NAIST-IS-DT0161027

Doctoral Thesis

High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion

Tomoki Toda

March 24, 2003

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctoral Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Tomoki Toda

Thesis committee: Kiyohiro Shikano, Professor
Yuji Matsumoto, Professor
Nick Campbell, Professor
Hiroshi Saruwatari, Associate Professor

High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*

Tomoki Toda

Abstract

Text-to-Speech (TTS) is a useful technology that converts any text into a speech signal. It can be utilized for various purposes, e.g. car navigation, announcements in railway stations, response services in telecommunications, and e-mail reading. Corpus-based TTS makes it possible to dramatically improve the naturalness of synthetic speech compared with the early TTS. However, no general-purpose TTS has been developed that can consistently synthesize sufficiently natural speech. Furthermore, there is not yet enough flexibility in corpus-based TTS.

This thesis addresses two problems in speech synthesis. One is how to improve the naturalness of synthetic speech in corpus-based TTS. The other is how to improve control of speaker individuality in order to achieve more flexible speech synthesis. To deal with the former problem, we focus on two factors: (1) an algorithm for selecting the most appropriate synthesis units from a speech corpus, and (2) an evaluation measure for selecting the synthesis units. Moreover, we focus on a voice conversion technique to control speaker individuality to deal with the latter problem.

Since various vowel sequences appear frequently in Japanese, it is not realistic to prepare long units that include all possible vowel sequences to avoid vowel-to-vowel concatenation, which often produces auditory discontinuity. In order to address this problem, we propose a novel segment selection algorithm based on both phoneme and diphone units that does not avoid concatenation of

*Doctoral Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0161027, March 24, 2003.

vowel sequences but alleviates the resulting discontinuity. Experiments testing concatenation of vowel sequences clarify that better segments can be selected by considering concatenations not only at phoneme boundaries but also at vowel centers. Moreover, the results of perceptual experiments show that speech synthesized using the proposed algorithm has better naturalness than that using the conventional algorithms.

A cost is established as a measure for selecting the optimum waveform segments from a speech corpus. In order to achieve high-quality segment selection for concatenative TTS, it is important to utilize a cost that corresponds to perceptual characteristics. We first clarify the correspondence of the cost to the perceptual scores and then evaluate various functions to integrate local costs capturing the degradation of naturalness in individual segments. From the results of perceptual experiments, we find a novel cost that takes into account not only the degradation of naturalness over the entire synthetic speech but also the local degradation. We also clarify that the naturalness of synthetic speech can be slightly improved by utilizing this cost and investigate the effect of using this cost for segment selection.

We improve the voice conversion algorithm based on the Gaussian Mixture Model (GMM), which is a conventional statistical voice conversion algorithm. The GMM-based algorithm can convert speech features continuously using the correlations between source and target features. However, the quality of the converted speech is degraded because the converted spectrum is excessively smoothed by the statistical averaging operation. To overcome this problem, we propose a novel voice conversion algorithm that incorporates Dynamic Frequency Warping (DFW) technique. The experimental results reveal that the proposed algorithm can synthesize speech with a higher quality while maintaining equal conversion-accuracy for speaker individuality compared with the GMM-based algorithm.

Keywords:

Text-to-Speech, Naturalness, speaker individuality, Segment selection, Synthesis unit, Measure for selection, Voice conversion

Acknowledgments

I would like to express my deepest appreciation to Professor Kiyohiro Shikano of Nara Institute of Science and Technology, my thesis advisor, for his constant guidance and encouragement through my master's course and doctoral course.

I would also like to express my gratitude to Professor Yuji Matsumoto, Professor Nick Campbell, and Associate Professor Hiroshi Saruwatari, of Nara Institute of Science and Technology, for their invaluable comments to the thesis.

I would sincerely like to thank Dr. Nobuyoshi Fugono, President of ATR, and Dr. Seiichi Yamamoto, Director of ATR Spoken Language Translation Research Laboratories, for giving me the opportunity to work for ATR Spoken Language Translation Research Laboratories as an Intern Researcher.

I would especially like to express my sincere gratefulness to Dr. Hisashi Kawai, Supervisor of ATR Spoken Language Translation Research Laboratories, for his continuous support and valuable advice through the doctoral course. The core of this work originated with his pioneering ideas in speech synthesis, which led me to a new research idea. This work could not have been accomplished without his direction. I could learn many lessons from his attitude toward study. I have always been happy to carry out research with him.

I would like to thank Assistant Professor Hiromichi Kawanami and Assistant Professor Akinobu Lee of Nara Institute of Science and Technology, for their beneficial comments. I would also like to thank former Associate Professor Satoshi Nakamura, who is currently Head of Department 1 at ATR Spoken Language Translation Research Laboratories, and former Assistant Professor Jinlin Lu, who is currently an Associate Professor at Aichi Prefectural University, for their helpful discussions. I want to thank all members of the Speech and Acoustics Laboratory and Applied Linguistics Laboratory in Nara Institute of Science and

Technology for providing fruitful discussions. I would especially like to thank Dr. Toshio Hirai, Senior Researcher of Arcadia Inc., for providing thoughtful advice and discussions on speech synthesis techniques. Also, I owe a great deal to Ryuichi Nishimura, doctoral candidate of Nara Institute of Science and Technology, for his support in the laboratories.

I greatly appreciate Dr. Hideki Tanaka, the Head of Department 4 at ATR Spoken Language Translation Research Laboratories, for his encouragement. I would sincerely like to thank Minoru Tsuzaki, Senior Researcher, and Dr. Jinfu Ni, a Researcher, of ATR Spoken Language Translation Research Laboratories, for providing lively and fruitful discussions about speech synthesis. I would also like to thank my many other colleagues at ATR Spoken Language Translation Research Laboratories.

I am indebted to many Researchers and Professors. I would especially like to express my gratitude to Dr. Masanobu Abe, Associate Manager, Senior Research Engineer of NTT Cyber Space Laboratories, Professor Hideki Kawahara of Wakayama University, Associate Professor Keiichi Tokuda of Nagoya Institute of Technology, and Professor Yoshinori Sagisaka of Waseda University, for their valuable advice and discussions. I would also like to express my gratitude to Professor Fumitada Itakura, Associate Professor Kazuya Takeda, Associate Professor Syoji Kajita, and Research Associate Hideki Banno, of Nagoya University, and Associate Professor Mikio Ikeda of Yokkaichi University, for their support, guidance, and having recommended that I enter Nara Institute of Science and Technology.

Finally, I would like to acknowledge my family and friends for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Problem Definition

Speech is the ordinary way for most people to communicate. Moreover, speech can convey other information such as emotion, attitude, and speaker individuality. Therefore, it is said that speech is the most natural, convenient, and useful means of communication.

In recent years, computers have come into common use as computer technology advances. Therefore, it is important to realize a man-machine interface to facilitate communication between people and computers. Naturally, speech is focused on as a medium for such communication. In general, two technologies for processing speech are needed. One is speech recognition, and the other is speech synthesis. Speech recognition is a technique for information input. Necessary information, e.g. message information, is extracted from input speech that includes diverse information. Thus, it is important to find a method to extract only useful information. On the other hand, speech synthesis is a technique for information output. This procedure is the reverse of speech recognition. Output speech includes various types of information, e.g. sound information and prosodic information, and is generated from input information. Moreover, other information such as speaker individuality and emotion is needed in order to realize smoother communication. Thus, it is important to find a method to generate the various types of paralinguistic information that are not processed in speech recognition.

Text-to-Speech (TTS) is one of the speech synthesis technologies. TTS is

a technique to convert any text into a speech signal [67], and it is very useful in many practical applications, e.g. car navigation, announcements in railway stations, response services in telecommunications, and e-mail reading. Therefore, it is desirable to realize TTS that can synthesize natural and intelligible speech, and research and development on TTS has been progressing.

The current trend in TTS is based on a large amount of speech data and statistical processing. This type of TTS is generally called corpus-based TTS. This approach makes it possible to dramatically improve the naturalness of synthetic speech compared with the early TTS. Corpus-based TTS can be used for practical purposes under limited conditions [15]. However, no general-purpose TTS has been developed that can synthesize sufficient natural speech consistently for any input text. Furthermore, there is not yet enough flexibility in corpus-based TTS. In general, corpus-based TTS can synthesize only speech having the specific style included in a speech corpus. Therefore, in order to synthesize other types of speech, e.g. speech of various speakers, emotional speech, and other speaking styles, various speech samples need to be recorded in advance. Moreover, large-sized speech corpora are needed to synthesize speech with sufficient naturalness. Speech recording is hard work, and it requires an enormous amount of time and expensive costs. Therefore, it is necessary to improve the performance of corpus-based TTS.

1.2 Thesis Scope

This thesis addresses two problems in speech synthesis shown in **Figure 1.1**. One is how to improve the naturalness of synthetic speech in corpus-based TTS. The other is how to improve control of speaker individuality in order to achieve more flexible speech synthesis.

1.2.1 Improvement of naturalness of synthetic speech

In corpus-based TTS, three main factors determine the naturalness of synthetic speech: (1) a speech corpus, (2) an algorithm for selecting the most appropriate synthesis units from the speech corpus, and (3) an evaluation measure to select the synthesis units. We focus on the latter two factors.

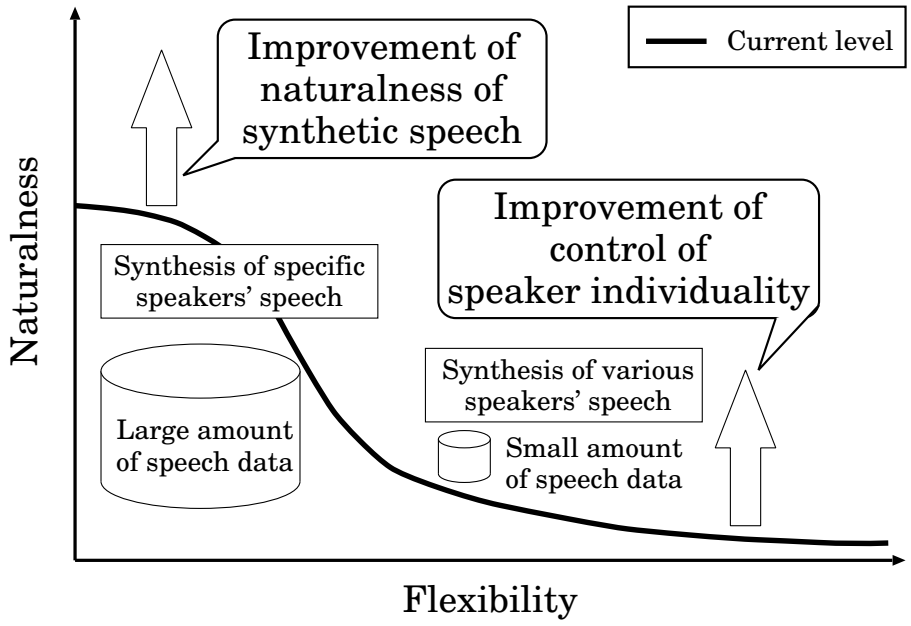

Figure 1.1. Problems addressed in this thesis.

In a speech synthesis procedure, the optimum set of waveform segments, i.e. portions of speech utterances included in the corpus, are selected, and the synthetic speech is generated by concatenating the selected waveform segments. This selection is performed based on synthesis units. Various units, phonemes, diphones, syllables, and so on have been proposed. In Japanese speech synthesis, syllable units are often used since the number of Japanese syllables is small and transition in the syllables is important for intelligibility. However, syllable units cannot avoid vowel-to-vowel concatenation, which often produces auditory discontinuity, because various vowel sequences appear frequently in Japanese. In order to alleviate this discontinuity, we propose a novel selection algorithm based on two synthesis unit definitions.

Moreover, in order to realize high and consistent quality of synthetic speech, it is important to use an evaluation measure that corresponds to perceptual characteristics in the selection of the most suitable waveform segments. Although a measure based on acoustic measures is often used, the correspondence of such a measure to the perceptual characteristics is indistinct. Therefore, we clarify

the correspondence of the measure utilized in our TTS by performing perceptual experiments on the naturalness of synthetic speech. Moreover, we improve this measure based on the results of these experiments.

1.2.2 Improvement of control of speaker individuality

We focus on a voice conversion technique to control speaker individuality. In this technique, conversion rules between two speakers are extracted in advance using a small amount of training speech data. Once training has been performed, any utterance of one speaker can be converted to sound like that of another speaker. Therefore, we can easily synthesize speech of various speakers from only a small amount of speech data of the speakers by using the voice conversion technique.

However, the performance of conventional voice conversion techniques is inadequate. The training of the conversion rules is performed based on statistical methods. Although accurate conversion rules can be extracted from a small amount of training data, important information influencing speech quality is lost. In order to avoid quality degradation caused by losing the information, we introduce Dynamic Frequency Warping (DFW) technique into the statistical voice conversion. From the results of perceptual experiments, we show that the proposed voice conversion algorithm can synthesize converted speech more naturally while maintaining equal conversion-accuracy on speaker individuality compared with a conventional voice conversion algorithm.

1.3 Thesis Overview

The thesis is organized as follows.

In **Chapter 2**, a corpus-based TTS system and conventional voice conversion techniques are described. We describe the basic structure of the corpus-based TTS system. Then some techniques in each module are reviewed, and we briefly introduce the techniques applied to the TTS system under development in ATR Spoken Language Translation Research Laboratories. Moreover, some conventional voice conversion algorithms are reviewed and conversion functions of the algorithms are compared.

In **Chapter 3**, we propose a novel segment selection algorithm for Japanese speech synthesis. Not only the segment selection algorithms but also our measure for selection of optimum segments are described. Results of perceptual experiments show that the proposed algorithm can synthesize speech more naturally than conventional algorithms.

In **Chapter 4**, the measure is evaluated based on perceptual characteristics. We clarify correspondence of the measure to the perceptual scores determined from the results of perceptual experiments. Moreover, we find a novel measure having better correspondence and investigate the effect of using this measure for segment selection. We also show the effectiveness of increasing the size of a speech corpus.

In **Chapter 5**, control of speaker individuality by voice conversion is described. We propose a novel voice conversion algorithm and perform an experimental evaluation on it. The results of experiments show that the proposed algorithm has better performance compared with a conventional algorithm.

In **Chapter 6**, we summarize the contributions of this thesis and offer suggestions for future work.

Chapter 2

Corpus-Based Text-to-Speech and Voice Conversion

Corpus-based TTS is the main current direction in work on TTS. The naturalness of synthetic speech has been improved dramatically by the transition from the early rule-based TTS to corpus-based TTS. In this chapter, we describe the basic structure of corpus-based TTS and the various techniques used in each module. Moreover, we review conventional voice conversion algorithms that are useful for flexibly synthesizing speech of various speakers, and then we compare various conversion functions.

2.1 Introduction

The early TTS was constructed based on rules that researchers determined from their objective decisions and experience [67]. In general, this type of TTS is called rule-based TTS. The researcher extracts the rules for speech production by the Analysis-by-Synthesis (A-b-S) method [6]. In the A-b-S method, parameters characterizing a speech production model are adjusted by performing iterative feedback control so that the error between the observed value and that produced by the model is minimized. Such rule determination needs professional expertise since it is difficult to extract consistent and reasonable rules. Therefore, the rule-based TTS systems developed by researchers usually have different performances. Moreover, synthetic speech by rule-based TTS has an unnatural quality because a

speech waveform is generated by a speech production model, e.g. terminal analog speech synthesizer, which generally needs some approximations in order to model the complex human vocal mechanism [67].

On the other hand, the current TTS is constructed based on a large amount of data and a statistical process [43][89]. In general, this type of TTS is called corpus-based TTS in contrast with rule-based TTS. This approach has been developed through the dramatic improvements in computer performance. In corpus-based TTS, a large amount of speech data are stored as a speech corpus. In synthesis, optimum speech units are selected from the speech corpus. An output speech waveform is synthesized by concatenating the selected units and then modifying their prosody. Corpus-based TTS can synthesize speech more naturally than rule-based TTS because the degradation of naturalness in synthetic speech can be alleviated by selecting units satisfying certain factors, e.g. a mismatch of phonetic environments, difference in prosodic information, and discontinuity produced by concatenating units. If the selected units need little modification, natural speech can be synthesized by concatenating speech waveform segments directly. Furthermore, since the corpus-based approach has hardly any dependency on the type of language, we can apply the approach to other languages more easily than the rule-based approach.

If a large-sized speech corpus of a certain speaker can be used, corpus-based TTS can synthesize high-quality and intelligible speech of the speaker. While, not only quality and intelligibility but also speaker individuality are important for smooth and full communication. Therefore, it is important to synthesize the speech of various speakers as well as the speech of a specific speaker. One of approaches for flexibly synthesizing speech of various speakers is speech modification by a voice conversion technique used to convert one speaker's voice into another speaker's voice [68].

In voice conversion, it is important to extract accurate conversion rules from a small amount of training data. This problem is associated with a mapping between features. In general, an extraction method of conversion rules is based on statistical processing, and it is often used in speaker adaptation for speech recognition.

This chapter is organized as follows. In **Section 2.2**, we describe the basic

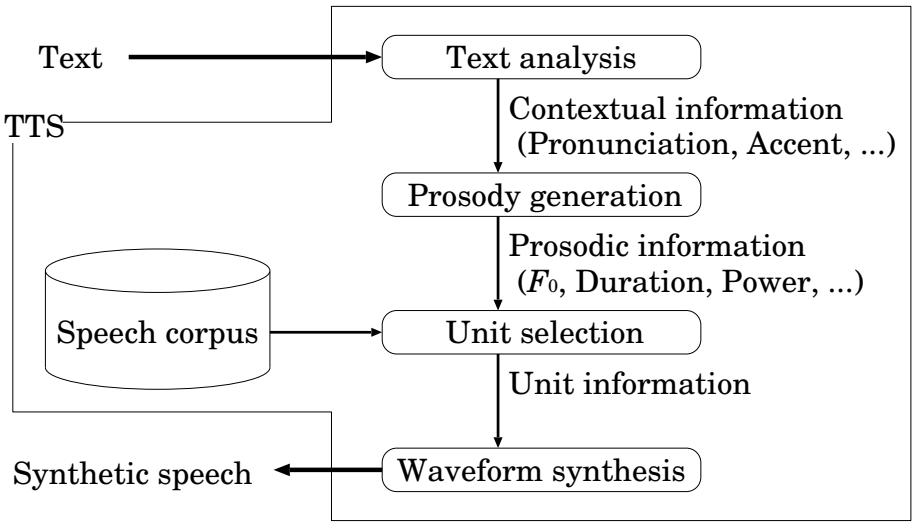

Figure 2.1. Structure of corpus-based TTS.

structure of corpus-based TTS and review various techniques in each module. In **Section 2.3**, conventional voice conversion algorithms and comparison of mapping functions of the algorithms are described. Finally, we summarize this chapter in **Section 2.4**.

2.2 Structure of Corpus-Based TTS

In general, corpus-based TTS is comprised of five modules: text analysis, prosody generation, unit selection, waveform synthesis, and speech corpus. The structure of corpus-based TTS is shown in **Figure 2.1**.

2.2.1 Text analysis

In the text analysis, an input text is converted into contextual information, i.e. pronunciation, accent type, part-of-speech, and so on, by natural language processing [91][96]. The contextual information plays an important role in the quality and intelligibility of synthetic speech because prediction accuracy on this information affects all of the subsequent procedures.

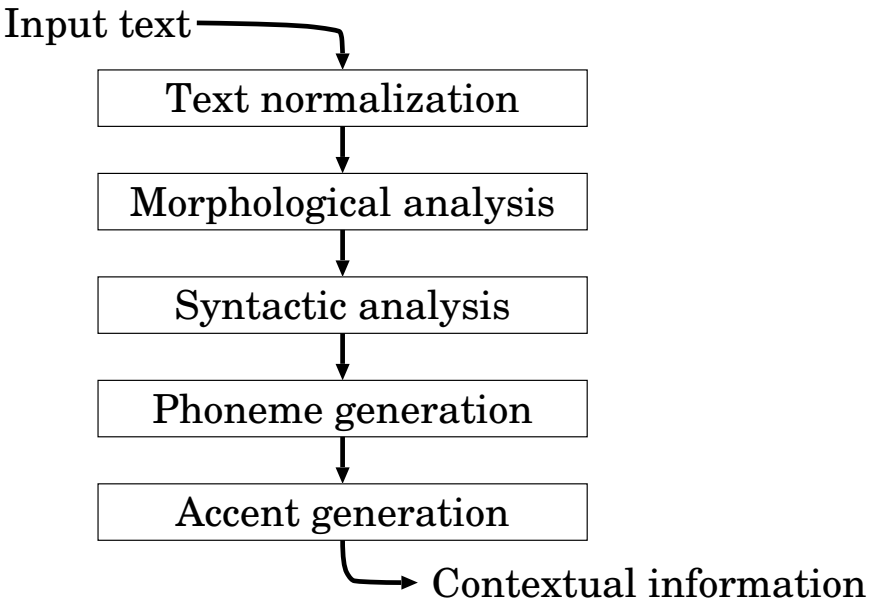

Figure 2.2. Schematic diagram of text analysis.

First, various obstacles, such as unreadable marks like HTML tags and e-mail headings, are removed if the input text includes these obstacles. This processing is called text normalization.

The normalized text is then divided into morphemes, which are minimum units of letter strings having linguistic meaning. These morphemes are tagged with their parts of speech, and a syntactic analysis is performed. Then, the module determines phoneme and prosodic symbols, e.g. accent nucleus, accentual phrases, boundaries of prosodic clauses, and syntactic structure. Reading rules and accentual rules for word concatenation are often applied to the determination of this information [77][88]. Especially in Japanese, the accent information is crucial to achieving high-quality synthetic speech. In some TTS systems, especially English TTS systems, ToBI (Tone and Break Indices) labels [97] or Tilt parameters [108] are predicted [17][38].

A schematic diagram of text analysis is shown in **Figure 2.2**.

2.2.2 Prosody generation

In the prosody generation, prosodic features such as F_0 contour, power contour, and phoneme duration are predicted from the contextual information output from the text analysis. This prosodic information is important for the intelligibility and naturalness of synthetic speech.

Fujisaki's model has been proposed as one of the models that can represent the F_0 contour effectively [39]. This model decomposes the F_0 contour into two components, i.e. a phrase component that decreases gradually toward the end of a sentence and an accent component that increases and decreases rapidly at each accentual phrase. Fujisaki's model is often used to generate the F_0 contour from the contextual information in rule-based TTS, particularly in Japanese TTS [45][61]. Then the rules arranged by experts are applied. In recent years, automatic extraction algorithms of control parameters and rules from a large amount of data with statistical methods have been proposed [41][46].

Many data-driven algorithms for prosody generation have been proposed. In the F_0 contour control model proposed by Kagoshima et al. [56], an F_0 contour of a whole sentence is produced by concatenating segmental F_0 contours, which are generated by modifying vectors that are representative of typical F_0 contours. The representative vectors are selected from an F_0 contour codebook with contextual information. The codebook is designed so that the approximation error between F_0 contours generated by this model and real F_0 contours extracted from a speech corpus is minimized. Isogai et al. proposed using not the representative vectors but natural F_0 contours selected from a speech corpus in order to generate an F_0 contour of a sentence [50]. In this algorithm, if there is an F_0 contour having equal contextual information to the predicted contextual information in the speech corpus, the F_0 contour is selected and used without modification. In all other cases, the F_0 contour that most suits the predicted contextual information is selected and used with modification. Moreover, algorithms for predicting the F_0 contour from the ToBI labels or Tilt parameters have been proposed [13][37].

As a powerful data-driven algorithm, HMM-based (Hidden Markov model) speech synthesis has been proposed by Tokuda et al. [111][112][117]. In this method, the F_0 contour, the mel-cepstrum sequence including the power contour, and phoneme duration are generated directly from HMMs trained by a

decision-tree based on a context clustering technique. The F_0 is modeled by multi-space probability distribution HMMs [111], and the duration is modeled by multi-dimensional Gaussian distribution HMMs in which each dimension shows the duration in each state of the HMM. The mel-cepstrum is modeled by either multi-dimensional Gaussian distribution HMMs or multi-dimensional Gaussian mixture distribution HMMs. Decision-trees are constructed for each feature. The decision-tree for the F_0 and that for the mel-cepstrum are constructed in each state of the HMM. As for the duration, one decision-tree is constructed. All training procedures are performed automatically. In synthesis, the smooth parameter contours, which are static features, are generated from the HMMs by maximizing the likelihood criterion while considering the dynamic features of speech [112].

Some TTS systems do not perform the prosody generation [24]. In these systems, contextual information is used instead of prosody information for the next procedure, unit selection.

In our corpus-based TTS under development, HMM-based speech synthesis is applied to a prosody generation module. A schematic diagram of HMM-based prosody generation is shown in **Figure 2.3**.

2.2.3 Unit selection

In the unit selection, an optimum set of units is selected from a speech corpus by minimizing the degradation of naturalness caused by various factors, e.g. prosodic difference, spectral difference, and a mismatch of phonetic environments [52][89]. Various types of units have been proposed to alleviate such degradation.

Nakajima et al. proposed an automatic procedure called Context Oriented Clustering (COC) [80]. In this technique, the optimum synthesis units are generated or selected from a speech corpus of a single speaker in advance in order to alleviate degradation caused by spectral difference. All segments of a given phoneme in the speech corpus are clustered in advance into equivalence classes according to their preceding and succeeding phoneme contexts. The decision trees that perform the clustering are constructed automatically by minimizing the acoustic differences within the equivalence classes. The centroid segment of each cluster is saved as a synthesis unit. In the speech synthesis phase, the optimum synthesis units are selected from leaf clusters that most suit given pho-

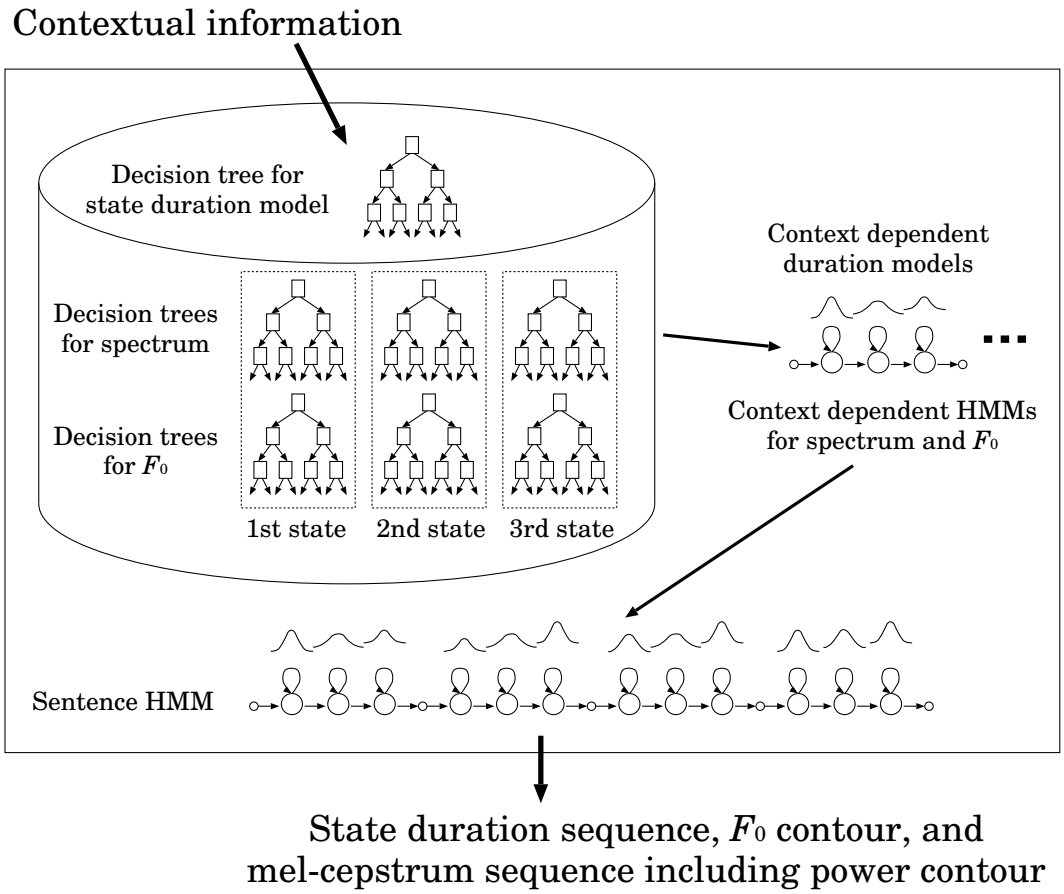

Figure 2.3. Schematic diagram of HMM-based prosody generation.

netic contexts. As the synthesis units, either spectral parameter sequences [40] or waveform segments [51] are utilized.

Kagoshima and Akamine proposed an automatic generation method of synthesis units with Closed-Loop Training (CLT) [5][55]. In this approach, an optimum set of synthesis units is selected or generated from a speech corpus in advance to minimize the degradation caused by synthesis processing such as prosodic modification. A measure capturing this degradation is defined as the difference between a natural speech segment prepared as training data cut off from the speech corpus and a synthesized speech segment with prosodic modification. The selection or generation of the best synthesis unit is performed on the basis of the evaluation

and minimization of the measure in each unit cluster represented by a diphone [87]. Although the number of diphone waveform segments used as synthesis units is very small (only 302 segments), speech with natural and smooth sounding quality can be synthesized.

There are many types of basic synthesis units, e.g. phoneme, diphone, syllable, VCV units [92], and CVC units [93]. The units comprised of more than two phonemes can preserve transitions between phonemes. Therefore, the concatenation between phonemes that often produces perceptual discontinuity can be avoided by utilizing these units. The diphone units have unit boundaries at phoneme centers [32][84][87]. In the VCV units, concatenation points are vowel centers in which formant trajectories are stabler and clearer than those in consonant centers [92]. While, in the CVC units, concatenation is performed at the consonant centers in which waveform power is often smaller than that in the vowel centers [93]. In Japanese, CV (C: Consonant, V: Vowel) units are often used since nearly all Japanese syllables consists of CV or V.

In order to use stored speech data effectively and flexibly, Sagisaka et al. proposed Non-Uniform Units (NUU) [52][89][106]. In this approach, the specific units are not selected or generated from a speech corpus in advance. An optimum set of synthesis units is selected by minimizing a cost capturing the degradation caused by spectral difference, difference in phonetic environment, and concatenation between units in a synthesis procedure. Since it is possible to use all phoneme subsequences as synthesis units, the selected units, i.e. NUU, have variable lengths. The ATR ν-talk speech synthesis system is based on the NUU represented by a spectral parameter sequence [90]. Hirokawa et al. proposed that not only factors related to spectrum and phonetic environment but also prosodic difference are considered in selecting the optimum synthesis units [44]. In this approach, speech is synthesized by concatenating the selected waveform segments and then modifying their prosody. Campbell et al. also proposed utilization of prosodic information in selecting the synthesis units [19][20]. Based on these works, Black et al. proposed a general algorithm for unit selection by using two costs [11][22][47]. One is a target cost, which captures the degradation caused by prosodic difference and difference in phonetic environment, and the other is a concatenation cost, which captures the degradation caused by concatenating

units. In this algorithm, the sum of the two costs is minimized using a dynamic programming search based on phoneme units. By introducing these techniques to ν-talk, CHATR is constructed as a generic speech synthesis system [10][12][21]. Since the number of considered factors increases, a larger-sized speech corpus is utilized than that of ν-talk. If the size of a corpus is large enough and it's possible to select waveform segments satisfying target prosodic features predicted by the prosody generation, it is not necessary to perform prosody modification [23]. Therefore, natural speech without degradation caused by signal processing can be synthesized by concatenating the waveform segments directly. This waveform segment selection has become the main current of corpus-based TTS systems for any language. In recent years, Conkie proposed a waveform segment selection based on half-phoneme units to improve the robustness of the selection [28].

CV* units [60] and multiform units [105] have been proposed as synthesis units by Kawai et al. and Takano et al., respectively. These units can preserve the important transitions for Japanese, i.e. V-V transitions, in order to alleviate the perceptual discontinuity caused by concatenation. The units are stored in a speech corpus as sequences of phonemes with phonetic environments. A stored unit can have multiple waveform segments with different F_0 or phoneme duration. Therefore, optimum waveform segments can be selected while considering both the degradation caused by concatenation and that caused by prosodic modification. In general, the number of concatenations becomes smaller by utilizing longer units. However, the longer units cannot always synthesize natural speech, since the number of candidate units becomes small and the flexibility of prosody synthesis is lost.

Shorter units have also been proposed. Donovan et al. proposed HMM state-based units [33][34]. In this approach, decision-tree state-clustered HMMs are trained automatically with a speech corpus in advance. In order to determine the segment sequence to concatenate, a dynamic programming search is performed over all waveform segments aligned to each leaf of the decision-trees in synthesis. In the HMM-based speech synthesis proposed by Yoshimura et al. [117], the optimum HMM sequence is selected from decision-trees by utilizing phonetic and prosodic context information.

In our corpus-based TTS, the waveform segment selection technique is applied

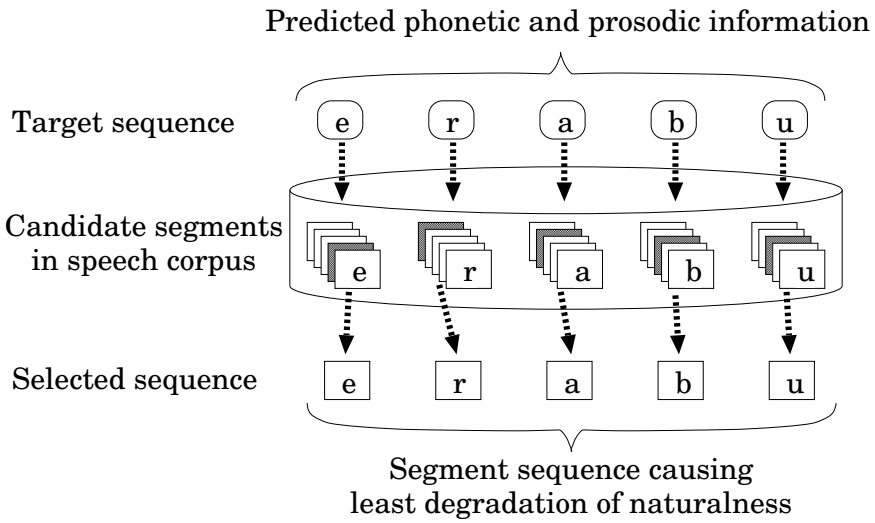

Figure 2.4. Schematic diagram of segment selection.

to a unit selection module. A schematic diagram of the segment selection is shown in **Figure 2.4**.

2.2.4 Waveform synthesis

An output speech waveform is synthesized from the selected units in the last procedure of TTS. In general, two approaches to waveform synthesis have been used. One is waveform concatenation without speech modification, and the other is speech synthesis with speech modification.

In the waveform concatenation, speech is synthesized by concatenating waveform segments selected from a speech corpus using prosodic information to remove need for signal processing [23]. In this case, instead of performing prosody modification, raw waveform segments are used. Therefore, synthetic speech has no degradation caused by signal processing. However, if the prosody of the selected waveform segments is different from the predicted target prosody, degradation is caused by the prosodic difference [44]. In order to alleviate the degradation, it is necessary to prepare a large-sized speech corpus that contains abundant wave-

form segments. Although synthetic speech by waveform concatenation sounds very natural, the naturalness is not always consistent.

In the speech synthesis, signal processing techniques are used to generate a speech waveform with the target prosody. The Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) algorithm is often used for prosody modification [79]. TD-PSOLA does not need any analysis algorithm except for determination of pitch marks throughout the segments. Speech analysis-synthesis methods can also modify the prosody. In the HMM-synthesis method, a mel-cepstral analysis-synthesis technique is performed [117]. Speech is synthesized from a mel-cepstrum sequence generated directly from the selected HMMs and the excitation source by utilizing a Mel Log Spectrum Approximation (MLSA) filter [49]. A vocoder type algorithm such as this can modify speech easily by varying speech parameters, i.e. spectral parameter and source parameter [36]. However, the quality of the synthetic speech is often degraded. As a high-quality vocoder type algorithm, Kawahara et al. proposed the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) analysis-synthesis method [58]. STRAIGHT uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region to remove signal periodicity and designs an excitation source based on phase manipulation. Moreover, STRAIGHT can manipulate such speech parameters as pitch, vocal tract length, and speaking rate while maintaining high reproductive quality. Stylianou proposed the Harmonic plus Noise Model (HNM) as a high-quality speech modification technique [100]. In this model, speech signals are represented as a time-varying harmonic component plus a modulated noise component. Speech synthesis with these modification algorithms is very useful in the case of a small-sized speech corpus. Synthetic speech by this speech synthesis sounds very smooth, and the quality is consistent. However, the naturalness of the synthetic speech is often not as good as that of synthetic speech by waveform concatenation.

In our corpus-based TTS, both the waveform concatenation technique and STRAIGHT synthesis method are applied in the waveform synthesis module. In the waveform concatenation, we control waveform power in each phoneme segment by multiplying the segment by a certain value so that average power in a

phoneme segment selected from a speech corpus becomes equal to average target power in the phoneme. When the segments modified by this power are concatenated, an overlap-add technique is applied in the frame-pair with the highest correlation around a concatenation boundary between the segments. A schematic diagram of the waveform concatenation is shown in **Figure 2.5**. In the other synthesis method based on STRAIGHT, speech waveforms in voiced phonemes are synthesized with STRAIGHT by using a concatenated spectral sequence, a concatenated aperiodic energy sequence, and target prosodic features. In unvoiced phonemes, we use original waveforms modified only by power. A schematic diagram of the speech synthesis with prosody modification by STRAIGHT is shown in **Figure 2.6**.

2.2.5 Speech corpus

A speech corpus directly influences the quality of synthetic speech in corpus-based TTS. In order to realize a consistently high quality of synthetic speech, it is important to prepare a speech corpus containing abundant speech segments with various phonemes, phonetic environments, and prosodies, which should be recorded while maintaining high quality.

Abe et al. developed a Japanese sentence set in which phonetic coverage is controlled [1]. This sentence set is often used not only in the field of speech synthesis but also in speech recognition. Kawai et al. proposed an effective method for designing a sentence set for utterances by taking into account prosodic coverage as well as phonetic coverage [62]. This method selects the optimum sentence set from a large number of sentences by maximizing the measure of coverage. The size of the sentence set, i.e. the number of sentences, is decided in advance. The coverage measure captures two factors, i.e. (1) the distributions of F_0 and phoneme duration predicted by the prosody generation and (2) perceptual degradation of naturalness due to the prosody modification.

In general, the degradation of naturalness caused by a mismatch of phonetic environments and prosodic difference can be alleviated by increasing the size of the speech corpus. However, variation in voice quality is caused by recording the speech of a speaker for a long time in order to construct the large-sized corpus [63]. Concatenation between speech segments with different voice qualities produces

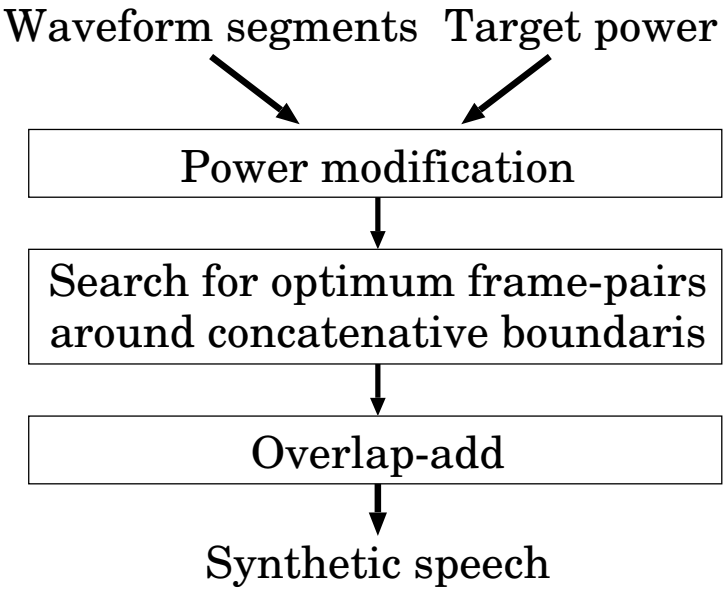

Figure 2.5. Schematic diagram of waveform concatenation.

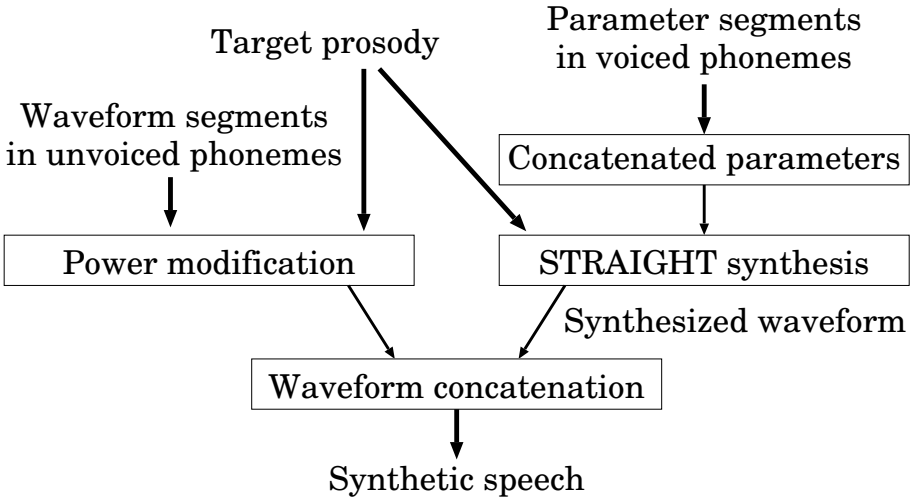

Figure 2.6. Schematic diagram of speech synthesis with prosody modification by STRAIGHT.

audible discontinuity. To deal with this problem, previous studies have proposed using a measure capturing the difference in voice quality to avoid concatenation between such segments [63] and normalization of power spectral densities [94].

In our corpus-based TTS, a large-sized corpus of speech spoken by a Japanese male who is a professional narrator is under construction. The maximum size of the corpus used in this thesis is 32 hours. A sentence set for utterances is extracted from TV news articles, newspaper articles, phrase books for foreign tourists, and so on by taking into account prosodic coverage as well as phonetic coverage.

2.3 Statistical Voice Conversion Algorithm

The code book mapping method has been proposed as a voice conversion method by Abe et al. [2]. In this method, voice conversion is formulated as a mapping problem between two speakers' codebooks, and this idea has been proposed as a method for speaker adaptation by Shikano et al. [95]. This algorithm has been improved by introducing the fuzzy Vector Quantization (VQ) algorithm [81]. Moreover, the fuzzy VQ-based algorithm using difference vectors between mapping code vectors and input code vectors has been proposed in order to represent various spectra beyond the limitation caused by the codebook size [75][76]. These VQ-based algorithms are basic to statistical voice conversion. Abe et al. has also proposed a segment-based approach [4]. In this approach, speech segments of a source speaker selected by HMM are replaced by the corresponding segments of a target speaker. Both static and dynamic characteristics of speaker individuality can be preserved in the segment.

A voice conversion algorithm using Linear Multivariate Regression (LMR) has been proposed by Valbret et al. [114]. In the LMR algorithm, a spectrum of a source speaker is converted with a simple linear transformation for each class. In the algorithm using the Dynamic Frequency Warping (DFW) [114], a frequency warping algorithm is performed in order to convert the spectrum. As a similar conversion method to the DFW algorithm, modification of formant frequencies and spectral intensity has been proposed [78]. Moreover, an algorithm using Neural Networks has been proposed [82].

A voice conversion algorithm by speaker interpolation has been proposed by Iwahashi and Sagisaka [53]. The converted spectrum is synthesized by interpolating the spectra of multiple speakers. In HMM-based speech synthesis, HMM parameters among representative speakers' HMM sets are interpolated [118]. Speech with the voice quality of various speakers can be synthesized from the interpolated HMMs directly [49][112]. Moreover, some speaker adaptation methods, e.g. VFS (Vector Field Smoothing) [83], MAP (Maximum A Posteriori) [69], VFS/MAP [104], and MLLR (Maximum Likelihood Linear Regression) [71], can be applied to HMM-based speech synthesis [74][107]. In the voice conversion by HMM-based speech synthesis, the average voice is often used in place of the source speaker's voice. Although the quality of the synthetic speech is not adequate, this attractive approach has the ability to synthesize various speakers' speech flexibly.

In this thesis, we focus on a voice conversion algorithm based on Gaussian Mixture Model (GMM) proposed by Stylianou et al. [98][99]. In this algorithm, the feature space can be represented continuously by multiple distributions, i.e. a Gaussian mixture model. Utilization of correlation between features of two speakers is indeed characteristic of this algorithm. The VQ-based algorithms mentioned above and the GMM-based algorithm are described in the following section. In these algorithms, only the speech data of the source speaker and target speaker are needed, and both training procedures and conversion procedures are performed automatically.

2.3.1 Conversion algorithm based on Vector Quantization

In the codebook mapping method, the converted spectrum is represented by a mapping codebook, which is calculated as a linear combination of the target speaker's code vectors [2]. A code vector $\boldsymbol{C}_i^{(map)}$ of class i in the mapping codebook for a code vector $\boldsymbol{C}_i^{(x)}$ of class i in the source speaker's codebook is generated as follows:

$$\boldsymbol{C}_i^{(map)} = \sum_{j=1}^{m} w_{i,j} \cdot \boldsymbol{C}_j^{(y)}, \tag{2.1}$$

$$w_{i,j} = \frac{h_{i,j}}{\sum_{k=1}^{m} h_{i,k}}, \tag{2.2}$$

where $\boldsymbol{C}_j^{(y)}$ denotes a code vector of class j in the target speaker's codebook having m code vectors. $h_{i,j}$ denotes a histogram for the frequency of the correspondence between the code vector $\boldsymbol{C}_i^{(x)}$ and the code vector $\boldsymbol{C}_j^{(y)}$ in the training data. In the conversion-synthesis step, the source speaker's speech features are vector-quantized into a code vector sequence with the source speaker's codebook, and then each code vector is mapped into a corresponding code vector in the mapping codebook. Finally, the converted speech is synthesized from the mapped code vector sequence having characteristics of the target speaker. Since representation of the features is limited by the codebook size, i.e. the number of code vectors, quantization errors are caused by vector quantization. Therefore, the converted speech includes unnatural sounds.

The quantization errors are decreased by introducing a fuzzy VQ technique that can represent various kinds of vectors beyond the limitation caused by the codebook size [81]. The vectors are represented not as one code vector but as a linear combination of several code vectors. The fuzzy VQ is defined as follows:

$$\boldsymbol{x}' = \sum_{i=1}^{k} w_i^{(f)} \cdot \boldsymbol{C}_i^{(x)}, \tag{2.3}$$

$$w_i^{(f)} = \frac{(u_i)^f}{\sum_{j=1}^{k} (u_j)^f}, \tag{2.4}$$

where $\boldsymbol{x}'$ denotes a decoded vector of an input vector $\boldsymbol{x}$. k denotes the number of the nearest code vectors to the input vector. u_i denotes a fuzzy membership function of class i and is given by

$$u_i = \frac{1}{\sum_{j=1}^{k} \left(\frac{d_i}{d_j}\right)^{\frac{1}{f-1}}}, \tag{2.5}$$

$$d_i = ||\boldsymbol{x} - \boldsymbol{C}_i^{(x)}||, \tag{2.6}$$

where f denotes fuzziness. The conversion is performed by replacing the source speaker's code vectors with the mapping code vectors. The converted vector

$\boldsymbol{x}^{(map)}$ is given by

$$\boldsymbol{x}^{(map)} = \sum_{i=1}^{k} w_i^{(f)} \cdot \boldsymbol{C}_i^{(map)}. \tag{2.7}$$

Furthermore, it is possible to represent various additional input vectors by introducing difference vectors between the mapping code vectors and the code vectors in the fuzzy VQ-based voice conversion algorithm [75]. In this algorithm, the converted vector $\boldsymbol{x}^{(map)}$ is given by

$$\boldsymbol{x}^{(map)} = \boldsymbol{D}^{(map)} + \boldsymbol{x}, \tag{2.8}$$

$$\boldsymbol{D}^{(map)} = \sum_{i=1}^{k} w_i^{(f)} \cdot \boldsymbol{D}_i, \tag{2.9}$$

where $w_i^{(f)}$ is given by Equation (2.4) and $\boldsymbol{D}_i$ denotes the difference vector as follows:

$$\boldsymbol{D}_i = \boldsymbol{C}_i^{(map)} - \boldsymbol{C}_i^{(x)}. \tag{2.10}$$

2.3.2 Conversion algorithm based on Gaussian Mixture Model

We assume that p-dimensional time-aligned acoustic features $\boldsymbol{x}\{[x_0,\ x_1,\ \ldots,\ x_{p-1}]^{\mathrm{T}}\}$ (source speaker's) and $\boldsymbol{y}\{[y_0,\ y_1,\ \ldots,\ y_{p-1}]^{\mathrm{T}}\}$ (target speaker's) are determined by Dynamic Time Warping (DTW), where T denotes transposition of the vector.

In the GMM algorithm, the probability distribution of acoustic features $\boldsymbol{x}$ can be written as

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{i=1}^{m} \alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \\ \text{subject to} \quad & \sum_{i=1}^{m} \alpha_i = 1, \ \ \alpha_i \geq 0, \end{aligned} \tag{2.11}$$

where α_i denotes a weight of class i, and m denotes the total number of Gaussian mixtures. $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ and is given by

$$N(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{u})\right]. \tag{2.12}$$

The mapping function [98][99] converting acoustic features of the source speaker to those of the target speaker is given by

$$\begin{aligned} F(\boldsymbol{x}) &= E[\boldsymbol{y}|\boldsymbol{x}] \\ &= \sum_{i=1}^{m} h_i(\boldsymbol{x})[\boldsymbol{\mu}_i^{(y)} + \boldsymbol{\Sigma}_i^{(yx)} \left(\boldsymbol{\Sigma}_i^{(xx)}\right)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i^{(x)})], \qquad (2.13) \\ h_i(\boldsymbol{x}) &= \frac{\alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i^{(x)}, \boldsymbol{\Sigma}_i^{(xx)})}{\sum_{j=1}^{m} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j^{(x)}, \boldsymbol{\Sigma}_j^{(xx)})}, \qquad (2.14) \end{aligned}$$

where $\boldsymbol{\mu}_i^{(x)}$ and $\boldsymbol{\mu}_i^{(y)}$ denote the mean vector of class i for the source and that for the target speakers, respectively. $\boldsymbol{\Sigma}_i^{(xx)}$ denotes the covariance matrix of class i for the source speaker. $\boldsymbol{\Sigma}_i^{(yx)}$ denotes the cross-covariance matrix of class i for the source and target speakers.

In order to estimate parameters α_i, $\boldsymbol{\mu}_i^{(x)}$, $\boldsymbol{\mu}_i^{(y)}$, $\boldsymbol{\Sigma}_i^{(xx)}$, and $\boldsymbol{\Sigma}_i^{(yx)}$, the probability distribution of the joint vectors $\boldsymbol{z} = [\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}$ for the source and target speakers is represented by the GMM [57] as follows:

$$\begin{aligned} p(\boldsymbol{z}) &= \sum_{i=1}^{m} \alpha_i N(\boldsymbol{z}; \boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma}_i^{(z)}), \\ \text{subject to} &\quad \sum_{i=1}^{m} \alpha_i = 1, \ \alpha_i \geq 0, \qquad (2.15) \end{aligned}$$

where $\boldsymbol{\Sigma}_i^{(z)}$ denotes the covariance matrix of class i for the joint vectors and $\boldsymbol{\mu}_i^{(z)}$ denotes the mean vector of class i for the joint vectors. These are given by

$$\boldsymbol{\Sigma}_i^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(xx)} & \boldsymbol{\Sigma}_i^{(xy)} \\ \boldsymbol{\Sigma}_i^{(yx)} & \boldsymbol{\Sigma}_i^{(yy)} \end{bmatrix}, \quad \boldsymbol{\mu}_i^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(x)} \\ \boldsymbol{\mu}_i^{(y)} \end{bmatrix}. \qquad (2.16)$$

These parameters are estimated by the EM algorithm [30]. In this thesis, we assume that the covariance matrices, $\boldsymbol{\Sigma}_i^{(xx)}$ and $\boldsymbol{\Sigma}_i^{(yy)}$, and the cross-covariance matrices, $\boldsymbol{\Sigma}_i^{(xy)}$ and $\boldsymbol{\Sigma}_i^{(yx)}$, are diagonal.

2.3.3 Comparison of mapping functions

Figure 2.7 shows various mapping functions: that in the codebook mapping algorithm ("VQ"), that in the fuzzy VQ mapping algorithm ("Fuzzy VQ"), that

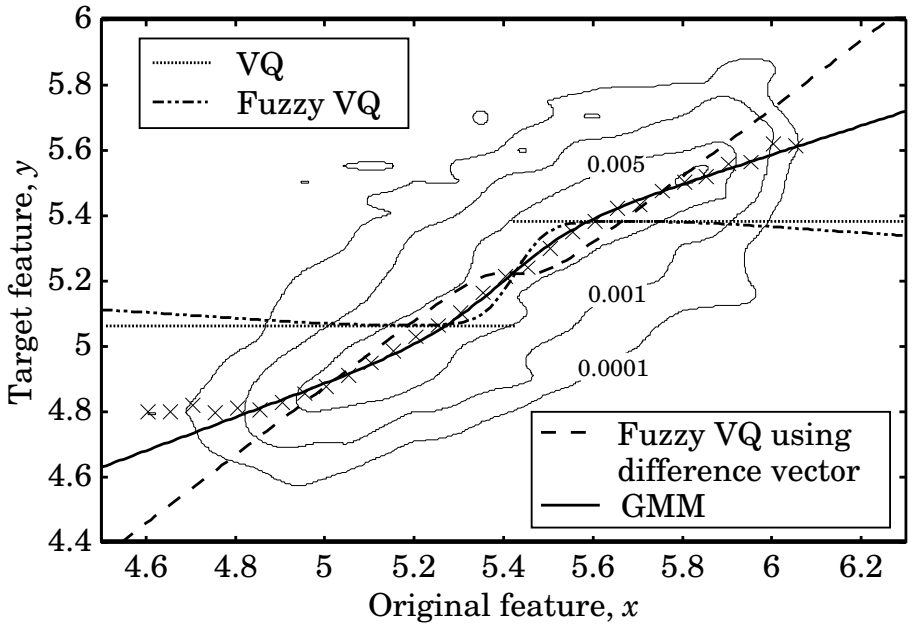

Figure 2.7. Various mapping functions. The contour line denotes frequency distribution of training data in the joint feature space. "x" denotes the conditional expectation $E[y|x]$ calculated in each value of original feature x.

in the fuzzy VQ mapping algorithm using the difference vector ("Fuzzy VQ using difference vector"), and that in the GMM-based algorithm ("GMM"). The number of classes or the number of Gaussian mixtures is set to 2. We also show values of the conditional expectation $E[y|x]$ calculated directly from the training samples.

The mapping function in the codebook mapping algorithm is discontinuous because hard decision clustering is performed. The mapping function becomes continuous by performing fuzzy clustering in the fuzzy VQ algorithm. However, the accuracy of the mapping function is bad because the mapping function seems to be different from the conditional expectation values. Although the mapping function nearly approximates the conditional expectation by introducing the difference vector, the accuracy is not high enough. On the other hand, it is shown that the mapping function in the GMM-based algorithm is close to the conditional expectation because the correlation between the source feature and the target fea-

ture can be utilized. Moreover, Gaussian mixtures can represent the probability distribution of features more accurately than the VQ-based algorithms, since the covariance can be considered in the GMM-based algorithm. Therefore, the mapping function in the GMM-based algorithm is the most reasonable and has the highest conversion-accuracy among the conventional algorithms.

The GMM-based algorithm can convert spectrum more smoothly and synthesize converted speech with higher quality compared with the codebook mapping algorithm [109].

2.4 Summary

This chapter described the basic structure of corpus-based Text-to-Speech (TTS) and reviewed the various techniques in each module. We also introduced some techniques applied to the corpus-based TTS under development.

Moreover, conventional voice conversion algorithms are described. From the results of comparing various mapping functions, the mapping function of the voice conversion algorithm based on the Gaussian Mixture Model (GMM) is the most practical and has the highest conversion-accuracy among the conventional algorithms.

Corpus-based TTS improve the naturalness of synthetic speech dramatically compared with rule-based TTS. However, its naturalness is still inadequate, and flexible synthesis has not yet been achieved.

Chapter 3

A Segment Selection Algorithm for Japanese Speech Synthesis Based on Both Phoneme and Diphone Units

This chapter describes a novel segment selection algorithm for Japanese TTS systems. Since Japanese syllables consist of CV (C: Consonant or consonant cluster, V: Vowel or syllabic nasal /N/) or V, except when a vowel is devoiced, and these correspond to symbols in the Japanese 'Kana' syllabary, CV units are often used in concatenative TTS systems for Japanese. However, speech synthesized with CV units sometimes has discontinuities due to V-V or V-semivowel concatenation. In order to alleviate such discontinuities, longer units, e.g. CV* units, have been proposed. However, since various vowel sequences appear frequently in Japanese, it is not realistic to prepare long units that include all possible vowel sequences. To address this problem, we propose a novel segment selection algorithm that incorporates not only phoneme units but also diphone units. The concatenation in the proposed algorithm is allowed at the vowel center as well as at the phoneme boundary. The advantage of considering both types of units is examined by experiments on concatenation of vowel sequences. Moreover, the results of perceptual evaluation experiments clarify that the proposed algorithm outperforms the conventional algorithms.

3.1 Introduction

In Japanese, a speech corpus can be constructed efficiently by considering CV (C: Consonant or consonant cluster, V: Vowel or syllabic nasal /N/) syllables as synthesis units, since Japanese syllables consist of CV or V except when a vowel is devoiced. CV syllables correspond to symbols in the Japanese 'Kana' syllabary and the number of the syllables is small (about 100). It is also well known that transitions from C to V, or from V to V, are very important in auditory perception [89][105]. Therefore, CV units are often used in concatenative TTS systems for Japanese. On the other hand, other units are often used in TTS systems for English because the number of syllables is enormous (over 10,000) [67]. In recent years, an English TTS system based on CHATR has been adapted for diphone units by AT&T [7]. Furthermore, the NextGen TTS system based on half-phoneme units has been constructed [8][28][102], and this system has proved to be an improvement over the previous system.

In Japanese TTS, speech synthesized with CV units has discontinuities due to V-V or V-semivowel concatenation. In order to alleviate these discontinuities, Kawai et al. extended the CV unit to the CV* unit [60]. Sagisaka proposed non-uniform units to use stored speech data effectively and flexibly [89]. In this algorithm, optimum units are selected from a speech corpus to minimize the total cost calculated as the sum of some sub-costs [52][90][106]. As a result of dynamic programming search based on phoneme units, various sized sequences of phonemes are selected [11][22][47]. However, it is not realistic to construct a corpus that includes all possible vowel sequences, since various vowel sequences appear frequently in Japanese. The frequency of vowel sequences is described in Appendix A. If the coverage of prosody is also to be considered, the corpus becomes enormous [62]. Therefore, the concatenation between V and V is unavoidable.

Formant transitions are more stationary at vowel centers than at vowel boundaries. Therefore, concatenation at the vowel centers tends to reduce audible discontinuities compared with that at the vowel boundaries. VCV units are based on this view [92], which has been supported by our informal listening test. As typical Japanese TTS systems that utilize the concatenation at the vowel centers, TOS Drive TTS (Totally Speaker Driven Text-to-Speech) has been constructed

by TOSHIBA [55] and Final Fluet has been constructed by NTT [105]. The former TTS is based on diphone units. In the latter TTS, diphone units are used if the desirable CV* units are not stored in the corpus. Thus, both TTS systems take into account only the concatenation at the vowel centers in vowel sequences. However, concatenation at the vowel boundaries is not always inferior to that at the vowel centers. Therefore, both types of concatenation should be considered in vowel sequences. In this chapter, we propose a novel segment selection algorithm incorporating not only phoneme units but also diphone units. The proposed algorithm permits the concatenation of synthesis units not only at the phoneme boundaries but also at the vowel centers. The results of evaluation experiments clarify that the proposed algorithm outperforms the conventional algorithms.

The chapter is organized as follows. In **Section 3.2**, cost functions for segment selection are described. In **Section 3.3**, the advantage of performing concatenation at the vowel centers is discussed. In **Section 3.4**, the novel segment selection algorithm is described. In **Section 3.5**, evaluation experiments are described. Finally, we summarize this chapter in **Section 3.6**.

3.2 Cost Function for Segment Selection

The cost function for segment selection is viewed as a mapping, as shown in **Figure 3.1**, of objective features, e.g. acoustic measures and contextual information, into a perceptual measure. A cost is considered the perceptual measure capturing the degradation of naturalness of synthetic speech. In this thesis, only phonetic information is used as contextual information, and the other contextual information is converted into acoustic measures by the prosody generation.

The components of the cost function should be determined based on results of perceptual experiments. A mapping of acoustic measures into a perceptual measure is generally not practical except when the acoustic measures have simple structure, as in the case of F_0 or phoneme duration. Acoustic measures with complex structure, such as spectral features that are accurate enough to capture perceptual characteristics, have not been found so far [31][66][101][115].

On the other hand, a mapping of phonetic information into perceptual measures can be determined from the results of perceptual experiments [64]. There-

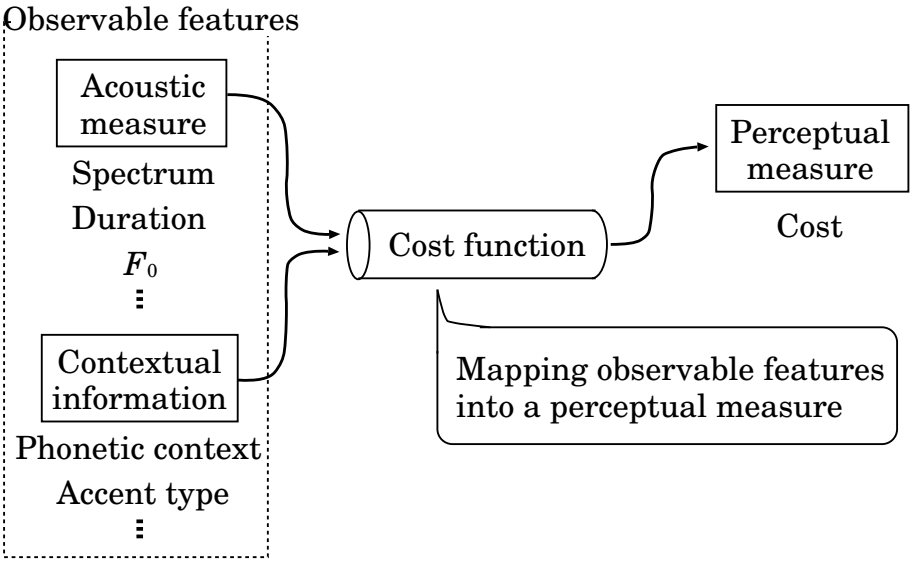

Figure 3.1. Schematic diagram of cost function.

fore, it is possible to capture the perceptual characteristics by utilizing such a mapping. However, acoustic measures that can represent the characteristic of each segment are still necessary, since phonetic information can only evaluate the difference between phonetic categories.

Therefore, we utilize both acoustic measures and perceptual measures determined from the results of perceptual experiments.

3.2.1 Local cost

The local cost shows the degradation of naturalness caused by utilizing an individual candidate segment. The cost function is comprised of five sub-cost functions shown in **Table 3.1**. Each sub-cost reflects either source information or vocal tract information.

The local cost is calculated as the weighted sum of the five sub-costs. The local cost $LC(u_i, t_i)$ at a candidate segment u_i is given by

$$\begin{aligned} LC(u_i, t_i) &= w_{pro} \cdot C_{pro}(u_i, t_i) \\ &\quad + w_{F_0} \cdot C_{F_0}(u_i, u_{i-1}) \end{aligned}$$

Table 3.1. Sub-cost functions

Source information	Prosody (F_0, duration)	C_{pro}
	F_0 discontinuity	C_{F_0}
Vocal tract information	Phonetic environment	C_{env}
	Spectral discontinuity	C_{spec}
	Phonetic appropriateness	C_{app}

$$
\begin{aligned}
& + w_{env} \cdot C_{env}(u_i, u_{i-1}) \\
& + w_{spec} \cdot C_{spec}(u_i, u_{i-1}) \\
& + w_{app} \cdot C_{app}(u_i, t_i), \qquad (3.1)
\end{aligned}
$$

$$
w_{pro} + w_{F_0} \quad + \quad w_{env} + w_{spec} + w_{app} = 1, \qquad (3.2)
$$

where t_i denotes a target phoneme. All sub-costs are normalized so that they have positive values with the same mean. These sub-cost functions are described in the following subsections. w_{pro}, w_{F_0}, w_{env}, w_{spec}, and w_{app} denote the weights for individual sub-costs. In this thesis, these weights are equal, i.e. 0.2. The preceding segment u_{i-1} shows a candidate segment for the $(i-1)$-th target phoneme t_{i-1}. When the candidate segments u_{i-1} and u_i are connected in the corpus, concatenation between the two segments is not performed. **Figure 3.2** shows targets and segments used to calculate each sub-cost in the calculation of the cost of a candidate segment u_i for a target t_i.

3.2.2 Sub-cost on prosody: C_{pro}

This sub-cost captures the degradation of naturalness caused by the difference in prosody (F_0 contour and duration) between a candidate segment and the target.

In order to calculate the difference in the F_0 contour, a phoneme is divided into several parts, and the difference in an averaged log-scaled F_0 is calculated in each part. In each phoneme, the prosodic cost is represented as an average of the

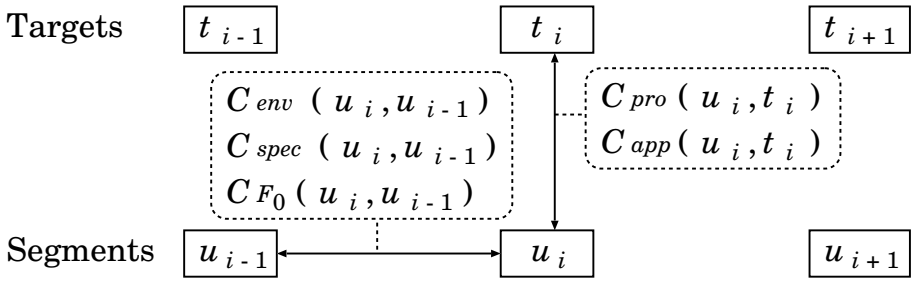

Figure 3.2. Targets and segments used to calculate each sub-cost in calculation of the cost of a candidate segment u_i for a target t_i. t_i and u_i show phonemes considered target and candidate segments, respectively.

costs calculated in these parts. The sub-cost C_{pro} is given by

$$C_{pro}(u_i, t_i) = \frac{1}{M} \sum_{m=1}^{M} P(D_{F_0}(u_i, t_i, m), D_d(u_i, t_i)), \tag{3.3}$$

where $D_{F_0}(u_i, t_i, m)$ denotes the difference in the averaged log-scaled F_0 in the m-th divided part. In the unvoiced phoneme, D_{F_0} is set to 0. D_d denotes the difference in the duration, which is calculated for each phoneme and used in the calculation of the cost in each part. M denotes the number of divisions. P denotes the nonlinear function and is described in Appendix B.

The function P was determined from the results of perceptual experiments on the degradation of naturalness caused by prosody modification, assuming that the output speech was synthesized with prosody modification. When prosody modification is not performed, the function should be determined based on other experiments on the degradation of naturalness caused by using a different prosody from that of the target.

3.2.3 Sub-cost on F_0 discontinuity: C_{F_0}

This sub-cost captures the degradation of naturalness caused by an F_0 discontinuity at a segment boundary. The sub-cost C_{F_0} is given by

$$C_{F_0}(u_i, u_{i-1}) = P(D_{F_0}(u_i, u_{i-1}), 0), \tag{3.4}$$

where D_{F_0} denotes the difference in log-scaled F_0 at the boundary. D_{F_0} is set to 0 at the unvoiced phoneme boundary. In order to normalize a dynamic range of the sub-cost, we utilize the function P in Equation (3.3). When the segments u_{i-1} and u_i are connected in the corpus, the sub-cost becomes 0.

3.2.4 Sub-cost on phonetic environment: C_{env}

This sub-cost captures the degradation of naturalness caused by the mismatch of phonetic environments between a candidate segment and the target. The sub-cost C_{env} is given by

$$C_{env}(u_i, u_{i-1}) = \{S_s(u_{i-1}, E_s(u_{i-1}), u_i) + S_p(u_i, E_p(u_i), u_{i-1})\}/2, \quad (3.5)$$

$$= \{S_s(u_{i-1}, E_s(u_{i-1}), t_i) + S_p(u_i, E_p(u_i), t_{i-1})\}/2, \quad (3.6)$$

where we turn Equation (3.5) into Equation (3.6) by considering that a phoneme for u_i is equal to a phoneme for t_i and a phoneme for u_{i-1} is equal to a phoneme for t_{i-1}. S_s denotes the sub-cost function that captures the degradation of naturalness caused by the mismatch with the succeeding environment, and S_p denotes that caused by the mismatch with the preceding environment. E_s denotes the succeeding phoneme in the corpus, while E_p denotes the preceding phoneme in the corpus. Therefore, $S_s(u_{i-1}, E_s(u_{i-1}), t_i)$ denotes the degradation caused by the mismatch with the succeeding environment in the phoneme for u_{i-1}, i.e. replacing $E_s(u_{i-1})$ with the phoneme for t_i, and $S_p(u_i, E_p(u_i), t_{i-1})$ denotes the degradation caused by the mismatch with the preceding environment in the phoneme u_i, i.e. replacing $E_p(u_i)$ with the phoneme for t_{i-1}. The sub-cost functions S_s and S_p are determined from the results of perceptual experiments described in Appendix C.

Even if a mismatch of phonetic environments does not occur, the sub-cost does not necessarily become 0 because this sub-cost reflects the difficulty of concatenation caused by the uncertainty of segmentation. When the segments u_{i-1} and u_i are connected in the corpus, this sub-cost is set to 0.

3.2.5 Sub-cost on spectral discontinuity: C_{spec}

This sub-cost captures the degradation of naturalness caused by the spectral discontinuity at a segment boundary. This sub-cost is calculated as the weighted

sum of mel-cepstral distortion between frames of a segment and those of the preceding segment around the boundary. The sub-cost C_{spec} is given by

$$C_{spec}(u_i, u_{i-1}) = c_s \cdot \sum_{f=-w/2}^{w/2-1} h(f) MCD(u_i, u_{i-1}, f), \tag{3.7}$$

where h denotes the triangular weighting function of length w. $MCD(u_i, u_{i-1}, f)$ denotes the mel-cepstral distortion between the f-th frame from the concatenation frame ($f = 0$) of the preceding segment u_{i-1} and the f-th frame from the concatenation frame ($f = 0$) of the succeeding segment u_i in the corpus. Concatenation is performed between the -1-th frame of u_{i-1} and the 0-th frame of u_i. c_s is a coefficient to normalize the dynamic range of the sub-cost. The mel-cepstral distortion calculated in each frame-pair is given by

$$\frac{20}{\ln 10} \cdot \sqrt{2 \cdot \sum_{d=1}^{40} (mc_\alpha^{(d)} - mc_\beta^{(d)})^2}, \tag{3.8}$$

where $mc_\alpha^{(d)}$ and $mc_\beta^{(d)}$ show the d-th order mel-cepstral coefficient of a frame α and that of a frame β, respectively. Mel-cepstral coefficients are calculated from the smoothed spectrum analyzed by the STRAIGHT analysis-synthesis method [58][59]. Then, the conversion algorithm proposed by Oppenheim et al. is used to convert cepstrum into mel-cepstrum [85]. When the segments u_{i-1} and u_i are connected in the corpus, this sub-cost becomes 0.

3.2.6 Sub-cost on phonetic appropriateness: C_{app}

This sub-cost denotes the phonetic appropriateness and captures the degradation of naturalness caused by the difference in mean spectra between a candidate segment and the target. The sub-cost C_{app} is given by

$$C_{app}(u_i, t_i) = c_t \cdot MCD(CEN(u_i), CEN(t_i)), \tag{3.9}$$

where CEN denotes a mean cepstrum calculated at the frames around the phoneme center. MCD denotes the mel-cepstral distortion between the mean cepstrum of the segment u_i and that of the target t_i. c_t is a coefficient to normalize the dynamic range of the sub-cost. The mel-cepstral distortion is given

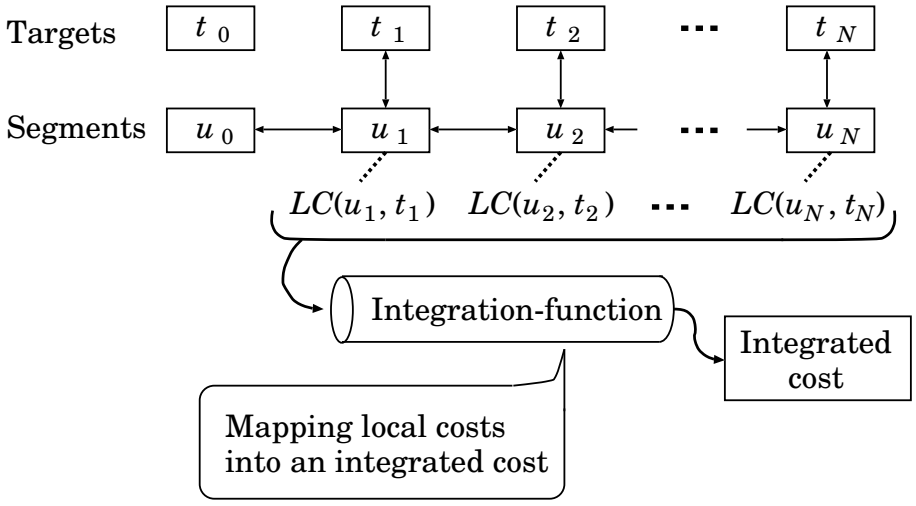

Figure 3.3. Schematic diagram of function to integrate local costs LC.

by Equation (3.8). We utilize the mel-cepstrum sequence output from context-dependent HMMs in the HMM synthesis method [117] in calculating the mean cepstrum of the target $CEN(t_i)$. In this thesis, this sub-cost is set to 0 in the unvoiced phoneme.

3.2.7 Integrated cost

In segment selection, the optimum set of segments is selected from a speech corpus. Therefore, we integrate local costs for individual segments into a cost for a segment sequence as shown in **Figure 3.3**. This cost is defined as an integrated cost. The optimum segment sequence is selected by minimizing the integrated cost.

The average cost AC is often used as the integrated cost [11][22][25][47][102], and it is given by

$$AC = \frac{1}{N} \cdot \sum_{i=1}^{N} LC(u_i, t_i), \tag{3.10}$$

where N denotes the number of targets in the utterance. t_0 (u_0) shows the pause before the utterance and t_N (u_N) shows the pause after the utterance. The sub-

costs C_{pro} and C_{app} are set to 0 in the pause. Minimizing the average cost is equivalent to minimizing the sum of the local costs in the selection.

3.3 Concatenation at Vowel Center

Figure 3.4 compares spectrograms of vowel sequences concatenated at a vowel boundary and at a vowel center. At vowel boundaries, discontinuities can be observed at the concatenation points. This is because it is not easy to find a synthesis unit satisfying continuity requirements for both static and dynamic characteristics of spectral features at once in a restricted-sized speech corpus. At vowel centers, in contrast, finding a synthesis unit involves only static characteristics, because the spectral characteristics are nearly stable. Therefore, it is expected that more synthesis units reducing the spectral discontinuities can be found. As a result, the formant trajectories are continuous at the concatenation points, and their transition characteristics are well preserved.

In order to investigate the instability of spectral characteristics in the vowel, the distances of static and dynamic spectral features were calculated between centroids of individual vowels and all segments of each vowel in a corpus described in the following subsection. As the spectral feature, we used the mel-cepstrum described in **Section 3.2.5**. The results are shown in **Figure 3.5**. It is obvious that the spectral characteristics are stabler around the vowel center than those around the boundary.

From these results, it is assumed that the discontinuities caused by concatenating vowels can be reduced if the vowels are concatenated at their centers. In order to clarify this assumption, we need to investigate the effectiveness of concatenation at vowel centers in segment selection. However, it is difficult to directly show the effectiveness achieved by using the concatenation at vowel centers since various factors are considered in segment selection. Therefore, we first investigate this effectiveness in terms of spectral discontinuity, which is one of the factors considered in segment selection.

In this section, we compare concatenation at vowel boundaries with that at vowel centers by the mel-cepstral distortion. When a vowel sequence is generated by concatenating one vowel segment and another vowel segment, the mel-cepstral

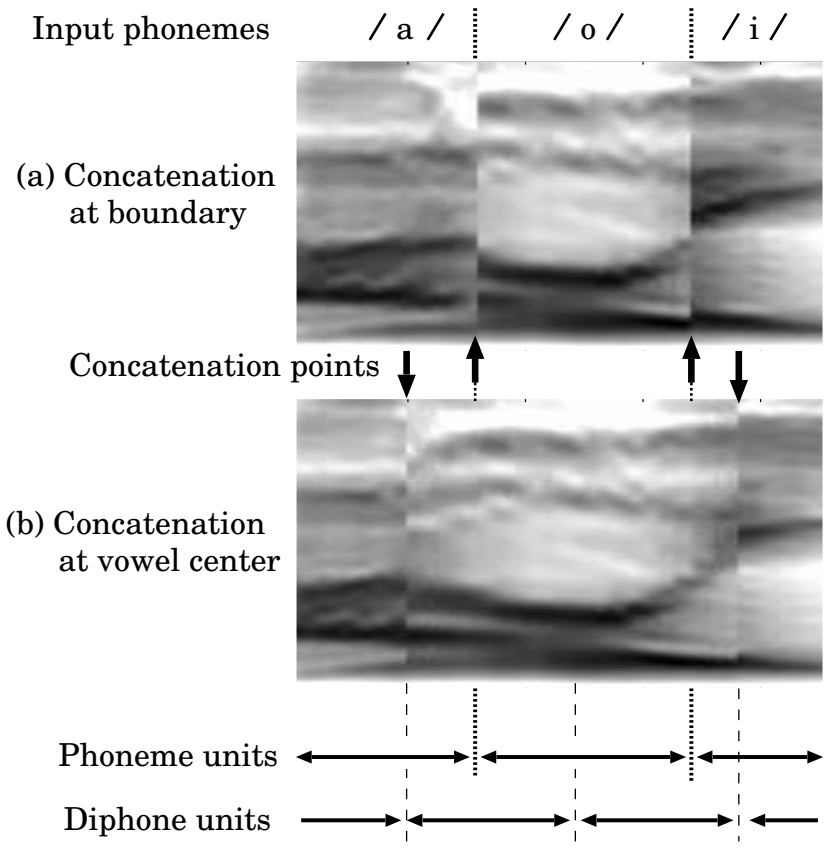

Figure 3.4. Spectrograms of vowel sequences concatenated at (a) a vowel boundary and (b) a vowel center.

distortion caused by the concatenation at vowel boundaries and that at vowel centers are calculated. The vowel center shows a point of a half duration of each vowel segment.

3.3.1 Experimental conditions

The concatenation methods at a vowel boundary and at a vowel center are shown in **Figure 3.6**. We used a speech corpus comprising Japanese utterances of a male speaker, where segmentation was performed by experts and F_0 was revised by hand. The utterances had a duration of about 30 minutes in total (450 sentences).

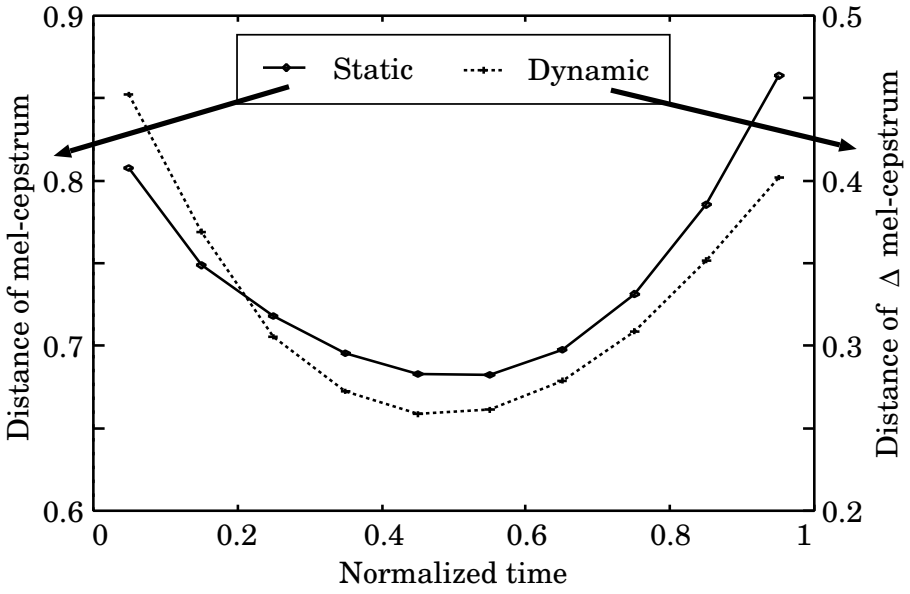

Figure 3.5. Statistical characteristics of static feature and dynamic feature of spectrum in vowels. "Normalized time" shows the time normalized from 0 (preceding phoneme boundary) to 1 (succeeding phoneme boundary) in each vowel segment.

The sampling frequency was 16,000 Hz. The concatenation at vowel boundaries and that at vowel centers were performed by using all of the vowel sequences in the corpus. In each segment-pair, the weighted sum of the mel-cepstral distortion given by Equation (3.7) was calculated, and then the coefficient c_s was set to 1.

3.3.2 Experiment allowing substitution of phonetic environment

In this experiment, substitution of phonetic environments was not prohibited. All segments of "V1" having a vowel as the succeeding phoneme in the corpus were used, i.e. "V*" in **Figure 3.6** shows all vowels.

Figure 3.7 shows frequency distribution of distortion caused by concatenation. Concatenation at vowel centers ("Vowel center") can generally reduce the discontinuity caused by the concatenation compared with that at vowel bound-

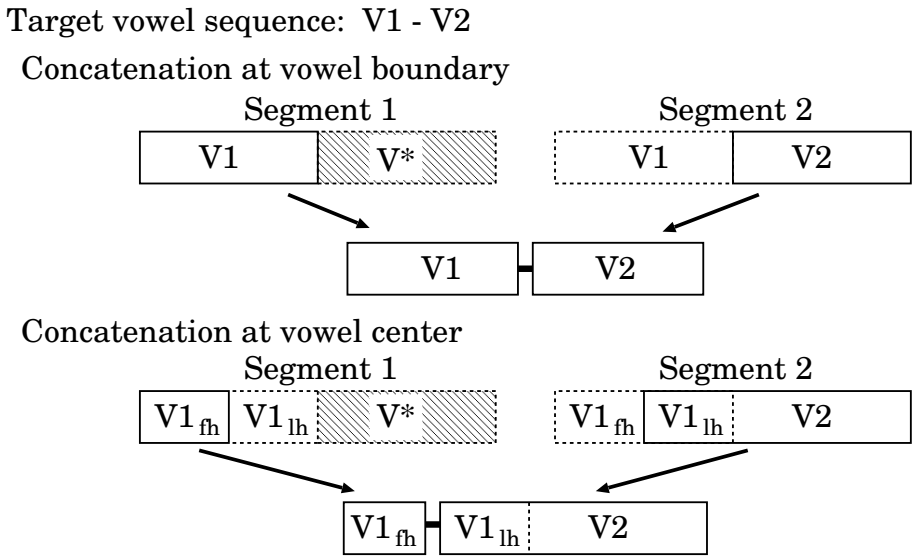

Figure 3.6. Concatenation methods at a vowel boundary and a vowel center. "V*" shows all vowels. "V_{fh}" and "V_{lh}" show the first half-vowel and the last half-vowel, respectively.

aries ("Vowel boundary"). In the segment selection, it is important to select the segments that can reduce not only the spectral discontinuity but also the distortion caused by various factors, e.g. prosodic distance. Therefore, as the frequency distribution shifts to the left side, the number of segments that can reduce distortion increases. From this point of view, it was found that segment selection was better when using concatenation at vowel centers along with substitution of phonetic environments.

3.3.3 Experiment prohibiting substitution of phonetic environment

In this experiment, substitution of phonetic environments was prohibited. All segments of "V1" having "V2" as the succeeding phoneme in the corpus were used, i.e. "V*" = "V2."

Figure 3.8 shows the frequency distribution of distortion caused by the concatenation between vowels that have the same phonetic environment. The distortion caused by the concatenation at vowel centers is almost equal to that at vowel boundaries. Therefore, the performance of concatenation at vowel centers is the same as that at vowel boundaries when substitution of phonetic environments is not performed.

Next, we selected the best type of concatenation in each segment-pair by allowing both vowel center and vowel boundary concatenations. Frequency distribution of distortion in this case is shown in **Figure 3.8**. This approach ("Vowel boundary & Vowel center") can reduce the discontinuity in concatenation compared with concatenation performed only at vowel boundaries or only at vowel centers. This shows that the number of better segments increases by considering both types of concatenation.

These results clarify that better segment selection can be achieved by considering not only the concatenation at vowel centers but also that at vowel boundaries in vowel sequences.

3.4 Segment Selection Algorithm Based on Both Phoneme and Diphone Units

Motivated by the considerations in **Section 3.3**, we propose a novel segment selection algorithm based on both phoneme and diphone units. Here, we also describe the conventional segment selection algorithm based on phoneme unit for comparison with the proposed algorithm.

3.4.1 Conventional algorithm

An input sentence, i.e. a target phoneme sequence, is divided into phonemes. The local costs of candidate segments for each target phoneme are calculated by Equation (3.1). The optimum set of segments is selected from a speech corpus by minimizing the average cost given by Equation (3.10), i.e. the sum of the local costs. As a result, non-uniform units based on phoneme units can be used as synthesis units.

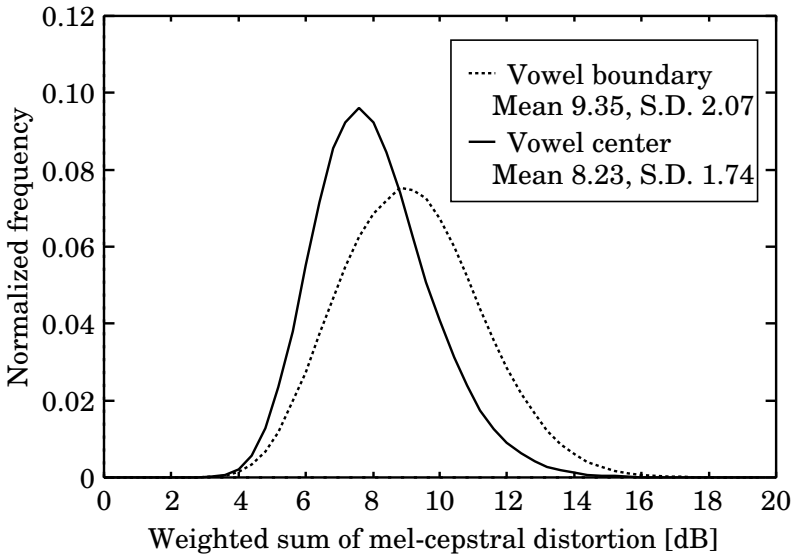

Figure 3.7. Frequency distribution of distortion caused by concatenation between vowels in the case of allowing substitution of phonetic environment. "S.D. " shows standard deviation.

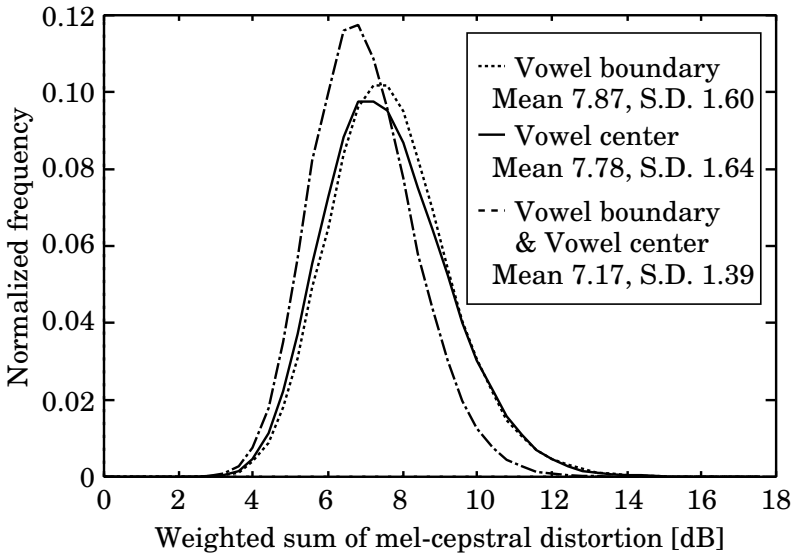

Figure 3.8. Frequency distribution of distortion caused by concatenation between vowels that have the same phonetic environment.

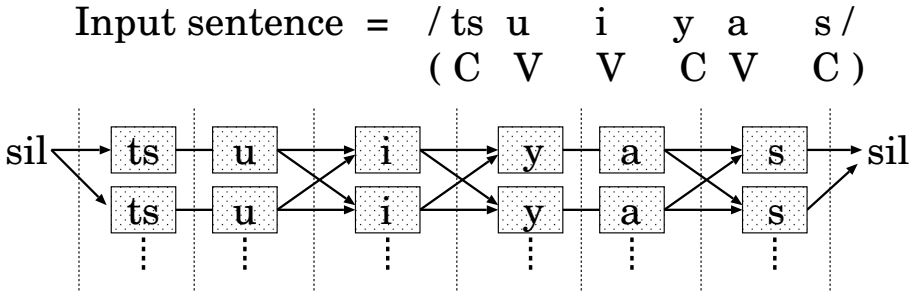

Figure 3.9. Example of segment selection based on phoneme units. The input sentence is "tsuiyas" ("spend" in English). Concatenation at C-V boundaries is prohibited.

Figure 3.9 shows an example of the conventional segment selection based on phoneme units. In this thesis, we do not allow C-V concatenation since the transition from C to V is very important in auditory perception to the intelligibility in Japanese [89][105]. Therefore, the segment sequences comprised of non-uniform units based on the syllable were selected.

3.4.2 Proposed algorithm

When concatenation is allowed at the vowel centers, the half-vowel segments dividing the vowel segments are utilized to take account of diphone units. Each half-vowel segment has a half duration of the original vowel segment.

We assumed that candidate segments u_i^f and u_i^l are the first half-vowel segment of the original vowel segment u_{1i} and the last half-vowel segment of the original vowel segment u_{2i}, respectively. Sub-costs for a target phoneme t_i, which is divided into the first half-phoneme t_i^f and the last half-phoneme t_i^l, are calculated as follows:

- The C_{pro} sub-cost is calculated as the weighted sum of the sub-costs calculated at the half-vowel segments and is given by

$$w_f \cdot C_{pro}(u_i^f, t_i^f) \quad + \quad w_l \cdot C_{pro}(u_i^l, t_i^l), \tag{3.11}$$

$$w_f = \frac{dur(u_i^f)}{dur(u_i^f) + dur(u_i^l)}, \qquad w_l = \frac{dur(u_i^l)}{dur(u_i^f) + dur(u_i^l)}, \tag{3.12}$$

where the weights w_f and w_l are defined according to durations of the segments. $dur(u_i^f)$ and $dur(u_i^l)$ denote duration of the first half-vowel segment u_i^f and that of the last half-vowel segment u_i^l, respectively. In calculating the C_{pro} for the half-vowel segments, each half-vowel segment is divided into $M/2$ parts.

- The C_{F_0} sub-cost is calculated as the sum of the sub-costs at a phoneme boundary and a vowel center and is given by

$$C_{F_0}(u_i^f, u_{i-1}) + C_{F_0}(u_i^l, u_i^f). \tag{3.13}$$

- The C_{env} sub-cost is calculated as the sum of the sub-costs at a phoneme boundary and a vowel center and is given by

$$\begin{aligned} C_{env}(u_i^f, u_{i-1}) &+ C_{env}(u_i^l, u_i^f) \\ &= C_{env}(u_{1i}, u_{i-1}) + \{S_s^d(u_{1i}, E_s(u_{1i}), t_{i+1}) \\ &\quad + S_p^d(u_{2i}, E_p(u_{2i}), t_{i-1})\}/2, \end{aligned} \tag{3.14}$$

  where the phonetic environments of the half-vowel segment are equal to those of the original vowel segment divided into the half-vowel segments. On the other hand, the sub-cost functions, S_s^d and S_p^d, for the concatenation at vowel centers are not equal to the sub-cost functions, S_s and S_p, for the concatenation at phoneme boundaries, which are given by Equation (3.6).

- The C_{spec} sub-cost is calculated as the sum of the sub-costs at a phoneme boundary and a vowel center and is given by

$$C_{spec}(u_i^f, u_{i-1}) + C_{spec}(u_i^l, u_i^f). \tag{3.15}$$

- The C_{app} sub-cost is calculated as the weighted sum of the sub-costs calculated at the original vowel segments, u_{1i} and u_{2i} divided into the half-vowel segments, u_i^l and u_i^f, and is given by

$$w_f \cdot C_{app}(u_{1i}, t_i) \quad + \quad w_l \cdot C_{app}(u_{2i}, t_i), \tag{3.16}$$

  where the weights w_f and w_l are given by Equation (3.12).

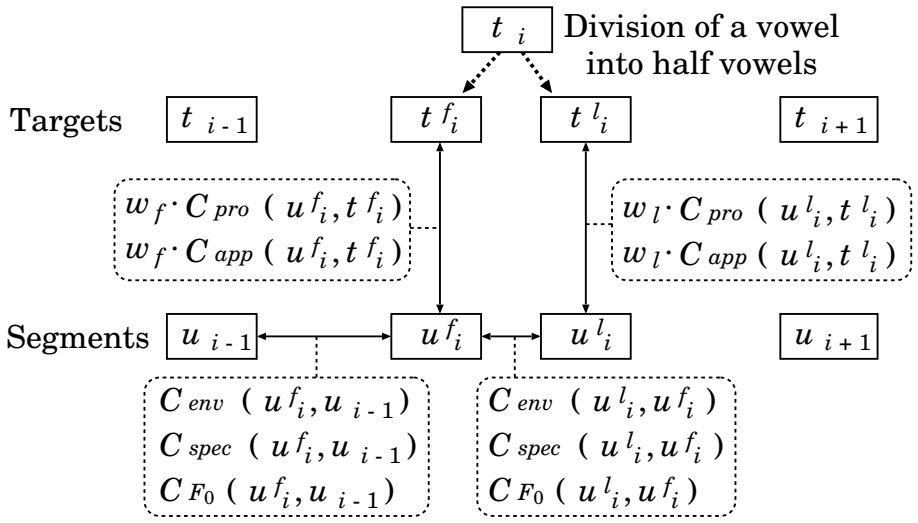

Figure 3.10. Targets and segments used to calculate each sub-cost in calculation of the cost of candidate segments u_i^f, u_i^l for a target t_i.

Figure 3.10 shows targets and segments used to calculate each sub-cost in the calculation of the cost of candidate segments u_i^f, u_i^l for a target t_i. If a diphone unit is used, the sub-costs, $C_{env}(u_i^f, u_{i-1})$, $C_{spec}(u_i^f, u_{i-1})$, and $C_{F_0}(u_i^f, u_{i-1})$, become 0, since u_{i-1} and u_i^f are connected in the corpus. Furthermore, if a phoneme unit is used, the sub-costs, $C_{env}(u_i^l, u_i^f)$, $C_{spec}(u_i^l, u_i^f)$, and $C_{F_0}(u_i^l, u_i^f)$, also become 0, since u_i^f and u_i^l are connected in the corpus.

In the other phonemes where concatenations are not allowed at their centers, the costs are calculated in the same way as the conventional algorithm. The optimum segment sequence is selected from a speech corpus by minimizing the average cost. As a result, non-uniform units based on both phoneme units and diphone units can be used as synthesis units.

Phoneme units and diphone units have their own advantages and disadvantages. The advantage of using phoneme units is the ability to preserve the characteristics of phonemes. Moreover, it might be assumed that slight spectral discontinuities in the transitions in which spectral features change rapidly are hard to perceive. However, it is not easy to find synthesis units that can reduce the spectral discontinuities since static and dynamic characteristics of spectral fea-

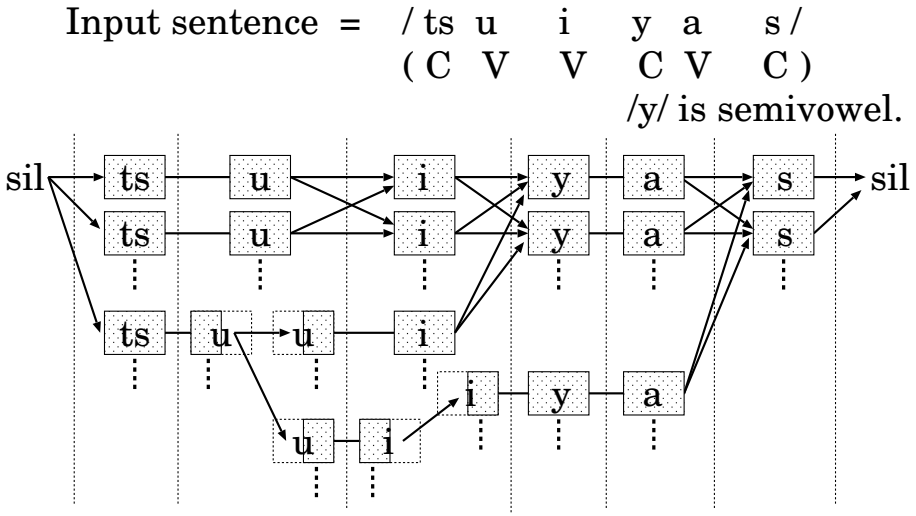

Figure 3.11. Example of segment selection based on phoneme units and diphone units. Concatenation at C-V boundaries and selection of isolated half-vowels are prohibited.

tures should be considered in the transitions. On the other hand, the advantage of using diphone units is the ability to preserve transitions between phonemes and to concatenate at steady parts in phonemes. Therefore, more synthesis units reducing the spectral discontinuities can be found. However, it might be assumed that spectral discontinuities in steady parts are easy to perceive. In the proposed algorithm, the cost is used to determine which of the two units is better.

We allow concatenations at vowel centers not only in transitions from V to V but also in transitions from V to a semivowel or a nasal. In the transitions from V to a semivowel or a nasal, diphone units that start from the center of a vowel in front of consonants are used. In this thesis, the half-vowel segments are not used except for the segments having silences as phonetic environments. Therefore, minimum units preserve either the important transitions between phonemes or the characteristics of Japanese syllables.

An example of the proposed segment selection algorithm is shown in **Figure 3.11**. Diphone units such as /ts-u/, /u-i/, and /i-y/ as well as phoneme units are considered in the segment selection.

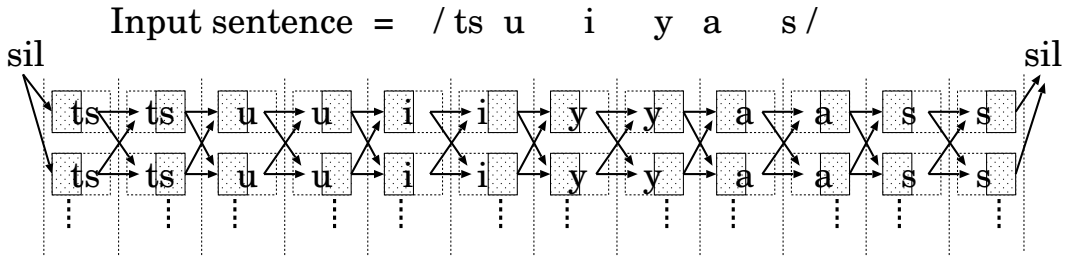

Figure 3.12. Example of segment selection based on half-phoneme units.

3.4.3 Comparison with segment selection based on half-phoneme units

Our TTS under development is mainly for Japanese speech synthesis. Since the number of Japanese syllables is much smaller than that of English syllables, we can construct a speech corpus containing all of the syllables. Therefore, we restrict minimum synthesis units to syllables or diphones in order to preserve important transitions. Namely, we don't use half-phoneme units as used in the AT&T NextGen TTS system [28]. An example of a segment selection algorithm based on half-phoneme units is shown in **Figure 3.12**.

The proposed algorithm can be considered an algorithm based on half-phoneme units adapted for Japanese speech synthesis by restricting some types of concatenation. However, it might be assumed that half-phoneme units would also work well for Japanese speech synthesis. Therefore, we compared the proposed algorithm with the conventional algorithm based on half-phoneme units. In order to make a fair comparison, the weight w_{env} in Equation (3.1) was set to 0 since we have not yet determined S_s and S_p in sub-cost C_{env} for almost all of the half-phoneme units. As a result of a preference test on the naturalness of synthetic speech, the 95% confidence interval of the proposed algorithm's preference score was $66.67 \pm 3.98\%$. This result shows that as the length of units used in segment selection becomes short, the risk of causing more audible discontinuities by excessive concatenations becomes high, although more combinations of units can be considered to reduce the prosodic difference. Therefore, it is important to use a cost that is accurate enough to capture the audible discontinuities, especially

in segment selection based on short units.

3.5 Experimental Evaluation

In order to evaluate the performance of the proposed algorithm, we compared the proposed algorithm with the conventional algorithm, which allows concatenation only at phoneme boundaries. Moreover, we also compared the proposed algorithm with another conventional algorithm, which allows concatenation only at vowel centers in V-V, V-S, and V-N sequences. We call the former comparison Experiment A and the latter comparison Experiment B.

3.5.1 Experimental conditions

We used the speech corpus comprising Japanese utterances of a male speaker, which is described in **Section 3.3.1**.

A preference test was performed with synthesized speech of 10 Japanese sentences. The sentences used in Experiment A were different form those in Experiment B. These sentences were not from the speech corpus used in the segment selection. The speech was synthesized by the proposed segment selection algorithm or the conventional algorithm in each experiment. The natural prosody and the mel-cepstrum sequences extracted from the original utterances were used to investigate the performance of the segment selection algorithms. In Experiment A, all of the synthesized speech were comprised of 366 phonemes, and the number of concatenations in each algorithm is shown in **Table 3.2**. In the other experiment, Experiment B, all of the synthesized speech were comprised of 453 phonemes, and the number of concatenations in each algorithm is shown in **Table 3.3**.

The speech was synthesized with prosody (F_0 contour, duration, and power) modification by using STRAIGHT. Ten listeners participated in the experiment. In each trial, a pair of utterances synthesized with the proposed algorithm and the conventional algorithm was presented in random order, and the listeners were asked to choose either of the two types of synthetic utterances as sounding more natural.

Table 3.2. Number of concatenations in experiment comparing proposed algorithm with segment selection based on phoneme units. "S" and "N" show semivowel and nasal. "Center" shows concatenation at vowel center

Concatenation	V-C	V-V	V-S	V-N	Center
Proposed algorithm	125	6	3	11	25
Conventional algorithm	124	16	3	20	-

Table 3.3. Number of concatenations in experiment comparing proposed algorithm with segment selection allowing only concatenation at vowel center in V-V, V-S, and V-N sequences

Concatenation	V-C	V-V	V-S	V-N	Center
Proposed algorithm	137	10	6	23	22
Conventional algorithm	143	-	-	-	62

3.5.2 Experimental results

Results of Experiment A and those of Experiment B are shown in **Figure 3.13**. The preference score of the proposed algorithm in Experiment A was 69.25%, and that in Experiment B was 64.25%. In both experiments, the preference scores of the proposed algorithm exceeded 50% by a large margin. These results demonstrate that the proposed algorithm can synthesize speech more naturally than the conventional algorithms.

3.6 Summary

In this chapter, we proposed a novel segment selection algorithm for Japanese speech synthesis with both phoneme and diphone units to avoid the degradation of naturalness caused by concatenation at perceptually important transi-

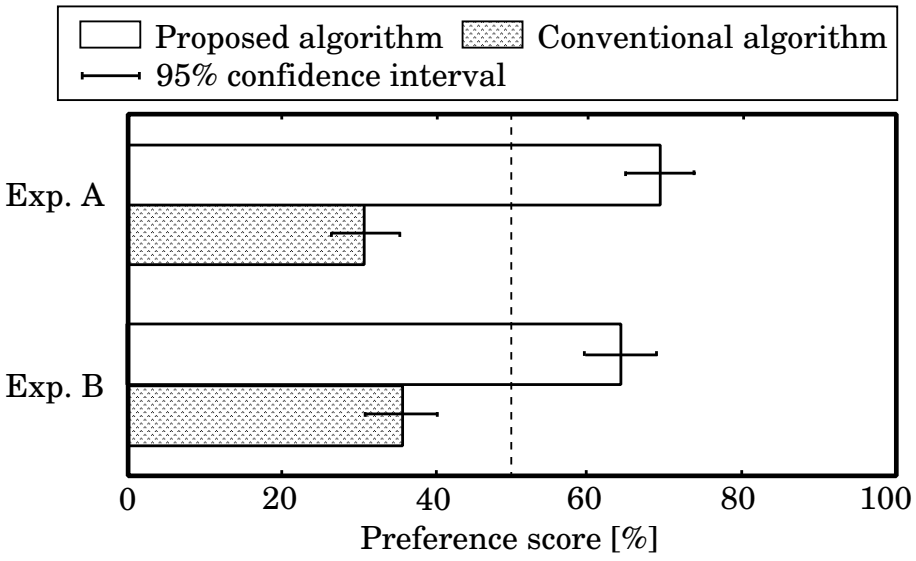

Figure 3.13. Results of comparison with the segment selection based on phoneme units ("Exp. A") and those of comparison with the segment selection allowing only concatenation at vowel center in V-V, V-S, and V-N sequences ("Exp. B").

tions between phonemes. In the proposed algorithm, non-uniform units allowing concatenation not only at phoneme boundaries but also at vowel centers can be selected from a speech corpus. The experimental results of concatenation of vowel sequences clarified that better segments reducing the spectral discontinuities increases by considering both types of concatenation. We also performed perceptual experiments. The results showed that speech synthesized with the proposed algorithm has better naturalness than that of the conventional algorithms.

Although we restrict minimum synthesis units to syllables or diphones, these units are not always the best units. The best unit definition is expected to be determined according to various factors, e.g. corpus size, correspondence of the cost to perceptual characteristics, synthesis methods, and the kinds of languages. Consequently, we need to investigate various units based on this view. Moreover, we need to clarify the effectiveness of searching optimal concatenation frames [27] after segment selection, although we have no acoustic measure that is accurate enough to capture perceptual characteristics.

Chapter 4

An Evaluation of Cost Capturing Both Total and Local Degradation of Naturalness for Segment Selection

In this chapter, we evaluate various costs for a segment sequence in terms of correspondence of the cost to perceptual scores determined from results of perceptual experiments on the naturalness of synthetic speech. The results show that the conventional average cost, which shows the degradation of naturalness over the entire synthetic speech, has better correspondence to the perceptual scores than the maximum cost, which shows the local degradation of naturalness. Furthermore, it is shown that RMS (Root Mean Square) cost, which is affected by both the average cost and the maximum cost, has the best correspondence. We also clarify that the naturalness of synthetic speech can be slightly improved by utilizing the RMS cost. Then, we investigate the effect of using the RMS cost for segment selection. From the results of experiments comparing this approach with segment selection based on the average cost, it is found that (1) in segment selection based on the RMS cost a larger number of concatenations causing slight local degradation are performed to avoid concatenations causing greater local degradation, and (2) the effect of the RMS cost has little dependence on the size of the corpus.

4.1 Introduction

In corpus-based concatenative TTS, speech synthesis based on segment selection has recently become the focus of much work on synthesis [24][102]. In segment selection, the optimum set of segments is selected from a speech corpus by minimizing the integrated cost for a segment sequence, which is described in **Section 3.2.7**. Therefore, it is important to utilize a cost that corresponds to the perceptual characteristics to synthesize speech naturally [70][113]. However, such a cost has not been found so far [31][66][103][115]. To realize TTS with high quality and robustness, it is necessary to improve this cost function [25][86].

In the design process of the cost function, however, it is doubtful whether this correspondence is preserved, since there are some approximations, e.g. utilization of acoustic measures that are not accurate enough to capture perceptual characteristics [64][101], and independence among various factors. Moreover, it might be assumed that local degradation of naturalness would have a great effect on the naturalness of synthetic speech, although the average cost, which shows the degradation of naturalness over the entire synthetic speech, is often used [11][22][47]. Therefore, a direct investigation of the relationship between a perceptual measure and the cost is worthwhile.

Here, we clarify the correspondence of our cost described in **Section 3.2** to the perceptual scores. Then various functions to integrate the costs are evaluated in terms of correspondence to the MOS (Mean Opinion Score) determined from the results of perceptual experiments on the naturalness of synthetic speech. As a result, we show that the RMS (Root Mean Square) cost, which is affected by both the average cost and the maximum cost showing the local degradation of naturalness, has the best correspondence to the perceptual scores. We also clarify that the naturalness of synthetic speech can be slightly improved by using the RMS cost in segment selection.

In order to investigate the effect of considering not only the total degradation of naturalness of synthetic speech but also the local degradation in segment selection, we compare segment selection based on the RMS cost with that based on the average cost. Selected segment sequences are analyzed from various points of view to clarify how the local degradation of naturalness can be alleviated by utilizing the RMS cost. We also clarify the relationship between the effectiveness

of the RMS cost and the size of the corpus.

This chapter is organized as follows. In **Section 4.2**, various integrated costs for the segment selection are described. In **Section 4.3**, we present perceptual evaluations of the costs. In **Section 4.4**, the effectiveness of utilizing the RMS cost in segment selection is discussed. Finally, we summarize this chapter in **Section 4.5**.

4.2 Various Integrated Costs

In the conventional segment selection [11][22][25][47][102], the optimum set of segments is selected from a speech corpus by minimizing the average cost AC given by Equation (3.10). The average cost shows the degradation of naturalness over the entire synthetic utterance. Therefore, a segment with a large cost can be included in the output sequence of segments even if it is optimal in view of the average cost.

It might be assumed that the largest cost in the sequence, i.e. the local degradation of naturalness, would have much effect on the degradation of naturalness in synthetic speech. To investigate this issue, let us define the maximum cost MC as the integrated cost given by

$$MC = \max_i \{WC(u_i, t_i)\}, \quad 1 \leq i \leq N, \tag{4.1}$$

where N denotes the number of targets in the utterance.

In order to evaluate the various integrated costs, we utilize the norm cost, NC_p, given by

$$NC_p = \left[\frac{1}{N} \cdot \sum_{i=1}^{N} \{WC(u_i, t_i)\}^p \right]^{\frac{1}{p}}, \tag{4.2}$$

where p denotes a power coefficient. When p is set to 1, the norm cost is equal to the average cost. When p is set to infinity, the norm cost is equal to the maximum cost. Thus, this norm cost takes into account both the mean value and the maximum value by varying the power coefficient. In this chapter, we find an optimum value of the power coefficient from the results of perceptual experiments.

4.3 Perceptual Evaluation of Cost

4.3.1 Correspondence of cost to perceptual score

We performed an opinion test on the naturalness of the synthetic speech. In order to select a proper set of test stimuli, a large number of utterances were synthesized by varying the corpus size from 0.5 to 32 hours (0.5, 0.7, 1, 1.4, 2, 2.8, 4, 5.7, 8, 11.3, 16, 22.6, 32). Each utterance consisted of a part of a sentence that was divided by a pause. We synthesized 14,926 utterances that were not included in the corpus, from which we selected a set of 140 stimuli so that the set covers a wide field in terms of both average cost and maximum cost. This selection was performed under the restriction that the number of phonemes in an utterance, the duration of an utterance, and the number of concatenations are roughly equal among the selected stimuli. The distribution of the average cost and the maximum cost for all synthetic utterances and selected test stimuli are shown in **Figure 4.1** and **Figure 4.2**, respectively.

Natural prosody and the mel-cepstrum sequences extracted from the original utterances were used as input information for segment selection. In order to alleviate audible discontinuity at the boundary between vowel and voiced phoneme, concatenation at the preceding vowel center is also allowed in the segment selection. In the waveform synthesis, signal processing for prosody modification was not performed except for power control. Therefore, we should use a different subcost on prosody C_{pro} from that described in **Section 3.2.2**, where the function P has been determined in the case of performing prosody modification. In the case of not performing prosody modification, however, we have not determined any function P. Therefore, we approximate C_{pro} by the function P described in Appendix C.

Eight listeners participated in the experiment. They evaluated the naturalness on a scale of seven levels, namely 1 (very bad) to 7 (very good). These levels were determined by each listener, so the scores were distributed widely among the stimuli. The perceptual score, here the MOS, was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

Figure 4.3 shows the correlation coefficient between the norm cost and

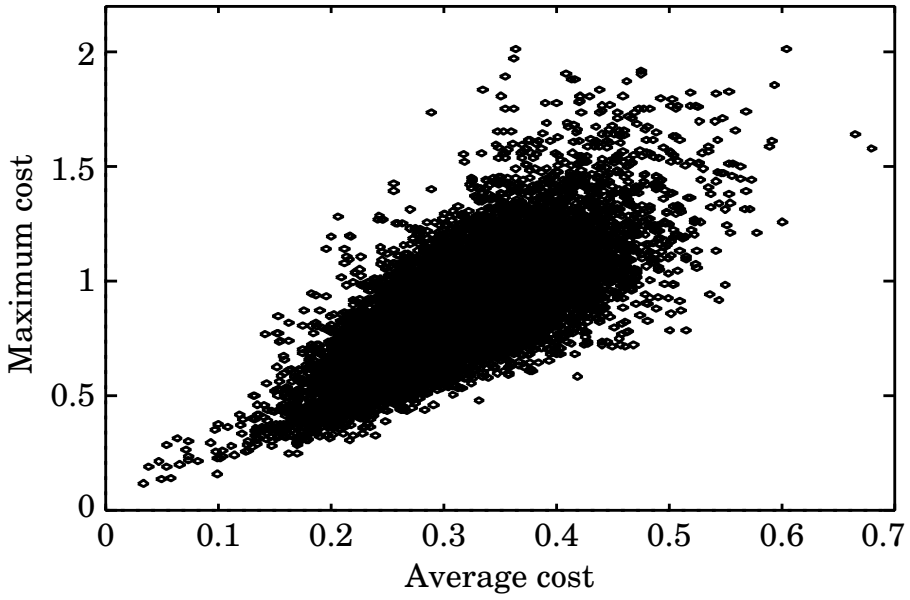

Figure 4.1. Distribution of average cost and maximum cost for all synthetic utterances.

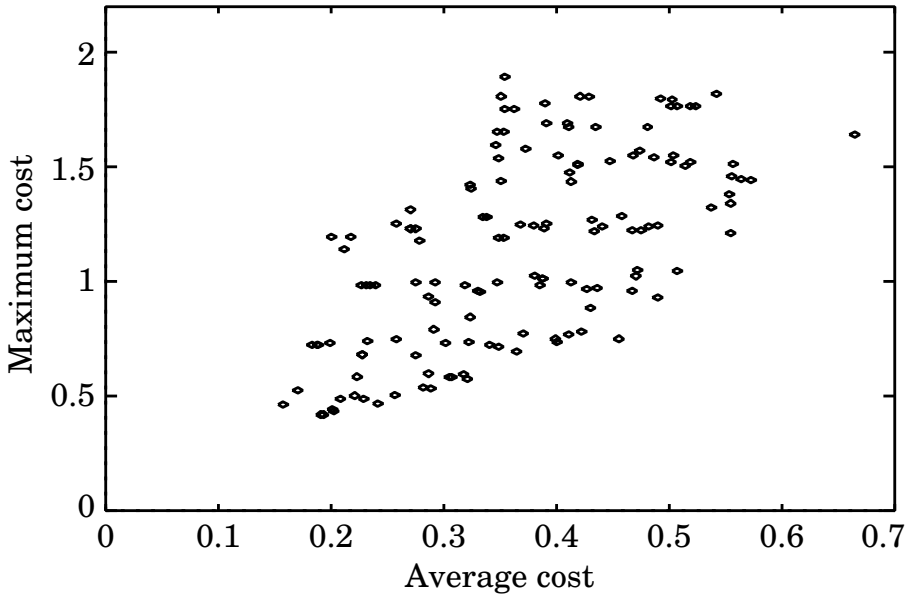

Figure 4.2. Scatter chart of selected test stimuli.

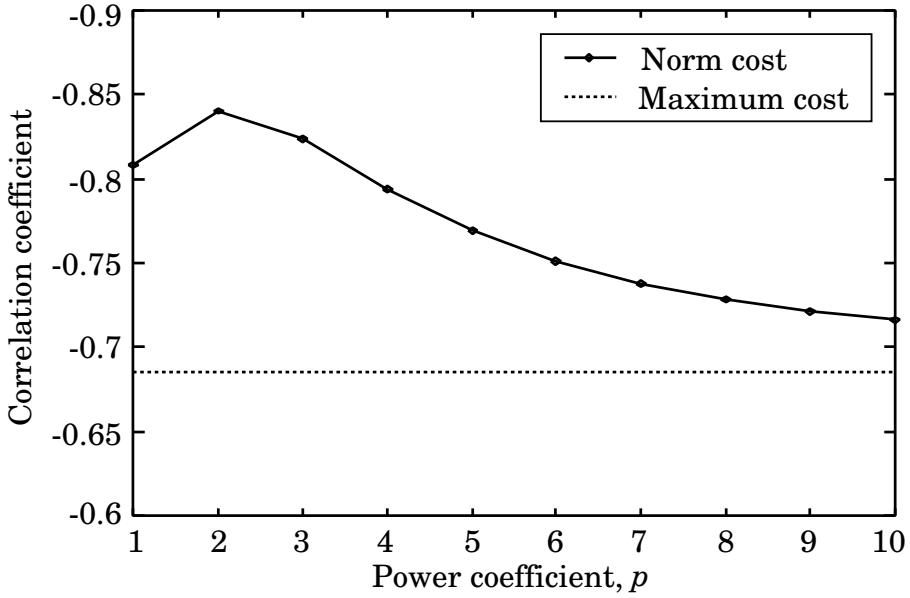

Figure 4.3. Correlation coefficient between norm cost and perceptual score as a function of power coefficient, p.

the perceptual score as a function of the power coefficient. The average cost ($p = 1$) has better correspondence to the perceptual scores (correlation coefficient $= -0.808$) than does the maximum cost (correlation coefficient $= -0.685$). Therefore, the naturalness of synthetic speech is better estimated by the degradation of naturalness over the entire synthetic utterance than by using only the local degradation of naturalness. **Figure 4.4** shows the correlation between the average cost and the perceptual score, and **Figure 4.5** shows the correlation between the maximum cost and the perceptual score.

Moreover, when the power coefficient is set to 2, the norm cost, called the Root Mean Square (RMS) cost, has the best correspondence to the perceptual scores (correlation coefficient $= -0.840$). The absolute value of the correlation coefficient in the case of the RMS cost is statistically larger than those in the cases of the average cost ($t = 2.4696, df = 137, p < 0.05$). Therefore, the naturalness of synthetic speech is better estimated by considering both the degradation of naturalness over the entire synthetic speech and the local degradation of natu-

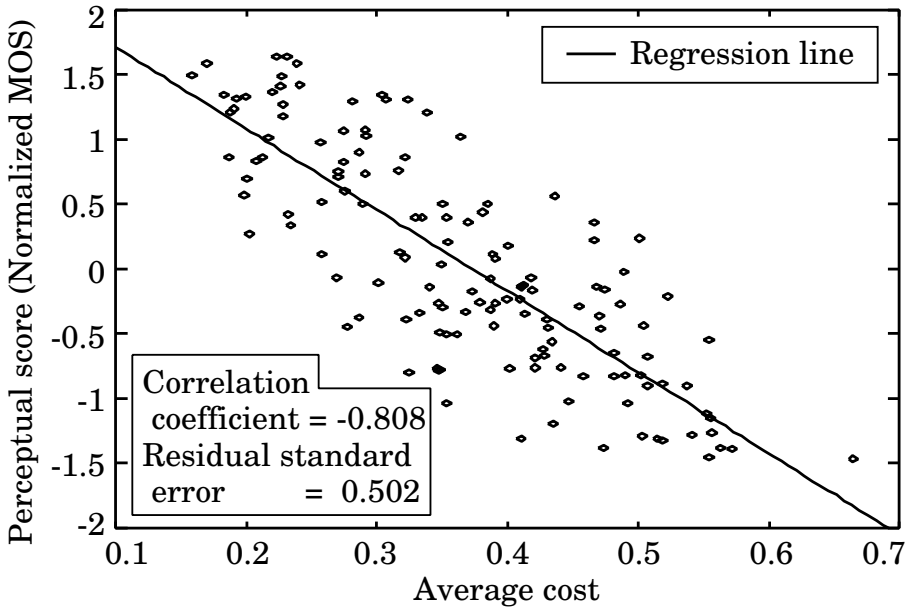

Figure 4.4. Correlation between average cost and perceptual score.

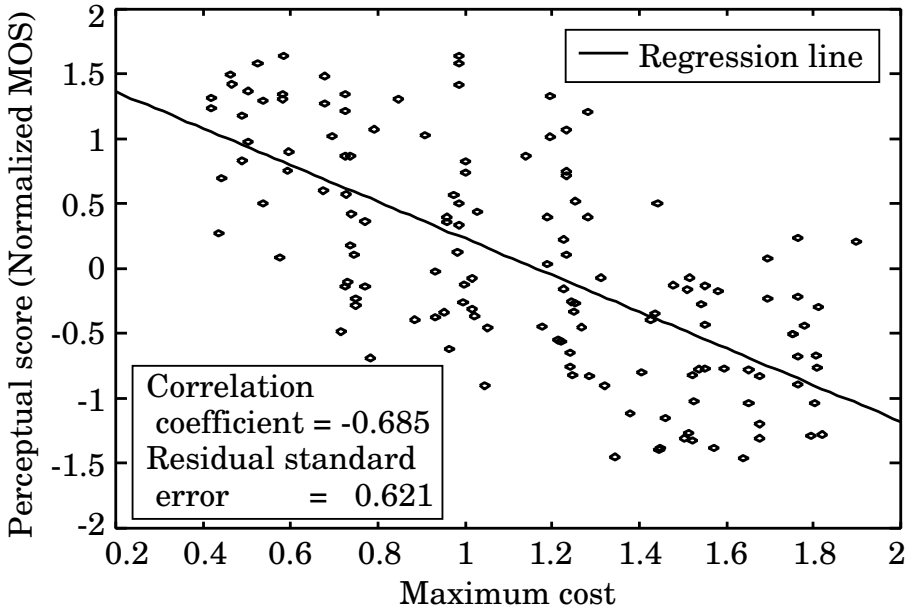

Figure 4.5. Correlation between maximum cost and perceptual score.

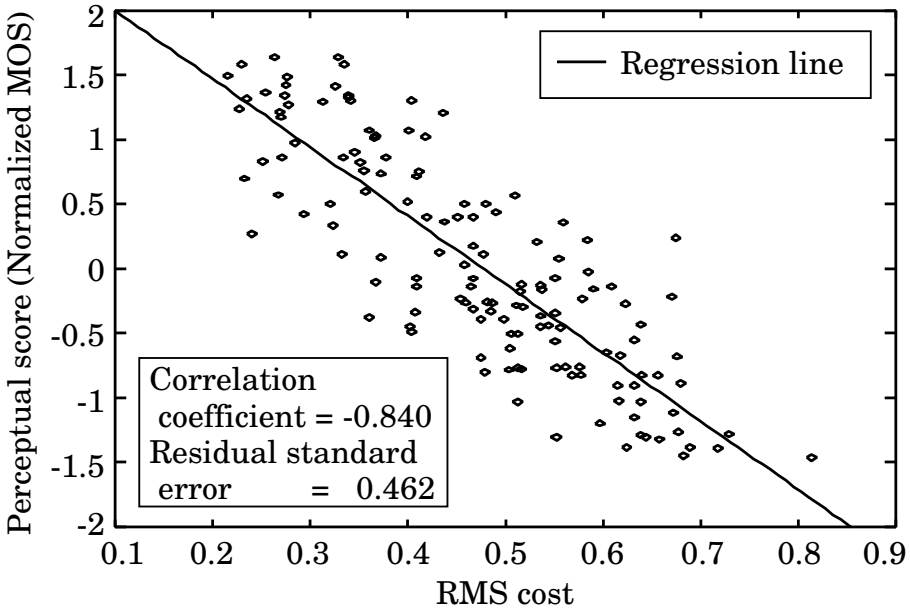

Figure 4.6. Correlation between RMS cost and perceptual score. The RMS cost can be converted into a perceptual score by utilizing the regression line.

ralness, since the RMS cost is affected by both types of degradation. **Figure 4.6** shows the correlation between the RMS cost and the perceptual score. Moreover, **Figure 4.7** shows the correlation coefficient between the RMS cost and the normalized opinion score for each listener. **Figure 4.8** shows the best correlation and the worst correlation in the results of all listeners.

In order to estimate the perceptual scores more accurately, we also performed multiple linear regression analysis by utilizing the norm costs while varying the power coefficient from 1 to 10 and the maximum cost as predictor variables. As a result, the correspondence to the perceptual scores was not improved statistically (multiple correlation coefficient = 0.846) compared with the correspondence of the RMS cost.

Although the RMS cost is the best integrated cost in this experiment, it is expected that the best power coefficient depends on the correspondence of the local cost to perceptual characteristics. It is worth noting that the naturalness of synthetic speech is better estimated by using both the total degradation of

naturalness and the local degradation than by using only the total degradation.

4.3.2 Preference test on naturalness of synthetic speech

Figure 4.9 shows an example of local costs of a segment sequence selected by the conventional average cost and that of another segment sequence selected by the novel RMS cost. Some large local costs surrounded by circles are shown in the case of the average cost. On the other hand, such large local costs are alleviated in the case of the RMS cost. In order to clarify which of the two costs could select the best segment sequence, we performed a preference test on the naturalness of synthetic speech. The corpus size was 32 hours, and utterances used as test stimuli were not included in the corpus. Natural prosody and the mel-cepstrum sequences extracted from the original utterances were used as input information for segment selection. Signal processing for prosody modification was not performed except for power control.

The naturalness of synthetic speech was expected to be nearly equal between segment sequences having similar costs. Therefore, we used pairs of segment sequences that had greater cost differences in the test. Each pair is comprised of the segment sequence selected by the RMS cost and that by the average cost. In order to fairly compare the performances of these two costs, we selected the pairs with the larger differences in the average cost as well as those with the larger differences in the RMS cost. A scatter chart of the test stimuli is shown in **Figure 4.10**. The difference in the RMS cost and that in the average cost were calculated as follows:

$$RMSC_{ACsel} - RMSC_{RMSCsel}, \tag{4.3}$$

$$AC_{ACsel} - AC_{RMSCsel}, \tag{4.4}$$

where $RMSC_{ACsel}$ and AC_{ACsel} denote the RMS cost and the average cost of the segment sequence selected by minimizing the average cost, respectively. $RMSC_{RMSCsel}$ and $AC_{RMSCsel}$ denote the RMS cost and the average cost of the segment sequence selected by minimizing the RMS cost, respectively. "Sub-set A" includes stimulus pairs with larger differences in the RMS cost. On the other hand, "sub-set B" includes stimulus pairs with larger differences in average cost.

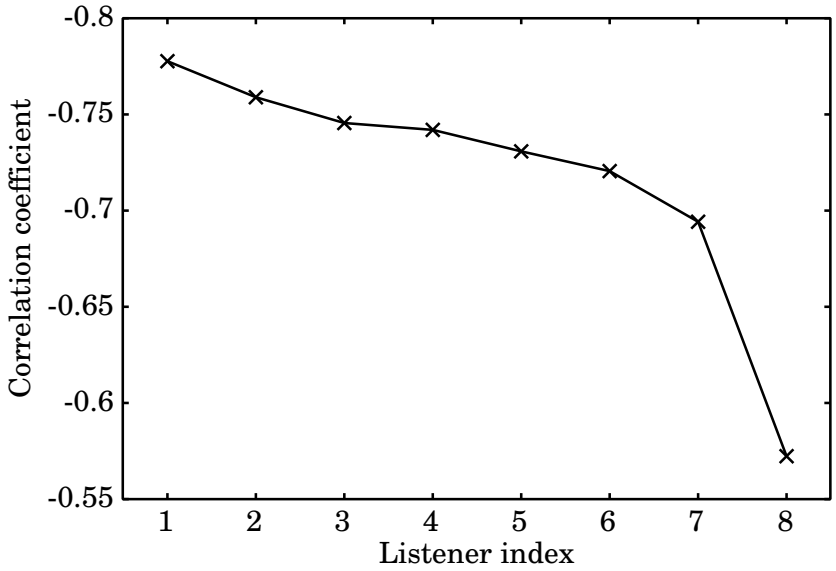

Figure 4.7. Correlation coefficient between RMS cost and normalized opinion score for each listener.

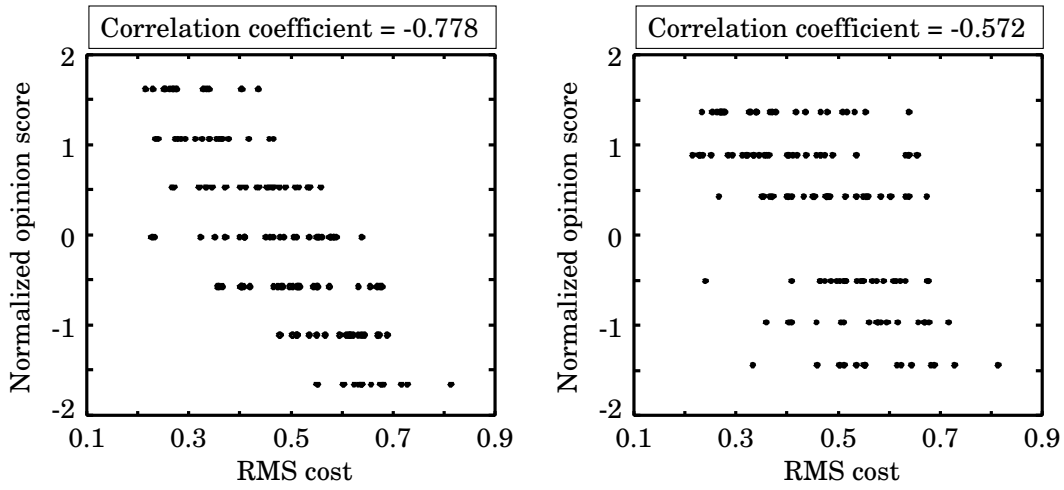

Figure 4.8. Best correlation between RMS cost and normalized opinion score (left figure) and worst correlation between RMS cost and normalized opinion score (right figure) in results of all listeners.

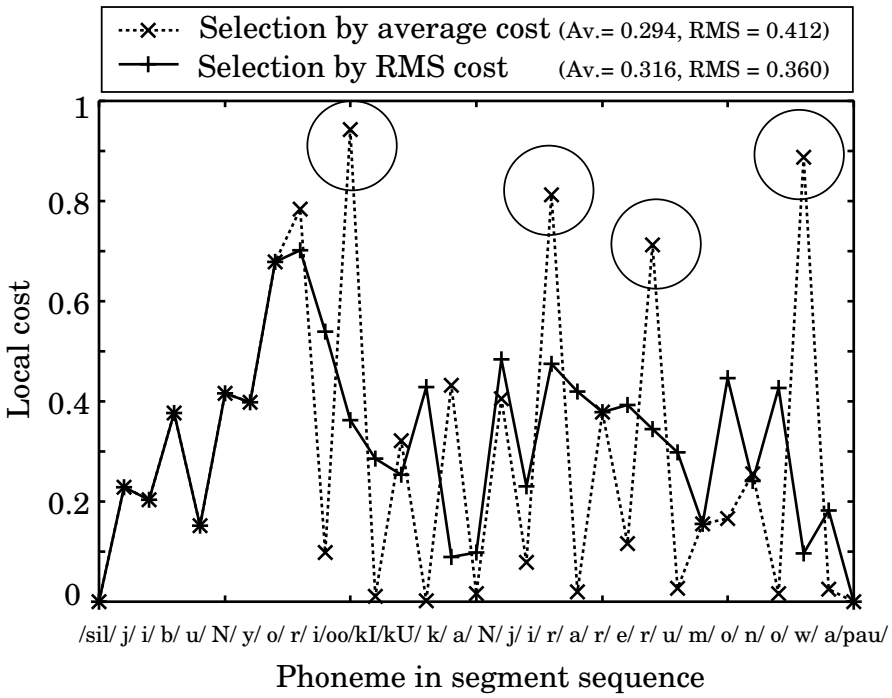

Figure 4.9. Examples of local costs of segment sequences selected by the average costs and by the RMS cost. "Av." and "RMS" show the average and the root mean square of local costs, respectively.

There were 20 stimulus pairs in each sub-set, and the total number of stimulus pairs was 35, since 5 pairs were included in both sub-sets.

Eight Japanese listeners participated in the experiment. In each trial, synthetic speech by the segment selection based on the average cost and that by the segment selection based on the RMS cost were presented in random order, and listeners were asked to choose either of the two types of synthetic speech as sounding more natural.

The results in **Figure 4.11** show that the segment selection based on the RMS cost can synthesize speech more naturally than that based on the average cost in all cases: utilizing all stimuli, stimuli in sub-set A only, and stimuli in sub-set B only. However, this improvement is only slight.

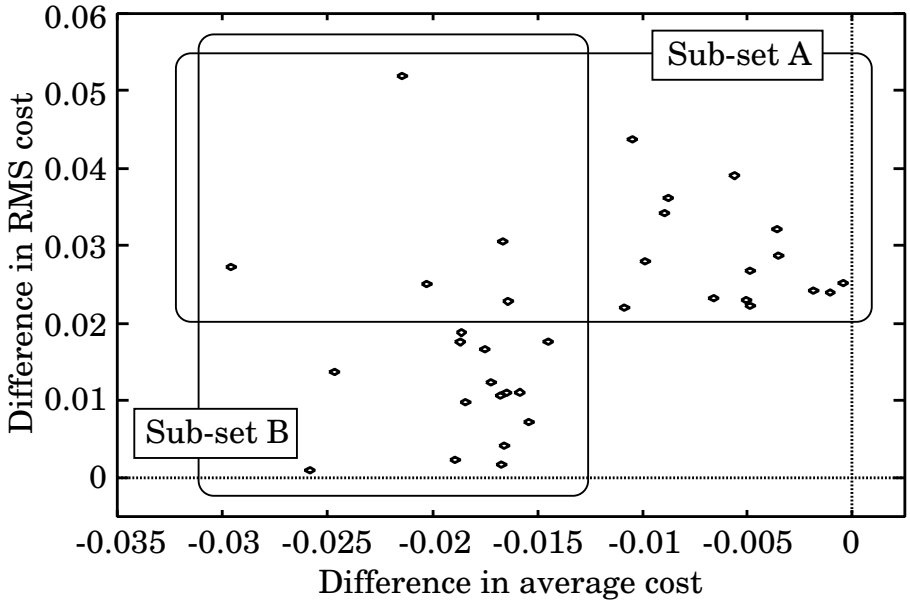

Figure 4.10. Scatter chart of selected test stimuli. Each dot denotes a stimulus pair.

4.3.3 Correspondence of RMS cost to perceptual score in lower range of RMS cost

We clarified the correspondence of the RMS cost to the perceptual scores when the size of the corpus was varied in **Section 4.3.1**. However, our TTS system utilizes a large-sized corpus that includes many segments with high coverage on both phonetic environment and prosody to synthesize speech more naturally and consistently. In the case of a large-sized corpus, the RMS costs are expected to be distributed not in a wide range but in a lower range, since segments causing only a slight degradation of naturalness can usually be selected. Thus, it is worthwhile to investigate the correspondence to the perceptual scores in a range of lower RMS costs.

We performed an opinion test on the naturalness of the synthetic speech to clarify the correspondence of the RMS cost to the perceptual scores in a lower range. Test stimuli were included in the region covered in utilizing the 32-hour

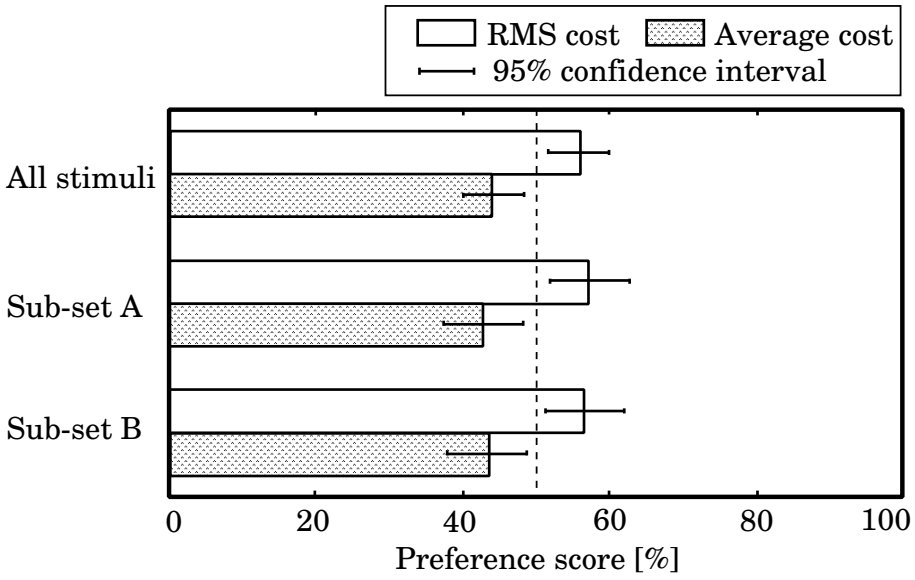

Figure 4.11. Preference score.

corpus, in which the RMS costs were less than 0.4. They were selected from a large number of utterances synthesized by varying the corpus size. This selection was performed under the restriction that the number of phonemes in an utterance, the duration of an utterance, and the number of concatenations were roughly equal among the selected stimuli. The number of selected stimuli was 160. Eight Japanese listeners participated in the experiment. They evaluated the naturalness on a scale of seven levels. These levels were determined by each listener, so the scores were distributed widely among the stimuli. The perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The correspondence of the RMS cost to the perceptual scores is shown in **Figure 4.12**. The correspondence is much worse (correlation coefficient = -0.400) than that in the case of utilizing stimuli that cover a wide range of the cost (correlation coefficient = -0.840). Therefore, it is obvious that the correspondence of the RMS cost is inadequate and that we should improve the cost function.

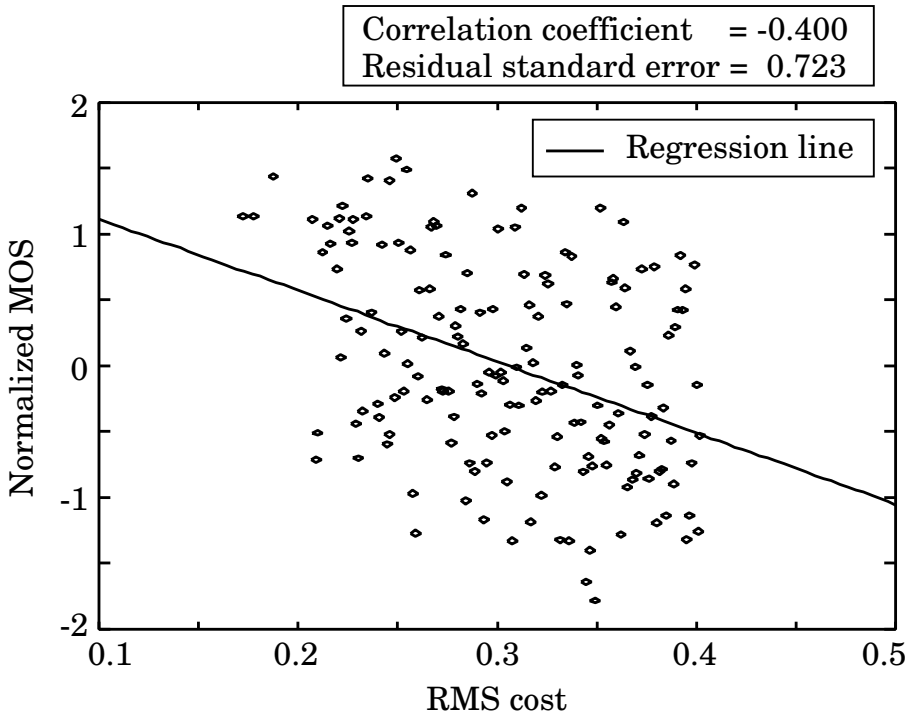

Figure 4.12. Correlation between RMS cost and perceptual score in lower range of RMS cost.

4.4 Segment Selection Considering Both Total Degradation of Naturalness and Local Degradation

In the conventional segment selection based on the average cost, the optimum segment sequence is selected while taking into account only the total degradation. In order to consider not only the total degradation but also the local degradation, we incorporated the RMS cost in segment selection. In this selection, the RMS cost $RMSC$ is minimized and given by

$$RMSC = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \{LC_i(u_i, t_i)\}^2}. \tag{4.5}$$

Actually, only the sum of the square local costs is calculated for the selection.

In this section, in order to represent the costs more easily, we divide the five sub-costs described in **Section 3.2** into two commonly used costs, i.e. a target cost C_t and a concatenation cost C_c [11][22][47][102]. These costs are given by

$$\begin{aligned} C_t(u_i, t_i) &= w_{pro}/w_t \cdot C_{pro}(u_i, t_i) \\ &\quad + w_{app}/w_t \cdot C_{app}(u_i, t_i), \qquad (4.6) \\ C_c(u_i, u_{i-1}) &= w_{env}/w_c \cdot C_{env}(u_i, u_{i-1}) \\ &\quad + w_{spec}/w_c \cdot C_{spec}(u_i, u_{i-1}) \\ &\quad + w_{F_0}/w_c \cdot C_{F_0}(u_i, u_{i-1}), \qquad (4.7) \\ w_t &= w_{pro} + w_{app}, \qquad (4.8) \\ w_c &= w_{F_0} + w_{env} + w_{spec}, \qquad (4.9) \end{aligned}$$

and then the local cost is written as

$$\begin{aligned} LC(u_i, t_i) &= w_t \cdot C_t(u_i, t_i) + w_c \cdot C_c(u_i, u_{i-1}), \qquad (4.10) \\ w_t &+ w_c = 1. \qquad (4.11) \end{aligned}$$

We compared the segment selection based on the RMS cost with that based on the average cost. We utilized 1,131 utterances as an evaluation set. These utterances were not included in the corpus used for segment selection. In the segment selection, concatenations at certain phoneme centers, i.e. preceding vowel centers of voiced phonemes and unvoiced fricative centers, were also allowed in order to alleviate audible discontinuity.

4.4.1 Effect of RMS cost on various costs

We investigated the effect of the RMS cost on the local cost. **Figure 4.13** shows the local costs as a function of corpus size. In the segment selection based on the RMS cost ("RMS cost"), the standard deviation of the local cost is smaller than that of the segment selection based on the average cost ("Average cost"), although the mean of this cost is slightly worse. This is a consequence of the large penalty imposed on a segment with a large local cost in the case of the RMS cost.

In order to clarify what causes the decrease in the standard deviation, we investigated the effects of the RMS cost on both the target cost and the concatenation cost. The target cost is shown in **Figure 4.14** as a function of corpus size.

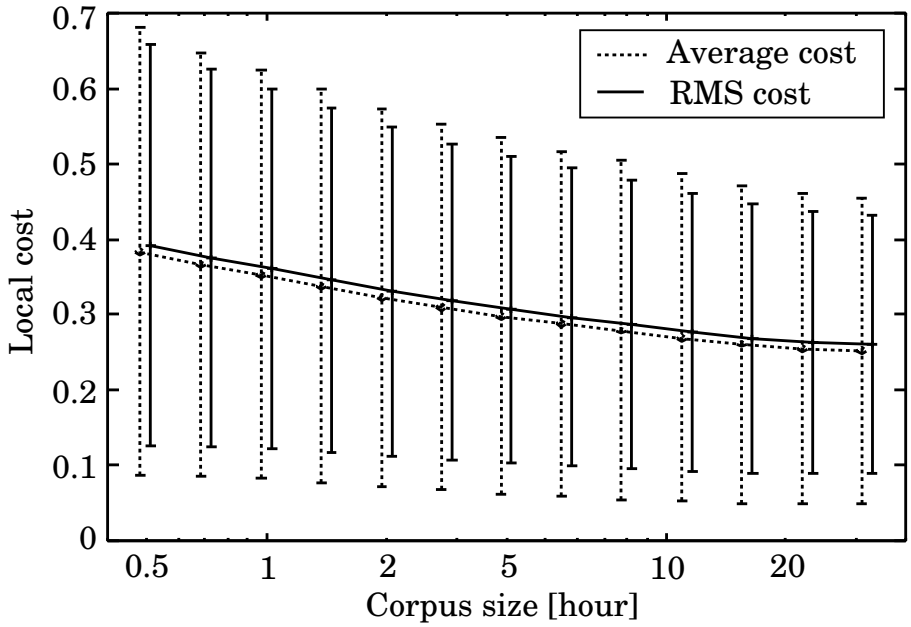

Figure 4.13. Local costs as a function of corpus size. Mean and standard deviation are shown.

The mean of the target costs is degraded, and the standard deviation increases slightly by utilizing the RMS cost. **Figure 4.15** shows the concatenation cost as a function of corpus size. Although the means of concatenation costs are equal between the average cost and the RMS cost, the standard deviation becomes smaller by utilizing the RMS cost. The increase in the standard deviation of the target cost is much smaller than the decrease in the standard deviation of the concatenation cost. These results show that the effectiveness of decreasing the standard deviation of the local cost is dependent on the concatenation cost. On the other hand, the mean of the local cost is slightly worse as a consequence of the degradation of the target cost.

However, it might be assumed that these results would be influenced by the weights for sub-costs rather than by the local degradation, since we utilized a weight set in which the weight for the target cost was smaller than that for the concatenation cost, i.e. $w_t = 0.4$, $w_c = 0.6$ in Equation (4.10). Therefore, we tried to analyze the effects of utilizing other weight sets. The same results

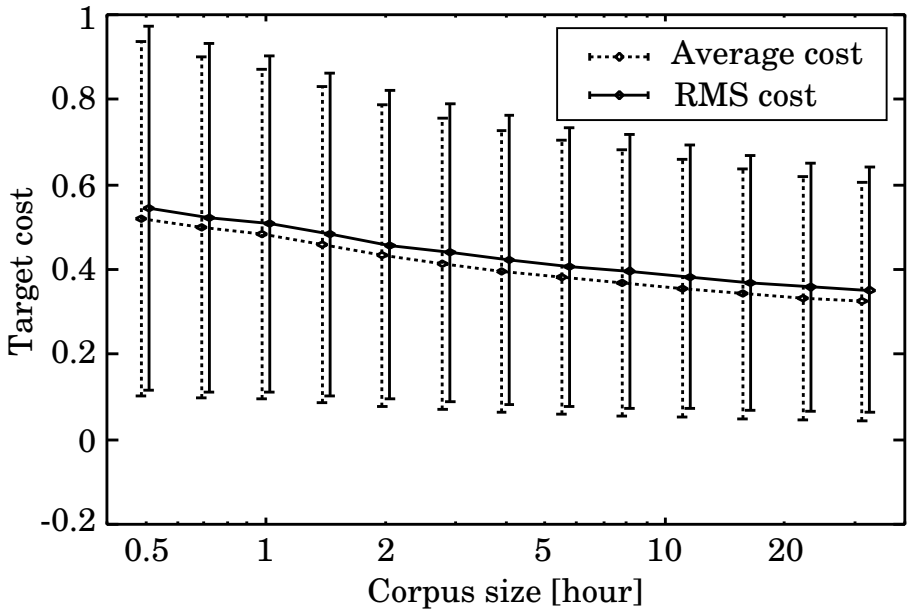

Figure 4.14. Target cost as a function of corpus size.

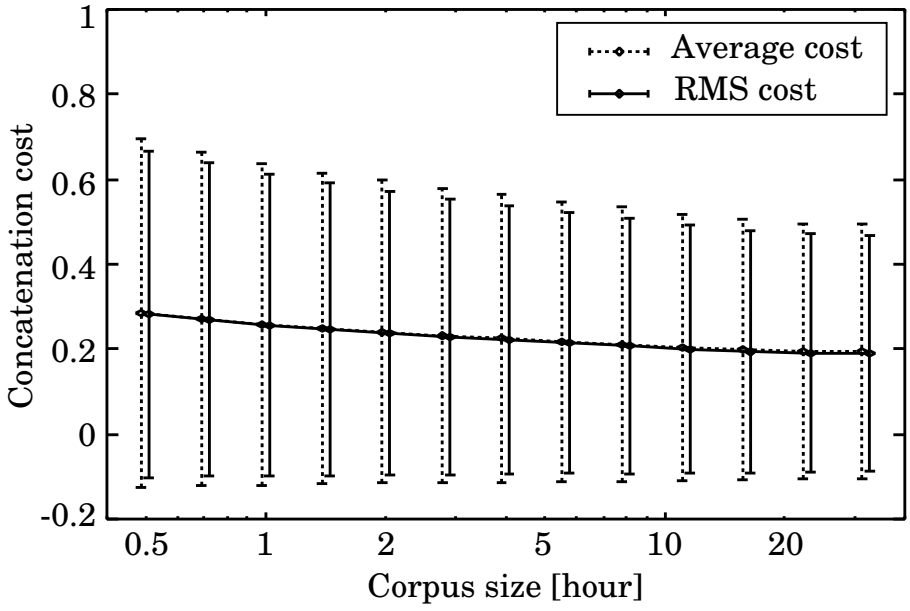

Figure 4.15. Concatenation cost as a function of corpus size.

were obtained for all weight sets, in which the ratios of the target cost to the concatenation cost were set to 1 to 2, 1 to 1.5, 1 to 1, 1.5 to 1, and 2 to 1. Therefore, the effectiveness mentioned above depends not on the weights for sub-costs but on the function used to integrate the local costs, which can take account of the local degradation as well as the the total degradation.

4.4.2 Effect of RMS cost on selected segments

In order to clarify what causes the decrease in the standard deviation of the concatenation cost, we investigated characteristics of the segments selected by minimizing the RMS cost. The segment length in the number of phonemes is shown in **Figure 4.16** as a function of corpus size. The segment length is shorter in the selection with the RMS cost compared with that in the case of the average cost, while the standard deviation is nearly equal.

Moreover, **Figure 4.17** shows the segment length in number of syllables as a function of corpus size for reference, since our segment selection is essentially based on syllable units, i.e. concatenations at C-V boundaries (C: Consonant and V: Vowel) are prohibited. In calculating the segment length, the number of syllables for half-phonemes in a diphone unit is set to 0.5 when the syllable is V, or 0.25 when the syllable is comprised of CV. It can be seen that the segment length is unexpectedly short, only less than 1.4 syllables, although the standard deviation of the segment length is large. In the segment selection, it is not necessarily best to select the longest segments, since such segments often cause a decrease in the number of candidate segments and thus cannot always synthesize speech naturally. The important point is to select the best segment sequence by considering not only the degradation caused by concatenation but also that caused by various factors, e.g. prosodic distance. It is also shown that the segment length becomes shorter as the corpus size increases to a level over 20 hours. This result is caused by pruning candidate segments to reduce the computational complexity of segment selection. We perform the pruning process, called pre-selection [29], by considering the target cost and the mismatch of phonetic environments. Namely, we do not consider whether segments are connected in the corpus. Therefore, when we use the large-sized corpus that includes many candidate segments having target phonetic environments, remaining candidate segments do not always

connected in the corpus even if these segments have target phonetic environments.

The rate of increase in the number of concatenations is shown in **Figure 4.18** as a function of corpus size. This rate is calculated by dividing the number of concatenations in the case of utilizing the RMS cost by that in the case of utilizing the average cost. By utilizing the RMS cost, the concatenation at a boundary between any phoneme and a voiced consonant ("* - Voiced consonant") decreases in any corpus size. However, the concatenations at both a phoneme center ("Phoneme center") and a boundary between any phoneme and an unvoiced consonant ("* - Unvoiced consonant") increase. **Figure 4.19** shows the concatenation cost in each type of concatenation when the corpus size is 32 hours. The concatenation between any phoneme and an unvoiced consonant can often reduce the concatenation cost compared with that between any phoneme and a voiced consonant, since the former type of concatenation has no discontinuity caused by concatenating F_0s at a segment boundary. It was also found that the concatenation at the phoneme center tends to reduce the concatenation cost compared with the other types of concatenation, although the number of concatenations is small.

These results show a tendency to avoid performing concatenations that cause much local degradation of naturalness by instead performing more concatenations that cause slight audible discontinuity. As a whole, the number of concatenations increases rather than decreases. Therefore, a larger number of segments with shorter lengths, which only cause slight local degradation, are selected by utilizing the RMS cost.

4.4.3 Relationship between effectiveness of RMS cost and corpus size

In order to clarify the relationship between the effectiveness of using RMS cost and the corpus size, we investigated the differences in average costs, RMS costs, and maximum costs between the segment sequences selected by utilizing the average cost and those by utilizing the RMS cost.

The results are shown in **Figure 4.20**. The cost differences are calculated by subtracting the costs of the segment sequences selected by utilizing the RMS cost from those of the segment sequences selected by utilizing the average cost

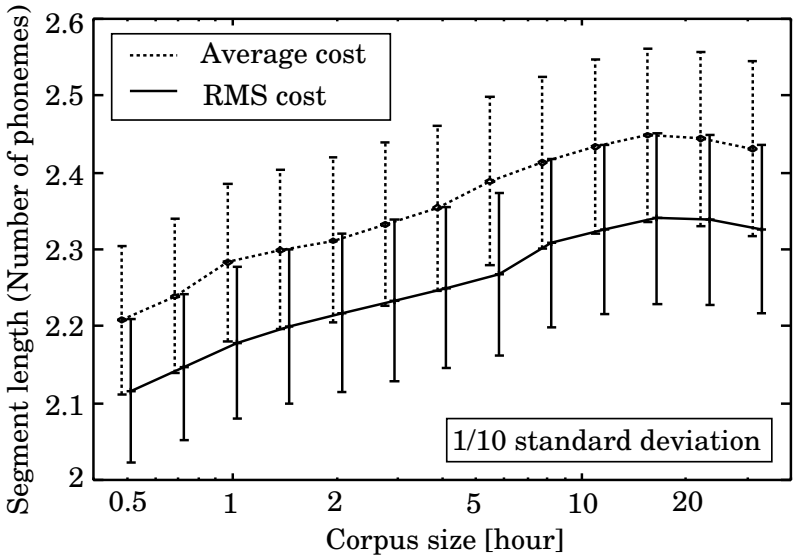

Figure 4.16. Segment length in number of phonemes as a function of corpus size.

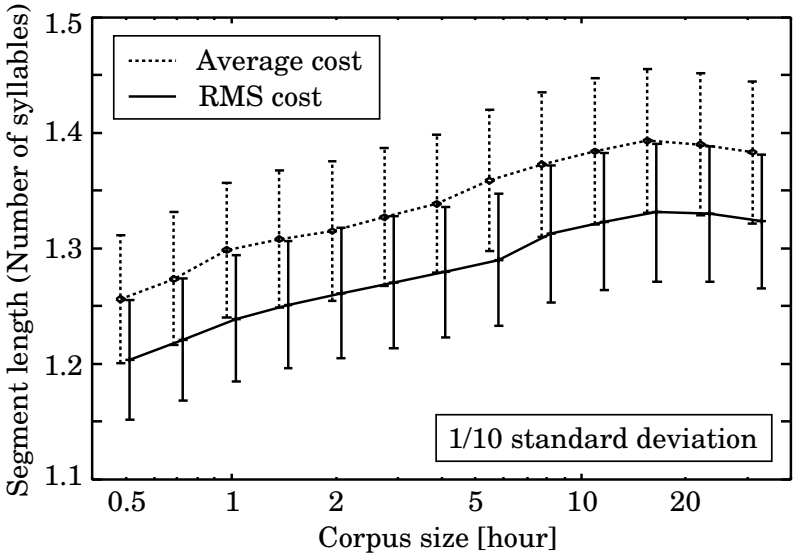

Figure 4.17. Segment length in number of syllables as a function of corpus size.

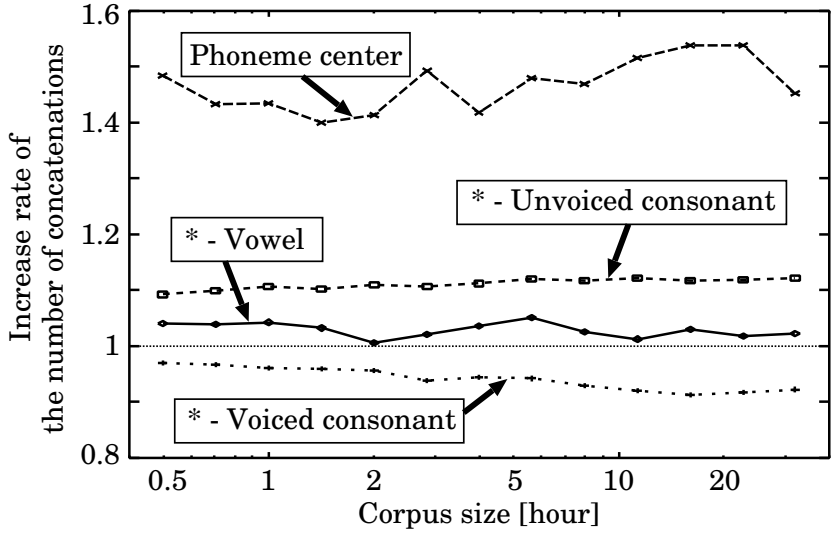

Figure 4.18. Increase rate in the number of concatenations as a function of corpus size. "*" denotes any phoneme.

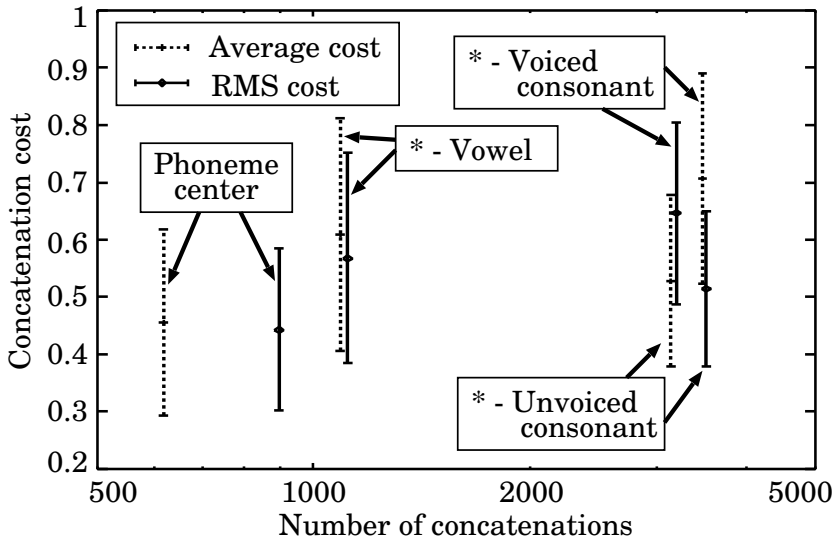

Figure 4.19. Concatenation cost in each type of concatenation. The corpus size is 32 hours.

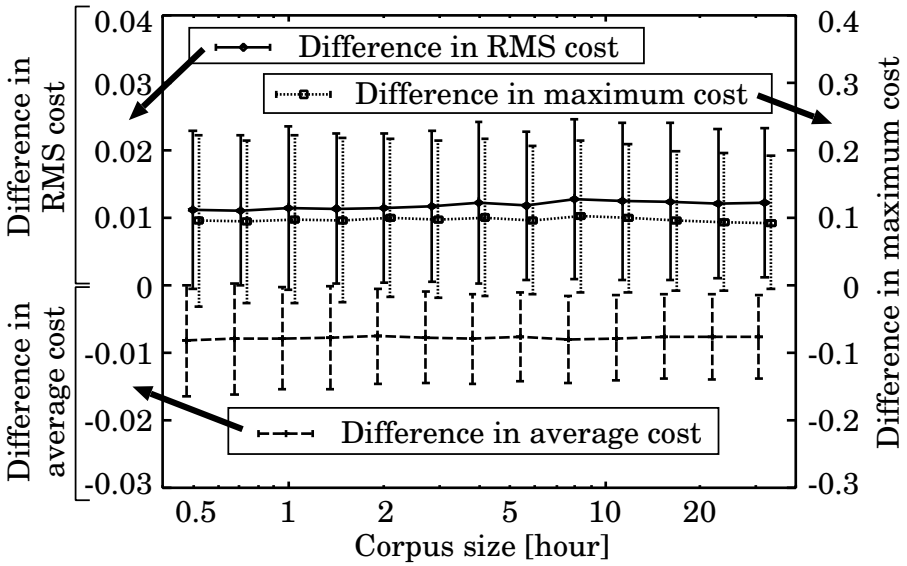

Figure 4.20. Differences in costs as a function of corpus size.

as described in **Section 4.3.2**. From the results, the RMS cost works well for alleviating the local degradation of naturalness, since the maximum cost becomes small, i.e. the differences in the maximum cost are positive. Moreover, the differences in all costs have little dependence on the corpus size. Therefore, the effectiveness of utilizing the RMS cost can be found in a corpus of any size.

4.4.4 Evaluation of segment selection by estimated perceptual score

The performance of segment selection is shown in the cost. However, it is difficult to estimate the naturalness of synthetic speech from the cost value directly. In order to indicate the performance of segment selection in a more intuitive quantity than the cost, we converted the cost value into a perceptual score by using the regression line on the RMS cost shown in **Figure 4.6**. Chu and Peng have also estimated the MOS from their cost [25][86].

Estimated perceptual score is shown in **Figure 4.21** as a function of corpus size. Small-sized corpora have been constructed so that various phonetic contexts

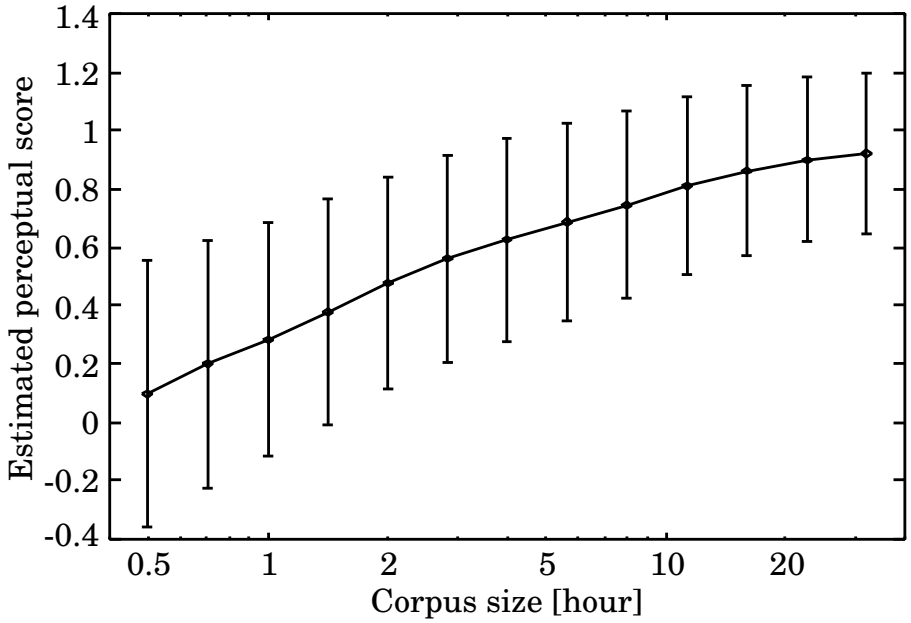

Figure 4.21. Estimated perceptual score as a function of corpus size.

are included. As the corpus size becomes larger, the estimated perceptual score becomes higher and its standard deviation becomes smaller. This result means that the quality of the segment selection is higher and more consistent by utilizing the larger corpus.

4.5 Summary

In segment selection for concatenative TTS, it is important to utilize a cost that corresponds to the perceptual characteristics. In this chapter, we evaluated a cost for segment selection based on a comparison with perceptual scores determined from the results of perceptual experiments on the naturalness of synthetic speech. As a result, we clarified that the average cost, which captures the total degradation, has a better correspondence to the perceptual scores than does the maximum cost, which captures the local degradation. Furthermore, we found that the RMS (Root Mean Square) cost, which takes into account both the average cost and the maximum cost, has the best correspondence. We also clarified

that the naturalness of synthetic speech could be slightly improved by utilizing the RMS cost.

We investigated the effect of considering not only the degradation of naturalness over the entire synthetic speech but also local degradation in segment selection. In this selection, the optimum segment sequences are selected by minimizing the RMS cost instead of the conventional average cost. From the results of experiments comparing this approach with segment selection based on the average cost, it was found that segment selection based on RMS cost performed a larger number of concatenations that caused slight local degradation in order to avoid concatenations causing greater local degradation. Namely, a larger number of segments with shorter units were selected. Moreover, the effectiveness of this selection was found for any size of corpus.

When the RMS costs were distributed widely, the correspondence of the RMS cost seemed to be good. Namely, it was possible to accurately estimate the perceptual score from the RMS cost for synthetic speech of various qualities. Therefore, we evaluated the performance of segment selection by the estimated perceptual score by varying the corpus size. As a result, the quality of the segment selection was higher and more consistent by utilizing the larger corpus.

We also performed the perceptual evaluation of the RMS cost in a lower range of the RMS cost. The results clarified that the correspondence of the RMS cost to the perceptual scores is inadequate in this case. Therefore, it is obvious that the RMS cost is not accurate enough for making comparisons between similar segments, which is naturally a difficult problem. However, since our TTS does not consistently synthesize sufficiently natural speech, we should further improve the cost function based on perceptual characteristics. In particular, it is necessary to determine the optimum weight set for sub-costs. We will determine this weight set from the results of perceptual experiments on the naturalness of synthetic speech with a set of stimuli covering a wide range in terms of individual sub-costs.

Chapter 5

A Voice Conversion Algorithm Based on Gaussian Mixture Model with Dynamic Frequency Warping

Speech of various speakers can be synthesized by utilizing a voice conversion technique that can control speaker individuality. The voice conversion algorithm based on the Gaussian Mixture Model (GMM), which is a conventional statistical voice conversion algorithm, can convert speech features continuously by using the correlation between a source feature and a target feature. However, the quality of the converted speech is degraded because the converted spectrum is excessively smoothed by the statistical averaging operation. In this chapter, we propose a GMM-based algorithm with Dynamic Frequency Warping (DFW) to avoid such over-smoothing. In the proposed algorithm, the converted spectrum is calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum to avoid deterioration of spectral conversion-accuracy. Results of evaluation experiments clarify that the converted speech by the proposed algorithm has a higher quality than that by the GMM-based algorithm and that the conversion-accuracy for speaker individuality is the same as that by the GMM-based algorithm.

5.1 Introduction

The speech of various speakers can be synthesized by preparing speech corpora of the various speakers. In order to synthesize natural speech consistently, each corpus must include a sufficient amount of waveform segments having various phonetic environments and prosodic variety. Therefore, it is necessary to record a large number of speech utterances for each speaker. However, recording speech is very difficult and expensive work. Furthermore, it needs an enormous amount of time [63].

Another approach for synthesizing speech of various speakers is speech modification by a voice conversion technique used to convert one speaker's voice into another speaker's voice [68]. In voice conversion, conversion rules are extracted in advance through training data comprised of utterance pairs of a source speaker and a target speaker. By utilizing the conversion rules, any utterance of the source speaker can be converted into that of the target speaker. A large amount of speech data is not required in the training. Therefore, speech of any speaker who is regarded as the target speaker can be synthesized from speech of a specific speaker who is regarded as the source speaker by only preparing a small number of the target speaker's utterances to extract the conversion rules. Namely, by applying voice conversion to corpus-based TTS, it is possible to easily synthesize any utterance of any speaker easily. However, the quality of the converted speech is degraded considerably. Therefore, we aim to develop a high-quality voice conversion system.

In voice conversion, it is important to realize both high speech quality and high conversion-accuracy for speaker individuality. Although we can synthesize high-quality speech using a simple conversion rule, e.g. converting only F_0, such converted speech cannot sound like that of the target speaker. Moreover, even if the converted speech sounds like that of the target speaker, the speech quality is not always high, e.g. in telephone speech, the speech quality is degraded although speaker individuality is kept. Therefore, we need to evaluate the performance of voice conversion in terms of both speech quality and conversion-accuracy for speaker individuality.

To continuously represent the acoustic space of a speaker, a voice conversion algorithm based on the Gaussian Mixture Model (GMM) has been proposed

by Stylianou et al. [98][99]. In this GMM-based algorithm, the acoustic space is modeled by the GMM without the use of vector quantization, and acoustic features are converted from a source speaker to a target speaker by a mapping function based on the feature-parameter correlation between two speakers.

The quality of an analysis-synthesis method is also important for achieving a high-quality voice conversion since voice conversion is usually performed with an analysis-synthesis method. As a high-quality analysis-synthesis method, STRAIGHT has been proposed by Kawahara et al. [58][59]. This algorithm provides a high-quality vocoder type algorithm.

In the GMM-based voice conversion algorithm applied to STRAIGHT, the quality of converted speech is degraded because the converted spectrum is excessively smoothed by statistical processing. In this chapter, we propose a GMM-based algorithm with Dynamic Frequency Warping (DFW) [114] to avoid over-smoothing. However, the spectral conversion-accuracy of the DFW algorithm is worse than that of the GMM-based algorithm because spectral intensity cannot be converted by the DFW. Therefore, the converted spectrum in the proposed algorithm is calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum in order to avoid the deterioration of spectral conversion-accuracy. We clarify that the proposed algorithm can improve the naturalness of synthetic speech by experimental evaluation.

This chapter is organized as follows. In **Section 5.2**, the GMM-based voice conversion algorithm applied to STRAIGHT is described. In **Section 5.3**, the proposed voice conversion algorithm is described, and the effect of mixing converted spectra is described in **Section 5.4**. In **Section 5.5**, experimental evaluation is described. Finally, we summarize this chapter in **Section 5.6**.

5.2 GMM-Based Conversion Algorithm Applied to STRAIGHT

The mel-cepstrum of the smoothed spectrum analyzed by STRAIGHT [58] is used as an acoustic feature. In this chapter, sampling frequency is 16000 Hz and the order of the mel-cesptrum is set to 40, i.e. quefrency is 2.5 ms. In order to perform voice conversion, the 1st to 40-th order cepstral coefficients are converted.

The 0-th order cepstral coefficient is converted so that the power of the converted speech becomes equal to that of the original speech. As a feature on source information, the log-scaled fundamental frequency extracted by STRAIGHT [59] is used.

In the training of the mapping function, we use 50 utterance-pairs comprised of the source speaker's speech and the target speaker's speech. The GMM parameters for spectral conversion are trained as follows:

Step 1: All utterances are converted into the mel-cepstrum sequences. Then, silence frames are removed by performing silence detection based on power information. In each pair, the correspondence between mel-cepstrum sequences extracted from two speakers' utterances is determined by using Dynamic Time Warping (DTW).

Step 2: GMM parameters for spectral conversion are trained with the time-aligned mel-cepstrum sequences.

Step 3: Differences between the source speaker's speech and the target speaker's speech have a bad influence on the accuracy of time-alignment. Therefore, the source speaker's utterances are converted into the target speaker's in order to perform more accurate time-alignment between the mel-cepstrum sequences. Then, mel-cepstral distortion between two speakers is calculated.

Step 4: Steps 2 and **3** are repeated to refine the GMM parameters until the mel-cepstral distortion calculated in **Step 3** is saturated.

The GMM parameters for F_0 conversion are trained as follows:

Step 1: F_0s are extracted from all utterances. In each pair, the correspondence of F_0 sequences from two speakers is determined by using time-alignment of the mel-cepstrum sequences determined in the training of spectral conversion. Then, frame-pairs including unvoiced frames are removed from the time-aligned F_0 sequences.

Step 2: GMM parameters for F_0 conversion are trained with the time-aligned F_0 sequences.

5.2.1 Evaluation of spectral conversion-accuracy of GMM-based conversion algorithm

We evaluated spectral conversion-accuracy of the GMM-based algorithm by the mel-cepstral distortion given by Equation (3.8). The evaluation set was comprised of 50 utterances that were not included in the training data. We performed voice conversions between male A and male B, between female A and female B, between male A and female A, and between male B and female B. In each case, we performed two kinds of conversion by alternating the speakers as a source speaker and a target speaker. The mel-cepstral distortion was calculated in every frame-pair in which the silence frames were removed. The frame shift is set to 5 ms. The number of Gaussian mixtures is set to 64.

The results (**Figure 5.1**) clarify that the GMM-based voice conversion can decrease the acoustic distance, that is, the mel-cepstral distortion, between the source speaker's speech and the target speaker's speech. The distances between speakers of opposite sex are larger than those between speakers of same sex in the case of original speech shown as "Before conversion." However, it can be seen that the distances become almost equal in all cases of voice conversion shown as "After conversion."

5.2.2 Shortcomings of GMM-based conversion algorithm

In the GMM-based algorithm applied to STRAIGHT, the quality of the converted speech is degraded because the converted spectrum is excessively smoothed by the statistical averaging operation. **Figure 5.2** shows an example of the GMM-based converted spectrum and the target speaker's spectrum. As shown in this figure, there is over-smoothing on the GMM-based converted spectrum.

As mentioned in **Section 2.3**, most statistical approaches to speaker adaptation can be applied to voice conversion. In speech recognition, it is not necessary to preserve the information removed by statistical processing. However, in speech synthesis, such information is important for synthesizing natural speech. Therefore, it is necessary to improve the adaptation algorithms in order to achieve high-quality voice conversion.

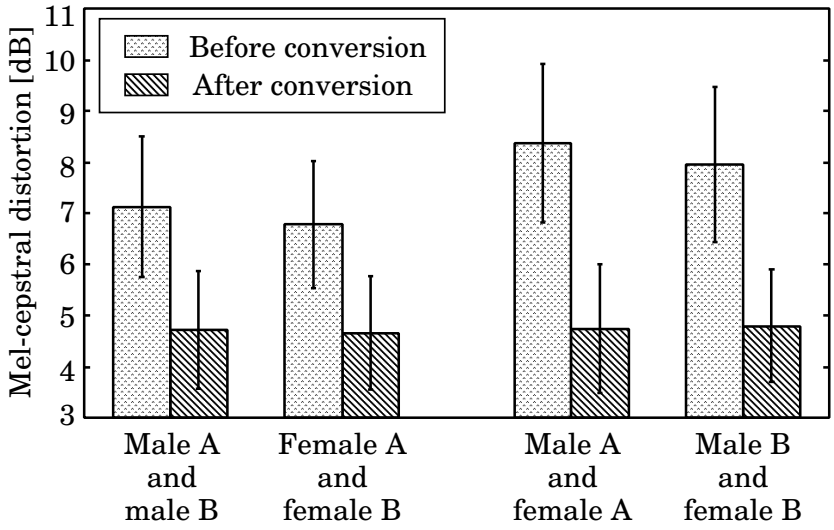

Figure 5.1. Mel-cepstral distortion. Mean and standard deviation are shown.

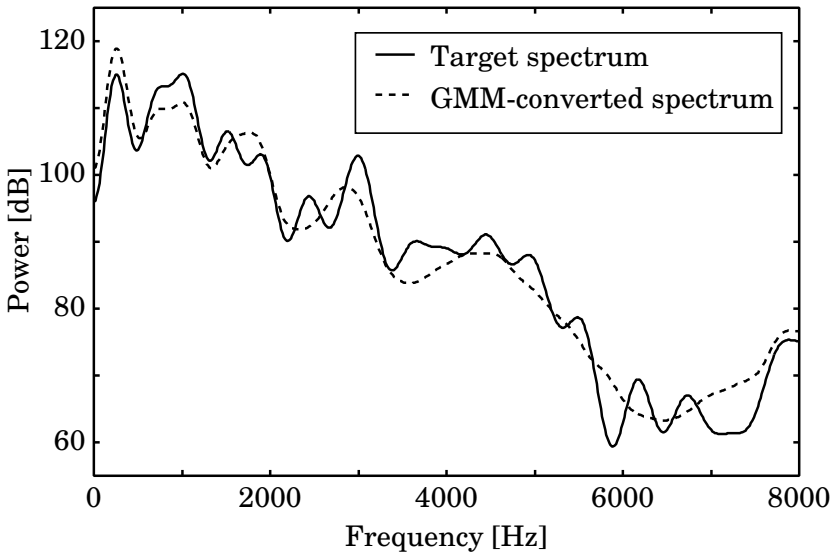

Figure 5.2. Example of spectrum converted by GMM-based voice conversion algorithm ("GMM-converted spectrum") and target speaker's spectrum ("Target spectrum").

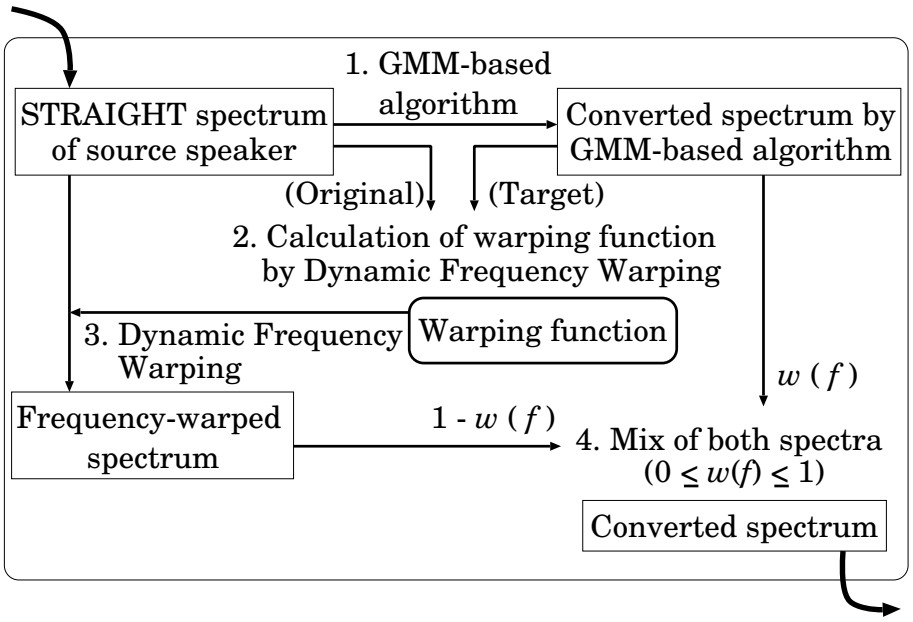

Figure 5.3. GMM-based voice conversion algorithm with Dynamic Frequency Warping.

5.3 Voice Conversion Algorithm Based on Gaussian Mixture Model with Dynamic Frequency Warping

To avoid over-smoothing, we propose a GMM-based algorithm with DFW. An overview of the proposed algorithm is shown in **Figure 5.3**. This procedure is performed in each frame.

5.3.1 Dynamic Frequency Warping

Spectral conversion is performed with DFW [72][114] to avoid over-smoothing of the converted spectrum. In this technique, the correspondence between an original frequency axis and a converted frequency axis is represented by a warping function. This function is calculated as a path minimizing a normalized mel-spectral distance between the STRAIGHT logarithmic mel-spectrum of the source

speaker and the GMM-based converted logarithmic mel-spectrum.

The normalized mel-spectral distance $D(i,j)$ is given by

$$D(i,j) = \sqrt{\frac{g(i,j)}{i+j}}, \tag{5.1}$$

$$g(i,j) = \min \begin{bmatrix} g(i-1,j-2) + 2d(i,j-1) + d(i,j), \\ g(i-1,j-1) + 2d(i,j), \\ g(i-2,j-1) + 2d(i-1,j) + d(i,j), \end{bmatrix}, \tag{5.2}$$

where $g(i,j)$ denotes an accumulated distance in the grid (i,j). $d(i,j)$ denotes a mel-spectral distance in a grid (i,j), and it is given by

$$d(i,j) = \left(20\log_{10}|S_x(i)| - 20\log_{10}|S_y(i)|\right)^2, \tag{5.3}$$

where S_x denotes a mel-spectrum of the source speaker at the i-th bin, and S_y denotes a mel-spectrum converted by the GMM-based algorithm at the j bin. In this chapter, the number of FFT points is set to 1024, and the number of bins to calculate the mel-spectral distance is 513. **Figure 5.4** shows an example of the frequency warping function for a certain frame.

Unnatural sound is often caused by synthesizing with DFW-based converted spectra, since such spectra have unnatural changing parts. These parts can be removed by performing smoothing or liftering. Therefore, we perform the smoothing processing used in STRAIGHT [58].

5.3.2 Mixing of converted spectra

The spectral conversion-accuracy of DFW is slightly worse than that of the GMM-based algorithm because spectral intensity cannot be converted. Therefore, we also propose that the converted spectrum be calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum in order to avoid the deterioration of the spectral conversion-accuracy. In the proposed algorithm, the converted spectrum $S_{\mathrm{c}}(f)$ is written as

$$\begin{aligned} |S_{\mathrm{c}}(f)| &= \exp[w(f)\ln|S_{\mathrm{g}}(f)| + \{1-w(f)\}\ln|S_{\mathrm{d}}(f)|], \\ \text{subject to } & \ 0 \le w(f) \le 1, \end{aligned} \tag{5.4}$$

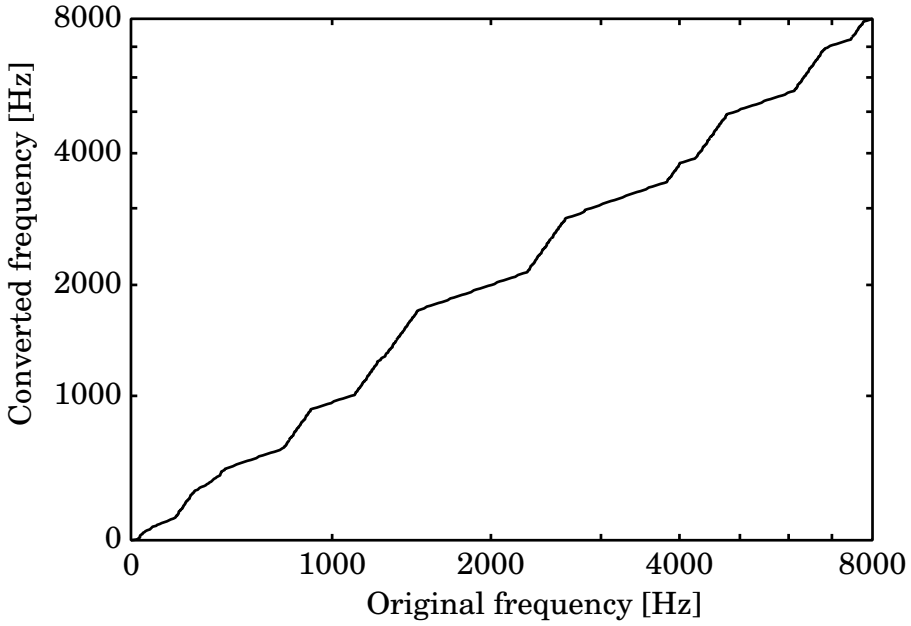

Figure 5.4. Example of frequency warping function.

where $S_{\mathrm{d}}(f)$ and $S_{\mathrm{g}}(f)$ denote the DFW-based converted spectrum and the GMM-based converted spectrum, respectively. $w(f)$ denotes the weight for mixing spectra. As the mixing-weight is closer to 1, the converted spectrum is closer to the GMM-based converted spectrum.

The results of preliminary experiments clarified that the quality of the converted speech is degraded considerably when a spectrum is excessively smoothed in the low-frequency regions [110]. Therefore, we use the mixing-weight as follows:

$$\begin{aligned} w(f) &= \left| \frac{2\pi f}{f_s} + 2\tan^{-1}\left(\frac{a\sin(2\pi f/f_s)}{1 - a\cos(2\pi f/f_s)} \right) \right| \Big/ \pi, \\ &\text{subject} \quad \text{to} \quad -1 < a < 1, \quad -f_s/2 \leq f \leq f_s/2, \end{aligned} \tag{5.5}$$

where f_s denotes the sampling frequency, and a denotes the parameter that changes the mixing-weight. **Figure 5.5** shows the variations of the mixing-weights that correspond to the different parameters a. **Figure 5.6** shows an example of converted spectra by the GMM-based algorithm ("GMM"), the GMM-based algorithm with DFW ("GMM & DFW"), and the GMM-based algorithm

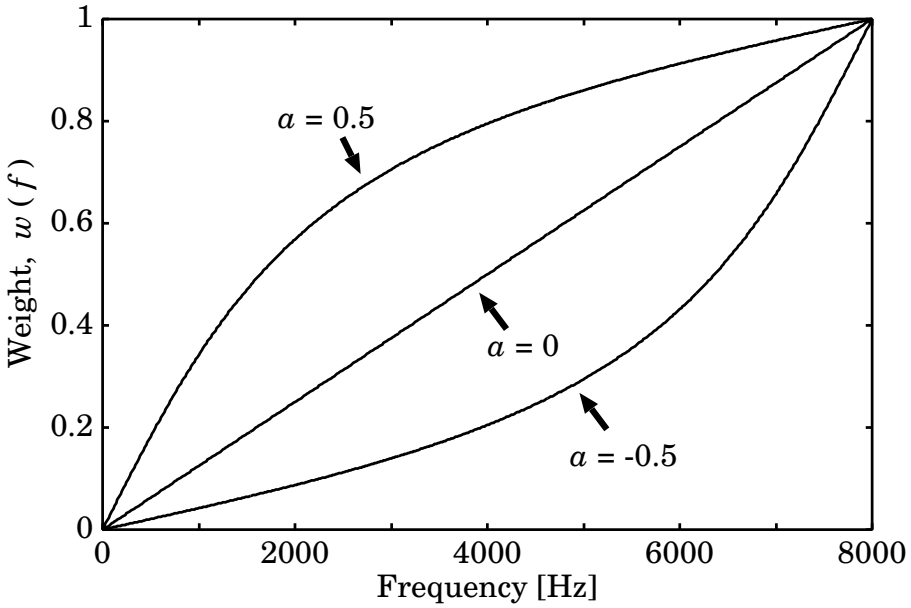

Figure 5.5. Variations of mixing-weights that correspond to the different parameters a.

with DFW and the mixing of converted spectra when the parameter a is set to 0 ("GMM & DFW & Mix of spectra"). The converted spectrum by the GMM-based algorithm with DFW has clearer peaks than that by the GMM-based algorithm. The mixed converted spectrum is close to the converted spectrum by the GMM-based algorithm with DFW in the low-frequency regions but close to that by the GMM-based algorithm in the high-frequency regions.

5.4 Effectiveness of Mixing Converted Spectra

Evaluation experiments were performed to investigate the effects of the mixing-weight. We performed objective experiments on spectral conversion-accuracy and subjective experiments on speech quality and speaker individuality. The amount of training data was set to 58 sentences. The total duration of this data is about 4 or 5 minutes. The male-to-male voice conversion and female-to-female voice conversion were performed, and 10 listeners participated in each subjective

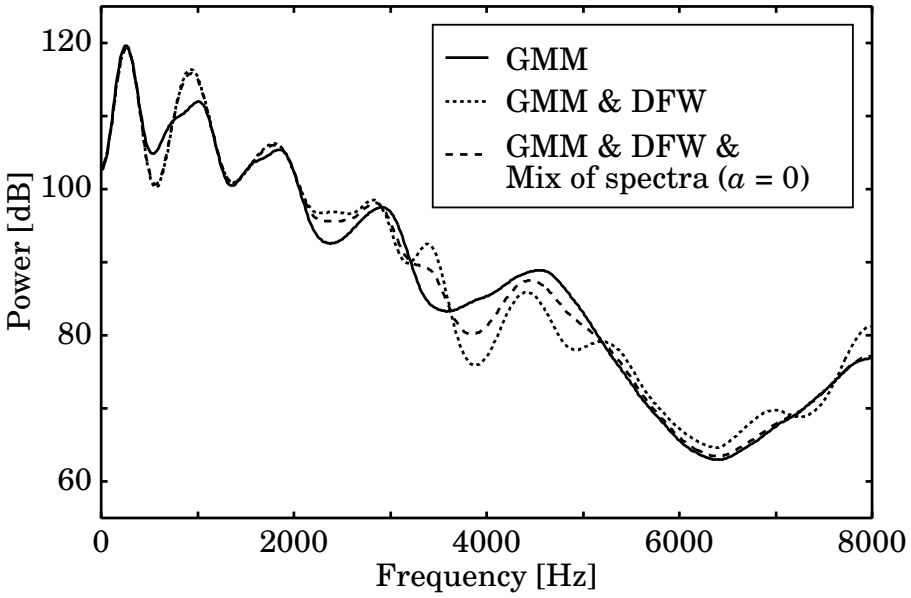

Figure 5.6. Example of converted spectra by the GMM-based algorithm ("GMM"), the proposed algorithm without the mix of the converted spectra ("GMM & DFW"), and the proposed algorithm with the mix of the converted spectra ("GMM & DFW & Mix of spectra").

experiment.

5.4.1 Effect of mixing-weight on spectral conversion-accuracy

In order to investigate the effects of the mixing-weight on spectral conversion-accuracy, we performed the experimental evaluation on the mel-cepstral distortion given by Equation (3.8) between the converted speech and the target speech. Ten sentences that were not included in the training data were used in the evaluation.

The experimental results shown in **Figure 5.7** clarify that the conversion-accuracy of the GMM-based algorithm with DFW ("GMM & DFW") is worse than that of the GMM-based algorithm ("GMM"). It is also shown that the deterioration of the conversion-accuracy can be alleviated by mixing the converted spectra.

As the parameter a of the mixing-weight becomes closer to 1, the mel-cepstral

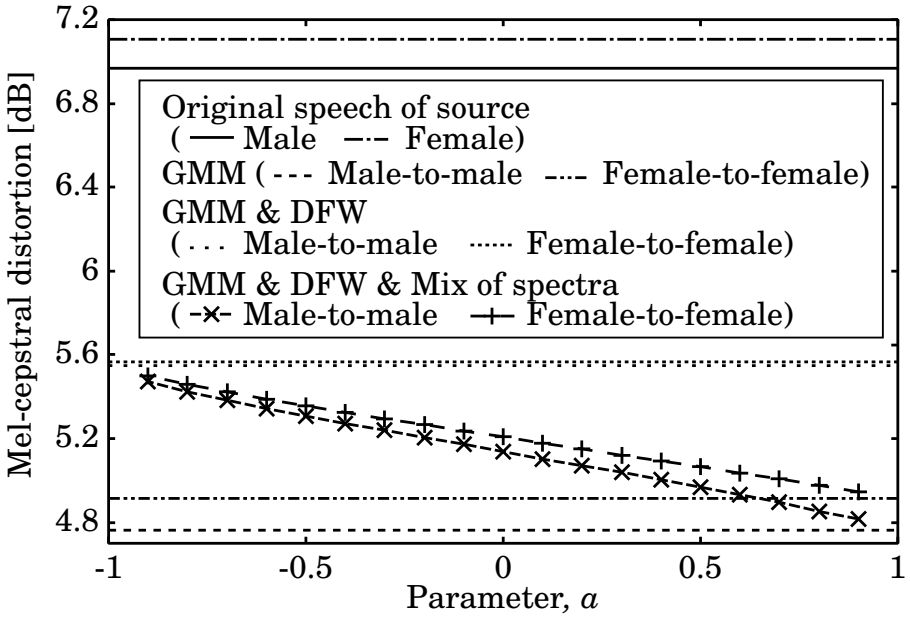

Figure 5.7. Mel-cepstral distortion as a function of parameter a of mixing-weight. "Original speech of source" shows the mel-cepstral distortion before conversion.

distortion in the GMM-based algorithm with DFW and the mixing of converted spectra ("GMM & DFW & Mix of spectra") becomes small. This is because the converted spectra by the proposed algorithm includes more components of the GMM-based converted spectra than those of the DFW-based converted spectra when the parameter a is near 1.

5.4.2 Preference tests on speaker individuality

In order to evaluate the relationship between the conversion-accuracy for speaker individuality and the parameter a of the mixing-weight, the preference (XAB) test was performed. Two sentences that were not included in the training data were used in the evaluation. In the XAB test, X was the synthesized speech by converting F_0 and replacing the source speaker's spectra with those of the target speaker, i.e. the perfect spectral conversion. A and B were the converted speech by the voice conversion algorithms. In each trial, X was presented first, and A

and B were presented in random order. Listeners were asked to choose either A or B as being the most similar to X.

The preference score of the proposed voice conversion algorithm compared with the GMM-based algorithm is shown in **Figure 5.8**. The conversion-accuracy for speaker individuality of the GMM-based algorithm with DFW ("GMM & DFW") tends to be slightly worse than that of the GMM-based algorithm. However, the conversion-accuracy becomes equal to that of the GMM-based algorithm by mixing the converted spectra ("GMM & DFW & Mix of spectra"). The tendency of these results is similar to that of the experimental results for the mel-cepstral distortion. However, even if the mel-cepstral distortion is degraded, the conversion-accuracy for speaker individuality is not always degraded. This demonstrates that the mel-cepstral distortion cannot capture the conversion-accuracy for speaker individuality with sufficient accuracy.

5.4.3 Preference tests on speech quality

In order to evaluate the relationship between the quality of the converted speech and the parameter a of the mixing-weight, the preference test was performed. Four sentences that were not included in the training data were used in the evaluation. In each trial, stimuli synthesized by the two voice conversion algorithms were presented in random order. Listeners were asked to choose the converted speech having better speech quality.

The experimental results are shown in **Figure 5.9**. The converted speech quality of the GMM-based algorithm with DFW is significantly better than that of the GMM-based algorithm ("GMM"). As the mixing-weight becomes close to 1, the converted speech quality of the GMM-based algorithm with DFW and the mix of the converted spectra ("GMM & DFW & Mix of spectra") becomes worse compared with the GMM-based algorithm with DFW. These results shows the opposite tendency to that of the conversion-accuracy for speaker individuality.

The results of the two experiments clarify that both the converted speech quality and the conversion-accuracy for speaker individuality show good performance when the parameter a of the mixing-weight is set to 0 or -0.5.

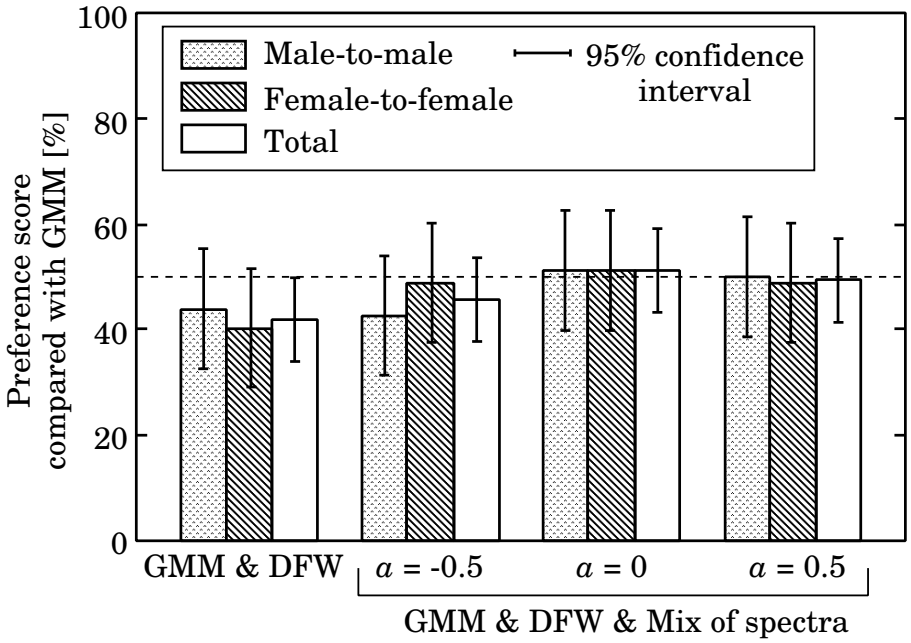

Figure 5.8. Relationship between conversion-accuracy for speaker individuality and parameter a of mixing-weight. A preference score of 50% shows that the conversion-accuracy is equal to that of the GMM-based algorithm, which provides good performance in terms of speaker individuality.

5.5 Experimental Evaluation

In order to evaluate the performance of the proposed algorithm, we performed subjective evaluation experiments on speaker individuality and speech quality. The experimental conditions were the same as those described in the previous section. The parameter a was set to 0 and -0.5 in the male-to-male and the female-to-female voice conversions, respectively.

5.5.1 Subjective evaluation of speaker individuality

In order to evaluate the conversion-accuracy for speaker individuality of the proposed algorithm, a preference (ABX) test was performed. In the ABX test, A and B were the source and target speaker's speech, respectively, and X was one

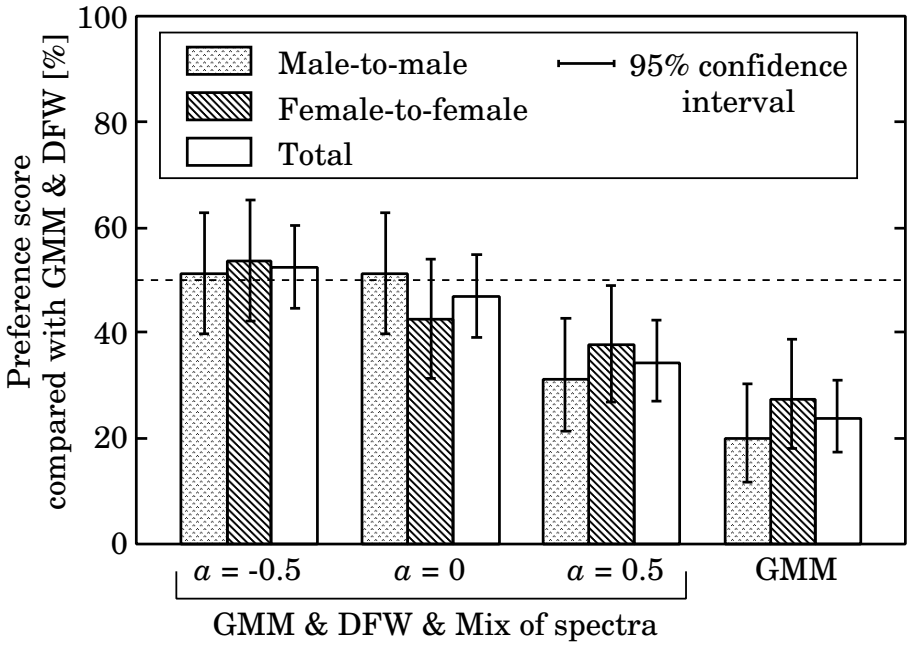

Figure 5.9. Relationship between converted speech quality and parameter a of mixing-weight. A preference score of 50% shows that the converted speech quality is equal to that of the GMM-based algorithm with DFW, which provides good performance in terms of speech quality.

of the types of converted speech as follows:

- converted speech by the proposed algorithm $\cdots$ "Proposed algorithm",
- converted speech by the GMM-based algorithm $\cdots$ "GMM",
- synthesized speech by performing only the F_0 conversion $\cdots$ "F_0 only",
- synthesized speech by performing the F_0 conversion and replacing the source speaker's spectrum with that of the target speaker $\cdots$ "F_0 & Spectrum",
- target speaker's speech synthesized by STRAIGHT $\cdots$ "STRAIGHT."

"F_0 & Spectrum" were used to evaluate the conversion-accuracy for speaker individuality when the spectral conversion was performed perfectly. "STRAIGHT"

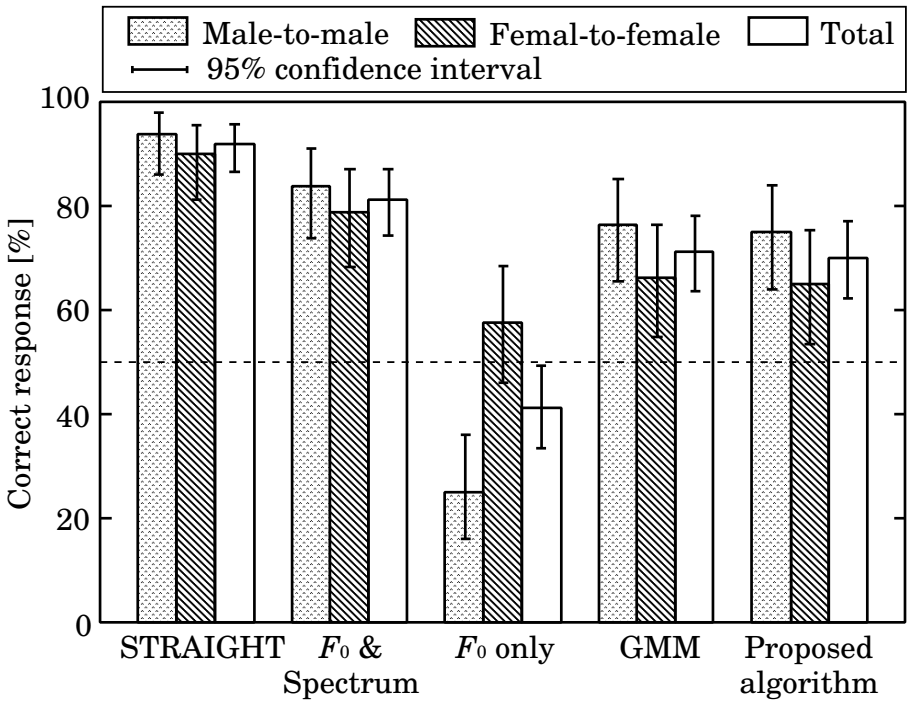

Figure 5.10. Correct response for speaker individuality.

was used to evaluate the conversion-accuracy when both the spectral conversion and the conversion of source information were performed perfectly. In each trial, A and B were presented in random order, and then X was presented. Listeners were asked to choose either A or B as being the most similar to X. Two sentences that were not included in the training data were used in the evaluation.

The experimental results are shown in **Figure 5.10**. The conversion-accuracy for speaker individuality of the proposed algorithm ("Proposed algorithm") is the same as that of the GMM-based algorithm ("GMM"). The results clarify that the conversion-accuracy of only the F_0 conversion ("F_0 only") is insufficient, and it can be improved by converting spectra. The difference in the conversion-accuracy between the "F_0 & Spectrum" and the proposed algorithm shows that the spectral conversion is not sufficient. Moreover, the difference between STRAIGHT and "F_0 & Spectrum" shows that conversion of prosody should be improved in order to achieve better performance.

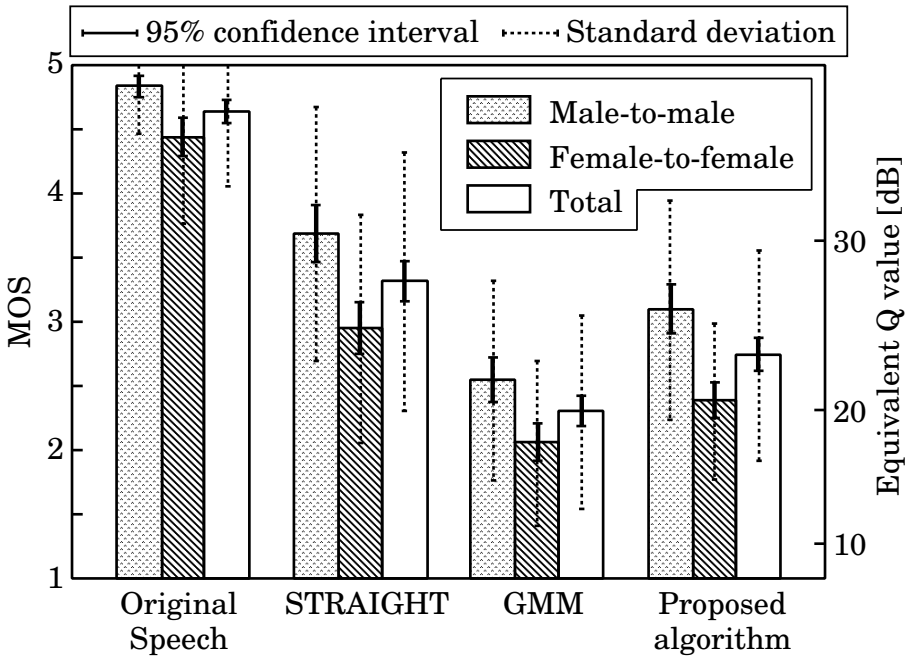

Figure 5.11. Mean Opinion Score ("MOS") for speech quality.

5.5.2 Subjective evaluation of speech quality

In order to evaluate the quality of the converted speech by the proposed algorithm, an opinion test was performed. An opinion score for evaluation was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Four sentences that were not included in the training data were used in the evaluation.

The experimental results are shown in **Figure 5.11**. The converted speech quality by the proposed algorithm ("Proposed algorithm") is significantly better than that of the GMM-based algorithm ("GMM") because over-smoothing of the converted spectrum can be avoided in the proposed algorithm. However, the converted speech quality is degraded compared with that of the synthetic speech by STRAIGHT analysis-synthesis. Moreover, it is shown that the quality of STRAIGHT cannot reach that of the original speech. Therefore, it is necessary to improve both the quality of the voice conversion algorithm and that of the analysis-synthesis method in order to achieve higher-quality voice conversion.

The results of the two subjective evaluations clarify that the proposed algo-

rithm can improve the naturalness of synthetic speech while maintaining equal conversion-accuracy on speaker individuality compared with the GMM-based algorithm.

5.6 Summary

In this chapter, we proposed a voice conversion algorithm based on the Gaussian Mixture Model (GMM) with Dynamic Frequency Warping (DFW) of the STRAIGHT spectrum. We also proposed a converted spectrum calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum. The GMM-based voice conversion algorithm, which is a statistical algorithm, can convert speech features continuously with the correlation between a source feature and a target feature. However, the converted spectrum is excessively smoothed by statistical processing. The proposed algorithm can avoid the typical over-smoothing of a converted spectrum, which causes the degradation of speech quality. In order to evaluate the proposed algorithm, we performed experimental evaluations of speech quality and speaker individuality by comparison with the conventional GMM-based algorithm. The experimental results revealed that the proposed algorithm can improve converted speech quality while maintaining equal conversion-accuracy for speaker individuality compared with the GMM-based algorithm.

However, the performance of the proposed voice conversion is not high enough. In order to reach a practical level of performance, it is necessary to improve not only the conversion algorithm but also the speech analysis-synthesis method. Moreover, the conversion of source information, i.e. prosody conversion, is needed since prosody also contains important information on speaker individuality.

Chapter 6

Conclusions

6.1 Summary of the Thesis

Corpus-based Text-to-Speech (TTS) enables us to dramatically improve the naturalness of synthetic speech over that of rule-based TTS. However, so far no general-purpose TTS has been developed that can consistently synthesize sufficiently natural speech. Furthermore, the flexibility of TTS is still inadequate.

In order to improve the performance of corpus-based TTS, we addressed two problems in speech synthesis in this thesis. One is how to improve the naturalness of synthetic speech in corpus-based TTS. The other is how to improve the control of speaker individuality in order to realize a more flexible speech synthesis.

We first described the structure of a corpus-based TTS system in **Chapter 2**. Almost all corpus-based TTS systems have been developed on the basis of this structure. The various techniques in each module were reviewed. We also described some conventional statistical voice conversion algorithms and compared conversion functions of the algorithms.

In **Chapter 3**, we proposed a novel segment selection algorithm for Japanese speech synthesis in order to improve the naturalness of synthetic speech. Since Japanese syllables consist of CV (C: Consonant or consonant cluster, V: Vowel or syllabic nasal /N/) or V, except when a vowel is devoiced, CV units are often used in concatenative TTS systems for Japanese. However, speech synthesized with CV units sometimes have auditory discontinuity due to V-V and V-semivowel concatenations. Since various vowel sequences appear frequently in Japanese, it

is not realistic to prepare long units that include all possible vowel sequences to avoid V-V concatenation. In order to address this problem, we proposed a novel segment selection algorithm that does not avoid the concatenation of the vowel sequences but alleviates the discontinuity by utilizing both phoneme and diphone units. In the proposed algorithm, non-uniform units allowing concatenation not only at phoneme boundaries but also at vowel centers can be selected from a speech corpus. The experiments on concatenation of vowel sequences clarified that the number of better candidate segments increases by considering concatenations both at phoneme boundaries and at vowel centers. We also performed perceptual experiments. The results showed that speech synthesized with the proposed algorithm has better naturalness than that of the conventional algorithms. We also compared the proposed algorithm with the algorithm based on half-phoneme units. Moreover, a cost function for selecting the optimum waveform segments in our TTS system was described.

In **Chapter 4**, we performed a perceptual evaluation of costs for segment selection. In order to achieve high-quality segment selection for concatenative TTS, it is important to utilize a cost that corresponds to the perceptual characteristics. From the results of perceptual experiments, we clarified the correspondence of the cost to the perceptual scores and then evaluated various functions to integrate local costs capturing degradation in individual segments. As a result, it was clarified that the average cost, which captures the degradation of naturalness over the entire synthetic speech, has better correspondence to the perceptual scores than the maximum cost, which captures the local degradation of naturalness. Furthermore, the RMS (Root Mean Square) cost taking account of both average cost and maximum cost has the best correspondence. We also showed that the naturalness of synthetic speech can be slightly improved by utilizing the RMS cost. Then, we investigated the effect of using the RMS cost for segment selection. From the results of experiments comparing this approach with segment selection based on the conventional average cost, it was found that (1) in segment selection based on the RMS cost, a larger number of concatenations causing slight local degradation were performed in order to avoid concatenations causing greater local degradation, and (2) the effect of the RMS cost has little dependence on the size of the corpus. We also clarified that utilizing a larger corpus improves the quality and

consistency of the segment selection.

In **Chapter 5**, control of speaker individuality by voice conversion was described. The voice conversion algorithm based on the Gaussian Mixture Model (GMM), which is one of the conventional algorithms, can convert speech features continuously with the correlation between a source feature and a target feature. However, the quality of the converted speech is degraded because the converted spectrum is excessively smoothed by the statistical averaging operation. To overcome this problem, we proposed the GMM-based algorithm with Dynamic Frequency Warping (DFW). The converted spectrum in the proposed algorithm is calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum in order to avoid deterioration of spectral conversion-accuracy. In order to evaluate the proposed algorithm, we performed experimental evaluation of speech quality and speaker individuality by comparison with the conventional GMM-based algorithm. The experimental results revealed that the proposed algorithm can improve converted speech quality while maintaining equal conversion-accuracy for speaker individuality compared with the GMM-based algorithm.

In summary, we confirmed that the proposed segment selection algorithm and the proposed cost function based on perceptual evaluation are effective for improving the naturalness of synthetic speech. In addition, the proposed voice conversion algorithm can help improve the control of speaker individuality.

6.2 Future Work

Although we have improved the corpus-based TTS system, a number of problems still remain to be solved.

Effective search algorithm: The computational complexity of segment selection needs to be reduced while maintaining the naturalness of synthetic speech. As an approach to this problem, some clustering algorithms have been proposed to decrease the number of candidate segments [14][35]. Decision trees are constructed by utilizing target features in segment selection in advance in order to cluster similar candidate segments into the same classes.

Other approaches have proposed pre-selection by sub-costs with small computational cost [29], which is applied in our TTS, and construction of a practical and efficient cache of sub-costs with high computational cost [9]. Moreover, a pruning algorithm that considers concatenation between candidate segments has been proposed [18]. It is expected that a larger-sized corpus will be used in the future to synthesize high-quality speech more consistently. More approaches from various viewpoints are needed to address this problem.

Measures against voice-quality variation: Variation in voice quality is caused by recording the speech of a speaker for a long time. Some approaches as described in **Section 2.2.5** have been proposed [63][94]. However, much more research is needed to solve this problem.

Utilization of multiple targets: Almost all algorithms use the most suitable target information predicted from contextual information. However, the predicted target information is not always the best in cases where only a small number of candidate segments having the predicted target exist in the corpus. It is assumed that segment sequences with smooth concatenations cannot be selected under such conditions. As an interesting approach to this problem, Bulyko et al. proposed a selection algorithm that considers multiple targets [15]. Moreover, Hirai et al. proposed a selection algorithm based on acceptable prosodic targets in Japanese speech synthesis [42]. Especially in Japanese, since accent information is crucial, it is important to avoid selecting segment sequences with unacceptable accent information. Moreover, it might be promising to utilize not only accent information but also other contextual information, e.g. syntactic structure in segment selection.

Improvement of cost: Cost functions for segment selection should be improved based on perceptual characteristics [25][86]. In this thesis, we did not determine the optimum weight set for sub-costs. Although a determination algorithm based on linear regression has been proposed, it uses an acoustic measure, e.g. cepstral distortion, as an objective measure in the regression [47]. However, the correspondence to perceptual characteristics of acoustic

measures is not sufficient. Therefore, it is necessary to explore a measure having better correspondence to perceptual characteristics.

Various applications of TTS: Not only general-purpose TTS but also limited domain TTS have been studied [15][26]. As an effective approach in the case of a small-sized corpus, spectral modification has been applied to concatenation speech synthesis [116]. Moreover, it is necessary to synthesize not only speech in a reading style but also speech in various speaking styles [54][65] as well as expressive speech [16][48] in order to realize rich communication between man and machine. A number of these problems still remain to be solved.

Further improvement of voice conversion: Voice conversion is an attractive technique. However, no practical voice conversion systems has been developed yet. Further improvement in performance is indispensable. Moreover, a prosodic conversion algorithm is also needed since prosody is an important factor for representing speaker individuality.

Cross-language voice conversion: As an interesting application of voice conversion, cross-language voice conversion has been proposed by Abe et al. [3]. This technique can synthesize other-language speech uttered by a speaker who cannot naturally speak the language. We applied our voice conversion algorithm to this type of cross-language voice conversion and evaluated the performance of the conversion [73]. More research is needed to advance cross-language voice conversion.

Appendix

A Frequency of Vowel Sequences

Table A.1 shows the frequency of vowel sequences in newspaper articles comprised of 571,283 sentences. Phoneme sequences were divided into CV* units [60], and then we calculated the frequency of vowel sequences while ignoring consonants, e.g. both /kai/ (CVV) and /ai/ (VV) are considered /ai/ (length = 2). Semivowels were considered vowels. The length of long vowels is set to 1, and that of devoiced vowels is set to 0.

Table A.1. Frequency of vowel sequences

Length	Frequency	Normalized frequency [%]	Number of different sequences
0	518553	2.7590	-
1	16069643	85.4999	11
2	1677866	8.9272	99
3	459538	2.4450	575
4	57090	0.3038	1154
5	10306	0.0548	1155
6	1582	0.0084	473
7	306	0.0016	146
8	28	0.0002	30
9	4	0.0000	4

B Definition of the Nonlinear Function P

We performed perceptual experiments on the degradation of naturalness caused by prosody modification with STRAIGHT [58][59] in order to define a sub-cost function on prosodic difference. Listeners evaluated the degradation on a scale of seven levels, namely 1 (very bad) to 7 (very good). Perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The perceptual scores were modeled by a nonlinear function z as follows:

$$z(x,y) = \min\{s, g(x,y)\}, \tag{B.1}$$

$$g(x,y) = a \cdot \exp\left[-\left\{\left(\frac{x}{S_x}\right)^2 - 2\cdot r\cdot\left(\frac{x}{S_x}\right)\cdot\left(\frac{y}{S_y}\right) + \left(\frac{y}{S_y}\right)^2\right\}\right] + b, \tag{B.2}$$

$$S_x = \begin{cases} S_{xp} & (x>0) \\ S_{xm} & (x\le 0) \end{cases}, \quad S_y = \begin{cases} S_{yp} & (y>0) \\ S_{ym} & (y\le 0) \end{cases}, \tag{B.3}$$

where, x and y denote F_0 modification ratio and duration modification ratio by octave, respectively. Parameters a, b, r, s, S_{xp}, S_{xm}, S_{yp}, and S_{ym} are given by

$$\begin{array}{rclrcl} a &=& 2.649368, & b &=& -1.574549, \\ r &=& 0.062317, & s &=& 0.942505, \\ S_{xp} &=& 0.400522, & S_{xm} &=& 0.568179, \\ S_{yp} &=& 0.924295, & S_{ym} &=& 0.913813. \end{array} \tag{B.4}$$

These parameters were estimated by minimizing the error between the perceptual score and that estimated by the function $z(x,y)$. The nonlinear function P in Equation (3.3) is defined as follow:

$$P(x,y) = -z(x,y) + s. \tag{B.5}$$

The function P is shown as the contour line in **Figure B.1**.

In the cases of 1) the sub-cost on F_0 discontinuity as described in Equation (3.4) and 2) no signal processing for prosody modification in waveform synthesis,

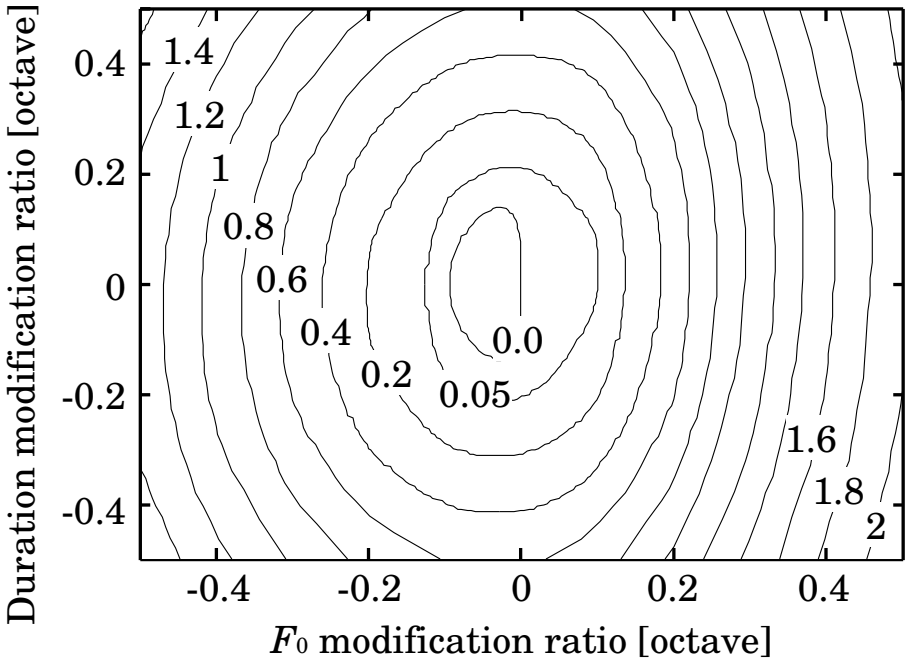

Figure B.1. Nonlinear function P for sub-cost on prosody.

the parameters are given by

$$\begin{array}{rclrcl} a &=& 2.649368, & b &=& -1.574549, \\ r &=& 0.0, & s &=& 1.074819, \\ S_{xp} &=& 0.400522, & S_{xm} &=& 0.400522, \\ S_{yp} &=& 0.913813, & S_{ym} &=& 0.913813. \end{array} \tag{B.6}$$

C Sub-Cost Functions, S_s and S_p, on Mismatch of Phonetic Environment

The sub-cost functions were determined from results of perceptual experiments, in which listeners evaluated the degradation of naturalness by listening to the speech stimuli synthesized by concatenating phonemes extracted from various phonetic environments. The experimental method is described in [64]. Listeners evaluated the degradation on a scale of seven levels, namely 1 (very bad) to 7 (very good). Perceptual score was calculated as an average of the normalized score calculated as a Z-score (mean = 0, variance = 1) for each listener in order to equalize the score range among listeners.

The cost S_s of capturing the degradation caused by a mismatch with the succeeding environment is given by

$$S_s(Ph, Ph_e, Ph_u) = -z(Ph, Ph_e, Ph_u) + b, \tag{C.1}$$

where $z(Ph, Ph_e, Ph_u)$ denotes the perceptual score in the case of a mismatch of the succeeding environment in a phoneme Ph, i.e. replacing a phoneme Ph_e with a phoneme Ph_u. b is set to 1.5, which is utilized in order to convert the perceptual score into a positive value. The cost S_p of capturing the degradation caused by a mismatch with the preceding environment is determined in a similar way.

References

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

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.

[3] M. Abe, K. Shikano, and H. Kuwabara. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *J. Acoust. Soc. Am.*, Vol. 90, No. 1, pp. 76–82, 1991.

[4] M. Abe and S. Sagayama. A voice conversion based on phoneme segment mapping. *J. Acoust. Soc. Jpn. (E)*, Vol. 13, No. 3, pp. 131–139, 1992.

[5] M. Akamine and T. Kagoshima. Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS Drive TTS). *Proc. ICSLP*, pp. 1927–1930, Sydney, Australia, Dec. 1998.

[6] C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House. Reduction of speech spectra by analysis-by-synthesis techniques. *J. Acoust. Soc. Am.*, Vol. 33, No. 12, pp. 1725–1736, 1961.

[7] M. Beutnagel, A. Conkie, and A.K. Syrdal. Diphone synthesis using unit selection. *Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 185–190, Jenolan Caves, Australia, Nov. 1998.

[8] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A.K. Syrdal. The AT&T Next-Gen TTS system. *Joint Meeting of ASA, EAA, and DAGA*,

Berlin, Germany, Mar. 1999.
http://www.research.att.com/projects/tts/pubs.html

[9] M. Beutnagel, M. Mohri, and M. Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. *Proc. EUROSPEECH*, pp. 607–610, Budapest, Hungary, Sep. 1999.

[10] A.W. Black and P. Taylor. CHATR: a generic speech synthesis system. *Proc. COLING*, pp. 983–986, Kyoto, Japan, Aug. 1994.

[11] A.W. Black and N. Campbell. Optimizing selection of units from speech database for concatenative synthesis. *Proc. EUROSPEECH*, pp. 581–584, Madrid, Spain, Sep. 1995.

[12] A.W. Black. Building practical speech synthesis systems. *ATR Technical Report*, TR-IT-0163, 1996.

[13] A.W. Black and A.J. Hunt, Generating F_0 contours from ToBI labels using linear regression. *Proc. ICSLP*, pp. 1385–1388, Philadelphia, U.S.A., Oct. 1996.

[14] A.W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. *Proc. EUROSPEECH*, pp. 601–604, Rhodes, Greece, Sep. 1997.

[15] A.W. Black and K. Lenzo. Limited domain synthesis. *Proc. ICSLP*, Vol. 2, pp. 411–414, Beijing, China, Sep. 2000.

[16] M. Bulut, S.S. Narayanan, A.K. Syrdal. Expressive speech synthesis using a concatenative synthesizer. *Proc. ICSLP*, pp. 1265–1268, Denver, U.S.A., Sep. 2002.

[17] I. Bulyko and M. Ostendorf. Efficient integrated response generation from multiple targets using weighted finite state transducers. *Computer Speech and Language*, Vol. 16, No. 3–4, pp. 533–550, 2002.

[18] I. Bulyko, M. Ostendorf, and J. Bilmes. Robust splicing costs and efficient search with BMM models for concatenative speech synthesis. *Proc. ICASSP*, pp. 461–464, Orlando. U.S.A., May 2002.

[19] W.N. Campbell and C.W. Wightman. Prosodic encoding of syntactic structure for speech synthesis. *Proc. ICSLP*, pp. 369–372, Banff, Canada, Oct. 1992.

[20] W.N. Campbell. Prosody and the selection of units for concatenation synthesis. *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis*, pp. 61–64, New York, U.S.A., Sep. 1994.

[21] W.N. Campbell. CHATR: a high-definition speech re-sequencing system. *Proc. Joint Meeting of ASA and ASJ*, pp. 1223–1228, Hawaii, U.S.A., Dec. 1996.

[22] W.N. Campbell and A.W. Black. Prosody and the selection of source units for concatenative synthesis. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 279–292, 1997.

[23] W.N. Campbell. Processing a speech corpus for CHATR synthesis. *Proc. ICSP*, pp. 183–186, Seoul, Korea, Aug. 1997.

[24] M. Chu, H. Peng, H. Yang, and E. Chang. Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. *Proc. ICASSP*, pp. 785–788, Salt Lake City, U.S.A., May 2001.

[25] M. Chu and H. Peng. An objective measure for estimating MOS of synthesized speech. *Proc. EUROSPEECH*, pp. 2087–2090, Aalborg, Denmark, Sep. 2001.

[26] M. Chu, C. Li, H. Peng, and E. Chang. Domain adaptation for TTS systems. *Proc. ICASSP*, pp. 453–456, Orlando. U.S.A., May 2002.

[27] A. Conkie and S. Isard. Optimal coupling of diphones. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 293–304, 1997.

[28] A. Conkie. Robust unit selection system for speech synthesis. *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, Mar. 1999.
http://www.research.att.com/projects/tts/pubs.html

[29] A. Conkie, M. Beutnagel, A.K. Syrdal, and P.E. Brown. Preselection of candidate units in a unit selection-based text-to-speech synthesis system. *Proc. ICSLP*, Vol. 3, pp. 279–282, Beijing, China, Oct. 2000.

[30] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc., Vol. 39, No. 1, pp. 1–38, 1977.

[31] W. Ding, K. Fujisawa, and N. Campbell. Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification. *Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 191–194, Jenolan Caves, Australia, Nov. 1998.

[32] N.R. Dixon and H.D. Maxey. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Trans. Audio Electroacoust.*, Vol. 16, No. 1, pp. 40–50, 1968.

[33] R.E. Donovan and E.M. Eide. The IBM trainable speech synthesis system. *Proc. ICSLP*, pp. 1703–1706, Sydney, Australia, Dec. 1998.

[34] R.E. Donovan and P.C. Woodland. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, Vol. 13, No. 3, pp. 223-241, 1999.

[35] R.E. Donovan. Segment pre-selection in decision-tree based speech synthesis systems. *Proc. ICASSP*, pp. 937–940, Istanbul, Turkey, June 2000.

[36] H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169-177, 1939.

[37] K.E. Dusterhoff and A.W. Black. Generating F_0 contours for speech synthesis using the Tilt intonation theory. *Proc. ESCA Workshop on Intonation*, pp. 107–110, Athens, Greece, Sep. 1997.

[38] K.E. Dusterhoff, A.W. Black, and P. Taylor. Using decision trees within the Tilt intonation model to predict F_0 contours. *Proc. EUROSPEECH*, Budapest, Hungary, pp. 1627–1630, Sep. 1999.

[39] H. Fujisaki, K. Hirose. Analysis of voice fundamental frequency contours for declarative sentence of Japanese. *J. Acoust. Soc. Jpn. (E)*, Vol. 5, No. 4, pp. 233–242, 1984.

[40] K. Hakoda, S. Nakajima, T. Hirokawa, and H. Mizuno. A new Japanese text-to-speech synthesizer based on COC synthesis method. *Proc. ICSLP*, pp. 809–812, Kobe, Japan, Nov. 1990.

[41] T. Hirai, N. Iwahashi, N. Higuchi, and Y. Sagisaka. Automatic extraction of F_0 control rules using statistical analysis. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 333–346, 1997.

[42] T. Hirai, S. Tenpaku, and K. Shikano. Speech unit selection based on target values driven by speech data in concatenative speech synthesis. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[43] T. Hirokawa. Speech synthesis using a waveform dictionary. *Proc. EUROSPEECH*, pp. 140–143, Paris, France, Sep. 1989.

[44] T. Hirokawa and K. Hakoda. Segment selection and pitch modification for high quality speech synthesis using waveform segments. *Proc. ICSLP*, pp. 337–340, Kobe, Japan, Nov. 1990.

[45] K. Hirose, H. Fujisaki, and H. Kawai. Generation of prosodic symbols for rule-synthesis of connected speech of Japanese. *Proc. ICASSP*, pp. 2415–2418, Tokyo, Japan, Apr. 1986.

[46] K. Hirose, M. Eto, and N. Minematsu. Improved corpus-based synthesis of fundamental frequency contours using generation process model. *Proc. ICSLP*, pp. 2085–2088, Denver, U.S.A., Sep. 2002.

[47] A.J. Hunt and A.W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. ICASSP*, pp. 373–376, Atlanta, U.S.A., May 1996.

[48] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura. A speech synthesis system with emotion for assisting communication. *Proc. ISCA*

Workshop on Speech and Emotion, pp. 167–172, Belfast, Northern Ireland, Sep. 2000.

[49] S. Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (MLSA) filter for speech synthesis. *IEICE Trans.*, Vol. J66-A, No. 2, pp. 122–129, 1983 (in Japanese).

[50] M. Isogai and H. Mizuno. A new F_0 contour control method based on vector representation of F_0 contour. *Proc. EUROSPEECH*, pp. 727–730, Budapest, Hungary, Sep. 1999.

[51] K. Itoh, S. Nakajima, and T. Hirokawa. A new waveform speech synthesis approach based on the COC speech spectrum. *Proc. ICASSP*, pp. 577–580, Adelaide, Australia, Apr. 1994.

[52] N. Iwahashi, N. Kaiki, and Y. Sagisaka. Speech segment selection for concatenative synthesis based on spectral distortion minimization. *IEICE Trans. Fundamentals*, Vol. E76-A, No. 11, pp. 1942–1948, 1993.

[53] N. Iwahashi and Y. Sagisaka. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. *Speech Communication*, Vol. 16, No. 2, pp. 139–151, 1995.

[54] K. Iwano, M. Yamada, T. Togawa, and S. Furui, Speech-rate-variable HMM-based Japanese TTS system. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[55] T. Kagoshima and M. Akamine. Automatic generation of speech synthesis units based on closed loop training. *Proc. ICASSP*, pp. 963–966, Munich, Germany, Apr. 1997.

[56] T. Kagoshima and M. Akamine. An F_0 contour control model for totally speaker driven text to speech system. *Proc. ICSLP*, pp. 1975–1978, Sydney, Australia, Dec. 1998.

[57] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. ICASSP*, Seattle, U.S.A., pp. 285–288, May 1998.

[58] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[59] H. Kawahara, H. Katayose, A.de Cheveigné, and R.D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity. *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sep. 1999.

[60] H. Kawai, N. Higuchi, T. Shimizu, and S. Yamamoto. Development of a text-to-speech system for Japanese based on waveform splicing. *Proc. ICASSP*, pp. 569–572, Adelaide, Australia, Apr. 1994.

[61] H. Kawai, K. Hirose, and H. Fujisaki. Rules for generating prosodic features for text-to-speech synthesis of Japanese. *J. Acoust. Soc. Jpn. (J)*, Vol. 50, No. 6, pp. 433–442, 1994 (in Japanese).

[62] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu. A design method of speech corpus for text-to-speech synthesis taking account of prosody. *Proc. ICSLP*, Vol. 3, pp. 420–425, Beijing, China, Oct. 2000.

[63] H. Kawai and M. Tsuzaki. A study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[64] H. Kawai and M. Tsuzaki. Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis. *Proc. ICSLP*, pp. 2621–2624, Denver, U.S.A., Sep. 2002.

[65] H. Kawanami, T. Masuda, T. Toda, and K. Shikano. Designing Japanese speech database covering wide range in prosody. *Proc. ICSLP*, pp. 2425–2428, Denver, U.S.A., Sep. 2002.

[66] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 39–51, 2001.

[67] D.H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, Vol. 82, No. 3, pp. 737–793, 1987.

[68] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication*, Vol. 16, No. 2, pp. 165–173, 1995.

[69] C.H. Lee, C.H. Lin, and B.H. Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Trans. Signal Processing*, Vol. 39, No. 4, pp. 806–814, 1991.

[70] M. Lee. Perceptual cost functions for unit searching in large corpus-based concatenative text-to-speech. *Proc. EUROSPEECH*, pp. 2227–2230, Aalborg, Denmark, Sep. 2001.

[71] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, Vol. 9, No. 2, pp. 171–185, 1995.

[72] N. Maeda, H. Banno, S. Kajita, K. Takeda, and F. Itakura. Speaker conversion through non-linear frequency warping of STRAIGHT spectrum. *Proc. EUROSPEECH*, Budapest, Hungary, pp. 827–830, Sep. 1999.

[73] M. Mashimo, T. Toda, H. Kawanami. K. Shikano, and N. Campbell. Cross-language voice conversion evaluation using bilingual databases. *IPSJ Journal*, Vol. 43, No. 7, pp. 2177–2185, 2002.

[74] T. Masuko, K. Tokuda, T. Kobayashi, and Satoshi Imai. Voice characteristics conversion for HMM-based speech synthesis system. *Proc. ICASSP*, pp.1611–1614, Munich, Germany, Apr. 1997.

[75] H. Matsumoto and H. Inoue. A minimum distortion spectral mapping applied to voice quality conversion. *Proc. ICSLP*, pp. 161–164, Kobe, Japan, Nov. 1990.

[76] H. Matsumoto and Y. Yamashita. Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function. *J. Acoust. Soc. Jpn. (E)*, Vol. 14, No. 5, pp. 353–361, 1993.

[77] N. Minematsu, R. Kita, and K. Hirose. Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[78] H. Mizuno and M. Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Communication*, Vol. 16, No. 2, pp. 153–164, 1995.

[79] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, Vol. 9, No. 5–6, pp. 453–467, 1990.

[80] S. Nakajima and H. Hamada. Automatic generation of synthesis units based on context oriented clustering. *Proc. ICASSP*, pp. 659–662, New York, U.S.A., Apr. 1988.

[81] S. Nakamura and K. Shikano. Speaker adaptation applied to HMM and neural networks. *Proc. ICASSP*, pp. 89–92, Glasgow, Scotland, May 1989.

[82] M. Narendranath, H.A. Murthy, S. Rajendran, B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, Vol. 16, No. 2, pp. 207–216, 1995.

[83] K. Ohkura, M. Sugiyama, and S. Sagayama. Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs. *Proc. ICSLP*, pp. 369–372, Banff, Canada, Oct. 1992.

[84] J.P. Olive. Rule synthesis of speech from diadic units. *Proc. ICASSP*, pp. 568–570, Hartford, U.S.A., May. 1977.

[85] A.V. Oppenheim and D.H. Johnson. Discrete representation of signals. *Proc. IEEE*, Vol. 60, No. 6, pp. 681–691, 1972.

[86] H. Peng, Y. Zhao, and M. Chu. Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation. *Proc. ICSLP*, pp. 2613–2616, Denver, U.S.A., Sep. 2002.

[87] G. Peterson, W. Wang, and E. Silvertsen. Segmentation techniques in speech synthesis. *J. Acoust. Soc. Am.*, Vol. 30, No. 8, pp. 739–742, 1958.

[88] Y. Sagisaka and H. Satoh. Accentuation rules for Japanese word concatenation. *IEICE Trans.*, Vol. J66-D, No. 7, pp. 849–856, 1983 (in Japanese).

[89] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proc. ICASSP*, pp. 679–682, New York, U.S.A., Apr. 1988.

[90] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR ν-talk speech synthesis system. *Proc. ICSLP*, pp. 483–486, Banff, Canada, Oct. 1992.

[91] Y. Sagisaka, Natural language processing in speech synthesis. *IPSJ Journal*, Vol. 34, No. 10, pp. 1281–1286, 1993 (in Japanese).

[92] H. Sato. Speech synthesis on the basis of PARCOR-VCV concatenation units. *IEICE Trans.*, Vol. J61-D, No. 11, pp. 858–865, 1978 (in Japanese).

[93] H. Sato. Speech synthesis using CVC concatenation units and excitation waveform elements. *Trans. Comm. Speech Res., Acoust. Soc. Jpn.*, S83-69, pp. 541–546, 1984 (in Japanese).

[94] Y. Shi, E. Chang, H. Peng, and M. Chu. Power spectral density based channel equalization of large speech database for concatenative TTS system. *Proc. ICSLP*, pp. 2369–2372, Denver, U.S.A., Sep. 2002.

[95] K. Shikano, K.F. Lee, and R. Reddy. Speaker adaptation through vector quantization. *Proc. ICASSP*, pp. 2643–2646, Tokyo, Japan, Apr. 1986.

[96] T. Shimizu, N. Higuchi, H. Kawai, and S. Yamamoto. A morphological analyzer for a Japanese text-to-speech system based on the strength of connection between words. *J. Acoust. Soc. Jpn. (J)*, Vol. 51, No. 1, pp. 3–13, 1995 (in Japanese).

[97] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labeling English prosody. *Proc. ICSLP*, pp. 867–870, Banff, Canada, Oct. 1992.

[98] Y. Stylianou, O. Cappé, and E. Moulines. Statistical methods for voice quality transformation. *Proc. EUROSPEECH*, Madrid, Spain, pp. 447–450, Sep. 1995.

[99] Y. Stylianou, and O. Cappé. A system voice conversion based on probabilistic classification and a harmonic plus noise model. *Proc. ICASSP*, Seattle, U.S.A., pp. 281–284, May 1998.

[100] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 21–29, 2001.

[101] Y. Stylianou and A.K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. ICASSP*, pp. 837–840, Salt Lake City, U.S.A., May 2001.

[102] A.K. Syrdal, C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay. Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. ICSLP*, Vol. 3, pp. 410–415, Beijing, China, Oct. 2000.

[103] A.K. Syrdal. Phonetic effects on listener detection of vowel concatenation. *Proc. EUROSPEECH*, pp. 979–982, Aalborg, Denmark, Sep. 2001.

[104] J. Takahashi and S. Sagayama. Vector-field-smoothed Bayesian learning for incremental speaker adaptation. *Proc. ICASSP*, pp. 696–699, Detroit, U.S.A., May 1995.

[105] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima. A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 3–10, 2001.

[106] K. Takeda, K. Abe, and Y. Sagisaka. On the basic scheme and algorithms in non-uniform unit speech synthesis. *Talking machines: theories, models, and designs*, G. Bailly, C. Benoit, and T. Sawallis, North-Holland, Elsevier Science, pp. 93–105, 1992.

[107] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. *Proc. ICASSP*, pp. 805–808, Salt Lake City, U.S.A., May 2001.

[108] P. Taylor and A.W. Black. Synthesizing conversational intonation from a linguistically rich input. *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis*, pp. 175–178, New York, U.S.A., Sep. 1994.

[109] T. Toda, J. Lu, H. Saruwatari, and K. Shikano. STRAIGHT-based voice conversion algorithm based on Gaussian mixture model. *Proc. ICSLP*, Vol. 3, pp. 279–282, Beijing, China, Oct. 2000.

[110] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. *Proc. ICASSP*, pp. 841–844, Salt Lake City, U.S.A., May 2001.

[111] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. *Proc. ICASSP*, pp. 229–232, Phoenix, U.S.A., May 1999.

[112] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[113] M. Tsuzaki and H. Kawai. Feature extraction for unit selection in concatenative speech synthesis: comparison between AIM, LPC, and MFCC. *Proc. ICSLP*, pp. 137–140, Denver, U.S.A., Sep. 2002.

[114] H. Valbret, E. Moulines and J.P. Tubach. Voice transformation using PSOLA technique. *Speech Communication*, Vol. 11, No. 2–3, pp. 175–187, 1992.

[115] J. Wouters and M.W. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis. *Proc. ICSLP*, pp. 2747–2750, Sydney, Australia, Dec. 1998.

[116] J. Wouters and M.W. Macon. Control of spectral dynamics in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp. 30–38, 2001.

[117] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *Proc. EUROSPEECH*, pp. 2347–2350, Budapest, Hungary, Sep. 1999.

[118] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Speaker interpolation for HMM-based speech synthesis system. *J. Acoust. Soc. Jpn. (E)*, Vol. 21, No. 4, pp. 199–206, 2000.

List of Publications

Journal Papers

1. T. Toda, H. Banno, S. Kajita, K. Takeda, F. Itakura, and K. Shikano. Improvement of STRAIGHT method under noisy conditions based on lateral inhibitive weighting. *IEICE Trans.*, Vol. J83-D-II, No. 11, pp. 2180–2189, Nov. 2000 (in Japanese).

2. T. Toda, J. Lu, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping. *IEICE Trans.*, Vol. J84-D-II, No. 10, pp. 2181–2189, Oct. 2001 (in Japanese).

3. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. A segment selection algorithm for Japanese concatenative speech synthesis based on both phoneme unit and diphone unit. *IEICE Trans.*, Vol. J85-D-II, No. 12, pp. 1760–1770, Dec. 2002 (in Japanese).

4. M. Mashimo, T. Toda, H. Kawanami. K. Shikano, and N. Campbell. Cross-language voice conversion evaluation using bilingual databases. *IPSJ Journal*, Vol. 43, No. 7, pp. 2177–2185, July 2002.

International Conferences

1. T. Toda, J. Lu, H. Saruwatari, and K. Shikano. STRAIGHT-based voice conversion algorithm based on Gaussian mixture model. *Proc. ICSLP*, Vol. 3, pp. 279–282, Beijing, China, Oct. 2000.

2. T. Toda, J. Lu, S. Nakamura, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model applied to STRAIGHT. *Proc. WESTPRAC VII*, pp. 169–172, Kumamoto, Japan, Oct. 2000.

3. T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. *Proc. ICASSP*, pp. 841–844, Salt Lake City, U.S.A., May 2001.

4. T. Toda, H. Saruwatari, and K. Shikano. High quality voice conversion based on Gaussian mixture model with dynamic frequency warping. *Proc. EUROSPEECH*, pp. 349–352, Aalborg, Denmark, Sep. 2001.

5. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit. *Proc. ICASSP*, pp. 465–468, Orlando, U.S.A., May 2002.

6. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Perceptual evaluation of cost for segment selection in concatenative speech synthesis. *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

7. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Segment Selection Considering Local Degradation of Naturalness in Concatenative Speech Synthesis. *Proc. ICASSP*, Hong Kong, China, Apr. 2003 (accepted).

8. M. Mashimo, T. Toda, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. *Proc. EUROSPEECH*, pp. 361–364, Aalborg, Denmark, Sep. 2001.

9. H. Kawanami, T. Masuda, T. Toda, and K. Shikano. Designing speech database with prosodic variety for expressive TTS system. *Proc. LREC*, pp. 2039–2042, Las Palmas, Spain, May 2002.

10. M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion using bilingual and non-bilingual databases. *Proc. ICSLP*, pp. 293–296, Denver, U.S.A., Sep. 2002.

11. H. Kawanami, T. Masuda, T. Toda, and K. Shikano. Designing Japanese speech database covering wide range in prosody. *Proc. ICSLP*, pp. 2425–2428, Denver, U.S.A., Sep. 2002.

Technical Reports

1. T. Toda, H. Banno, S. Kajita, K. Takeda, and F. Itakura. Improvement of STRAIGHT method under noisy conditions based on lateral inhibitive weighting. *IEICE Tech. Rep.*, EA99-7, pp. 25–32, Aichi, May 1999 (in Japanese).

2. T. Toda, J. Lu, S. Nakamura, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model applied to STRAIGHT. *IEICE Tech. Rep.*, SP2000-7, pp. 49–54, Kyoto, May 2000 (in Japanese).

3. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. A unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit. *IEICE Tech. Rep.*, SP2001-120, pp. 45–52, Shiga, Jan. 2002 (in Japanese).

4. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Perceptual evaluation of cost for segment selection in concatenative text-to-speech synthesis. *IEICE Tech. Rep.*, SP2002-69, pp. 19–24, Fukuoka, Aug. 2002 (in Japanese).

5. T. Masuda, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. A study on the speech synthesis method by using database with variety of speech-rate. *IEICE Tech. Rep.*, SP2001-122, pp. 61–68, Shiga, Jan. 2002 (in Japanese).

6. T. Shiraishi, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. A visual speech synthesis method using real image database. *IPSJ SIG Notes*, 2002-SLP-42, pp. 23–29, Gunma, July 2002 (in Japanese).

7. H. Kawai and T. Toda. An evaluation of automatic phoneme segmentation for concatenative speech synthesis. *IEICE Tech. Rep.*, SP2002-170, pp. 5–10, Shiga, Jan. 2003 (in Japanese).

8. Y. Iwami, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. Emotional speech synthesis using GMM-based voice conversion technique. *IEICE Tech. Rep.*, SP2002-171, pp. 11–16, Shiga, Jan. 2003 (in Japanese).

Meetings

1. T. Toda, J. Lu, S. Nakamura, and K. Shikano. Application of voice conversion algorithm based on Gaussian mixture model to STRAIGHT. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 1-7-6, pp. 207–208, Chiba, Mar. 2000 (in Japanese).

2. T. Toda, J. Lu, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian Mixture Model with dynamic frequency warping. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 2-1-6, pp. 185–186, Iwate, Sep. 2000 (in Japanese).

3. T. Toda, H. Saruwatari, and K. Shikano. Study on conversion-accuracy on speaker individuality of voice conversion algorithm with dynamic frequency warping. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 1-6-5, pp. 237–238, Ibaraki, Mar. 2001 (in Japanese).

4. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. A unit selection algorithm for Japanese speech synthesis allowing concatenation at vowel center. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 2-10-2, pp. 267–268, Kanagawa, Mar. 2002 (in Japanese).

5. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Evaluation based on perceptual characteristic of cost for segment selection in concatenative speech synthesis. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 1-10-6, pp. 235–236, Akita, Sep. 2002 (in Japanese).

6. T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. Evaluation of segment selection considering local degradation of naturalness in concatenative speech synthesis. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 1-6-13, pp. 247–248, Tokyo, Mar. 2003 (in Japanese).

7. T. Masuda, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. STRAIGHT-based prosody modification of CHATR output. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 1-2-20, pp. 245–246, Ohita, Oct. 2001 (in Japanese).

8. M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and Nick Campbell. Adaptation of voice conversion method based on GMM and STRAIGHT to cross-language speech synthesis. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 1-P-17, pp. 389–390, Ohita, Oct. 2001 (in Japanese).

9. M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and Nick Campbell. Cross-language voice conversion and vowel pronunciation. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 1-10-16, pp. 261–262, Kanagawa, Mar. 2002 (in Japanese).

10. T. Masuda, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. Designing and evaluation of speech database with prosodic variety. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 2-10-13, pp. 289–290, Kanagawa, Mar. 2002 (in Japanese).

11. Y. Iwami, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. Evaluation of emotional control with GMM-based voice conversion. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 1-10-24, pp. 277–278, Akita, Sep. 2002 (in Japanese).

12. T. Shiraishi, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. A visual speech synthesis method using real image corpus. *Proc. Fall Meeting, Acoust. Soc. Jpn.*, 2-10-15, pp. 315–316, Akita, Sep. 2002 (in Japanese).

13. H. Kawai and T. Toda. An evaluation of a design method of speech corpus for concatenative speech synthesis. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 1-6-20, pp. 261–262, Tokyo, Mar. 2003 (in Japanese).

14. Y. Iwami, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. Emotion control of speech using GMM-based voice conversion. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 1-6-23, pp. 267–268, Tokyo, Mar. 2003 (in Japanese).

15. T. Shiraishi, T. Toda, H. Kawanami, H. Saruwatari, and K. Shikano. Corpus-based visual speech synthesis using perceptual definition of viseme. *Proc. Spring Meeting, Acoust. Soc. Jpn.*, 2-Q-13, pp. 399–400, Tokyo, Mar. 2003 (in Japanese).

Master's Thesis

1. T. Toda. High quality voice conversion based on STRAIGHT analysis-synthesis method. *Master's thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology*, NAIST-IS-MT9951077, Feb. 2001 (in Japanese).