NAIST-IS-DD0761009

Doctoral Thesis

Techniques for Improving Voice Conversion Based on Eigenvoices

Yamato Ohtani

March 31, 2010

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

A 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

Yamato Ohtani

Thesis Committee:

Professor Kiyohiro Shikano	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Associate Professor Hiroshi Saruwatari	(Co-supervisor)
Assistant Professor Tomoki Toda	(Co-supervisor)

Techniques for Improving Voice Conversion Based on Eigenvoices*

Yamato Ohtani

Abstract

Voice conversion (VC) is a technique for converting a source speaker's voice into another speaker's voice without changing linguistic information. As a typical approach to VC, a statistical method based on the Gaussian mixture model (GMM) is widely used. A GMM is trained as a conversion model using a parallel data set composed of many utterance-pairs of source and target speakers. Although this framework works reasonably well, the converted speech quality is still insufficient and the training process of the conversion model is less flexible.

Eigenvoice conversion (EVC) is an effective method for making the training process more flexible. An eigenvoice GMM (EV-GMM) is trained in advance with multiple parallel data sets consisting of the single pre-defined speaker and many pre-stored speakers. Then, a conversion model for a new speaker is flexibly built by adapting the EV-GMM using a few arbitrary utterances of the new speaker. Two main frameworks have been proposed based on EVC: 1) one-to-many EVC, which allows the conversion from a single source speaker's voice into an arbitrary target speaker's voice; and 2) many-to-one EVC, which allows the conversion in reverse. Although these frameworks achieve much higher flexibility than the traditional VC, there remain limitations in building the conversion model between an arbitrary speaker-pair. In addition, the conversion performance of the EVC is significantly degraded because the EV-GMM captures acoustic variations among the pre-stored target speakers. To make VC applications more practical, it is indispensable to improve the conversion performance and develop a more flexible training framework.

*Doctoral Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0761009, March 31, 2010.

This thesis addresses two issues in the traditional methods: 1) the insufficient converted speech quality and 2) the insufficient flexibility in model training. To address the former issue, we first improve the excitation modeling accuracy in the traditional VC framework. The traditional VC employs the simple excitation model based on selecting a phase-manipulated pulse train and a noise signal. This excitation model is too simple to capture acoustic characteristics of the excitation signal. To address this issue, we introduce a more precise excitation model, i.e., STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) mixed excitation (STME), which is generated by frequency-dependent-weighted sum of a phase-manipulated pulse train and a noise signal. Then, we also improve the conversion performance of the EV-GMM. The inferior conversion performance is caused by the inter-speaker acoustic variations captured by the EV-GMM. To alleviate this problem, we propose an adaptive training method for the EV-GMM to effectively reduce the inter-speaker acoustic variations. Moreover, we develop an enhanced EVC system by integrating the above proposed methods and the conversion algorithm considering global variance (GV) into the conventional EVC system. The experimental results demonstrate that each proposed method yields significant improvements in the conversion performance, and the enhanced EVC system dramatically outperforms the conventional one in terms of both converted speech quality and conversion accuracy for speaker individuality.

To address the latter issue, we propose many-to-many EVC for achieving a very flexible training process of the conversion model. Many-to-many EVC is a technique for converting an arbitrary source speakers' voice into an arbitrary target speaker's voice. This framework is achieved by performing many-to-one EVC and one-to-many EVC sequentially with a single EV-GMM through a reference speaker's voice, which is considered as a hidden variable. Moreover, we also propose a refining method of the EV-GMM using non-parallel data sets by extending the many-to-many EVC method. The experimental results demonstrate the effectiveness of these proposed methods.

Keywords:

speech synthesis, voice conversion, Gaussian mixture model, eigenvoice, many-to-many

Acknowledgements

I deeply appreciate the help of my main thesis adviser, Professor Kiyohiro Shikano of Nara Institute of Science and Technology for his constant guidance and encouragement throughout my master's course and doctoral course. I would also like to express my gratitude to Professor Yuji Matsumoto and Associate Professor Hiroshi Saruwatari of Nara Institute of Science and Technology for their invaluable comments to the thesis.

Especially, I would like to express my appreciation to Assistant Professor Tomoki Toda of Nara Institute of Science and Technology, for his continuous support and valuable advices through the master and doctor course. The core of my work was originated with his ideas and he let me suggest research ideas. All of my work could not have been achieved without his direction. I have enjoyed implementing my research with him.

I would sincerely like to thank Dr. Satoshi Nakamura, who is currently a leader of NICT Spoken Language Communication Group, for giving me the opportunity to work for ATR Spoken Language Communication Research Laboratories as an Intern Researcher.

I an very grateful Dr. Tatsuo Yotsukura, who is currently a software engineer at OLM Digital, Inc. Research and Development Division, Dr. Shin-ich Kawamoto, Dr. Seigo Enomoto and Dr. Yusuke Ikeda who is currently a researcher at NICT Spoken Language Communication Group, for their valuable advices on my work in ATR.

I would like to thank Professor Shigeo Morishima of Waseda University and Professor Yasushi Yagi of Osaka University, for their valuable comments on my work in ATR.

I would like to thank Assistant Professor Hiromichi Kawanami of Nara Insti-

tute of Science and Technology, for his beneficial comments.

Finally, I would like to acknowledge the support of my family and friends.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. General background and problem definition

Speech is one of the principal ways for people to communicate. Speech conveys not only textual information but also emotion, a speaking style, speaker individuality, and so on. Therefore, speech plays an important role in human communication.

Recently, our access to computers has increased with the development of computer technology. It is necessary to develop a useful man-machine interface to support communication between people and computers. As one of the methods of achieving this, speech interfaces have been studied for several decades. There are two important technologies for developing man-machine speech interfaces, i.e., speech recognition and speech synthesis. Speech recognition is a technique for information input. In this technology, the textual information is extracted from speech, and it is converted into data that a computer can understand. On the other hand, speech synthesis is a technique for information output. This process is the reverse of speech recognition. Output speech includes diverse information such as linguistic content, prosody, emotion, and speaker individuality. These technologies are essential for developing a more natural and usable communication system between people and computers.

In this thesis, we focus on voice conversion (VC) [1], which is one of the speech synthesis techniques. VC is a technique for converting a source speaker's voice quality into another speaker's voice quality without changing linguistic information. There have been various proposed applications using this technique, e.g.,

cross-language conversion [2][3], bandwidth extension for mobile phones [4][5] and conversion from body-conducted speech to air-conducted speech [6] with a non-audible murmur microphone [7].

In recent years, a statistical method using the Gaussian mixture model (GMM) as a conversion model has been widely used [8] in the VC framework. In this method, joint probability density of source and target acoustic features is modeled by a GMM [9]. The GMM is trained with a parallel data set consisting of utterance-pairs of source and target speakers. The trained GMM allows us to convert the source features into the target features based on minimum mean square error (MMSE) [8] or the maximum likelihood (ML) criterion [10]. Although the traditional VC framework works reasonably well, it is difficult to develop practical VC applications because there are still many problems to be solved.

The insufficient converted speech quality is caused by the improper excitation model, and the training process of the conversion model is less flexible due to the use of a large amount of parallel data.

To ameliorate the lower flexibility of the conversion model training, eigenvoice conversion (EVC) has been proposed [11]. EVC is one effective approach to using voices of other speakers as prior knowledge for building a conversion model for a new speaker. This method has brought novel VC frameworks, i.e., one-to-many EVC and many-to-one EVC [12]. The one-to-many EVC framework allows the conversion from a specific source speaker's voice into an arbitrary target speaker's voice and many-to-one EVC allows the conversion in reverse. In the one-to-many EVC framework, an eigenvoice GMM (EV-GMM) is trained in advance with multiple parallel data sets consisting of a single source speaker and many pre-stored target speakers. The GMM between the source speaker and an arbitrary target speaker is flexibly developed by estimating a small number of free parameters of the EV-GMM, i.e., weights for eigenvectors, using only a few utterances from the adapted speaker in a text-independent manner. Although these EVC frameworks are very flexible compared to the conventional VC framework, there are still some issues to be addressed as follows: 1) it is still hard to flexibly perform the conversion between arbitrary speaker-pairs; 2) the converted speech quality of the EVC is still not high enough; and 3) the use of parallel data is inevitable for the EV-GMM training. Therefore, it is necessary to work out these problems.

1.2. Thesis scope

This thesis describes two main approaches to addressing the issues of the conventional EVC framework as shown in Figure 1.1. One is an approach to improving the converted speech quality and the other is an approach to improving the flexibility of building the conversion model.

1.2.1 Improvement of converted speech quality

The converted speech quality is affected by the following elements: 1) the quality of the excitation model, 2) the performance of the conversion algorithm, and 3) the quality of the conversion model. The excitation model strongly affects the converted speech quality. The STRAIGHT (Speech Transformation and Representation of weiGHTed spectrum) simple excitation (STSE) model based on switching a phase-manipulated pulse train and white noise [13] is often used but is too simple to model the human excitation signal appropriately. Moreover, the spectral conversion algorithm without considering global variance (GV) [10] employed in the conventional EVC frameworks usually makes the converted spectral parameters over-smoothed. The use of the EV-GMM with the target-speaker-independent GMM (TI-GMM) parameters [11] also causes quality degradation of the adapted conversion model because it improperly captures acoustic variations among many pre-stored target speakers. These techniques often make the converted speech sound buzzy and muffled. In this thesis, three effective techniques for addressing these issues are applied to the standard VC framework and the EVC framework in order to improve the converted speech quality. Experimental results of objective and subjective evaluations demonstrate that the proposed techniques yield significant quality improvements in the converted speech.

1.2.2 Improvement of flexibility of model training

The conventional EVC framework can perform only one-to-many VC, which is the conversion from a specific source speaker to arbitrary target speakers, or many-to-one VC, which is the conversion from arbitrary source speakers to a specific target speaker. This is because we use parallel data sets between a specified

speaker and various pre-stored speakers in the EV-GMM training. In order to achieve many-to-many VC capable of the conversion between arbitrary source and target speakers, we propose an approach based on many-to-one EVC and one-to-many EVC. In our proposed approach, many-to-one EVC and one-to-many EVC are performed sequentially using a single EV-GMM. In addition, in order to make many-to-many EVC more effective, we consider voices of a single reference speaker (i.e., the target speaker in many-to-one EVC and the source speaker in one-to-many EVC) as hidden variables.

We also propose a method for refining the canonical EV-GMM using various non-parallel data sets by developing the basic idea of the many-to-many EVC. In this method, the initial EV-GMM is trained using the existing multiple parallel data sets, and then it is refined using only non-parallel data sets including a larger number of speakers while considering speech data of the single reference speaker in the existing multiple parallel data sets as hidden variables. Note that these non-parallel data sets are much more easily available than the multiple parallel data sets. Therefore, the proposed method allows us to extract more informative prior knowledge from a much larger number of speakers in EV-GMM training. The experimental results of objective and subjective evaluations demonstrate that our proposed methods are very effective.

1.3. Thesis overview

This thesis is organized as follows.

In Chapter 2, the traditional VC frameworks based on the statistical approaches are described. We describe state-of-the-art conversion methods for the standard VC framework. We also describe the EVC framework and review the conventional one-to-many EVC system. Finally, we describe the problems of the conventional one-to-many EVC framework.

In Chapter 3, we address the improvement of the excitation model in the traditional VC framework. As the proposed excitation model, we introduce STRAIGHT mixed excitation (STME) [14] to the standard VC framework based on a GMM. Objective and subjective results show that the proposed method significantly outperforms the conventional method.

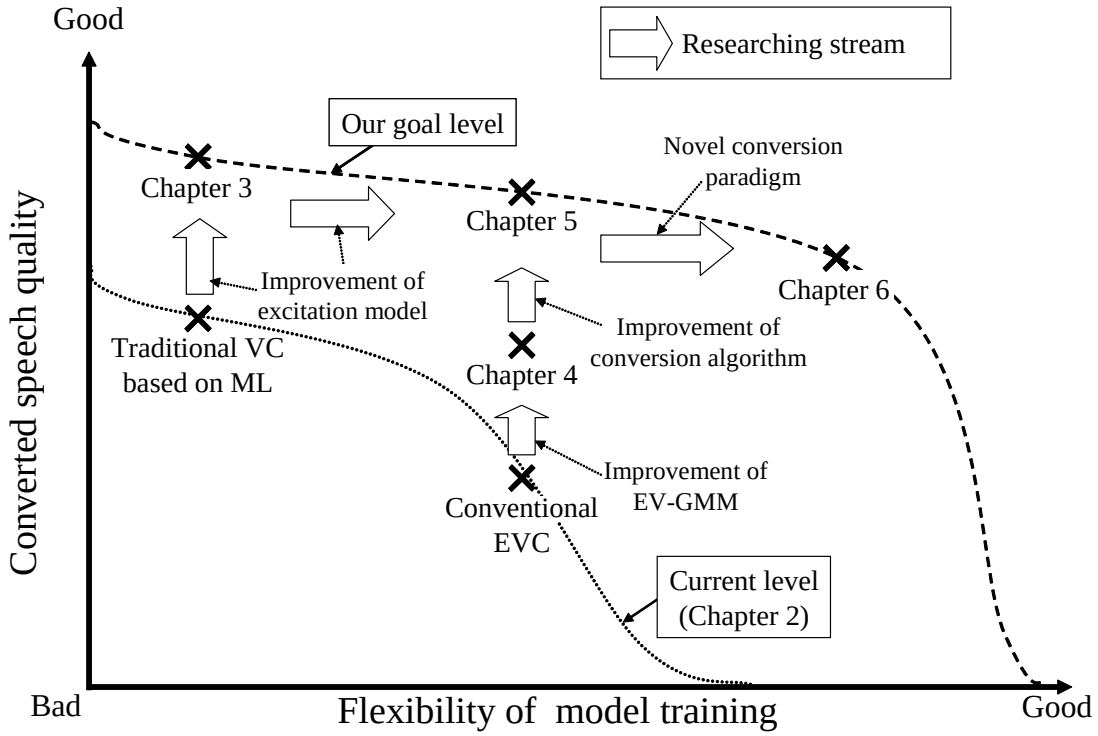


Figure 1.1. Overview of thesis scope.

In Chapter 4, we propose the adaptive training method of the EV-GMM. This proposed method is based on speaker adaptive training (SAT) [15], which has been developed for speech recognition. Moreover, we describe a problem of the proposed adaptive training method and present methods for alleviating it. Experimental results demonstrate that the proposed training method yields quality improvements of the adapted conversion model in EVC.

In Chapter 5, we describe the proposed one-to-many system into which some promising techniques such as the state-of-the-art conversion method, STME and adaptive training of the EV-GMM are integrated. Experimental results of objective and subjective evaluations demonstrate that our proposed system outperforms the conventional system.

In Chapter 6, we propose a novel EVC framework, many-to-many EVC. In addition, we describe the proposed method for refining the EV-GMM using non-

parallel data sets. Objective and subjective results demonstrate the effectiveness of our proposed methods.

In Chapter 7, we summarize the contributions of this thesis and suggest future work.

Chapter 2

Traditional Techniques for Voice Conversion

Voice conversion (VC) is a technique that allows us to convert a source speaker's voice into another speaker's voice. Currently, VC frameworks based on a statistical model are studied widely. Although the traditional frameworks work reasonably well, they have limitations of the conversion model training. To ameliorate these limitations, eigenvoice conversion (EVC) has been proposed as one of the VC frameworks using model adaptation techniques. In this chapter, we review the traditional VC techniques and present some problems to be solved.

2.1. Introduction

VC is a technique for modifying non-linguistic information such as voice characteristics without changing linguistic information. Two main approaches have been studied. One is the rule-based approach and the other is the statistical approach. In the rule-based approach, voice quality is changed by directly modifying acoustic parameters such as spectral envelope and fundamental frequency. As a typical application, voice morphing [16][17][18] has been proposed. This technique can generate intermediate voices among some different speakers' voices by linear interpolation of the acoustic parameters of those speakers. Moreover, the extended voice morphing method to generate a specified speaker's voice has been proposed [19][20][21]. In this approach, we need to determine rules for mod-

ifying the acoustic parameters but it is difficult to find proper and generic rules. Therefore, the performance of the rule-based framework depends on developers' abilities.

On the other hand, in the statistical approach, voice quality is changed based on the statistical model constructed with a large amount of speech data. In the early statistical VC approach, the codebook mapping method was proposed by Abe et al. [1]. This technique is based on a speaker adaptation method using vector quantization (VQ) [22]. In this method, the mapping codebook representing the correspondence between source and target speakers' codebooks is constructed. Then the source speaker's voice is converted into the target speaker's voice with that mapping codebook. As another VC framework using a statistical model, Kim et al. have proposed VC using the hidden Markov model (HMM) [23]. In this framework, we train the target speaker's HMM and the mapping function of two state-dependent codebooks [24] in advance. These are called recognition-codebook and synthesis-codebook respectively. In the conversion process, the recognition-codewords are generated from the source feature sequence with the recognition-codebook, and a state sequence for the target speaker's HMM is determined by translating these recognition-codewords into the synthesis-codewords with the mapping function. Then, the converted feature sequence is generated from the target speaker's HMM based on this state sequence.

In recent years, VC using the Gaussian mixture model (GMM) has been proposed by Stylianou et al. [8]. In this framework, the GMM is trained with a source speaker's acoustic features. The conversion function from the source features to the target features is determined based on the minimum mean square error (MMSE) criterion [8] with the source utterance set and its counterpart set of the target speaker. In the conversion process, arbitrary utterances of the source speaker's voice are converted into those of the target speaker's voice frame-by-frame with the trained conversion function. Kain et al. have proposed an improved GMM training method [9]. In their proposed method, a GMM of joint probability density of source and target acoustic features is trained using a parallel data set consisting of utterance-pairs of the source and target speakers. The conversion function constructed from the trained GMM can transform the source features into the target features based on the MMSE criterion. Toda

et al. have applied STRAIGHT (Speech Transformation and Representation of weiGHTed spectrum) [13] to the GMM-based VC [25] to improve the converted speech quality. STRAIGHT is a high quality analysis-synthesis system. In this method, a GMM is trained with acoustic features analyzed by STRAIGHT.

The GMM-based VC using the MMSE criterion has two essential problems. One is inappropriate spectral movements often caused by the MMSE-based conversion because inter-frame feature correlation is ignored in the conversion process. The other is the over-smoothing of the converted acoustic feature sequence because the statistical modeling often removes the detail of spectral structures. In order to consider inter-frame correlation of the converted features, Toda et al. have proposed a conversion method based on the maximum likelihood criterion [10] inspired by the parameter generation algorithm for speech synthesis based on the HMM [26][27]. In this conversion method, an acoustic feature sequence is converted considering dynamic features [28][29][30]. In order to alleviate the over-smoothing of the converted features, Toda et al. have also proposed an ML-based conversion method considering global variance (GV) [10]. The GV is defined as a variance over a time sequence of the acoustic features. These two ideas achieve a significant quality improvement of the converted acoustic features.

Although the GMM-based VC framework works reasonably well, this training framework using the parallel data set causes many limitations on VC applications due to its lower flexibility. To relax the use of the parallel data set, several approaches to flexibly building a GMM for a desired speaker-pair by effectively using another GMM for a different speaker-pair have been studied. Mouchtaris et al. [31] have proposed an unsupervised training method based on ML constrained adaptation [32]. In this training method, a previously trained GMM between certain source and target speakers is adapted to different source and target speakers with linear transformation for the source speaker and that for the target speaker, which are estimated independently. Lee et al. [33] have proposed a training method based on maximum a posteriori (MAP) [34]. In this method, the mean parameters of the reference speaker included in a GMM between source and reference speakers are updated to a new target speaker's mean parameters by unsupervised MAP estimation.

In order to use more informative prior knowledge extracted from many other

speakers, we have proposed eigenvoice conversion (EVC) [11]. The eigenvoice technique [35] was originally proposed as a model adaptation technique in the speech recognition area. In the speech synthesis area, eigenvoice has been applied to HMM-based text-to-speech (TTS) in order to achieve speaker adaptation using a small amount of speech data [36]. This technique also achieves HMM-based TTS capable of voice quality control [37] or speaking style control [38]. EVC is similar to the speaker interpolation technique proposed by Iwahashi and Sagisaka [39] and the speaker generation system based on STRAIGHT morphing proposed by Ohtani et al. [19][20][21] in terms of using the information of various pre-stored speakers. The speaker interpolation technique can convert only feature segments included in the pre-stored database, and the speaker generation system requires expert knowledge to give pre-stored spectra anchor-points for morphing. On the other hand, EVC is capable of converting any source utterance into a target one because speaker interpolation is performed on the model parameter space.

EVC has brought novel VC frameworks, i.e., many-to-one EVC and one-to-many EVC [12]. Many-to-one EVC framework is a technique for converting an arbitrary source speaker's voice into a pre-determined target speaker's voice. It is possible to rapidly adapt the EV-GMM to a new source speaker using only an input utterance to be converted. The rapid adaptation performance is significantly improved by applying MAP adaptation to the unsupervised weight estimation [40]. On the other hand, the one-to-many EVC framework enables the conversion from a pre-determined source speaker's voice into an arbitrary target speaker's voice. One of the interesting applications of one-to-many EVC is voice quality control [41]. This application allows us to intuitively control the converted voice quality by manipulating voice quality control scores capturing voice characteristics represented by several primitive words such as gender and age. As another application for each EVC framework, cross-language EVC has been proposed [42]. In this method, the EV-GMM is adapted to a new speaker whose language is different from that used in training of the EV-GMM. The adapted EV-GMM allows the conversion between the pre-determined speaker's voice and the new speaker's voice while keeping the language of input speech unchanged. Therefore, even if languages of two speakers are different from each other, this method can effectively build the conversion model between them.

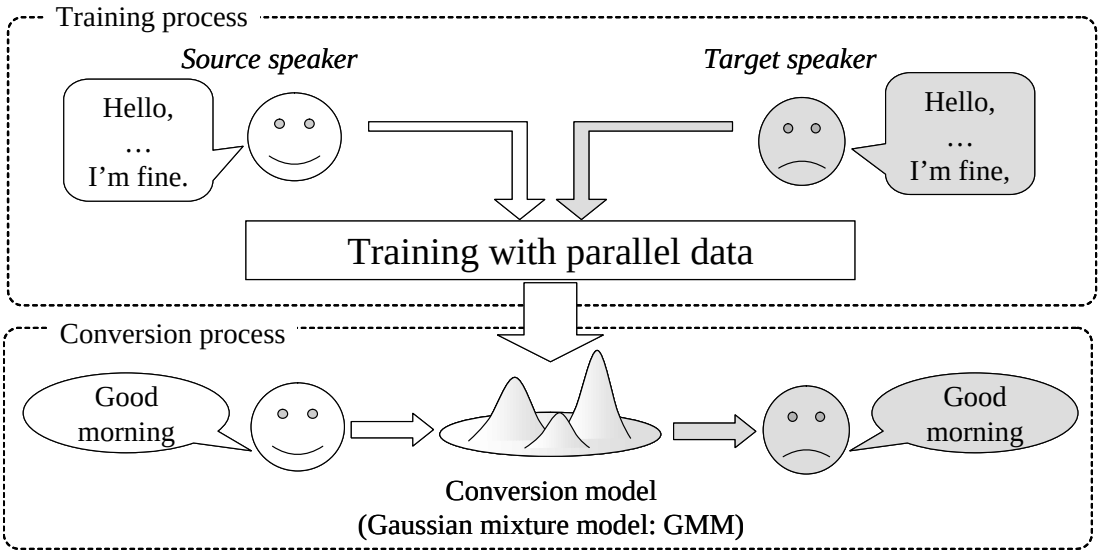


Figure 2.1. Overview of GMM-based voice conversion.

This chapter has described various traditional voice conversion (VC) frameworks such as statistical VC based on the Gaussian mixture model (GMM). In section 2.2, we describe the basic GMM-based VC frameworks. In section 2.3, the basic one-to-may EVC framework is described. In section 2.4, problems of one-to-many EVC system are described. Finally, we summarize this chapter in section 2.5.

2.2. Voice conversion based on Gaussian mixture model

Figure 2.1 shows the overview of the GMM-based VC, which includes the training process and the conversion process. In this section, we describe two types of the GMM-based VCs such as VC based on MMSE and VC based on maximum likelihood estimation (MLE).

2.2.1 Voice conversion based on minimum mean square error

We use D-dimensional acoustic features, the source speaker's feature $\boldsymbol{x}_t$ at the t^{th} frame and a target speaker's feature $\boldsymbol{y}_t$ at the t^{th} frame. A GMM models joint probability density of time-aligned source and target features determined by dynamic time warping (DTW) and is described as follows:

$$P(\boldsymbol{x}_t, \boldsymbol{y}_t|\lambda) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\Big(\left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top\right]^\top ; \boldsymbol{\mu}_m^{(x,y)},\ \boldsymbol{\Sigma}_m^{(x,y)} \Big), \tag{2.1}$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. M represents the number of mixture components and α_m is weight for the m^{th} component of the GMM. λ is parameter set of GMM, which includes α_m, $\boldsymbol{\mu}_m^{(x,y)}$ and $\boldsymbol{\Sigma}_m^{(x,y)}$. And $\top$ denotes transposition of the vector. The m^{th} mean vector $\boldsymbol{\mu}_m^{(x,y)}$ and $\boldsymbol{\Sigma}_m^{(x,y)}$ are written as follows:

$$\boldsymbol{\mu}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \tag{2.2}$$

where, $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ denote the m^{th} source mean vector and target mean vector, respectively. $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are m^{th} covariance matrices of source and target speakers, respectively. $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ represent the m^{th} cross-covariance matrices for source and target speakers, respectively. This GMM is trained with the EM algorithm [43] using parallel data set composed of utterance-pairs of source and target speakers as follows:

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{t=1}^{T} P(\boldsymbol{x}_t, \boldsymbol{y}_t|\lambda)\,. \tag{2.3}$$

In the conversion process based on MMSE [8][9], the converted target feature $\hat{\boldsymbol{y}}_t$ is determined by the following conversion function formulated as the conditional expectation $E[\boldsymbol{y}_t|\boldsymbol{x}_t]$:

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= E[\boldsymbol{y}_t|\boldsymbol{x}_t] \\ &= \int P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)\, \boldsymbol{y}_t d\boldsymbol{y}_t, \end{aligned} \tag{2.4}$$

where $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)$ is the conditional probability density of $\boldsymbol{y}_t$ given $\boldsymbol{x}_t$ which modeled by a GMM as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \lambda)\, P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \lambda)\,, \tag{2.5}$$

$$P(m|\boldsymbol{x}_t, \lambda) = \frac{\alpha_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}, \tag{2.6}$$

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \lambda) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m_t,t}^{(y)}, \boldsymbol{D}_{m_t}^{(y)}\right). \tag{2.7}$$

The mean vector $\boldsymbol{E}_{m,t}^{(y)}$ and covariance matrix $\boldsymbol{D}_m^{(y)}$ of the m^{th} conditional distribution are written as follows:

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_{m_t}^{(y)} + \boldsymbol{\Sigma}_{m_t}^{(yx)} \boldsymbol{\Sigma}_{m_t}^{(xx)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{m_t}^{(x)}\right), \tag{2.8}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_{m_t}^{(yy)} - \boldsymbol{\Sigma}_{m_t}^{(yx)} \boldsymbol{\Sigma}_{m_t}^{(xx)^{-1}} \boldsymbol{\Sigma}_{m_t}^{(xy)}. \tag{2.9}$$

Therefore, Eq. (2.4) is rewritten as follows:

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= \int \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \lambda)\, P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \lambda)\, \boldsymbol{y}_t d\boldsymbol{y}_t, \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \lambda)\, \boldsymbol{E}_{m_t,t}^{(y)}. \end{aligned} \tag{2.10}$$

2.2.2 Voice conversion based on maximum likelihood estimation

In the MMSE-based method, the conversion process is performed frame-by-frame. Therefore, the converted acoustic feature sequence often includes discontinuities. In contrast, the MLE-based conversion alleviates these discontinuities by converting feature vectors in all frames over a time sequence.

In the MLE-based conversion, we use $2D$-dimensional acoustic features, a source speaker's feature $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and a target speaker's feature $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$, consisting of D-dimensional static and dynamic features. In the

same manner as the MMSE-based VC, we models joint probability density of source and target speakers with a GMM as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\Big(\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top ; \boldsymbol{\mu}_m^{(X,Y)},\ \boldsymbol{\Sigma}_m^{(X,Y)} \Big), \tag{2.11}$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}. \tag{2.12}$$

In this conversion, a time sequence of converted static feature vectors $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top,\ \hat{\boldsymbol{y}}_2^\top,\ \cdots,\ \hat{\boldsymbol{y}}_T^\top\right]^\top$ are obtained as follows:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \lambda) \\ &= \arg\max_{\boldsymbol{y}} \sum_{\text{all}\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \lambda)\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \lambda), \\ &= \arg\max_{\boldsymbol{y}} \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \lambda)\, P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda), \end{aligned} \tag{2.13}$$

$$\text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{2.14}$$

where $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ are a time sequence of the source features and that of the target features, respectively. $\boldsymbol{m} = \{m_1, m_2, \cdots, m_T\}$ is a mixture component sequence. Then, $\boldsymbol{W}$ denotes the matrix to extend the static feature sequence to the static and dynamic feature sequence. In this thesis, we employ $\boldsymbol{W}$ written as follows

$$\boldsymbol{W} = \left[\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \cdots, \boldsymbol{w}_T^\top\right]^\top, \tag{2.15}$$

where

$$\boldsymbol{w}_t = \begin{bmatrix} \boldsymbol{0}_{D\times(t-1)D} & \boldsymbol{I} & \boldsymbol{0}_{D\times(T-t)D} \\ \boldsymbol{0}_{D\times(t-2)D} & -0.5\boldsymbol{I}\ \ \boldsymbol{0}_{D\times D}\ \ 0.5\boldsymbol{I} & \boldsymbol{0}_{D\times(T-t-1)D} \end{bmatrix}, \tag{2.16}$$

and the matrix $\boldsymbol{I}$ is a $D \times D$ identity matrix. Figure 2.2 shows the relationship between $\boldsymbol{y}$ and $\boldsymbol{Y}$. The conditional probability density $P(\boldsymbol{Y}|\boldsymbol{X}, \lambda)$ is modeled as a GMM and the m^{th} component weight $P(\boldsymbol{m}|\boldsymbol{X}_t, \lambda)$ and conditional probability

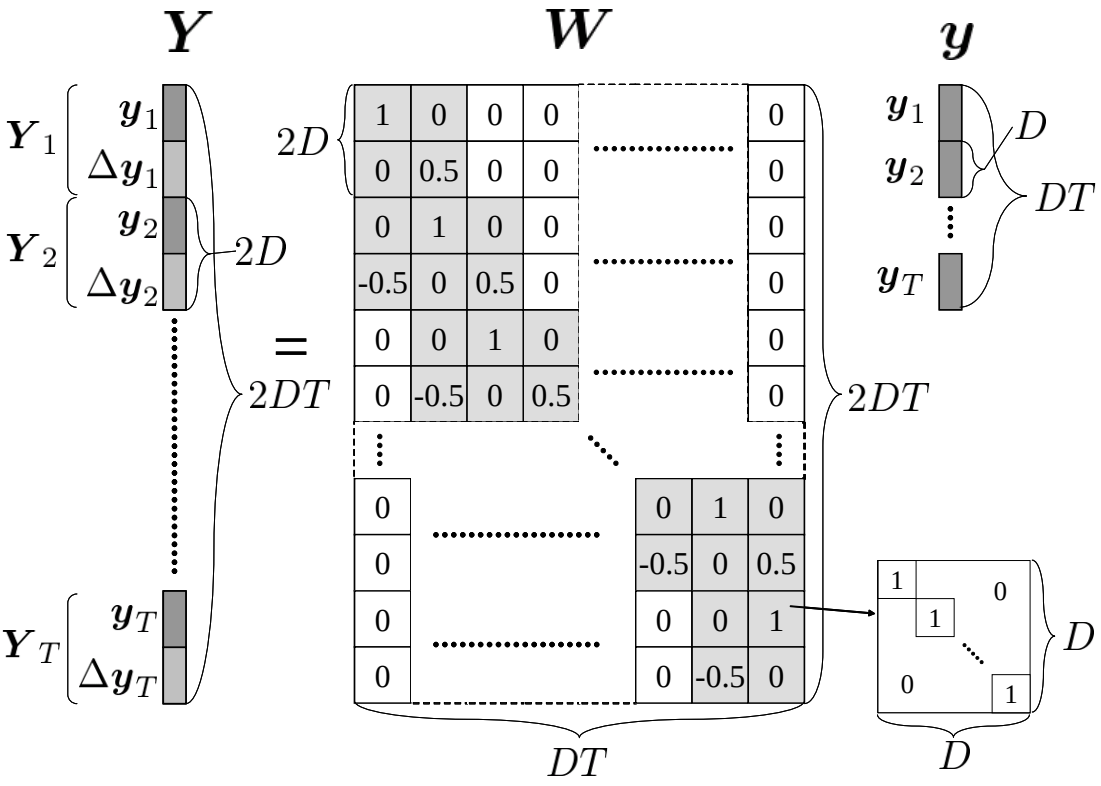


Figure 2.2. Overview of relationship between a sequence of the static feature vectors $\boldsymbol{y}$ and that of the static and dynamic feature vectors $\boldsymbol{Y}$.

$P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{m}, \lambda)$ at the t^{th} frame are given as follows:

$$P(m|\boldsymbol{X}_t, \lambda) = \frac{\alpha_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}, \tag{2.17}$$

$$P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}\right), \tag{2.18}$$

where

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_{m_t}^{(Y)} + \boldsymbol{\Sigma}_{m_t}^{(YX)} \boldsymbol{\Sigma}_{m_t}^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_{m_t}^{(X)}\right), \tag{2.19}$$

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_{m_t}^{(YY)} - \boldsymbol{\Sigma}_{m_t}^{(YX)} \boldsymbol{\Sigma}_{m_t}^{(XX)^{-1}} \boldsymbol{\Sigma}_{m_t}^{(XY)}. \tag{2.20}$$

A converted static feature sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}}, \tag{2.21}$$

where

$$\overline{\boldsymbol{D}^{(Y)^{-1}}} = \text{diag}\left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}}, \overline{\boldsymbol{D}_2^{(Y)^{-1}}}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}}\right], \tag{2.22}$$

$$\overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)^{-1}} \boldsymbol{E}_1^{(Y)}}^\top, \overline{\boldsymbol{D}_2^{(Y)^{-1}} \boldsymbol{E}_2^{(Y)}}^\top, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}} \boldsymbol{E}_T^{(Y)}}^\top\right]^\top, \tag{2.23}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}}, \tag{2.24}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}} \boldsymbol{E}_t^{(Y)}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)}, \tag{2.25}$$

$$\gamma_{m,t} = P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \lambda) \tag{2.26}$$

.

In addition, this likelihood function is approximated with the suboptimum mixture component sequence $\hat{\boldsymbol{m}} = [\hat{m}_1,\ \hat{m}_2,\ \cdots,\ \hat{m}_T]$, which is determined by

$$\hat{\boldsymbol{m}} = \arg\max_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \lambda). \tag{2.27}$$

We determined the converted static feature sequence $\hat{\boldsymbol{y}}$ as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda\right). \tag{2.28}$$

The conditional probability $P\left(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\lambda\right)$ is written as

$$P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\lambda)=\prod_{t=1}^{T}\mathcal{N}\left(\boldsymbol{Y}_t;\boldsymbol{E}_{\hat{m}_t,t}^{(Y)},\boldsymbol{D}_{\hat{m}_t}^{(Y)}\right), \tag{2.29}$$

$$\boldsymbol{E}_{\hat{m}_t,t}^{(Y)}=\boldsymbol{\mu}_{\hat{m}_t}^{(Y)}+\boldsymbol{\Sigma}_{\hat{m}_t}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}_t}^{(XX)^{-1}}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_{\hat{m}_t}^{(X)}\right), \tag{2.30}$$

$$\boldsymbol{D}_{\hat{m}_t}^{(Y)}=\boldsymbol{\Sigma}_{\hat{m}_t}^{(YY)}-\boldsymbol{\Sigma}_{\hat{m}_t}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}_t}^{(XX)^{-1}}\boldsymbol{\Sigma}_{\hat{m}_t}^{(XY)}. \tag{2.31}$$

The converted static feature sequence $\hat{\boldsymbol{y}}$ is given by

$$\hat{\boldsymbol{y}}=\left(\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}, \tag{2.32}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}=\left[\boldsymbol{E}_{m_1,1}^{(Y)},\ \boldsymbol{E}_{m_2,2}^{(Y)},\ \cdots,\ \boldsymbol{E}_{m_T,T}^{(Y)}\right], \tag{2.33}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}=\mathrm{diag}\left[\boldsymbol{D}_{m_1}^{(Y)^{-1}},\ \boldsymbol{D}_{m_2}^{(Y)^{-1}},\ \cdots,\ \boldsymbol{D}_{m_T}^{(Y)^{-1}}\right]. \tag{2.34}$$

Figure 2.3 shows the graphical representation in the MLE-based conversion process. In the MLE-based conversion without considering dynamic features, the converted feature at a certain frame is independently modified by only source feature and mixture component information at the same frame. Therefore, this conversion is performed frame-by-frame and inconsistencies between converted features are arisen. On the other hand, in the MLE-based conversion considering dynamic features, the converted static feature at a certain frame is modified by using all of source static and dynamic features and mixture components information. Consequently, this conversion is performed in each trajectory.

2.2.3 MLE-based conversion considering global variance

Although the discontinuity of converted feature sequence has been improved by the MLE-based conversion using dynamic feature, the problem of over-smoothed converted features still remains. To alleviate this problem, GV is introduced to the MLE-based conversion. The GV of the target static feature vectors over a time sequence is written as

$$\boldsymbol{v}_y=\left[v_y(1),\ v_y(2),\ \cdots,\ v_y(D)\right]^{\top}, \tag{2.35}$$

$$v_y(d)=\frac{1}{T}\sum_{t=1}^{T}\left(y_t(d)-\frac{1}{T}\sum_{t=1}^{T}y_t(d)\right)^2, \tag{2.36}$$

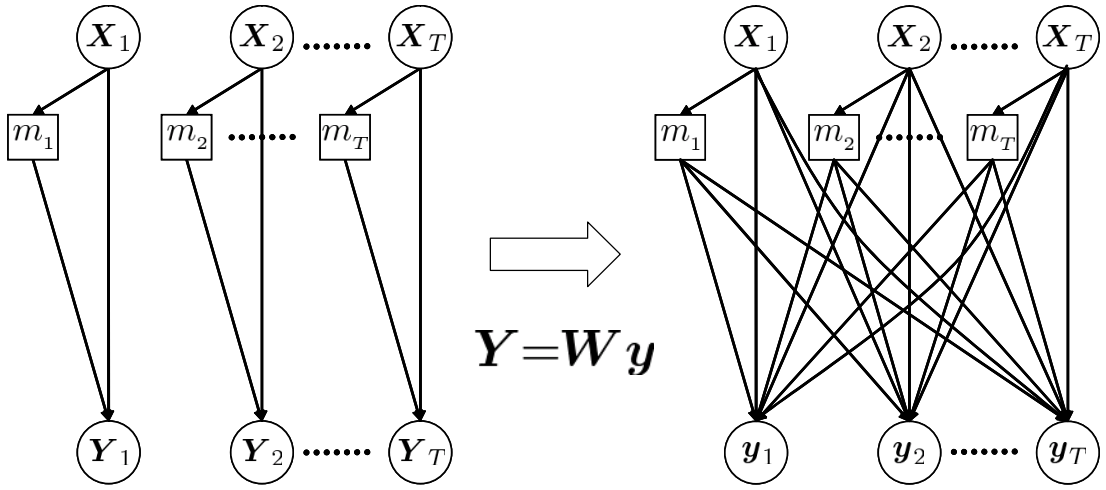


Figure 2.3. Graphical representation of relationship between individual variables in MLE-based conversion process. Left graph is MLE-based conversion without considering dynamic features and right figure is MLE-based conversion considering dynamic features.

where, $y_t(d)$ is the d^{th} component of the target static feature at the t^{th} frame. We calculate the GV utterance by utterance. In this conversion, the converted target static feature sequence $\hat{\boldsymbol{y}}$ is determined by maximizing the following likelihood function:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \left\{ \sum_{\text{all}\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \lambda)\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) \right\}^{\omega} P(\boldsymbol{v}_y|\lambda_v)\,, \tag{2.37}$$

where $P(\boldsymbol{v}_y|\lambda_v)$ is modeled by the single Gaussian which includes a mean vector $\boldsymbol{\mu}^{(v)}$ and a covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ as follows:

$$P(\boldsymbol{v}_y|\lambda_v) = \mathcal{N}\Big(\boldsymbol{v}_y; \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\Big)\,. \tag{2.38}$$

Then, ω is a weight parameter for controlling the two likelihoods. In this thesis, ω is set $\frac{1}{2T}$ which is determined by the ratio of the number of dimensions between target acoustic feature sequence and GV.

The same as in section 2.2.2, Eq. (2.37) can be approximated with the sub-

optimum mixture component sequence $\hat{\boldsymbol{m}}$ as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \mathcal{L}, \tag{2.39}$$

$$\begin{aligned}\mathcal{L} &= \omega \log P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda) + \log P(\boldsymbol{v}_y|\lambda_v) \\ &= \omega \left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} \right) \\ &\quad - \frac{1}{2}\boldsymbol{v}_y^\top \boldsymbol{\Sigma}^{(vv)^{-1}} \boldsymbol{v}_y + \boldsymbol{v}_y^\top \boldsymbol{\Sigma}^{(vv)^{-1}} \hat{\boldsymbol{\mu}}^{(v)} + K,\end{aligned} \tag{2.40}$$

where K denotes the independent invariable of $\boldsymbol{Y}$. In this conversion method, the converted feature $\hat{\boldsymbol{y}}$ is updated by using steepest descent method as follows:

$$\boldsymbol{y}^{(i+1)^{\text{th}}} = \boldsymbol{y}^{(i)^{\text{th}}} + \theta \cdot \delta \boldsymbol{y}^{(i)^{\text{th}}}, \tag{2.41}$$

where θ is the step size. $\delta \boldsymbol{y}^{(i)^{\text{th}}}$ is the first derivative of $\mathcal{L}$ described as follows:

$$\delta \boldsymbol{y}^{(i)^{\text{th}}} = \left. \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} \right|_{\boldsymbol{y}=\boldsymbol{y}^{(i)^{\text{th}}}}, \tag{2.42}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} &= \omega \left(-\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} \right) \\ &\quad + \left[\boldsymbol{v}_1'^\top,\ \boldsymbol{v}_2'^\top, \cdots, \boldsymbol{v}_T'^\top \right]^\top,\end{aligned} \tag{2.43}$$

$$\boldsymbol{v}_t' = \left[v_t'(1),\ v_t'(2), \cdots,\ v_t'(D) \right]^\top, \tag{2.44}$$

$$v_t'(d) = -\frac{2}{T} \boldsymbol{p}_v^{(d)^\top} \left(\boldsymbol{v}_y - \hat{\boldsymbol{\mu}}^{(v)} \right) \left(y_t(d) - \bar{y}(d) \right), \tag{2.45}$$

where $\boldsymbol{p}_v^{(d)}$ is the d^{th} column vector of $\boldsymbol{\Sigma}^{(vv)^{-1}}$.

2.3. Eigenvoice conversion

In the EVC framework, there are two frameworks, one-to-many EVC and many-to-one EVC [12]. In this thesis, we describe overall EVC framework by using one-to-many EVC framework. Figure 2.4 shows an overview of one-to-many EVC framework. EVC has three main processes: training, adaptation and conversion. In the training process, an eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets consisting of the specified source speaker and

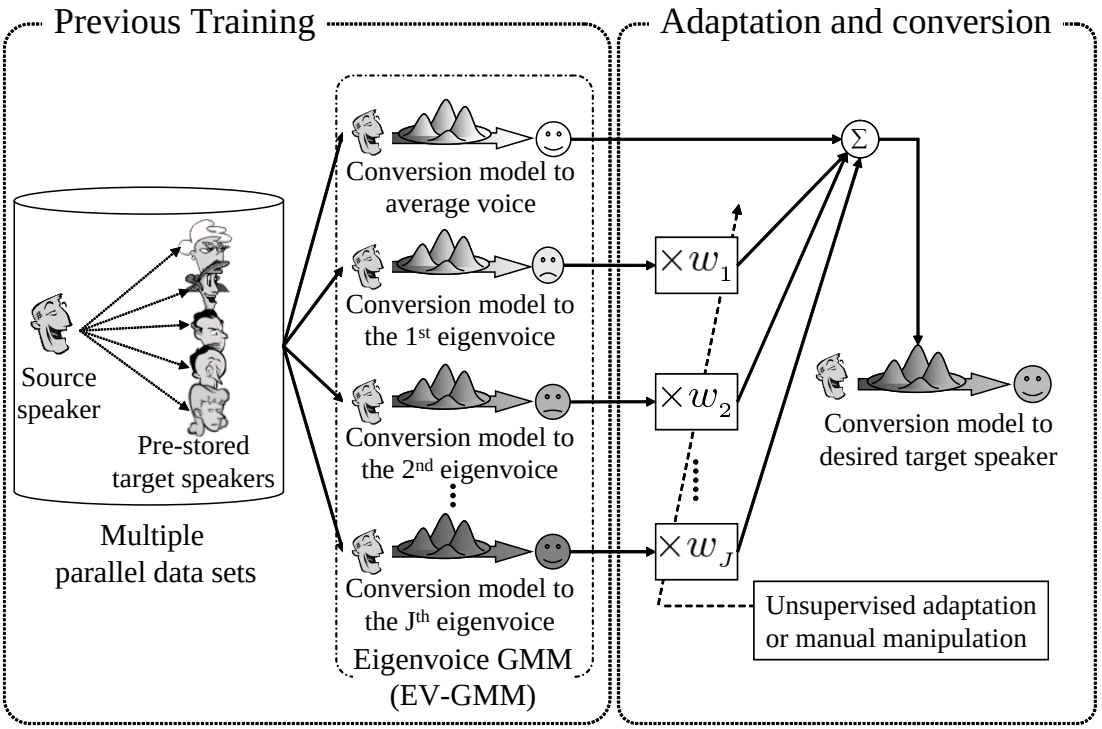


Figure 2.4. Overview of one-to-many EVC framework.

many pre-stored target speakers. In the adaptation process, the trained EV-GMM is capable of being flexibly adapted to a new target speaker using only a few arbitrary utterances of the target speaker. A few adaptive parameters are estimated in a completely text-independent manner. Moreover, the EV-GMM allows us to control voice quality of the converted speech by manipulating those adaptive parameters. In the conversion process, arbitrary utterances of the source speaker's voice are converted into those of the new target speaker's voice using adapted EV-GMM.

In following subsections, the EV-GMM's structure and above processes are described in detail.

2.3.1 One-to-many eigenvoice Gaussian mixture model

We use $2D$-dimensional acoustic features, source speaker's feature $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and the s^{th} target speaker's feature $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}\right]^\top$, consisting of D-dimensional static and dynamic features. The joint probability density of time-aligned source and target features determined by DTW is modeled with EV-GMM as follows:

$$P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}\right]^\top; \boldsymbol{\mu}_{m,s}^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right), \tag{2.46}$$

$$\boldsymbol{\mu}_{m,s}^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)} \end{bmatrix}, \tag{2.47}$$

In EV-GMM, a target mean vector is modeled as linear combination with the bias vector $\boldsymbol{b}_m^{(0)}$, representative vectors $\boldsymbol{B}_m = \left[\boldsymbol{b}_m^{(1)}, \boldsymbol{b}_m^{(2)}, \cdots, \boldsymbol{b}_m^{(J)}\right]$ and the weight vector $\boldsymbol{w}^{(s)}$. Voice characteristics of various target speakers are effectively modeled by setting $\boldsymbol{w}^{(s)}$ to appropriate values. The other parameters $\lambda^{(EV)}$, such as mixture component weights, source mean vectors, bias vectors, representative vectors and covariance matrices, are tied for every target speaker.

2.3.2 Training of EV-GMM based on principal component analysis

First, a target-speaker-independent GMM (TI-GMM) $\lambda^{(0)}$ is trained with multiple parallel data sets consisting of utterance-pairs of the source speaker and multiple pre-stored target speakers as follows:

$$\hat{\lambda}^{(0)} = \arg\max_{\lambda^{(0)}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \lambda^{(0)}\Big). \tag{2.48}$$

Then, using only the parallel data set for the s^{th} pre-stored target speaker, the s^{th} target-speaker-dependent GMM (TD-GMM) $\lambda^{(s)}$ is trained by only updating

target mean vectors $\boldsymbol{\mu}^{(Y)}(s)$ of $\lambda^{(0)}$ based on maximum likelihood (ML) estimate as follows:

$$\hat{\lambda}^{(s)} = \arg\max_{\boldsymbol{\mu}^{(Y)}(s)} \prod_{t=1}^{T_s} P\Big(\boldsymbol{X}, \boldsymbol{Y}_t^{(s)} | \lambda^{(0)}\Big). \tag{2.49}$$

After training the TD-GMMs for all pre-stored target speakers, a $2DM$-dimensional supervector $\boldsymbol{SV}^{(s)} = \Big[\boldsymbol{\mu}_1^{(Y)^\top}(s),\ \boldsymbol{\mu}_2^{(Y)^\top}(s),\ \cdots,\ \boldsymbol{\mu}_M^{(Y)^\top}(s)\Big]^\top$ is constructed for each pre-stored target speaker by concatenating the updated target mean vectors $\Big\{\boldsymbol{\mu}_1^{(Y)}(s), \boldsymbol{\mu}_2^{(Y)}(s), \cdots, \boldsymbol{\mu}_M^{(Y)}(s)\Big\}$ of $\lambda^{(s)}$. Finally, bias vector $\boldsymbol{b}_m^{(0)}$ and representative vectors $\boldsymbol{B}_m$ are extracted by performing principal component analysis (PCA) for the supervectors for all pre-stored target speakers $\Big\{\boldsymbol{SV}^{(1)},\ \boldsymbol{SV}^{(2)},\ \cdots,\ \boldsymbol{SV}^{(S)}\Big\}$, where S denotes the number of pre-stored target speakers. Each supervector is approximated as follows:

$$\boldsymbol{SV}^{(s)} \simeq \big[\boldsymbol{B}_1^\top,\ \boldsymbol{B}_2^\top,\ \cdots,\ \boldsymbol{B}_M^\top\big]^\top \boldsymbol{w}^{(s)} + \Big[\boldsymbol{b}_1^{(0)^\top},\ \boldsymbol{b}_1^{(0)^\top},\ \cdots,\ \boldsymbol{b}_M^{(0)^\top}\Big]^\top, \tag{2.50}$$

$$\boldsymbol{b}_m^{(0)} = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\mu}_m^{(Y)}(s), \tag{2.51}$$

where $\boldsymbol{w}^{(s)}$ is $J(< S \ll 2DM)$ principle components for the s^{th} target speaker.

2.3.3 Unsupervised adaptation of trained EV-GMM

We adapt the EV-GMM to an arbitrary target speaker by estimating the optimum weight vector for their given speech samples without any linguistic information. We apply maximum likelihood eigen-decomposition (MLED) [35] to the weight vector estimation. In one-to-many EVC, the weight vector $\boldsymbol{w}$ is estimated so that likelihood of the marginal distribution for a time sequence of the given target features $\Big\{\boldsymbol{Y}_1^{(tar)}, \boldsymbol{Y}_2^{(tar)}, \cdots, \boldsymbol{Y}_T^{(tar)}\Big\}$ is maximized [11] as follows:

$$\begin{aligned}\hat{\boldsymbol{w}} &= \arg\max_{\boldsymbol{w}} \prod_{t=1}^{T} \int P\Big(\boldsymbol{X}_t,\ \boldsymbol{Y}_t^{(tar)} | \lambda^{(EV)}, \boldsymbol{w}\Big)\, d\boldsymbol{X}_t \\ &= \arg\max_{\boldsymbol{w}} \prod_{t=1}^{T} P\Big(\boldsymbol{Y}_t^{(tar)} | \lambda^{(EV)}, \boldsymbol{w}\Big).\end{aligned} \tag{2.52}$$

This adaptation process is performed with EM algorithm by maximizing the following auxiliary function:

$$Q\left(\boldsymbol{w}, \hat{\boldsymbol{w}}\right) = \sum_{t=1}^{T}\sum_{m=1}^{M} P\Big(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}, \boldsymbol{w}\Big) \log P\Big(\boldsymbol{Y}_t^{(tar)}, m|\lambda^{(EV)}, \hat{\boldsymbol{w}}\Big). \quad (2.53)$$

The ML estimate of the weight vector for a target speaker $\hat{\boldsymbol{w}}$ is determined as follows:

$$\hat{\boldsymbol{w}} = \left\{\sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{B}_m^{\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m\right\}^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^{\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)}, \quad (2.54)$$

where

$$\overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} P\Big(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}, \boldsymbol{w}\Big), \quad (2.55)$$

$$\overline{\boldsymbol{Y}}_t^{(tar)} = \sum_{t=1}^{T} P\Big(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}, \boldsymbol{w}\Big) \Big(\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}\Big). \quad (2.56)$$

Note that this process is completely unsupervised adaptation using only arbitrary utterances of the target speaker. We also need to use only a small amount of adaptation data because there are a few parameters to be adapted.

2.3.4 Conversion with adapted EV-GMM

We use the conversion method based on MLE considering dynamic features [10] described in section 2.2.2. Converted static feature vectors $\hat{\boldsymbol{y}}$ are obtained as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \sum_{\text{all } \boldsymbol{m}} P\big(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(EV)}\big) P\big(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}\big). \quad (2.57)$$

Furthermore, using the approximated MLE-based conversion method, the converted static feature sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(EV)}, \hat{\boldsymbol{w}}\right), \quad (2.58)$$

where $P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(EV)}, \hat{\boldsymbol{w}}\right)$ is written as

$$P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(EV)}, \hat{\boldsymbol{w}}\right) = \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{E}^{(Y)}_{\hat{m}_t,t}, \boldsymbol{D}^{(Y)}_{\hat{m}_t}\right), \tag{2.59}$$

$$\boldsymbol{E}^{(Y)}_{\hat{m}_t,t} = \boldsymbol{B}_{\hat{m}_t}\hat{\boldsymbol{w}} + \boldsymbol{b}^{(0)}_{\hat{m}_t} + \boldsymbol{\Sigma}^{(YX)}_{\hat{m}_t}\boldsymbol{\Sigma}^{(XX)^{-1}}_{\hat{m}_t}\left(\boldsymbol{X}_t - \boldsymbol{\mu}^{(X)}_{\hat{m}_t}\right), \tag{2.60}$$

$$\boldsymbol{D}^{(Y)}_{\hat{m}_t} = \boldsymbol{\Sigma}^{(YY)}_{\hat{m}_t} - \boldsymbol{\Sigma}^{(YX)}_{\hat{m}_t}\boldsymbol{\Sigma}^{(XX)^{-1}}_{\hat{m}_t}\boldsymbol{\Sigma}^{(XY)}_{\hat{m}_t}. \tag{2.61}$$

The converted static feature sequence $\hat{\boldsymbol{y}}$ is given by

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \boldsymbol{D}^{(Y)^{-1}}_{\hat{\boldsymbol{m}}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}^{(Y)^{-1}}_{\hat{\boldsymbol{m}}} \boldsymbol{E}^{(Y)}_{\hat{\boldsymbol{m}}}, \tag{2.62}$$

where,

$$\boldsymbol{E}^{(Y)}_{\hat{\boldsymbol{m}}} = \left[\boldsymbol{E}^{(Y)}_{m_1,1},\ \boldsymbol{E}^{(Y)}_{m_2,2},\ \cdots,\ \boldsymbol{E}^{(Y)}_{m_T,T}\right], \tag{2.63}$$

$$\boldsymbol{D}^{(Y)^{-1}}_{\hat{\boldsymbol{m}}} = \mathrm{diag}\left[\boldsymbol{D}^{(Y)^{-1}}_{m_1},\ \boldsymbol{D}^{(Y)^{-1}}_{m_2},\ \cdots,\ \boldsymbol{D}^{(Y)^{-1}}_{m_T}\right]. \tag{2.64}$$

2.4. Issues of the conventional one-to-many EVC system

Figure 2.5 shows an overview of the conventional one-to-many EVC system that includes training, adaptation and conversion processes. In this system, we train the EV-GMM only for spectral features and employ the adapted EV-GMM for the spectral conversion.

In the conversion process for fundamental frequency, we convert source fundamental frequency F_0 to target one as follows:

$$\log \tilde{F}_0 = \frac{\sigma^{(y)}}{\sigma^{(x)}}\left(\log F_0 - \mu^{(x)}\right) + \mu^{(y)}, \tag{2.65}$$

where $\mu^{(x)}$ and $\sigma^{(x)}$ denote mean and standard deviation of log-scaled source F_0, and $\mu^{(y)}$ and $\sigma^{(y)}$ denote those of log-scaled target F_0. In this system, these statistics for the source speaker are calculated from the training data and those for the target speaker are calculated from the adaptation data.

In the process of synthesizing converted speech, the excitation signal is generated with STRAIGHT simple excitation (STME) [13]. Figure 2.6 shows the

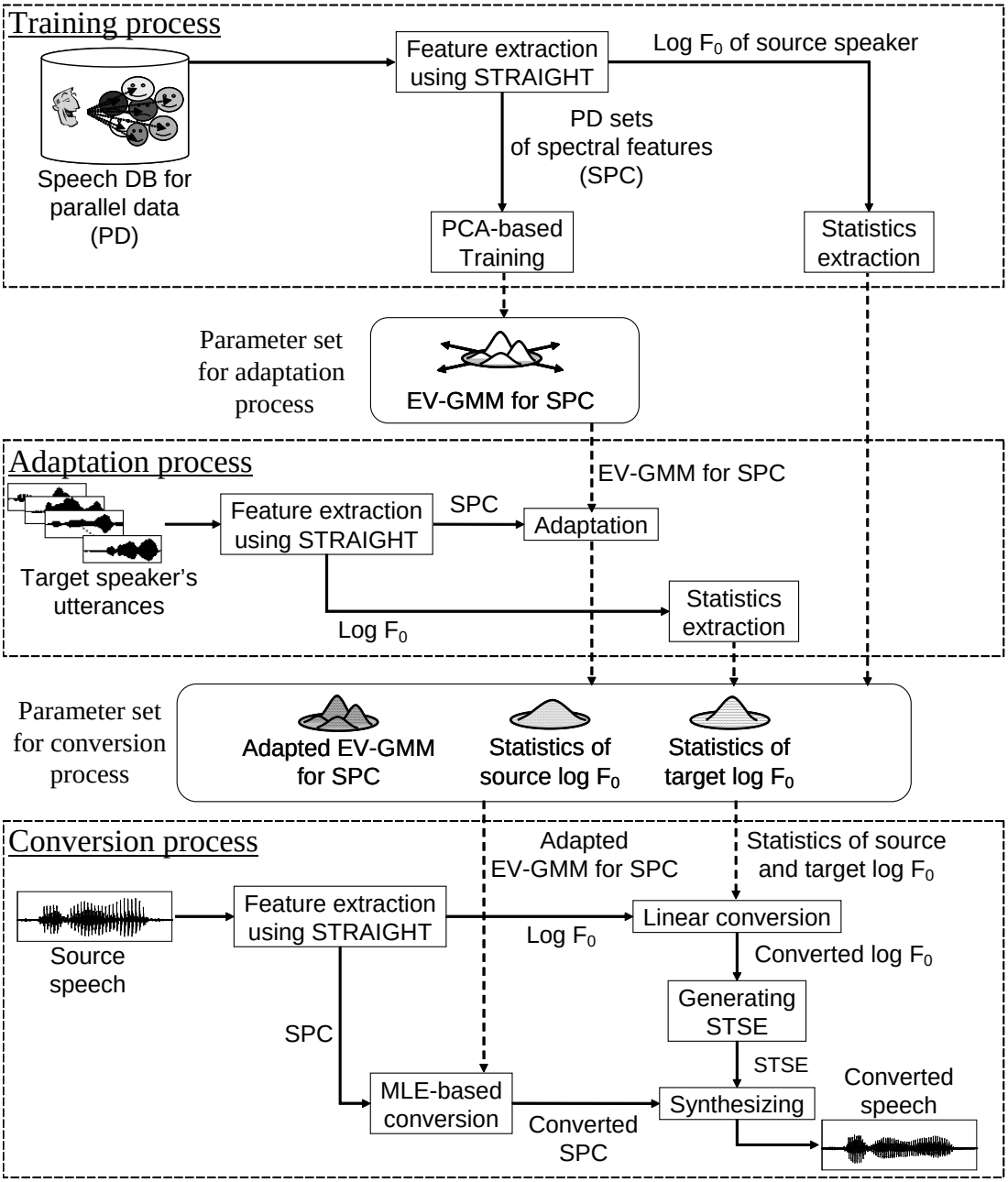


Figure 2.5. Conventional one-to-many EVC system which includes training, adaptation and conversion processes.

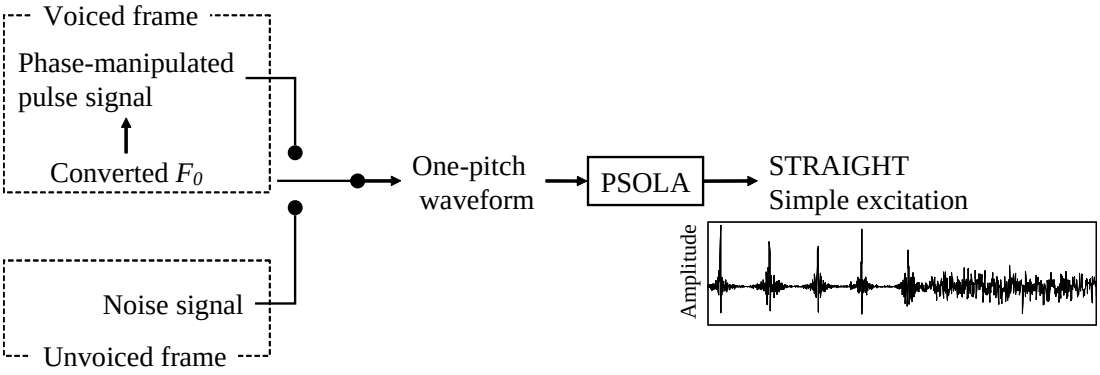


Figure 2.6. Generation process of STRAIGHT simple excitation.

generation process of STSE. In order to alleviate buzz sounds caused by using a pulse train for generating a voiced excitation signal, phase components in high-frequency bands (e.g., over 3 kHz) are dispersed with an all-pass filter [44]. A one-pitch waveform is generated by selecting the phase-manipulated pulse train based on the converted F_0 for voiced segments or white noise for unvoiced segments. Then, an excitation signal is generated by PSOLA (Pitch Synchronous OverLap Add) technique [45]. Then, the converted speech is synthesized by filtering the generated excitation with the converted spectral sequence.

Although this one-to-many EVC system achieves the flexible training of conversion model, this converted speech quality is still not high enough. This degradation of the converted speech quality is caused by employing three materials, i.e., the excitation model, the conversion algorithm and the EV-GMM.

2.4.1 Problem of excitation model

Figure 2.7 shows a comparison of the STSE constructed from F_0 information and the residual signal extracted from natural speech. In the unvoiced segment, the simple excitation models that residual signal well because the simple excitation is modeled as white noise train and then the residual signal looks noisy. In the voiced segment, while the simple excitation is composed of only phase-manipulated pulse train, the residual signal includes noise and pulse-like signals. This means that the

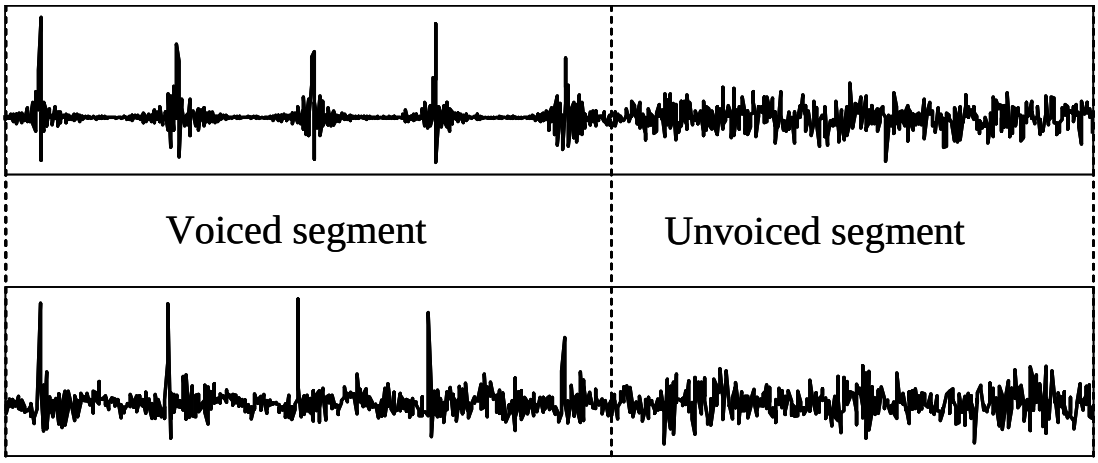


Figure 2.7. Comparison of excitation signals of the identical speaker, in which top is STRAIGHT simple excitation and bottom is the residual signal.

simple excitation cannot model the voice segment of the residual signal accurately and causes the buzzy sound of the converted speech. Therefore, we need to employ better excitation model than the simple excitation.

2.4.2 Problem of conversion algorithm

Figure 2.8 shows the comparison between the converted spectrum of the conventional one-to-many EVC system and the correspondent target spectrum. The converted spectral shape is not clearer than the target spectral one. This is because the over-smoothing is caused by employing the MLE-based conversion not considering the GV in the conventional one-to-many EVC system. Therefore, the conventional one-to-many EVC system gives us a muffled speech. In order to alleviate the over-smoothing, we need to use the MLE-based conversion algorithm considering the GV.

2.4.3 Problem of EV-GMM

The tied-parameters of the PCA-based EV-GMM are from the TI-GMM modeling joint probability density of acoustic features of the source speaker and all

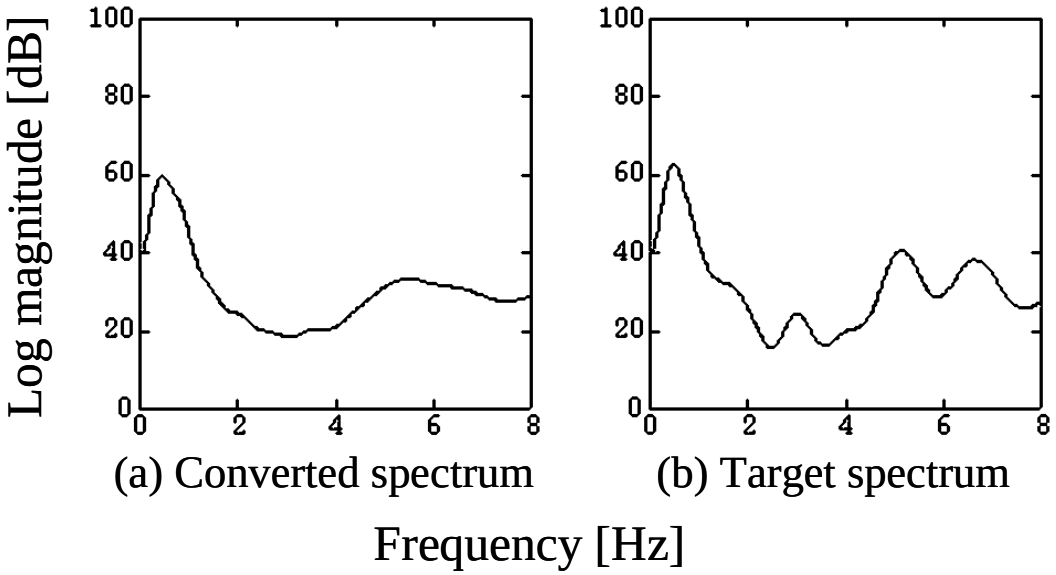


Figure 2.8. Comparison of converted spectrum and target spectrum.

pre-stored target speakers as mentioned. They capture not only intra-speaker acoustic variations but also inter-speaker acoustic variations of pre-stored target speakers. To demonstrate the impact of these tied-parameters on the EV-GMM adaptation, Figure 2.9 shows marginal distributions of the TI-GMM, two TD-GMMs and two adapted EV-GMMs on the 2^{nd} dimensional coefficient of target features. We can see that variance values of the adapted EV-GMMs are much larger than those of the TD-GMMs although their mean values are close to those of the TD-GMMs. These mismatches in modeling probability density would cause performance degradation of the adapted EV-GMM.

2.5. Summary

This chapter has described various traditional voice conversion (VC) frameworks and we have focused on statistical VC based on the Gaussian mixture model (GMM). We have reviewed some conversion algorithms used in the GMM-based VC framework. Also, we have described eigenvoice conversion (EVC) capable of making the conversion model training more flexible compared with the traditional

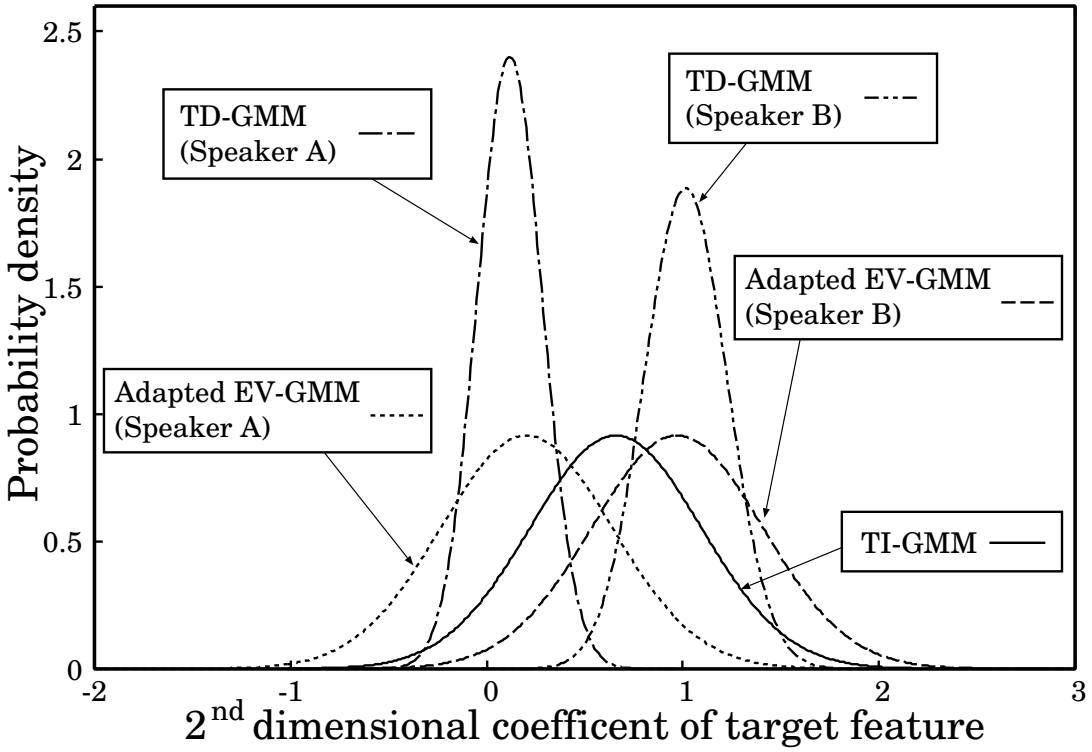


Figure 2.9. Marginal distributions for the 2^{nd} dimensional coefficient of target features of TI-GMM, TD-GMMs and adapted EV-GMMs.

GMM-based VC. The EVC framework brings two novel VC frameworks, i.e., one-to-many EVC and many-to-one EVC. Moreover, we have also explained some problems in the conventional one-to-many EVC system: i.e., quality degradation of the converted speech is caused by the use of the simple excitation model, the conversion algorithm not considering the global variance, and the training method of the EV-GMM.

Chapter 3

Voice Conversion with STRAIGHT Mixed Excitation

In this chapter, we describe the improved traditional VC framework, which is achieved by introducing STRAIGHT mixed excitation (STME). The performance of the traditional VC has been improved by applying spectral feature conversion to the MLE-based conversion algorithm with GV. However, the converted speech still contains traces of artificial sounds. To alleviate this, it is necessary to statistically model a source sequence as well as a spectral sequence. Thus, we introduce STME to the traditional VC framework. STME is generated by frequency-dependent weighted sum of white noise and a pulse train with phase manipulation based on aperiodic component, which represents noise barometer included in spectral features. In this framework, aperiodic component is modeled by GMM and converted aperiodic components are employed for STME building. Moreover, objective and subjective evaluation results demonstrate that the proposed source conversion produces strong improvements in both the converted speech quality and the conversion accuracy for speaker individuality.

3.1. Introduction

The converted speech quality of the GMM-based VC has been improved by applying spectral features to MLE-based conversion [10]. However, artificial sound is still evident in the converted speech. This problem is caused by employing not

so good excitation model, e.g., STSE shown in Figure 2.7. In speech synthesis, an excitation plays a very important role to generate more natural synthesized speeches. In VC frameworks, it is also important to build the better target excitation model. Therefore, it is necessary to statistically model a source sequence as well as a spectral sequence in order to alleviate the artificial sound.

Several researchers have proposed the source conversion methods such as the residual codebook [46], residual selection [47], and phase prediction [48]. The residual codebook, proposed by Kain et al., uses speech coders with a speaker-dependent excitation codebook. In training process, one-pitch-period residual waveforms, which are extracted target training data, are clustered by correspondent spectral information. Then, each centroid is designed by the averaged residual spectrum and pitch information extracted from the nearest sample. In the conversion process, the converted excitation is generated from weighted mean of centroid residual spectra based on the converted spectrum and centroid phase selected from the ML class. Residual selection, proposed by Sündermann et al., is a refinement of the residual codebook. This method selects appropriate residuals from a database extracted from the target speaker's training data. In phase prediction proposed by Ye et al., the codebook which consists of one-pitch-period waveforms based on spectral information is determined by MMSE. Then the required phases are obtained from the predicted waveform shapes of converted spectra. In these methods, when analyzing, we need to extract one-pitch-period waveforms accurately because we directly treat waveform information, i.e., spectral features and phase features. Moreover, these frameworks are based on selecting optimal residual waveforms from training data. Therefore, these performances strongly depend on training data size and it is difficult to introduce statistical methods.

In order to achieve the statistical solution of excitation problem, we employ STRAIGHT mixed excitation (STME) as a source model. Advantages of STME are that 1) the extracted features are statistically modeled in the same manner as that for spectral modeling, and 2) robust feature extraction is possible without pitch marks because of not using phase information. This source model is also used in the Nitech HTS system [49]. In this chapter, we introduce STME to maximum likelihood voice conversion based on GMM [10]. We convert both

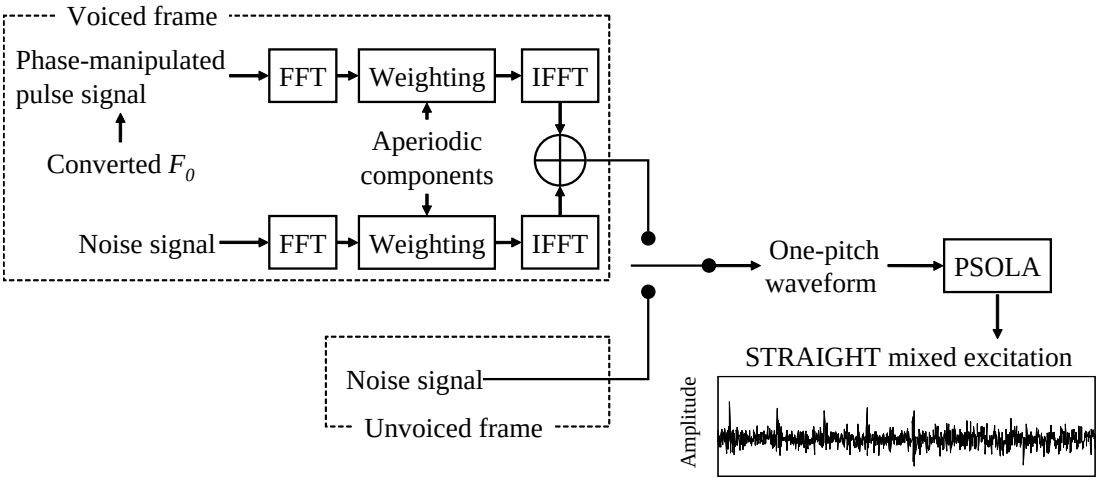


Figure 3.1. Generation process for STRAIGHT mixed excitation.

spectral and source feature sequences by MLE-based conversion algorithm. The proposed conversion's effectiveness is demonstrated through objective and subjective evaluations.

This chapter is organized as follows. In Section 3.2, we describe STME. In Section 3.3, the VC introducing STME is briefly explained, and in Section 3.4, we evaluate the experimental results. Finally, we summarize this chapter in Section 3.5.

3.2. STRAIGHT mixed excitation

Figure 3.1 shows a generation process of the STME [13]. Voiced segments of STME are defined as the frequency-dependent weighted sum of white noise and a phase-manipulated pulse train based on the converted F_0. The weight is determined based on an aperiodic component in each frequency bin [14]. On the other hand, unvoiced segments are generated white noise. Finally, an excitation signal is generated by PSOLA technique [45].

3.2.1 Aperiodic component analysis

Aperiodic component [14] is defined as the ratio between a periodic component and a noise component in each frequency band. The process of aperiodic component extraction is performed as follows: 1) A waveform which includes constant fundamental frequency is generated from an original speech waveform by DTW, 2) the time-warped waveform is transformed to frequency domain with the window function which is set zero-point between harmonic components of spectra generated by short-time Fourier transform, and 3) Log-scaled power spectra generated in 2) are liftered by suppressing quefrency components higher than fundamental period.

Figure 3.2 depicts aperiodic component extraction from a liftered power spectrum keeping the periodicity. The aperiodic component is calculated as a subtraction of an upper spectral envelope from a lower spectral envelope, where the upper one shows periodic components and the lower one represents noise components. Because the subtracted value should be less than 0 dB, the range of the aperiodic component is between 0 and 1. In the figure, aperiodicity is large when the lower envelope is close to the upper one. In order to reduce the dimensionality of the parameter to be statistically modeled, the aperiodic components are averaged on five frequency sub-bands, i.e., 0 to 1, 1 to 2, 2 to 4, 4 to 6, and 6 to 8 kHz, in the same manner as described in [49]. Figure 3.3 shows a normalized frequency distribution of aperiodic components in each frequency band. These normalized frequency distributions are calculated by using utterance data of a female speaker (FKN) which is included in A subset of ATR phoneme-balanced sentences [50]. The horizontal axis of this figure represents values of aperiodic components. Periodicity is stronger on pints closer to zero and noise component is stronger on points near to one. There is a noticeable tendency indicating that periodicity is dominant in the lower frequency bands and that aperiodicity is dominant in the higher ones.

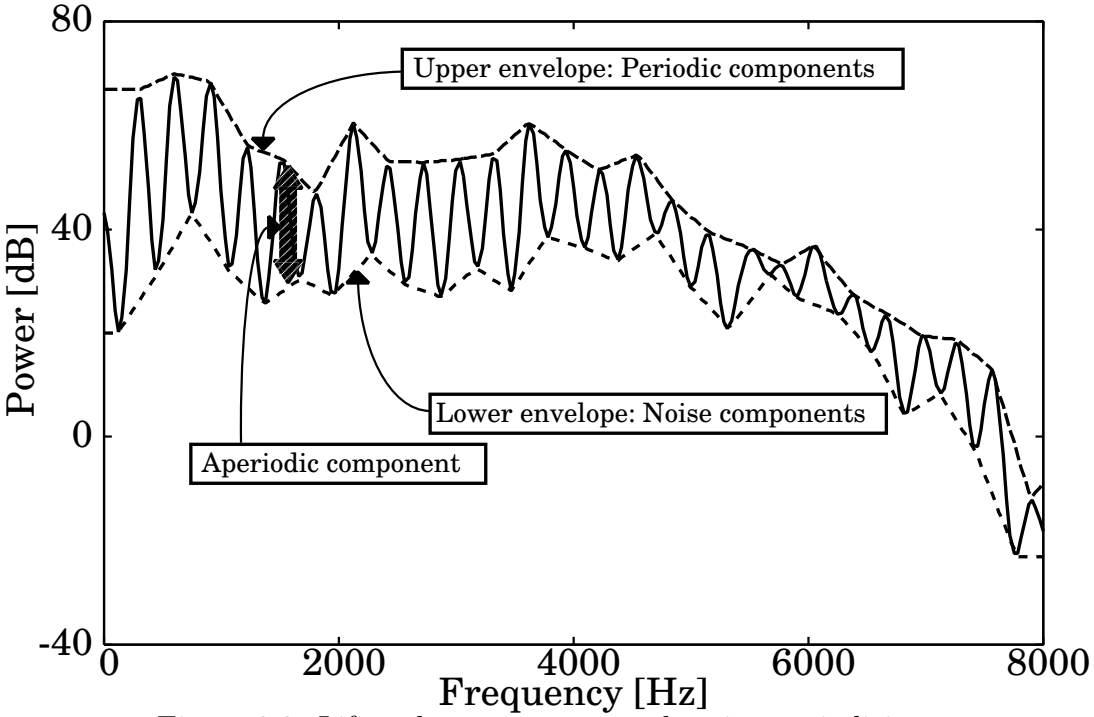


Figure 3.2. Liftered power spectrum keeping periodicity.

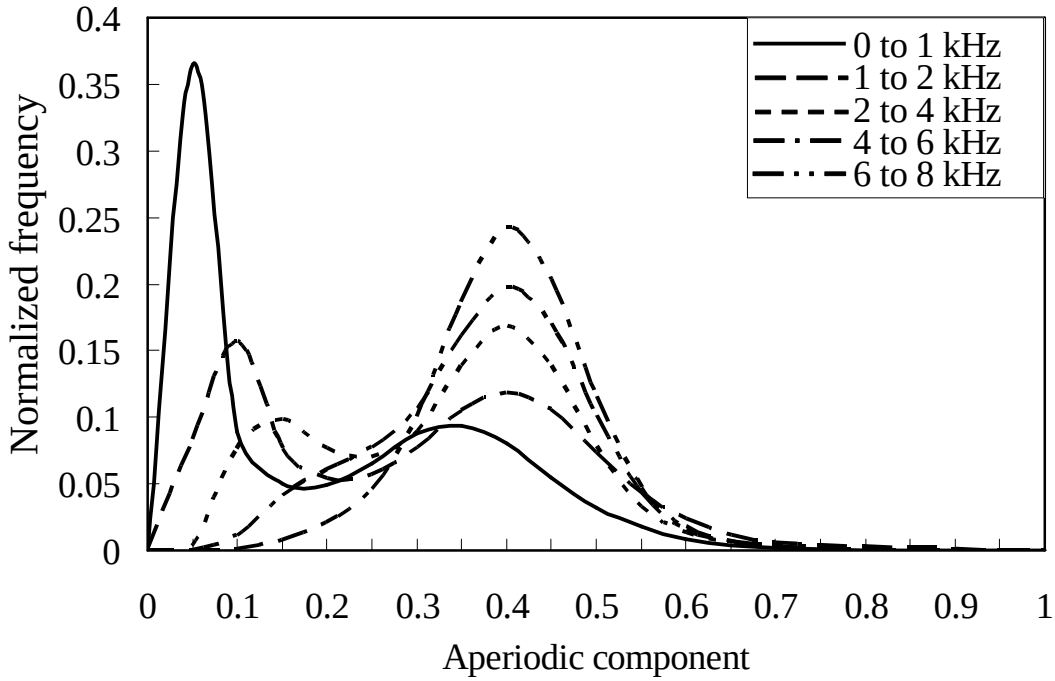


Figure 3.3. Normalized frequency distribution of aperiodic component on each frequency band.

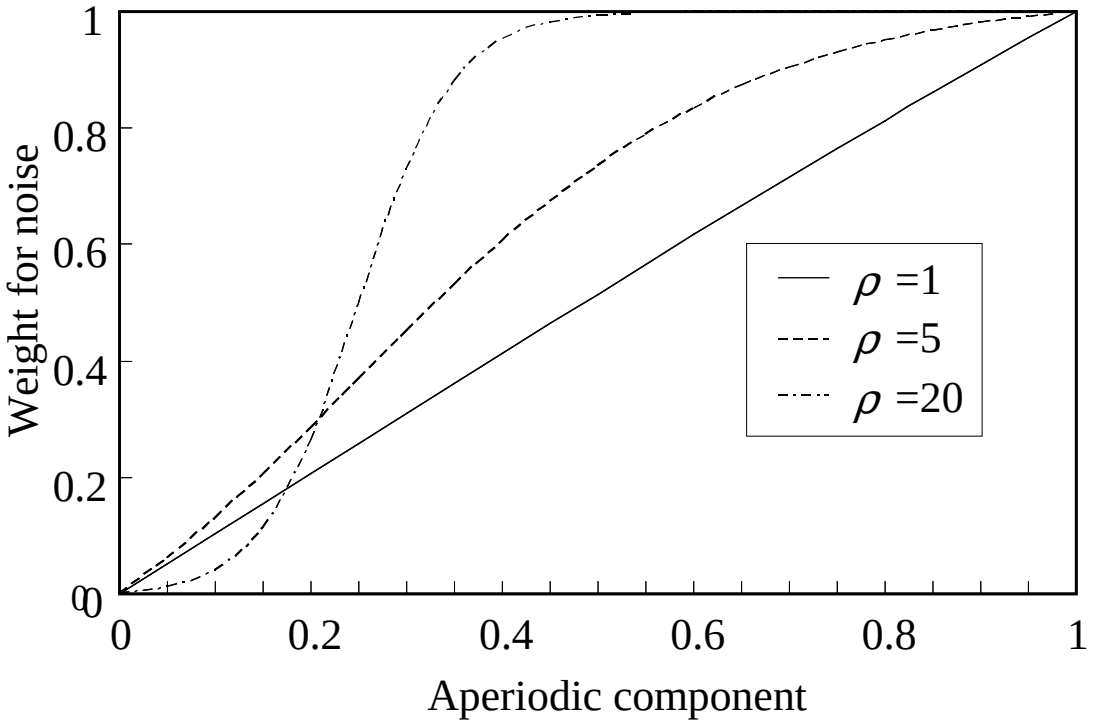


Figure 3.4. Mapping function from an aperiodic component into weight for noise when varying the mapping parameter ρ.

3.2.2 Design of excitation

The aperiodic component at each frequency bin is converted to the weight for a noise signal used in the mixed excitation as follows:

$$s(a_f) = \frac{1}{1 + \exp\{-\rho\,(a_f - 0.25)\}}, \tag{3.1}$$

$$\Omega(a_f) = \frac{s\,(a_f) - s\,(0)}{s\,(1) - s\,(0)}, \tag{3.2}$$

where a_f denotes the aperiodic component at each frequency bin and $\Omega(a_f)$ is a mapping function. This mapping function varies according to the mapping parameter ρ as shown in Figure 3.4. As the mapping component ρ is larger, the aperiodic component is mapped onto the larger weight.

The mixed excitation is defined as follows:

$$S(f) = \sqrt{1 - \Omega(a_f)^2}\,\widetilde{P}(f) + \Omega(a_f)\,N(f), \tag{3.3}$$

where $\widetilde{P}(f)$ denotes a pulse train with phase manipulation [13], and $N(f)$ denotes a white noise signal.

Figure 3.5 shows an example of a residual signal, STSE, STME and each excitation parameter, i.e., an F_0 contour and aperiodic component sequences in individual frequency sub-band. STSE is quite different from the residual signal especially at voiced segments with less periodicity, e.g., around 1.7–2.0 [sec]. Because a voiced excitation signal is generated using only the phase-manipulated pulse train, the strength of periodicity depends on the pre-defined all-pass filter for phase dispersion and it doesn't vary frame-by-frame. On the other hand, STME is more similar to the residual signal than STSE because STME is capable of modeling the strength of periodicity at voiced frames. We can observe that aperiodic components in lower frequency sub-bands are inversely correlated with the strength of periodicity of the residual signal. These properties of aperiodic components are very helpful for generating an excitation signal with more similar characteristics to the residual signal.

3.3. Voice conversion with STRAIGHT mixed excitation

In order to introduce STME to the GMM-based VC framework, aperiodic components are modeled by the GMM of joint probability density between source and target speakers as defined in Eq. (2.11).

Figure 3.6 shows the process of the proposed voice conversion. Our proposed method employs two GMMs. One is used for the spectral conversion and the other is for the aperiodic conversion. Both conversions are performed with MLE. We consider GV only in the spectral conversion because GV does not cause any large difference to the converted speech in the aperiodic conversion. We synthesize the mixed excitation from the converted aperiodic components. Finally, we synthesize the convert speech by filtering the excitation with the converted spectra.

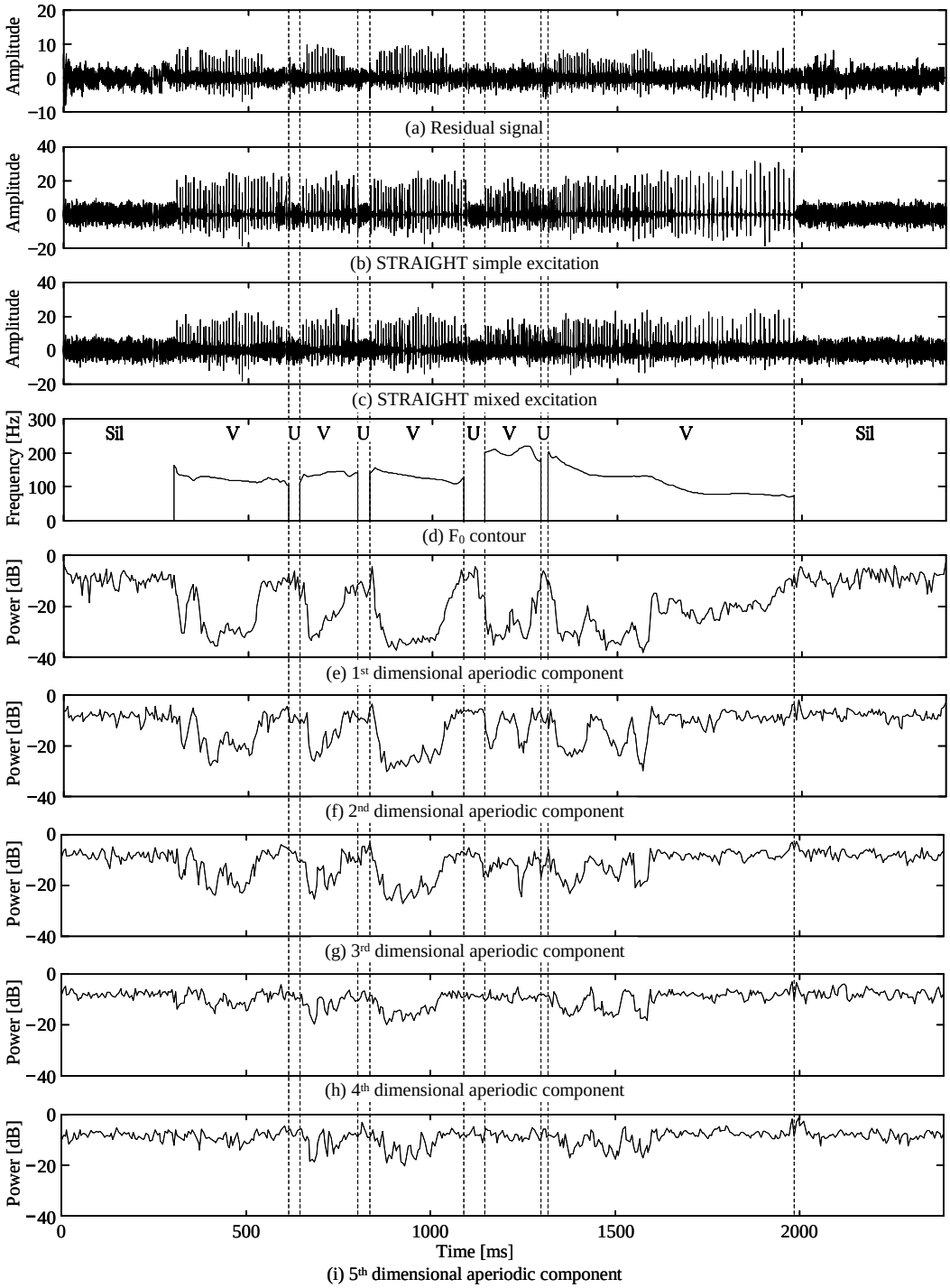


Figure 3.5. An example of (a) residual signal, (b) STRAIGHT simple excitation (STSE), (c) STRAIGHT mixed excitation (STME), (d) F_0 contour, and (e)–(i) aperiodic components. "Sil", "V" and "U" denote silent, voiced and unvoiced segments, respectively.

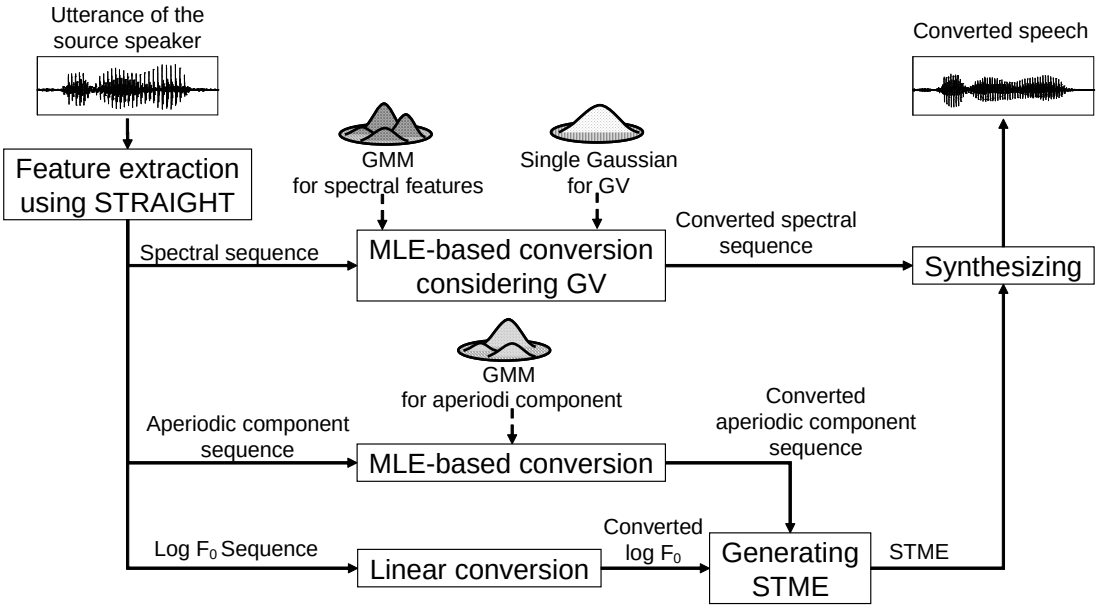


Figure 3.6. Process of the proposed voice conversion.

3.4. Experimental evaluation

3.4.1 Experimental conditions

We used the speech data of two male speakers and two female speakers from ATR's phonetically balanced sentence database [50]. We considered 50 sentences for training data, and another 50 sentences for the evaluation. The total number of combinations of source and target speakers was 12.

For the spectral feature, we take the first through the 24^{th} mel-cepstral coefficients from the STRAIGHT smoothed spectrum. For the aperiodic feature, we used average dB values of the aperiodic components on five frequency bands described in Section 3.2.1.

In each feature conversion, we used full covariance matrices, and set the number of mixtures for the spectral conversion to 32 based on our preliminary experiment.

As an objective metric for aperiodic component, we employ RMSE between

two aperiodic component sequences written as follows:

$$\text{RMSE [dB]} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{5}\sum_{f=1}^{5}\left(a_{f,t}^{(X)} - a_{f,t}^{(Y)}\right)^2}, \tag{3.4}$$

where $a_{f,t}^{(X)}$ and $a_{f,t}^{(Y)}$ denote time-aligned source and target aperiodic components on the f^{th} frequency band at t frame, respectively.

3.4.2 Optimization of mapping parameter

To optimize the mapping parameter ρ for each speaker, we evaluated the aperiodic component distortion between natural speech and analysis-synthesized speech. Figure 3.7 shows the aperiodic component distortion as a function of the mapping parameter ρ. It is apparent that 8 is the optimal value for every speaker, thus we designed STRAIGHT mixed excitation using this value.

To demonstrate the effectiveness of the mixed excitation in the analysis-synthesis, we evaluated the speech quality of natural speech, analysis-synthesized speech without mixed excitation, and analysis-synthesized speech with mixed excitation. Figure 3.8 shows the result of a preference test. The number of listeners in this case was five. The figure shows that the speech quality of analysis-synthesized speech using mixed excitation is higher than that without mixed excitation

3.4.3 Objective evaluation

We evaluated the distortion between the target aperiodic component and the converted one. Figure 3.9 shows RMSE on aperiodic components as a function of the number of mixtures. The aperiodic conversion causes a reduction of the aperiodic distortion. Therefore, the conversion causes the source signal to have characteristics much more similar to those of the target speaker, compared to those of the source speaker. The optimum number of mixtures is 32. However, it is shown that the conversion performance is not very sensitive to the number of mixtures.

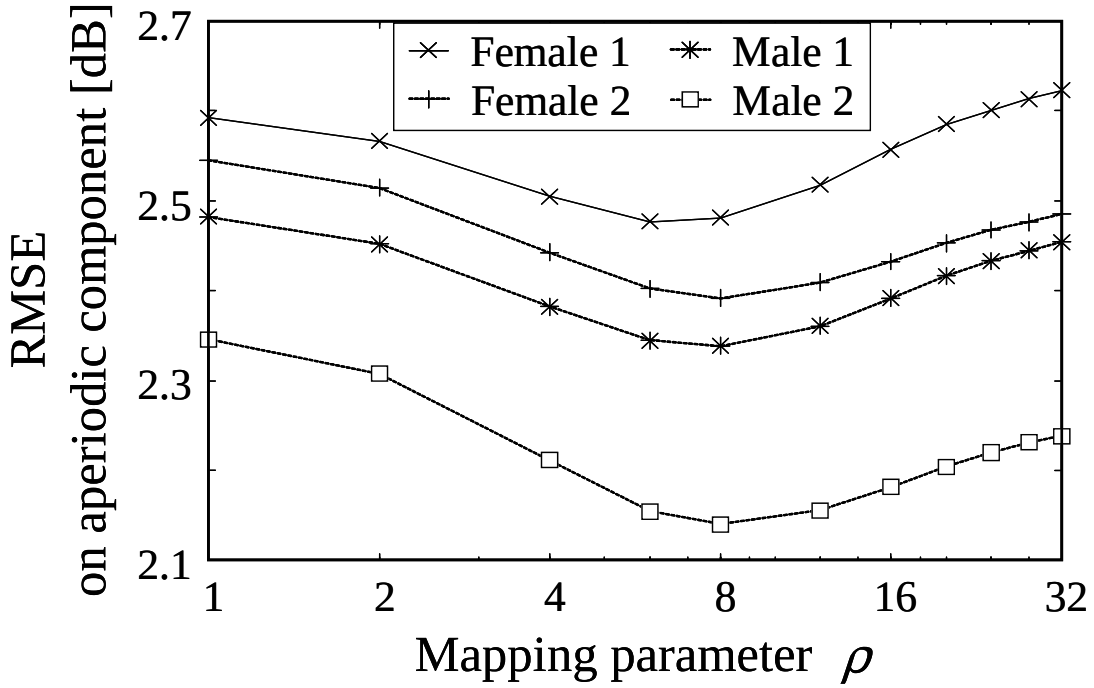


Figure 3.7. The aperiodic component distortion as a function of the mapping parameter ρ.

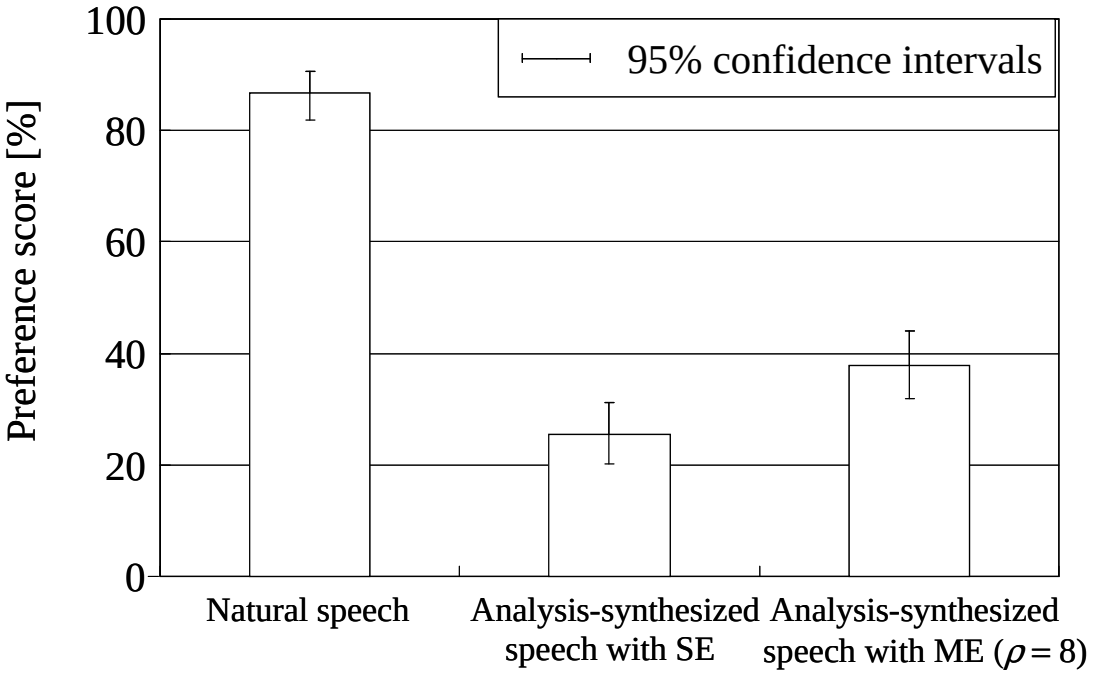


Figure 3.8. Result of preference test on speech quality comparing natural speech, analysis-synthesized speech includes STSE, with that includes STME.

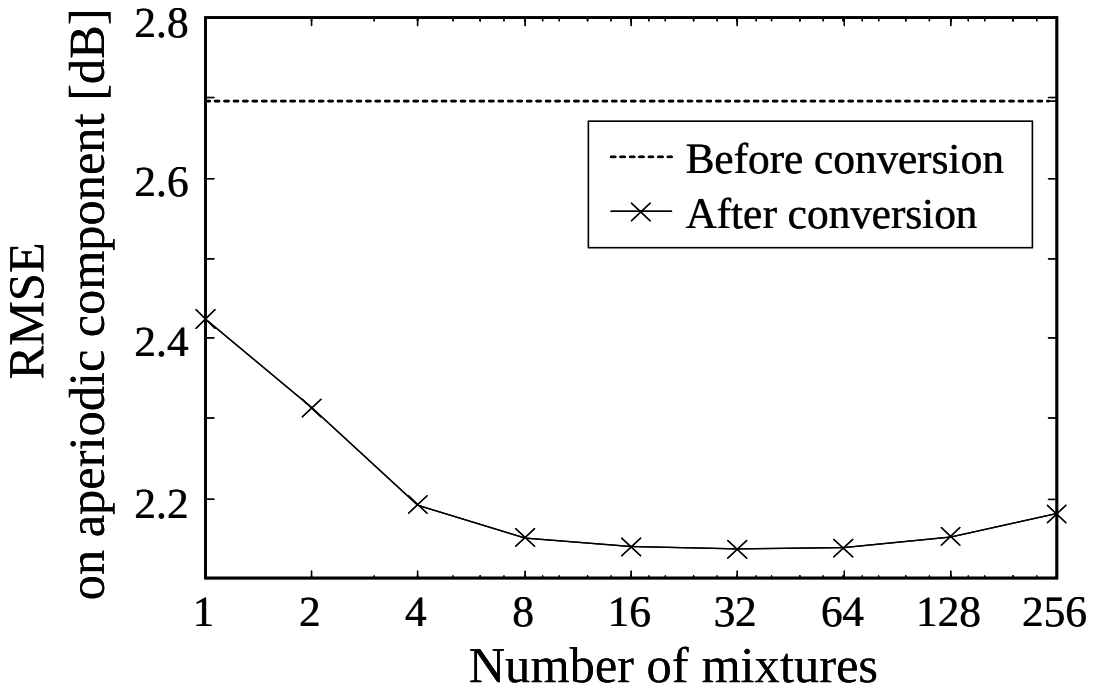


Figure 3.9. The aperiodic component distortion as a function of the number of mixture components.

3.4.4 Subjective evaluation

We subjectively evaluated the converted speech quality and the conversion accuracy for the speaker individuality. In this evaluation, we employed the following converted voices: 1) converted voice with STSE; 2) converted voice with STME based on source speaker's aperiodic component; and 3) converted voice with the mixed excitation based on the converted aperiodic component.

In the preference test for speech quality, we randomly presented a pair of voices from three kinds of voice to listeners. In the XAB test on speaker individuality, we presented the target speaker's voice and after that a pair of converted voices randomly. Then we asked listeners which converted voice is similar to the target speaker's. The number of listeners was eight and each listener evaluated 72 sample-pairs which combine 12 types of voices and three types of conversion methods.

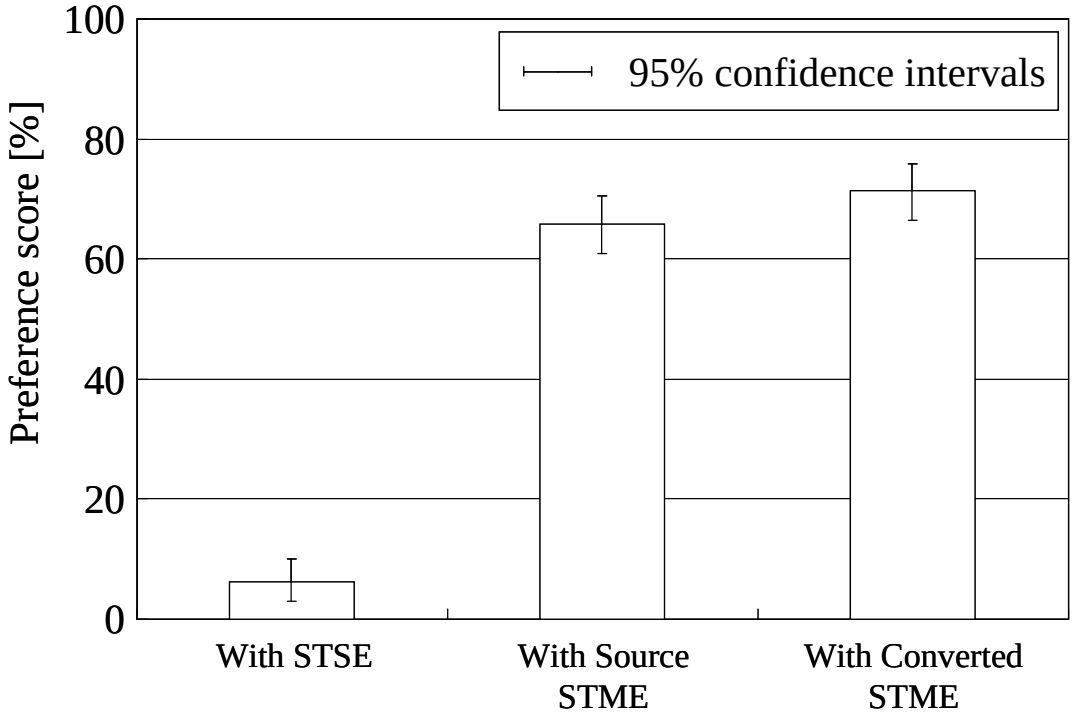


Figure 3.10. Result of preference test on speech quality on STME evaluation.

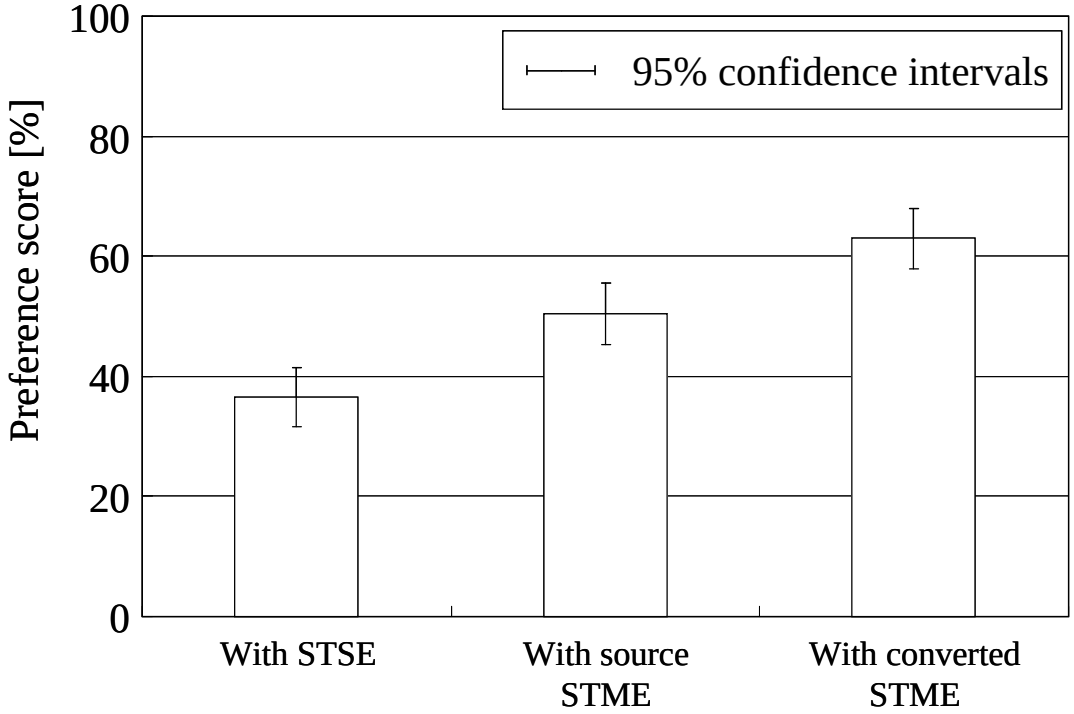


Figure 3.11. Result of preference test on conversion accuracy for speaker individuality on STME evaluation.

Figure 3.10 shows the result of the preference test. The STRAIGHT mixed excitation greatly improved speech quality when using mixed excitation. Moreover, the results reveal that the aperiodic conversion slightly improves the converted speech quality.

Figure 3.11 shows the result of the XAB test. We can see that the conversion accuracy for speaker individuality was also improved by using STRAIGHT mixed excitation. In addition, we can improve it further by converting the aperiodic components.

From these results, it is possible that the improvement of the converted speech quality using STME affects that of conversion accuracy, because the converted speech with source STME, which is not reflected target voice quality, outperforms that with STSE, which is also not reflected target voice quality. However, there are significant differences between converted speech with source STME and with converted STME in the speech quality test. Therefore, we consider that the significant improvement of conversion accuracy for speaker individuality is caused by the factor which differs from speech quality. To bring out the improving factor, we compared converted speeches with converted STME with those with source STME by a preference test. In this result, preference score of converted speeches with converted STME is 61.46% (95% confidence interval is 54.18% to 68.38%). Therefore, the converted STME outperforms the source STME significantly on conversion accuracy for speaker individuality. Thus, aperiodic components include speaker-dependent characteristics and the statistical conversion of aperiodic components achieves the generation of the more proper excitation model. Objective and subjective results demonstrate that converted speech is improved by introducing STME to the GMM-based VC framework.

3.5. Summary

In this chapter, we have introduced STRAIGHT mixed excitation (STME) to MLE-based voice conversion with a Gaussian Mixture model (GMM) in order to improve the converted speech quality and the conversion accuracy for speaker individuality. The STME is constructed by the weighted sum of white noise and a phase-manipulated pulse train in each frequency bin. The weight is determined

based on an aperiodic component which represents the ratio between a periodic component and a noise component in each frequency band. We have statistically converted an aperiodic component sequence as well as a spectral sequence.

In addition, we have subjectively evaluated the proposed conversion method, finding that the proposed method improved both converted speech quality and conversion accuracy for speaker individuality.

Chapter 4

Adaptive Training for Eigenvoice Conversion

In this chapter, we describe a novel model training method for EVC. In the conventional one-to-many EVC framework, the canonical EV-GMM is based on the target-speaker-independent GMM, which includes inter-speaker variation of the pre-stored target speakers. Therefore, this is one of the factors that degrade conversion performance in the conventional EVC. In order to improve the conversion performance in one-to-many EVC, we propose an adaptive training method of the EV-GMM. In the proposed training method, both of the fixed parameters and the adaptive parameters are optimized by maximizing a total likelihood function of the EV-GMMs adapted to individual pre-stored target speakers. Moreover, we also propose improved adaptive training methods to alleviate the problem which is caused with a small number of representative vectors. We conducted objective and subjective evaluations to demonstrate the effectiveness of the proposed training method. The experimental results show that the proposed adaptive training yields significant quality improvements in the converted speech.

4.1. Introduction

As a method to alleviate limitations of the conversion model training, we have proposed EVC [11][12]. In the conventional training method of the EV-GMM, we build the EV-GMM using parameters from a TI-GMM. These parameters are

strongly affected by acoustic variations among many pre-stored target speakers. They usually cause significant degradation of the conversion performance of the adapted EV-GMM.

This is a well-known problem often observed in speech recognition. In general, a speaker-dependent acoustic model is often constructed by adapting speaker-independent model to a desired speaker using maximum likelihood linear regression (MLLR) [51], MAP adaptation [34], and so on, because we need a large amount of speech data for achieving better speech recognition. Although the speaker-independent model includes various acoustic information, it also includes acoustic variations among speakers, which is a factor that degrades speech recognition ratio. One of the alleviating methods is to use a pseudo-normalized speaker model, called the canonical model, rather than a speaker-independent model as an initial model for speaker adaptation. It has been reported that adaptive training, such as speaker adaptive training (SAT) [15] or cluster adaptive training (CAT) [52], is a very effective paradigm for training the canonical model. In SAT, proposed by Anastasakos et al., the canonical model parameters are estimated by maximizing a total likelihood of adapted models, which are adapted to pre-stored speakers for training by using MLLR in a supervised manner. CAT, proposed by Gales, is an extension method of speaker cluster scheme. In this framework, the canonical model parameters are determined by maximizing likelihoods of adapted models, which are constructed from data included in each cluster. Thus, CAT framework includes SAT framework.

Inspired by these studies, we propose an adaptive training method of the EV-GMM in one-to-many EVC.Moreover, we propose methods for alleviating the local optimum problem often caused in the proposed adaptive training when the number of adaptive parameters is set to be low. The experimental results of objective and subjective evaluations demonstrate that the proposed adaptive training yields significant quality improvements in converted speech.

This chapter is organized as follows. In Section 4.2, we describe basic algorithms of the proposed adaptive training method for the EV-GMM. In Section 4.3, the problem of the basic adaptive training is described. In Section 4.4, we describe improved adaptive training methods. In Section 4.5, we discuss our proposed adaptive training method by comparing with other adaptive training

methods in speech recognition area. In Section 4.6, we describe experimental evaluations. Finally, this chapter is summarized in Section 4.7.

4.2. Basic adaptive training algorithm

In order to alleviate the mismatch issues observed in the PCA-based EV-GMM, we propose an adaptive training method for the one-to-many EV-GMM. A canonical EV-GMM for the EV-GMM adaptation is trained in the adaptive training paradigm so that the performance of the adapted EV-GMMs is improved.

The canonical EV-GMM is trained by maximizing the total likelihood of the adapted EV-GMMs for individual pre-stored target speakers with respect to both canonical EV-GMM parameters (i.e., the tied-parameters of the EV-GMM $\lambda^{(EV)}$) and target-speaker adaptive parameters (i.e., the weight vector $\boldsymbol{w}^{(s)}$ for each pre-stored target speaker) as follows:

$$\left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S\right\} = \arg\max_{\lambda^{(EV)}, \boldsymbol{w}_1^S} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{X}_t \boldsymbol{Y}_t^{(s)} | \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right), \tag{4.1}$$

where $\boldsymbol{w}_1^S$ is a set of weight vectors for individual pre-stored target speakers $\left\{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \cdots, \boldsymbol{w}^{(S)}\right\}$. The training process is performed with the EM algorithm [43] by maximizing the following auxiliary function,

$$\begin{aligned} &Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S\right\}\right) \\ &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right) \log P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m|\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(s)}\right), \end{aligned} \tag{4.2}$$

where the first variable in the auxiliary function (i.e., $\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}$ in Eq. (6.18)) is a parameter set used in E-step for calculating the following posterior probability $P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right)$ and the other variable in it (i.e., $\left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S\right\}$ in Eq. (6.18)) is a parameter set to be updated in M-step, which is used for calculating the log-scaled likelihood $\log P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m|\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(s)}\right)$. It is difficult to update all parameters simultaneously because some of them depend on each

other. Therefore, each parameter of EV-GMM is updated as follows:

$$
\begin{aligned}
&Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \ \left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}\right) \\
&\quad \leq Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\alpha_m, \boldsymbol{\Sigma}_m^{(X,Y)}, \boldsymbol{B}_m, \boldsymbol{b}_m^{(0)}, \boldsymbol{\mu}_m^{(X)}, \hat{\boldsymbol{w}}_1^S\right\}\right) \\
&\quad \leq Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\alpha_m, \boldsymbol{\Sigma}_m^{(X,Y)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}, \hat{\boldsymbol{w}}_1^S\right\}\right) \\
&\quad \leq Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\hat{\alpha}_m, \hat{\boldsymbol{\Sigma}}_m^{(X,Y)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}, \hat{\boldsymbol{w}}_1^S\right\}\right). \qquad (4.3)
\end{aligned}
$$

It is sufficient to ensure these updates to satisfy

$$
\begin{aligned}
&\prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}, m | \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right) \\
&\quad \leq \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}, m | \left\{\alpha_m, \boldsymbol{\Sigma}_m^{(ZZ)}, \boldsymbol{B}_m, \boldsymbol{b}_m^{(0)}, \boldsymbol{\mu}_m^{(X)}\right\}, \hat{\boldsymbol{w}}^{(s)}\right) \\
&\quad \leq \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}, m | \left\{\alpha_m, \boldsymbol{\Sigma}_m^{(ZZ)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}\right\}, \hat{\boldsymbol{w}}^{(s)}\right) \\
&\quad \leq \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}, m | \left\{\hat{\alpha}_m, \hat{\boldsymbol{\Sigma}}_m^{(ZZ)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}\right\}, \hat{\boldsymbol{w}}^{(s)}\right). \qquad (4.4)
\end{aligned}
$$

These update processes are iteratively performed in each M-step for improving parameter estimation accuracy. ML estimate of the weight vector for the s^{th} pre-stored target speaker is written as

$$
\begin{aligned}
\hat{\boldsymbol{w}}^{(s)} = &\left(\sum_{m=1}^{M} \overline{\gamma}_m^{(s)} \boldsymbol{B}_m^\top \boldsymbol{P}_m^{(YY)} \boldsymbol{B}_m\right)^{-1} \\
&\times \left[\sum_{m=1}^{M} \boldsymbol{B}_m^\top \left\{\boldsymbol{P}_m^{(YX)}\left(\overline{\boldsymbol{X}}_m^{(s)} - \overline{\gamma}_m^{(s)} \boldsymbol{\mu}_m^{(X)}\right) + \boldsymbol{P}_m^{(YY)}\left(\overline{\boldsymbol{Y}}_m^{(s)} - \overline{\gamma}_m^{(s)} \boldsymbol{b}_m^{(0)}\right)\right\}\right], \qquad (4.5)
\end{aligned}
$$

where

$$\overline{\gamma}_m^{(s)} = \sum_{t=1}^{T_s} P\Big(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big), \tag{4.6}$$

$$\overline{\boldsymbol{X}}_m^{(s)} = \sum_{t=1}^{Ts} P\Big(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)\, \boldsymbol{X}_t, \tag{4.7}$$

$$\overline{\boldsymbol{Y}}_m^{(s)} = \sum_{t=1}^{Ts} P\Big(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)\, \boldsymbol{Y}_t^{(s)}, \tag{4.8}$$

$$\boldsymbol{\Sigma}_m^{(X,Y)^{-1}} = \begin{bmatrix} \boldsymbol{P}_m^{(XX)} & \boldsymbol{P}_m^{(XY)} \\ \boldsymbol{P}_m^{(YX)} & \boldsymbol{P}_m^{(YY)} \end{bmatrix}. \tag{4.9}$$

ML estimates of the tied-parameters for mean vectors are written as

$$\hat{\boldsymbol{\nu}}_m = \left(\sum_{s=1}^{S} \overline{\gamma}_m^{(s)} \hat{\boldsymbol{W}}_s{}^\top \boldsymbol{\Sigma}_m^{(X,Y)^{-1}} \hat{\boldsymbol{W}}_s\right)^{-1} \left(\sum_{s=1}^{S} \hat{\boldsymbol{W}}_s{}^\top \boldsymbol{\Sigma}_m^{(X,Y)^{-1}} \overline{\boldsymbol{Z}}_m^{(s)}\right), \tag{4.10}$$

where

$$\overline{\boldsymbol{Z}}_m^{(s)} = \left[\overline{\boldsymbol{X}}_t^{(s)^\top}, \overline{\boldsymbol{Y}}_t^{(s)^\top}\right]^\top \tag{4.11}$$

$$\hat{\boldsymbol{\nu}}_m = \left[\hat{\boldsymbol{\mu}}_m^{(X)^\top}, \hat{\boldsymbol{b}}_m^{(0)^\top}, \hat{\boldsymbol{b}}_m^{(1)^\top}, \ \cdots, \ \hat{\boldsymbol{b}}_m^{(J)^\top}\right]^\top, \tag{4.12}$$

$$\hat{\boldsymbol{W}}_s = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \hat{w}_1^{(s)}\boldsymbol{I} & \hat{w}_2^{(s)}\boldsymbol{I} & \cdots & \hat{w}_J^{(s)}\boldsymbol{I} \end{bmatrix}. \tag{4.13}$$

Then, mixture component weights and covariance matrices are determined as follows:

$$\hat{\alpha}_m = \frac{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}{\sum_{m=1}^{M}\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}, \tag{4.14}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(X,Y)} = \frac{1}{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}} \sum_{s=1}^{S} \left\{ \overline{\boldsymbol{V}}_{m,s}^{(X,Y)} + \overline{\gamma}_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)^\top} - \left(\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} \overline{\boldsymbol{Z}}_m^{(s)^\top} + \overline{\boldsymbol{Z}}_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)^\top} \right) \right\}, \tag{4.15}$$

where

$$\overline{\boldsymbol{V}}_{m,s}^{(X,Y)} = \sum_{t=1}^{T_s} P\Big(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big) \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}\right] \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}\right]^\top, \quad (4.16)$$

$$\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} = \hat{\boldsymbol{W}}_s \, \hat{\boldsymbol{\nu}}_m = \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(X)} \\ \hat{\boldsymbol{B}}_m \hat{\boldsymbol{w}}^{(s)} + \hat{\boldsymbol{b}}_m^{(0)} \end{bmatrix}. \quad (4.17)$$

Note that each estimating process is described Appendix A in detail.

4.3. Local optimum problem of adaptive training

In the canonical EV-GMM training, it is essential to estimate the representative vectors spanning a sub-space effectively modeling acoustic variations among all of pre-stored target speakers in each mixture component. To achieve this, individual mixture-component occupancies calculated by Eq. (4.6) have to be largely enough for every pre-stored target speakers. Figure 4.1 shows an example of the occupancies for one pre-stored target speaker, which are sorted in descending order. We can see that the occupancies calculated with the TI-GMM are more biased compared with those calculated with the TD-GMM for the same pre-stored target speaker. Because the TI-GMM needs to model wide varieties of acoustic features of all pre-stored target speakers, some mixture components model only acoustic features of a part of pre-stored target speakers. Consequently, a larger number of mixture components with lack of occupancies are observed in the TI-GMM, compared to the TD-GMM. Figure 4.1 also shows the occupancies of the canonical EV-GMMs with 159 representative vectors and with only one representative vector, which are iteratively updated from the TI-GMM with the EM algorithm. Even if the occupancies are biased in the first E-step as observed in the TI-GMM, the trained canonical EV-GMM with a largely enough number of representative vectors has an occupancy distribution similar to that of the TD-GMM. However, if the number of representative vectors is very small, the occupancy distribution of the trained canonical EV-GMM remains biased. Namely, the initial occupancies strongly affect the final canonical EV-GMM unless the number of representative vectors is large enough to precisely model acoustic features of individual pre-stored target speakers. This problem is critical because there are some situations

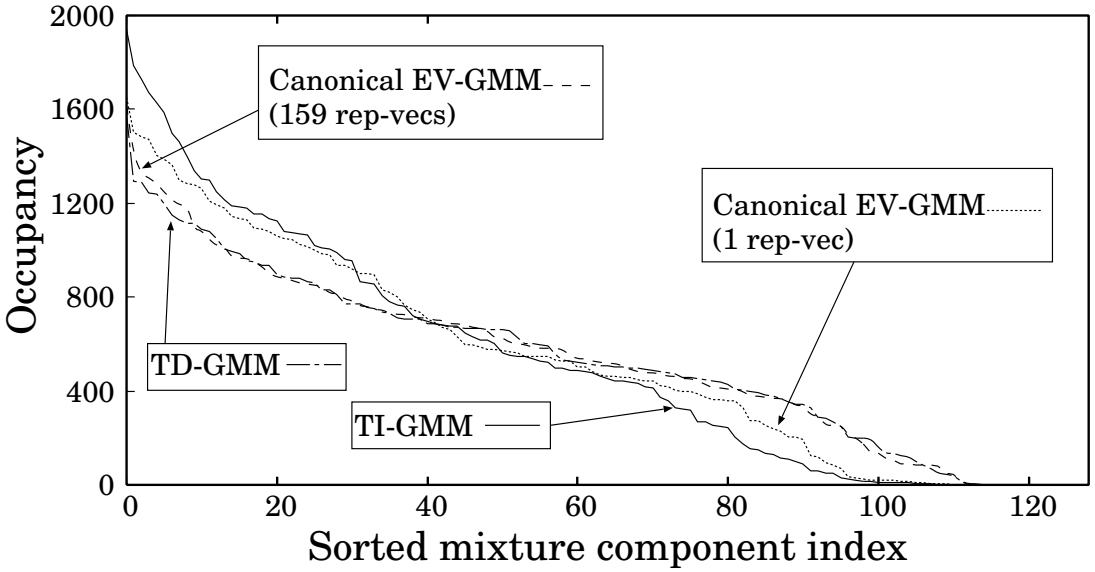


Figure 4.1. Mixture-component occupancies for one pre-stored speaker using several models.

where we prefer to train the EV-GMM with a small number of representative vectors, e.g., for keeping the computational cost of the adaptation process as low as possible or for reducing the model size as much as possible. In order to alleviate this local optimum problem, we propose two approaches: 1) the TD-GMMs using for calculating the occupancies in the first E-step; and 2) the deterministic annealing EM (DAEM) algorithm using for the EV-GMM.

4.4. Improved adaptive training of alleviating local optimum problem

4.4.1 First E-step approximation with target-speaker-dependent models

Each of the TD-GMMs models the joint probability density for the source speaker and each pre-stored target speaker. Therefore, they generally yield more unbiased

occupancies in each mixture component for every pre-stored speaker as shown in Figure 4.1. The use of these occupancies for estimating the representative vectors is supposed to be helpful for alleviating the local optimum problem. Therefore, we propose an occupancy approximation method using the TD-GMMs for calculating the occupancies only in the first E-step. Note that the correspondence of each mixture component between every TD-GMM and the PCA-based EV-GMM is known because only target mean vectors are updated with parallel data in building each TD-GMM in order to preserve the correspondence of each mixture component in a phonemic space [12]. We update the PCA-based EV-GMM parameters in the first M-step based on the occupancies calculated with the TD-GMMs. Therefore, the first E-step and M-step are no longer regarded as the EM algorithm. In all steps that follow, we use the updated EV-GMM parameters for calculating the occupancies as in the EM algorithm.

4.4.2 Deterministic annealing EM algorithm

In order to alleviate the local optimum problem, we apply the deterministic annealing EM (DAEM) algorithm [53] to the adaptive training of the EV-GMM. The DAEM algorithm reformulates a maximization process of a likelihood function as a minimization process of free energy. In adaptive training of the EV-GMM based on the DAEM algorithm, parameters are estimated as follows:

$$\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S = \arg\min_{\lambda^{(EV)}, \boldsymbol{w}_1^S} F_\beta \tag{4.18}$$

$$F_\beta = -\frac{1}{\beta} \log \prod_{s=1}^{S} \prod_{t=1}^{T_s} \sum_{m=1}^{M} P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m | \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)^\beta, \tag{4.19}$$

where $\frac{1}{\beta}$ is called the "temperature." The free energy given by Eq. (4.19) is minimized by maximizing the following auxiliary function,

$$\begin{aligned} &Q_\beta\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S\right\}\right) \\ &\quad = \sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{m=1}^{M} \gamma_{m,t,\beta}^{(s)} \log P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m | \hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(s)}\Big), \end{aligned} \tag{4.20}$$

where

$$\gamma_{m,t,\beta}^{(s)} = \frac{P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m|\lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)^{\beta}}{\sum_{m=1}^{M} P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m|\lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)^{\beta}}. \tag{4.21}$$

In the algorithm, we gradually decrease the temperature and deterministically optimize the function at each temperature. First β is set to nearly zero so that the free energy function has a single global maximum, and the canonical model parameters and the adaptive parameters are iteratively updated by maximizing the auxiliary function. Then, we decrease $\frac{1}{\beta}$ and update the parameters while fixing β. As β gradually changes from nearly zero to 1, the posterior distribution shown in Eq. (4.21) also gradually changes from a nearly uniform distribution to the original posterior distribution. Finally, we decrease $\frac{1}{\beta}$ to 1 and update the parameters by maximizing the auxiliary function in Eq. (4.20), which is equivalent to the auxiliary function used in the EM algorithm.

4.5. Discussion

CAT [52] has been proposed as an adaptive training method for HMM-based speech recognition using the eigenvoice technique. There are some differences between CAT and the proposed method. The proposed adaptive training adapts only parameters on a part of the model space, i.e., the target feature space rather than the joint space. Consequently, the update formulas shown in Eq. (4.5) and (4.10) are different from those used in CAT. Moreover, the proposed method uses a GMM rather than HMMs. This makes an adaptation process much easier because the decoding process is inevitable if using HMMs. Consequently, the proposed method enables unsupervised adaptation in a linguistically independent manner.

Table 4.1. Number of pre-stored target speakers uttering each subset A, B, $\cdots$, or G. Each subset consists of 50 phonetically balanced sentences

Sub-sets	A	B	C	D	E	F	G	Total
Number of male speakers	15	11	15	13	15	11	0	80
Number of female speakers	15	11	15	13	12	0	14	80

4.6. Experimental evaluations

4.6.1 Experimental conditions

We objectively and subjectively evaluated the conversion performance of the proposed canonical EV-GMM compared with that of the conventional PCA-based EV-GMM in one-to-many EVC. We used parallel data sets of a single source male speaker and 160 pre-stored target speakers consisting of 80 male and 80 female speakers for training the EV-GMM. These speakers were included in the Japanese Newspaper Article Sentences (JNAS) database [54]. Each pre-stored target speaker uttered 50 phoneme-balanced sentences included in one of seven subsets as shown in Table 4.1. The source male speaker was not included in JNAS and uttered all of the seven subsets and an additional subset used for evaluation. We prepared parallel data sets between the source and each pre-stored target speaker by performing DTW automatically.

In evaluation, we used 10 target speakers consisting of five male and five female speakers not included in the pre-stored target speakers. We used 1 to 32 utterances for adapting the EV-GMM, and 21 utterances for evaluation. In the first E-step to estimate the adaptive parameters, i.e., the weight vector, we used TI-GMM as an initial model.

We used 24-dimensional mel-cepstrum as a spectral feature, which was extracted from smoothed spectrum analyzed by STRAIGHT [13]. We trained several EV-GMMs while changing the number of representative vectors as shown in Table 4.2. The number of mixture components was 128.

Table 4.2. Relationship between number of representative vectors and contribution rate

Number of representative vectors	Contribution rate [%]
1	22.49
3	42.85
8	61.95
26	80.00
159	100.0

4.6.2 Objective evaluations

To investigate the effectiveness of the proposed adaptive training for alleviating the mismatches of probability density as mentioned in Section 2.4.3, we compared static feature components of the target covariance matrices $\mathbf{\Sigma}_m^{(YY)}$ of the conventional PCA-based EV-GMM with those of the proposed canonical EV-GMM. In the proposed adaptive training, we set the number of representative vectors to 159 and used the TI-GMM as an initial model. Figure 4.2 shows mean values of those covariance components over all mixture components. It also shows those values averaged over traditional GMMs separately trained using individual parallel data sets of the source and the target speakers. The covariance values of the PCA-based EV-GMM are larger than those of the traditional GMMs because the PCA-based EV-GMM models acoustic variations among many pre-stored target speakers. On the other hand, the covariance values of the canonical EV-GMM are almost equal to those of the traditional GMMs. This result shows that the proposed adaptive training is capable of effectively reducing the influence of the inter-speaker variations on the EV-GMM training.

To demonstrate the effectiveness of the proposed adaptive training in spectral conversion accuracy, we evaluated mel-cepstral distortion between the target and converted features when using the conventional PCA-based EV-GMM and the

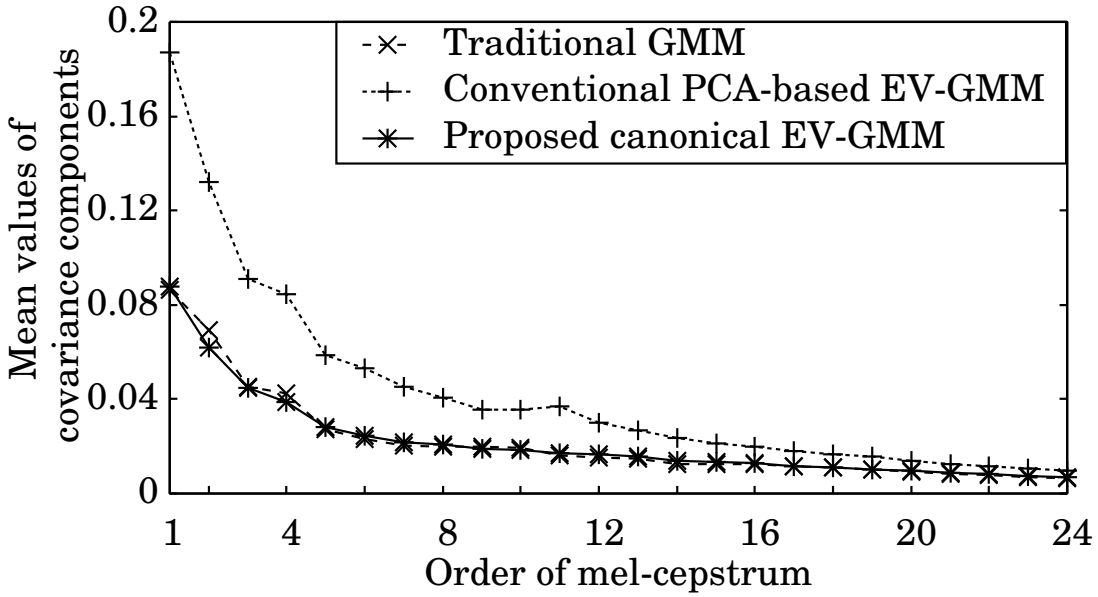


Figure 4.2. Mean values of target covariance components of individual GMMs.

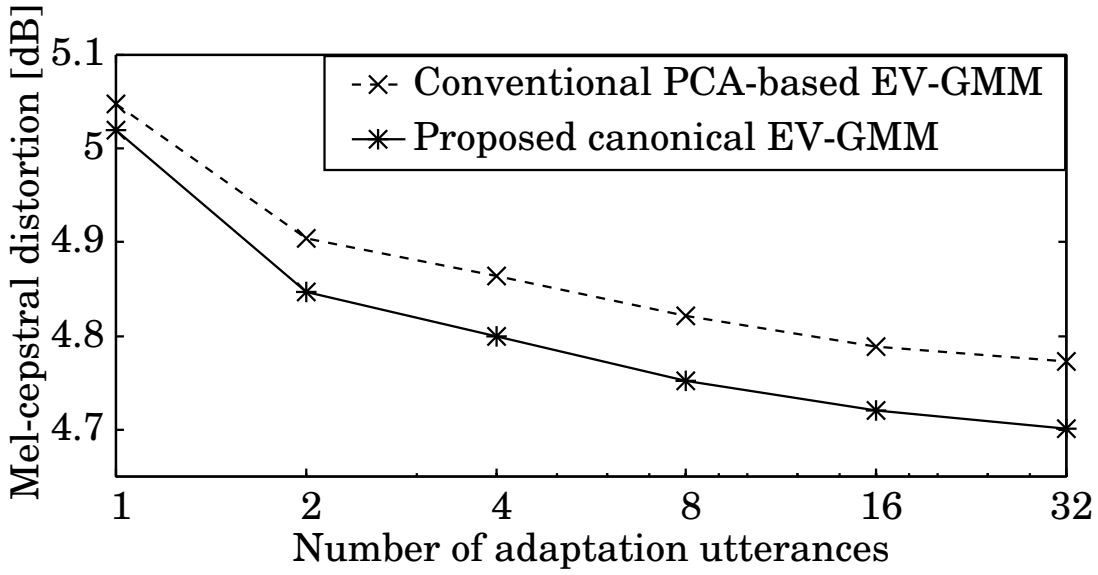


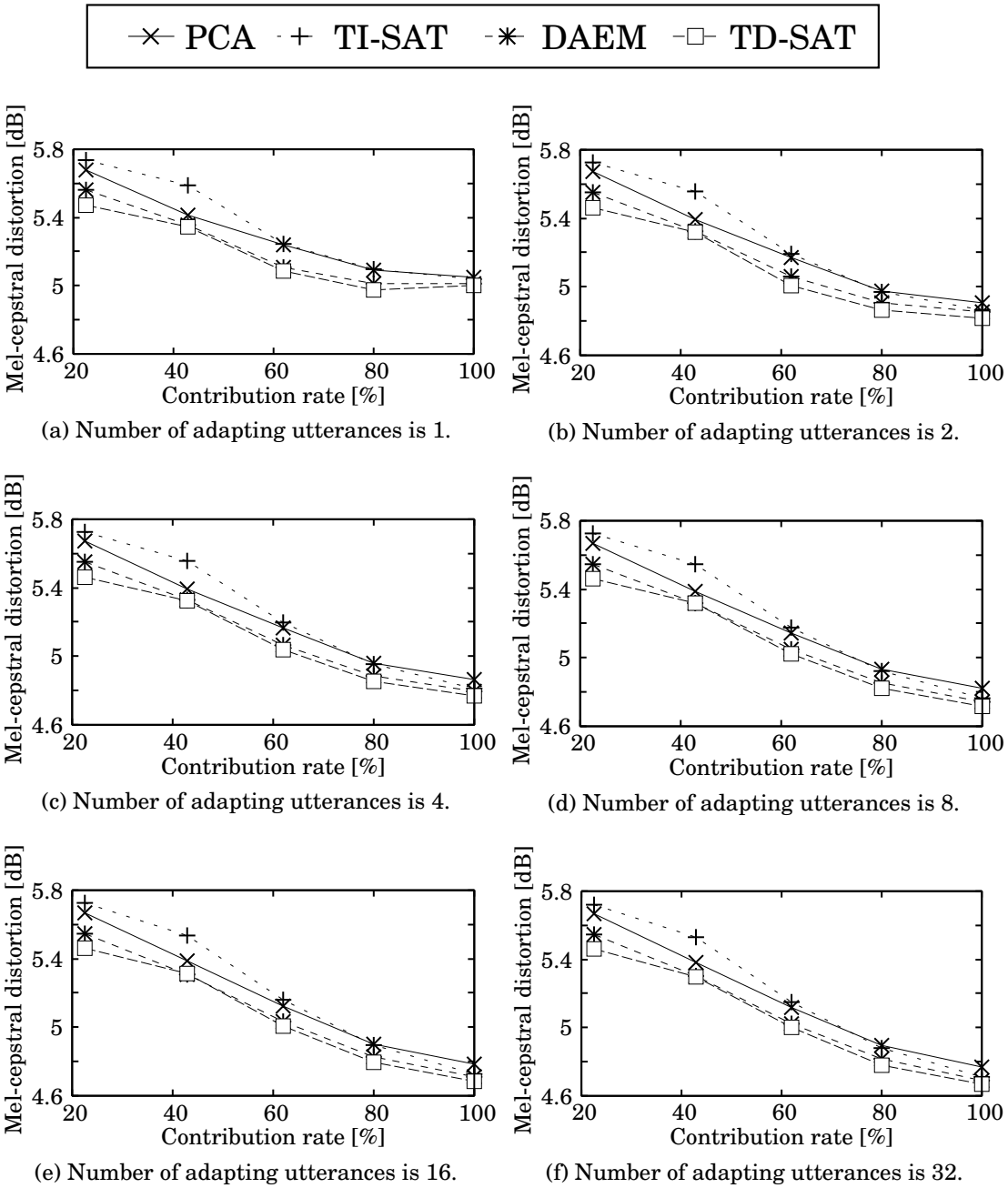
Figure 4.3. Mel-cepstral distortion as a function of the number of adaptation utterances.

proposed canonical EV-GMM. Mel-cepstral distortion is calculated as

$$\text{Mel-CD [dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(X)} - mc_d^{(Y)}\right)}, \tag{4.22}$$

where $mc_d^{(X)}$ and $mc_d^{(Y)}$ represent the d^{th} dimensional component of converted mel-cepstrum and that of target mel-cepstrum, respectively. Note that the average of mel-cepstral distortion between source and target speakers is 8.11 [dB]. These EV-GMMs were the same as those evaluated in Figure 4.2. Figure 4.3 shows mel-cepstral distortion as a function of the number of adaptation utterances. We can see that the proposed canonical EV-GMM always outperforms the conventional PCA-based EV-GMM. Therefore, the proposed adaptive training is effective for improving spectral conversion accuracy in one-to-many EVC.

To investigate the influence of a local optimum problem as described in Section 4.3, we evaluated spectral conversion accuracy using the four EV-GMMs: 1) PCA-based EV-GMM "Conventional"; 2) the canonical EV-GMM "Proposed (TI-GMM)" trained using the TI-GMM as an initial model; 3) the canonical EV-GMM "Proposed (TD-GMM)" trained with the occupancy approximation as mentioned in Section 4.4.1; and 4) the canonical EV-GMM "Proposed (DAEM)" trained with DAEM algorithm. Figure 4.4 shows results of mel-cepstral distortion as a function of the contribution rate of the representative vectors as shown in Table 4.2. Although the "Proposed (TI-GMM)" method yields performance improvement when the contribution rate is 100%, it causes performance degradation when the contribution rate is less than 60%. If using a largely enough number of the representative vectors so that acoustic characteristics of individual pre-stored target speakers are modeled well, the EV-GMM training doesn't cause a severe local optimum problem. On the other hand, if the number of representative vectors is too small to span a subspace modeling those acoustic characteristics precisely, the EV-GMM training significantly suffers from a local optimum problem. This problem is effectively addressed by introducing the occupancy approximation "Proposed (TD-GMM)" or the DAEM algorithm "Proposed (DAEM)" to the EV-GMM training. It is interesting that the occupancy approximation outperforms the DAEM algorithm in particular when the contribution rate is 20%. It would be expected that the mixture-component oc-



(a) Number of adapting utterances is 1.

(b) Number of adapting utterances is 2.

(c) Number of adapting utterances is 4.

(d) Number of adapting utterances is 8.

(e) Number of adapting utterances is 16.

(f) Number of adapting utterances is 32.

Figure 4.4. Mel-cepstral distortion as a function of the number of contribution rate in each proposed adaptive training method and the conventional training method.

cupancies calculated by the TD-GMM would give better ML estimates of model parameters since they are quite similar to those calculated by reasonable model parameters such as the canonical EV-GMM with 159 representative vectors as shown in Figure 4.1. Although the occupancy approximation is a heuristic approach not supported mathematically, it is more computationally efficient than the DAEM algorithm. Therefore, it is a very effective approach for alleviating the local optimum problem.

4.6.3 Subjective evaluations

To demonstrate the effectiveness of the proposed method, we conducted a preference test on speech quality and an XAB test on conversion accuracy for speaker individuality. In these tests, the proposed canonical EV-GMM evaluated in Figure 4.2 was compared with the conventional PCA-based EV-GMM. In the preference test, a pair of two different types of the converted speech was presented to listeners, and then they were asked which voice sounded better. In the XAB test, a pair of two different types of the converted speech was presented to them after presenting the target speech as a reference. Then, they were asked which voice sounded more similar to the reference target. The number of listeners was five and each speaker evaluated 60 sample-pairs. The number of adaptation utterances was set to two in each evaluation.

Figure 4.5 shows the results of each subjective evaluation. We can see that 1) the proposed method yields significant improvement in speech quality and 2) conversion accuracy for speaker individuality by the proposed method is almost equal to that by the conventional method. These results suggest that the proposed adaptive training is very effective for improving the performance of one-to-many EVC.

We also conducted another preference test on speech quality to evaluate the effectiveness of the proposed methods for alleviating the local optimum problem. The four EV-GMMs evaluated in Figure 4.4 were compared with each other. In this test, the number of representative vectors was set to one and the number of adaptation sentences was set to two. The number of listeners was 10 and each speaker evaluated 48 sample-pairs.

Figure 4.6 shows the result of the preference test on speech quality. “Pro-

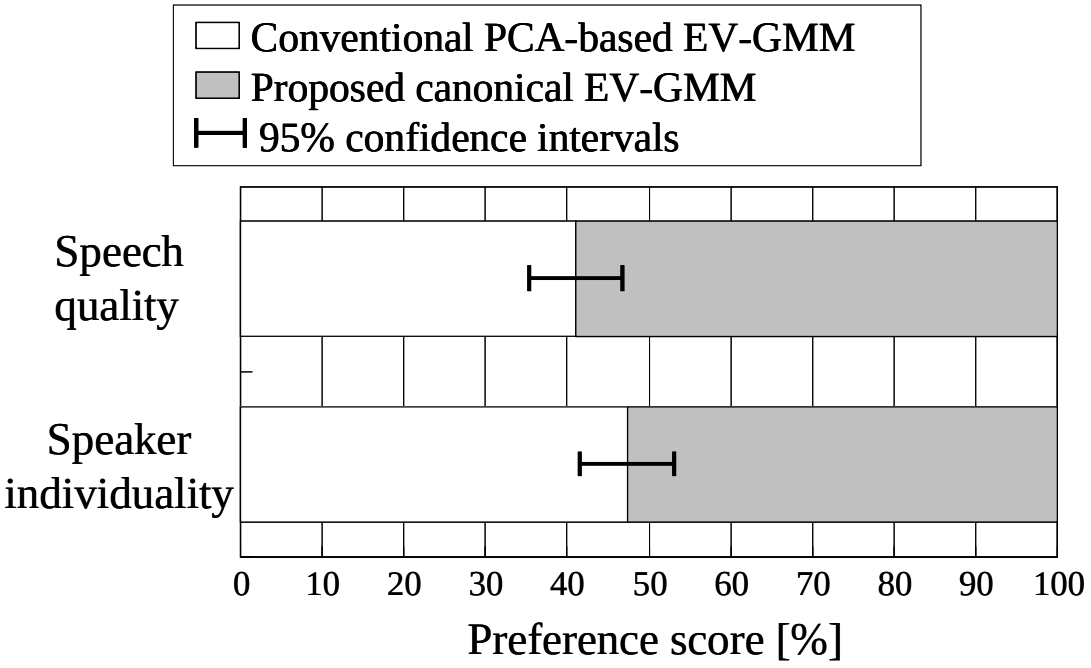


Figure 4.5. Results of subjective evaluation in adaptive training for EV-GMM.

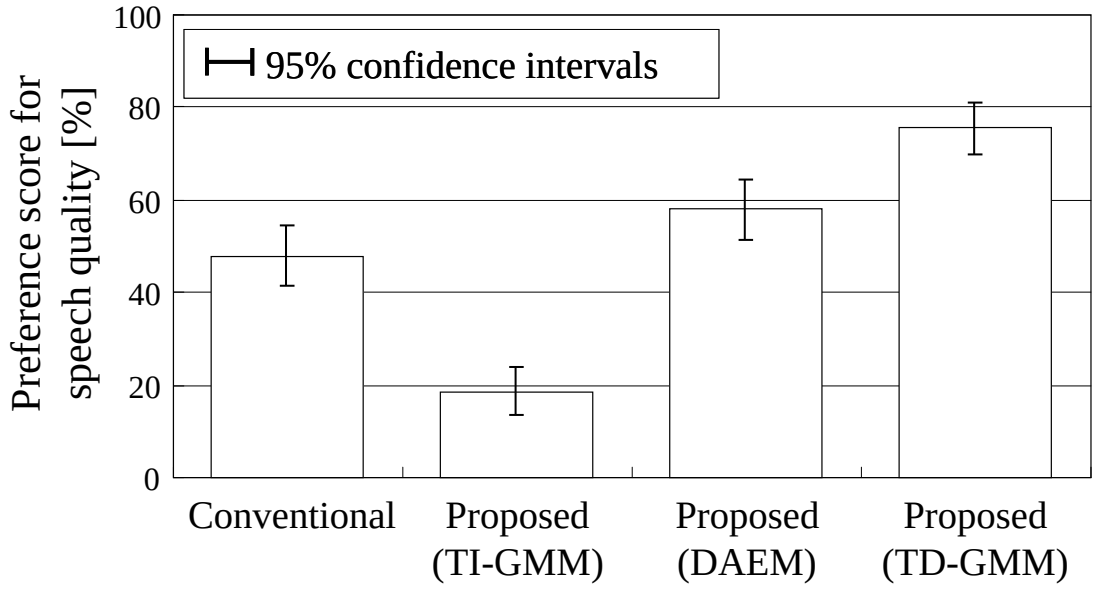


Figure 4.6. Result of preference test on speech quality when the contribution rate is set to 20%.

posed (TI-GMM)" causes performance degradation because of a local optimum problem. On the other hand, the local optimum problem is effectively alleviated by "Proposed (TD-GMM)" or "Proposed (DAEM)". These results are consistent with those observed in Figure 4.4.

4.7. Summary

In order to improve the performance of one-to-many eigenvoice conversion (EVC), we have proposed an adaptive training method for the eigenvoice Gaussian mixture model (EV-GMM). The conventional EV-GMM, parameters were affected by inter-speaker acoustic variations because they were determined based on a target-speaker-independent GMM (TI-GMM). These parameters often caused the degradation of the conversion performance. To address this problem, we have applied the proposed adaptive training to the EV-GMM. In the proposed method, we can construct the canonical EV-GMM which includes the parameters of the pseudo-normalized speaker.

Moreover, we have also proposed two approaches to alleviate the local optimum problem observed in the EV-GMM training using a small number of eigenvoices. One is the first E-step approximation method in which we calculate the occupancies of the first E-step with target-speaker-dependent GMMs. And the other is adaptive training using Deterministic annealing EM algorithm which reformulates a maximization process of a likelihood function as a minimization process of free energy.

We have evaluated the effectiveness of the proposed methods objectively and subjectively. The experimental results have demonstrated that the proposed training method is very effective for improving the performance of the one-to-many EVC.

Chapter 5

Improvements of One-to-Many Eigenvoice Conversion System

In this chapter, we describe the proposed one-to-many EVC system. In Chapter 2, It is mentioned that the conventional one-to-many EVC system includes three factors causing the degradation of the converted speech quality, i.e., STRAIGHT simple excitation (STSE), the conversion algorithm not considering the global variance (GV), and inter-speaker variations of the EV-GMM. In order to solve these problems of the conventional one-to-many EVC framework, three promising techniques are introduced, i.e., STRAIGHT mixed excitation (STME), conversion algorithm considering the GV and adaptive training method for EVC frameworks which have been respectively described in Chapter 2, 3 and 4. Experimental results demonstrate that the proposed system causes significant improvements in the performance of EVC.

5.1. Introduction

Although EVC frameworks [11][12] achieve more flexible conversion training, these converted speech qualities are not still high enough because these techniques often make the converted speech buzzing and muffled. In the one-to-many EVC, the factor of the insufficient converted speech quality is caused by employing the following techniques:

- STRAIGHT simple excitation (STSE) based on switching a phase-manipulated pulse train and white noise [13], which is too simple to model the excitation signal appropriately;

- The EV-GMM based on the target-speaker-independent GMM (TI-GMM), which usually causes the conversion model improperly capturing acoustic variations among many pre-stored target speakers;

- The spectral conversion algorithm not considering the GV, which often causes over-smoothed spectral parameters.

Therefore, there remains room to improve the conventional one-to-many EVC system.

This thesis has already described the following promising techniques:

- In Chapter 3, the converted speech quality of the traditional VC was improved by employing STRAIGHT mixed excitation (STME), which represents the actual excitation signal more properly than the STME;

- The conversion performance of one-to-many EVC was improved by applying the adaptive training for the EV-GMM described in Chapter 4, which can reduce the inter-speaker variations included in the EV-GMM;

- In the research described in Chapter 2, the conversion algorithm considering the GV was introduced, which improves the over-smoothing of converted spectral features.

Using these promising techniques, we improve the one-to-many EVC system. Moreover, we clear up which elements contribute the improvement of our proposed one-to-may EVC in the experimental evaluations.

This chapter is organized as follows. In Section 5.2, we describe the many-to-many EVC conversion algorithms. In Section 5.3, the proposed EV-GMM training with non-parallel data sets is described. Finally, we summarize this chapter in Section 5.4

5.2. Improved one-to-many EVC system

In order to improve the converted speech quality of EVC, we apply STME, the conversion algorithm considering the GV and the adaptive training method to the conventional one-to-many EVC system.

5.2.1 STRAIGHT mixed excitation for one-to-many EVC

We have proposed the conversion of aperiodic components based on a GMM in order to apply STME to VC and have demonstrated its effectiveness in the conventional VC framework in Chapter 3. In this chapter, aperiodic components are modeled by EV-GMM for applying STME to the one-to-many EVC system. The EV-GMM for aperiodic components is defined as the same described in Section 2.3.1. Then, the proposed adaptive training addressed in Chapter 4 is applied to this EV-GMM.

5.2.2 MLE-based conversion considering GV

Figure 5.1 shows a time sequence of the 7^{th} mel-cepstral coefficient extracted from the target speech and that of the converted coefficient by the conventional EVC system, respectively. We can observe that the GV of the converted mel-cepstral sequence is smaller than that of the target one. This is because the over-smoothing is caused through a statistical modeling process. It has been reported that both the converted speech quality and conversion accuracy for speaker individuality are dramatically improved by considering the GV of the converted parameters in the conversion process [10].

Eigenvoice single Gaussian distribution for GV

In the MLE-based conversion considering the GV, we employ the single Gaussian distribution, which is modeled with the probability density of the target GV. In the one-to-many EVC system, we cannot build the single Gaussian distribution of the desired target GV if we obtain only one utterance of that speaker because the GV is calculated utterance by utterance. Therefore, in the proposed one-to-many EVC system, the probability density of the GV is modeled by an eigenvoice

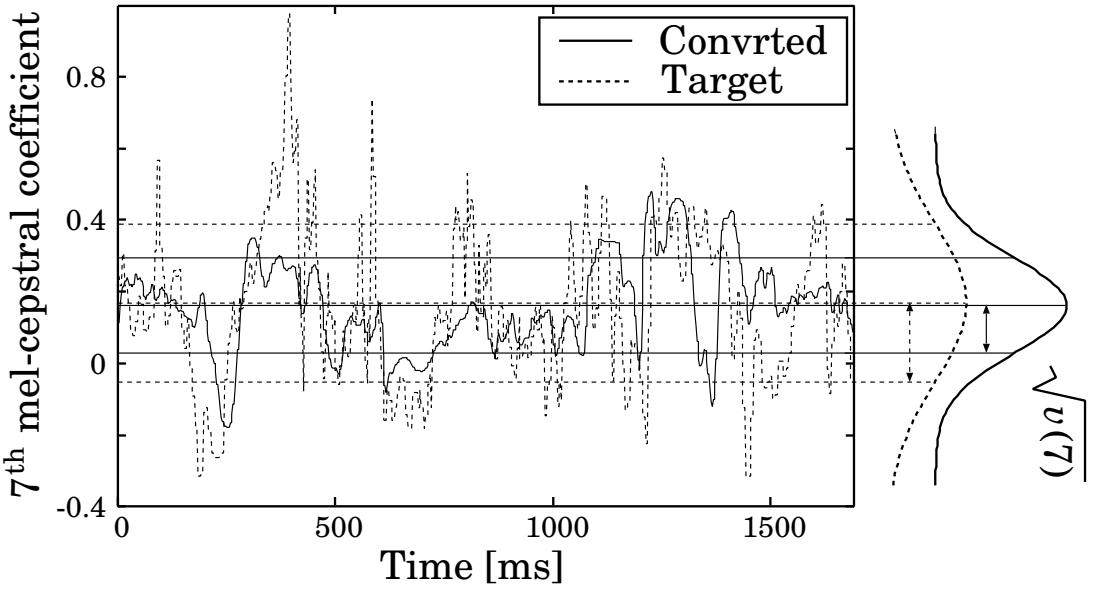


Figure 5.1. Target mel-cepstrum sequence and converted sequence in the conventional EVC system. Bidirectional arrows show square root of GV extracted from each sequence. Note that duration of converted sequence is different from that of target one.

single Gaussian distribution (EV-SG) as follows:

$$P\big(\boldsymbol{v}_y^{(s)}|\lambda_v^{(EV)}, \boldsymbol{w}_v^{(s)}\big) = \mathcal{N}\left(\boldsymbol{v}_y^{(s)}; \boldsymbol{\mu}(\boldsymbol{w}_v^{(s)}), \boldsymbol{\Sigma}^{(vv)}\right), \tag{5.1}$$

$$\boldsymbol{\mu}(\boldsymbol{w}_v^{(s)}) = \boldsymbol{B}_v \boldsymbol{w}_v^{(s)} + \boldsymbol{b}_v^{(0)}, \tag{5.2}$$

where $\boldsymbol{v}_y^{(s)}$ denotes the GV vector of the s^{th} target speaker. A tied-parameter set $\lambda_v^{(EV)}$ includes representative vectors $\boldsymbol{B}_v$, a bias vector $\boldsymbol{b}_v^{(0)}$ and a covariance matrix $\boldsymbol{\Sigma}^{(vv)}$. The weight vector of the s^{th} target speaker is $\boldsymbol{w}_v^{(s)}$.

Training and adaptation of EV-SG

The PCA-based EV-SG is trained using all of GV vectors extracted from individual utterances of every pre-stored target speaker in the similar manner as written in Section 2.3.2: 1) we train a speaker-independent single Gaussian distribution

(SI-SG) with all of pre-stored target speakers' GV vectors; 2) we calculate GV mean vectors of each pre-stored target speaker using each target speaker's GV vectors; and 3) bias vector $\boldsymbol{b}_v^{(0)}$ and representative vectors $\boldsymbol{B}_v$ of GV are extracted from GV's supervector, which is constructed by concatenating pre-stored target GV mean vectors, by performing PCA.

In the same case as the EV-GMM, we apply this adaptive training to the PCA-based EV-SG. Therefore, in the proposed system, we also apply the adaptive training to the EV-SG for the GV. The canonical EV-SG parameters are estimated by maximizing a total likelihood of the adapted EV-SGs for individual pre-stored target speakers' GVs as follows:

$$\left\{\hat{\lambda}_v^{(EV)}, \hat{\boldsymbol{w}}_{v1}^{S}\right\} = \underset{\lambda_v^{(EV)}, \boldsymbol{w}_{v1}^{S}}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{n=1}^{N_s} P\left(\boldsymbol{v}_{y,n}^{(s)} | \lambda^{(EV)}, \boldsymbol{w}_v^{(s)}\right), \tag{5.3}$$

where $\boldsymbol{v}_{y,n}^{(s)}$ denotes the GV vector extracted from the n^{th} utterance of the s^{th} pre-stored target speaker and $\boldsymbol{w}_{v1}^{S}$ is a set of weight vectors for EV-SG. We do not have to perform EM algorithm because there is no hidden variable in the EV-SG. However, it is still difficult to update all parameters simultaneously for the same reason as in the EV-GMM. Therefore, individual parameters of the EV-SG are updated iteratively in the same manner as mentioned in Chapter 4.

MLE-based conversion with adapted EV-SG

The converted speech features are determined by maximizing with respect to $\boldsymbol{y}$ as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \left\{ \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(EV)}\right) P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}\right) \right\}^{\omega} P\left(\boldsymbol{v}_y | \lambda_v^{(EV)}, \hat{\boldsymbol{w}}_v\right), \tag{5.4}$$

This conversion algorithm is solved using the same manner described in Section 2.2.2. Note that we again approximate the objective function with the suboptimum mixture component sequence in Section 2.2.3.

Figure 5.2 shows an example of the converted trajectories with/without the GV. By considering the GV, many trajectory movements are dramatically emphasized. However, other trajectory movements are almost same. This is because

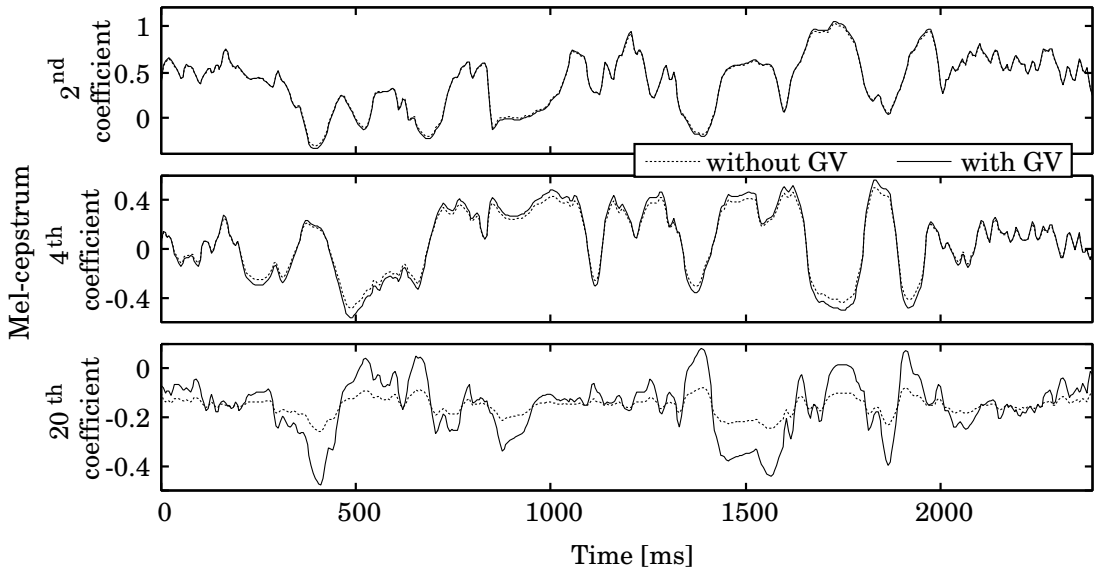


Figure 5.2. Examples of converted spectral trajectories with/without GV in the one-to-many EVC.

the degree of emphasis is determined by the likelihood written in Eq. (5.4). Thus, in the one-to-many EVC system, GV is effectively improved over-smoothing problem.

5.2.3 Overview of the proposed one-to-many EVC system

Figure 5.3 shows an overview of the proposed one-to-many EVC system.

In the training process, spectral features, aperiodic components, and GV vectors are extracted from speech samples of multiple parallel data sets, and joint feature vectors are constructed for spectral features and for aperiodic components. We build the EV-GMM for spectral features, the EV-GMM for aperiodic components, and the EV-SG for the GV using the PCA-based training process described in Chapter 4 and Section 5.2.2, respectively. Then, these models are independently optimized with the adaptive training based on the approximated method. We also calculate mean and variance values of log-scaled F_0 of the source speaker.

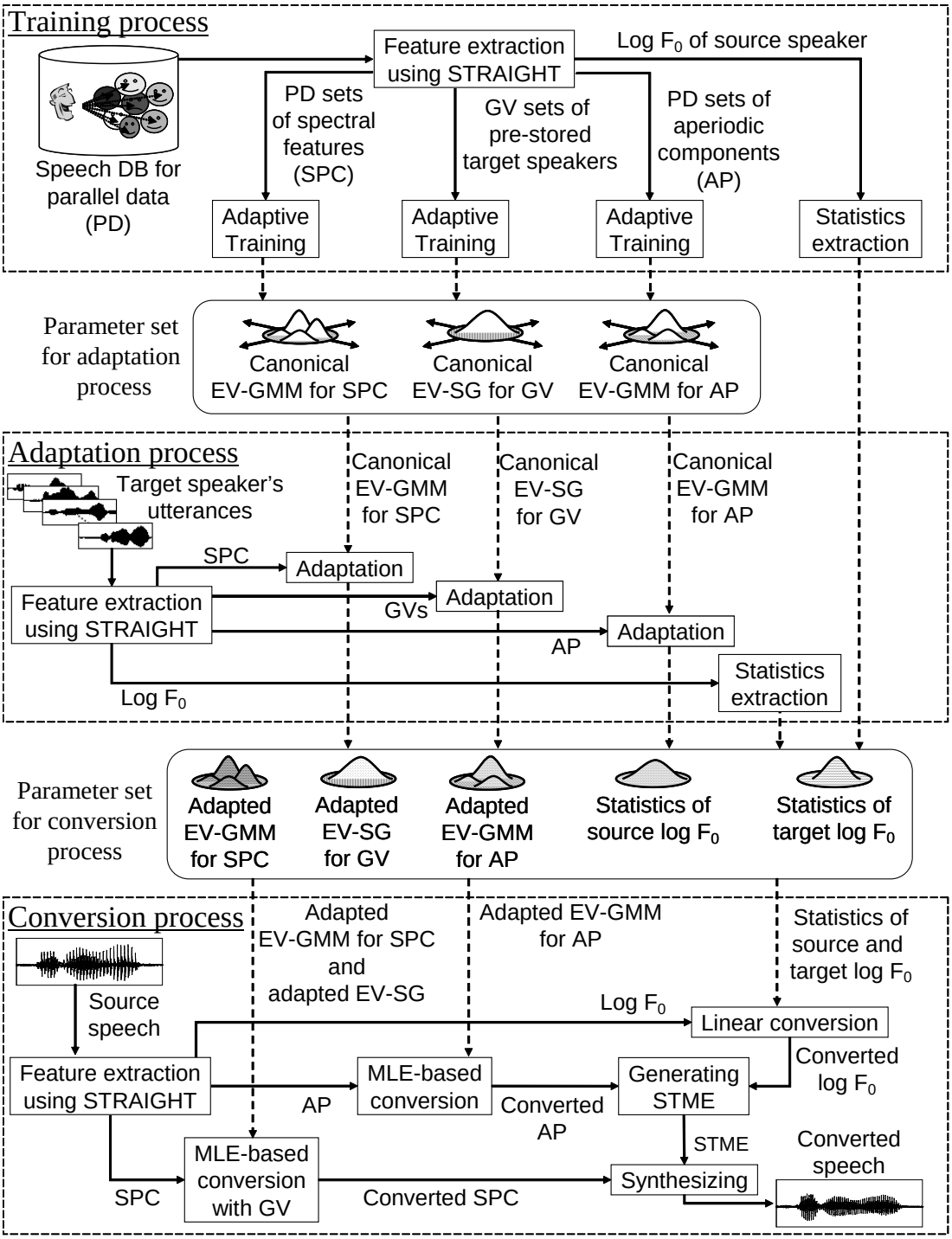


Figure 5.3. Overview of proposed one-to-many EVC system.

In the adaptation process, spectral features, aperiodic components, GV vectors, and F_0 values are extracted from adaptation data of a new target speaker. Then, the weight vectors for the EV-GMMs for the spectral features and for the aperiodic components are independently estimated as shown in Eq. (2.52). The weight vector for the EV-SG is also estimated by maximizing the likelihood of the EV-SG in Eq. (5.1) for given the GV vectors. Mean and variance values of log-scaled F_0 of the target speaker are also calculated.

In the conversion process, spectral features are converted by MLE-based conversion considering the GV. On the other hand, aperiodic components are converted by the conventional MLE-based conversion without the GV because quality improvements yielded by considering the GV in the aperiodic conversion are not significant. The converted F_0 values are determined in Eq. (2.65). Finally, the excitation signal is generated based on STME with the converted F_0 values and the converted aperiodic components, and then synthetic speech is generated by filtering the excitation signal with the converted spectral features.

5.3. Experimental evaluations

5.3.1 Experimental conditions

We objectively and subjectively compared the performance of the proposed one-to-many EVC system with that of the conventional one. In this evaluation, we employ same training data set and evaluation data set in Section 4.6. We used 24-dimensional mel-cepstrum as a spectral feature, which were extracted from smoothed spectrum analyzed by STRAIGHT [13], and aperiodic components that were averaged on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 kHz) the same as in Chapter 3. The number of representative vectors was 159 for mel-cepstrum, 64 for aperiodic components and 4 for GV, respectively. The number of mixture components was 128 for spectral features and 64 for the aperiodic features, respectively. These parameters were optimized so that the best conversion accuracy for each feature was obtained in the evaluation data.

5.3.2 Objective evaluations

We evaluated the effectiveness of the proposed adaptive training method in the adaptation of the EV-GMM. Note that the effectiveness of the proposed adaptive training method in the spectral conversion has already been described in Chapter 4. As evaluation measures, we used RMSE on aperiodic components and the likelihood of the adapted EV-SG for GVs in the evaluation data. Before the conversion, RMSE on aperiodic components between the source and target speakers was 2.70 [dB] and the log-scaled likelihood of EV-SG was 80.64. When we used STSE for synthesizing the converted speech, RMSE on aperiodic components between the converted speech and the target speech was 3.05 [dB].

Figure 5.4 and 5.5 show RMSE on aperiodic components and the log-scaled likelihood of the EV-SG for GV, respectively. We can see that the adaptation performance of both the EV-GMM for aperiodic components and the EV-SG for the GV is significantly improved by applying the proposed adaptive training. Moreover, these improvements are always observed even if varying the amount of adaptation data. These results demonstrate the effectiveness of applying the proposed adaptive training to modeling of the aperiodic components and the GV.

5.3.3 Subjective evaluations

We conducted a preference test and an opinion test on speech quality and an XAB test on conversion accuracy for speaker individuality. It has been reported that a combination of STME and GV yields significant improvements in naturalness of synthetic speech in HMM-based speech synthesis [49]. In this chapter, in order to further demonstrate the effectiveness of a combination of STME, GV and adaptive training, we evaluated several types of converted speech shown in Table 5.1. In the preference test, a pair of two different types of the converted speech was presented to listeners, and then they were asked which voice sounded better. In the opinion test, each listener evaluated speech quality of the converted voices using a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In the XAB test, a pair of two different types of the converted speech was presented to them after presenting the target speech as a reference. Then, they were asked which voice sounded more similar to the reference. Each listener evaluated every

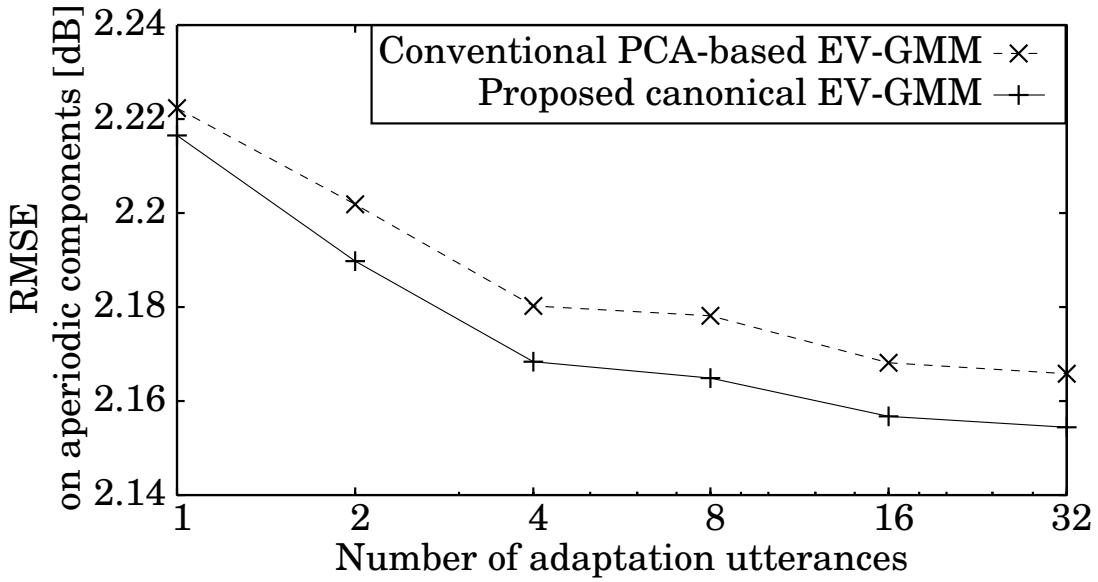


Figure 5.4. Result of objective evaluation by RMSE on aperiodic components for one-to-many EVC system.

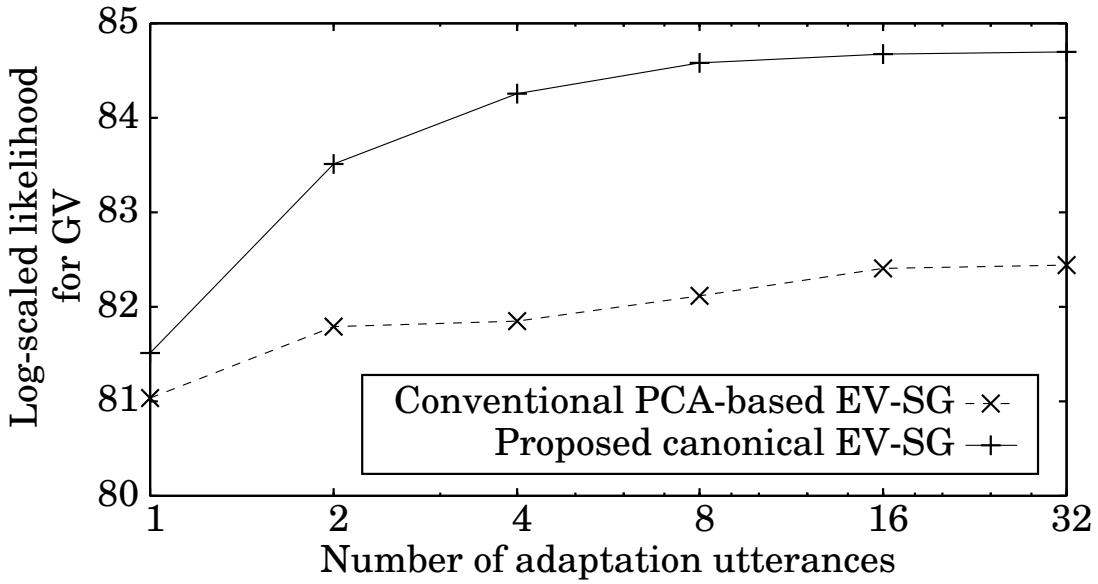


Figure 5.5. Result of objective evaluation by log-scaled likelihood of the EV-SG for GV.

Table 5.1. Combinations of improving methods for generating converted speech. "Y" means using method and "N" means not using method

Method \ Type	i	ii	iii	iv	v	vi	vii	viii
Adaptive training	N	Y	N	N	Y	Y	N	Y
STME	N	N	Y	N	Y	N	Y	Y
Conversion with GV	N	N	N	Y	N	Y	Y	Y

pair-combination of all types of the converted speech. The number of listeners was 15 and each listener evaluated 56 sample pairs in the preference test and the XAB test. In the opinion test, the number of listeners was 10 and each listener evaluated 180 samples.

Figure 5.6 shows the result of the preference test on speech quality. Note that type i is equivalent to the conventional one-to-many EVC system. We can see that speech quality of the converted speech is significantly improved by applying each of the adaptive training (type ii), STME (type iii), and the conversion with the GV (type iv) to the conventional system. Especially, the conversion with the GV achieves the largest quality improvement. Furthermore, further quality improvements are obtained by combining these methods. Consequently, the proposed one-to-many EVC system (type viii) is capable of synthesizing the converted speech with much higher speech quality compared with the conventional system.

Figure 5.7 shows the result of the opinion test on speech quality. We can see the same tendency as observed in the preference test shown in Figure 5.6. Significant improvements in the converted speech quality are yielded by applying each of the adaptive training, STME, and the conversion with the GV to the conventional system, and the best quality is yielded by the type viii system.

Figure 5.8 shows the result of the XAB test of conversion accuracy for speaker individuality. The adaptive training (type ii) does not contribute to the improvement of conversion accuracy for speaker individuality. This result is consistent to the result reported in Chapter 4. On the other hand, STME (type iii) or the con-

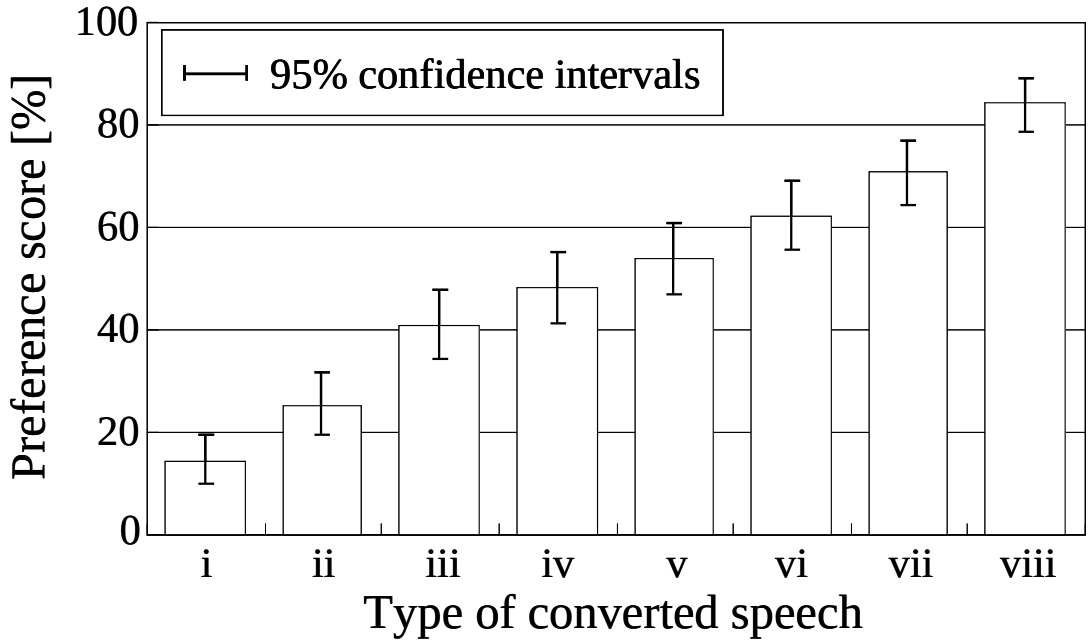


Figure 5.6. Subjective result of preference test on speech quality for one-to-many EVC system.

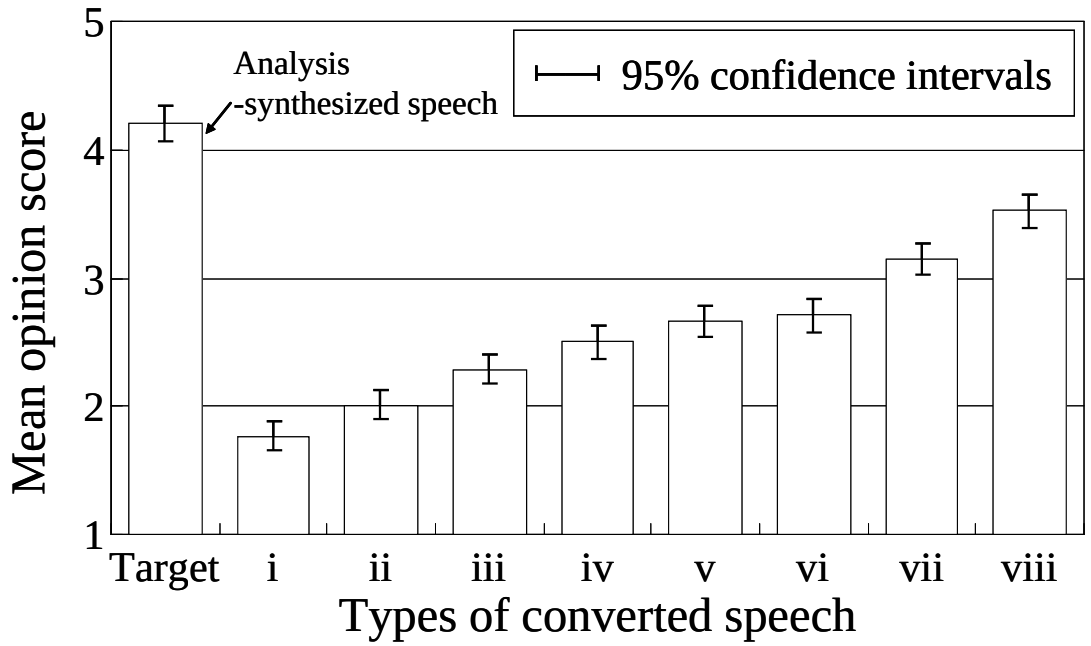


Figure 5.7. Subjective result of opinion test on speech quality for one-to-many EVC system.

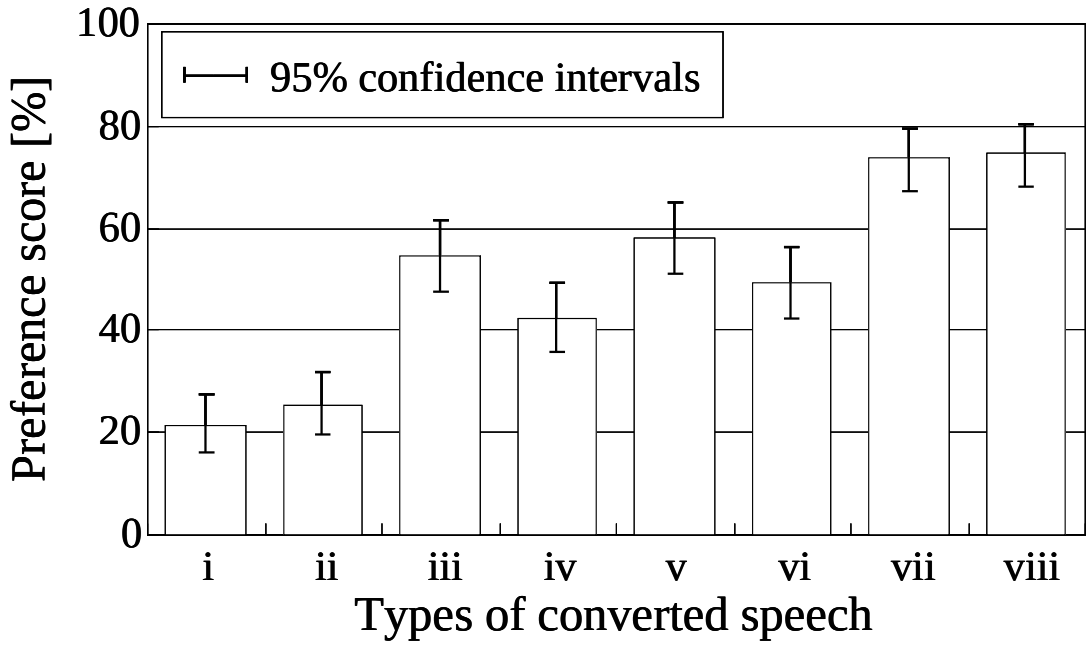


Figure 5.8. Subjective result of conversion accuracy for speaker individuality fro one-to-many EVC system.

version with the GV (type iv) yields significant improvements of the conversion accuracy, and STME is the most effective. As observed in the result of speech quality, further improvements are yielded by combining these effective methods.

These results suggest that the proposed system yields dramatic improvements in the performance of the one-to-many EVC system.

5.4. Summary

In order to improve converted speech quality and conversion accuracy for speaker individuality of the one-to-many eigenvoice conversion (EVC) system, we have applied three promising techniques, i.e., STRAIGHT mixed excitation, the conversion algorithm considering global variance (GV) and the adaptive training method of the eigenvoice Gaussian mixture model (EV-GMM) to the conventional

system. In the proposed one-to-many EVC system, we train two EV-GMMs for spectral features and for aperiodic components and an eigenvoice single Gaussian distribution for the GV separately. These models are effectively adapted to a new target speaker using a very small amount of adaptation data in a completely text-independent manner.

The results of objective evaluation have demonstrated that the proposed adaptive training has improved adaptation performance of both the EV-GMM for aperiodic components and the EV-SG for the GV. In the results of subjective evaluations, we have been able to see that the conversion algorithm considering the GV is the most effective method for improving speech quality and the STME contributes to the improvement of conversion accuracy for speaker individuality most. Moreover, the proposed one-to-many EVC system considerably outperforms the conventional one in view of both converted speech quality and conversion accuracy for speaker individuality.

Chapter 6

Many-to-Many Eigenvoice Conversion

In this chapter, we describe a novel EVC paradigm, i.e., many-to-many EVC. This framework achieves the conversion from an arbitrary source speaker's voice to an arbitrary target speaker's voice. In this framework, we perform many-to-one EVC and one-to-many EVC sequentially through the specified speaker, what we call "reference speaker", using the single EV-GMM. Moreover, inspired by this conversion framework, we propose the refining EV-GMM method using non-parallel utterance data set. In each proposed method, we conduct objective and subjective experimental evaluations. These results demonstrate the effectiveness of the proposed many-to-many conversion algorithms and refining EV-GMM method.

6.1. Introduction

In Chapter 5, the converted speech quality for the EVC framework is significantly improved. However, it is still hard to flexibly perform the conversion between arbitrary speaker-pairs because the conventional EVC framework has achieved one-to-many and many-to-one VC frameworks.

In previous work, VC using ML constrained adaptation [31] achieves many-to-many VC. In this method, unsupervised adaptation is applied to both side of GMM based on joint probability density between specified source and target

speaker. Masuda et al., have proposed multistep VC [55] for reducing costs of training conversion models modeled by GMMs. If we want to achieve the conversion from N source speakers to M target speakers in the traditional VC framework, we must train $N \times M$ conversion models. On the other hand in the multistep VC framework, we train only $N + M$ conversion models between individual speakers and a specified pre-defined speaker, which is called "reference speaker." In the conversion process, we convert a source speaker's voice into a target speaker's voice through the reference voice. Therefore, from another point of view, this framework is a successful framework to achieve many-to-many VC.

Inspired by multistep VC framework, we propose many-to-many VC as a much more flexible VC framework by extending the conventional EVC frameworks. In the proposed framework, an EV-GMM between the reference speaker and pre-stored speakers is trained in advance in the same manner as the conventional EVC framework. The GMMs between the reference speaker and an arbitrary source/target speaker is flexibly developed by estimating a small amount of free parameters of the trained EV-GMM, i.e., weights for eigenvectors, using only a few utterances of the adapted speaker in an unsupervised manner. In the conversion process, the proposed framework is to sequentially perform many-to-one EVC and one-to-many EVC through the reference speaker. In this framework, two many-to-many EVC algorithms are investigated; one is the sequential conversion based on multistep VC [55], and the other is the sequential conversion sharing mixture components between many-to-one EVC and one-to-many EVC by considering the reference speaker's voice as hidden variable.

Moreover, inspired by this proposed conversion algorithm, we propose a method of refining the EV-GMM by additionally using any arbitrary utterance sets of a larger number of pre-stored speakers, i.e., non-parallel data sets from various speakers, in order to relax the use of parallel data sets in EV-GMM training. In the proposed training method, the initial EV-GMM is trained using the existing multiple parallel data sets. Then it is refined using only non-parallel data sets including a larger number of speakers while considering the reference voices corresponding to those data sets as hidden variables. Note that these non-parallel data sets are generally much more easily available than the multiple parallel data sets. Therefore, the proposed method allows us to extract more informative prior

knowledge from a much larger number of speakers in EV-GMM training.

This chapter is organized as follows. In Section 6.2, we describe the many-to-many EVC conversion algorithms. In Section 6.3, the proposed EV-GMM training with non-parallel data sets is described. Section 6.4 describes experimental evaluations for many-to-many EVC algorithms and the proposed training method. Finally, we summarize this chapter in Section 6.5.

6.2. Many-to-many conversion algorithm based on eigenvoices

In order to achieve many-to-many EVC, we employ many-to-one EVC and one-to-many EVC. Figure 6.1 shows a schematic image of the proposed many-to-many EVC process. We use only one EV-GMM because the EV-GMM models joint probability density, the one-to-many EV-GMM can also be used as the many-to-one EV-GMM by just switching the source and the target features. Note that this chapter employs a one-to-many EV-GMM in this chapter. There, the source speaker $\boldsymbol{X}$ included in the one-to-many EV-GMM is regarded as the reference speaker and $\boldsymbol{Y}^{(i)}$ and $\boldsymbol{Y}^{(o)}$ represent source and target speakers in this chapter, respectively.

Given a small amount of adaptation data of arbitrary source and target speakers, ML estimates of the weight vectors, $\hat{\boldsymbol{w}}^{(i)}$ and $\hat{\boldsymbol{w}}^{(o)}$, for the source $\boldsymbol{Y}^{(i)}$ and the target $\boldsymbol{Y}^{(o)}$ are determined by Eq. (2.52), respectively. Then, the arbitrary source speaker's voice is converted into the reference voice with many-to-one EVC. After that, the converted reference speaker's voice is further converted into the arbitrary target speaker's voice with one-to-many EVC.

6.2.1 Conversion algorithm based on multistep VC

Many-to-many EVC based on multistep VC [55] simply performs two conversion processes. In the first step, we convert the source voice into the reference voice using the EV-GMM adapted to the source speaker. The ML estimate of a static and dynamic feature sequence of the reference voice $\hat{\boldsymbol{X}}$ is determined by

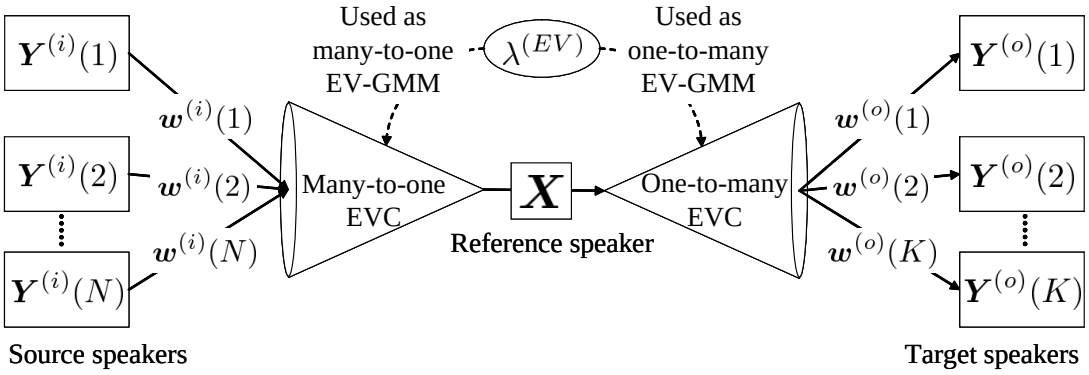


Figure 6.1. Overview of many-to-many EVC

maximizing the following likelihood function:

$$\hat{\boldsymbol{X}} = \arg\max_{\boldsymbol{X}} \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{Y}^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right) P\left(\boldsymbol{X}|\boldsymbol{Y}^{(i)}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right). \quad (6.1)$$

Since this conversion is not considering dynamic features, a converted joint feature vector at frame t $\hat{\boldsymbol{X}}_t$ is written as

$$\hat{\boldsymbol{X}}_t = \sum_{m=1}^{M} P\left(\boldsymbol{m}|\boldsymbol{Y}_t^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right) \boldsymbol{E}_{m,t}^{(X)}, \quad (6.2)$$

where

$$\boldsymbol{E}_{m,t}^{(X)} = \boldsymbol{\mu}_m^{(X)} + \boldsymbol{\Sigma}_m^{(XY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \left(\boldsymbol{Y}_t^{(i)} - \boldsymbol{B}_m \hat{\boldsymbol{w}}^{(i)} - \boldsymbol{b}_m^{(0)}\right). \quad (6.3)$$

In the manner described in section 2.2.2, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(i)}$ is determined as follows:

$$\hat{\boldsymbol{m}}^{(i)} = \underset{\boldsymbol{m}}{\text{argmax}}\, P\left(\boldsymbol{m}|\boldsymbol{Y}^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right). \quad (6.4)$$

Then, the converted reference voice is determined as follows:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X}}{\text{argmax}}\, P\left(\boldsymbol{X}|\boldsymbol{Y}^{(i)}, \hat{\boldsymbol{m}}^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right). \quad (6.5)$$

In this case, we can obtain $\boldsymbol{E}_{\hat{m}^{(i)},t}^{(X)}$ as the converted reference feature $\hat{\boldsymbol{X}}_t$.

In the second step, we convert the converted reference voice into the target voice using EV-GMM adapted to the target speaker. We estimate a target static feature sequence $\hat{\boldsymbol{y}}^{(o)}$ by maximizing the following likelihood function:

$$\hat{\boldsymbol{y}}^{(o)} = \arg\max_{\boldsymbol{y}} \sum_{\text{all } \boldsymbol{m}} P\Big(\boldsymbol{m}|\hat{\boldsymbol{X}}, \lambda^{(EV)}\Big) P\Big(\boldsymbol{Y}^{(o)}|\hat{\boldsymbol{X}}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(o)}\Big). \tag{6.6}$$

In the same as the first step, using approximation conversion algorithm, the converted static feature sequence $\hat{\boldsymbol{y}}^{(o)}$ is determined as follows:

$$\hat{\boldsymbol{y}}^{(o)} = \arg\max_{\boldsymbol{y}^{(o)}} P\left(\boldsymbol{Y}^{(o)}|\hat{\boldsymbol{X}}, \hat{\boldsymbol{m}}^{(o)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(o)}\right), \tag{6.7}$$

$$\hat{\boldsymbol{m}}^{(o)} = \arg\max_{\boldsymbol{m}} P\left(\boldsymbol{m}|\hat{\boldsymbol{X}}, \lambda^{(EV)}\right). \tag{6.8}$$

In addition, we can perform the conversion algorithm with the GV by applying Eq. (5.4) to the second step of the conversion process. Note that the mixture component sequence in many-to-one EVC shown by Eq. (6.4) is not always the same as that in one-to-many EVC shown by Eq. (6.8). It is possible that this inconsistency of the mixture component sequences causes the conversion between different phonemic spaces through the sequential conversion process.

6.2.2 Conversion algorithm with shared mixture components

To avoid the inconsistency of the mixture component sequences, we propose a sequential conversion method while sharing the same mixture component sequence in both many-to-one EVC and one-to-many EVC.

The converted static feature sequence $\hat{\boldsymbol{y}}^{(o)}$ is determined by maximizing the following likelihood function,

$$\hat{\boldsymbol{y}}^{(o)} = \arg\max_{\boldsymbol{y}} \sum_{\text{all } \boldsymbol{m}} P\Big(\boldsymbol{m}|\boldsymbol{Y}^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\Big) P\Big(\boldsymbol{Y}^{(o)}|\boldsymbol{Y}^{(i)}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}\Big), \tag{6.9}$$

where $P\Big(\boldsymbol{Y}^{(o)}|\boldsymbol{Y}^{(i)}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}\Big)$ represents conditional probability density of target features $\boldsymbol{Y}^{(o)}$ given source features $\boldsymbol{Y}^{(i)}$ modeled by a single Gaussian

distribution as follows:

$$
\begin{aligned}
&P\Big(\boldsymbol{Y}^{(o)}|\boldsymbol{Y}^{(i)}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}\Big) \\
&\quad = \int P\Big(\boldsymbol{Y}^{(o)}|\boldsymbol{X}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(o)}\Big)\, P\Big(\boldsymbol{X}|\boldsymbol{Y}^{(i)}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\Big)\, d\boldsymbol{X} \\
&\quad = \prod_{t=1}^{T} \mathcal{N}\Big(\boldsymbol{Y}_t^{(o)}; \tilde{\boldsymbol{E}}_{m,t}^{(Y)}, \tilde{\boldsymbol{D}}_m^{(Y)}\Big),
\end{aligned} \tag{6.10}
$$

$$
\tilde{\boldsymbol{E}}_{m,t}^{(Y)} = \boldsymbol{B}_m \hat{\boldsymbol{w}}^{(o)} + \boldsymbol{b}_m^{(0)} + \boldsymbol{A}_m \boldsymbol{\Sigma}_m^{(YY)^{-1}} \Big(\boldsymbol{Y}_t^{(i)} - \boldsymbol{B}_m \hat{\boldsymbol{w}}^{(i)} - \boldsymbol{b}_m^{(0)}\Big), \tag{6.11}
$$

$$
\tilde{\boldsymbol{D}}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{A}_m^{\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{A}_m, \tag{6.12}
$$

$$
\boldsymbol{A}_m = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_{\hat{m}}^{(XY)}. \tag{6.13}
$$

The feature vector sequence of the reference speaker $\boldsymbol{X}$ is regarded as a hidden variable. Therefore, this algorithm effectively converts the source speaker's voice into the target speaker's voice. Also, this algorithm can be applied as the approximation method described in section 2.2.2. The suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined by

$$
\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}}\, P\left(\boldsymbol{m}|\boldsymbol{Y}^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\right). \tag{6.14}
$$

Then, we determine the converted static feature sequence $\boldsymbol{y}^{(o)}$ as follows:

$$
\hat{\boldsymbol{y}}^{(o)} = \underset{\boldsymbol{y}^{(o)}}{\operatorname{argmax}}\, P\left(\boldsymbol{Y}^{(o)}|\boldsymbol{Y}^{(i)}, \hat{\boldsymbol{m}}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}\right). \tag{6.15}
$$

In addition, in order to perform this conversion algorithm considering GV, we determine the converted static feature sequence $\hat{\boldsymbol{y}}^{(o)}$ by maximizing the following likelihood function,

$$
\begin{aligned}
\hat{\boldsymbol{y}}^{(o)} = \arg\max_{\boldsymbol{y}} \Bigg\{ \sum_{\text{all } \boldsymbol{m}} & P\Big(\boldsymbol{m}|\boldsymbol{Y}^{(i)}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}\Big)\, P\Big(\boldsymbol{Y}^{(o)}|\boldsymbol{Y}^{(i)}, \boldsymbol{m}, \lambda^{(EV)}, \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}\Big) \Bigg\}^{\omega} \\
& \times P\big(\boldsymbol{v}_{y^{(o)}}|\lambda_v^{(EV)}, \boldsymbol{w}_v^{(o)}\big),
\end{aligned} \tag{6.16}
$$

where $\boldsymbol{v}_{y^{(o)}}$ and $\boldsymbol{w}_v^{(o)}$ are target GV vector and target weight vector for representative vectors of the EV-SG, respectively. In the same as the proposed conversion method not considering the GV, we can also approximate Eq. (6.16) with the

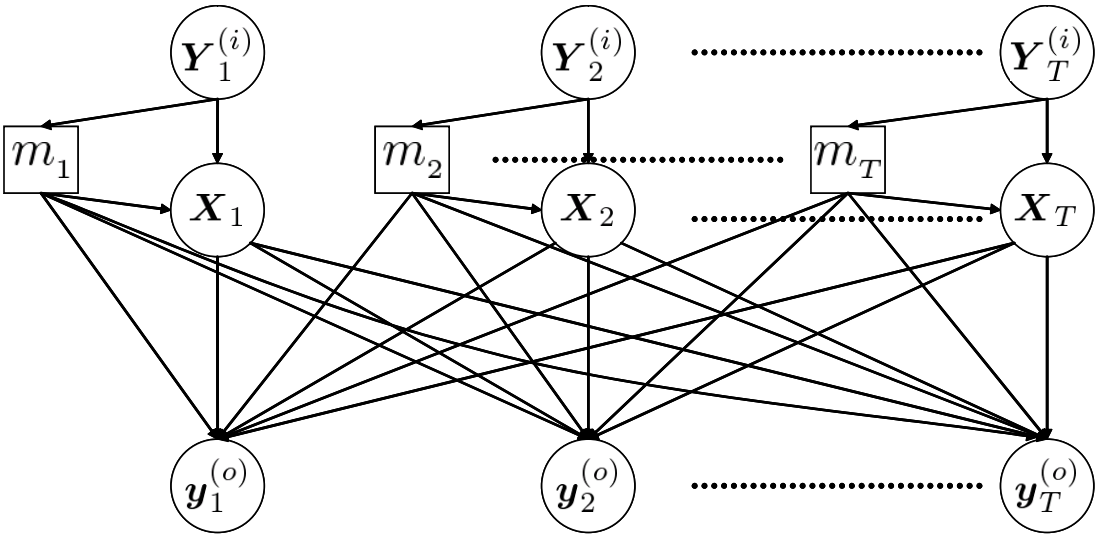


Figure 6.2. Graphical representation of relationship among individual variables in many-to-many EVC with reference voice

suboptimum mixture component sequence using the same technique described in Section 2.2.3. Figure 6.2 shows a graphical representation of the relationship among individual variables in the conversion process. This conversion algorithm effectively models correlations among the target feature sequence. Consequently, all of the source and reference feature vectors affect the determination of each converted feature vector.

6.3. Non-parallel training for EV-GMM of many-to-many EVC

Inspired by the conversion process in many-to-many EVC, we propose a new training method of the EV-GMM considering the reference voice as a hidden variable. Figure 6.3 shows an overview of the proposed training process. In the first step, we train the initial EV-GMM using the existing multiple parallel data sets between a single reference speaker and many pre-stored speakers in the same manner as described in the previous section. In the second step, we refine the EV-

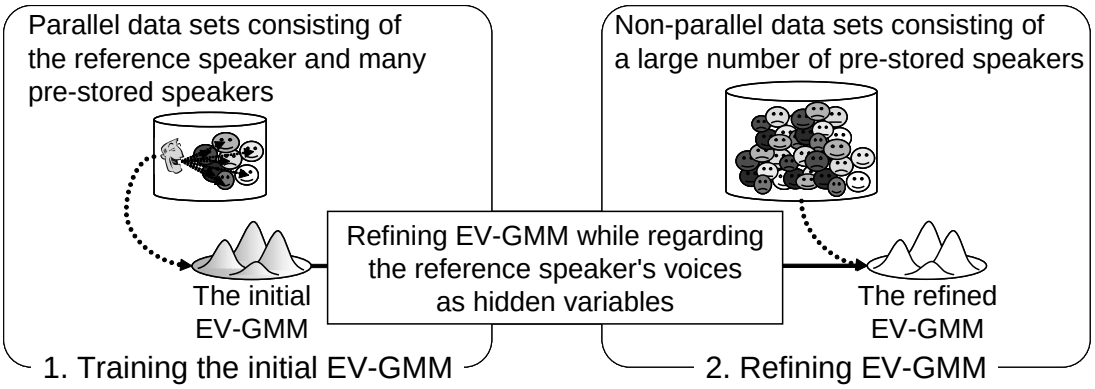


Figure 6.3. Overview of proposed EV-GMM training process

GMM using non-parallel data including a larger number of pre-stored speakers while regarding the reference features corresponding to those non-parallel data as hidden variables. Because this process is performed in a completely text-independent manner, any pre-stored speech data, i.e., any utterance set of any speaker, can be used for refining the EV-GMM. Therefore, the proposed training method allows us to use a larger amount of training data including more varieties of texts and speakers.

In the second training process, we update the EV-GMM parameters by maximizing the following marginal likelihood:

$$\begin{aligned}\left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S\right\} &= \arg\max_{\lambda^{(EV)}, \boldsymbol{w}_1^S} \prod_{s=1}^{S}\prod_{t=1}^{T_s} \int P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)\, d\boldsymbol{X}_t \\ &= \arg\max_{\lambda^{(EV)}, \boldsymbol{w}_1^S} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\Big(\boldsymbol{Y}_t^{(s)} | \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big). \end{aligned} \tag{6.17}$$

The training process is achieved with EM algorithm [43] by maximizing the following auxiliary function,

$$\begin{aligned} &Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S)\right\}\right) \\ &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M} P\Big(m | \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big) \log P\Big(\boldsymbol{Y}_t^{(s)}, m | \hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{(s)}\Big). \end{aligned} \tag{6.18}$$

In this proposed training, we can update the speaker-dependent weight vector $\boldsymbol{w}^{(s)}$ and the EV-GMM parameters related to only the pre-stored speaker's features, i.e., the mixture component weight α_m, the representative vectors $\boldsymbol{B}_m$, the bias vector $\boldsymbol{b}_m^{(0)}$, and the covariance matrix of the pre-stored speakers $\boldsymbol{\Sigma}_m^{(YY)}$ for each mixture component. The ML estimates of these parameters are given by

$$\hat{\boldsymbol{w}}^{(s)} = \left(\sum_{m=1}^{M} \overline{\gamma}_m^{(s)} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m \right)^{-1} \sum_{m=1}^{M} \left\{ \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \left(\overline{\boldsymbol{Y}}_m^{(s)} - \overline{\gamma}_m^{(s)} \boldsymbol{b}_m^{(0)} \right) \right\}, \tag{6.19}$$

$$\hat{\alpha}_m = \frac{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}{\sum_{m=1}^{M} \sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}, \tag{6.20}$$

$$\hat{\boldsymbol{\nu}}_m = \left(\sum_{s=1}^{S} \overline{\gamma}_m^{(s)} \hat{\boldsymbol{W}}_s^{(Y)^\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \hat{\boldsymbol{W}}_s^{(Y)} \right)^{-1} \left(\sum_{s=1}^{S} \hat{\boldsymbol{W}}_s^{(Y)^\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(s)} \right), \tag{6.21}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(YY)} = \frac{1}{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}} \sum_{s=1}^{S} \left\{ \overline{\boldsymbol{V}}_{m,s}^{(YY)} + \overline{\gamma}_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(Y)} \hat{\boldsymbol{\mu}}_{m,s}^{(Y)^\top} - \left(\hat{\boldsymbol{\mu}}_{m,s}^{(Y)} \overline{\boldsymbol{Y}}_m^{(s)^\top} + \overline{\boldsymbol{Y}}_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(Y)^\top} \right) \right\}, \tag{6.22}$$

where

$$\hat{\boldsymbol{\nu}}_m = \left[\hat{\boldsymbol{b}}_m^{(0)^\top},\ \hat{\boldsymbol{b}}_m^{(1)^\top},\ \cdots,\ \hat{\boldsymbol{b}}_m^{(J)^\top} \right]^\top, \tag{6.23}$$

$$\hat{\boldsymbol{W}}_s^{(Y)} = \left[\boldsymbol{I},\ \hat{w}_1^{(s)} \boldsymbol{I},\ \hat{w}_2^{(s)} \boldsymbol{I},\ \cdots,\ \hat{w}_J^{(s)} \boldsymbol{I} \right], \tag{6.24}$$

$$\hat{\boldsymbol{\mu}}_{m,s}^{(Y)} = \hat{\boldsymbol{W}}_s^{(Y)} \hat{\boldsymbol{\nu}}_m. \tag{6.25}$$

The sufficient statistics for these estimates are given by

$$\overline{\gamma}_m^{(s)} = \sum_{t=1}^{Ts} P\left(m | \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)} \right), \tag{6.26}$$

$$\overline{\boldsymbol{Y}}_{m}^{(s)} = \sum_{t=1}^{Ts} P\Big(m|\boldsymbol{Y}_{t}^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)\, \boldsymbol{Y}_{t}^{(s)}, \tag{6.27}$$

$$\overline{\boldsymbol{V}}_{m,s}^{(YY)} = \sum_{t=1}^{T_s} P\Big(m|\boldsymbol{Y}_{t}^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big)\, \boldsymbol{Y}_{t}^{(s)} \boldsymbol{Y}_{t}^{(s)^{\top}}. \tag{6.28}$$

Note that, in the same as the adaptive training described in Chapter 4, it is difficult to estimate each parameter independently. Therefore, we need to perform these updates iteratively in each M-step for improving parameter estimation accuracy.

6.4. Experimental evaluations

6.4.1 Experimental conditions

In experimental evaluations, we evaluated the effectiveness of proposed many-to-many EVC algorithms and compared performance of the proposed non-parallel training with that of the conventional parallel training separately. For the experimental evaluation of the proposed conversion algorithms, we employed the proposed canonical one-to-many EV-GMM for spectral features and for aperiodic components described in Section 5.3 to perform many-to-many EVC. For the experimental evaluation for proposed training method, we trained the EV-GMM with the conventional parallel training using one male speaker as the reference speaker and 27 pre-stored speakers including 13 male and 14 female speakers selected from JNAS [54]. The trained EV-GMM was further refined with the proposed non-parallel training using 160 pre-stored speakers including 80 male and 80 female speakers. To demonstrate the effectiveness of increasing the number of pre-stored speakers used in the proposed non-parallel training, we varied the number of pre-stored speakers from 27 consisting of the same pre-stored speakers as used in the conventional parallel training to 160.

In the evaluations, we used eight speaker pairs (two male-to-male pairs, two female-to-female pairs, two male-to-female pairs, and two female-to-male pairs) selected from four male and five female speakers that were not included in the pre-stored speakers. We used 1 to 32 utterances for the adaptation, and the other 21 utterances for the evaluations.

We used 24-dimensional mel-cepstrum analyzed by STRAIGHT [13] as a spectral feature and aperiodic components [14] that were averaged on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 kHz) as an excitation feature to construct STME. Two one-to-many EV-GMMs were trained separately for these two features. The number of representative vectors was 159 for mel-cepstrum and 64 for aperiodic components, respectively. The number of mixture components was 128 for mel-cepstrum and 64 for the aperiodic components, respectively.

In evaluation for conversion algorithms, we compared the proposed many-to-many EVC algorithms based on multistep VC "M-to-M (multistep)" and based on shared mixture components "M-to-M (shared)" with traditional VC with the parallel training "Traditional." Note that the unsupervised adaptation was performed in the many-to-many EVC while the supervised training using parallel data was performed in the traditional VC.

6.4.2 Objective evaluations for conversion algorithms

We evaluated the conversion performance using mel-cepstral distortion for the spectral conversion and RMSE on aperiodic component for the aperiodicity conversion. Note that, average values of mel-cepstral distortion and RMSE between source and target speakers are 7.23 [dB] and 2.75 [dB], respectively. Figures 6.4 and 6.5 show results when varying the number of adaptation sentences (or the number of parallel training sentences in "Traditional"). The performance of both "M-to-M" methods is significantly better than "Traditional" when using a small amount of adaptation data. Moreover, we can observe that "M-to-M (shared)" outperforms "M-to-M (multistep)." When using more then 16 adaptation sentences, "Traditional" overcomes the proposed methods because a large enough amount of parallel data to train the GMM is available. Incidentally, we have never observed significant differences of the conversion performance between the within-gender conversion and the cross-gender conversion.

6.4.3 Subjective evaluations for conversion algorithms

We conducted a preference test on speech quality and an XAB test on conversion accuracy for speaker individuality. In the preference test, a pair of two different

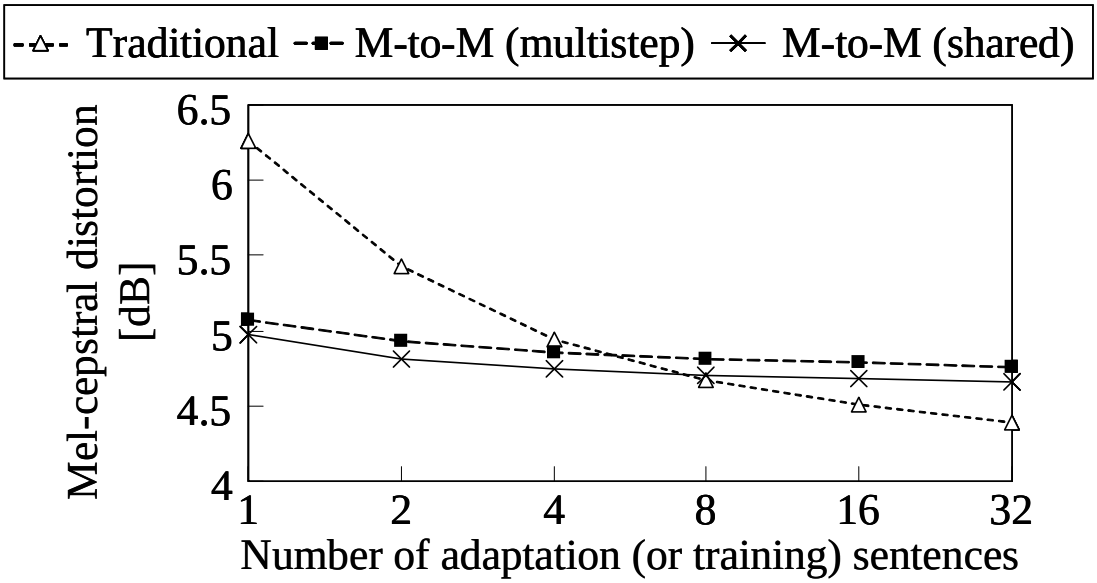


Figure 6.4. Result of objective evaluation by mel-cepstral distortion for many-to-many EVC algorithms.

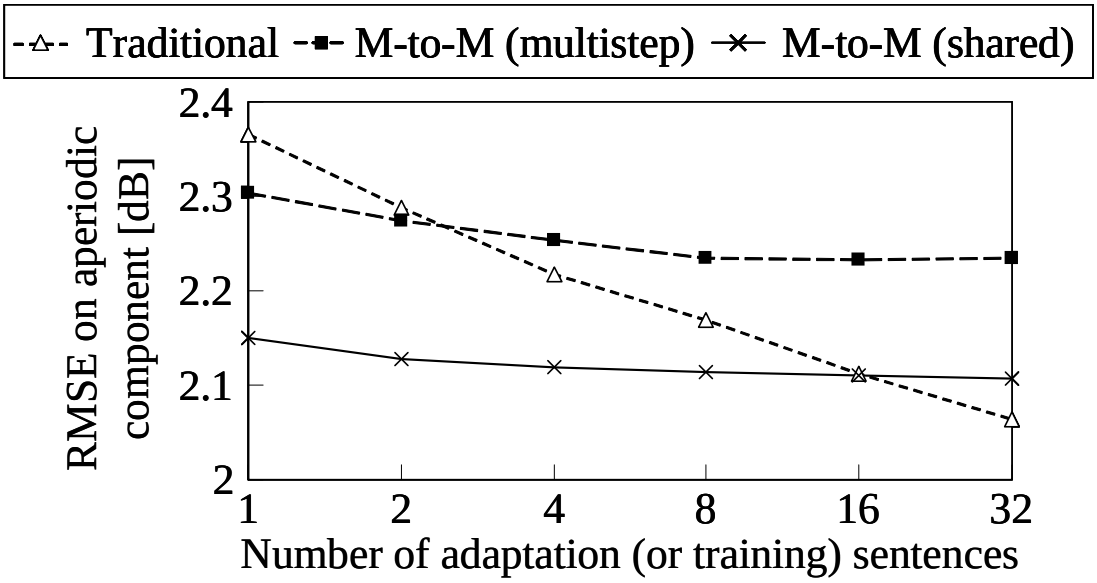


Figure 6.5. Result of objective evaluation by RMSE on aperiodic components for many-to-many EVC algorithms.

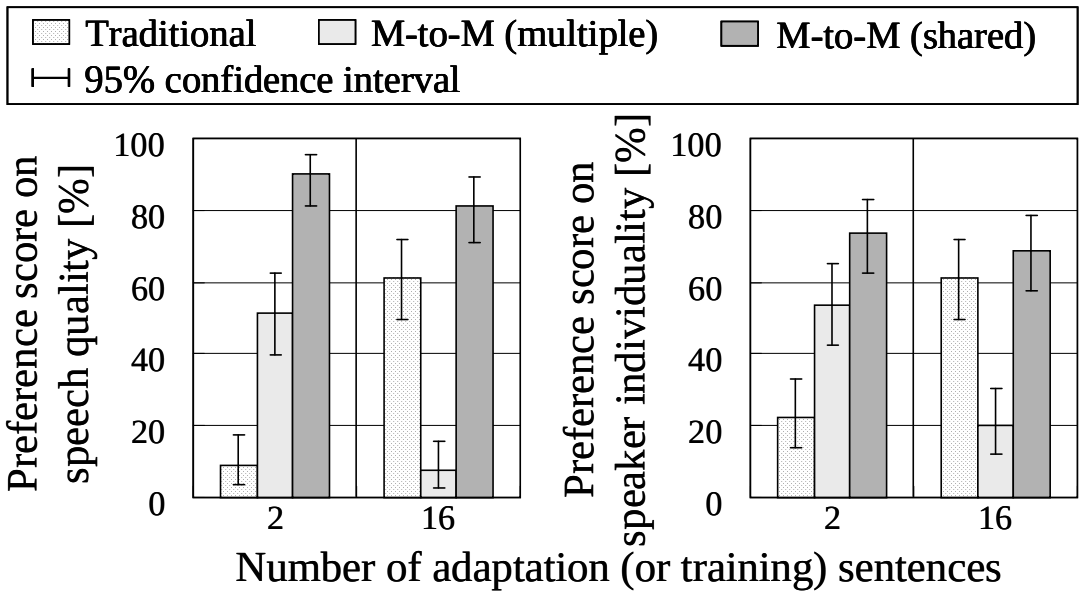


Figure 6.6. Results of subjective evaluations for many-to-many EVC algorithms.

types of the converted speech was presented to listeners, and then they were asked which voice sounded better. In the XAB test, a pair of two different types of the converted speech was presented to them after presenting the target speech as a reference. Then, they were asked which voice sounded more similar to the reference. Each listener evaluated every pair-combination of all types of the converted speech. The number of listeners was ten and the number of sample pairs evaluated by each listener was 48 in each test. Note that these converted utterances were generated by each conversion algorithm with GV.

Figure 6.6 shows the results. When using two adaptation sentences, both "M-to-M" algorithms outperform "Traditional". Moreover, "M-to-M (shared)" significantly outperforms "M-to-M (multistep)". When using 16 adaptation sentences, "M-to-M (shared)" still has the best performance in both speech quality and conversion accuracy for speaker individuality. Although "Traditional" would outperform the proposed methods when using a larger amount of training data, parallel data are still indispensable. These results suggest that the proposed many-to-many EVC with shared mixture components is very effective for flexibly

developing conversion models for arbitrary speaker-pairs.

6.4.4 Objective evaluation for proposed training method

We evaluated spectral conversion performance using mel-cepstral distortion. Figure 6.7 shows the result when varying the number of pre-stored speakers used in the proposed non-parallel training. When using the same 27 pre-stored speakers as used in the conventional parallel training, the proposed non-parallel training method causes degradation of conversion performance. This would be reasonable because the non-parallel training data sets are less informative than the parallel training data sets in this case. It is observed that the proposed non-parallel training yields better conversion performance as the number of pre-stored speakers in the non-parallel data increases. This is because the proposed training method effectively updates the EV-GMM parameters so that the EV-GMM models well voice characteristics of a larger number of speakers; e.g., representative vectors are updated so that a sub-space spanned by them widely covers more varieties of speakers. Consequently, the proposed non-parallel training provides significant improvements in conversion performance when using a much larger number of pre-stored speakers than that used in the conventional parallel training.

6.4.5 Subjective evaluations for proposed training method

We compared the converted speech samples of the proposed non-parallel training with those of the conventional parallel training. We used 27, 80 and 160 speakers for the proposed non-parallel training. Note that we generated these converted speech data using STSE and conversion algorithm not considering the GV described in Section 6.2.2. We conducted a preference test on speech quality and an XAB test on conversion accuracy for speaker individuality. In the preference test, a pair of two different types of the converted speech was presented to listeners, and then they were asked which voice sounded better. In the XAB test, a pair of two different types of the converted speech was presented to them after presenting the target speech as a reference. Then, they were asked which voice sounded more similar to the reference target. The number of listeners was ten and the number of sample-pairs evaluated by each listener was 48 in each test.

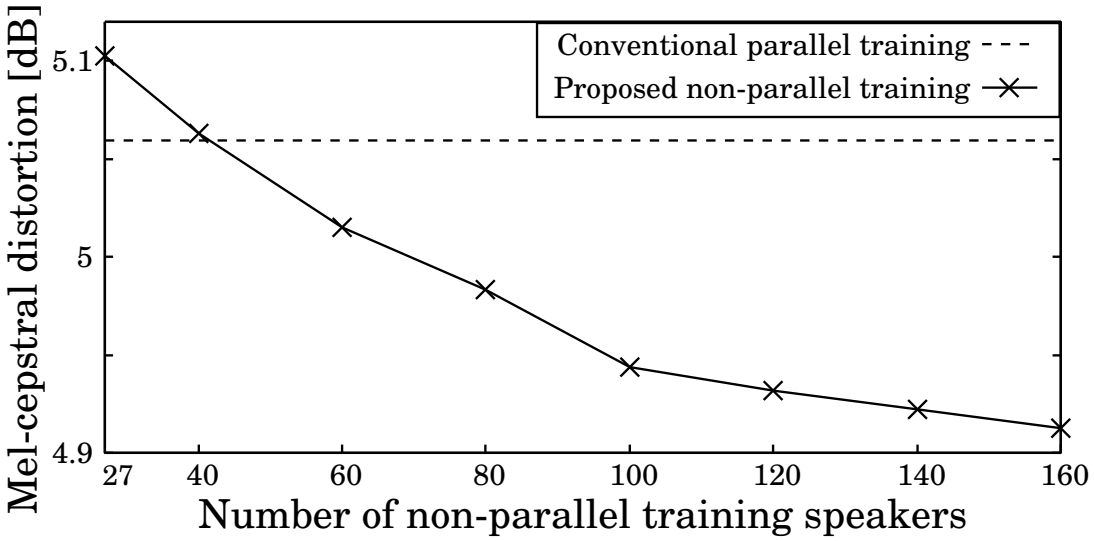


Figure 6.7. Mel-cepstral distortion as a function of the number of non-parallel training speakers.

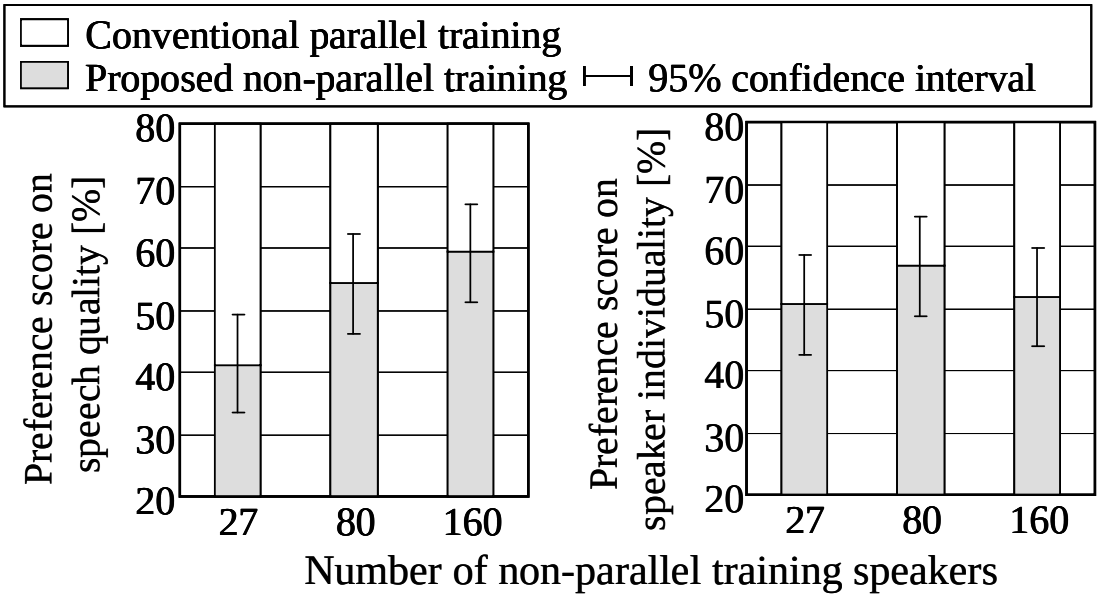


Figure 6.8. Results of subjective evaluations for proposed refining EV-GMM.

Figure 6.8 shows the experimental results. In the speech quality test, we can see almost the same tendency as observed in the previous objective evaluation; i.e., better speech quality is obtained by increasing the number of pre-stored speakers in the proposed non-parallel training; and the proposed non-parallel training yields significant quality improvements in converted speech compared with the conventional parallel training when using 160 pre-stored speakers. On the other hand, in the speaker individuality test, conversion accuracy of the proposed non-parallel training is kept almost equal to that of the conventional parallel training.

These results suggest that the proposed non-parallel training yields better converted speech quality than the conventional parallel training without degradation of conversion accuracy for speaker individuality by additionally using non-parallel data including a larger number of pre-stored speakers.

6.5. Summary

In this chapter, we have described a novel EVC framework, many-to-many EVC. Many-to-one EVC and one-to-many EVC has been integrated into many-to-many EVC by using the single EV-GMM and sharing mixture components between two EVCs. The proposed many-to-many EVC allows us to develop a conversion model from an arbitrary source speaker to an arbitrary target speaker by the unsupervised adaptation using a small amount of adaptation data of the arbitrary source and target speakers. When converting, a source speaker's voice is effectively converted into a target speaker's voice by considering the reference speaker's voice as hidden variable.

Moreover, this chapter has described the EV-GMM training method using non-parallel data sets for many-to-many EVC. This method has been inspired by above many-to-many conversion algorithm. In the proposed training method, an initial EV-GMM is trained using parallel data sets consisting of a single reference speaker and multiple pre-stored speakers. And then, the initial EV-GMM is further refined additionally using non-parallel data sets consisting of a larger number of pre-stored speakers while considering the reference speaker's voices as hidden variables.

The experimental results have demonstrated that our proposed algorithm has effectively performed and the proposed training method has yielded significant quality improvements in converted speech by effectively using non-parallel data sets including a larger number of pre-stored speakers.

Chapter 7

Conclusions

7.1. Summary of thesis

Voice conversion (VC) is a technique for converting the source speaker's voice into a target speaker's voice without changing linguistic information using a statistical conversion model. As a statistical conversion model, the Gaussian mixture model (GMM) is trained with a parallel data set consisting of utterance-pairs of source and target speakers. Although this framework works reasonably well, the converted speech quality is still insufficient and the training process of the conversion model is less flexible. Recently, eigenvoice conversion (EVC) was proposed in order to make the conversion model training process more flexible. EVC has brought two new conversion paradigms, i.e., one-to-many VC performing the conversion from a single source speaker to arbitrary target speakers, and many-to-one VC performing the conversion from arbitrary source speakers to a single target speaker. However, the converted speech quality of the EVC is not high enough. Moreover, it is desired to further improve flexibility of the conversion model training to make VC applications more practical. In this thesis, in order to improve the performance of EVC, we have proposed two approaches to improving the converted speech quality of EVC and achieving a more flexible conversion paradigm.

First, we described the traditional VC and EVC frameworks in **Chapter 2**. We explained the GMM training process of the traditional VC framework and two main conversion algorithms, i.e., the conversion algorithm based on minimum

mean square error and that based on maximum likelihood (ML) estimation. As the state-of-the art conversion algorithms, we also explained the maximum likelihood conversion algorithm considering global variance (GV) that is defined as variances of the feature vectors over a time sequence. This conversion algorithm has dramatically improved the converted speech quality. Next, we described the one-to-many EVC framework in detail. In this framework, the eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets consisting of the pre-defined source speaker and various pre-stored target speakers. The conversion model from the source speaker to the desired target speaker is constructed by adapting the trained EV-GMM to the target speaker using a small number of arbitrary target utterances. In the conversion process, arbitrary utterances of the source speaker are converted into those of the target speaker. We also described issues of the converted speech quality in the conventional one-to-many EVC system. The severe degradation of the converted speech quality is often caused by STRAIGHT simple excitation, inter-speaker variations captured by the EV-GMM, and the conversion algorithm not considering the GV. In addition, we reviewed various VC techniques related to the traditional VC and EVC.

In **Chapter 3**, we improved the quality of the source excitation model in the traditional GMM-based VC. We conventionally employed the STRAIGHT simple excitation model that generates an excitation signal by switching a phase-manipulated pulse train or white noise based on F_0 information. This excitation is too simple to precisely model an excitation signal because it generally consists of both pulse and noise signals. In order to improve the excitation model, we introduced the STRAIGHT mixed excitation (STME) model to the traditional GMM-based VC. STME generates an excitation signal based on the weighted sum of a phase-manipulated pulse train and white noise. The weighting values vary according to aperiodic components capturing the strength of noise components in each frequency bin. In our proposed method, joint probability density of aperiodic components between source and target speakers is modeled by a GMM. The aperiodic components converted from the source aperiodic components are used for generating an excitation signal with STME. In objective evaluations, we determined the optimal mapping parameter from the aperiodic components into the weighting values. The experimental result has demonstrated that the

converted aperiodic components are more similar to the target aperiodic components compared with the source aperiodic components. The result of subjective evaluation has demonstrated that the proposed GMM-based VC with STME significantly outperforms the conventional GMM-based VC in view of conversion accuracy for speaker individuality.

In **Chapter 4**, we proposed a novel training method for the eigenvoice GMM (EV-GMM). In the conventional one-to-many EVC, most parameters of the EV-GMM are from target-speaker-independent GMM (TI-GMM). TI-GMM models not only intra-speaker variations but also inter-speaker variations. Therefore, the adapted EV-GMM is also affected by the inter-speaker variations, and this causes significant degradation of conversion performance. In order to improve the conversion performance, we proposed an adaptive training method for the EV-GMM. This training method effectively reduces inter-speaker variations. On the other hand, a local optimum problem is caused because the EM algorithm is employed in this training. The noticeable quality degradation in the converted speech is caused by this problem especially when the number of adaptation parameters of the EV-GMM decreases. In order to ameliorate this problem, we proposed two methods, approximation of occupancy probabilities and adaptive training based on deterministic annealing EM (DAEM). In the former method, target-speaker-dependent GMMs (TD-GMMs) are used for calculating occupancy probabilities in the first E-step. In the latter method, the DAEM algorithm instead of the normal EM algorithm is applied to the adaptive training. We evaluated our proposed adaptive training objectively and subjectively. In objective evaluations, we first investigated covariance values of the EV-GMM and we confirmed that the covariance values of the proposed EV-GMM were almost equal to those of the GMM trained between a single speaker-pair. These results show that the proposed adaptive training effectively reduces the impact of inter-speaker variations on the EV-GMM. We have also clarified that the proposed EV-GMM outperforms the conventional EV-GMM in terms of spectral conversion accuracy, and the proposed methods for alleviating the local optimum problem are effective. In the subjective evaluation, we compared the converted speech quality of the conventional EV-GMM with that of the proposed EV-GMM. The experimental results have demonstrated that our proposed adaptive training significantly improves the

converted speech quality.

In **Chapter 5**, we have developed an improved one-to-many EVC system. In this system, we introduced three promising techniques, i.e., STME, MLE-based conversion considering the GV, and the adaptive training for the EV-GMM. For using STME, we modeled aperiodic components by the EV-GMM built with the proposed adaptive training. In the conversion algorithm, the GV was modeled by an eigenvoice single Gaussian distribution (EV-GS) also built with the adaptive training. We evaluated our proposed system objectively and subjectively. In the objective evaluation, we clarified the effectiveness of the adaptive training applied to the EV-GMM for aperiodic components and the EV-SG for the GV. In the subjective evaluation, we demonstrated that the converted speech generated by our proposed system has much better quality and conversion accuracy for speaker individuality compared to that by the conventional system.

In **Chapter 6**, we proposed many-to-many EVC as a novel EVC framework. Many-to-many EVC is a technique for converting an arbitrary source speaker's voice into an arbitrary target speaker's voice. This framework is achieved by performing many-to-one EVC and one-to-many EVC sequentially with a single EV-GMM between the reference and many pre-stored speakers. In this conversion process, using two conversion models that are constructed by respectively adapting the single EV-GMM to source and target speakers, we convert the source speaker's voice into the target speaker's voice through the reference speaker's voice considered as a hidden variable. Moreover, inspired by this conversion algorithm, we proposed refining the EV-GMM method. In this method, the canonical EV-GMM for many-to-many EVC is retrained with a large amount of non-parallel data sets by considering the reference voice as a hidden variable. Objective and subjective evaluations demonstrated the effectiveness of our proposed methods in many-to-many EVC.

In summary, our proposed one-to-many EVC system has improved the converted speech quality and conversion accuracy for speaker individuality by introducing STME, the conversion algorithm considering the GV and adaptive training. Moreover, the effectiveness of our proposed many-to-many EVC algorithm and EV-GMM refinement algorithm was confirmed by objective and subjective evaluations.

7.2. Future work

We have improved the flexibility of the conversion model training and the converted speech quality. However, there are still several problems to be solved for EVC applications. In this section, we describe several issues and proposals for the many-to-many EVC framework.

• Improvement of training the EV-GMM for many-to-many voice conversion

In Section 6.3, we have proposed to refine the EV-GMM method using non-parallel data sets. Although the proposed training improves the EV-GMM performance, mismatches between the reference and pre-stored speakers' parameters are caused because this training method cannot update parameters related to the reference speaker. Therefore, we need to consider the EV-GMM refining method and alleviate these mismatches. In addition, we have another model training problem. In our proposed conversion algorithm, the mapping parameter written as in Eq. (6.13) depends on the reference speaker's information. That is, the reference speaker's voice characteristics affect conversion performance directly. Therefore, we need to consider how to select the proper reference speaker.

• Improvement of speaker individuality

Speaker individuality depends not only on voice quality but also on prosody and duration. However, in the many-to-many EVC framework, prosody and duration of the converted speech are based on those of the source speech. Therefore, we need to apply statistical modeling of prosody and duration to this framework in order to improve speaker individuality. In previous studies of the traditional GMM-based framework, Nankaku et al. proposed VC based on GMM including time sequence matching [56] for solving the duration problem, and Uto et al. have proposed F_0 conversion using GMM based on multi-space probability distribution [57]. We consider that speaker individuality of the many-to-many EVC can be improved by applying these ideas.

- **Various applications using many-to-many EVC**

For the practical use of VC techniques, it is important to achieve real-time processing and usability of voice quality control. For example, VC applications for conversations such as band extension for mobile phones [4] and speech communication aid systems [58] demand real-time conversion. Also, entertainment content using voice such as singing and instant casting movies [20][21] demand voice quality control. For the former requirement, Muramatsu et al. have proposed real-time VC considering dynamic features [59] by applying a time-recursive algorithm [60][61]. For the letter requirement, Ohta et al. have proposed a voice-control system for one-to-many EVC [41][62]. By applying these frameworks to many-to-many EVC, VC applications are available to more varied fields.

Appendix

A. Parameter estimations of adaptive training for EV-GMM

In this appendix, we describe the solution of the EV-GMM's parameters estimated by adaptive training, addressed in Chapter 4. In this adaptive training, these parameters are estimated by maximizing the expectation of the log-scaled likelihood written as Eq. (6.18) based on EM algorithm. Eq. (6.18) is developed as follows:

$$
\begin{aligned}
&Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}_1^S)\right\}\right) \\
&= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M} P\Big(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\Big) \left\{\log \hat{\alpha}_m - D\log 2\pi - \frac{1}{2}\log\left|\hat{\boldsymbol{\Sigma}}_m^{(X,Y)}\right|\right. \\
&\qquad \left. - \frac{1}{2}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix} - \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)}\right)^{\top} \hat{\boldsymbol{\Sigma}}_m^{(X,Y)} \left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix} - \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)}\right)\right\}.
\end{aligned}
\tag{A.1}
$$

Note that this parameter estimation is updated based on Eq. (4.3), i.e., pre-stored target speaker's weight vectors, tied-parameter set corresponded to source and target mean vectors, mixtures weight and covariance matrices in that order.

A.1 Estimation of weight vector for each pre-stored target speaker

We estimate the s^{th} target speaker's weight for representative vectors. The optimal weight vector $\hat{\boldsymbol{w}}^{(s)}$ is determined by maximizing the following auxiliary

function for weight vectors:

$$\frac{\partial}{\partial \hat{\boldsymbol{w}}^{(s)}} Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\lambda^{(EV)}, \hat{\boldsymbol{w}}_1^S\right\}\right)$$
$$= -\frac{1}{2}\sum_{t=1}^{T_s}\sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right)\left\{2\boldsymbol{B}_m^\top \boldsymbol{P}_m^{(YX)}\left(\boldsymbol{X}_t - \hat{\boldsymbol{\mu}}_m^{(X)}\right)\right.$$
$$\left. - 2\boldsymbol{B}_m^\top \boldsymbol{P}_m^{(YY)}\left(\boldsymbol{Y}_t^{(s)} - \boldsymbol{b}_m^{(0)} - \hat{\boldsymbol{B}}_m \hat{\boldsymbol{w}}^{(s)}\right)\right\} = \boldsymbol{0}. \tag{A.2}$$

Therefore, the estimated weight vector $\hat{\boldsymbol{w}}^{(s)}$ is written as

$$\hat{\boldsymbol{w}}^{(s)} = \left(\sum_{m=1}^{M} \overline{\gamma}_m^{(s)} \boldsymbol{B}_m^\top \boldsymbol{P}_m^{(YY)} \boldsymbol{B}_m\right)^{-1}$$
$$\times \left[\sum_{m=1}^{M} \boldsymbol{B}_m^\top \left\{\boldsymbol{P}_m^{(YX)}\left(\overline{\boldsymbol{X}}_m^{(s)} - \overline{\gamma}_m^{(s)} \boldsymbol{\mu}_m^{(X)}\right) + \boldsymbol{P}_m^{(YY)}\left(\overline{\boldsymbol{Y}}_m^{(s)} - \overline{\gamma}_m^{(s)} \boldsymbol{b}_m^{(0)}\right)\right\}\right]. \tag{A.3}$$

A.2 Estimations of tied-parameter set for mean vectors

We simultaneously estimate tied-parameters related to source and target mean vectors, i.e., source mean vectors, bias vectors and representative vectors. We set $\boldsymbol{\nu}_m$ by concatenating source mean vector, bias vector and each representative vector included in the m^{th} mixture component. We determine optimal source mean vectors, bias vectors and representative vectors by partially differentiating Eq. (A.1) with respect to $\hat{\boldsymbol{\nu}}_m$ as follows:

$$\frac{\partial}{\partial \hat{\boldsymbol{\nu}}_m} Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\alpha_m, \boldsymbol{\Sigma}_m^{(X,Y)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}, \hat{\boldsymbol{w}}_1^S\right\}\right)$$
$$= -\frac{1}{2}\sum_{s=1}^{S}\sum_{t=1}^{T_s} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right)$$
$$\times \left\{2\hat{\boldsymbol{W}}_s \boldsymbol{\Sigma}_m^{(X,Y)^{-1}}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix} - \hat{\boldsymbol{W}}_s \hat{\boldsymbol{\nu}}_m\right)\right\} = \boldsymbol{0}, \tag{A.4}$$

where $\boldsymbol{W}_s$ is the matrix which includes the s^{th} speaker's weight components for eigenvoices written as Eq. (4.13). And then, the concatenated vector $\hat{\boldsymbol{\nu}}_m$ is written

as

$$\hat{\boldsymbol{\nu}}_m = \left(\sum_{s=1}^{S} \overline{\gamma}_m^{(s)} \hat{\boldsymbol{W}}_s^{\top} \boldsymbol{\Sigma}_m^{(X,Y)^{-1}} \hat{\boldsymbol{W}}_s \right)^{-1} \left(\sum_{s=1}^{S} \hat{\boldsymbol{W}}_s^{\top} \boldsymbol{\Sigma}_m^{(X,Y)^{-1}} \overline{\boldsymbol{Z}}_m^{(s)} \right). \tag{A.5}$$

In Eq. (A.5), we need to calculate $\{D\,(J+2)\} \times \{D\,(J+2)\}$-sized inverse matrix. When using diagonal covariance matrices for $\boldsymbol{\Sigma}^{(XX)}$, $\boldsymbol{\Sigma}^{(XY)}$, $\boldsymbol{\Sigma}^{(YX)}$ and $\boldsymbol{\Sigma}^{(YY)}$, the computational cost is significantly reduced by calculating the ML estimate of $\boldsymbol{\nu}_m$ dimension by dimension as follows:

$$\hat{\boldsymbol{\nu}}_m^{(d)} = \left(\sum_{s=1}^{S} \overline{\gamma}_m^{(s)} \hat{\boldsymbol{W}'}_s^{\top} \boldsymbol{P}_{m,d}^{(X,Y)} \hat{\boldsymbol{W}'}_s \right)^{-1} \left(\sum_{s=1}^{S} \hat{\boldsymbol{W}'}_s^{\top} \boldsymbol{P}_{m,d}^{(X,Y)} \overline{\boldsymbol{Z}}_{m,d}^{(s)} \right), \tag{A.6}$$

where

$$\hat{\boldsymbol{\nu}}_m^{(d)} = \left[\hat{\mu}_{m,d}^{(X)}, \hat{b}_{m,d}^{(0)}, \hat{b}_{m,d}^{(1)}, \cdots, \hat{b}_{m,d}^{(J)} \right]^{\top}, \tag{A.7}$$

$$\hat{\boldsymbol{W}'}_s = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \hat{w}_1^{(s)} & \hat{w}_2^{(s)} & \cdots & \hat{w}_J^{(s)} \end{bmatrix}, \tag{A.8}$$

$$\boldsymbol{P}_{m,d}^{(X,Y)} = \begin{bmatrix} p_{m,d}^{(XX)} & p_{m,d}^{(XY)} \\ p_{m,d}^{(YX)} & p_{m,d}^{(YY)} \end{bmatrix}, \tag{A.9}$$

$$\overline{\boldsymbol{Z}}_{m,d}^{(s)} = \left[\overline{X}_{m,d}^{(s)}, \overline{Y}_{m,d}^{(s)} \right]^{\top}. \tag{A.10}$$

The vector $\boldsymbol{\nu}_m^{(d)}$ consists of the d^{th} dimensional components of $\boldsymbol{\nu}_m$, and $\mu_{m,d}^{(X)}$ and $b_{m,d}^{(j)}$ are the d^{th} components of $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{b}_m^{(j)}$, respectively. In Eq. (A.9), $p_{m,d}^{(XX)}$, $p_{m,d}^{(XY)}$, $p_{m,d}^{(YX)}$ and $p_{m,d}^{(YY)}$ are the d^{th} diagonal components of $\boldsymbol{P}_m^{(XX)}$, $\boldsymbol{P}_m^{(XY)}$, $\boldsymbol{P}_m^{(YX)}$ and $\boldsymbol{P}_m^{(YY)}$, respectively. In Eq. (A.10), $\overline{X}_{m,d}^{(s)}$ and $\overline{Y}_{m,d}^{(s)}$ are the d^{th} components of $\overline{\boldsymbol{X}}_m^{(s)}$ and $\overline{\boldsymbol{Y}}_m^{(s)}$, respectively. We need to calculate $\{J+2\} \times \{J+2\}$-sized inverse matrices for each dimension.

A.3 Estimations of covariance matrices for canonical EV-GMM

The optimal covariance matrix $\hat{\boldsymbol{\Sigma}}_m^{(X,Y)}$ is determined by, the partial differential equation of Eq. (A.1) by $\hat{\boldsymbol{\Sigma}}_m^{(X,Y)}$ written as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \hat{\boldsymbol{\Sigma}}_m^{(X,Y)}} Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\alpha_m, \hat{\boldsymbol{\Sigma}}_m^{(X,Y)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}, \hat{\boldsymbol{w}}_1^S\right\}\right) \\
&= \sum_{s=1}^{S}\sum_{t=1}^{T_s} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}\right) \\
&\qquad \times \left\{-\frac{1}{2}\hat{\boldsymbol{\Sigma}}_m^{(X,Y)^{-1}} + \frac{1}{2}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)}\end{bmatrix} - \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)}\right)\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)}\end{bmatrix} - \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)}\right)^\top \right. \\
&\qquad\qquad \left. + \hat{\boldsymbol{\Sigma}}_m^{(X,Y)^{-1}}\hat{\boldsymbol{\Sigma}}_m^{(X,Y)^{-1}{}^\top}\right\} = \boldsymbol{0}.
\end{aligned} \tag{A.11}$$

This solution is written as

$$\hat{\boldsymbol{\Sigma}}_m^{(X,Y)} = \frac{1}{\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}}\sum_{s=1}^{S}\left\{\overline{\boldsymbol{V}}_{m,s}^{(X,Y)} + \overline{\gamma}_m^{(s)}\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)}\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)^\top} - \left(\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)}\overline{\boldsymbol{Z}}_m^{(s)^\top} + \overline{\boldsymbol{Z}}_m^{(s)}\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)^\top}\right)\right\}. \tag{A.12}$$

A.4 Estimations of weights for mixture components

Each mixture component weight is determined by Lagrange's method of undetermined multipliers. We define the function of $\{\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_m\}$ and Λ, which is a variable except zero, as follows:

$$\begin{aligned}
&F\left(\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_m, \Lambda\right) \\
&\quad = Q\left(\left\{\lambda^{(EV)}, \boldsymbol{w}_1^S\right\}, \left\{\hat{\alpha}_m, \hat{\boldsymbol{\Sigma}}_m^{(X,Y)}, \hat{\boldsymbol{B}}_m, \hat{\boldsymbol{b}}_m^{(0)}, \hat{\boldsymbol{\mu}}_m^{(X)}, \hat{\boldsymbol{w}}_1^S\right\}\right) - \Lambda\left(\sum_{m=1}^{M}\hat{\alpha}_m - 1\right),
\end{aligned} \tag{A.13}$$

$$\text{Subject to} \quad \sum_{m=1}^{M}\hat{\alpha}_m = 1. \tag{A.14}$$

This function is partially differentiated by $\hat{\alpha}_m$ as follows:

$$\frac{\partial}{\partial \hat{\alpha}_m} F\left(\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_m, \Lambda\right) = \frac{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}{\hat{\alpha}_m} - \Lambda = 0. \tag{A.15}$$

We multiply both sides of Eq. (A.15) by $\hat{\alpha}_m$ and sum up all of Eq. (A.15) related to each mixture component, and then we obtain Λ as follows:

$$\Lambda = \sum_{s=1}^{S} \sum_{m=1}^{M} \overline{\gamma}_m^{(s)}. \tag{A.16}$$

Thus, the optimal weight for the m^{th} mixture component $\hat{\alpha}_m$ is determined as

$$\hat{\alpha}_m = \frac{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}{\sum_{s=1}^{S} \sum_{m=1}^{M} \overline{\gamma}_m^{(s)}}. \tag{A.17}$$

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustic Society of Japan (E)*, vol. 1, no. 2, pp. 71–76, 1990.

[2] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *The Journal of the Acoustic Society of America*, vol. 90, no. 1, pp. 76–82, 1991.

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

[4] K.-H. Park and H.S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," *Proc ICASSP*, vol. 3, pp. 1843–1846, Istanbul, Turkey, 2000.

[5] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, and K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with GMM," *2008 RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08)*, pp. 283–286, Gold Coast, Australia, March 2008.

[6] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano, "Technologies for processing body-conducted speech detected with non-audible murmur microphone," *Proc. Interspeech 2009 - Eurospeech*, pp. 632–635, Brighton, U.K., September 2009.

[7] Y. Nakajima, H. Kashioka, N. Campppbell, and K. Shikano, "Non-audible murmur (NAM) recognition," *IEICE Trans. Info. and Syst.*, vol. E89-D, no. 1, pp. 1–8, 2006.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

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

[10] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, November 2007.

[11] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," *Proc. Interspeech 2006 - ICSLP*, pp. 2446–2449, Pittsburgh, U.S.A., September 2006.

[12] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, vol. 4, pp. 1249–1252, Hawaii, U.S.A., April 2007.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and 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.

[14] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA 2001*, Firentze, Italy, September 2001.

[15] T. Anastasakos, J. McDonough, Schwartz R., and J. Makhoul, "A compact model for speaker-adaptive training," *Proc. ICSLP*, vol. 2, pp. 1137–1140, Atlanta, U.S.A, 1996.

[16] M. Slaney, M. Covell, and B Lassiter, "Automatic audio morphing," *Proc. ICASSP*, pp. 1001–1004, Atlanta, U.S.A, 1996.

[17] H. Kawahara and H Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," *Proc. ICASSP*, vol. 1, Hong Kong, PRC, 256–259 2003.

[18] T. Takahashi, M. Nishi, T. Irino, and H Kawahara, "Average voice synthesis using multiple speech morphing," *Proc. 2006 Spring Meeting of Acoustic Society of Japan*, pp. 229–230, March 2006 (in Japanese).

[19] Y. Ohtani, S. Kawamoto, T. Toda, S. Nakamura, and K. Shikano, "Specific speech generation based on STRAIGHT morphing," *Proc. 2008 Spring Meeting of Acoustic Society of Japan*, pp. 309–310, March 2008 (in Japanese).

[20] S. Kawamoto, Y. Adachi, Y. Ohtani, T. Yotsukura, S. Morishima, and S. Nakamura, "Scenario speech assignment technique for instant casting movie system," *ACCV2009 Invited workshop on Vision Based Human Modeling and Synthesis*, Xi'an, China, September 2009.

[21] S. Kawamoto, Y. Adachi, Y. Ohtani, T. Yotsukura, S. Morishima, and S. Nakamura, "Voice output system considering personal voice for instant casting movie," *IPSJ Journal*, vol. 51, no. 2, pp. 1234–1248, February 2010 (in Japanese).

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

[23] E.K. Kim, S. Lee, and Y.H. Oh, "Hidden Markov model based voice conversion using dynamic characteristics of speaker," *Proc. Eurospeech*, vol. 5, pp. 2519–2522, Rhodes, Greece, September 1997.

[24] S.J. Yun and Y.H Oh, "Performance improvement of speaker recognition system for small training data," *Proc. ICSLP*, pp. 1863–1866, 1994.

[25] 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, October 2000.

[26] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to english," *Proc. IEEE Workshop Speech Synth.*, pp. 227–230, September 2002.

[27] 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, September 1999.

[28] K. Tokuda, K. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP*, pp. 660–663, Detroit, U.S.A., May 1995.

[29] K. Tokuda, T. Masuko, T. Yamada, K. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," *Proc. Eurospeech*, pp. 757–760, September 1995.

[30] 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 2000.

[31] A. Mouchtaris, J.V. der Spiegel, and P. Mueller, "Non-parallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 952–963, May 2006.

[32] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 5, pp. 389–406, September 1997.

[33] C.-H. Lee and C.-H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," *Proc. Interspeech 2006 - ICSLP*, pp. 2254–2257, Pittsburgh, U.S.A., September 2006.

[34] GJ.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[35] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, April 2000.

[36] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T Kitamura, "Eigenvoice for HMM-based speech synthesis," *Proc. Interspeech 2002 - ICSLP*, pp. 1269–1272, Denver, U.S.A., September 2002.

[37] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, "A technique for controlling voice quality of synthetic speech using multiple regression HSMM," *Proc. Interspeech 2006 - ICSLP*, pp. 2438–2441, Pittsburgh, U.S.A., September 2006.

[38] T. Nose, M. Tachibana, and T. Kobayash, "HMM-based style control for expressive speecch synthesis with arbitrary speaker's voice using model adaptation," *IEICE Trans. Inf. and Syst.*, vol. Vol. E92-D, no. 3, pp. 489–497, 2009.

[39] 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.

[40] D. Tani, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Maximum a posteriori adaptation for many-to-one eigenvoice conversion," *Proc. Interspeech 2008 - ICSLP*, pp. 1461–1464, Brisbane, Australia, September 2008.

[41] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Regression approaches to voice quality control based on one-to-many eigenvoice conversion," *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 101–106, Bonn, Germany, August 2007.

[42] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," *Proc. Interspeech 2009 - Eurospeech*, pp. 1635–1638, Brighton, U.K., September 2009.

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

[44] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," *Proc. ICASSP*, vol. 2, pp. 1303–1306, April 1997.

[45] E. Moulines, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–206, 1995.

[46] A. Kain and M.W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," *Proc. ICASSP*, pp. 813–816, Salt Lake City, U.S.A., May 2001.

[47] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A study on residual prediction techniques for voice conversion," *Proc. ICASSP*, pp. 13–16, Philadelphia, U.S.A., March 2005.

[48] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1301–1312, May 2006.

[49] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. and syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[50] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database," *ATR technical report*, , no. TR-I-0166, September 1990.

[51] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[52] M. J. F. Gales, "Cluster adaptive training for hidden Markov models," *IEEE Trans. Speech and Audio Process*, vol. 8, no. 4, pp. 417–428, 2000.

[53] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *IEICE Trans. Inf. and syst.*, vol. J80-D-2(1), pp. 267–276, October 1997 (in Japanese).

[54] "JNAS: Japanese Newspaper Article Sentences," *http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html*.

[55] T. Masuda and M. Shozakai, "Cost reduction of training mapping function based on multistep voice conversion," *Proc. ICASSP*, vol. 4, pp. 693–696, Hawaii, U.S.A., April 2007.

[56] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, "Spectral conversion based on statistical models including time-sequence matching," *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 333–338, Bonn, Germany, August 2007.

[57] Y. Uto, Y. Nankaku, A. Lee, and K. Tokuda, "Simultaneous modeling of spectrum and F_0 for voice conversion," *IEICE technical report, Speech*, vol. 107, no. 406 (SP2007-113), pp. 103–108, December 2007 (in Japanese).

[58] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "A speech communication aid system for total laryngectomees using voice conversion of body transmitted articial speech," *IEICE Trans. Inf. and syst.*, vol. J90-D, no. 3, pp. 780–787, March 2007 (in Japanese).

[59] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Proc. Interspeech 2008 - ICSLP*, pp. 1309–1312, September 2008.

[60] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP*, vol. 1, pp. 660–663, May 1995.

[61] K. Koishida, K. Tokuda, T. Masuko, and T. Kobayashi, "Vector quantization of speech spectral parameters using statistics of dynamic features," *IEICE Trans. Inf. and syst.*, vol. E84-D, no. 10, pp. 1427–1434, October 2001.

[62] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Enhanced voice quality control method based on one-to-many eigenvoice conversion," *Proc. 2008 Spring Meeting of Acoustic Society of Japan*, pp. 345–346, Chiba, Japan, March 2008 (in Japanese).

List of Publications

Journal papers

1. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Maximum likelihood voice conversion based on Gaussian mixture model with STRAIGHT mixed excitation," *IEICE Transactions Information and systems*, vol. J91-D, no. 4, pp. 1082–1091, April 2008 (in Japanese).

2. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Adaptive training for voice conversion based on eigenvoices," *IEICE Transactions Information and Systems*, vol. E93-D, no. 6, June 2010 (accept).

3. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Improvements of the one-to-many eigenvoice conversion system," *IEICE Transactions Information and Systems*, September 2010 (accept).

4. S. Kawamoto, Y Adachi, Y. Ohtani, T. Yotsukura, S. Morishima and S. Nakamura, "Voice output system considering personal voice for instant casting movie," *IPSJ Journal*, vol. 51, no. 2, pp. 1234–1248, February 2010 (in Japanese).

International conference

1. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Proc. Interspeech 2006 - ICSLP*, pp. 2266–2269, Pittsburgh, U.S.A., September, 2006.

2. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Evaluation of eigenvoice conversion based on Gaussian mixture model," *Proc. ASA/ASJ Joint Meeting*, Hawaii, U.S.A., November, 2006.

3. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model," *Proc. Interspeech 2007 - Eurospeech*, pp. 1981–1984, Antwerp, Belgium, August 2007.

4. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "An Improved one-to-many eigenvoice conversion system," *Proc. Interspeech 2008 - ICSLP*, pp. 1080–1083, Brisbane, Australia, September 2008.

5. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," *Proc. Interspeech 2009 - Eurospeech*, pp. 1623–1626, Brighton, U.K., September 2009.

6. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," *Proc. ICASSP*, pp. 4822–4825, Dallas, U.S.A., March 2010.

7. T. Toda, Y. Ohtani and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," *Proc. Interspeech 2006 - ICSLP*, pp. 2446–2449, Pittsburgh, U.S.A., September, 2006.

8. T. Toda, Y. Ohtani and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, vol. 4, pp. 1249–1252, Hawaii, U.S.A., April 2007.

9. K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Regression approaches to voice quality control based on one-to-many Eigenvoice Conversion," *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 101–106, Bonn, Germany, August 2007.

10. D. Tani, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "An Evaluation of many-to-one voice conversion algorithms with pre-stored speaker data sets," *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 101–106, Bonn, Germany, August 2007.

11. T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Proc. Interspeech 2008 - ICSLP*, pp.1076–1079, Brisbane, Australia, September 2008.

12. D. Tani, T. Toda, Y. Ohtani, H. Saruwatari and K. Shikano, "Maximum a posteriori adaptation for many-to-one eigenvoice conversion," *Proc. Interspeech 2008 - ICSLP*, pp.1461–1464, Brisbane, Australia, September 2008.

13. M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," *Proc. Interspeech 2009 - Eurospeech*, pp. 1635–1638, Brighton, U.K., September 2009.

14. S. Kawamoto, Y Adachi, Y. Ohtani, T. Yotsukura, S. Morishima and S. Nakamura, "Scenario speech assignment technique for instant casting movie system", *ACCV2009 Invited workshop on Vision Based Human Modeling and Synthesis*, Xi'an, China, September 2009.

Technical reports

1. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Speaker adaptive training for voice conversion based on eigenvoice," *IEICE Technical Report*, SP2006-40, pp. 31–36, August 2006 (in Japanese).

2. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Many-to-many eigenvoice conversion algorithms with a reference speaker," *IEICE Technical Report*, SP2008-140, pp. 85–90, January 2009 (in Japanese).

3. T. Toda, Y. Ohtani and K. Shikano, "A voice conversion algorithm based on eigenvoice," *IEICE Technical Report*, SP2006-39, pp. 25–30, August 2006 (in Japanese).

4. D. Tani, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Many-to-one voice conversion algorithms with pre-stored speaker data sets," *IEICE Technical Report*, SP2007-81, pp. 61–66, October 2007 (in Japanese).

5. K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Evaluation of voice quality control based on one-to-many eigenvoice conversion," *IEICE Technical Report*, SP2007-82, no. 282, pp. 67–72, October 2007 (in Japanese).

6. T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Low-delay voice conversion algorithm based on maximum likelihood estimation of spectral parameter trajectory," *IEICE Technical Report*, SP2008-141, pp. 91–96, January 2009 (in Japanese).

7. T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Diagonalizing covariance matrices for reducing computation cost of voice conversion based on Gaussian mixture model," *ISPJ SIG Notes*, 2008-SLP-75, pp. 33–38, February 2009 (in Japanese).

Meetings

1. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Proc. 2006 Spring Meeting of Acoustic Society of Japan*, 1-4-11, pp. 233-234, Tokyo, Japan, March 2006 (in Japanese).

2. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Applying speaker adaptive training to voice conversion based on eigenvoice," *Proc. 2006 Autumn Meeting of Acoustic Society of Japan*, 1-6-14, pp. 181–182, Ishikawa, Japan, September 2006 (in Japanese).

3. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Voice conversion based on eigenvoices considering source features and global variance," *Proc. 2007 Spring Meeting of Acoustic Society of Japan*, 1-8-12, pp. 215–216, Tokyo, Japan, March 2007 (in Japanese).

4. Y. Ohtani, S. Kawamoto, T. Toda, S. Nakamura, and K. Shikano, "Specific speech generation based on STRAIGHT morphing", *Proc. 2008 Spring Meeting of Acoustic Society of Japan*, 1-11-29, pp. 309–310, Chiba, Japan, March 2009 (in Japanese).

5. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "A study of the initial models in speaker adaptive training for eigenvoice conversion," *Proc. 2008 Autumn Meeting of Acoustic Society of Japan*, 2-P-23, pp. 409–410, Fukuoka, Japan, September 2008 (in Japanese).

6. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Many-to-many voice conversion based on eigenvoices with a reference voice," *Proc. 2009 Autumn Meeting of Acoustic Society of Japan*, 2-1-1, pp. 285–286, Fukushima, Japan, September 2009 (in Japanese).

7. Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Canonical model training using non-parallel data sets for many-to-many eigenvoice conversion," *Proc. 2010 Spring Meeting of Acoustic Society of Japan*, 1-7-17, pp. 317–318, Tokyo, Japan, March 2010 (in Japanese).

8. T. Toda, Y. Ohtani and K. Shikano, "A voice conversion/control algorithm based on eigenvoice," *Proc. 2006 Autumn Meeting of Acoustic Society of Japan*, 1-6-13, pp.179–180, Ishikawa, Japan, September 2006 (in Japanese)

9. K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Preliminary Evaluation of voice quality control based on one-to-many eigenvoice conversion," *Proc. 2007 Autumn Meeting of Acoustic Society of Japan*, 1-4-13, pp. 317–318, Yamanashi, Japan, September 2007 (in Japanese).

10. D. Tani, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "An evaluation of many-to-one voice conversion algorithms based on speaker selection and eigenvoice," *Proc. 2007 Autumn Meeting of Acoustic Society of Japan*, 1-4-14, pp. 318–319, Yamanashi, Japan, September 2007 (in Japanese).

11. K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Enhanced voice quality control method based on one-to-many eigenvoice conversion," *Proc. 2008 Spring Meeting of Acoustic Society of Japan*, 2-11-5, pp. 345–346, Chiba, Japan, March 2008, (in Japanese).

12. D. Tani, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Many-to-one eigenvoice conversion method rubust to the amount of adaptation data,"

Proc. 2008 Spring Meeting of Acoustic Society of Japan, 3-Q-11, pp. 397–398, Chiba, Japan, March 2009 (in Japanese).

13. T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Proc. 2008 Autumn Meeting of Acoustic Society of Japan*, 3-4-9, pp. 299–300, Fukuoka, Japan, September 2008 (in Japanese).

14. T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "Diagonalizing covariance matrices for reducing computation cost of voice conversion based on Gaussian mixture model," *Proc. 2009 Spring Meeting of Acoustic Society of Japan*, 1-6-10, pp. 309–310, Tokyo, Japan, March 2008 (in Japanese).

15. C. Hayashida, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "An evaluation of many-to-one voice conversion with linear regression-based adaptation," *Proc. 2009 Autumn Meeting of Acoustic Society of Japan*, 1-2-13, pp. 251–252, Fukushima, Japan, September 2009 (in Japanese).

16. C. Hayashida, Y. Ohtani, T. Toda, H. Saruwatari and K. Shikano, "An evaluation of method of model adaptation with linear regression for many-to-one voice conversion," *Proc. 2010 Spring Meeting of Acoustic Society of Japan*, Tokyo, Japan, 1-7-18, pp. 319–320, March 2010 (in Japanese).

Master's thesis

1. Y. Ohtani, "High quality one-to-many voice conversion with mixed excitation and eigenvoices," *Master's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology*, NAIST-IS-MT0551027, Februray 2007 (in Japanese).