalness. Therefore, over-smoothness is the main remaining factor reducing the naturalness of HMMSS. Over-smoothness is mainly affected by the accuracy of model estimates and that of the training algorithm. These factors are affected by the amount of training data. The larger the amount of training data, the more accurate the model estimates and the training algorithm, and the lesser the over-smoothness in synthesized speech. As a result, the effect of over-smoothness becomes more serious in a situation with limited training data and it is difficult to ensure the naturalness of HMMSS under limited data conditions. Therefore, the third issue of this research is to reduce the over-smoothness in HMMSS by transform speech synthesized by HMMSS into the original speech, in order to adjust synthesized speech to obtain an "appropriate smoothness". This issue was presented in chapter 5.

In ultra-limited data conditions, specifically in tonal languages, the number of phonetic units for synthesizing can be not sufficient. This problem prevents to synthesize speech with any input text content. Since speech modification or transformation can convert a speech unit into other units, it can reduce the number of speech units required for synthesizing. Therefore, by using tone transformations in SS of tonal languages, the transformed tonal speech units with an "appropriate smoothness" can be used instead of the missing original speech. It was the fourth issue of this research and were presented in chapter 6.

For the first issue presented in chapter 3, a speech smoothness measure, named DGSM, based on the square sum of the variance of the delta-delta sequence in both time and spectral domains was proposed. The proposed DGSM was shown to be reliable and efficient to measure the smoothness in different kinds of speech in both time and spectral domains. Therefore, the proposed DGSM was later used to evaluate the smoothness in the speech synthesized / transformed / modified by the methods given in chapters 4, 5, and 6.

For the second issue presented in chapter 4, methods for reducing mismatch-context errors in CSS based on MRTD, including an method of speech modification, were proposed. The experimental results with Vietnamese datasets revealed that the proposed speech modification method outperformed the conventional method in terms of producing an "appropriate smoothness" in synthesized speech, the proposed CSS also convincingly outperformed the conventional CSS in terms of both speech intelligibility and naturalness. As a consequence, the proposed methods efficiently reduced the mismatch-context errors in CSS, and the proposed CSS ran efficiently under limited data conditions.

For the third issue presented in chapter 5, a hybrid SS between HMMSS and MRTD, named HTD, that uses MRTD to transform speech synthesized by HMMSS into the original speech was proposed to solve the over-smoothness problem of HMMSS under limited data conditions. The experimental results with Vietnamese datasets revealed that the proposed HTD could synthesize speech with an "appropriate smoothness", and it articulated efficiently under limited data conditions in terms of both speech intelligibility and naturalness. Additionally, the proposed synthesizer had a small footprint, had small computational load, and could be flexible for voice transformation.

For the fourth issue presented in chapter 6, a method of tone transformation based

on MRTD-GMM framework was proposed. The experimental results with a Vietnamese dataset revealed that a significant number of tonal units could be reduced by using the transformed tonal units instead of using the original ones, while the transformed tonal units are natural and high-intelligible with an "appropriate smoothness". As a consequence, the proposed lexical tone transformation was efficient for the SS of tonal languages.

Ensuring synthesized speech with an "appropriate smoothness" is the core common concept of the proposed methods in this research. With a determination of event functions close to the concept of co-articulation in speech, MRTD can synthesize smooth speech. The smoothness in synthesized speech can be adjusted by modifying event targets of MRTD. Therefore, MRTD can be used to synthesize speech with an "appropriate smoothness". As a result, MRTD was used throughout all proposed methods in this research.

Consequently, four issues were solved in order to improve the naturalness of synthesized speech under limited data conditions. The experimental results show that speech synthesized / transformed / modified by the proposed methods were intelligible and natural with an "appropriate smoothness". Therefore, the unified purpose of this research was mostly done and the motivation of this research was mostly archived. The remaining problem of this research is the lack of implementations and evaluations with several speakers and languages to confirm the speaker independence, the language independence and the generality of the proposed solutions and methods.

Based on the results of this dissertation, SS can be built with lower cost and lesser time-consuming while its quality can be still preserved and ensured. The results of this research can be applied for some classes of languages such as tonal languages with the proposed tone transformation or the general language with the proposed modification methods for CSS or with the proposed hybrid HTD synthesizer. The research results contribute to research fields of speech processing by introducing a new concept of "appropriate smoothness" in speech. The research results also contribute to research fields and development fields of SS in order to make SS more convenient and more efficient for human-machine interaction systems. The research results also contribute to the research fields of speech modification and voice transformation to make these systems more flexible.

7.2 Future works

Based on the results obtained from this dissertation, some future works and future research directions are specified.

First, the concepts of the proposed methods are speaker-independent but they were just implemented and evaluated with a single-speaker speech corpus. Therefore, the proposed methods will be evaluated with multi-speaker speech corpora in the future.

Second, the concepts of the proposed methods are language-independent, in which the proposed methods in chapter 6 can be applied for all tonal languages and the proposed methods in chapters 4 and 5 can be applied for the general language. However, these proposed methods were evaluated with only Vietnamese. Therefore, the proposed methods will be developed to other languages to increase their generalities and their language-independence.

Third, the institution that the author belong to is one member of the Asian network-based speech-to-speech translation project (A-STAR) [9], and the contributions of this dissertation can directly contribute to the tasks related to Vietnamese TTS, one component of the translation system.

Fourth, the main topics studied and solved in this work is SS under limited data conditions and these conditions are popular in under-resourced languages such as languages of the minorities. Therefore, one interesting research direction is to develop the proposed methods for languages of the minorities.

Firth, the proposed hybrid HTD in chapter 5 can be developed to synthesize speech with multiple target voices with a few amounts of target data. This result can contribute to solve the challenges of voice transformations, where people usually need a significant amount of target data. Therefore, this is a promising research direction in the future.

Sixth, the proposed speech smoothness measure DGSM in chapter 3 will be studied deeper to find out the relations between physical smoothness in speech and human perception of speech.

Last but not least, an important and interesting characteristic of TD, which was confirmed in this research, is that the event functions are closely related to linguistic information while the event targets convey non-linguistic information such as naturalness or speaking styles. Separating linguistic and non-linguistic information can help to process and control each component more accurately and efficiently. Therefore, one of possible future research direction is to improve the TD for this task.

Appendix A

Vietnamese speech synthesis

A.1 Vietnamese phonology

Vietnamese is the national, official language of Vietnam. It is the native language of approximately 90 million Vietnamese people, and of about three million Vietnamese residing elsewhere.

A brief review of Vietnamese phonology and phonetics is presented in this sub-section following the works by D. T. Thuat [137] and H. Phe [138].

Structure of Vietnamese syllables

Vietnamese is a typical monosyllabic tonal language [137, 138]. The total of pronounceable distinct syllables in Vietnamese is approximately 19000, but the syllables used in practice are only around 7000 and it is reduced to around 1200 when tone discriminations are discarded [138]. The structure of a Vietnamese syllable is depicted in Table. A.1. Each syllable could be considered as a combination of Initial, Final and Tone. There are 22 Initials and 155 Finals in Vietnamese [140].

The Initial component is generally a consonant, but can be omitted in some syllables. The Final can be decomposed into three parts, Onset, Nucleus and Coda. The Onset and Coda are optional and may not exist in a syllable. The Nucleus is a vowel or a diphthong, and the Coda is a consonant or a semi-vowel. There is one Onset, 16 Nuclei and 8 Codas.

Vietnamese Tones

Tone is a super-segmental feature which uniquely exists in a tonal language. There are six distinct tones in Vietnamese as described in Table. A.2 and Fig. A.1. Each tone has a distinct F0 contour shape. In Vietnamese, two kinds of syllable can be distinguished "open" and "closed" syllable. Closed syllables with the codas /p,/ /t/, /k/ could only be combined with tone rising and tone drop while open and other closed syllables could be combined with all six tones to become a meaningful tonal syllable.

A.2 Development of Vietnamese speech corpora

Research in Vietnamese TTS and ASR began two decades ago. However, only a few works have been carried out. One of the main reason is that there are not standard and reliable Vietnamese speech corpora.

In Vietnam, each research group has made their speech corpora itself and due to limitations of budget, time and knowledge, these corpora are always small and not high

Table A.1: Structure of a Vietnamese syllable

Tone			
Initial	Final		
	Onset	Nucleus	Coda

Tone Index	Vietnamese Name	English Name
1	Ngang	Level
2	Huyen	Falling
3	Nga	Broken
4	Hoi	Curve
5	Sac	Rising
6	Nang	Drop

Table A.2: Six Vietnamese Tones

quality. In addition, isolated corpora have made co-operation among Vietnamese research groups very difficult or even impossible. As a result, Vietnamese is still an under-resourced language and building a large-scale and high-quality Vietnamese corpora for common use becomes an urgent demand [72].

In this section, we review current Vietnamese speech corpora that have been used in speech researches.

Telephone Number Corpus

This corpus was created by Institute of Information Technology of Vietnam (IoIT) [72]. The audio format is WAV with 8 kHz sampling rate and 16-bit resolution. The size of corpus is quite small with 1541 words that describe telephone numbers. The first session contains 170 speakers including 94 males and 76 females from various localities in the north of Vietnam. The second session contains 208 speakers including 130 males and 78 female from mostly from the south of Vietnam. The corpus was labeled at phonetic levels.

Broadcasting Speech Corpus VOV (Voice of Vietnam)

This corpus was constructed by IoIT [72]. The corpus, including stories, mailbags, reports, etc. broadcast on the VOV, was collected from 15 speakers with standard Vietnamese accent. Sound files in RealAudio format were collected from VOV website and converted into WAV format (bit rate 256 kbps, mono channel, sampling rate 16 kHz).

The corpus contains 29062 utterances of on average 10-syllable length. The number of distinct syllables with tone is 4379 while the number of distinct syllables without tone is 1646 covering most of Vietnamese syllables. The total capacity of the corpus in WAV format is about 2.5GB.

The corpus is not phonetically balanced for each speaker and each session and it was manually labeled at syllable level only. Although this is a large-scale speech corpus, it is difficult to use this corpus for speech researches due to the lack of label at phonetic levels.

TTS Corpus DEMEN567

This corpus was constructed by IoIT [72, 98, 99]. The text content of this corpus was extracted from a short story shortly named DEMEN. Speaker was a female with standard

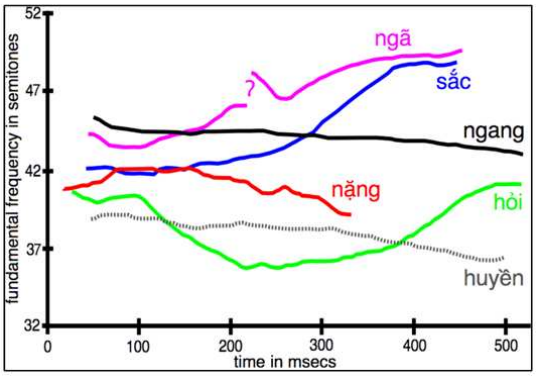

Figure A.1: General F0 contours of Six Vietnamese Tones: tone 1 (ngang - level), tone 2 (huyen - falling), tone 3 (nga - broken), tone 4 (hoi - curve), tone 5 (sac - rising) and tone 6 (nang - drop), adopted from [69]. The sign ? in tone 3 indicates that F0 contours with this tone are not consistent among the samples in the central region

.

Vietnamese accent. Recordings are of WAV format, 11025 Hz sampling rate, and 16-bit resolution. The corpus contains 567 utterances of an average length of 15 syllables. The size of this corpus is approximately 70 MB and the duration of speech is less than one hour.

This corpus was manually labeled in both syllable and phoneme levels with tone information. Although this is a small corpus, it can be one of high-quality Vietnamese speech corpora due to the careful design for the text script that ensures the phonetically balance and the good labeling at phonetic level.

MICA VNSpeechCorpus

This corpus was constructed by the International Research Center MICA of Hanoi University of Technology [141].

The corpus was spoken by 50 speakers including 25 females and 25 males with the age from 15 to 45 years old. The speakers were most educated at university level and represent three major dialect regions: the south, the north, and the central of Vietnam. Each speaker was asked for recording about 60 minutes of speech. The sampling frequency was 16 kHz.

It was shown that the corpus is acceptable and correctly balanced in terms of acoustic units and tones [141].

Although this corpus can be considered as a large-scale speech corpus, the phonetic transcription was performed by automatically forced-alignment labeling that reduces its accuracy and limited its applications.

Other corpora

Just recently, there are some research groups attempting to collect speech materials to build large-scale speech corpora, such as the Vietnamese broadcast news corpus (VNBN) with 40-hour and a spontaneous speech corpus with 11 hours of speech [142]. However, these corpora have not been labeled or just labeled at syllable level. Therefore, these corpora have not widely used for research.

A.3 Development of Vietnamese speech synthesis

Vietnamese SS has been just studied and developed for the last decade.

The two earliest ones are *VnSpeech* [143], which is a formant synthesizer, and *VnVoice* [144, 145], which is a unit-based concatenation synthesizer. The quality of these two systems is not high. Therefore, they are almost not alive at present.

There are some commercial Vietnamese TTSs such as *Voice of southern Vietnam (VOS)* [142], *Sao Mai Voice* [146], and *Hoa sung* [147]. These TTSs are based on concatenation at syllable or word levels, which can synthesize single neutral voices. The cores of these systems are text processing and there are almost none acoustical processing tasks. Therefore, these systems cannot be customized to enhanced features such as synthesizing multiple individual or emotional voices. Additionally, these kinds of TTS require a huge amount of resources and they can only run in high-performance servers instead of personal computers or devices.

The state-of-the-art Vietnamese TTS is a HMM-based Vietnamese TTS [135] that can be balanced between its quality and its open features for research and development. Therefore, this TTS was used in this research as a baseline system for comparison and for building the hybrid TTS.

Bibliography

[1] V.J. Santen, S. W. Richard, O.P. Joseph, H. Julia, Progress in Speech Synthesis, Springer, (1997).

[2] T. Dutoit, "An Introduction to Text-to-Speech Synthesis," *Text, Speech and Language Technology,* Vol. 3, p. 286, (1997).

[3] D. Klatt, "Review of Text-to-Speech Conversion for English," *Journal of the Acoustical Society of America,* vol. 82 (3), pp. 737-793, (1987).

[4] T. Portele, J. Kramer, "Adapting a TTS System to a Reading Machine for the Blind," *Proc. ICSLP 96*, (1996).

[5] W. Hess, "Speech Synthesis - A Solved Problem?," *Proc. EUSIPCO 92,* pp. 37-46, (1992).

[6] I. Murray, J. Arnott, N. Alm, and A. Newell, "A Communication System for the Disabled with Emotional Synthetic Speech Produced by Rule," *Proc. Eurospeech 91,* pp. 311-314, (1991).

[7] E. Abadjieva, I. Murray, and J. Arnott, "Applying Analysis of Human Emotion Speech to Enhance Synthetic Speech," *Proc. Eurospeech 93,* pp. 909-912, (1993).

[8] C. Hori, B. Zhao, S. Vogel, A. Waibel, H. Kashioka and S. Nakamura, "Consolidation-based Speech Translation and Evaluation Approach," *IEICE Transactions on Information and Systems,* Vol. E92-D , pp. 477-488, (2008).

[9] S. Sakti, N. Kimura, M. Paul, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C.M. Luong, and H. Li, "The Asian network-based speech-to-speech translation system," *Proc. ASRU 2009,* pp. 507-512, (2009).

[10] O. Turk, and L. M. Arslan, "Subband Based Voice Conversion," *Proc. ICSLP 2002,* Vol. 1, pp. 289-292, (2002).

[11] O. Turk, Cross-lingual Voice Conversion, PhD Thesis, Bogazici University, Turkey, (2007).

[12] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP'98,* pp. 655-658, (1998).

[13] Y. Stylianou, 0. CappC, and E. Moulines, "Statistical methods for voice quality transformation," *Proc. EUROSPEECH,* pp. 447-450, (1995).

[14] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc.ICASSP'98,* pp. 285-288, (1998).

[15] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals,* vol.E86-A, no.8, pp. 1956–1963, (2003).

[16] T. Toda, A. Black, K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Language Process,* v15-i8, pp. 2222-2235, (2007).

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

[18] P.B. Nguyen and M. Akagi, "Phoneme-based spectral voice conversion using temporal decomposition and Gaussian mixture mode," *Proc. ICCE-08,* pp. 224-229,(2008).

[19] 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.*

[20] B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," *Proc. Interspeech'07,* pp. 538-541, (2007).

[21] B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *Proc. NCSP'07,* pp. 481-484, (2007).

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

[23] V. Popa, J. Nurminen, M. Gabbouj, "A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models," *Interspeech 2009,* pp. 2655-2658, (2009).

[24] M. Akagi, T. Saitou, and C.F. Huang, "Voice conversion to add non-linguistic information into speaking voices," *JCA2007*, (2007).

[25] K. Sawamura, J. Dang, D. Erickson, A. Li, K. Sakaraba, N.Minematsu, and K. Hirose, "Common factors in emotion perception among different cultures," *Proc. ICPhS2007,* pp. 2113-2116, (2007).

[26] C.F. Huang and M. Akagi, "A rule-based speech morphing for verifying an expressive speech perception model," *Proc. Interspeech2007* pp. 2661-2664, (2007).

[27] W. Verhelst, T. Ceyssens and P. Wambacq, "On Inter-Signal Transplantation Of Voice Characteristics," *Proc. 3rd IEEE Benelux Signal Processing Symposium (SPS 2002,* pp. 137-140, (2002).

[28] T. Saitou, M. Unoki, and M. Akagi,"Development of an F0 control Model Based on F0 Dynamic Characteristics for Singing-Voice Synthesis," *Speech Communication,* 46, pp. 405-417, (2005).

[29] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices," *Proc. WASPAA2007,* pp. 215-218, (2007).

[30] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using STRAIGHT," *Proc. Interspeech'07*, pp. 4005-4006, (2007).

[31] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-Singing Synthesis System: Vocal Conversion from Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices," *NCMMSC2009*, (2009).

[32] K. Hideki and O. Hayato, "VOCALOID - Commercial singing synthesizer based on sample concatenation," *Proc. Interspeech 2007*

[33] N.B. Pinto, D.G. Childers, A.L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Trans. on Audio, Speech, and Language Proc,* vol.37, no.12, pp. 1870-1887, (1989).

[34] A. Acero, "Formant analysis and synthesis using hidden Markov models," *Proc. Eurospeech'99,* pp. 1047-1050, (1999).

[35] M.M. Sondhi, J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. on Acoustics, Speech and Signal Proc.,* vol.35, no.7, pp. 955-967, (1987).

[36] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *Journal of Acoustical Society of America,* 115(2), pp. 853-870, (2004).

[37] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units,*Proc. ICASSP-88,* pp. 679-682, (1988).

[38] A. Hunt, A. Black and W. Alan, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP-96*, **1**, pp. 373–376 (1996).

[39] A.W. Black, P.A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Proc. Eurospeech97,* volume 2, pp. 601-604, (1997).

[40] T. Shoham, D. Malah, and S. Shechtman, "Quality preserving compression of a concatenative text-to-speech acoustic database," *IEEE Trans. on Audio, Speech, and Lang. Proc.,* 20(3), pp. 1056–1068, (2012).

[41] D. Suendermann, H. H 묘 ge, A. Black, "Challenges in Speech Synthesis," *Speech Technology 2010,* pp 19-32, (2010).

[42] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute," *Tech Report CMU-LTI-03-177,* (2003).

[43] Y. Qian, F. K. Soong, Z. Yan, "A Unified Trajectory Tiling Approach to High Quality Speech Rendering," *IEEE Trans. on Audio, Speech, and Language Proc.,* Vol. 21, No. 2, pp. 280–290, (2013).

[44] H. Zen, T. Toda "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," *Proc. Interspeech,* (2005).

[45] K. Tokuda, H. Zen, A.W. Black, "An HMM-based speech synthesis system applied to English," *Proc. of 2002 IEEE SSW,* (2002).

[46] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP,* pp.13151318, (2000).

[47] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans on Information and Systems,* Vol.E90-D, No.5 pp. 816-824, (2007).

[48] M. Zhang, J. Tao, H. Jia, X. Wang, "Improving HMM Based Speech Synthesis by Reducing Over-Smoothing Problems," *Proc. ISCSLP '08,* pp. 1–4, (2008).

[49] K. Hashimoto, S. Takaki, K. Oura, K. Tokuda, "Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2011," *Blizzard Challenge 2011.*

[50] M. Plumpe, A. Acero, H.W. Hon, X.D. Huang, "HMM-based smoothing for concatenative speech synthesis," *Proc. ICSLP 1998,* pp. 27512754.

[51] J.H. Yang, Z.W. Zhao, Y.H. Jiang, X.R. Wu, "Multi-tier nonuniform unit selection for corpus-based speech synthesis," *Proc. Blizzard Challenge Workshop 2006.*

[52] Z.H. Ling, R.H. Wang, "Minimum unit selection error training for HMM-based unit selection speech synthesis system," *Proc. ICASSP 2008,* pp. 39493952.

[53] T. Hirai and S. Tenpaku, "Using 5ms segments in concatenative speech synthesis," *Proc. 5th Speech Synth,* (2004).

[54] L.F. Lamel, R.H. Kassel, S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," *Proc. Speech I/O assessment and speech databases*, (1989).

[55] J. Matousek, J. Psutka, J. Krouta, "Design of Speech Corpus for Text-to-Speech Synthesis," *Proc. Eurospeech 2001*, (2001).

[56] A. Iida, N. Cambell, "Speech Database Design for a Concatenative Text-to-Speech Synthesis System for Individuals with Communication Disorders," *Intl. Journal of Speech Technology,* vol. 6, pp. 379392, (2003).

[57] B. Bozkurt, O. Ozturk, T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection*Proc. Eurospeech 2003*, (2003).

[58] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of Speech Corpus for Mandarin Text to Speech," *The Blizzard Challenge 2008 workshop,* (2008).

[59] S. Kiruthiga and K. Krishnamoorthy, "Design Issues in Developing Speech Corpus for Indian Languages A survey," *Proc. ICCCI -2012,* (2012).

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

[61] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations, *Proc. Eurospeech: ESCA,* pp. 10291032, (1995).

[62] J. Wouters, and M.W. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Trans. on Audio, Speech, and Lang. Proc.,* 9(1), pp. 30–38, (2001).

[63] 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).

[64] A. Kain, Q. Miao, and J. van Santen, "Spectral control in concatenative speech synthesis," *Proc. ISCA Workshop on Speech Synthesis,* (2007).

[65] R. McArdle, R.H. Wilson, "Speech Perception in Noise: The Basics," *Perspectives on Hearing and Hearing Disorders: Research and Diagnostics,* No. 13, pp. 4-13, (2009).

[66] P. Lieberman, S.B. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *Journal of the Acoustical Society of America,* Vol. 34, pp. 922927, (1962).

[67] S. Luksaneeyanawin, Tone transformation, *Proc. SNLP'95*, pp. 345-360, (1995).

[68] Z. Wu, T. Kinnunen, E.S Chng, H. Li, "Text-Independent F0 Transformation with Non-Parallel Data for Voice Conversion," *Proc. Interspeech 2010,* pp. 1732-1735, (2010).

[69] Nguyen, VL., Edmondson, A., "Tones and voice quality in modern northern Vietnamese: Instrumental case studies," *Mon-Khmer Studies* 28: pp. 1-18, (1998).

[70] Z. Wu, Y. Qian, F.K. Soong, B. Zhang, "Modeling and Generating Tone Contour with Phrase Intonation for Mandarin Chinese Speech," *ISCSLP '08,* pp. 1-4, (2008).

[71] M. Wang, M. Wen, K.Hirose, N. Minematsu, "Emotional Voice Conversion for Mandarin using Tone Nucleus Model Small Corpus and High Efficiency," *Proc. Speech Prosody 2012.*

[72] L.C. Mai and D.N. Duc, "Design of Vietnamese speech corpus and current status," *Proc. ISCSLP-06*, pp. 748–758 (2006).

[73] P. Delattre, "Co-articulation and the Locus theory," *Studia Linguistica*, 23(1), 1–26 (1969).

[74] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "An investigation on speech perception over coarticulation," *Proc. ICSAP 2011*, pp. 507-511, Singapore (2011).

[75] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "An investigation on perceptual line spectral frequency (PLP-LSF) target stability against the vowel neutralization phenomenon," *Proc. ICSAP 2011*, pp.512-514, Singapore (2011).

[76] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "On the Stability of Spectral Targets under Effects ofCoarticulation," *J. Computer & Electrical Engineering,* 4(4), pp. 537-541, (2012) (selected from ICSAP 2011).

[77] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "An Investigation on Speech Perception Under Effects of Coarticulation," *J. Computer & Electrical Engineering,* 4(4), pp. 532-536, (2012) (selected from ICSAP 2011).

[78] A.M. Liberman, I.G. Mattingly, "The motor theory of speech perception revised," *Cognition,* 21 (1), pp. 1-36, (1985).

[79] 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).

[80] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE Publication,* pp. 561-580, (1975).

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

[82] 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).

[83] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. ICASSP'84,* pp. 1.10.1-1.10.4, (1984).

[84] H. Kawahara, "STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci & Tech.*, 27(6), 349–353 (2006).

[85] H. Kawahara, I. Masuda-Katsuse, A. de Cheveign 辿, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication,* **27**(3-4), 187-207 (1999).

[86] 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.

[87] P.C. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," *IEICE Trans. Inf. and Syst.*, E86-D(3), pp. 397-405, (2003).

[88] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP-83*, pp. 81–84 (1983).

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

[90] M. Niranjan and F. Fallside, "Temporal decomposition: A framework for enhanced speech recognition," *Proc. ICASSP'89*, pp. 655-658, (1989).

[91] S. Kim and Y. Oh, "Efficient quantisation method for LSF parameters based on restricted temporal decomposition," *Electronics Letters*, 35(12), 962-964 (1999).

[92] P.C. Nguyen, M. Akagi, T.B. Ho, "Temporal decomposition: A promising approach to VQ-based speaker identification," *Proc. ICASSP-03,* pp. 184–187, (2003).

[93] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "A low-cost concatenative TTS for monosyllabic languages," *IEICE Tech. Rep.*, Vol.112, No.81 , IEICE-SP2012-35, pp. 13-18, Tokyo, Japan (2012).

[94] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "Transformation of F0 contours for lexical tones in concatenative speech synthesis of tonal languages," *Proc. Oriental COCOSDA 2012*, pp.129 - 134, Macau (2012).

[95] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "A concatenative speech synthesis for monosyllabic languages with limited data," *Proc. APSIPA ASC 2012*, pp.1 - 10, USA (2012).

[96] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "Improving the flexibility of unit selection TTS with Temporal Decomposition," *ASJ Spring Meeting 2013*, Tokyo, Japan (2013).

[97] Trung-Nghia Phung, Tuan Dinh, Thang Tat Vu, Mai Chi Luong, Masato Akagi, "Constructing natural concatenative speech synthesis under limited data conditions," *submitted to Acoust. Sci & Tech,* (2013).

[98] Trung-Nghia Phung, Thanh-Son Phan, Thang Tat Vu, Mai Chi Luong, Masato Akagi, "Improving the naturalness of HMM-based TTS under limited data conditions," *submitted to IEICE Trans. Inf. and Syst,* (2013).

[99] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi, "A Hybrid TTS between Unit Selection and HMM-based TTS in limited data conditions," *Proc. ISCA Speech Synthesis Workshop (SSW8),* August, Spain (2013) (submitted).

[100] Trung-Nghia Phung, Masashi Unoki, Masato Akagi, "Comparative evaluation of bone-conducted-speech restoration based on linear prediction scheme," *IEICE Tech. Rep.* vol. 110, no. 71, EA2010-31, pp. 53-58, Hokkaido, Japan (2010).

[101] Trung-Nghia Phung, Masashi Unoki, Masato Akagi, "Improving Bone-Conducted Speech Restoration in noisy environment based on LP scheme," *APSIPA Annual Summit and Conference 2010*, p. 10, Singapore (2010).

[102] Masashi Unoki, Phung Nghia Trung, Masato Akagi, "Comparative evaluation and improvement of bone-conducted-speech restoration method based on linear prediction," *ASJ Spring Meeting 2011.* 1-Q-31, pp. 819-822, Japan (2011) (In Japanese).

[103] Phung Nghia Trung, Masashi Unoki, Masato Akagi, "A Study on Restoration of Bone-Conducted Speech in Noisy Environment with LP-based Model and Gaussian Mixture Model," *J. Signal Processing,* 16(5), pp.409-417, 2012-09.

[104] R.B. Chicote, J. Yamagishi, S. King, J.M. Montero, J.M. Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech communication,* Vol. 52, pp. 394404, (2010).

[105] L. Pols, "Multilingual Synthesis Evaluation Methods," *Proc. ICSLP 92,* (1), pp. 181-184, (1992).

[106] V. Botinhao, C. Yamagishi, S. King, "Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise," *Proc. ICASSP-11,* pp. 5112-5115, (2011).

[107] M. Goldstein, "Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener," *Speech Communication,* vol. 16, pp. 225-244, (1995).

[108] C. Benoit, M. Crice, V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication,* vol. 18, issue. 4, pp. 381-392, (1996).

[109] J. Li, D. Sityaev, and J. Hao, "Sentence Level Intelligibility Evaluation for Mandarin Text-to-Speech Systems Using Semantically Unpredictable Sentences," *Proc. Interspeech 2007,* pp. 1350–1353, (2007).

[110] H.C. Nusbaum, A.L. Francis, A.S. Henly, "Measuring the naturalness of synthetic speech," *International Journal of Speech Technology,* Volume 2, Issue 1, pp 7-19, (1997).

[111] L. Neovius, P. Raghavendra, "Comprehension of KTH Text-to-Speech with Listening Speed Program," *Proc. Eurospeech 93,* (3), pp. 1687-1690, (1993).

[112] http://festvox.org/blizzard/

[113] S. King and V. Karaikos, "The Blizzard Challenge 2012."

[114] Y.Y. Chen, T.W. Kuan, C.Y. Tsai, J.F. Wang, C.H. Chang, "Speech variability compensation for expressive speech synthesis," *Proc. ICOT 2013,* pp. 210-213, (2013).

[115] G. Beller, N. Obin, X. Rodet, "Articulation Degree as a Prosodic Dimension of Expressive Speech," *Proc. Speech prosody 2008,* (2008).

[116] S. Lee, E. Bresch, S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," *Proc. 7th Int. Seminar Speech Production,* pp. 11-17, (2006).

[117] S.W. Lee, S.T. Ang, M. Dong, and H. Li, "Generalized F0 modelling with absolute and relative pitch features for singing voice synthesis," *Proc. ICASSP 2012,* pp. 429 - 432, (2012).

[118] J.F. Pitrelli, R. Bakis, E.M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English," *IEEE Trans. Audio, Speech, and Language Proc.,* Vol. 14, No. 4, (2006).

[119] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *Proc. Eusipco 94,* pp. 1182-1185, (1994).

[120] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Audio, Speech, and Language Proc.,* Vol. 9, No. 5, pp. 504 - 512, (2001).

[121] E. Mehlum, C. Tarrou, "Invariant smoothness measures for surfaces," *Advances in Computational Mathematics,* vol. 8, (1998).

[122] E.B. Dagum, M. Morry, "Basic Issues on the Seasonal Adjustment of the Canadian Consumer Price Index," *Journal of Business & Economic Statistics,* Vol. 2, No. 3, pp. 250-259, (1984).

[123] S. Singh, "Pattern Recognition and Image Analysis," *Proc. Advances in Pattern Recognition ICAPR 2005.*

[124] J.T. Schwartz, M. Sharir, "Algorithmic motion planning in robotics," *Algorithms and Complexity,* pp. 391-430, (1990).

[125] B. Efron, "Defining the curvature of a statistical problem (with applications to second order efficiency," *The Annals of Statistics,* (1975).

[126] G Kindlmann, R Whitaker, T Tasdizen, "Curvature-based transfer functions for direct volume rendering: Methods and applications," *Proc. VIS 2003,* (2003).

[127] D.L. Thomson, R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features," *Proc. ICASSP 98,* vol. 1, pp. 21 - 24, (1998).

[128] J.G. Wilpon, H. Murray, C.H. Lee, L. Rabiner, "Improvements in connected digit recognition using higher order spectral and energy features," *Proc. ICASSP 91,* vol. 1, pp. 349 - 352, (1991).

[129] S. Rangachari, P.C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech communication,* Vol. 48, Issue. 2, pp. 220231, (2006).

[130] Kent R.D and R. Charles, "The acoustic analysis of speech," *San Diego: Singular Publishing Group,* ISBN 1-879105-43-8(1992).

[131] H. W. Strube, R. Wilhelms, "Synthesis of unrestricted German speech from interpolated log-area-ratio coded transitions," *Speech Communication,* 93-102, (1982).

[132] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, 80(4), 1016–1025 (1986).

[133] T.V. Do, D.D. Tran, and T.T. Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," *Proc. SoICT '11,* pp. 165-171, (2011).

[134] J.B. Tenenbaum, and W.T. Freeman, "Separating Style and Content with Bilinear Models," *Neural Computation,* 12(6),pp. 1247-1283, (2000).

[135] TT. Vu, MC. Luong and S. Nakamura, "An HMM-based Vietnamese speech synthesis system, Speech Database and Assessments," *Proc. COCOSDA-2009*, pp. 116–121 (2009).

[136] D. Erro, A. Moreno, A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora," *IEEE Trans. Audio, Speech, and Language Proc.,* vol.18, no.5, pp.944-953, (2010).

[137] D.T. Thuat, Vietnamese Phonology, Vietnamese National Publisher, Second edition, (2003) (In Vietnamese).

[138] H. Phe, Vietnamese Grammar, Da Nang Publisher, (2003) (In Vietnamese).

[139] E. Helander, J. Nurminen, M. Gabbouj, " LSF mapping for voice conversion with very small training sets," *Proc. ICASSP 2008,* pp. 4669 - 4672, (2008).

[140] T.T Vu, D.T. Nguyen, M.C. Luong, and J.P. Hosom, "Vietnamese large vocabulary continuous speech recognition," *Proc. INTERSPEECH 2005,* pp. 1689-1692, (2005).

[141] V.B. Le, D.D. Tran, L. Besacier, E. Castelli, and J.F. Serignat, "First steps in building a large vocabulary continuous speech recognition system for Vietnamese," *Proc. RIVF05,* pp. 330-333, pp. 21-24, (2005).

[142] Quan VU, "VOS: The Corpus-based Vietnamese Text-to-speech System," *Journal on Information, Technologies, anh Communications,* (2010).

[143] H.M. Le, "Some results in Research and Development of Text To Speech conversion system for Vietnamese language based on formant synthesis," *Proc. ICT.RDA 2003* (2003).

[144] H. Mixdorff, D.T. Nguyen and T.W. Nghia, "Duration Modeling in a Vietnamese Text-to-Speech System," *Proc. SPECOM 2005* (2005).

[145] DT. Nguyen, H. Mixdorff, MC. Luong, HH. Ngo, and BK. Vu, "Fujisaki Model based F0 contours in Vietnamese TTS," *Proc. ICSLP 2004* (2004).

[146] Sao Mai Computer Center for the Blind SMCC, Sao Mai Voice, [Online] http://www.saomaicenter.org/.

[147] International Research Center MICA, Hoa Sung, [Online] http://www.mica.edu.vn/tts/.

Publications

Journal Articles

[1] Trung-Nghia Phung, Thanh-Son Phan, Thang Tat Vu, Mai Chi Luong, Masato Akagi. Improving the naturalness of HMM-based TTS under limited data conditions. *IEICE Trans. Inf. and Syst.* Vol.E96-D, No.11, (2013).

[2] Phung Nghia Trung, Masashi Unoki, Masato Akagi. A Study on Restoration of Bone-Conducted Speech in Noisy Environment with LP-based Model and Gaussian Mixture Model. *J. Signal Processing.* 16(5), pp.409-417, (2012).

[3] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. An Investigation on Speech Perception Under Effects of Coarticulation. *J. Computer & Electrical Engineering.* 4(4), pp. 532-536, (2012) (selected from ICSAP 2011).

[4] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. On the Stability of Spectral Targets under Effects ofCoarticulation. *J. Computer & Electrical Engineering.* 4(4), pp. 537-541, (2012) (selected from ICSAP 2011).

International Conference Articles

[5] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. A Hybrid TTS between Unit Selection and HMM-based TTS in limited data conditions. *the 8th ISCA Speech Synthesis Workshop (SSW8),* August, Spain (2013) .

[6] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. Transformation of F0 contours for lexical tones in concatenative speech synthesis of tonal languages. *International Conference on Speech Database and Assessments (Oriental COCOSDA) 2012*, pp.129 - 134 , December, Macau (2012).

[7] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. A concatenative speech synthesis for monosyllabic languages with limited data. *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC) 2012*, pp.1 - 10 , December, USA (2012).

[8] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. An investigation on perceptual line spectral frequency (PLP-LSF) target stability against the vowel neutralization phenomenon. *3rd International Conference on Signal Acquisition and Processing (ICSAP 2011)*, pp.512-514, Singapore (2011).

[9] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. An investigation on speech perception over coarticulation. *3rd International Conference on Signal Acquisition and Processing (ICSAP 2011)* - pp. 507-511, March, Singapore (2011).

[10] Trung-Nghia Phung, Masashi Unoki, Masato Akagi. Improving Bone-Conducted Speech Restoration in noisy environment based on LP scheme. *APSIPA Annual Summit and Conference 2010*, p. 10, Dec, Singapore (2010).

Domestic Conference Articles

[11] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. Improving the naturalness of speech synthesized by HMMSS by producing an appropriate smoothness. *ASJ Autumn Meeting 2013.*, Japan, (2013).

[12] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. Improving the flexibility of unit selection TTS with Temporal Decomposition. *ASJ Spring Meeting 2013.*, March, Tokyo, Japan, (2013).

[13] Trung-Nghia Phung, Mai Chi Luong, Masato Akagi. A low-cost concatenative TTS for monosyllabic languages. *IEICE Tech. Rep.*, Vol.112, No.81 , IEICE-SP2012-35, pp. 13-18, June, Tokyo, Japan, (2012).

[14] Masashi Unoki, Phung Nghia Trung, Masato Akagi. Comparative evaluation and improvement of bone-conducted-speech restoration method based on linear prediction. *ASJ Spring Meeting 2011.* 1-Q-31, pp. 819-822, March, Japan (2011) (In Japanese).

[15] Trung-Nghia Phung, Masashi Unoki, Masato Akagi. Comparative evaluation of bone-conducted-speech restoration based on linear prediction scheme. *IEICE Tech. Rep.* vol. 110, no. 71, EA2010-31, pp. 53-58, June, Hokkaido, Japan (2010).