NAIST-IS-DD1061014

Doctoral Dissertation

Augmented speech production beyond physical constraints using statistical voice conversion –Alaryngeal speech enhancement and singing voice quality control –

Hironori Doi

March 31, 2013

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

A Doctoral Dissertation
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

Hironori Doi

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Kiyohiro Shikano	(Co-supervisor)
Associate Professor Tomoki Toda	(Co-supervisor)
Dr. Masataka Goto	(AIST)

Augmented speech production beyond physical constraints using statistical voice conversion –Alaryngeal speech enhancement and singing voice quality control – *

Hironori Doi

Abstract

Speech is used as one of the principal means of communication between people. Moreover, singing voices, which are produced by adding rhythm and tone to speech, are widely used by many people as one of the musical instruments. Although people can control their voice quality to some degree to produce speech sound expressively, the varieties of voice quality that can be produced by individuals are limited owing to physical constraints in the speech production mechanism.

This limitation of voice quality caused by physical constraints cannot be overcome by the effort and skill of the speaker, which can cause several problems. In this thesis, two of those problems are focused on. One problem is vocal disorders caused by a total laryngectomy, which is a surgical operation to remove the larynx including the vocal folds owing to laryngeal cancer. People who have undergone a total laryngectomy, called laryngectomees, must produce alaryngeal speech by alternative speaking methods using residual organs or medical devices instead of vocal fold vibration to generate excitation sounds. Although alaryngeal speech allows laryngectomees to speak, they suffer from unnatural sound and lack of speaker individuality. The other problem is limited expression in singing. Although, singers can change the voice quality of their singing voice to sing expressively, they cannot sing with a voice quality beyond their physical constraints.

* Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1061014, March 31, 2013.

It is desirable to achieve new singing styles to produce more expressive singing voices beyond the naturally achievable varieties.

To address these problems, new alaryngeal speech enhancement and singing voice quality control methods based on statistical voice conversion (VC) are proposed as techniques to augment speech production beyond physical constraints in this thesis. VC is a technique that is capable of converting a speaker's voice into another speaker's voice using Gaussian mixture model (GMM) that models the relationship between the acoustic features of the source and target speaker's voices. Moreover, the eigenvoice conversion (EVC) technique, which allows us to convert an arbitrary speaker's voice into another arbitrary speaker's voice, has been proposed recently. In EVC, an eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets consisting of speech uttered by a single predefined speaker, called the reference speaker, and many prestored target speakers. A conversion model that converts the source speaker's voice into the target speaker's voice can be easily constructed by the adaptation of an EV-GMM using a few of their given voice samples in text-independent manner.

The proposed alaryngeal speech enhancement method based on VC, called AL-to-Speech in this thesis, allows laryngectomees to produce naturally sounding speech by converting alaryngeal speech into normal speech uttered by non-laryngectomees. Moreover, in this thesis one-to-many EVC, which is one of the EVC techniques, is incorporated into AL-to-Speech to make it possible for laryngectomees to speak with their desired voice quality. In particular, the enhancement method for three types of alaryngeal speech, i.e., esophageal speech, electrolaryngeal speech, and silent electrolaryngeal speech, is focused on. The experimental results demonstrate that our proposed method recovers the sound quality and speaker individuality of alaryngeal speech.

The proposed singing voice conversion method allows singers to freely control their voice quality by converting their singing voice into an arbitrary singer's singing voice using EVC techniques. Furthermore, to easily develop multiple parallel data sets from the nonparallel singing voice data sets of many singers, a technique for efficiently and effectively generating parallel data sets using a singing-to-singing synthesis system is proposed to artificially generate voices of the reference speaker, called the reference singer in singing voice conversion. The

experimental results demonstrate that the proposed method allows a singer to sing with voice quality of an arbitrary singer.

Keywords:

statistical voice conversion, eigenvoice conversion, laryngectomees, alaryngeal speech, singing voice, singing-to-singing synthesis

統計的声質変換を用いた身体的制約を超えた音声生成
–無喉頭音声強調及び歌声の声質制御–*

土井　啓成

内容梗概

音声は，言語情報だけでなく感情や話者性といったパラ言語/非言語情報も相手にリアルタイムに伝えることが可能である．また，音声をリズムや音楽にのせることで，音声は歌声となり，曲を構成する一つの要素として扱われる．しかしながら，声の質，すなわち声質は，話者ないし歌手の身体的特徴によるところが大きく，人は，自身の身体的制約を超えて，他者の声質で話すことや歌うことはできない．このことは，いくつかの問題を引き起こす原因となる．

本論文では，声質表現の限界により生じる二つの問題に着目する．一つ目の問題は，全喉頭摘出手術を受けたことによる発声障害である．喉頭癌等の切除のために声帯を含む喉頭領域全体を摘出した喉頭摘出者は通常の発声法が行えなくなり，代用発声法により生成される無喉頭音声で音声コミュニケーションを行う．しかしながら，無喉頭音声の音質は総じて低く，話者の判別も付かない．これは，「声帯が無い」という身体的特徴によって，健常者とは比べ物にならない程の大きな制約を受けるためである．二つ目の問題は，歌声における表現の限界である．歌声は，最も多くの人々に親しまれている楽器の一つと言えるが，人一人が扱える声質（音色）には限界がある．これは，音楽表現に制限をかける大きな要因の一つと言える．

本論文では，無喉頭音声の強調や歌声の声質を操作可能にすることで，それらの音声にかかる制約・制限を排し，新たな発声や表現を取得することを目的とする．目的の達成のため，本論文では統計的声質変換技術を用いる．統計的声質変換とは，ある話者の声を別の話者の声へと統計モデルに基づき変換する技術である．また近年では，この統計的声質変換の枠組みを拡張した固有声変換も提案

* 奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DD1061014, 2013年3月31日.

されている．固有声変換では，少量の音声サンプルを用いて，変換モデルを元話者ないし目標話者に適応させることが可能であるため，変換モデルの構築が容易に行える．ただし，そのための事前準備として，参照話者と呼ばれる特定の一人の話者と多数の事前収録目標話者との間で同一内容発話を用意する必要がある．本論文では，この統計的声質変換及び固有声変換を応用し，無喉頭音声強調及び，歌声の声質制御を実現する．

無喉頭音声強調では，統計的声質変換を用いて，無喉頭音声を健常者の音声へと変換することで，その音質の改善を試みる．さらに，本論文では，上記の統計的無喉頭音声強調法に，固有声変換の技術を導入することで，強調音声の声質を制御可能にし，喉頭摘出者が自身の望む声で発声できる枠組みを提案する．提案法では，喉頭摘出者の喉頭摘出以前の音声が保持されていれば，かつての音声と類似した声質を持つ音声を取得することが可能である．提案法を，食道音声，電気音声，微弱電気音声の 3 つの無喉頭音声に導入し，総合的に比較評価を行う．実験結果から，本手法が各無喉頭音声の自然性改善に有効であること，また，話者性の回復が可能であることを示す．

歌声の声質制御では，固有声変換を用いて，利用者の歌声を任意の別の歌手の歌声へと変換することで，歌声における表現の拡張を試みる．この固有声変換に基づく歌声声質変換は，ユーザの歌唱技法を変化させることなく，変換音声の声質のみを目標歌手の声質へと変換するため，利用者の歌い回し等の技術による表現を損なうことなく，声質による表現のみを拡張することが可能である．さらに，学習データ収集を容易にするため，参照歌手（話声における参照話者に相当）の歌声として，singing-to-singing synthesis により生成される合成音声を用いる手法を提案する．singing-to-singing synthesis により，目標歌手の歌声から，自動的に特定の声質を持つ歌声が生成されるため，システム構築者は，多数の目標歌手の歌声を収集するだけで本システムを構築することが可能になり，システム構築における負担が劇的に軽減される．実験結果から，提案法によって，利用者の歌声の声質を，容易に任意の歌手の声質へと変換可能になることを示す．

キーワード

統計的声質変換, 固有声変換, 喉頭摘出者, 無喉頭音声, 歌声, singing-to-singing synthesis

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background and problem definition

People communicate with each other by various methods such as speech, songs, letters, facial expressions, and so forth. Among them, speech is one of the simplest and most popular methods. Moreover, singing voices, which are produced by adding rhythm and tone to speech, are widely used by many people as a means of producing music. Speech and singing voices usually convey not only linguistic information but also emotion, speaking style, and speaker individuality in real time. The information that is conveyed by speech can be categorized in three types of information [1]. One is linguistic information, which is symbolic information representing words, text, and context. The next is paralinguistic information, which is not inferable from the written text. Paralinguistic information is information that the speaker can consciously control, such as emotions and speaking styles. Speakers can deliberately add this information into an utterance to express different intentions and attitudes. The other types of information is nonlinguistic information, which concerns factors such as the physical characteristics, physical condition, and emotional state of the speaker. These factors cannot generally be controlled by the speaker. Thus, nonlinguistic information is information that is conveyed regardless of the intentions of the speaker, such as speaker individuality.

Timbre and voice quality, which are feature of voice, play an important role in conveying paralinguistic information and nonlinguistic information. The American National Standards Institute (ANSI) defines timbre as "that attribute of auditory sensation in terms of which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different." In other words, timbre is defined as everything that is not loudness, pitch, or spatial perception. Although the voice quality usually represents timbre, all features that capture the short-term characteristics conveying paralinguistic and nonlinguistic information except for the pitch is called voice quality in this thesis. Figure 1 shows an image explaining voice quality and paralinguistic and nonlinguistic information. People can expressively control their voice quality, pitch, and speaking style to some degree to vary the paralinguistic information conveyed. Moreover,

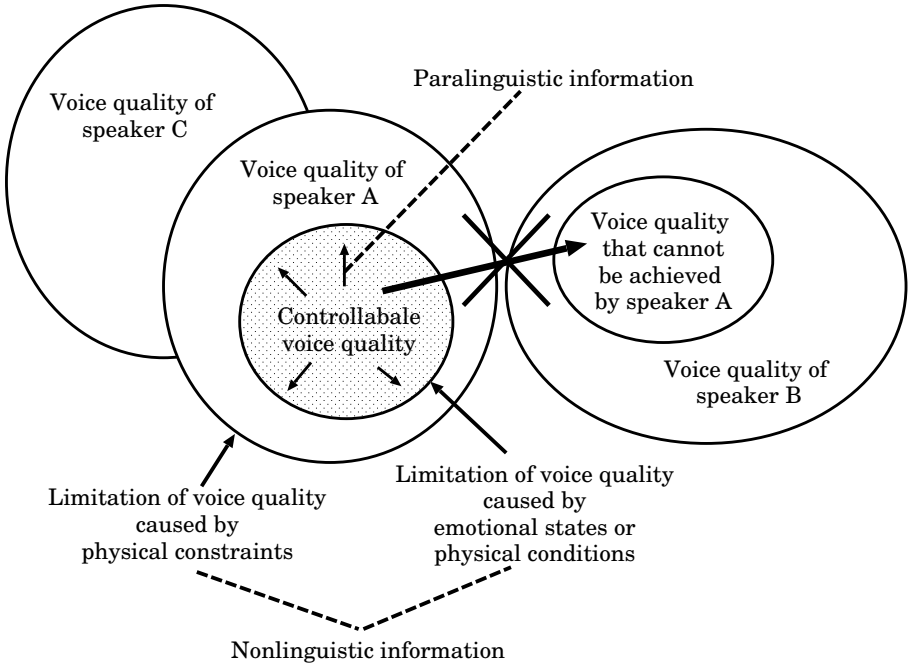

Figure 1. Schematic image explaining voice quality and paralinguistic and nonlinguistic information.

actors and showmen often imitate another person's speech very well by controlling these characteristics. Although the pitch and speaking style of another person can be imitated by a speaker, the voice quality cannot. The varieties of voice quality that can be produced by an individual person are limited owing to physical characteristics of the speaker.

The voice quality is limited by physical constraints that cannot be overcome by the effort and skill of the speaker, which can cause several problems. The aim of this thesis is to overcome such physical constraints to make it possible for people to speak with voice quality that cannot be achieved by natural speech production. The concept of this research is augmentation of the expressiveness of speech beyond physical constraints. In this thesis, augmented speech production means speech production that increase ability of speaker beyond physical constraints. Therefore, this thesis basically considers only information that cannot be controlled by a speaker. As a technique that allows us to speak with a voice quality overcoming physical constraints, statistical voice conversion (VC)

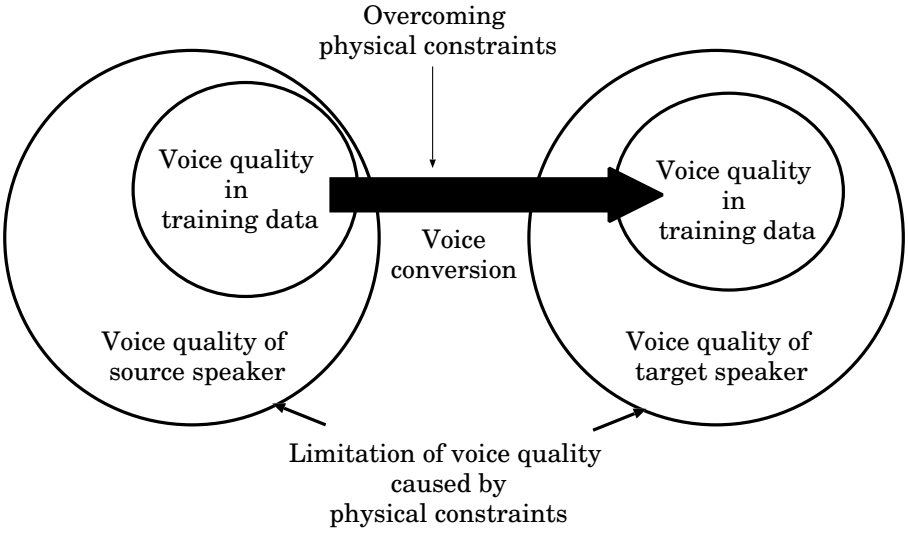

Figure 2. Voice conversion.

has been proposed [2]. Figure 2 shows the concept of VC. Statistical VC is a technique that converts the voice quality of a source speaker into that of another speaker, called the target speaker, while keeping the linguistic information unchanged. Recently, a statistical method based on a Gaussian mixture model (GMM) [3], which is employed as a conversion model, has been used widely in the VC framework. In this technique, a GMM of the joint probability density of an acoustic feature between the source speaker's voice and the target speaker's voice is trained in advance using a special data set, called *parallel data*, which consists of utterance pairs of these two speakers. The trained model is capable of converting the acoustic features of the source speaker's voice into those of the target speaker's voice in any utterance while keeping the linguistic information unchanged. In statistical VC, because the converted acoustic feature is generated from statistics calculated from the acoustic feature of the target speaker, this probabilistic conversion process allows us to produce speech beyond physical constraints that is difficult to achieve by simple modification. Moreover, in statistical VC, because the controllable information of speaker, such as the linguistic information and speaking style, are basically not converted, speaker's intentions are directly reflected in the converted speech. Therefore, statistical VC does not constrain the expression of the speaker. Moreover, low-delay VC, which is capa-

ble of converting voice quality in real time and makes it possible to use VC in man-to-man speech communication, has been proposed [4].

Although statistical VC is capable of generating converted speech with similar voice quality to that of the target speaker, it is difficult to freely control the voice quality of converted speech because a large amount of parallel data is needed to train the conversion model. As a technique that enables us to freely control the voice quality of the converted speech and is capable of rapidly adapting the conversion model to an arbitrary speaker, eigenvoice conversion (EVC) has been proposed [5], which is an effective approach that uses the voices of other speakers as prior knowledge. In this technique, as the conversion model, an eigenvoice GMM (EV-GMM) [5] is trained in advance using multiple parallel data sets that consist of a single predefined speaker, called a reference speaker, and many pre-stored target speakers. The EV-GMM can be easily adapted to the voice quality of the source, target, or both using a few of their given voice samples in a text-independent manner. Additionally, because the voice quality of the converted speech can be controlled using a few parameters, EVC makes it possible to intentionally control speaker individuality like paralinguistic information.

This thesis deals with two problems caused by physical constraints by using statistical VC and EVC techniques. One problem is vocal disorders caused by a total laryngectomy, which is a surgical operation to remove the larynx including the vocal folds owing to laryngeal cancer. The other problem is limited expression in singing.

1.1.1 Problems of alaryngeal speech

People who have undergone a total laryngectomy, called laryngectomees, cannot speak in the same manner as non-laryngectomees because they cannot generate excitation sounds using vocal fold vibration. Therefore, they use several alternative speaking methods for speech communication, which generate excitation sounds using residual organs or a medical device instead of vocal folds, so that they can reintegrate into their individual, social, and regular activities. Speech produced by an alternative speaking method without vocal fold vibration is called alaryngeal speech.

There are various alternative speaking methods such as the esophageal speak-

ing method (ES) and electrolaryngeal speaking method (EL). ES is one of the major alternative speaking methods that generate esophageal speech (ES speech) using residual organs. In ES, laryngectomees generate excitation sounds by releasing gas from or through the esophagus by articulatory movement. Although ES allows laryngectomees to speak without any equipment, ES speech has less intelligibility than normal speech uttered by non-laryngectomees. EL is the most popular alternative speaking method among those using medical devices. Alternative excitation sounds are produced using an electrolarynx, which is a medical device that mechanically generates sound source signals. It is much easier to learn how to speak using the electrolarynx than to learn how to produce ES speech. However, the EL speech sounds mechanical and artificial because the generated fundamental frequency (F_0) contour is fixed. Additionally, because the electrolarynx must generate sufficiently loud sound source signals to make the produced EL speech sufficiently audible, the sound source signals are readily emitted outside, disturbing speech communication. To resolve the issue of emitted sound source signals in the speaking method for EL speech, a new speaking method for silent EL speech has been proposed [6]. A new sound source unit is used to generate less audible sound source signals. Since the produced speech also becomes less audible, it is detected using a nonaudible murmur (NAM) microphone [7], which is a body-conductive microphone capable of detecting extremely soft speech from the neck below the ear. The detected speech signals are presented outside as silent EL speech while keeping the external sound source signals sufficiently silent.

Although these three types of alaryngeal speech allow laryngectomees to speak again, their sound quality and intelligibility are severely degraded compared with those of normal speech. Moreover, alaryngeal speech sounds have a similar voice quality regardless of the speaker because the production mechanism of the sound source signals in each type of alaryngeal speech strongly affects the voice quality of the proposed speech. These problems are caused by the production mechanism of alaryngeal speech, i.e., physical constraints. Therefore, to alleviate these problems, a technique that allows laryngectomees to produce speech sounds beyond the physical constraints is needed.

1.1.2 Problems of singing voice

Singing voices are different from the sounds of musical instruments because they can convey linguistic information in the lyrics as well as the pitch, dynamics, and voice quality. Singing is highly familiar to not only professional singers but also many people regardless of age and gender. Moreover, it is considered as an artform and as entertainment for singers and listeners. Although, singers can change the voice quality of their singing voice to sing expressively, they cannot sing with a voice quality beyond their physical constraints. This limitation of the voice quality of singing voice seriously limits possibility of expression in music. Thus, it is desirable to achieve new singing styles to produce more expressive singing voices beyond the naturally achievable varieties.

Various singing synthesis systems that allow a user to generate a singing voice with voice quality different from that of the user, have been proposed. Several singing synthesis systems can artificially generate humanlike singing voices with different voice quality by changing the singing synthesis parameters. However, it is still difficult to generate singing voices with arbitrary and desired voice qualities. Moreover, because singing synthesis systems generally do not work in real time, singers cannot use them in singing.

1.2 Scope of thesis

This thesis describes two types of augmented speech production based on statistical voice conversion. One is AL-to-Speech, which is a method of enhancing alaryngeal speech. The other is singing voice quality control using singing voice conversion. Figure 3 and 4 show the concepts of AL-to-Speech and singing voice quality control, respectively.

1.2.1 Enhancement of alaryngeal speech

This thesis describes an enhancement method for three types of alaryngeal speech, i.e., ES speech, EL speech, and silent EL speech. Recently, to enhance EL speech and silent EL speech, EL-to-Speech and silent EL-to-Speech have been proposed. These methods, which convert EL or silent EL speech into normal speech uttered by non-laryngectomees based on statistical VC, are effective for enhancing these

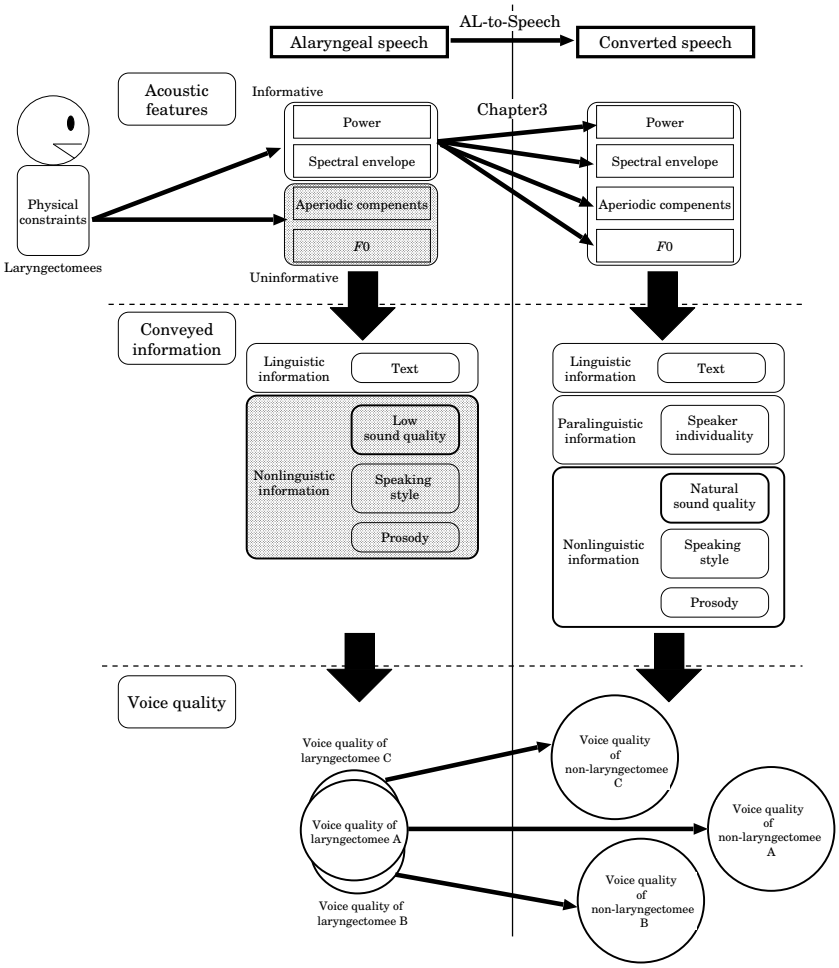

Figure 3. Concept of AL-to-Speech.

types of alaryngeal speech. Using this method, EL speech and silent EL speech obtain similar naturalness to normal speech.

This statistical approach is employed to enhance ES speech in this thesis. Thus, ES speech is converted into normal speech using statistical VC framework. This method is called ES-to-Speech in this thesis. In these three types of enhancement method, called alaryngeal speech-to-speech (AL-to-Speech), because converted speech is generated from the statistics of normal speech, the low naturalness and intelligibility caused by alaryngeal speech production are fundamentally not included in the voice quality of converted speech. Therefore, AL-to-Speech allows laryngectomees to speak with a natural voice quality similar

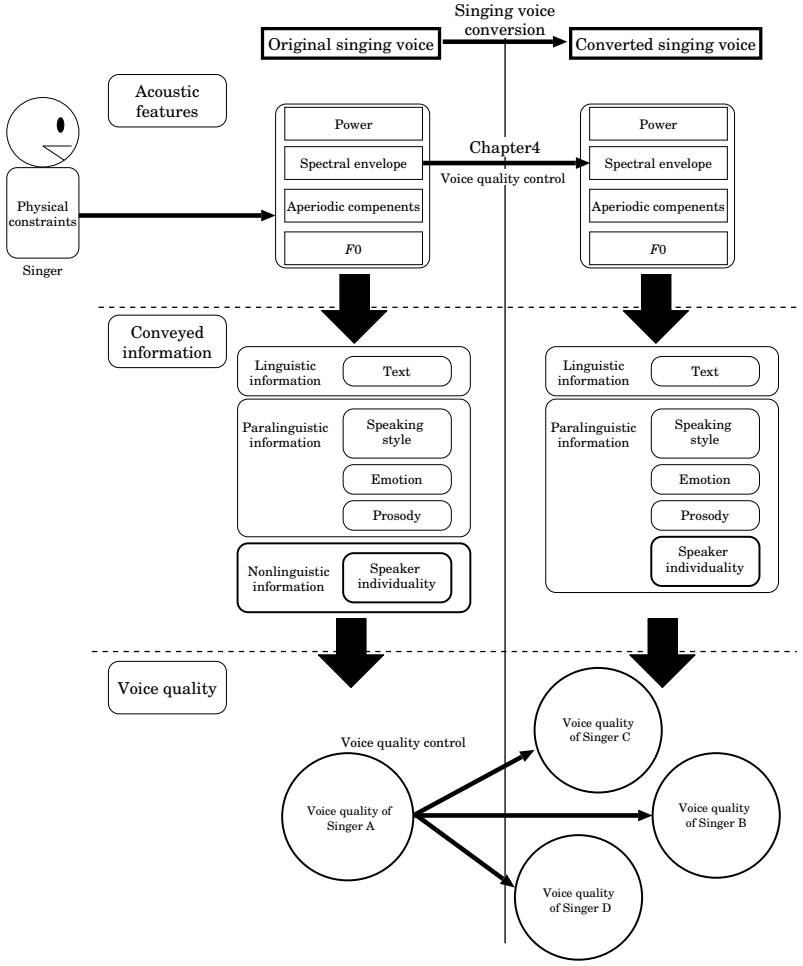

Figure 4. Concept of singing voice quality control.

to that of normal speech. However, these methods cannot recover speaker individuality because the voice quality of the converted speech is fixed and depends on the training data.

In this thesis, to recover the speaker individuality of alaryngeal speech, one-to-many EVC, which is capable of flexibly adapting the conversion model to a given target natural voice, is applied to AL-to-Speech. Because one-to-many EVC is capable of controlling the voice quality of converted speech by adapting the conversion model to the target speaker, laryngectomees can recover their original voice quality provided from speech samples uttered before undergoing a total laryngectomy if they are available. Even if such speech samples do not exist, the

voice quality desired by the user can be created by manipulating the parameter of EV-GMM. In this thesis, the spectral feature and power of alaryngeal speech are converted into the acoustic features of normal speech using EVC techniques because spectral feature and power are informative among the acoustic features of alaryngeal speech as shown Fig. 3. The results of objective and subjective evaluations demonstrate that the proposed methods yield significant improvements of the speech quality and producing a converted voice with voice quality similar to that of target voice.

1.2.2 Singing voice quality control

This thesis describes a control method of singing voice quality. To make it possible for people to directly sing with a different specific voice quality, and thus overcome physical constraints, singing voice conversion has been proposed [8]. Statistical VC techniques are used to convert the singing voice quality of a source singer into that of a target singer. The trained model is capable of converting the acoustic features of the source singer's singing voice into those of the target singer's singing voice in any song while keeping the linguistic information of the lyrics unchanged. Additionally, because this VC-based system can work in real time by applying the low-delay conversion algorithm, the singer can use this system while singing.

Towards realizing a more flexible singing-voice conversion technique that enables singers to freely control the voice quality of converted singing voice and is capable of rapidly adapting the conversion model to arbitrary singers, a singing-voice conversion method based on two techniques is proposed in this thesis: singing voice conversion based on many-to-many EVC [9] and training data generation using a singing-to-singing synthesis system [10]. The many-to-many EVC is a technique of conversion from the voice of an arbitrary source singer into the voice of an arbitrary target singer. An EV-GMM is capable of easily adapting the source/target voice quality to that of a few of the given singing voice samples in a text-independent (lyrics-independent) manner. Moreover, the EVC technique allows singers to create a target voice quality by manipulating a few parameters of the EV-GMM. By controlling the voice quality of converted singing voice while singing, the singer can produce a new singing style which is movement of voice quality. In this thesis, to control the voice quality, only the spectral feature is

converted into those of target singer, as shown in Fig. 4 because the voice quality mainly depends on the spectral feature. The proposed system makes it possible for singers to control information that cannot be controlled by their own effort. Therefore, acoustic features that convey information controllable by the singer are not converted to maintain the singer's expression. Furthermore, to easily develop multiple parallel data sets from nonparallel singing voice data sets of many singers, a technique is proposed for efficiently and effectively generating parallel data sets using a singing-to-singing synthesis system called *VocaListener* [10] to artificially generate voices of the reference singer. The results of objective and subjective evaluations demonstrate that the proposed methods make it possible for a singer to freely control the voice quality.

1.3 Overview of thesis

This thesis is organized as follows. In Chapter 2, VC frameworks based on the statistical approach are described as well as state-of-the-art conversion methods for the VC frameworks. We also describe the EVC frameworks.

In Chapter 3, we address the enhancement methods for alaryngeal speech. First, we describe detail of laryngectomees and alaryngeal speech. In this chapter, several conventional speaking-aid systems for ES, EL, and silent EL are described. Then, we describe our proposed speaking-aid system based on statistical VC for ES speech, called ES-to-Speech. Moreover, we also describe an enhancement method based on one-to-many EVC for ES, EL, and silent EL, called AL-to-Speech. The effectiveness of ES-to-Speech is demonstrated in detail by objective and subjective evaluations. AL-to-Speech is also experimentally evaluated for the three types of alaryngeal speech.

In Chapter 4, we address singing voice quality control based on the EVC frameworks as a technique that allows us to overcome the limitation of a singer's voice quality. As the conventional method, three types of singing synthesis, i.e., text-to-singing synthesis, speech-to-singing synthesis, singing-to-singing synthesis are described, along with a conversion approach based on statistical VC. To make it possible for singers to freely control the voice quality of their singing voice, we apply many-to-many EVC, which is a technique of converting from the voice of an arbitrary source singer into the voice of an arbitrary target singer, to singing

voice conversion. Moreover, training data generation using singing-to-singing synthesis, which is one of the singing synthesis systems, is proposed to reduce the burden of training data generation. The effectiveness of these proposed methods is demonstrated by objective and subjective evaluations.

In Chapter 5, we summarize the contributions of this thesis and suggest future work.

2. Statistical voice conversion

This chapter describes the voice conversion (VC) and eigenvoice conversion (EVC) techniques based on a GMM used in this thesis. VC based on GMM is a technique that is capable of converting an arbitrary utterance of a source speaker into that of a target speaker with the GMM trained using parallel data that consists of a source speaker's voice and a target speaker's voice. Although it is well defined mathematically and its conversion performance is relatively high, it is difficult to freely control the converted voice quality owing to the use of a large amount of parallel data in the training process. To freely control the converted voice quality, EVC has been proposed. In the EVC technique, an eigenvoice GMM (EV-GMM), which is used as an initial model, is trained using multiple parallel data sets consisting of a single speaker's voice and many prestored speaker's voice. The conversion model between the source speaker's and target speaker's voices is constructed by adapting the EV-GMM to the source and target speakers with weight parameters estimated from a few speech samples of the source and target speaker.

2.1 Introduction

Speech has several important characteristics such as linguistic information, emotion, and speaker individuality. VC is a technique that converts a source speaker's voice quality including nonlinguistic information into those of another speaker while keeping the linguistic information unchanged. This technique is useful for many applications such as voice responses, text reader systems, and so forth.

Among the VC techniques, two main approaches have been studied. One is a rule-based approach [11, 12, 13] that directly modifies acoustic features such as the spectral envelope and fundamental frequency (F_0). In this approach, the performance of the conversion depends on the developers' abilities because the rules for modifying acoustic features are determined by developers. The other is a statistical approach that converts voice quality using a statistical model constructed from a large amount of speech data.

Many statistical approaches to VC have been studied. In an early statistical VC approach, the codebook mapping method based on hard clustering and dis-

crete mapping was proposed [2]. In this method, a converted acoustic feature is determined by quantizing the source speaker's acoustic feature to the nearest centroid feature of the source codebook and substituting it with the corresponding centroid feature of the mapping codebook. In recent years, statistical VC using a GMM has been widely used. For typical statistical VC, two GMM-based conversion methods have been proposed. One is a frame-based conversion method that converts features based on the minimum mean square error (MMSE) [3]. The other is a trajectory-based conversion method [14] that simultaneously converts a feature sequence over an utterance based on maximum likelihood estimation (MLE) [15]. Although the former method is capable of real-time conversion because source features at individual frames are converted independently of each other, this method sometimes causes feature fluctuation with inappropriate dynamic characteristics because the interframe feature correlation is ignored. On the other hand, the latter method provides converted feature sequences exhibiting appropriate dynamic characteristics by considering the dynamic features of the converted speech. Additionally, to alleviate the oversmoothing of the converted features due to statistical modeling, trajectory-based VC considering the global variance (GV), which is the variance of features over a time sequence, has also been proposed [14]. Although the trajectory-based conversion method results in significant quality improvements of the converted speech, it does not work in real time because the source features over an utterance need to be converted simultaneously to consider the interframe correlation. To achieve real-time conversion considering the dynamic characteristics of converted features, low-delay VC based on MLE has been proposed [4].

In the statistical VC approaches, it is possible to change the converted voice quality using different target voices but it is necessary to prepare training data consisting of utterance pairs of the source and target speakers, which is very laborious. Towards realizing a more flexible VC technique that enables a speaker to freely control the converted voice quality and is capable of rapidly adapting the conversion model to arbitrary speakers, EVC has been proposed [9]. EVC has been used in the development of several VC frameworks, i.e., one-to-many EVC, many-to-one EVC, and many-to-many EVC [16]. One-to-many, many-to-one, and many-to-many EVC allow us to adapt a conversion model to an arbi-

trary target, an arbitrary source, or to both speakers, respectively. In each EVC method, a conversion model is adapted using a weight parameter capturing voice quality, which is automatically estimated by maximum likelihood eigendecomposition (MLED) [17] using a few source or target speech samples. The performance of this adaptation is significantly improved by applying MAP [18] adaptation to the unsupervised weight estimation. Moreover, the weight parameter can be manipulated to freely control the voice quality of the converted speech.

This section is organized as follows. In Section 2.2, the framework of trajectory-based VC used in this thesis is explained. In Section 2.3, the framework of EVC is introduced. Finally, this chapter is summarized in Section 2.4.

2.2 Voice conversion based on GMM

This section describes a conversion method based on the MLE of speech parameter trajectories considering GV [14] as one of the state-of-the-art VC methods. This method converts acoustic features, such as the spectrum, F_0, and aperiodic components of the source speaker's voice, into those of the target speaker. This method consists of a training process and a conversion process.

2.2.1 Training process

In the training process, first, a large number of utterance pairs, consisting of the same sentence spoken by the source and target speakers, are recorded, then acoustic features are extracted from each utterance. Let us assume a source static feature vector $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D_x)]^\top$ and a target static feature vector $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D_y)]^\top$ at frame t, where $\top$ denotes the transposition of the vector. As an input speech parameter vector, $\boldsymbol{X}_t$ is used to capture contextual features of the source speech, for example, the joint feature vector of static and dynamic feature vectors or the concatenated feature vector from multiple frames. As an output speech feature vector, $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, consisting of static and dynamic features, is used. The dynamic feature vector at frame t is calculated from the static feature vectors as follows:

$$\Delta\boldsymbol{y}_t = -0.5\boldsymbol{y}_{t-1} + 0.5\boldsymbol{y}_{t+1}. \tag{1}$$

Using a parallel training data set consisting of time-aligned input and output parameter vectors $[\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top]^\top$, $[\boldsymbol{X}_2^\top, \boldsymbol{Y}_2^\top]^\top$, $\cdots$, $[\boldsymbol{X}_T^\top, \boldsymbol{Y}_T^\top]^\top$ determined by dynamic time warping (DTW), where T denotes the total number of frames, the joint probability density of the input and output parameter vectors is modeled by a GMM [15] as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right), \tag{2}$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \tag{3}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. A parameter set of the GMM is denoted by λ, which consists of weights α_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$, and covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components. The mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ consists of an input mean vector $\boldsymbol{\mu}_m^{(X)}$ and an output mean vector $\boldsymbol{\mu}_m^{(Y)}$. The covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ consists of input and output covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ and cross-covariance matrices $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$.

The GV is defined as the variance of features over one utterance. To consider the GV in the conversion, the probability density of the GV $\boldsymbol{v}(\boldsymbol{y})$ of the output static feature vectors $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ over an utterance is also modeled by a Gaussian distribution,

$$P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}), \tag{4}$$

where $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(D_y)]^\top$ is calculated by

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2. \tag{5}$$

A parameter set $\lambda^{(v)}$ consists of a mean vector $\boldsymbol{\mu}^{(v)}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}^{(v)}$.

2.2.2 Conversion process

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ be time sequences of the input and the output feature vectors, respectively. The con-

verted static feature vector sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined by maximizing the following objective function:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \lambda) \\
&= \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{all\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \lambda) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \lambda) \\
&= \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \lambda) P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda), \qquad (6) \\
&\text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \qquad (7)
\end{aligned}$$

where $\boldsymbol{W}$ is a window matrix used to extend the static feature vector sequence to the joint feature vector sequence of static and dynamic features [19]. In this thesis, we use

$$\boldsymbol{W} = \left[\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \cdots, \boldsymbol{w}_T^\top\right]^\top, \qquad (8)$$

where

$$\boldsymbol{w}_t = \begin{bmatrix} & \boldsymbol{0}_{D\times(t-1)D} & \boldsymbol{I} & \boldsymbol{0}_{D\times(T-t)D} & \\ \boldsymbol{0}_{D\times(t-2)D} & -0.5\boldsymbol{I} & \boldsymbol{0}_{D\times D} & 0.5\boldsymbol{I} & \boldsymbol{0}_{D\times(T-t-1)D}. \end{bmatrix} \qquad (9)$$

The matrix $\boldsymbol{I}$ is a $D \times D$ identity matrix. The m^{th} component weight $P(m|\boldsymbol{X}_t, \lambda)$ and conditional probability density $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda)$ at frame t are given as follows:

$$P(m|\boldsymbol{X}_t, \lambda) = \frac{\alpha_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}, \qquad (10)$$

$$P(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \lambda) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}\right), \qquad (11)$$

where

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right), \qquad (12)$$

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)}. \qquad (13)$$

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

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^{\top} \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^{\top} \overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}}, \tag{14}$$

where

$$\overline{\boldsymbol{D}^{(Y)^{-1}}} = diag \left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}}, \overline{\boldsymbol{D}_2^{(Y)^{-1}}}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}} \right], \tag{15}$$

$$\overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)^{-1}} \boldsymbol{E}_1^{(Y)}}^{\top}, \overline{\boldsymbol{D}_2^{(Y)^{-1}} \boldsymbol{E}_2^{(Y)}}^{\top}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}} \boldsymbol{E}_T^{(Y)}}^{\top} \right]^{\top}, \tag{16}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}}, \tag{17}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}} \boldsymbol{E}_t^{(Y)}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)}, \tag{18}$$

$$\gamma_{m,t} = P(m|\boldsymbol{X}_t, \lambda). \tag{19}$$

Moreover, the likelihood function given by eq. (6) is approximated with the sub-optimum mixture component sequence $\hat{\boldsymbol{m}} = [\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_T]$ given as follows:

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}}\, P(\boldsymbol{m}|\boldsymbol{X}, \lambda). \tag{20}$$

Using this mixture component sequence, the converted static feature sequence is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda). \tag{21}$$

Then, eq. (14) is rewritten as follows:

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^{\top} \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^{\top} \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}, \tag{22}$$

$$\tag{23}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} = \left[\boldsymbol{E}_{m_1,1}^{(Y)}, \boldsymbol{E}_{m_2,2}^{(Y)}, \cdots, \boldsymbol{E}_{m_T,T}^{(Y)} \right], \tag{24}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} = diag \left[\boldsymbol{D}_{m_1}^{(Y)^{-1}}, \boldsymbol{D}_{m_2}^{(Y)^{-1}}, \cdots, \boldsymbol{D}_{m_T}^{(Y)^{-1}} \right]. \tag{25}$$

One essential problem in MLE is that the estimated parameters tend to be oversmoothed. To address this problem, the GV has been introduced into trajectory-based conversion. The converted target static feature sequence $\hat{\boldsymbol{y}}$ is determined by maximizing the following likelihood function:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \lambda)^{\omega} P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)}). \tag{26}$$

The balance between $P(\boldsymbol{Y}|\boldsymbol{X}, \lambda)$ and $P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)})$ is controlled by the weight ω. In this thesis, ω is set to $\frac{1}{2T}$, which is determined by the ratio between the numbers of dimensions in the target acoustic feature sequence and the GV. Then, eq. (26) is approximated with the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ and an auxiliary function as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, L \tag{27}$$

$$L = \omega \log P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda) + \log P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)}) \tag{28}$$

$$\begin{aligned} &= \omega \left(-\frac{1}{2} \boldsymbol{y}^{\top} \boldsymbol{W} \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W} \boldsymbol{y} + \boldsymbol{y}^{\top} \boldsymbol{W} \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} \right) \\ &\quad - \frac{1}{2} \boldsymbol{v}(\boldsymbol{y})^{\top} \boldsymbol{\Sigma}^{(v)^{-1}} \boldsymbol{v}(\boldsymbol{y}) + \boldsymbol{v}(\boldsymbol{y})^{\top} \boldsymbol{\Sigma}^{(v)^{-1}} \boldsymbol{\mu}^{(v)} + K, \end{aligned} \tag{29}$$

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

$$\boldsymbol{y}^{(i+1)^{th}} = \boldsymbol{y}^{(i)^{th}} + \theta \cdot \delta \boldsymbol{y}^{(i)^{th}}, \tag{30}$$

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

$$\delta \boldsymbol{y}^{(i)^{th}} = \frac{\partial L}{\partial \boldsymbol{y}}, \tag{31}$$

$$\begin{aligned} &= \omega \left(-\boldsymbol{W} \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W} \boldsymbol{y} + \boldsymbol{W} \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} \right) \\ &\quad + \left[\acute{\boldsymbol{v}}(\boldsymbol{y})_1^{\top}, \acute{\boldsymbol{v}}(\boldsymbol{y})_2^{\top}, \cdots, \acute{\boldsymbol{v}}(\boldsymbol{y})_T^{\top} \right]^{\top}, \end{aligned} \tag{32}$$

$$\acute{\boldsymbol{v}}(\boldsymbol{y})_t^{\top} = [\acute{v}_t(1), \acute{v}_t(2), \cdots, \acute{v}_t(D)]^{\top}, \tag{33}$$

$$\acute{v}_t(d) = -\frac{2}{T} \boldsymbol{p}_v^{(d)^{\top}} \left(\boldsymbol{v}(\boldsymbol{y}) - \boldsymbol{\mu}^{(v)} \right) (y_t(d) - \overline{y}(d)). \tag{34}$$

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

2.3 Eigenvoice conversion

There are three EVC frameworks: one-to-many EVC, many-to-one EVC, and many-to-many EVC [16]. In this section, we first describe the overall EVC framework by using the one-to-many EVC framework. EVC consists of a training process, an adaptation process, and a conversion process. In the training process, an eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets consisting of the specified source speaker, called the reference speaker, and many prestored target speakers. In the adaptation process, the trained EV-GMM, which is a target-speaker-independent model, 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, which are weight parameters for eigenvoices, are estimated in a completely text-independent manner. Moreover, in this framework, the user can freely control the voice quality of the converted speech by manipulating these weight parameters. In the conversion process, an arbitrary utterance of the source speaker's voice is converted into that of the new target speaker's voice using the adapted EV-GMM, which is a target-speaker-dependent model.

In this framework, the trained EV-GMM, which is a target-speaker-independent model, is strongly affected by acoustic variations among the many prestored target speakers, which have been reported to cause significant degradation of the performance of the adapted EV-GMM. To alleviate this problem, a pseudonormalized speaker model, called the canonical model, based on speaker adaptive training (SAT) [20] has been proposed [16] as an initial model for speaker adaptation. In SAT, the canonical model parameters are estimated by maximizing the total likelihood of adapted models, which are adapted to prestored speakers used for training. The performance of the adapted EV-GMM is significantly improved by using the canonical EV-GMM trained with SAT as the initial model in the adaptation process.

In this section, we also describe many-to-many EVC, which converts an arbitrary source speaker's voice into an arbitrary target speaker's voice. Basically, many-to-many EVC performs a conversion process consisting of two steps. In the first step, an arbitrary speaker's voice is converted into a reference speaker's voice using many-to-one EVC. In the second step, the converted reference speaker's

voice is converted into an arbitrary speaker's voice using one-to-many EVC. In many-to-many EVC, only one EV-GMM is used because the one-to-many EV-GMM can also be used as the many-to-one EV-GMM by simply switching the source and target features. Although this voice conversion system allows users to convert their voice into an arbitrary speaker's voice, the conversion error is often larger than those of one-to-many EVC and many-to-one EVC. Because the mixture component sequence in many-to-one EVC and one-to-many EVC is not always the same, it is possible that the source feature is converted into a different phonemic space through the sequential conversion process. It has been reported that this inconsistency of the mixture component sequence is avoided by sharing the same mixture component sequence in both many-to-one EVC and one-to-many EVC [16]. This thesis focuses on many-to-many EVC with shared mixture components.

In this section, the structure of the EV-GMM, its training process with PCA and SAT, the adaptation process, and many-to-many EVC with shared mixture components are described.

2.3.1 EV-GMM

In one-to-many EVC, the EV-GMM is used as a conversion model. The EV-GMM is trained using multiple parallel data sets consisting of a single input speech data set and many output speech data sets including various speakers' voices. The EV-GMM models the joint probability density of the input and output parameter vectors as follows:

$$
\begin{aligned}
&P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda^{(EV)}, \boldsymbol{w}) \\
&\quad = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}), \boldsymbol{\Sigma}_m^{(X,Y)}\right) \qquad (35)
\end{aligned}
$$

$$
\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{A}_m \boldsymbol{w} + \boldsymbol{b}_m \end{bmatrix} \qquad (36)
$$

$$
\boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \qquad (37)
$$

where $\lambda^{(EV)}$ is the parameter set of the EV-GMM. It consists of target-speaker-independent parameters, i.e., α_m, $\boldsymbol{\mu}_m^{(X)}$, $\boldsymbol{\Sigma}_m^{(X,Y)}$, $\boldsymbol{A}_m$, and $\boldsymbol{b}_m$ for the m^{th} mixture component, where $\boldsymbol{b}_m$ and $\boldsymbol{A}_m = [\boldsymbol{a}_m(1), \cdots, \boldsymbol{a}_m(j), \cdots, \boldsymbol{a}_m(J)]$ are a bias vector and the eigenvectors $\boldsymbol{a}_m(j) = [a_m(1), \cdots, a_m(D_y)]^\top$, respectively. A J-dimensional weight vector $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$ is a target-speaker-dependent parameter for controlling target speaker individuality. The number of eigenvectors is J.

2.3.2 Training process of EV-GMM

The EV-GMM is trained with multiple parallel data sets consisting of acoustic features of the source speaker's voice and multiple prestored target speakers' voices. In this subsection, we first describe basic method [14] for the training EV-GMM, which is based on PCA. This method consists of training a target-speaker-independent GMM $\lambda^{(0)}$, a target-speaker-dependent GMM $\lambda^{(s)}$, a bias vector $\boldsymbol{b}_m^{(0)}$, and eigenvectors $\boldsymbol{a}_m^{(j)}$.

***Step* 1 :** A target-speaker-independent GMM $\lambda^{(0)}$ is trained using multiple parallel data sets as follows:

$$\lambda^{(0)} = \underset{\lambda}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \lambda) \tag{38}$$

where $\boldsymbol{Y}_t^{(s)}$ is the s^{th} prestored target speaker's feature vector at frame t. T_s denotes the total number of frames. The total number of prestored target speakers is S.

***Step* 2 :** The target-speaker-dependent GMMs are trained for each prestored target speaker. Using only the parallel data set for the s^{th} prestored target speaker, the target-speaker-dependent GMM $\lambda^{(s)}$ for the s^{th} prestored target speaker is trained by simply updating the target mean vectors $\boldsymbol{\mu}^{(s)}$ of $\lambda^{(0)}$ based on maximum likelihood (ML) estimation as follows:

$$\lambda^{(s)} = \underset{\lambda}{\operatorname{argmax}} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \lambda) \tag{39}$$

***Step* 3 :** A $2D_yM$-dimensional supervector $\boldsymbol{SV}^{(s)} = \left[\boldsymbol{\mu}_1^{(Y)}(S)^\top, \cdots, \boldsymbol{\mu}_M^{(Y)}(S)^\top\right]^\top$ is constructed for each prestored target speaker by concatenating the updated

target mean vector of $\lambda^{(s)}$. Finally, the bias vector $\boldsymbol{b}_m$ and eigenvectors $\boldsymbol{A}_m$ are extracted by performing PCA for the supervectors for all prestored target speakers. Each supervector is approximated as follows:

$$\boldsymbol{SV}^{(s)} \approx \left[\boldsymbol{A}_1^\top, \cdots, \boldsymbol{A}_M^\top\right]^\top \boldsymbol{w}^{(s)} + [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M] \tag{40}$$

$$\boldsymbol{b}_m = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\mu}_m^{(Y)}(s) \tag{41}$$

where $\boldsymbol{\omega}^{(s)}$ is s^{th} prestored target speaker's the $J(J < S \ll 2D_yM)$ dimensional weight parameter for the eigenvectors.

The phonetic space modeled by each mixture is integrated into the same space regardless of the target speaker because the probability density distribution of the input feature is common regardless of the target speaker. Thus, a supervector space is constructed in which phonemic information and speaker individuality are separated.

The adapted EV-GMM is unsuitable as a target-speaker-dependent model because the target-speaker-independent parameter of the PCA-based EV-GMM is strongly affected by acoustic variations among the many prestored target speakers. To alleviate this problem, EV-GMM training based on SAT has been proposed. The canonical EV-GMM is trained by maximizing the total likelihood of the adapted EV-GMMs for individual prestored target speakers as follows:

$$\left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{\Omega}}^{(S)}\right\} = \underset{\lambda^{(EV)}, \boldsymbol{\Omega}^{(S)}}{\operatorname{argmax}} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \lambda^{(EV)}, \boldsymbol{w}^{(s)}) \tag{42}$$

where $\boldsymbol{\Omega}^{(S)} = \left\{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \ldots, \boldsymbol{\omega}^{(S)}\right\}$ is a set of weight parameters for individual prestored target speakers. The training process is performed using the EM algorithm [21] by maximizing the following auxiliary function:

$$\begin{aligned} &Q\left(\left\{\hat{\lambda}^{(EV)}, \hat{\boldsymbol{\Omega}}^{(S)}\right\}, \left\{\lambda^{(s)}, \boldsymbol{\Omega}^{(S)}\right\}\right) \\ &= \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)}) \log P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m | \lambda^{(EV)}, \boldsymbol{w}^{(s)}). \end{aligned} \tag{43}$$

In the E-step, the posterior probability $P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{w}^{(s)})$ is calculated. Then, the target-speaker-independent parameter of the EV-GMM $\lambda^{(EV)}$ and the

speaker-dependent weight parameter set $\boldsymbol{\Omega}^{(S)}$ are updated in the M-step. It is difficult to update all parameters simultaneously because some of them depend on each other. Therefore, each parameter of the EV-GMM is updated in order. The update process is iteratively performed in each M-step to improve the parameter estimation accuracy.

First, the weight parameters for all prestored target speakers are updated. The ML estimate of the weight parameters for the s^{th} prestored target speaker is written as

$$\hat{\boldsymbol{\omega}}^{(s)} = \left(\sum_{m=1}^{M} \Gamma_{m,s} \boldsymbol{A}_m^\top \boldsymbol{P}_m^{(YY)} \boldsymbol{A}_m \right)^{-1} \times \left[\sum_{m=1}^{M} \boldsymbol{A}_m^\top \left\{ \boldsymbol{P}_m^{(YX)} (\overline{\boldsymbol{X}}_m^{(s)} - \Gamma_m^{(s)} \boldsymbol{\mu}_m^{(X)}) + \boldsymbol{P}_m^{(YY)} (\overline{\boldsymbol{Y}}_m^{(s)} - \Gamma_m^{(s)} \boldsymbol{b}_m) \right\} \right], \quad (44)$$

where

$$\Gamma_m^{(s)} = \sum_{t=1}^{T_S} \gamma_{m,t}^{(s)} = \sum_{t=1}^{T_S} P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \lambda^{(EV)}, \boldsymbol{\omega}^{(s)}) \quad (45)$$

$$\overline{\boldsymbol{X}}_m^{(s)} = \sum_{t=1}^{T_S} \gamma_{m,t}^{(s)} \boldsymbol{X}_t \quad (46)$$

$$\overline{\boldsymbol{Y}}_m^{(s)} = \sum_{t=1}^{T_S} \gamma_{m,t}^{(s)} \boldsymbol{Y}_t^{(s)} \quad (47)$$

$$\boldsymbol{\Sigma}_m^{(X,Y)^{-1}} = \begin{bmatrix} \boldsymbol{P}_m^{(XX)} & \boldsymbol{P}_m^{(XY)} \\ \boldsymbol{P}_m^{(YX)} & \boldsymbol{P}_m^{(YY)} \end{bmatrix}. \quad (48)$$

Next, the parameter set for the mean vector of the EV-GMM is updated as follows:

$$\hat{\boldsymbol{v}}_m = \left(\sum_{s=1}^{S} \Gamma_m^{(s)} \hat{\boldsymbol{W}}_s^\top \boldsymbol{\Sigma}_m^{(XY)^{-1}} \hat{\boldsymbol{W}}_s \right) \left(\sum_{s=1}^{S} \Gamma_m^{(s)} \hat{\boldsymbol{W}}_s^\top \boldsymbol{\Sigma}_m^{(XY)^{-1}} \overline{\boldsymbol{Z}}_m^{(s)} \right), \quad (49)$$

where

$$\overline{\boldsymbol{Z}}_m^{(s)} = \left[\overline{\boldsymbol{X}}_m^{(s)\top}, \overline{\boldsymbol{Y}}_m^{(s)\top} \right]^\top, \tag{50}$$

$$\hat{\boldsymbol{\upsilon}}_m = \left[\hat{\boldsymbol{\mu}}_m^{(X)\top}, \hat{\boldsymbol{b}}_m^\top, \hat{\boldsymbol{a}}_m(1)^\top, \hat{\boldsymbol{a}}_m(2)^\top, \cdots, \hat{\boldsymbol{a}}_m(J)^\top \right]^\top, \tag{51}$$

$$\hat{\boldsymbol{W}}_s = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \hat{\omega}_1^{(s)} \boldsymbol{I} & \hat{\omega}_2^{(s)} \boldsymbol{I} & \cdots & \hat{\omega}_J^{(s)} \boldsymbol{I} \end{bmatrix}. \tag{52}$$

Then, the weights for the mixture components are determined as follows:

$$\hat{\alpha}_m = \frac{\sum_{s=1}^S \Gamma_m^{(s)}}{\sum_{m=1}^M \sum_{s=1}^S \Gamma_m^{(s)}}. \tag{53}$$

Finally, the covariance matrix is updated as follows:

$$\boldsymbol{\Sigma}_m^{(X,Y)} = \frac{1}{\sum_{s=1}^S \Gamma_m^{(s)}} \sum_{s=1}^S \left\{ \overline{\boldsymbol{V}}_m^{(s)} + \Gamma_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)\top} - \left(\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} \overline{\boldsymbol{Z}}_m^{(s)\top} + \overline{\boldsymbol{Z}}_m^{(s)} \hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)\top} \right) \right\} \tag{54}$$

where

$$\overline{\boldsymbol{V}}_m^{(s)} = \sum_{t=1}^{T_s} \gamma_{m,t}^{(s)} \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top} \right] \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top} \right]^\top, \tag{55}$$

$$\hat{\boldsymbol{\mu}}_{m,s}^{(X,Y)} = \hat{\boldsymbol{W}}_s \overline{\boldsymbol{V}}_m^{(s)} = \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(X)} \\ \hat{\boldsymbol{A}}_m \hat{\boldsymbol{\omega}}^{(s)} + \hat{\boldsymbol{b}}_m \end{bmatrix}. \tag{56}$$

2.3.3 Adaptation process

The trained EV-GMM allows users to control the voice quality of the converted speech by manipulating the weight vector $\boldsymbol{w}$. If users have the target speech data, the GMMs for the source speech and a new target speech can be flexibly built by automatically determining the weight vector $\boldsymbol{w}$ using only a few arbitrary utterances of the target speech in a text-independent manner. The optimal weight vector $\hat{\boldsymbol{w}}$ is estimated by maximizing the likelihood of the following marginal

distribution [17]:

$$\begin{aligned}\hat{\boldsymbol{w}} &= \underset{w}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)}|\lambda^{(EV)}, \boldsymbol{w}) d\boldsymbol{X} \\ &= \underset{w}{\operatorname{argmax}} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(tar)}|\lambda^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}_t \\ &= \underset{w}{\operatorname{argmax}} \prod_{t=1}^{T} P(\boldsymbol{Y}_t^{(tar)}|\lambda^{(EV)}, \boldsymbol{w}). \end{aligned} \tag{57}$$

where $\boldsymbol{Y}^{(tar)}$ is the time sequence of the target features used in the adaptation. This adaptation process is performed using the EM algorithm by maximizing the following auxiliary function:

$$Q(\boldsymbol{\omega}, \hat{\boldsymbol{\omega}}) = \sum_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}, \boldsymbol{w}) \log P(\boldsymbol{Y}_t^{(tar)}, m|\lambda^{(EV)}, \boldsymbol{w}). \tag{58}$$

The weight parameter for the target is estimated as follows:

$$\hat{\boldsymbol{\omega}} = \left\{ \sum_{m=1}^{M} \Gamma_m^{(tar)} \boldsymbol{A}_m^\top {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} \boldsymbol{A}_m \right\}^{-1} \sum_{m=1}^{M} \boldsymbol{A}_m^\top {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} \overline{\boldsymbol{Y}}_m^{(tar)} \tag{59}$$

where

$$\Gamma_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}^{(tar)} = \sum_{t=1}^{T} P(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}, \boldsymbol{w}) \tag{60}$$

$$\overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}^{(tar)} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m). \tag{61}$$

This process is completely unsupervised adaptation using only arbitrary utterances of the target speaker. Because the number of estimated parameters, which are weight parameters, is small, only a small amount of adaptation data is needed for adaptation.

Although the unsupervised adaptation of the EV-GMM is successfully performed using only a small amount of adaptation data such as a few sentences, the conversion performance rapidly degrades owing to the overfitting problem when the amount of adaptation is very limited, for example, one utterance or less. In the worst case, it is possible that each mixture component is assigned

to an improper acoustic space. The MAP adaptation for EVC considering prior information of the weight vector has been proposed to improve the robustness of the EV-GMM adaptation against the amount of adaptation data [22]. As the prior distribution, the following Gaussian distribution is employed:

$$P\left(\boldsymbol{\omega}|\lambda^{(\omega)}\right) = \mathcal{N}\left(\boldsymbol{\omega}; \boldsymbol{\mu}^{(w)}, \boldsymbol{\Sigma}^{(\omega)}\right), \tag{62}$$

where $\lambda^{(\omega)}$ is a model parameter set consisting of the mean vector $\boldsymbol{\mu}^{(w)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(\omega)}$. This model parameter set is trained in advance using a set of weight vectors estimated for individual prestored target speakers in SAT as follows:

$$\hat{\lambda}^{(\omega)} = \underset{\lambda^{(s)}}{\operatorname{argmax}} \prod_{s=1}^{S} P\left(\hat{\boldsymbol{\omega}}|\lambda^{(\omega)}\right). \tag{63}$$

For the given adaptation data, the MAP adaptation of the EV-GMM is conducted as follows:

$$\hat{\lambda}^{(\omega)} = \underset{w}{\operatorname{argmax}}\, P\left(\boldsymbol{w}|\lambda^{(w)}\right)^{\tau} \prod_{t=1}^{T} P(\boldsymbol{Y}_t^{(tar)}|\lambda^{(EV)}, \boldsymbol{w}), \tag{64}$$

where τ is the hyperparameter controlling the balance between the prior distribution of weight parameters and the ML estimate. This estimation is conducted by iteratively maximizing the following auxiliary function using the EM algorithm:

$$\begin{aligned} Q(\boldsymbol{\omega}, \hat{\boldsymbol{\omega}}) = {} & \tau \log P\left(\boldsymbol{w}|\lambda^{(w)}\right) \\ & + \sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\boldsymbol{Y}_t^{(tar)}, \lambda^{(EV)}, \boldsymbol{w}) \log P(\boldsymbol{Y}_t^{(tar)}, m|\lambda^{(EV)}, \boldsymbol{w}). \end{aligned} \tag{65}$$

In the MAP estimation, the weight parameter for the target is estimated as follows:

$$\begin{aligned} \hat{\boldsymbol{\omega}} = {} & \left\{\tau{\boldsymbol{\Sigma}^{(w)}}^{-1} + \sum_{m=1}^{M} \Gamma_m^{(tar)} \boldsymbol{A}_m^{\top} {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} \boldsymbol{A}_m\right\}^{-1} \\ & \cdot \left\{\tau{\boldsymbol{\Sigma}^{(w)}}^{-1} + \sum_{m=1}^{M} \boldsymbol{A}_m^{\top} {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} \overline{\boldsymbol{Y}}_m^{(tar)}\right\} \end{aligned} \tag{66}$$

If the adaptation data is not given, the estimated weight parameter vector is equal to the prior mean vector. As the amount of adaptation data increases, the MAP estimate asymptotically approaches the ML estimate. Note that if the hyperparameter τ is set to 0, the MAP estimate is always equal to the ML estimate.

The conversion process is performed in the same manner as the conversion of statistical VC discussed in Section 2.2.2

2.3.4 Many-to-many EVC

In this section, many-to-many EVC with shared mixture components is described. Many-to-many EVC consists of training, adaptation, and conversion processes. In the training process, the canonical EV-GMM is trained in the same manner as described in Section 2.3.2. In the adaptation process, the source speaker's weight parameter $\boldsymbol{\omega}^{(i)}$ and target speaker's weight parameter $\boldsymbol{\omega}^{(o)}$ are independently estimated using a few utterances of each speaker. Then, the one-to-many EV-GMM and many-to-one EV-GMM are independently adapted using the estimated weight parameters. Then, the joint probability density of the acoustic features between the source speaker's voice and the target speaker's voice is derived as

$$\begin{aligned}
&P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}, \lambda^{(EV)}\right) \\
&= \sum_{m=1}^{M} P(m|\lambda^{(EV)}) \int P\left(\boldsymbol{Y}_t^{(i)} | \boldsymbol{X}_t, m, \boldsymbol{w}_t^{(i)}, \lambda^{(EV)}\right) \\
&\quad P\left(\boldsymbol{Y}_t^{(o)} | \boldsymbol{X}_t, m, \boldsymbol{w}_t^{(o)}, \lambda^{(EV)}\right) P\left(\boldsymbol{X}_t | m, \lambda^{(EV)}\right) d\boldsymbol{X}_t \\
&= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{y}^{(i)} \\ \boldsymbol{y}^{(o)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(i)} \\ \boldsymbol{\mu}_m^{(o)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right),
\end{aligned} \tag{67}$$

where

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)}, \tag{68}$$

$$\boldsymbol{\mu}_m^{(i)} = \boldsymbol{A}_m \boldsymbol{\omega}^{(i)} + \boldsymbol{b}_m, \tag{69}$$

$$\boldsymbol{\mu}_m^{(o)} = \boldsymbol{A}_m \boldsymbol{\omega}^{(o)} + \boldsymbol{b}_m. \tag{70}$$

In the conversion process, the converted static feature sequence vector is estimated using the adapted EV-GMM. Maximum likelihood estimation considering

the dynamic features and GV [14] is adopted.

2.4 Summary

This chapter has described statistical voice conversion (VC) based on the GMM. The dynamic characteristic has been introduced in order to capture certain correlations between frames over an utterance. Moreover, the statistic of GV has been introduced to suppress oversmoothing. Also, this chapter has described eigenvoice conversion (EVC), which allows users to more flexibly train the conversion model than GMM-based VC.

3. Enhancement of alaryngeal speech

This chapter describes a novel speaking-aid system based on VC/EVC frameworks for laryngectomees. This system allows laryngectomees to produce natural speech by overcoming the physical constraint of having no vocal folds. Although alaryngeal speech allows laryngectomees to utter speech sounds, it suffers from the lack of speech quality and speaker individuality. To improve the sound quality of alaryngeal speech, alaryngeal-speech-to-speech (AL-to-Speech) methods based on statistical VC have been proposed for electrolaryngeal speech (EL speech) and silent electrolaryngeal speech (silent EL speech). In this chapter, we apply this VC-based speaking-aid framework to esophageal speech (ES speech) enhancement. This approach, called ES-to-Speech, is capable of converting ES speech into normal speech uttered by non-laryngectomees. The effectiveness of ES-to-Speech is demonstrated by objective and subjective evaluations. Moreover, one-to-many EVC is capable of flexibly controlling the voice quality of the converted speech by adapting the conversion model to given target natural voices. This is further implemented in the AL-to-Speech methods to effectively recover speaker individuality for each type of alaryngeal speech. In this chapter, these proposed systems are compared with each other from various perspectives. The experimental results demonstrate that our proposed systems yield significant improvements in speech quality and speaker individuality for each type of alaryngeal speech.

3.1 Introduction

Unfortunately, speech is not always available to speech-disabled people who have difficulty with speech communication. A total laryngectomy is an operation to remove the vocal folds for reasons such as an accident or laryngeal cancer, and people who have undergone a total laryngectomy are called laryngectomees. They cannot produce speech sounds in the usual manner because their vocal cords have also been removed. Therefore, they require another method of producing speech sounds without vocal fold vibration. To accomplish this, alternative speaking methods of producing speech sounds using residual organs or medical devices instead of vocal cords have been used. Speech sounds generated by alternative speaking methods without vocal fold vibration are called alaryngeal speech.

There are various alternative speaking methods. This thesis focuses on three types of alternative speaking method and alaryngeal speech produced by them: the esophageal speaking method and esophageal speech (ES speech), the electrolaryngeal speaking method and electrolaryngeal speech (EL speech), and silent electrolaryngeal speaking method and silent electrolaryngeal speech (silent EL speech) [6]. Among them, the speaking methods for ES speech and EL speech are the most popular in Japan. The speaking method for ES speech is one of the major alternative speaking methods that generate alaryngeal speech using residual organs. ES speech is generated by modulating alternative excitation sounds that are produced by releasing gases from or through the esophagus by articulatory movement. This speaking method allows laryngectomees to speak without any equipment and ES speech sounds more natural than the other types of alaryngeal speech such as EL speech. However, its sound quality is inferior to that of normal speech uttered by non-laryngectomees. Although it generally takes a long time to learn the speaking method for ES speech, support for learning it is provided by many volunteers in Japan.

The speaking method for EL speech is the most popular alternative speaking method among those using medical devices. Alternative excitation sounds are produced using an electrolarynx, which is a medical device that mechanically generates sound source signals. The generated sound source signals are conducted as alternative excitation sounds into the oral cavity from the skin on the lower jaw. Then, the alternative excitation sounds are articulated to produce EL speech sounds. It is much easier to learn how to speak using the electrolarynx than to learn how to produce ES speech. Moreover, users need less physical power to produce EL speech compared with other types of alaryngeal speech, such as ES speech. However, the EL speech sounds mechanical and artificial because the generated fundamental frequency (F_0) contour is unnatural owing to the predefined frequency of the vibration (i.e., F_0 for the sound source signals). Additionally, because the electrolarynx must generate sufficiently loud sound source signals to make the produced EL speech sufficiently audible, the sound source signals are readily emitted outside, disturbing speech communication.

To resolve the issue of emitted sound source signals in the speaking method for EL speech, a new speaking method for silent EL speech has been proposed [6]. A

new sound source unit is used to generate less audible sound source signals. Since the produced speech also becomes less audible, it is detected using a nonaudible murmur (NAM) microphone [7], which is a body-conductive microphone capable of detecting extremely soft speech from the neck below the ear. The detected speech signals are presented outside as silent EL speech while keeping the external sound source signals sufficiently silent.

Although these three types of alaryngeal speech allow laryngectomees to speak again, their sound quality and intelligibility are severely degraded compared with those of normal speech. Moreover, alaryngeal speech sounds have a similar voice quality regardless of the speaker because the production mechanism of the sound source signals in each type of alaryngeal speech strongly affects the voice quality of the proposed speech. Consequently, alaryngeal speech suffers from the degradation of speaker individuality.

Several attempts to improve alaryngeal speech quality have been made. A new electrolarynx using an air pressure sensor has been developed to control F_0 for the sound source signals via the expiratory pressure. Although it is not easy to accurately control F_0 by adjusting the expiratory pressure, this device makes it possible for laryngectomees to produce more naturally sounding EL speech, the F_0 of which effectively varies over an utterance [23]. One weakness of this device is that both hands are needed to hold the electrolarynx and air pressure sensor while speaking. Moreover, it is still difficult to mechanically generate sound source signals similar to those naturally generated by vocal fold vibrations. Consequently, the produced EL speech quality is still different from the natural voices produced by non-laryngectomees.

As another attempt, speech enhancement methods based on the modification of acoustic features of ES speech using signal processing have been proposed, such as comb filtering [24], the smoothing of acoustic parameters [25], formant manipulation [26], and noise reduction based on auditory masking [27]. Although they are useful in alaryngeal speech enhancement, the improvement in quality is still limited since the acoustic features of alaryngeal speech exhibit markedly different properties from those of normal speech, and therefore, it is difficult to compensate for these acoustic differences using such a simple modification process.

Recently, statistical approaches to alaryngeal speech enhancement have been

proposed [23, 28] to convert alaryngeal speech into target normal speech while keeping the linguistic information unchanged. The statistical enhancement framework consists of training and conversion processes. In the training process, a function converting the acoustic features of alaryngeal speech into those of target normal speech is modeled using training data including utterance pairs of alaryngeal speech and normal speech. In the conversion process, an utterance of alaryngeal speech is converted to that of the target normal speech using the conversion function. This data-driven approach is capable of more complicated acoustic modifications to compensate for the large acoustic differences between alaryngeal speech and normal speech.

As typical conventional methods, the codebook mapping method [2] and a probabilistic conversion method based on GMMs [3] have been applied to alaryngeal speech enhancement [6, 28, 29]. The GMM-based conversion method is one of the most popular voice conversion methods. It is well defined mathematically and its conversion performance is relatively high. It has been reported that the alaryngeal speech enhancement method based on GMM-based voice conversion, which is called the alaryngeal speech-to-speech (AL-to-Speech) method, is highly effective for improving the naturalness and intelligibility of the different types of alaryngeal speech [6, 23, 29]. In this thesis, this statistical enhancement technique is applied into ES speech enhancements.

Although these conventional enhancement methods allow laryngectomees to speak in more natural voices than alaryngeal speech, recovering speaker individuality has hardly been considered. In fact, it is difficult to flexibly control the voice quality of enhanced alaryngeal speech by these methods. In the statistical voice conversion approaches, it is possible to change the voice quality of the converted speech using different target voices but it is necessary to prepare training data consisting of utterance pairs of the alaryngeal speech and each target voice, which is very laborious. In this thesis, to flexibly change the voice quality of the converted speech to recover speaker individuality or provide a unique voice for laryngectomees, one-to-many EVC [5] is applied to ES speech enhancement, an approach called esophageal speech-to-speech (ES-to-Speech). One-to-many EVC is a technique for converting a specific source speaker's voice into an arbitrary target speaker's voice. This method allows us to control the speaker individual-

ity of the converted speech by manipulating a small number of parameters or to flexibly adapt the conversion model to an arbitrary target speaker on the basis of a small number of given target speech samples in a text-independent manner. ES-to-Speech based on EVC helps laryngectomees speak in their desired voice or in their previous voice provided a few recorded samples are available.

In this thesis, we also develop AL-to-Speech systems capable of flexibly controlling the enhanced voice quality based on one-to-many EVC for not only ES speech but also EL speech and silent EL speech. The effectiveness of the proposed AL-to-Speech systems based on VC/EVC for the three types of alaryngeal speech is evaluated from various perspectives. The features of each AL-to-Speech system are demonstrated through various comparisons among the different AL-to-Speech systems.

This chapter is organized as follows. Laryngectomees and alaryngeal speech are described in Section 3.2 and Section 3.3, respectively. Several conventional speaking-aid systems for laryngectomees are described in Section 3.4. The enhancement of ES speech based on VC and EVC is described in Section 3.5. The enhancement of ES speech, EL speech, and silent EL speech based on one-to-many EVC is described in Section 3.6. The experimental evaluations of our proposed methods is described in Section 3.7 This chapter is summarized in Section 3.8

3.2 Laryngectomees

People who have undergone a total laryngectomy, called laryngectomees, cannot speak in the usual manner owing to the removal of their vocal folds. A total laryngectomy is a surgical operation that cuts off the connection between the trachea, pharynx, and larynx, and then the larynx including the vocal folds is enucleated. The main reason for undergoing a total laryngectomy is laryngeal cancer. Laryngeal cancer is categorized according to the location of the tumor into glottic cancer, supraglottic cancer, and subglottic cancer. It is easier to detect laryngeal cancer than other cancers because it causes hoarseness in speech production. In the early stage of laryngeal cancer, the larynx can be recovered fully by radiation therapy, which can cure the cancer while retaining the larynx and vocal folds. For the intermediate and final stages of laryngeal cancer, there are three main types of surgical procedure: partial laryngectomy, total laryngectomy, and supracricoid

laryngectomy with cricohyoid-epiglottopexy (SCL-CHEP) [30]. A partial laryngectomy removes past of the larynx to preserve the patient's voices. Although a partial laryngectomy does not severely a affect speech communication, it is no longer generally performed because of the high likelihood of the reappearance of the disease and the high frequency of aspiration. A total laryngectomy removes all surrounding areas including the epiglottis, hyoid bone, arytenoid cartilage, cricoid cartilage, thyroid cartilage, and vocal folds, and is the default surgical procedure for laryngeal cancer in the final stage. SCL-CHEP is a rather novel surgical procedure, which preserves the patient's voice even though the vocal folds are removed. Note that SCL-CHEP cannot always be performed depending on the state of the disease. The speech quality after undergoing SCL-CHEP also depends on the state of the disease. Moreover, SCL-CHEP is still not yet become popular, meaning that, many patients who suffer from laryngeal cancer become laryngectomees. The number of laryngectomees was estimated to be less than 20,000 more than 20 years ago [31]. Since then, the number of people with laryngeal cancer has tended to increase yearly with the increasing numbers of elderly people in Japan.

Figure 5 shows air flows from the lungs in non-laryngectomees and total laryngectomees. In laryngectomees, the trachea and the oral cavity connected to the esophagus are completely separated from each other to prevent food from entering the trachea. Therefore, laryngectomees cannot generate vocal fold vibrations or expire air through the oral cavity.

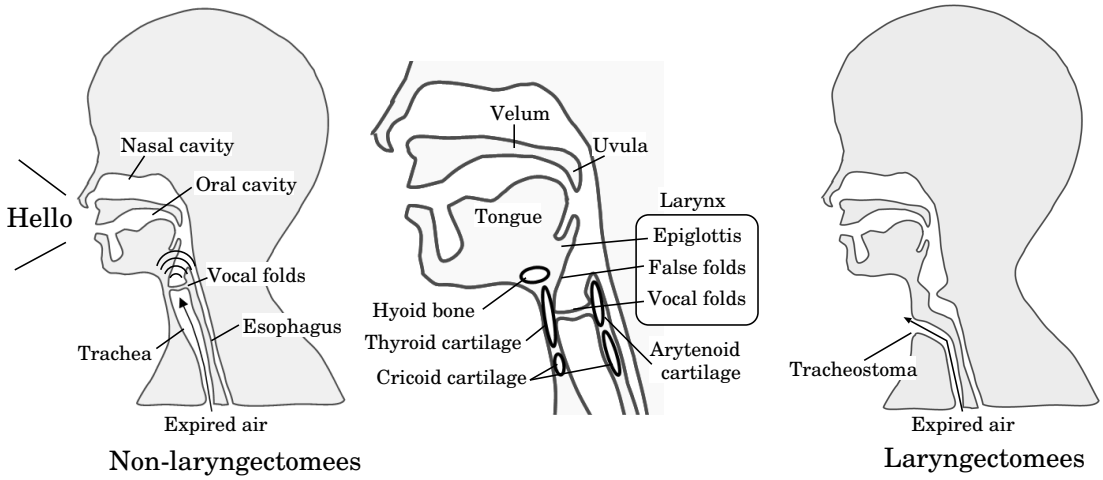

Figure 5. The air flows from lungs in non-laryngectomees and laryngectomees

3.3 Alaryngeal speech

In this section, we describe alternative speaking methods for producing speech sounds using residual organs or medical devices instead of vocal cords. Speech sounds generated by alternative speaking methods without vocal fold vibration are called alaryngeal speech. Laryngectomees have three main types of alternative speaking method which are different ways of obtaining the sound sources: 1) a method using an external unit such as an electrolarynx (EL) or whistle-larynx [32], 2) the esophageal speaking method (ES), 3) the tracheo-esophageal (TE) shunt speaking method. Figure 6 shows the different types of alaryngeal speech for laryngectomees.

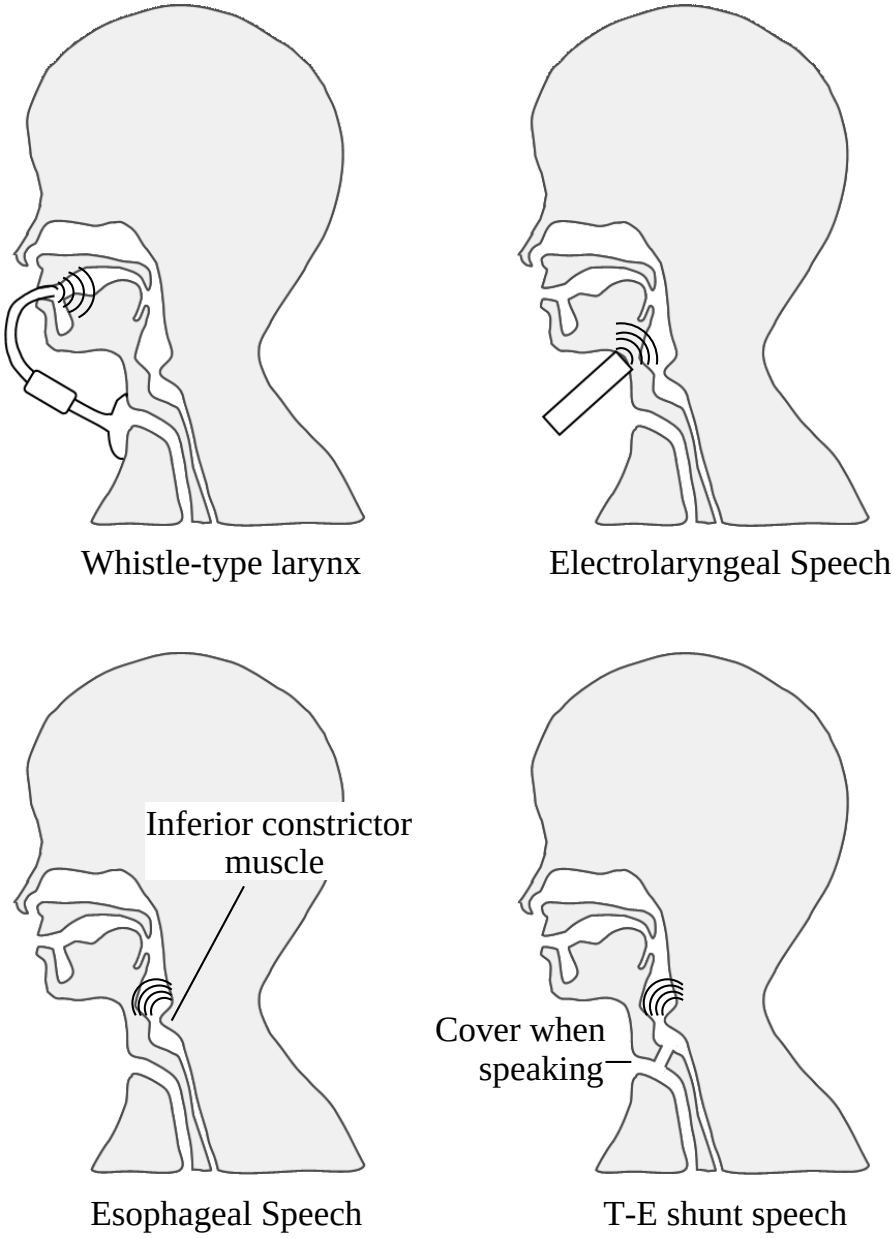

Figure 6. Alaryngeal speeches for laryngectomees

3.3.1 Whistle-larynx

One major external speaking device is the whistle-larynx which generates an excitation sound by the vibration of a rubber membrane using air released from the tracheostoma. Then, the excitation sound is delivered into the mouth by holding a whistle. The fundamental frequency (F_0) is manipulated via the expired air that flows from the tracheostoma. As a result, whistle-larynx enables laryngectomees to speak with natural speech compared with an EL. However, there are some syllables that are difficult to produce because articulation is encumbered owing to the speaker biting on the tube.

3.3.2 Electrolaryngeal speech

The other major external medical device is an EL, which generates an excitation sound by beating a vibration plate using electric power. An EL is pushed onto the lower jaw during speech, and it is switched on and off by a button. EL speech exhibits higher intelligibility than the other types of alaryngeal speech. Moreover, it is easier to produce speech by the EL method than other types of alaryngeal speech. Therefore, EL speech is the most popular alternative speaking method in Japan. On the other hand, the main defect of the EL is its fixed F_0, making EL speech sound mechanical and unnatural. Moreover, the sound source signals are noisy and may disturb people around the speaker especially in quiet situations.

3.3.3 Esophageal speech

ES speech is generated as follows. First, the laryngectomee pumps a certain amount of air from the mouth into the esophagus and the stomach, which play the role of the respiratory organs. Next, releasing the air from them, excitation sounds are generated by vibrating tissues around the entrance of the esophagus. Finally, esophageal speech is produced by articulating the generated excitation sounds in the same manner as performed by non-laryngectomees.

ES speech sounds more natural than the other types of alaryngeal speech. Moreover, ES speech has some other merits: it allows laryngectomees to speak without any equipment, and support for learning esophageal speech is provided by many volunteers in Japan. Thus, esophageal speech is one of the most popular

alternative speaking methods in Japan. On the other hand, although a speaker skilled in producing ES speech can control prosody using residual organs, the produced sound is constantly low in tone regardless the speaker. Moreover, specific unnatural sounds caused by producing the excitation sounds in the manner mentioned above are often observed. Because ES require strength for the speaker, some elderly people find it difficult to speak with ES.

3.3.4 TE shunt speech

The TE shunt speaking method is similar a speaking method to ES. The only difference from ES is the method of producing the air flow used to vibrate the vocal folds. In TE shunt speaking method, the air flow is delivered from the lungs and trachea into the esophagus through a voice prosthesis, which is a valve inserted between the trachea and esophagus. When speaking, laryngectomees block the tracheostoma to make the air flow to the esophagus through the prosthesis. The air vibrates tissues around the entrance of the esophagus, inducing sound source vibration similarly to in esophageal speech. It is easier to produce TE shunt speech than ES speech and the resulting speech is smoother because laryngectomees can use their breath in the same way as non-laryngectomees. Moreover, for the same reason, TE shunt speech has greater power than ES speech. Thus, TE shunt speech sounds more natural than ES speech and other alaryngeal speech. On the other hand, some elderly patients or those with a lung disease cannot undergo the operation required to enable TE shunt speech. Moreover, the voice prosthesis must be maintained every several years. Therefore, it is a less popular method in Japan.

3.4 Conventional speaking-aid systems for laryngectomees

The naturalness and speaker individuality of alaryngeal speech are less than those of normal speech uttered by non-laryngectomees. To alleviate this problem, many speaking-aid systems for each types of alaryngeal speech have been proposed. In this section, we describe conventional speaking-aid systems for EL and ES, which are the most popular methods in Japan.

3.4.1 Speaking-aid systems for esophageal speech

As the speaking-aid system for ES speech, modification of the acoustic features of the produced ES speech based on signal processing has been studied widely. Many methods based on this approach have been proposed, and their effectiveness has been reported. However, these enhancement methods may be difficult to use in face-to-face communication because the listener must listen to not only the enhanced speech but also the produced ES speech. On the other hand, enhancement methods are effective in situations that listeners listen to only the enhanced ES such as in telecommunication. There have been some attempts to enhance ES speech based on the modification of acoustic features such as using comb filtering [24], the smoothing of acoustic parameters [25], formant manipulation [26], and noise reduction based on auditory masking [27]. Although they have some efficacy in ES speech enhancement, it is difficult to compensate for the acoustic differences using those simple modification processes since the acoustic features of ES speech exhibit markedly different properties from these of normal speech. In particular, it is difficult to extract the F_0 corresponding to pitch information even if pitch information can be perceived from the ES speech.

Figure 7 shows waveforms, spectrograms, F_0 contours, and F_0 candidates of ES speech and normal speech. These acoustic features were extracted by STRAIGHT [33]. In this figure, we can see that the extraction of F_0 is not effective. Moreover, other acoustic features of the ES speech are also markedly different from those of the normal speech. Therefore, the enhancement of ES speech based on the modification of acoustic features does not achieve the naturalness of normal speech. To enhance the quality of ES speech, it is essential to develop more sophisticated techniques enabling the more complicated modification of ES speech parameters.

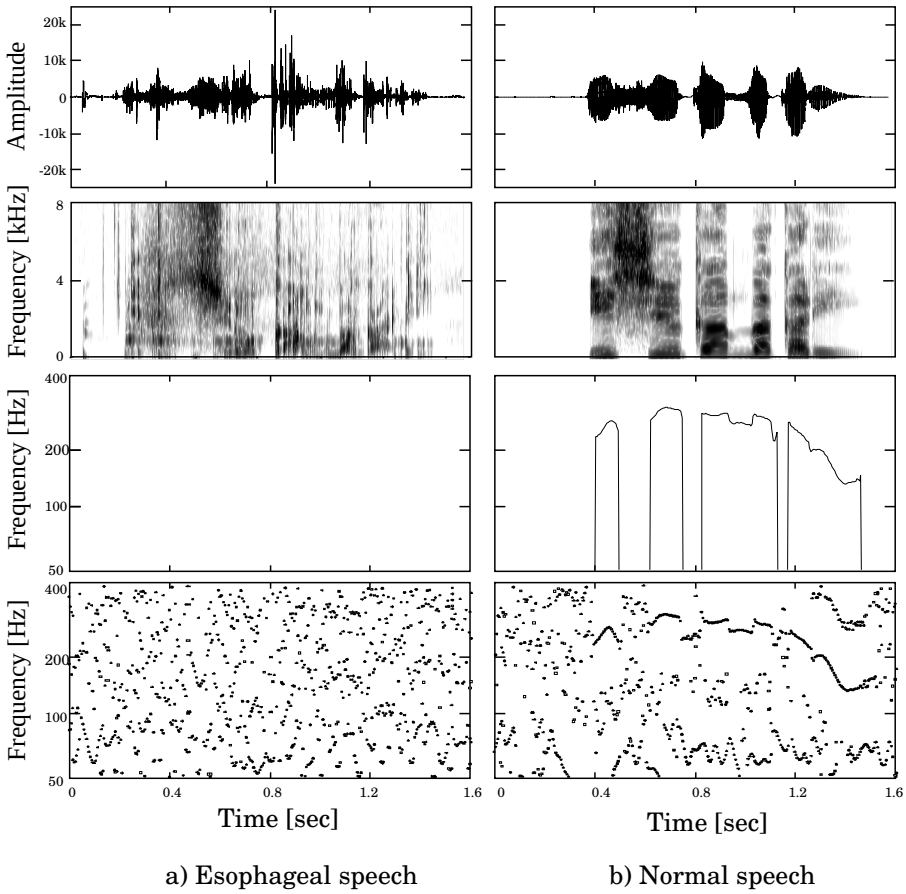

Figure 7. Example of waveforms, spectrograms, F_0 contours, and the 1st to 5th candidates of F_0 components of a) normal speech and b) esophageal speech for the sentence fragment /i q sh u: k a n b a k a r i/ which means "for about one week" in Japanese.

3.4.2 Speaking-aid systems for electrolaryngeal speech

As the speaking-aid systems for EL speech, three main types of approach have been proposed. One approach is to improve the artificial larynx. EL speech generated by a basic EL does not have inflection because a basic EL can only generate a monotone pitch. To alleviate this problem, two types of EL that can generate inflections have been proposed. One is an EL using an air-pressure

sensor. In this method, the laryngectomee places an air-pressure sensor in the tracheostoma. Then, inflection is generated according to the breath flowing from the tracheostoma. This EL allows laryngectomees to control inflection, resulting in the generated speech sounding more natural than that using a basic EL. However, it is difficult to freely control the air flow while speaking. Moreover, the convenience of speaking using an external device is reduced since the user needs both hands to hold the main body of the EL and the air-pressure sensor. In the other type of EL, F_0 increases and decreases over time [34]. In this EL, F_0 patterns are included in the EL in advance. Therefore, the naturalness of the generated EL speech is increased while maintaining usability. However, the essential problem of F_0 is not addressed since the F_0 pattern is fixed and cannot be changed by the uses.

Another approach for the enhancement of EL speech is modification of the acoustic features of EL speech. In this approach, the most popular method is noise reduction using spectral subtraction (SS) [35]. This method is based on the assumption that speech signals and additive noise are uncorrelated. In this method, the amount of radiated noise is significantly reduced, and the enhanced speech sound becomes clearer than EL speech. However, this method does not improve the naturalness of the speech.

Recently, an enhancement method based on statistical VC has been proposed [36]. This method, called EL-to-Speech, converts EL speech into normal speech based on statistical VC. This method consists of a training process and a conversion process. In the training process, the relationship between the acoustic features of EL speech and normal speech is modeled by a GMM. In this method, three GMMs are traind: a GMM that converts the spectrum of EL speech into a spectrum of normal speech, a GMM that converts the spectrum of EL speech into the F_0 of normal speech, and a GMM that converts the spectrum of EL speech into aperiodic components of normal speech. This is because F_0 and the aperiodic components of EL speech are not informative. In the conversion process, an arbitrary utterance of EL speech is converted into normal speech while keeping the linguistic information unchanged. This method significantly improves the naturalness of EL speech. However, the generated F_0 pattern is still unnatural because the relationship between F_0 and EL speech is not particularly strong. To

alleviate this problem, the conversion from EL speech into the whisper, such as that made by a non-laryngectomee has also been proposed [36]. In this method, called EL-to-Whisper, the problem of an unnatural F_0 does not occur because whispers do not have an F_0. Therefore, the listener finds the whisper more natural than the speech obtained from EL-to-Speech. However, there are many situations where a whispered voice is unsuitable for communication. Moreover, speaker individuality is fixed depending on the target speaker in EL-to-Speech and EL-to-Whisper.

3.4.3 Speaking-aid system for electrolaryngeal speech generated by low-power sound source unit

To reduce the noise in EL speech, a new EL that generates a low-power sound that it is almost too quiet for people to hear while speaking has been proposed [6]. In this thesis, this low-power sound source unit is called a silent EL. The silent EL generates an excitation sound without a large noise. However, the generated speech sound is also less audible. To detect a speech sound generated using the silent EL, a method using a NAM microphone has been proposed [6]. A NAM is defined as articulated respiratory sound without vocal fold vibration transmitted through the soft tissues of the head [37]. NAM microphone is designed to detect NAMs. Silent EL speech, which is less audible to surrounding people, is detected by a NAM microphone. The detected speech signals are presented outside as silent EL speech while keeping the external sound source signals sufficiently silent.

Silent EL speech allows the user to speak without the leakage of excitation sounds even with an EL. However, silent EL speech sounds much more unnatural than EL speech owing to its lower-power sound source signals and body conduction. It basically has similar acoustic characteristics to EL speech except that 1) the signal-to-noise ratio of silent EL speech is much lower than that of EL speech and 2) high-frequency components of over 3 or 4 kHz are severely attenuated by the lack of radiation characteristics from the lips and by the effect of the low-pass characteristics of the soft tissues.

To enhance silent EL speech, two conversion methods that convert it into normal speech and into a whispered voice have been proposed [36] similarly to EL-to-Speech and EL-to-Whisper. These methods, respectively called silent EL-

to-Speech and silent EL-to-Whisper, allows the user to speak more natural voice than silent EL while keeping the external sound source signal sufficiently silent, as shown in Fig. 8. Moreover, this method can be used in face-to-face communication because the listener only hears the converted speech sound. Thus, the communication of laryngectomees is significantly improved by these enhancement methods. However, the naturalness and intelligibility of the converted speech sound are less than those of other methods based on VC. This is because the acoustic features of silent EL speech are less informative than those of EL speech.

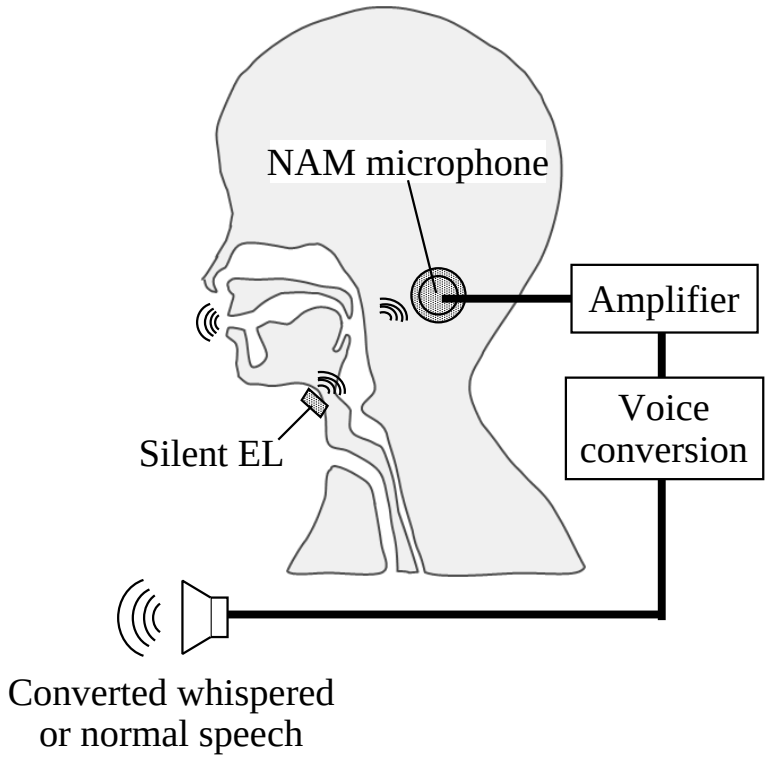

Figure 8. Overview of silent EL-to-Speech or silent EL-to-Whisper.

3.5 Proposed speaking-aid system based on VC from esophageal speech to speech (ES-to-Speech)

This section describes the speaking-aid system based on VC for ES speech. The enhancement method that converts ES speech into normal speech is referred to as ES-to-Speech. Because the converted speech parameters that smoothly vary over an utterance, as observed in normal speech, are basically determined according to the statistics extracted from the normal speech in a probabilistic manner, the specific sounds and unstable acoustic variations mentioned in Section 3.3 are effectively alleviated by the proposed conversion process. Furthermore, even if it is difficult to directly extract some speech parameters such as F_0 or unvoiced/voiced information from ES speech, VC enables the estimation of these parameters, which exhibit properties similar to those of normal speech, from another speech parameter robustly extracted from ES speech (e.g., spectral envelope). Therefore, this estimation process can be regarded as a statistical feature extraction process.

3.5.1 Feature extraction in ES-to-Speech

The spectral components of ES speech vary unstably and the spectral structures of some phonemes are often collapsed owing to the difficulty of producing them in ES speech. To address these issues, we use a spectral segment feature extracted from multiple frames as follows:

$$\boldsymbol{X}'_t = \boldsymbol{C}\boldsymbol{X}_t + \boldsymbol{d} \tag{71}$$

where $\boldsymbol{X}_t = [\boldsymbol{x}_{t-i}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+i}^\top]^\top$ is a joint vector generated by concatenating a spectral parameter vector $\boldsymbol{x}_t$ at the current frame and those at the i preceding and succeeding frames. Because this joint vector includes a significant amount of redundant information, dimension reduction by PCA is performed for the joint vector $\boldsymbol{X}_t$ to extract the spectral segment feature $\boldsymbol{X}'_t$ at frame t, where $\boldsymbol{C}$ and $\boldsymbol{d}$ are a transformation matrix and a bias vector extracted by PCA, respectively. In this feature extraction process, contextual information over several frames is effectively used to compensate for collapsed spectral features and alleviate unstable acoustic variations.

Although it is difficult to extract F_0 from ES speech (see Fig. 7), we usually perceive pitch information in ES speech. Assuming that relevant information

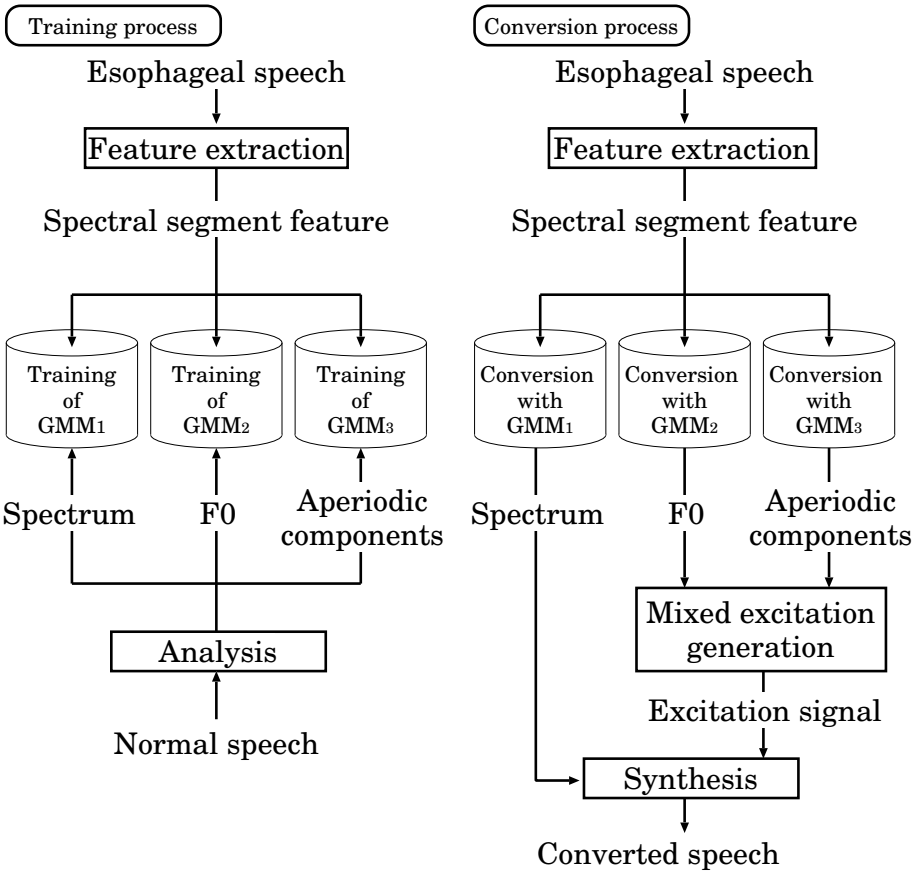

Figure 9. Training process and conversion process.

is included in the spectral parameters, we use the spectral segment feature as an input feature to estimate F_0 in the conversion process as described in [38]. Moreover, to make the estimated F_0 correspond to the perceived pitch information in ES speech, we record normal speech uttered by a non-laryngectomee so that its pitch sounds similar to that of esophageal speech and use its F_0 values as the target.

3.5.2 ES-to-Speech based on VC

The training process and conversion process are shown in Fig. 9.

In the training process, we record utterance pairs of esophageal speech and

normal speech uttered by a non-laryngectomee in the manner as mentioned above. In our proposed method, the spectral segment feature of esophageal speech is converted into three speech parameters of the target normal speech: 1) the spectral features, 2) log-scaled F_0, and 3) aperiodic components that capture the magnitude of noise of an excitation signal in each frequency band [39]. Therefore, we independently train three GMMs, modeling the joint probability densities of the spectral segment features of ES speech and individual target speech parameters using the corresponding joint feature vector sets. Note that we set a constant log-scaled F_0 value (e.g., zero) for unvoiced frames to construct the joint feature vectors of the spectral segment features and the log-scaled F_0.

In the conversion process, spectral segment features are extracted from ES speech. Then, individual converted speech parameters are independently estimated from the extracted spectral segment features using each of the trained GMMs; for example, the GMM modeling the joint probability density of the spectral segment feature of ESl speech and the target spectral features is used for the spectral estimation. After estimating the converted spectral features, the converted log-scaled F_0, and the converted aperiodic components, the excitation signal is generated using STRAIGHT mixed excitation based on the converted F_0 values and the converted aperiodic components [39]. Finally, the converted speech is synthesized by filtering the generated excitation signal with the converted spectral features.

3.5.3 ES-to-Speech based on one-to-many EVC

In ES-to-Speech based on basic VC, the voice quality of the converted speech is fixed to that of the target non-laryngectomee. To flexibly control the converted voice quality, we apply one-to-many EVC to ES-to-Speech. Figure 10 shows the training process and adaptation process in ES-to-Speech based on one-to-many EVC.

In the training process, we independently train two one-to-many EV-GMMs: one EV-GMM for estimating the converted spectral feature and the other EV-GMM for estimating the converted aperiodic components. To train these two EV-GMMs, we use multiple parallel data sets consisting of ES speech data uttered by the laryngectomee and prestored normal speech data uttered by many

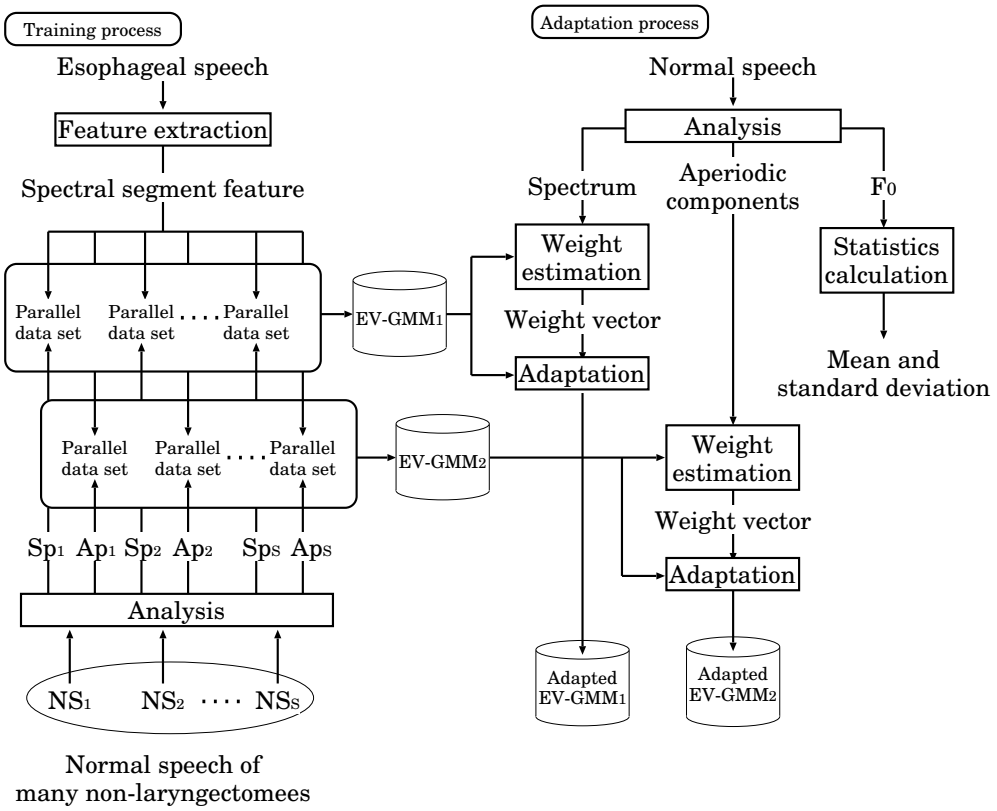

Figure 10. Training process and adaptation process in ES-to-Speech based on one-to-many EVC. "Sp_s" and "Ap_s" show spectral features and aperiodic components of normal speech of the s^{th} speaker NS_s, respectively.

non-laryngectomees. The EV-GMM for spectral estimation is trained using multiple joint feature vector sets consisting of the spectral segment features of the laryngectomee and the prestored spectral features of the non-laryngectomees. On the other hand, the EV-GMM for aperiodic component estimation is trained using multiple joint feature vector sets consisting of the spectral segment features of the laryngectomee and the prestored aperiodic components of the non-laryngectomees. In this section, we describe the PCA-based training method for the EV-GMM [5].

ES-to-Speech based on one-to-many EVC is capable of giving the converted voice similar quality to a specific voice quality if small amounts of speech data

are provided. In the adaptation process, the spectral features and aperiodic components are extracted from the given speech data, then, the weight vector of the EV-GMM for spectral estimation and that for aperiodic component estimation are independently estimated using the extracted parameters in the manner shown in eq. (57). Finally, the adapted EV-GMMs are generated on the basis of the estimated weight vectors. On the other hand, in the F_0 estimation, F_0 contours corresponding to the pitch of the input esophageal speech are first estimated using the same GMM as that used in ES-to-Speech based on basic VC. Then, the estimated log-scaled F_0 values denoted by $\log x$ are adjusted so that their mean μ_x and standard deviation σ_x are equal to those of the adaptation speech data as follows:

$$\log y = \frac{\sigma_y}{\sigma_x}(\log x - \mu_x) + \mu_y \tag{72}$$

where $\log y$ denotes the converted log-scaled F_0 value, and μ_y and σ_y denote the mean and standard deviation of log-scaled F_0 values extracted from the adaptation speech data, respectively. Because this adaptation process is performed in a completely text-independent manner and requires only a small amount of adaptation data [5], converted speech with voice quality similar to the laryngectomee's original voice quality is obtained provided recorded normal speech data of the laryngectomee before undergoing the total laryngectomy still exists. Therefore, the proposed method has great potential to provide a new speaking-aid system allowing laryngectomees to artificially recover their original voices. Furthermore, ES-to-Speech based on one-to-many EVC also allows the laryngectomee to manually control the voice quality of the converted speech by manipulating the weight vector for the eigenvectors.

In this chapter, we attempt to use the esophageal speech itself as the adaptation data for weight estimation. This approach is available even if no suitable speech data can be found for use as adaptation data. In the one-to-many EV-GMM, the voice quality of various non-laryngectomees is efficiently modeled on a subspace spanned by the eigenvectors. Therefore, this approach finds the optimum weight vector representing the normal speech of a hypothetical non-laryngectomee whose voice quality is similar to that of the ES speech in the sense of maximum likelihood. In other words, this process is regarded as a projection from an esophageal speech space into a normal speech space; the specific noises

and unstable acoustic variations are effectively alleviated while trying to keep the voice quality as similar as possible.

This process works reasonably well in the EV-GMM for spectral estimation. On the other hand, it does not work very well in the EV-GMM for aperiodic component estimation. The aperiodic component values extracted from the esophageal speech are high owing to the lower periodicity of its excitation signal. If the EV-GMM is adapted using these aperiodic components, the adapted EV-GMM always makes the converted aperiodic component values too high and the excitation signal generated on the basis of these values results in the converted speech being noisy. Because such a noisy excitation signal is caused by the production mechanism of the esophageal speech, it is regarded as a feature of the esophageal speech itself rather than a feature of individual laryngectomees. To address this issue, a multiple regression GMM (MR-GMM) [40] is used for aperiodic component estimation. The target mean vector for the m^{th} mixture component of the MR-GMM, $\boldsymbol{\mu}^{(Y)}_{m(ap)}$, is given by

$$\boldsymbol{\mu}^{(Y)}_{m(ap)}(\boldsymbol{w}_{(sp)}) = \boldsymbol{A}_m \boldsymbol{w}_{(sp)} + \boldsymbol{b}_m = \boldsymbol{A}'_m \boldsymbol{w}'_{(sp)}, \tag{73}$$

where $\boldsymbol{w}_{(sp)}$ is the weight vector of the EV-GMM for spectral estimation, $\boldsymbol{w}'_{(sp)} = [1, \boldsymbol{w}^{\top}_{(sp)}]^{\top}$, and $\boldsymbol{A}'_m = [\boldsymbol{b}_m, \boldsymbol{A}_m]$. In this thesis, the regression parameters $\boldsymbol{A}'_m$ are estimated using the multiple regression analysis approach [40], which minimizes the following error function with respect to the regression parameters:

$$\varepsilon = \sum_{s=1}^{S} \|(\boldsymbol{\mu}^{(Y)}_{m(ap)}(s) - \boldsymbol{A}'_m \boldsymbol{w}'_{(sp)}(s))\|^2 \tag{74}$$

where $\boldsymbol{w}'_{(sp)}(s)$ and $\boldsymbol{\mu}^{(Y)}_{m(ap)}(s)$ are respectively the weight vector and the m^{th} target mean vector for the s^{th} prestored target non-laryngectomee, and S is the number of prestored target non-laryngectomees.

Figure 11 shows the training process and adaptation process of this approach. The training and adaptation processes of the EV-GMM for spectral estimation are the same as those in Fig. 10. The weight vectors for individual target non-laryngectomees are determined through the PCA-based training process for this EV-GMM. These weight vectors are used in the training process of the MR-GMM for aperiodic component estimation. In the adaptation process of the MR-GMM,

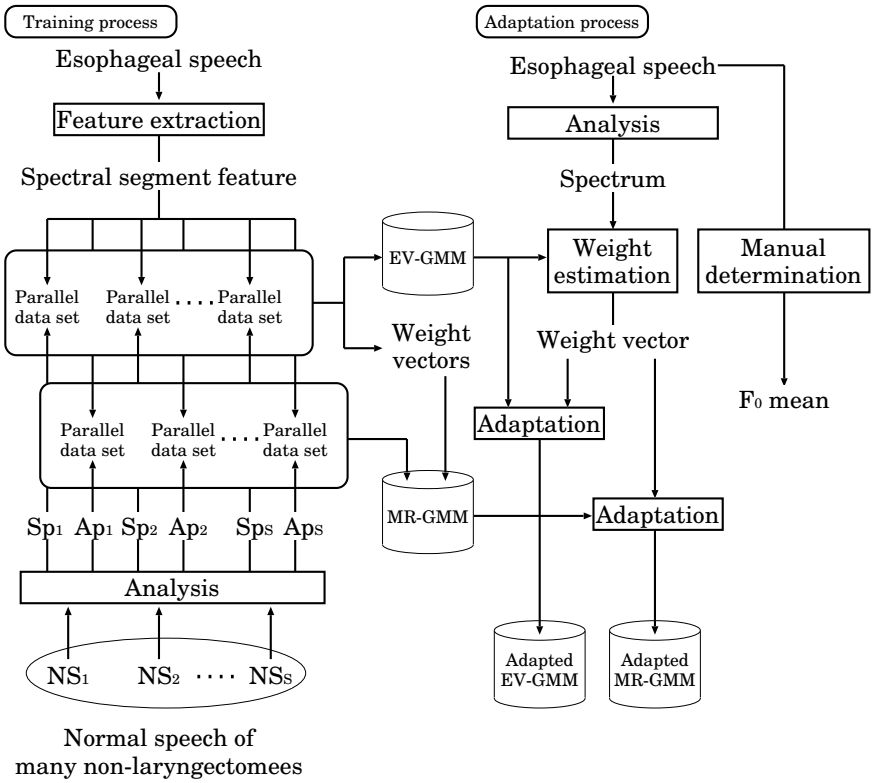

Figure 11. Training process and adaptation process in ES-to-Speech based on one-to-many EVC when using esophageal speech as adaptation data. "Sp_s" and "Ap_s" show spectral features and aperiodic components of normal speech of the s^{th} speaker NS_s, respectively.

the weight vector determined using the EV-GMM for spectral estimation is used. Note that the calculation of the F_0 statistics of the adaptation data is difficult because the extraction of F_0 from the ES speech does not work well. Therefore, the mean value of the converted speech is manually determined so that the pitch of the converted speech sounds similar to that of the esophageal speech or the converted speech sounds more natural. In this F_0 modification process, only mean values are considered in eq. (72) (i.e., we assume $\sigma_y = \sigma_x$).

3.6 Proposed speaking-aid system based on VC from alaryngeal speech to speech (AL-to-Speech)

Many types of enhancement method for the several types of alaryngeal speech have been proposed. Among them, the EVC-based methods can recover the speaker individuality of laryngectomees. Section 3.5 showed the effectiveness of the EVC-based method for ES speech. The EVC-based enhancement method allows laryngectomees to not only speak with a natural voice but also freely control the voice quality of the converted speech. Moreover, if laryngectomees record their own natural voice before undergoing a total laryngectomy, they can recover their own voice quality using the EVC-based enhancement method.

In this section, we focus on three types of alaryngeal speech: ES speech, EL speech, and silent EL speech. We apply EVC to the conventional VC-based enhancement method, i.e., EL-to-Speech and silent EL-to-Speech. This framework, which is the VC-based enhancement method for the several types of alaryngeal speech, is called AL-to-Speech in this thesis. Effectiveness of three types of EVC-based enhancement method, i.e., ES-to-Speech based on EVC, EL-to-Speech based on EVC, and silent EL-to-Speech based on EVC, are compared by experimental evaluations.

3.6.1 Feature extraction in AL-to-Speech

In AL-to-Speech, three types of acoustic feature of normal speech, namely, the spectrum, aperiodic components, and F_0, are separately estimated from the acoustic features of each type of alaryngeal speech. Then, the estimated acoustic features are used in vocoding to generate the converted speech. To estimate the acoustic features of normal speech by AL-to-Speech, we need to decide which acoustic features of each type of alaryngeal speech to use as the input feature. However, we have little choice because most of the acoustic features of alaryngeal speech are less informative. Table 1 and Fig. 12 shows the characteristics and samples of the acoustic features of each type of alaryngeal speech, respectively.

Because the spectrum of ES speech is the only informative acoustic feature of ES speech, even though it unstably varies, we use the spectrum of ES speech as an input feature to estimate the spectrum and aperiodic components of normal

Table 1. *Characteristics of acoustic features of alaryngeal speech.*

	ES	EL and silent EL
Spectrum	Unstably varying according to phoneme (still informative)	Varying according to phoneme (still informative)
Aperiodic components	Constantly noisy (less informative)	Strong due to mechanical excitation (less informative)
F_0	Difficult to extract (less informative)	Constant due to mechanical excitation (less informative)

speech, which smoothly vary according to the phoneme. On the other hand, for F_0 estimation, we assume that the pitch of ES speech can be included in the spectrum including the power as mentioned in Section 3.5. On the basis of this assumption, we also use the spectrum of ES speech as an input feature to estimate F_0 for normal speech. The target normal speech is uttered by a non-laryngectomee so that its pitch sounds similar to that of ES speech.

Although the spectra of EL speech and silent EL speech change according to the phoneme, they are significantly different from those of normal speech. The F_0 values of EL speech and silent EL speech are mechanically determined independently of the utterance content. Moreover, because an electrolarynx is driven during an utterance, the aperiodic components of EL speech and silent EL speech are not informative as the input feature. Therefore, in the acoustic features of EL speech and silent EL speech, only the spectrum is informative; thus, we use the spectra of EL speech and silent EL speech as an input feature to estimate the spectrum, aperiodic components, and the F_0 for normal speech.

In AL-to-Speech for these three types of alaryngeal speech, the spectrum of each types of alaryngeal speech is used as an input feature to estimate each acoustic feature of some types of alaryngeal speech, respectively. However, directly using the spectrum of alaryngeal speech causes the degradation of the converted speech because the spectrum structures of some phonemes of alaryngeal speech are often collapsed owing to the difficulty of producing them. To address these

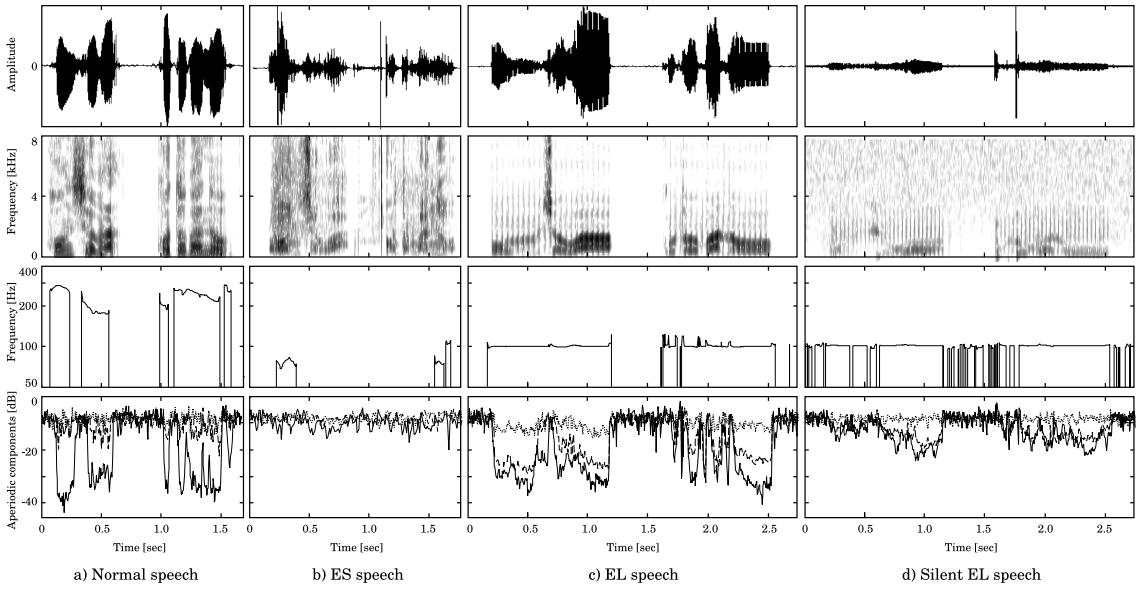

Figure 12. Example of acoustic features, i.e., waveforms, spectrograms, F_0 contours, and aperiodic components of a) normal speech, b) ES speech, c) EL speech, and d) silent EL speech in the same sentence fragment /h o N sy o w a k o t o b a n o/. In aperiodic components, the solid line, coarse broken line, and fine broken line represent averaged aperiodic components in low frequency band, middle frequency band, and high frequency band, respectively.

issues, we use a spectral segment feature extracted from multiple frames [41] eq. (71).

3.6.2 AL-to-Speech based on EVC

In AL-to-Speech based on basic VC, i.e., ES-to-Speech, EL-to-Speech, and silent EL-to-Speech, it is difficult to recover the speaker individuality of alaryngeal speech owing to the predefined voice quality of converted speech, although the sound quality of the converted speech is improved. To flexibly control the voice quality of converted speech, we apply one-to-many EVC to AL-to-Speech. The EVC technique allows users to manually control the voice quality of the converted speech. Furthermore, conversion models can be adapted using a small number

of utterances of target speech in a text-independent manner. Therefore, provided recorded normal speech data of the laryngectomee before undergoing the total laryngectomy still exist, AL-to-Speech based on EVC can estimate the converted speech that sounds similar to the laryngectomee's original voice. In this section, we describe the AL-to-Speech system assuming that a few utterances of the laryngectomee's original speech have been recorded. AL-to-Speech based on EVC consists of training, adaptation, and conversion processes as mentioned in Section 3.5.

Training process

In the training process, we independently train two one-to-many EV-GMMs for each alaryngeal speech: a one-to-many EV-GMM for estimating the converted spectral feature and a one-to-many EV-GMM for estimating the converted aperiodic components. To train these two EV-GMMs, we use multiple parallel data sets consisting of alaryngeal speech data uttered by the laryngectomee and prestored normal speech data uttered by many non-laryngectomees. The EV-GMM for spectral estimation is trained using multiple joint feature vector sets consisting of the spectral segment features of the laryngectomee and the prestored spectral features of the non-laryngectomees. On the other hand, the EV-GMM for aperiodic component estimation is trained using multiple joint feature vector sets consisting of the spectral segment features of the laryngectomee and the prestored aperiodic components of the non-laryngectomees. In this thesis, the training method for the EV-GMMs is based on speaker adaptive training (SAT) for the EV-GMM [16].

Each EV-GMM is adapted to a new target speaker by adjusting the weight vector so that the marginal likelihood for the given target speech features is maximized see eq. 57. This adaptation process is effective if the speaker-dependent characteristics are well captured by short-term features, such as the spectrum and aperiodic components. On the other hand, it is difficult to control speaker-dependent characteristics captured by long-term features, such as F_0 patterns. Therefore, instead of an EV-GMM, a well-trained speaker-dependent GMM is used to estimate the F_0 patterns from the spectral segment sequence of alaryngeal speech. In AL-to-Speech for ES speech, to develop the GMM for estimating the F_0 patterns corresponding to the perceived pitch information of ES speech,

we use F_0 values extracted from normal speech uttered by a non-laryngectomee in the training as the output features to imitate the prosody of ES speech. To develop a GMM for F_0 estimation in EL speech and silent EL speech, speaker-dependent GMMs are separately trained for all target speakers. Then, the GMM achieving the highest F_0 estimation accuracy is manually selected.

Adaptation and conversion processes

Using the few speech samples uttered by laryngectomees before undergoing a total laryngectomy as adaptation data, the EV-GMM is flexibly adapted to the target voice quality by automatically determining the weight vector in a text-independent manner [5]. The weight vectors of the EV-GMMs for spectral and aperiodic estimation are independently estimated using the spectral features and aperiodic components extracted from the given target speech samples, respectively. The converted spectral feature vectors and aperiodic components are independently estimated using the adapted EV-GMMs. In the F_0 estimation, the global speaker-dependent characteristics of F_0 patterns are controlled in a simple manner. A log-scaled F_0 sequence is first estimated with the selected speaker-dependent GMM, and then further converted so that its mean μ_x and standard deviation σ_x are equal to those of the adaptation speech data, μ_y and σ_y, using eq. 72.

3.7 Experimental evaluations

3.7.1 Evaluations of ES-to-Speech based on VC

To demonstrate the effectiveness of the ES-to-Speech based on VC, we conducted experimental evaluations using several criteria.

Experimental Conditions

We recorded 50 phoneme-balanced sentences of esophageal speech uttered by one Japanese male laryngectomee. We also recorded the same sentences of normal speech uttered by a Japanese male non-laryngectomee, who tried to imitate the prosody of the laryngectomee utterance-by-utterance as closely as possible. The sampling frequency was set to 16 kHz. We conducted a fivefold cross-validation test in which 40 utterance pairs were used for training and the remaining 10 utterance pairs were used for evaluation.

The 0^{th} to 24^{th} mel-cepstral coefficients extracted by STRAIGHT analysis [33] were used as the spectral parameter. As the source excitation features of normal speech, we used log-scaled F_0 extracted by STRAIGHT F_0 analysis [42] and aperiodic components [39] averaged over five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz [43]. The shift length was set to 5 ms.

We optimized several parameters such as the number of mixture components of each GMM and the number of frames used to extract the spectral segment feature to maximize the conversion accuracy of the evaluation data in the cross-validation test [38]. As a result, we set the number of mixture components to 32 for each of the three GMMs. For segment feature extraction, we used the current $\pm$ 8 frames in both the spectral estimation and the aperiodic estimation and the current $\pm$ 16 frames in F_0 estimation.

We conducted both objective and subjective evaluations. In the objective evaluations, to demonstrate the effectiveness of ES-to-Speech, we evaluated the mel-cepstral distortion between the target and converted spectral feature for each phoneme category. Mel-cepstral distortion (Mel-CD) is calculated as

$$Mel-CD\ [dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(X)} - mc_d^{(Y)}\right)}, \tag{75}$$

where $mc_d^{(X)}$ and $mc_d^{(Y)}$ represent the d^{th} dimensional component of the converted mel-cepstrum and that of the target mel-cepstrum, respectively. We also evaluated the estimation accuracy of each parameter. To objectively calculate the conversion accuracy between the converted and target F_0, the unvoiced or voiced decision error (U/V error) and the correlation coefficient only for voiced frames were used. U/V errors were calculated as the rate of the number of target U/V frames and source or converted U/V frames. As conversion accuracy of aperiodic components, the distortion of converted and target aperiodic components was calculated as the root mean square error (RMSE) as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} r_t}{T}} \tag{76}$$

$$r_t = \sum_{d=1}^{D}\left(AP_t^{(X)}(d) - AP_t^{(Y)}(d)\right)^2 \tag{77}$$

where $AP_t^{(X)}(d)$ and $AP_t^{(Y)}(d)$ represent the d^{th} aperiodic components at frame t of the converted data and target data, respectively.

In the subjective evaluations, we conducted two opinion-based tests on intelligibility and naturalness. The following six types of speech sample were evaluated by 10 listeners.

ES recorded esophageal speech

ES-AS analysis-synthesized esophageal speech

EstSpg synthetic speech using converted mel-cepstrum, converted aperiodic components, and extracted F_0 of esophageal speech

EstF_0 synthetic speech using extracted mel-cepstrum, extracted aperiodic components, and converted F_0

CV synthetic speech using converted mel-cepstrum, converted aperiodic components, and converted F_0

NS-AS analysis-synthesized normal speech

Each listener evaluated 120 samples in each of the two tests.

Experimental results

Figure 13 shows the mel-cepstral distortion for each phoneme category. The use of the spectral segment features is effective for improving the spectral estimation accuracy. We can see that the estimation accuracy is considerably improved for some phonemes such as liquids, unvoiced plosives, and voiced fricatives. Although an increase in the number of input frames tends to decrease the mel-cepstral distortion, the use of too many input frames increase the mel-cepstral distortion because the dimension reduction process by PCA starts to remove important acoustic characteristics at each frame to capture complicated acoustic variations over longer segments.

Tables 2 and 3 show the estimation accuracy of the spectrum, aperiodic components, and F_0. It is observed that the distortion between the parameters extracted from esophageal speech and the target parameters of normal speech is

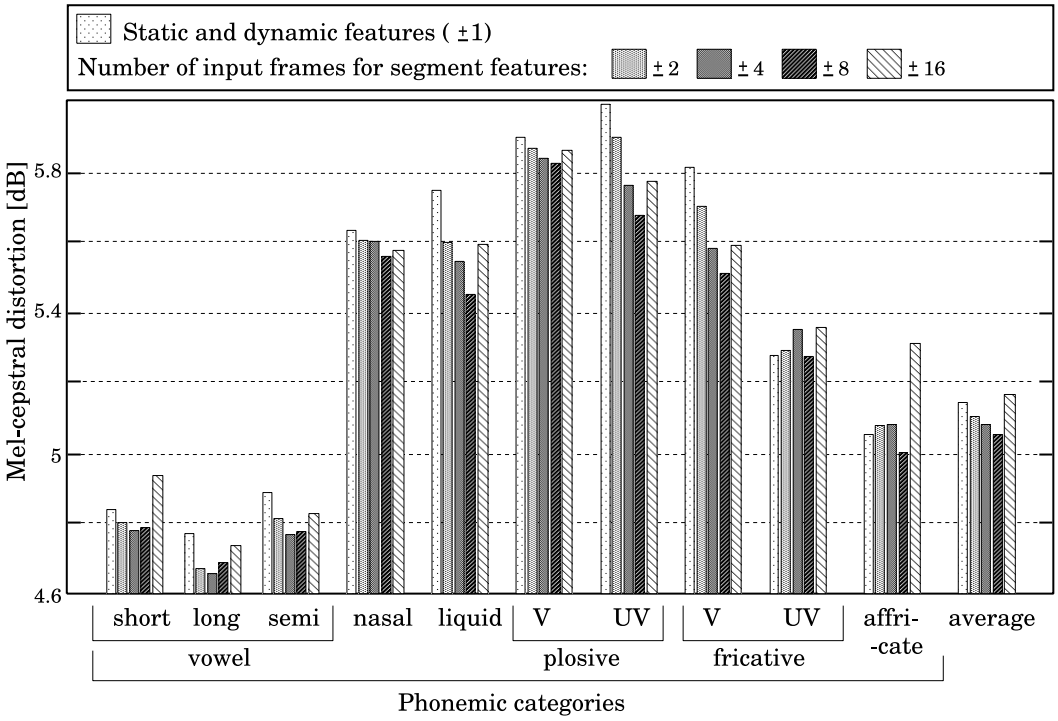

Figure 13. Estimation accuracy of mel-cepstrum on each phonemic category. The notation "V" denotes voiced phonemes, and "UV" denotes unvoiced phonemes.

large, and therefore, the acoustic features of esophageal speech are very different from those of normal speech. These large differences in the acoustic features are significantly reduced by ES-to-Speech. In the F_0 estimation, not only the correlation coefficient between the extracted/converted F_0 and the target F_0 extracted from normal speech but also the unvoiced/voiced decision error is also significantly improved by ES-to-Speech based on VC. These results demonstrate that the proposed conversion method is very effective for improving all the essential acoustic features, i.e., the spectrum, aperiodic components, and F_0.

Figures 14 and 15 show the results of the opinion test on intelligibility and naturalness, respectively. ES-AS causes significant intelligibility degradation compared with ES owing to the difficulty of acoustic feature extraction in esophageal speech. The specific sounds and unstable variations in the spectrogram of esophageal speech are significantly alleviated by using the estimated spectral features (Est-

Table 2. *Estimation accuracy of mel-cepstrum without power and aperiodic components. Mel-cepstral distortion with power (i.e., including the 0^{th} coefficient) is shown in parentheses.*

	Mel-cepstral distortion [dB]	Aperiodic distortion [dB]
Extracted	8.46 (12.95)	6.99
Converted	4.96 (6.26)	3.71

Table 3. *Correlation coefficient (Corr.) between extracted or converted F_0 and target F_0 and unvoiced/voiced (U/V) decision error. Correlation coefficients are calculated using only F_0 values at voiced frame pairs. "VU" shows the rate of estimating voiced frames as unvoiced ones and "UV" shows that of estimating unvoiced frames as voiced ones.*

	Corr.	U/V decision error [%]
Extracted	0.07	43.82 (VU: 42.60, UV: 1.22)
Converted	0.68	8.36 (VU: 4.06, UV: 4.30)

Spg). Moreover, converted speech exhibiting pitch information similar to that perceived in esophageal speech is generated using the estimated F_0 contour (EstF_0). These estimation processes effectively reduce the degradation of intelligibility caused by ES-AS. Although significant improvements in the intelligibility and naturalness of esophageal speech (ES) are not observed when using only one of these estimated features, we can see that the ES-to-Speech (CV) using all the acoustic features yields much more intelligible and natural speech than esophageal speech.

These results suggest that the proposed ES-to-Speech is very effective for improving both naturalness and the intelligibility of esophageal speech.

3.7.2 Evaluations of ES-to-Speech based on one-to-many EVC

Experimental Conditions

To train one-to-many EV-GMMs for ES-to-Speech, we recorded the same sentences as those used in the previous evaluations of normal speech uttered by 30 (22

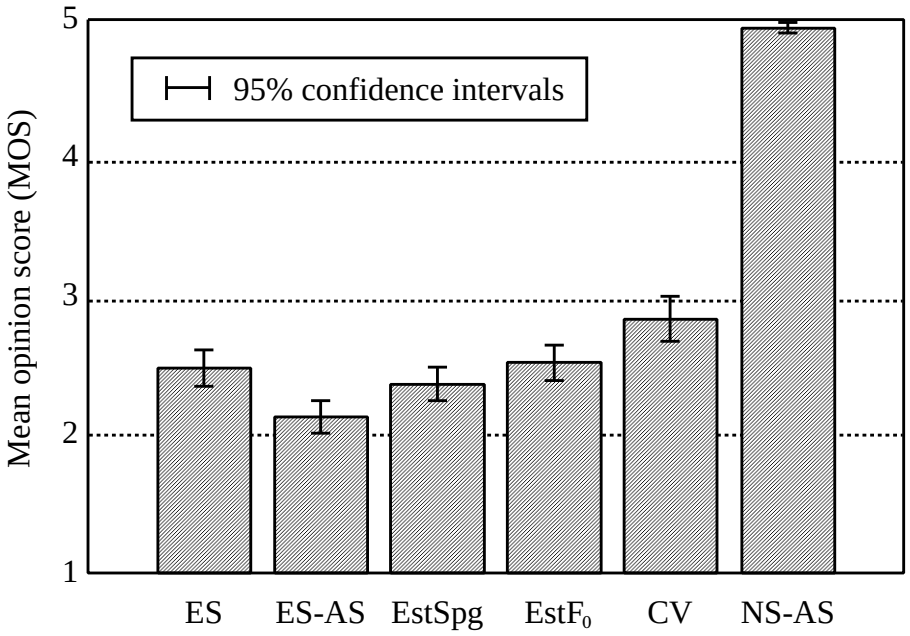

Figure 14. Mean opinion score on intelligibility.

male, 8 female) Japanese non-laryngectomees. We used the same ES speech data as that used in the previous evaluations. To make the automatic time-alignment process for constructing joint feature vectors easier, the non-laryngectomees uttered the sentence being recorded sentences so that the pause positions of the recorded normal speech were the same as those of the ES speech data. The one-to-many EV-GMM for estimation of the spectrum and the MR-GMM for estimation of the aperiodic components were trained using 30 parallel data sets developed by the recorded normal speech data and esophageal speech data. Forty utterance pairs consisting of a single input speech data set of ES speech were used; each of 30 output speech data sets of normal speech were used for training, and the remaining 10 utterance pairs were used for evaluation. The number of eigenvectors was set to 29. The other experimental conditions were the same as those described in Section 3.7.1.

We used the esophageal speech samples as the adaptation data for the EV-GMM and MR-GMM as described in Section 3.5.3. Using the adapted models,

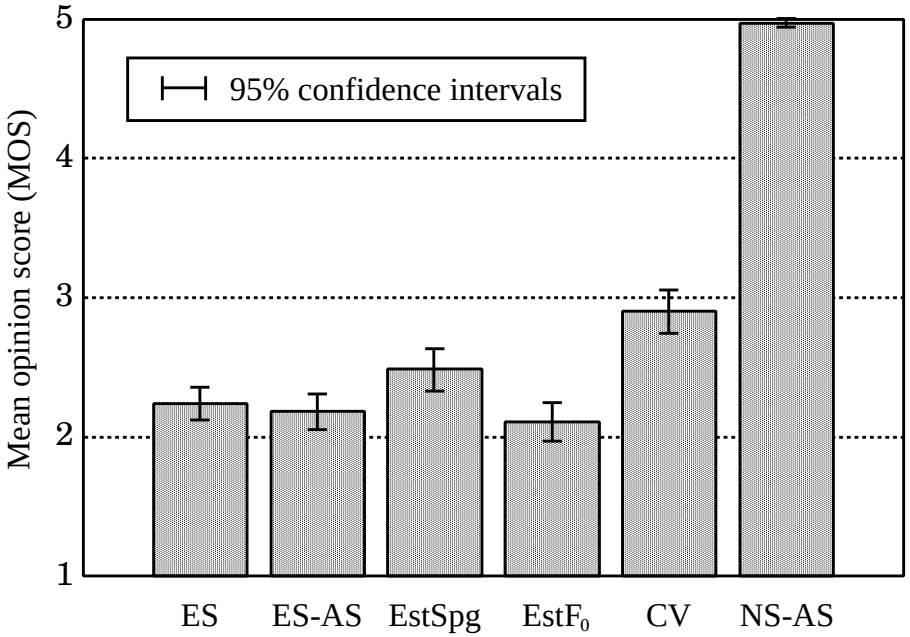

Figure 15. Mean opinion score on naturalness.

the spectrum and aperiodic components were estimated from the spectral segment feature of esophageal speech. For the F_0 estimation, we used the same GMM as that in the previous evaluations. Furthermore, we adjusted the estimated F_0 contour by shifting it so that the converted speech sounded more natural, i.e., more similar to non-laryngectomees' normal voices.

To demonstrate the effectiveness of ES-to-Speech based on one-to-many EVC, we conducted two preference tests on naturalness and intelligibility. In each preference test, a pair consisting of recorded esophageal speech and the converted speech was presented to listeners, who were asked which voice sounded better in terms of naturalness or intelligibility. The number of listeners was 10 and each listener evaluated 20 sample pairs in each test.

Experimental results

Figure 16 shows the results of the preference tests on intelligibility and naturalness. We can see that both the intelligibility and naturalness of esophageal

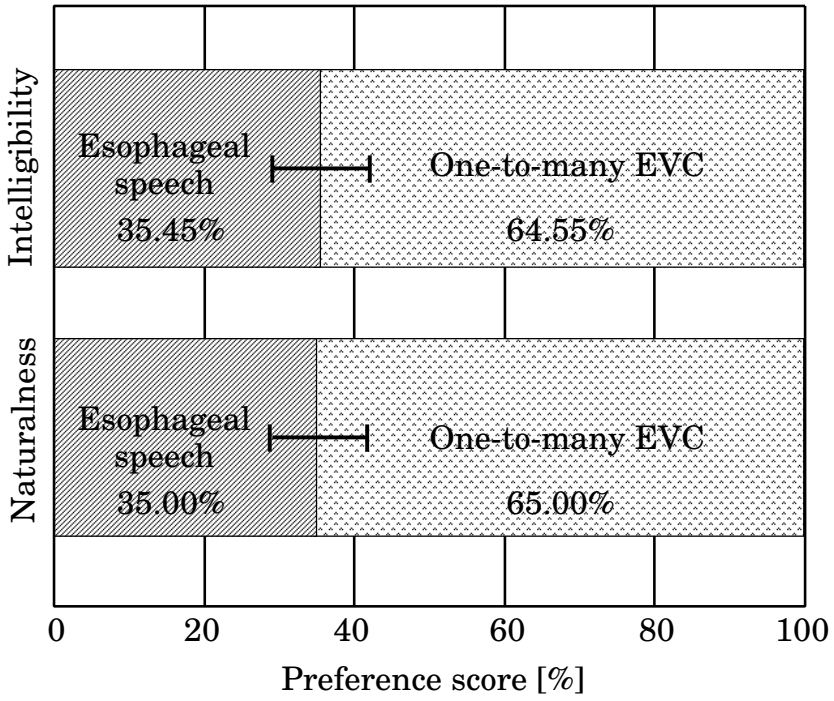

Figure 16. Result of preference tests of intelligibility and naturalness

speech are significantly improved by ES-to-Speech. Therefore, one-to-many EVC is also effective in ES-to-Speech.

Figure 17 shows an example of the spectrogram of each speech sample: a) recorded esophageal speech and b) speech converted by ES-to-Speech based on one-to-many EVC. Furthermore, to demonstrate the effectiveness of voice quality control by ES-to-Speech based on one-to-many EVC, we also show an example of the spectrogram of the converted speech after manipulating part of the adapted weights; we add 3σ to the weight value for the third eigenvector, where σ is the square root of the third principal component. We can see that the proposed conversion significantly reduces the unstable acoustic variations observed in esophageal speech and makes the spectral structures much clearer and similar to those observed in normal speech when the EV-GMMs are adapted to ES speech. Furthermore, we can observe that the spectral structures are changed by the weight manipulation; e.g., the spectral peaks around 4 kHz are shifted to a lower

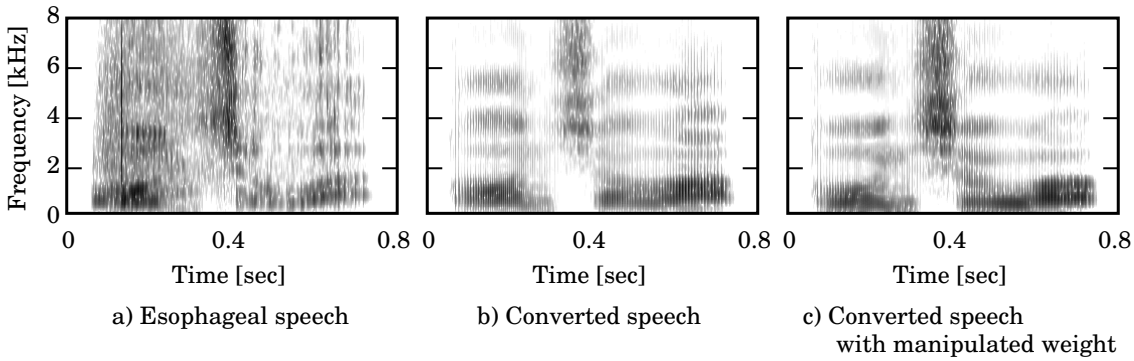

Figure 17. Example of spectrogram of each of a) recorded esophageal speech, b) converted speech by one-to-many EVC using the adapted weights, and c) converted speech by one-to-many EVC when further manipulating a part of the adapted weights for the sentence fragment /h o n sh o h a/ which means "This book" in Japanese.

frequency. In fact, the converted speech samples generated by the spectrograms in b) and c) sound like that voices of very different speakers. Therefore, ES-to-Speech based on one-to-many EVC allows laryngectomees to manually control the voice quality of converted speech.

3.7.3 Evaluations of AL-to-Speech based on one-to-many EVC

To demonstrate the effectiveness of AL-to-Speech based on VC/EVC methods, we conducted experimental evaluations using several criteria. Then, we explicitly indicate the advantage of each alaryngeal speech when applying AL-to-Speech. In this section, each evaluation is conducted assuming that recorded normal speech data of the laryngectomees before undergoing a total laryngectomy still exists. Note that this assumption is different from that in Section 3.7.2.

Experimental conditions

We recorded 50 phonetically balanced sentences of ES speech uttered by one Japanese male laryngectomee, those of EL speech and silent EL speech uttered by another Japanese male laryngectomee, and those of normal speech uttered by each

of 40 Japanese non-laryngectomees. The speech data of 30 non-laryngectomees was used for training and that of the other 10 non-laryngectomees was used as the target data for evaluation. From the 50 recorded sentences of each speaker, 40 were used as the training or adaptation data and the remaining 10 were used as the test data. The sampling frequency was set to 16 kHz.

The 0^{th} to 24^{th} mel-cepstral coefficients were used as spectral parameters. STRAIGHT analysis [33], which involved F_0 adaptive analysis to extract an accurate spectral envelope by effectively removing the effect of the periodicity of F_0 on the spectrum, was employed for normal speech. On the other hand, mel-cepstrum analysis [44] was employed for alaryngeal speech since F_0 for alaryngeal speech is not informative.

As the source excitation features of normal speech, we used log-scaled F_0 values and aperiodic components in five frequency bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz, which were used for generating mixed excitation. The frame shift was 5 ms. To extract the spectral segment feature of ES speech, the current and ± 8 frames were used for spectral and aperiodic component estimation and the current and ± 16 frames were used for F_0 estimation. For EL speech and silent EL speech, the current and ± 8 frames were used for each parameter estimation. These numbers of frames per segment were experimentally optimized [45].

The EV-GMMs for spectral and aperiodic component estimation were trained for each type of alaryngeal speech. The numbers of eigenvectors and mixture components were set to 29 and 64 in every EV-GMM, respectively. The EV-GMMs were adapted to the target speakers using 1, 2, 4, 8, 16, or 32 utterances of their normal speech data. For AL-to-Speech based on VC, the GMMs for spectral and aperiodic estimation were trained using a parallel dataset for each type of alaryngeal speech and the normal speech of each target speaker. The number of training utterance pairs was set to 1, 2, 4, 8, 16, or 32. The number of mixture components was optimized manually and depended on the training data size. Individual speaker-dependent GMMs for F_0 estimation were trained for all 40 non-laryngectomees. The GMM yielding the most natural F_0 pattern was then selected by listening to the converted speech. The same F_0 estimation process was performed for EVC-based AL-to-Speech and VC-based AL-to-Speech.

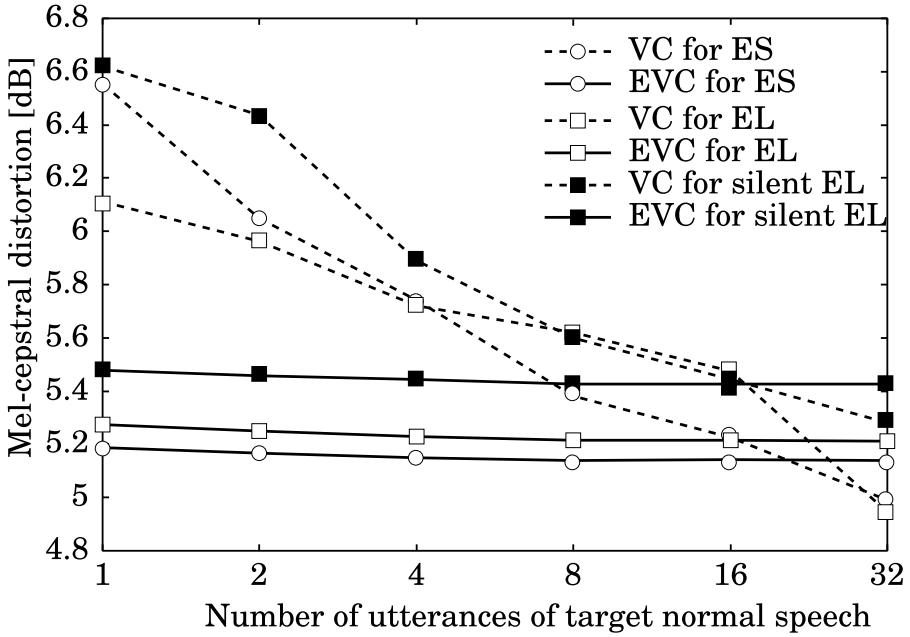

Figure 18. Mel-cepstral distortion as a function of the number of utterances of target normal speech (i.e., utterance pairs in VC or adaptation utterances in EVC).

Estimation accuracy of spectrum and aperiodic components

Figures 18 and 19 show the mel-cepstral distortion and RMSE of aperiodic components as a function of the number of adaptation utterances used in EVC or of the number of utterance pairs used in VC, respectively.

EVC exhibits significantly smaller mel-cepstral distortion and a significantly smaller RMSE than VC for each type of alaryngeal speech enhancement when the amount of target normal speech data is small. Even if only one arbitrary utterance of the target normal speech is available in EVC, its conversion performance is almost equivalent to or better than that of VC using 16 parallel utterance pairs. It is also observed that ES speech yields the highest conversion accuracy and silent EL speech yields the lowest accuracy among the three types of alaryngeal speech.

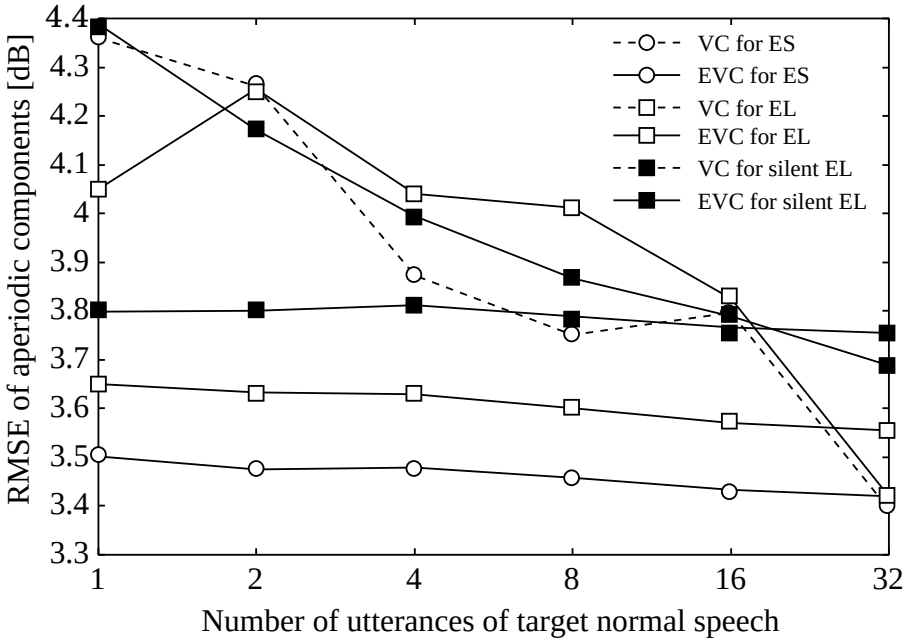

Figure 19. RMSE on aperiodic components as a function of the number of utterances of target normal speech (i.e., utterance pairs in VC or adaptation utterances in EVC).

F_0 estimation accuracy

We also evaluated the F_0 estimation accuracy in AL-to-Speech for each type of alaryngeal speech using the F_0 correlation coefficient and the unvoiced/voiced (U/V) error between the converted speech and target normal speech. To demonstrate the F_0 estimation accuracy for various speakers in AL-to-Speech, the results calculated using individual speaker-dependent GMMs for the 40 non-laryngectomees are shown in Table 4. For ES speech, the results for another non-laryngectomee who uttered normal speech so that its pitch sounded similar to that of ES speech are also shown in the table as "ES pitch." ES speech yields the highest estimation accuracy among the three types of alaryngeal speech. Additionally, the estimation accuracy is significantly improved using the GMM developed using the normal speech, the F_0 patterns of which correspond well to the pitch patterns of ES speech.

Table 4. *F_0 estimation accuracy for various target speakers using corresponding target-speaker-dependent GMMs.*

	Correlation	U/V error [%]
ES	0.58	12.39 ($V \to U : 6.59, U \to V : 5.80$)
EL	0.40	13.20 ($V \to U : 4.92, U \to V : 8.28$)
Silent EL	0.42	14.02 ($V \to U : 6.89, U \to V : 7.13$)
ES pitch	0.68	8.36 ($V \to U : 4.30, U \to V : 4.05$)

Table 5. *F_0 estimation accuracy for actual target speakers in evaluation using well-trained speaker-dependent GMMs.*

ES pitch	0.62	13.88 ($V \to U : 10.70, U \to V : 3.18$)
EL	0.51	12.05 ($V \to U : 7.13, U \to V : 4.92$)
Silent EL	0.45	13.78 ($V \to U : 8.92, U \to V : 4.86$)

The final results for the 10 target non-laryngectomees used as the test data are shown in Table 5. The GMM for "ES pitch" was used in ES speech enhancement, and manually selected speaker-dependent GMMs were used in the EL/silent EL speech enhancement. Namely, the speaker used in the model training was different from the target speakers. It is observed that, for EL speech and silent EL speech, the estimation accuracy of the selected GMMs is higher than that of the various speaker-dependent GMMs shown in Table 4, even though a speaker different from the target speakers was used in the training. To generate a natural F_0 pattern in AL-to-Speech, it is useful to select an optimum speaker for training rather than to directly use the actual target speaker since the F_0 estimation accuracy varies considerably among speakers. It is also observed that ES speech enhancement yields a better F_0 correlation than the other methods.

Opinion tests on speech quality and intelligibility

We conducted opinion tests on speech quality and intelligibility. In these tests, eight listeners evaluated nine types of speech including original alaryngeal speech and ES speech, EL speech, and silent EL speech converted with AL-to-Speech based on VC/EVC. VC-based AL-to-Speech used 32 utterance pairs for GMM training. On the other hand, only one utterance was used as adaptation data

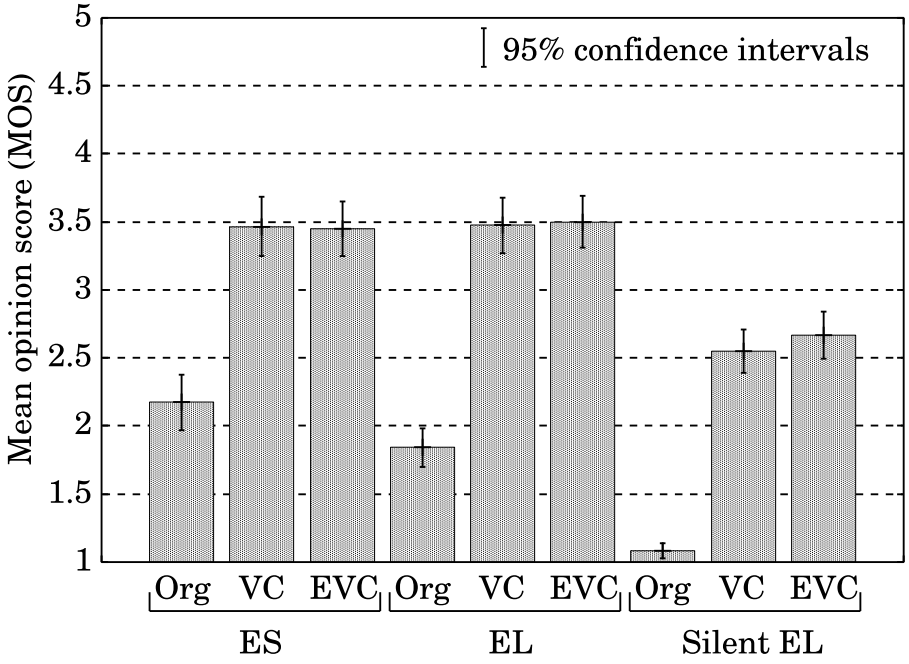

Figure 20. Result of opinion test of speech quality. "Org", "VC", and "EVC" indicate original alaryngeal speech, converted speech by AL-to-Speech based on VC trained with 32 utterance pairs, and converted speech by AL-to-Speech based on EVC adapted with one utterance of target speech, respectively.

for EVC-based AL-to-Speech. The GV was considered in the conversion process. For EVC-based AL-to-Speech, the mean vector of the GV probability density was set to the GV extracted from the adaptation utterance for each target non-laryngectomee and the covariance matrix was fixed to that calculated using the GVs extracted from all utterances of the non-laryngectomees used in the training of the EV-GMM. The opinion score in each test was set to a five point scale. We asked the listeners to evaluate speech samples and assign scores from 1 to 5; i.e., a higher score included higher speech quality or intelligibility. In each test, the individual listeners listened to several speech samples before starting the test to make their own score ranges as stable as possible. Each listener evaluated 135 speech samples in each test.

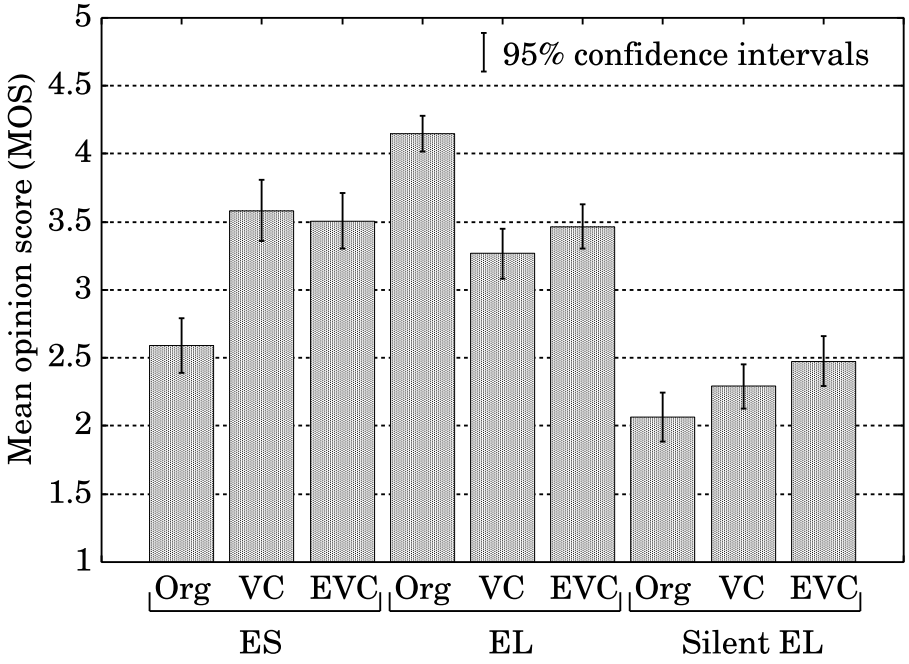

Figure 21. Result of opinion test of intelligibility.

Figures 20 and 21 show the results of the opinion tests of speech quality and intelligibility, respectively. All the AL-to-Speech methods yield significant improvements in speech quality compared with that of the original alaryngeal speech. The speech quality of the enhanced silent EL speech is lower than those of the enhanced ES speech and enhanced EL speech, but it is significantly higher than that of each type of original alaryngeal speech. The intelligibility of ES speech and silent EL speech is also improved by AL-to-Speech. On the other hand, the intelligibility of EL speech is slightly degraded from that of the original EL speech by AL-to-Speech. The enhancements of speech quality and intelligibility by EVC-based AL-to-Speech are almost equivalent to those enhanced by the VC-based AL-to-Speech. Note that the EVC-based method requires only one arbitrary utterance of the target normal speech, whereas the VC-based method requires 32 utterance pairs of alaryngeal speech and the target normal speech.

Table 6. *Result of dictation test*

	Word correct [%]	Word accuracy [%]	Number of replays
ES	87.76	84.3	2.23
EL	92.89	90.93	2.7
Silent EL	66.42	64.71	2.7
ES-EVC	79.90	76.96	2.93
EL-EVC	89.22	87.5	1.9
Silent EL-EVC	84.8	82.84	2.57

Dictation test

We conducted dictation test. In this test, six listeners evaluated six types of speech including original alaryngeal speech and converted speech by AL-to-Speech based on EVC. Only one utterance was used as adaptation data for the adaptation of EV-GMMs. We allowed listeners to replay the same stimulus time after time.

Table 6 shows word correct, word accuracy, and the average number of replays by listeners. Improvement of the word correct and word accuracy of silent EL speech by AL-to-speech based on EVC is significantly larger than that of ES speech. This result shows different tendency from the opinion test on intelligibility shown in fig 21. This is because that low word correct and word accuracy of ES speech are caused by different factor from that of silent EL speech. Although the intelligibility of silent EL speech is low owing to noisy sound caused by body conduction, phonemes and words are correctly uttered because movement of articulatory organs is same to that of EL speech. Therefore, the word correct and word accuracy of silent EL are improved by converting silent EL speech into normal speech. On the other hand, the word correct and word accuracy of ES speech are low because some phonemes and words are not correctly produced owing to difficulty of utterance of ES. Because the proposed method can convert ES speech into normal speech while keeping linguistic information unchanged, incorrect utterance caused by difficulty of producing ES speech does not be alleviated even if converted ES speech can be clearly heard by listener. Therefore, using the proposed method, the word correct and word accuracy of ES speech is not improved

even if its opinion score of intelligibility is improved. Moreover, the word correct and word accuracy of the converted ES and EL speech are a little degraded from those of original ES and EL speech owing to oversmoothing caused by statistical processing. These results suggest that although the proposed method is basically capable of improving intelligibility of alaryngeal speech, which has low intelligibility, by converting alaryngeal speech into normal speech, it is difficult to convert incorrect words into correct words.

Preference test on speaker individuality

We also conducted on XAB test as a preference test to evaluate speaker individuality. In the preference test, six listeners evaluated six types of speech speech consisting of ES speech, EL speech, and silent EL speech converted by VC/EVC-based AL-to-Speech. In this test, listeners heard one target speech sample and two speech samples from among the six types of converted speech, then, they chose the speech sample that had more similar speaker individuality to the target speech sample. The training data used in VC and the adaptation data used in EVC were the same as those used in the opinion tests.

Figure 22 shows the result of the preference test. We can observe the same tendency as that in Fig. 20. Enhanced ES speech yields the greatest speaker individuality and enhanced silent EL speech yields the least speaker individuality among the three types of alaryngeal speech. Even using only one arbitrary utterance of the target speaker in the EVC-based method, its performance is close to that of the VC-based method using 32 parallel utterances of the target speaker. This result shows that the EVC-based method is capable of effectively adjusting the speaker individuality of the converted speech.

Examples of speech converted by AL − to − Speech based on EVC

Figure 23 shows examples of the acoustic features of a target normal speech and the three types of alaryngeal speech converted by AL-to-Speech based on EVC.

These samples were converted from each alaryngeal speech shown in Fig. 23. We can see that the acoustic features of each converted speech were closer to those of the normal speech than to those of each type of alaryngeal speech. In the spectrogram, the spectral structure of the speech converted from ES speech

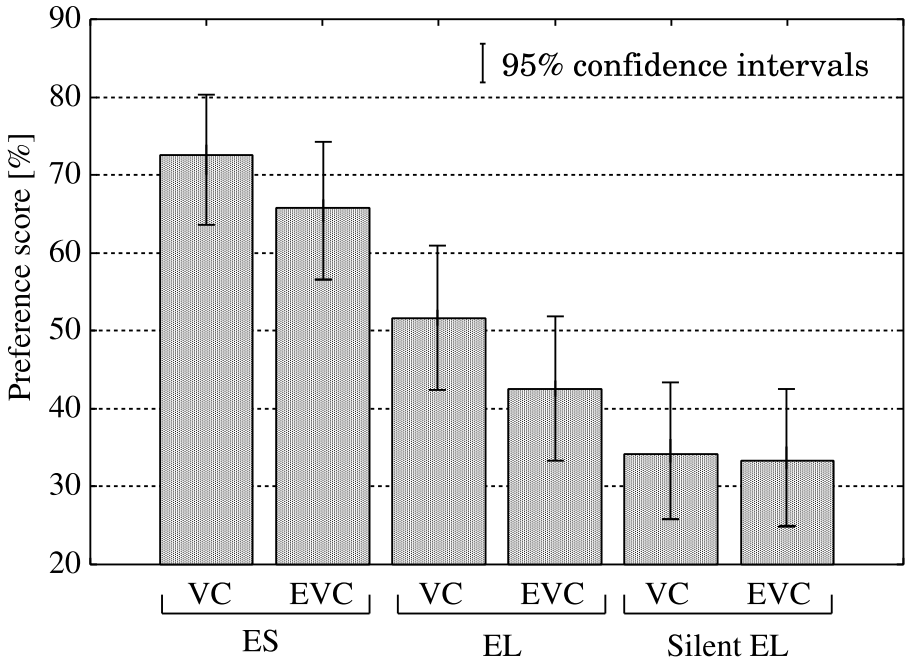

Figure 22. Result of preference test of speaker individuality.

became clearer and stabler than that of the ES speech. Moreover, the spectral structure at high frequencies that could not be observed in EL speech and silent EL speech is observed in the speech converted from EL speech and silent EL speech. F_0 for the speech converted from each type of alaryngeal speech can capture the coarse structure of the normal speech. Furthermore, although over-smoothing occurs, the aperiodic components of the converted speech are similar to those of the normal speech. Therefore, AL-to-Speech based on EVC is effective for the enhancement of alaryngeal speech.

3.8 Summary

To enhance ES speech, this chapter presented an enhancement method based on VC, namely, ES-to-Speech based on VC. This method converts a spectral segment feature of ES speech into a spectrum, aperiodic components, and F_0 for normal speech independently using different GMMs. The experimental results

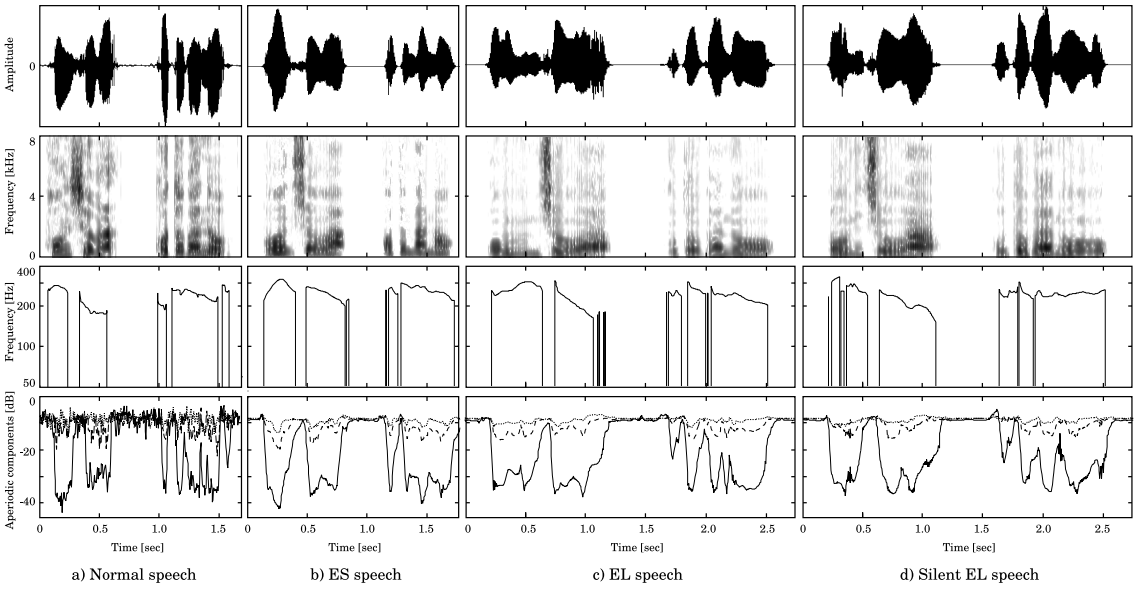

Figure 23. Example of acoustic features, i.e., waveforms, spectrograms, F_0 contours, and aperiodic components of a) normal speech, converted speech by AL-to-Speech based on EVC from b) ES speech, c) EL speech, d) silent EL speech in the same sentence fragment /h o N sy o h a k o t o b a n o/. In aperiodic components, the solid line, coarse broken line, and fine broken line represent low band, middle band, and high band of aperiodic components, respectively.

have demonstrated that ES-to-Speech yields significant improvements in the intelligibility and naturalness of esophageal speech. Moreover, we have also applied one-to-many EVC to ES-to-Speech to flexibly control the voice quality of the converted speech. The effectiveness of ES-to-Speech based on EVC using ES speech as adaptation data has been demonstrated.

We have also applied one-to-many EVC to conventional VC-based enhancement methods, i.e., EL-to-Speech and silent EL-to-Speech. Then, the effectiveness of three types of EVC-based enhancement method were compared by experimental evaluations. The experimental results suggested that 1) the proposed methods significantly improve the speech quality of each type of alaryngeal speech, 2) they also improve the intelligibility of ES speech and silent EL speech,

3) AL-to-Speech based on EVC is capable of effectively adjusting the voice quality of enhanced speech to the target voice quality using only one arbitrary utterance of the target voice, and 4) AL-to-Speech for ES speech is the best in terms of speech quality and speaker individuality. We expect that the proposed enhancement methods for ES speech and EL speech will be particularly effective in mobile speech communication because only the converted speech is presented to the receiver. Moreover, silent EL-to-Speech based on VC/EVC is can be used in face-to-face communication because the source of silent EL speech is not audible to listeners.

4. Singing voice quality control

This chapter describes a new singing voice quality control method using a singing voice conversion system based on many-to-many EVC and training data generation using singing-to-singing synthesis. First, as conventional augmented singing voice generation methods capable of overcoming the limited expression of a singer's own singing voice quality, three types of singing synthesis method and a singing voice conversion method using VC are described. These methods allow a user to generate or sing a song with a different voice quality from that of the user. However, in these methods, it is still difficult to freely control the voice quality of the singing voice. To freely control the voice quality, singing voice conversion based on many-to-many EVC is proposed in this chapter. Although this singing voice conversion based on many-to-many EVC allows the user to freely control the voice quality, the training of the conversion model requires a large amount of training data, a long time, and considerable effort. To alleviate this problem, training data generation using singing-to-singing synthesis is proposed. We conducted objective and subjective evaluations to demonstrate the effectiveness of the proposed methods.

4.1 Introduction

Singing voices are one of the musical instruments which are easily usable for the most people and one of the most important factors to create music expression. Singers can create their own musical expression by controlling the voice quality of their singing voice but the range of the singing voice quality that can be produced by individual singers is limited by physical constraints. It is desirable to be able to achieve new singing styles to produce singing voices beyond what is naturally achievable by singers. The aim of this thesis is to overcome the limitation of voice quality in singing.

To produce a singing voice beyond physical constraints, many approaches have been studied. One of the most popular approaches is the use of singing synthesis systems, which generate a singing voice from several pieces of information such as lyrics and the musical score. Using singing synthesis systems, people can artificially generate humanlike singing voices having different voice qualities by

changing the singing synthesis parameters. Since 2007, many people have started to use singing synthesis systems, such as Vocaloid2 [46] and Sinsy [47], to produce music, and the number of listeners enjoying synthesized singing voices has been increasing. Particularly in Japan, compact discs that include synthesized vocal tracks have often appeared in the popular music chart [48]. In addition, several techniques that can change the voice quality of a user's singing voice or a synthesized singing voice have been proposed [49, 50, 51]. However, the voice quality of the singing voice produced by these singing synthesis systems is fixed in each package. As another approach, a method of morphing between the acoustic features of two singers' singing voices has been proposed [52]. The auditory morphing method allows the user to control the voice quality between the two singers by manipulating the morphing rate parameter. In other words, this method is capable of creating the voice quality of a new singer from the singing voices of two source singers. These systems have enhanced opportunities for musical creation. As a result, many new songs have been created that have brought pleasure to not only music creators but also listeners. Although the singing synthesis and auditory morphing methods allow the user to produce augmented singing voices, it is still difficult to produce singing voices having any voice quality desired by the user.

Generally, these systems are systems for music creators. On the other hand, for singers, a singing voice quality control method using singing voice conversion based on VC has been proposed [8]. This method makes it possible for the singer to directly sing with a different specific voice quality. Statistical VC techniques [3, 14, 15] are used to convert the singing voice quality of a source singer into that of a target singer. In this technique, GMM of the joint probability density of an acoustic feature between the source singer's singing voice and the target singer's singing voice is trained in advance using a special data set, called *a parallel data set*, that consists of pairs of songs the two singers. The trained model is capable of converting the acoustic features of the source singer's singing voice into those of the target singer's singing voice for any song while keeping the linguistic information of the lyrics unchanged. Moreover, real-time singing voice conversion can also be achieved using the low-delay conversion algorithm [4].

Towards realizing a more flexible singing voice conversion technique that en-

ables singers to freely control the converted singing voice quality and is capable of rapidly adapting the conversion model to arbitrary singers, we propose a singing voice conversion method based on two techniques: many-to-many EVC [9] and training data generation using a singing-to-singing synthesis system [10]. Many-to-many EVC is a technique of converting from the voice of an arbitrary source singer into that of an arbitrary target singer. An eigenvoice GMM (EV-GMM) [5] is trained in advance using multiple parallel data sets that consist of a single predefined singer, called a reference singer in this thesis, and many prestored target singers. The EV-GMM is capable of easily adapting the source/target voice quality to that of its given voice samples in a text-independent (lyrics-independent) manner. Using this flexible voice conversion technique, the proposed method enables any singer or user to freely control their singing voice quality. Furthermore, to easily develop multiple parallel data sets from nonparallel singing voice data sets of many singers, we propose a technique for efficiently and effectively generating parallel data sets using a singing-to-singing synthesis system called *VocaListener* to artificially generate voices of the reference singer.

4.2 Conventional methods

This section describes singing synthesis systems and singing voice conversion based on VC as conventional methods. Table 7 shows characteristics of each conventional method.

4.2.1 Singing synthesis systems

Among the methods that allow a user to overcome the limitation of their voice quality, singing synthesis systems are the most popular. In this section, we describe three types of singing synthesis system.

Text − to − singing

In the text-to-singing approach, a singing voice is synthesized from note-level score information of the melody and its lyrics. Vocaloid2 [46] and Sinsy [47] are the most popular systems using this approach in Japan.

Vocaloid2 is a singing synthesis system based on speech synthesis by waveform concatenation. This system generates a singing voice by the concatenation of

Table 7. *Singing synthesis systems and singing voice conversion based on VC.*

	Text-to-singing	Speech-to-singing	Singing-to-singing	singing voice conversion
Voice quality of synthesized singing voice	Fixed depending on system	Source speaker	Fixed depending on system	Target singer
Input	Note-level score of the melody and lyrics	Speaking voice	Singing voice	Singing voice
Singing style	Controllable or fixed depending on system	Fixed depending on system	User's	User's
Main users	Music creators	Music creators	Music creators or Singers	Singers

samples from a prerecorded singer database. The singer database consists of high-quality audio recordings of selected diphone sequences sung uniformly within a specific pitch range. The number of samples is approximately 2000 per pitch. Vocaloid2 allows the user to produce original expressions because it enables a user to adjust singing synthesis parameters such as the pitch and dynamics. However, manual parameter adjustment is not easy and requires considerable time and effort.

Sinsy is a singing synthesis system based on hidden Markov models (HMMs). This system includes training and synthesis processes. In the training process, the spectrum and F_0 are extracted from the singing voice of a singer, then these features are modeled with using context-dependent HMMs. In the synthesis process, first, a musical score and lyrics are converted into a context-dependent label sequence. Next, according to the label sequence, an HMM corresponding to the song is constructed by concatenating the context-dependent HMMs. Finally, the spectrum and F_0 are generated by the HMM and a singing voice is synthesized from these features. Sinsy is easier to use its than Vocaloid2 because the user only has to prepare the musical score and lyrics. However, its sound quality is less than that of Vocaloid2.

In the text-to-singing approach, the voice quality is fixed and depends on the training data. Moreover, these methods do not work in real time.

Speech − to − singing [53]

In the speech-to-singing approach, a speaking voice reading the lyrics of a song is converted into a singing voice upon providing a musical score. In this approach, the spectrum and aperiodic components are first extracted from the speaking voice. Next, the spectrum envelope, aperiodic components, and duration are modified by a spectral control model and a duration control model. Then, F_0 is generated depending on the musical score by an F_0 control model. Finally, a singing voice is synthesized from these modified features. In this method, the users have to prepare speaking voices of the target singer and the musical score. Therefore, it is easy to use this system if speaking voices of the target singer are available. However, it is difficult to freely control the voice quality of synthesized singing voice.

Singing − to − singing [10]

In the singing-to-singing approach, a more naturally sounding singing voice is automatically synthesized by estimating a parameters of the text-to-singing system from a target singing voice. VocaListener [10] is a system used for estimation in singing-to-singing synthesis. VocaListener estimates the parameters of pitch and dynamics for the singing synthesis system so that the synthesized singing voice becomes more similar to the target singing voice. If a user's singing voice and the corresponding lyrics without any score information are available, VocaListener can synchronize them automatically to determine the musical note corresponding to each phoneme of the lyrics. Thus, a singing-to-singing synthesis system allows the users to easily, speedily, and effectively synthesize a singing voice.

4.2.2 Singing voice conversion based on VC

Singing voice conversion is a method based on VC that converts the source singer's singing voice into the target singer's singing voice. This method consists of a

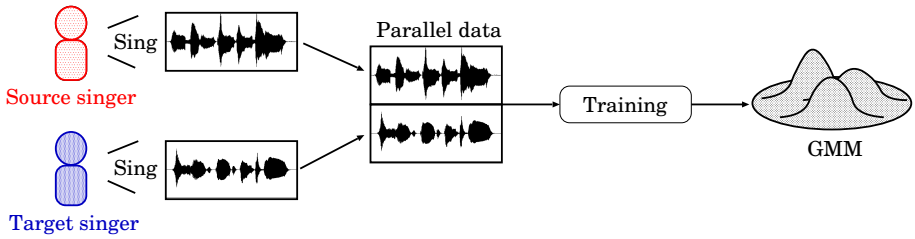

Figure 24. Training process of conventional singing voice conversion.

training process and a conversion process.

Figure 24 shows the training process of the singing voice conversion based on VC. In the training process, singing voice pairs of the source and target singers are first recorded as parallel data for training. Then, a GMM modeling the joint probability densities of acoustic features of the source and target singers' singing voices is trained using the parallel data. In the conversion process, acoustic features extracted from a new singing voice of the source singer are converted into those of the target singer using the trained GMM. Finally, the converted acoustic features are synthesized into the target singing voice.

This method allows the user to sing with the voice quality of the target singer. However, to train the conversion model, a large amount of parallel data is needed for training. Therefore, it is still not convenient for a user desiring to freely control the voice quality.

4.3 Proposed singing voice conversion based on many-to-many EVC

In singing voice conversion based on VC, the voice quality of the converted speech is fixed to that of the target singer. To flexibly control the voice quality of the converted singing voice, we apply many-to-many EVC to singing voice conversion. In this thesis, only spectral feature are converted using the proposed method because the voice quality strongly depends on the spectral feature. F_0 and the aperiodic components of the source singer are directly used to synthesize the converted singing voice.

Figure 25 shows the training process and adaptation process in singing voice conversion based on many-to-many EVC. In the training process, we train one-

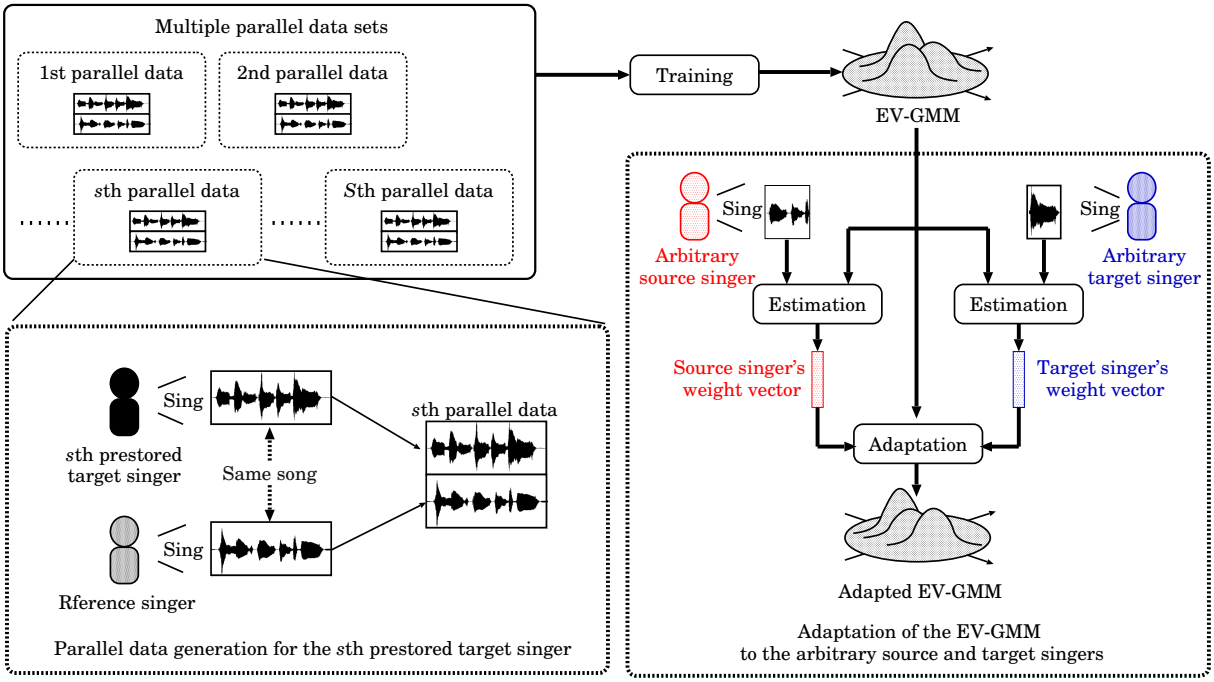

Figure 25. Training and adaptation processes of the singing voice conversion based on many-to-many EVC.

to-many EV-GMM to estimate the converted spectral feature. To train this EV-GMM, we use multiple parallel data sets consisting of the singing voice of the reference speaker (called the reference singer in this chapter) and those of many prestored target singers. The training of the EV-GMM is performed on the basis of SAT.

In the adaptation process, the EV-GMM is adapted to the user and the target singer via their weight parameters, which represent the voice quality of the user and target singer. The user's weight parameter is estimated using the spectral features extracted from the given user's singing voice samples in a text-independent manner. The adaptation method is based on MAP [18] for the EV-GMM [22]. To prepare the target weight parameter, there are two methods. One is the manipulation of weight parameter. Users can freely control the voice quality of converted singing voice by setting an arbitrary score to the weight parameter, although time and effort are required to create the voice quality desired by the user because the value of the weight parameter is not always according to the perception of the

user. The other method is adaptation to the actual target singer in the same manner as adaptation of the source singer. Although a few singing voice samples of the target singer are required in this method, the weight parameter of the target singer is automatically estimated.

On the other hand, the user can use both methods in combination. In this section, we describe two examples of this. Ons is a method that further manipulates the estimated weight parameter. In this method, the user prepares a singing voice sample that has a similar voice quality to the desired voice quality, then, the desired voice quality is created by manipulating the weight parameter estimated from the given singing voice sample. The other method involves linear interpolation between several the weight parameters that are estimated from the singing voices of individual singers. In this method, a new voice quality is easily generated using several weight parameters, making it easy for the user to imagine the voice quality that will be generated.

In the conversion process, the user's singing voice is converted into the target singer's singing voice using the adapted EV-GMM. Note that real-time singing voice conversion is also achieved using the low-delay conversion algorithm [4].

4.4 Proposed training data generation using singing-to-singing synthesis

Although the proposed singing voice conversion method based on many-to-many EVC effectively and rapidly develops a voice conversion model for arbitrary source and target singers, it is necessary to train the EV-GMM in advance using multiple parallel data sets consisting of the singing voice pairs of a single reference singer and many prestored target singers. To prepare a parallel data, the reference singer and target singer have to sing the same song. Note that it is not necessary that all the prestored target singers sing a same song. Therefore, there are two methods of preparing the parallel data sets. One method is for the reference singer to sing several songs and then the many prestored target singers sing the same as the reference singer. In this method, the burden of reference singer is small because the reference singer sings only a few songs. However, it is difficult to collect the singing voices of the many singers singing the same songs. The other

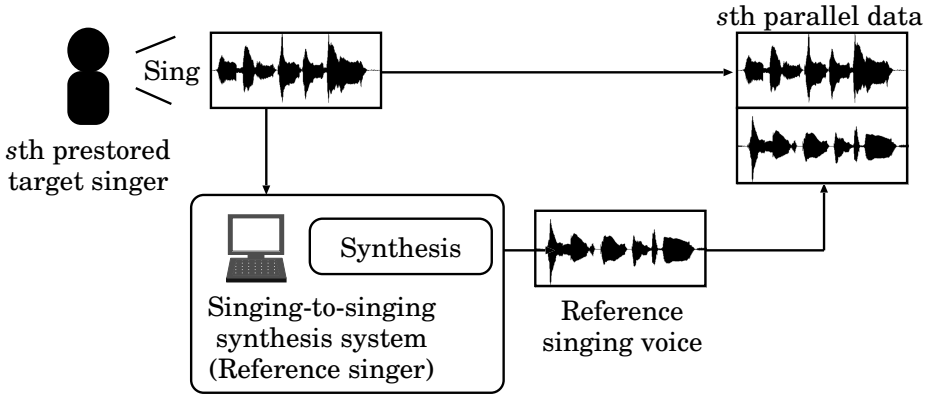

Figure 26. Training data generation using singing-to-singing synthesis.

method is for the reference singer to sing many songs that are independently sung by each prestored target singer. In this method, it is necessary to collect many songs sung by different singers. This has recently comparatively become feasible because there are many songs that are available without permission such as video-sharing web site. Moreover, databases such as the RWC Music Database [54] are available. However, in this method, the burden of the reference singer is large because of the need to sing many songs.

To address this issue, we artificially generate singing voices of the reference singer by applying a singing-to-singing synthesis system to the singing voices of many prestored target singers. In this approach, we only need to prepare the singing voices of many prestored target singers, who need not sing the same song. As shown in Fig. 26, for the singing voices of each prestored target singer, the corresponding singing voices of the reference singer are artificially generated using the singing-to-singing synthesis system. Then, a parallel data set is developed between the singing voices of each prestored target singer and the artificially generated singing voices of the system's singer as the reference singer. Finally the EV-GMM is trained using the developed multiple parallel data sets. Thus, this approach to training data generation can efficiently and effectively develop parallel data sets without recording the singing voices of the reference singer.

In the singing-to-singing synthesis, some acoustic parameters, such as phoneme duration and vibrato, of the synthesized singing voice are automatically tuned so that they are similar to those of the given target singing voice. Therefore, the

synthesized reference singing voice more closely corresponds to the target singing voice than the singing voice of a real singer. Moreover, the voice quality of the singing voices generated by the singing-to-singing synthesis system is more consistent, regardless of the song, genre of music, or the singer's physical condition than that of the singing voice of a real singer.

4.5 Experimental evaluations

This section describes experimental evaluations using several criteria carried out to investigate the following two items: 1) The effectiveness of the singing voice conversion based on many-to-many EVC and training data generation using singing-to-singing synthesis, 2) The effectiveness of the GMM and the EV-GMM trained using a singing voice. It has been reported that the acoustic features of a person's singing voice are significantly different from those of their speaking voice [55]. For example, the spectral envelope of a singing voice has a strong peak, called the singing formant, near 3 kHz. Therefore, a conversion model trained using a speaking voice is not suitable for converting a singing voice because the features converted by this model have the characteristics of a speaking voice even if singing voices are used as adaptation data. On the other hand, it is easier to collect speaking voices than singing voices. Therefore, in this thesis, an EV-GMM trained using singing voices is compared with an EV-GMM trained using a larger numbers of speaking voices for singing voice conversion.

In this evaluation, four types of conversion model are compared.

VC conventional singing voice conversion based on VC

EVC-human singing voice conversion based on many-to-many EVC with conventional training data generation using a human voice as the reference singer's voice

EVC-synth singing voice conversion based on many-to-many EVC with training data generation using singing-to-singing synthesis

EVC-speaking conventional many-to-many EVC for a speaking voice

4.5.1 Experimental conditions

Acoustic features

In this evaluation, only the spectral feature is converted in all conversion methods. The 1^{th} to 24^{th} mel-cepstral coefficients were used as a spectral feature. STRAIGHT analysis [33] was employed to extract these coefficients from singing voices. The shift length was 5 ms and the sampling frequency was 16000 Hz.

Training of EV − GMM

We used the solo singing voices of 30 Japanese songs in the RWC Music Database [54] as the prestored target singing voices. Because the lyrics of songs were all different, the balance of phonemes was not considered. For EVC-human, the solo singing voices of one male singer were used as the singing voices of the reference singer. In EVC-human, parallel data sets between the reference and prestored target singers were automatically aligned by performing DTW. For EVC-synth, singing voices synthesized using the singing-to-singing synthesis system *VocaListener* with a singer database called *Hatsune Miku* [56] based on Vocaloid2 were used as the reference singer. In EVC-synth, the duration of the synthesized reference singing voice was automatically aligned to those of the prestored singing voices by singing-to-singing synthesis. The EV-GMMs used for spectral conversion were trained from 30 parallel data sets consisting of the synthesized or recorded reference singing voices and the prestored target singing voices. The number of basis vectors of the EV-GMMs was set to 29 and the number of mixture components of the EV-GMMs was set to 128.

On the other hand, in EVC-speaking, we used parallel data sets of a single reference male speaker and 152 prestored target speakers to train the EV-GMM. These speakers were from the Japanese Newspaper Article Sentence (JNAS) database. Each prestored target speaker uttered 50 phoneme-balanced sentences included in one of seven subsets. Parallel data sets between the reference speaker and prestored target speakers were automatically aligned by performing DTW. The EV-GMM for spectral conversion was trained from 152 parallel data sets consisting of the recorded reference speaking voices and the prestored target speaking voices. The number of basis vectors of the EV-GMMs was set to 151 and the number of mixture components of the EV-GMMs was set to 128.

Adaptation of EV − GMM and training of GMM

For the adaptation and testing of the EV-GMMs and for the training and testing of the GMM, we selected two Japanese songs from the RWC Music Database (RWC-MDB-P-2001 No.46 and No.76), which were not included in the above 30 songs. Then, five singers (four male singers and one female singer) sang these two songs. Thus, as adaptation/training data and test data, we prepared 10 songs consisting of two songs sung by each singer.

As the training data for the VC-based method and the adaptation data for the EVC-based methods, 2, 4, 8, 16, 32, or 64% of the sung parts of songs sung by the source and target singers was used, then, the remaining 36% of data was used for the test. The GMM and EV-GMMs were prepared for all combinations of source and target speakers. Thus, for each method, 20 conversion models (10 models $\times$ 2 song) were prepared.

The weight parameters representing the voice quality of the source and target singer were independently estimated using the spectral features from the source and target singing voice samples. In this thesis, the adaptation method based on MAP adaptation was performed for the EV-GMMs. The hyperparameter of MAP adaptation was preliminarily optimized in each method. In this evaluation, the hyperparameter of MAP adaptation was set to 250, 1000, and 100 for EVC-human, EVC-synth, and EVC-speaking, respectively.

For VC, we also trained a standard GMM for spectral conversion using a parallel data set consisting of the source and target singing voices. The number of mixture components of the GMM was preliminarily optimized so that the spectral conversion accuracy was maximized in the test data.

4.5.2 Objective evaluation

We evaluated two conditions of song setting: 1) the same-song condition, where the same song is used in both the training/adaptation process and the test process, and 2) the different-song condition, where different songs are used in the training/adaptation process and the test process. The different-song condition is a more realistic situation than the same-song condition because the target singing voice of a song desired by the user is not always available.

To investigate the effectiveness of the proposed methods, we evaluated the

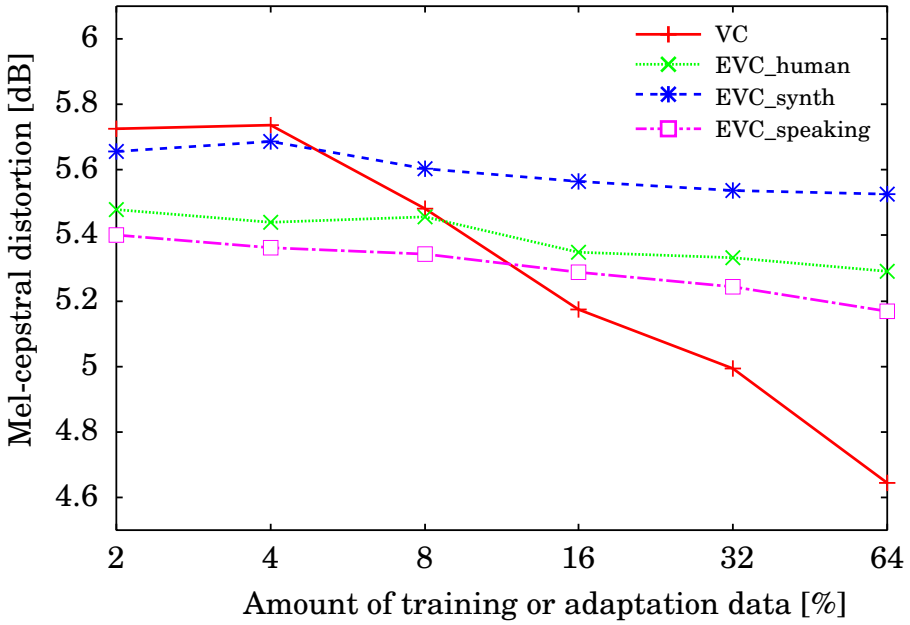

Figure 27. Mel-cepstral distortion as a function of amount of target singing voice data (i.e., singing voice pairs in VC-based method or singing voice adaptation data in EVC-based methods) under the same-song condition.

conversion accuracy of VC, EVC-human, EVC-synth, and EVC-speaking using the mel-cepstral distortion between the converted and target mel-cepstra as an evaluation metric. Figures 27 and 28 show mel-cepstral distortion as a function of the amount of the singing voice adaptation data used in the EVC-based methods or the amount of parallel data of the singing voice pairs used in the VC-based method. Figure 27 show the results for same-song condition and Fig. 28 shows those for different-song condition.

Under the same-song condition, we can see that the EVC-based methods yield higher spectral conversion accuracy than the VC-based method when using a small amount of the available data of the source/target singers. Note that the EVC-based methods do not require the use of parallel data in the adaptation, in contrast to the VC-based method. In the EVC-based methods, EVC-speaking exhibits the highest conversion accuracy. However, the differences in the conversion accuracy from that of EVC-human are not so large even if the amount of training data for EVC-speaking is significantly larger than that for EVC-human.

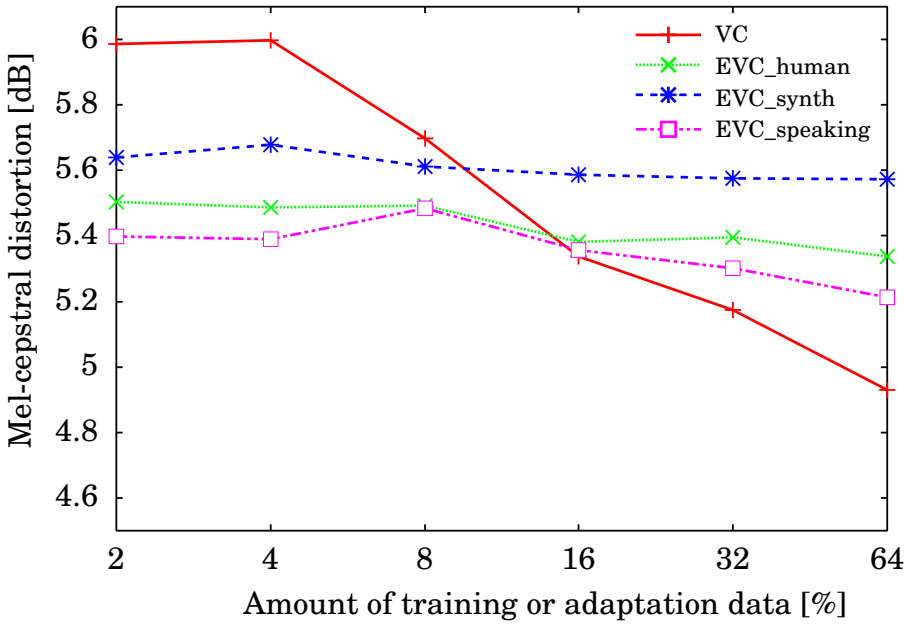

Figure 28. Mel-cepstral distortion as a function of amount of target singing voice data (i.e., singing voice pairs in VC-based method or singing voice adaptation data in EVC-based methods) under the different-song condition.

On the other hand, EVC-synth has lower conversion accuracy than EVC-human.

Under the different-song condition, the VC has much lower conversion accuracy than under the same-song condition. This is because the voice quality of the singing voice of a singer significantly changes depending on the song. On the other hand, it is observed that the EVC-based methods reduce this degradation. Since the EV-GMM is trained with many singers' voices, it is more robust against variations of the singing voice quality.

4.5.3 Subjective evaluation

We conducted an opinion test on the naturalness of the singing voice and a preference test on singer individuality. The opinion was expressed using a five point scale (i.e., 1 (very poor) to 5 (excellent)). In this test, 10 listeners heard 16 types of converted singing voice sample, then they judged the naturalness of each sample using the opinion score. In the preference test, listeners heard a target singing voice sample and two converted singing voice samples, then they chose the

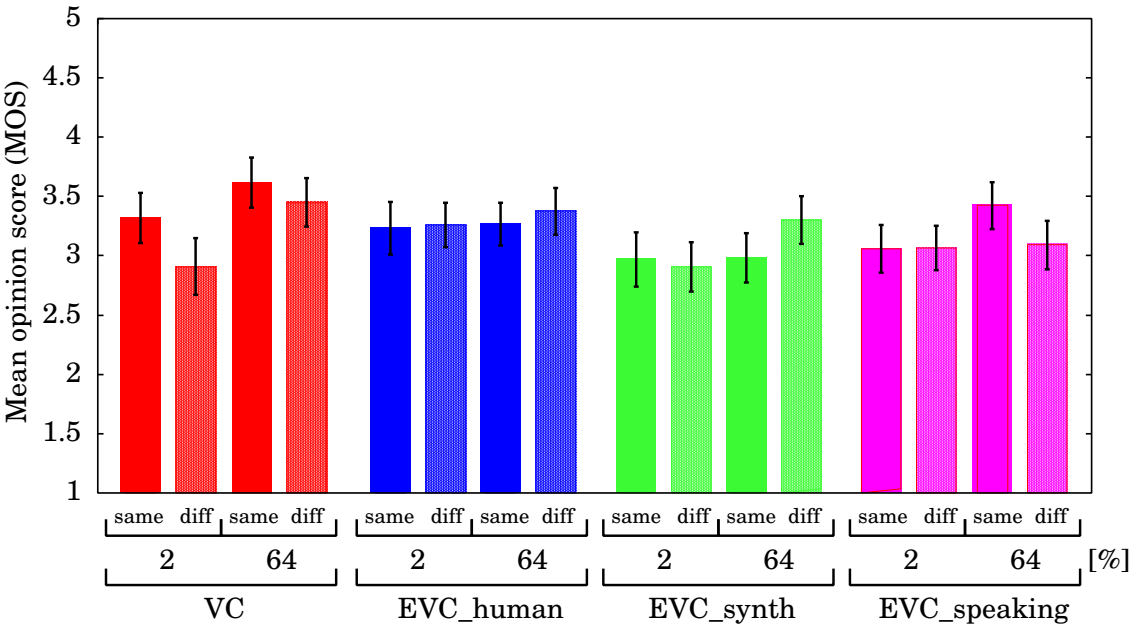

Figure 29. Result of opinion test on naturalness.

converted singing voice sample with more similar singer individuality to the target singing voice sample. The preference test was performed under the different-song condition because of its greater realism than same-song condition. In this tests, nine listeners evaluated eight types of the singing voice generated under the different-song condition for all combinations of 2% or 64% of training/adaptation data and the VC-based method or an EVC-based methods.

Figure 29 shows the result of the opinion test on the naturalness of the singing voice. Under the same-song condition, EVC-human using 2% adaptation data has similar naturalness to VC using 2% parallel training data. On the other hand, VC using 64% parallel training data sounds the most natural. Note that the EVC-based methods do not use the parallel data set of the source and target singers, in contrast to VC. The naturalness of EVC-speaking using 2% adaptation data is slightly lower than that of EVC-human even if a large amount of training data is used in the training process of EVC-speaking. This result suggests that the effect of differences in the acoustic features between a speaking voice and singing voice are sufficiently large to be measurable. On the other hand, the naturalness of EVC-synth is lower than that of EVC-human.

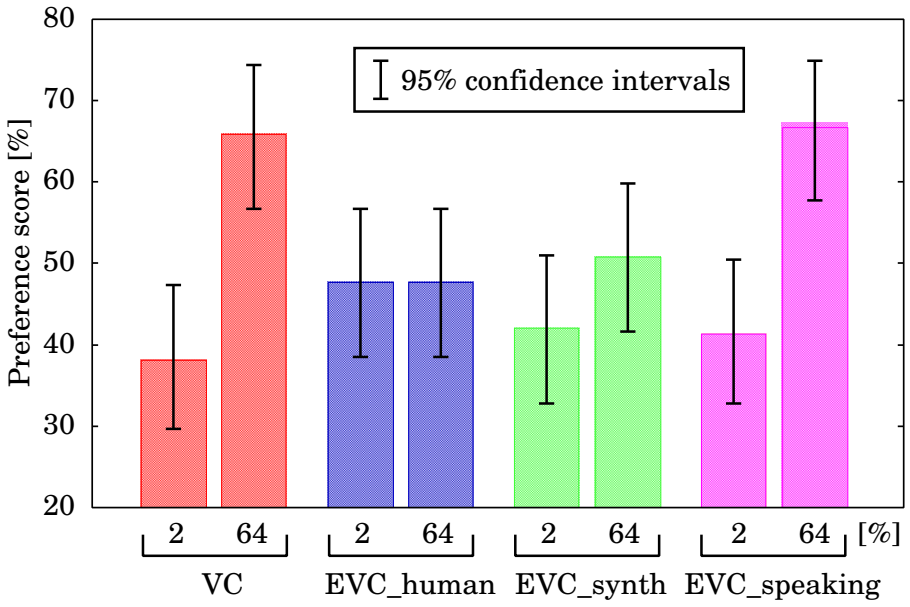

Figure 30. Result of preference test on singer individuality under the different-song condition.

Under the different-song condition, VC using 2% parallel data has lower naturalness than that for EVC-human. This is because the acoustic characteristics of the singing voices are considerably different between different songs. On the other hand, EVC-human and EVC-synth yield almost the same naturalness as those under the same-song condition. The proposed methods using a singing voice as training data are robust against variations of the singing voice quality. Moreover, EVC-synth using 2% adaptation data has similar naturalness to VC using 2% parallel training data. On the other hand, the naturalness of EVC-speaking is slightly lower than that of EVC-human.

Figure 30 shows the result of the preference test on singer individuality. The preference score was calculated as the ratio of the number of samples selected as having better singer individuality to the number of samples presented to the listeners. We can see the same tendency as that observed in the result on the opinion test of naturalness. EVC-human also yields a higher preference score when only 2% adaptation data is available compared with VC using 2% parallel training data. The preference score of EVC-synth using 64% adaptation data is

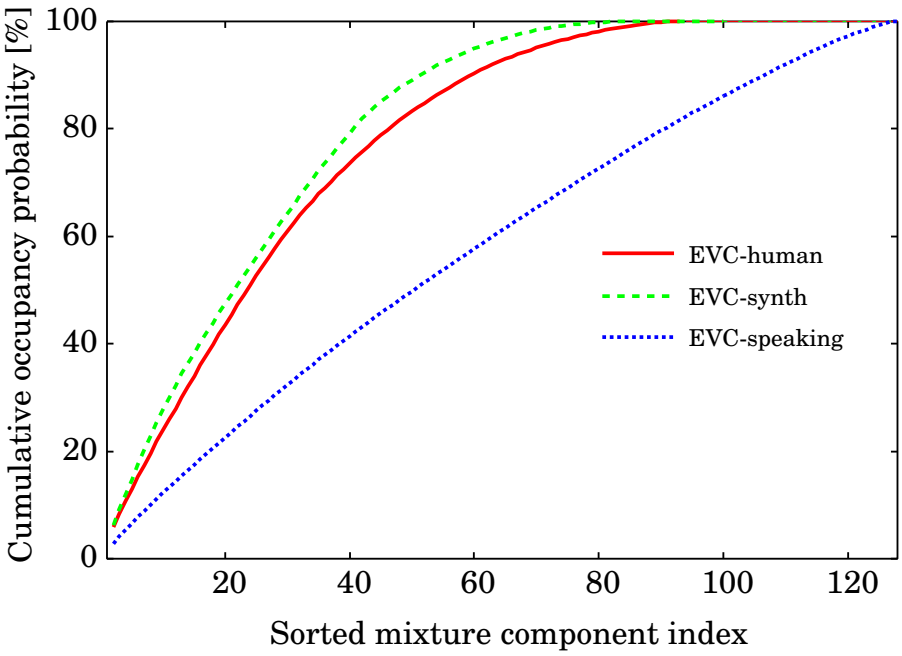

Figure 31. Cumulative occupancy probability for all parallel data set using several models.

similar to that of VC using 64% parallel training data. However, EVC-speaking using 2% data has a lower preference score than that for EVC-human even if the amount of training data for EVC-speaking is significantly larger than that for EVC-human. On the other hand, EVC-synth using 2% adaptation data has a similar performance to VC using 2% adaptation data.

4.5.4 Comparison of each EV-GMMs

In this section, we compared each EV-GMMs trained with speaking voice, singing voice, and synthesized singing voice. Figure 31 shows the cumulative distribution of occupancies of the canonical EV-GMM of EVC-human, EVC-synth, and EVC-speaking. These individual mixture component occupancies have been calculated by eq. (45) from all parallel data set in training process with SAT. In this figure, we can see that the occupancies of EVC-human and EVC-synth are more biased than that of EVC-speaking. Then, the occupancies of EVC-synth are the most biased. On the other hand, average Kullback-Leibler divergences (KLD), calculated

from average occupancy distribution of all prestored target singer and occupancy distribution of individual prestored target singer, of EVC-human, EVC-synth, and EVC-speaking are 0.37, 0.31, 0.05, respectively. These results show that although the EV-GMM needs to model wide varieties of acoustic features of all prestored target speakers, some mixture components of EVC-human and EVC-synth model only acoustic features of a part of prestored target speakers. Consequently, phonemic information and speaker individuality were not separated well in EVC-human and EVC-synth. This issue causes degradation of controllability of voice quality as shown in Fig. 30. On the other hand, Fig. 30 shows that EVC-speaking which separates phonemic information and speaker individuality well needs a large amount of adaptation data to control the voice quality. This result suggests that speaker individuality represented by eigenvoices of EVC-speaking is different from those of the singing voice. Therefore, to more correctly control the voice quality, constructing parallel data set considering phoneme balance and other information is needed for training EV-GMM.

The above results suggest that 1) the proposed methods are capable of effectively using nonparallel data of the source and target singers to rapidly develop a conversion model between the singers, 2) since the proposed methods are robust against variations of the singing voice quality often observed between different songs, they work reasonably well even when different songs are used in the adaptation and conversion processes, 3) although the conversion performance of EVC-synth is slightly lower than that of EVC-human, training data generation with singing-to-singing synthesis significantly reduces the burden of training data generation, 4) to convert a singing voice, it is more desirable to train an EV-GMM from a singing voice than from a speaking voice.

4.6 Summary

To make it possible for singers to produce augmented singing voices beyond their physical constraints, we have proposed a singing voice quality control method using singing voice conversion based on many-to-many EVC. Moreover, to reduce the burden of preparing training data, we have also proposed a method of training data generation using singing-to-singing synthesis.

Although, conventional singing voice conversion based on VC is capable of

converting the acoustic features of the source singer's singing voice into those of the target singer's singing voice in any song while keeping the linguistic information of the lyrics unchanged, it is difficult for a user to freely control the voice quality of converted singing voice. Moreover, the user has to prepare a large amount of training data when using conventional method.

Towards realizing a more flexible singing-voice conversion technique that enables singers to freely control the converted singing voice quality and is capable of rapidly adapting the conversion model to arbitrary singers, we have applied many-to-many EVC to singing voice conversion. Many-to-many EVC is a technique of converting the voice of an arbitrary source singer into that of an arbitrary target singer. Singing voice conversion based on many-to-many EVC has significantly improved usability. On the other hand, to construct the conversion system, a large amount of training data is needed in the training process of the singing voice conversion based on many-to-many EVC. To alleviate this problem, we applied singing-to-singing synthesis to training data generation in the singing voice conversion based on many-to-many EVC. This allows us to easily prepare a large amount of training data.

We have evaluated the effectiveness of the proposed methods objectively and subjectively, assuming that the singing voice of the target singer is available. These experimental results have demonstrated the singing voice conversion based on many-to-many EVC is very effective for increasing usability while maintaining a high conversion performance. Moreover, training data generation with singing-to-singing synthesis significantly reduces the burden of preparing parallel data although the conversion performance is slightly degraded. On the other hand, the proposed method is capable of controlling the singing voice quality by manipulating the weight parameter without a target singing voice. Thus, it is expected that a singing voice converted by the proposed method with manipulation of the weight parameter will have similar naturalness to the results of evaluations. Therefore, the proposed method allows singers to freely control their own voice quality beyond their physical constraints.

5. Conclusion

5.1 Summary of thesis

The voice quality of speakers and singers is limited by physical constraints in the speech production mechanism. The aim of this thesis has been to overcome such physical constraints to make it possible for people to speak with a voice quality that cannot be achieved in a manner of natural speech.

Two problems caused by physical constraints have been dealt with. One problem is vocal disorders caused by a total laryngectomy. Although laryngectomees can speak using an alternative speaking method such as the esophageal speaking method (ES) and electrolaryngeal speaking method (EL), the sound quality, naturalness, and speaker individuality of the produced alaryngeal speech are significantly lower than those of normal speech uttered by non-laryngectomees. Therefore, laryngectomees are subject to stronger physical constraints in speech production than non-laryngectomees. The other problem is limited expression in singing. Although, singers can change their voice quality to sing expressively, the variety of singing voice quality that can be produced by individual singers is limited owing to physical constraints. It is desirable to achieve new singing styles to produce more expressive singing voices beyond the naturally achievable varieties. To address these problems, new alaryngeal speech enhancement and singing voice quality control methods based on statistical voice conversion (VC) and eigenvoice conversion (EVC) have been proposed as techniques to augment speech production beyond physical constraints in this thesis. The alaryngeal speech enhancement method based on EVC is capable of increasing the naturalness and recovering the speaker individuality of alaryngeal speech by converting alaryngeal speech into arbitrary normal speech. The singing voice quality control method using singing voice conversion based on EVC allows singers to produce a singing voice with the desired voice quality singers want.

In Chapter 2, the VC and EVC used in this thesis were described, which employed a conversion method based on the maximum likelihood estimation (MLE) of speech parameter trajectories considering the global variance (GV). This trajectory-based VC is one of the state-of-the-art VC methods. This method consists of a training process and a conversion process. In the training process,

as the conversion model, GMM is trained using parallel data that consists of a single source speaker's voice and a single target speaker's voice. The probability density of the GV, which is defined as the variance of features over one utterance, of the target speaker is also modeled by a GMM. In the conversion process, source feature sequences are simultaneously converted into target feature sequences over an utterance. Then, the problem that features converted based on MLE tend to be oversmoothed is alleviated using the trained GV model. We also explained the EVC frameworks. In EVC-based conversion, there are three types of conversion method: one-to-many EVC, many-to-one EVC, and many-to-many EVC. We explained the overall EVC framework by using one-to-many EVC framework. In this framework, the eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets consisting of the predefined source speaker, called the reference speaker, and various prestored target speakers. Next, the canonical EV-GMM, which is target-speaker-independent model, is capable of being flexibly adapted to a new target speaker using only a few arbitrary utterance of the target speaker. Finally, an arbitrary utterance of the source speaker's voice is converted into that of target speaker's voice using the adapted EV-GMM. In this chapter, we also explained many-to-many EVC, which is a technique that converts an arbitrary target speaker's voice into another arbitrary target speaker's voice. Many-to-many EVC simply involves many-to-one EVC and one-to-many EVC. Moreover, to increase the conversion accuracy, the mixture component sequence is shared in both of types of EVC.

In Chapter 3, we described the enhancement of alaryngeal speech. First, we explained detail of laryngectomees, whose vocal cords have been removed by a total laryngectomy, and alaryngeal speech, which is generated by an alternative speaking method. Although alaryngeal speech allows laryngectomees to speak using residual organs or medical devices instead of vocal cords, the naturalness and speaker individuality are lower than those of normal speech uttered by non-laryngectomees. Esophageal speech (ES speech) and electrolaryngeal speech (EL speech) are the most popular alaryngeal speech in Japan. Next, we explained conventional enhancement method for ES speech and EL speech. Moreover, we also explained alternative speaking method using a small-power sound source unit, which is generate less audible sound source signals, called silent EL. To

enhance ES speech, we applied the VC-based enhancement method used for EL speech. This method makes it possible for laryngectomees to speak with a more natural voice than ES speech. To recover the speaker individuality of ES speech, we applied many-to-many EVC to ES-to-Speech. Moreover, we also applied the EVC-based enhancement method to EL speech and silent EL speech. By carrying out objective and subjective evaluations, we clarified the effectiveness of the proposed enhancement methods for ES speech, EL speech, and silent EL speech.

In Chapter 4, we described method that is capable of controlling voice quality of singing voice. First, we explained singing synthesis systems and singing voice conversion based on VC as conventional methods that overcome physical constraints in singing voice production. Although these conventional methods allow a user to sing or generate a singing voice with voice quality different from that of the user, the voice quality is basically fixed and depend on systems. To freely control the voice quality of a singing voice, we applied many-to-many EVC, which convert an arbitrary speaker's voice to another arbitrary speaker's voice, into singing voice conversion. In this method, an EV-GMM is trained in advance using multiple parallel data sets that consist of a single predefined singer, called a the reference singer, and many prestored target singers. Furthermore, to easily develop multiple parallel data sets from nonparallel singing voice data sets of many singers, we proposed a technique for efficiently and effectively generating parallel data sets using singing-to-singing synthesis to artificially generate singing voices of the reference singer. Objective and subjective evaluations demonstrated the effectiveness of our proposed methods.

In summary, AL-to-Speech based on EVC recovers the sound quality and speaker individuality of alaryngeal speech, which are lacking in original alaryngeal speech owing to physical constraints. On the other hand, singing voice conversion based on EVC allows a singer to freely control the voice quality beyond limitation of the singer's own voice quality. Therefore, the proposed methods make it possible for user to produce speech sound beyond physical constraints.

5.2 Future work

This thesis focused on two problems caused by physical constraints in speech production. To address these issues, we have proposed new alaryngeal speech enhancement and singing voice quality control methods based on statistical VC frameworks. However, several problems remain.

5.2.1 Enhancement of alaryngeal speech

To enable the practical use of AL-to-Speech based on one-to-many EVC, there are still problems to be solved, for example, the development of a user-friendly interface for voice quality control, the development of technologies for flexibly training the conversion models using a large amount of existing normal speech data, and further improvement of the conversion performance. The development of an interface for AL-to-Speech is particularly important. AL-to-Speech for ES and AL-to-Speech for EL are effective for telecommunication, which can convey only converted speech to listeners. Therefore, we need to develop an interface for our proposed systems for use with mobile phone such as smartphones. Moreover, the reduction of environmental noise is needed in practical daily-life situations because laryngectomees do not always use these systems in silent spaces.

AL-to-Speech for silent EL has many more problems than that for ES and EL. The acoustic features of silent ES are strongly affected by noise caused by the speaker's movement. This must be suppressed to enable the use of AL-to-Speech for silent EL in practical situations. On the other hand, the design of the system is also important. To use the system in daily life, the user must carry several items, i.e., a NAM microphone, amplifier, speaker, and the computer installed the system. Therefore, the development of a device suitable for use in daily life is required for each items.

On the other hand, it is expected that the number of patients undergoing a total laryngectomy well decrease with the popularization of SCL-CHEP discussed in Section 3.2. Therefore, we need to consider a method for enhancing alaryngeal speech of people who have undergone SCL-CHEP.

5.2.2 Singing voice quality control

In singing voice conversion based on VC/EVC, there are also many problems to be solved. An important problems is that the lyrics suitable for training conversion models in singing voice conversion were not been clarified. It has been reported that voice conversion methods for a speaking voice have high conversion accuracy when phoneme-balanced sentences are used as training data. However, for a singing voice, phoneme-balanced lyrics have not yet been clarified. Moreover, it is expected that several factors affect the acoustic features of a singing voice such as the genre of music, singing style, and singing skill of singer. To increase the accuracy of singing voice conversion, we need to investigate the effect of these factors and determine which lyrics and music should be used as training data.

As other approach, it is possible that speaking voices can be used to train conversion models for singing voices. Although the acoustic features of a speaking voice are different from those of a singing voice, it is easier to collect speaking voices than singing voice for use as training data. Therefore, if the acoustic features of speaking voice can be changed into those of a singing voice, the preparation of training data will become easier and the conversion accuracy will increase when a large amount of training data. To realize this, we need to investigate the differences between acoustic features of speaking and singing voices in singing voice conversion.

Acknowledgements

I would like to express my appreciation to Professor Satoshi Nakamura of Nara Institute of Science and Technology, my thesis advisor, for his guidance through my doctoral course.

I would also like to express my gratitude to Professor Kiyohiro Shikano of Nara Institute of Science and Technology for their invaluable comments to the thesis.

I would like to particularly express my deepest appreciation to Associate Professor Tomoki Toda of Nara Institute of Science and Technology, my thesis advisor, for his constant guidance and considerate encouragement through my master's and doctoral course. His great efforts for voice conversion play the most important key technique in this thesis. This work could not have been accomplished without his skills and carefully designed exact direction of this research.

I would sincerely like to express my appreciation to Dr. Masataka Goto, who is a prime senior researcher and leader of the Media Interaction Group of National Institute of Advanced Industrial Science and Technology, for his guidance and encouragement. I would also like to express my gratitude to Dr. Tomoyasu Nakano, who is a research scientist of the Media Interaction Group of National Institute of Advanced Industrial Science and Technology, for his guidance and considerate.

I would like to thank Associate Professor Hiroshi Saruwatari and Assistant Professor Hiromichi Kawakami of Nara Institute of Science and Technology, for their beneficial comments.

I would like to thank Dr. Keigo Nakamura, who is a software engineer at Rakuten, Inc. for his valuable advice on my work. I would also like to thank Dr. Yamato Ohtani, who is a research scientist of the TOSHIBA CORPORATION, for his valuable advice on my work.

I want to thank all members of Speech and Acoustics Laboratory in Nara Institute of Science and Technology for providing fruitful discussions. I want to thank all members of Augmented Human Communication Laboratory in Nara Institute of Science and Technology for providing fruitful discussions.

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

Appendix

A. Development of application for singing voice quality control

In this appendix, we describe development of application for singing voice quality controlling. This application has been developed as the system that is used by singer to enjoy singing. Figure 32 shows our developed application.

A.1 Basic function

This application has several functions as follows.

Recording

By clicking "録音" button, recording from microphone is started. And then, by clicking "録音" button again, recording is ended.

Adaptation

By clicking "適応" button, weight parameter for eigenvoices is estimated from singing voice selected by user, and then, estimated weight parameter is stored.

Selection of weight parameter of user

By clicking "某" button, user select own weight parameter from stored weight parameters, and then, EV-GMM is adapted with selected weight parameter.

Voice conversion

By clicking "変換" button, voice conversion is started. And then, by clicking "変換" button again, voice conversion is ended.

Selection of BGM

By clicking "楽曲選択" button, user selects back ground music (BGM), that is played while voice conversion works.

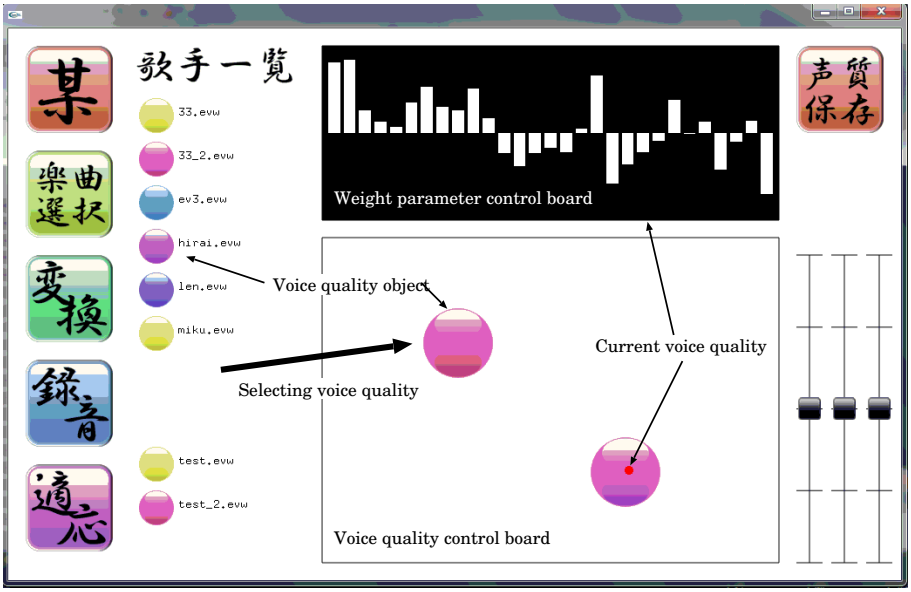

Figure 32. *Application for singing voice quality control.*

Voice quality control

By selecting circular button called voice quality object, which represent voice quality, i.e., weight parameter of eigenvoices, from "歌手一覧", user change voice quality of converted singing voice. Moreover, by drag bar which represent weight of each dimension of weight parameter, in weight parameter control board, user directly control weigh parameter. Detail of way of voice quality control is written later.

Storing weight parameter

By clicking "声質保存" button, directly controlled weight parameter shown in weight parameter control board is stored. And then, stored weight parameter is showed in "歌手一覧" as voice quality object.

A.2 Singing voice quality control

We describe way of singing voice quality control. There are several ways to control voice quality in this application. One is the control way that directly manipulates weight parameter on weight parameter control board in detail. This way allows user to create any voice quality that can be expressed by trained EV-GMM. However, because weight parameter of eigenvoices does not correspond to perception of human, it needs long time and effort of user to create voice quality that user wants. Thus, this way is effective for preinclination of voice quality.

The other is the control way that selects voice quality object. In this way, user right clicks voice quality object and drags this to voice quality control board. Any numerical voice quality object can be set to an arbitrary place of voice quality control board. When singing with voice conversion, user clicks voice quality object on voice quality control board, and then, voice quality of converted singing voice is changed into selected voice quality. Moreover, voice quality of converted singing voice can be controlled using several voice quality objects. When place that there is not voice quality object on voice quality control board is clicked, current weight parameter is calculated from weight parameter of each voice quality objects on voice quality depending on distance between clicked point and each voice quality objects on voice quality control board. Therefore, new voice quality is created by mixing voice quality of each voice quality object. These ways are easy for user to control voice quality while singing.

A.3 Condition of this application

In this application, only spectral feature is converted. The 1^{th} through 30^{th} mel-cepstral coefficients were used as a spectral parameter. STRAIGHT analysis [33] was employed to extract these coefficients from singing voices. The shift length was 5 ms and the sampling frequency was 44100 Hz. We used solo singing voices of 30 Japanese songs in the RWC Music Database [54] as the prestored target singing voices. Lyric of these songs is different from each other. And then, balance of phonemes is not considered. Singing voices synthesized using the singing-to-singing synthesis system *VocaListener* with a singer database called *Hatsune Miku* [56] based on Vocaloid2 are used as reference singing voice. This

EV-GMM for spectral conversion was trained from 30 parallel data sets consisting of the synthesized or recorded reference singing voices and the prestored target singing voices. The number of basis vectors of the EV-GMMs was set to 29. The number of mixture components of the EV-GMMs was set to 128. In conversion process, we employed low-delay VC [4] based on MLE to convert voice quality in real time.

References

[1] H. fujisaki, *Prosody, models, and spontaneous speech, Computing Prosody: Computationaly models for processing spontaneous speech*, Springer, 1997.

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

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[4] 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*, pp. 1076–1079, Sept. 2008.

[5] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, pp. 1249–1252, Apr. 2007.

[6] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Evaluation of extremely small sound source signals used in speaking-aid system with statistical voice conversion," *IEICE Trans. Inf. and Syst.*, vol. E93-D, no. 7, pp. 1909–1917, July 2010.

[7] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) Recognition," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 1, pp. 1–8, Jan. 2006.

[8] Y. Kawakami, H. Banno, and F. Itakura, "GMM voice conversion of singing voice using vocal tract area function," *IEICE technical report. Speech 110(297) (Japanese edition)*, pp. 71–76, Nov. 2010.

[9] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," *INTERSPEECH*, pp. 1623–1626, Sept. 2009.

[10] T. Nakano and M. Goto, "VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation," *Proc. SMC 2009*, pp. 343–348, May 2009.

[11] 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, Mar. 2008 (in Japanese).

[12] S. Kawamoto, Y. Adachi, Y. Ohtani, T. Yoshikuwa, S. Morishima, and S. Nakamura, "Scenario speech assignment technique for instant casting move system," *ACCV2009 Invited workshop on Vision Based Human Modeling and Synthesis*, Sept. 2009.

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

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

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

[16] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Adaptive training for voice conversion based on eigenvoices," *IEICE Trans. Inf. and Syst.*, vol. E93-D, no. 6, pp. 1589–1598, June 2010.

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

[18] 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.

[19] 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, June 2000.

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

[21] 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.

[22] D. Tani, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Maximum a posteriori adaptation for many-to-many eigenvoice conversion," *Proc. INTERSPEECH*, pp. 1461–1464, Sept. 2008.

[23] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "The use of air-pressure sensor in electrolaryngeal speech enhancement based on statistical voice conversion," *Proc. Interspeech*, pp. 1628–1631, Sept. 2010.

[24] A. Hisada and H. Sawada, "Real-time clarification of esophageal speech using a comb filter," *Proc. ICDVRAT*, pp. 39–46, Sept. 2002.

[25] K. Matui, N. Hara, N. Kobayashi, and H. Hirose, "Enhancement of esophageal speech using formant synthesis," *Proc. ICASSP*, pp. 1831–1834, May 1999.

[26] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, vol. 57, no. 10, Oct. 2010.

[27] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, vol. 53, no. 5, May 2006.

[28] G. Aguilar-Torres, M. Nakano-Miyatake, and H. Perez-Meana, "Enhancement and restoration of alaryngeal speech signals," *Proc. 16th IEEE Interna-*

tional Conference on Electronics, Communications and Computers (CONIELECOMP 2006), p. 30, Feb. 2006.

[29] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," *Proc. ICASSP*, pp. 5136–5139, May 2011.

[30] K. Nagahara, "Application and ranges of supracricoid laryngectomy with cricohyoidoepinglottopexy (SCL-CHEP)," *Journal of Otoloaryngology, Head and Neck Surgery (Japanese edition)*, vol. 18, no. 4, pp. 798–802, Apr. 2002.

[31] T. Takafuji, "Current situations of alaryngeal speech by laryngectomees," *Journal of Otolaryngology, Head and Neck Surgery (Japanese edition)*, vol. 2, no. 5, pp. 527–531, May 1986.

[32] S. E. Williams and J. B. Watoson, "Differences in speaking proficiencies in three laryngectomee groups," *Arch Otolaryngol*, vol. 111, pp. 216–219, Apr. 1985.

[33] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[34] SECOMMYVOICE,
"http://www.secom.co.jp/personal/medical/myvoice.html (in Japanese)," .

[35] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[36] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Electrolaryngeal speech enhancement based on statistical voice conversion," *Proc. Interspeech*, pp. 1431–1434, Sept. 2009.

[37] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Remodeling of the sensor for Non-Audible Murmur (NAM)," *Proc. INTERSPEECH*, pp. 389–392, Sept. 2005.

[38] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Enhancement of esophageal speech using statistical voice conversion," *APSIPA ASC*, pp. 805–808, Oct. 2009.

[39] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight," *Proc. MAVEBA*, Sept. 2001.

[40] 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, Aug. 2007.

[41] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum with a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, Mar. 2008.

[42] H. Kawahara, H. Katayose, A. Cheveigne, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f_0 and periodicity," *Proc. EUROSPEECH*, pp. 2781–2784, Sept. 1999.

[43] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on gaussian mixture model with straight mixed excitation," *IEICE Trans, Inf. and syst. (Japanese edition)*, vol. J91-D, no. 4, pp. 1082–1091, Apr. 2008.

[44] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," *ICSLP*, pp. 1043–1045, Sept. 1994.

[45] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models," *IEICE Trans. Inf. Syst.*, vol. E93-D, no. 9, pp. 2472–2482, Sept. 2010.

[46] H. Kenmochi and H. Ohshita, "VOCALOID – Commericial singing synthesizer based on sample concatenation," *Proc. INTERSPEECH*, pp. 4011–4012, Aug. 2007.

[47] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system - Sinsy," *SSW7*, pp. 211–216, Sept. 2010.

[48] H. Kenmochi, "VOCALOID and Hatsune Miku phenomenon in Japan," *Proc. InterSinging*, pp. 1–4, Oct. 2010.

[49] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," *INTERSPEECH*, pp. 2162–2165, Sept. 2010.

[50] T. Nakano and M. Goto, "Vocalistener2: A singing synthesis sytem able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," *Proc. ICASSP*, pp. 453–456, May 2012.

[51] Y. Yoshida, R. Nishimura, T. Irino, and H. Kawahara, "Vowel-based voice conversion and its application to singing-voice manipulation," *Proc. AES 35th International Conference: Audio for Games*, , no. 6, Feb. 2009.

[52] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP*, pp. 3905–3908, Apr. 2009.

[53] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voice," *Proc. WASPAA*, pp. 215–218, Oct. 2007.

[54] M. Goto, T. Nishimura, H. Hashiguchi, and R. Oka, "RWC Music Database: Music genre database and musical instrument sound database," *Proc. ISMIR*, pp. 229–230, Oct. 2003.

[55] J. Sundberg, "Articulatory interpretation of 'Singing Formant'," *J. Acoust. Soc. Am.*, vol. 55, pp. 838–844, 1974.

[56] Crypton Future Media, "What is the "HATSUNE MIKU movement"?," 2012.

List of publications

Journal papers

1. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models", IEICE Trans. Inf. Syst., Vol. E93-D, No. 9, pp.2472–2482, Sept. 2009.

International conference

1. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of esophageal speech using statistical voice conversion," APSIPA ASC, pp. 805–808, Oct. 2009.

2. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Statistical approach to enhancing esophageal speech based on Gaussian mixture models", Proc. ICASSP, pp. 4250–4253, Mar. 2010.

3. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Speaking-aid system based on one-to-many eigenvoice conversion for total laryngectomees", APSIPA ASC, pp. 498–501, Dec. 2010.

4. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques", Proc. ICASSP, pp. 5136–5139, May 2011.

5. H. Doi, T. Toda, T. Nakano, M. Goto, S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis systems", APSIPA ASC, Dec. 2012.

6. D. Deguchi, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Acoustic Compensation Method for Accepting Different Recording Devices in Body-Conducted Voice Conversion," Proc. APSIPA ASC, pp.502-505, Dec. 2010.

7. D. Deguchi, T. Toda, H. Doi, H. Saruwatari, K. Shikano, "Computationally efficient body-conducted voice conversion with original excitation signals," Proc. APSIPA ASC, Oct. 2011.

8. K. Yamamoto, T. Toda, H. Doi, H. Saruwatari, K. Shikano, "Statistical approach to voice quality control in esophageal speech enhancement," Proc. ICASSP, pp.4497–4500, Mar. 2012.

Technical reports

1. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of esophageal speech using statistical voice conversion," IPSJ SIG Technical Report, SIG-SLP, 2009-SLP-77, No. 18, pp. 1-6, July 2009 (in Japanese).

2. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Improvement of sound quality and speaker individuality for alaryngeal speech based on one-to-many eigenvoice conversion,"IPSJ SIG, Technical Report, 2010-SLP-82-15, no.2, pp.1-6, July 2010 (in Japanese).

3. D. Deguchi, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Automatic acoustic compensation method for accepting different recording devices in body-conducted voice conversion," IPSJ SIG Technical Report, 2010-SLP-82-15, no.1, pp.1-6, July 2010 (in Japanese).

4. Y. Yamamoto, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Voice Quality Control in Esophageal Speech Enhancement Using Statistical Voice Conversion," IPSJ SIG Technical Report, 2011-SLP-85, no.11, pp.1-6, Feb. 2011.

5. T. Toda, K. Nakamura, H. Doi, "Statistical voice vonversion techniques for alaryngeal speech enhancement," IEICE Technical Report, SP2010-58, pp.75-80, October 2010 (in Japanese).

Meetings

1. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of esophageal speech based on statistical voice conversion," Proc. 2009 Autumn Meeting of Acoustic Society of Japan, pp. 295–296, Sept. 2009 (in Japanese).

2. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of esophageal speech based on one-to-many eigenvoice conversion", Proc. 2010 Spring Meeting of Acoustic Society of Japan, pp. 349–350, Mar. 2010 (in Japanese).

3. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of alaryngeal speech based on one-to-many eigenvoice conversion", Proc. 2010 Autumn Meeting of Acoustic Society of Japan, pp. 295–296, Sept. 2010 (in Japanese).

4. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Evaluation of enhancement of alaryngeal speech based on statistical voice conversion", Proc. 2011 Spring Meeting of Acoustic Society of Japan, pp. 369–370, Mar. 2011 (in Japanese).

5. H. Doi, T. Toda, T. Nakano, M. Goto, S. Nakamura, "Singing voice conversion based on many-to-many eigenvoice conversion and training data generation with singing synthesis", Proc. 2010 Autumn Meeting of Acoustic Society of Japan, pp. 231–232, Sept. 2010 (in Japanese).

6. K. Yamamoto, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "An analysis of perceptual scores of primitive words expressing voice quality for voice quality control," Proc. 2010 Spring Meeting of Acoustic Society of Japan, pp. 479–480, Mar. 2010 (in Japanese).

7. K. Yamamoto, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Voice quality control in esophageal speech enhancement based on eigenvoice conversion," Proc. 2010 Autumn Meeting of Acoustic Society of Japan, pp. 319–320, Sept. 2010 (in Japanese).

8. D. Deguchi, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Acoustic compensation for body-conducted voice conversion using different recording devices," Proc. 2010 Autumn Meeting of Acoustic Society of Japan, pp. 289–290, Sept. 2010 (in Japanese).

9. D. Deguchi, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Reducing computation cost of body-conducted voice convesion by using residual waveform," Proc. 2011 Spring Meeting of Acoustic Society of Japan, pp. 329–330, Mar. 2011 (in Japanese).

10. K. Yamamoto, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Investigation of voice quality control in esophageal speech enhancement," Proc. 2011 Spring Meeting of Acoustic Society of Japan, pp. 375–376, Mar. 2011 (in Japanese).

11. M. Kishimoto, H. Doi, T. Toda, H. Saruwatari, K. Shikano, "Model training using training data including mismatched pause positions in statistical esophageal speech enhancement," Proc. 2012 Spring Meeting of Acoustic Society of Japan, pp. 367–368, Mar. 2012 (in Japanese).

Master's Thesis

1. H. Doi, "Quality Improvement of esophageal speech using statistical voice conversion," Master's thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT0851069, Feb., 2010.

Awards

1. Best student paper award, Proc. ICASSP, Mar. 2010.
2. Best presentation award, SIGMUS 96th regular meeting, Aug. 2012.
3. Student Award, 2012 Autumn Meeting of Acoustic Society of Japan, 2012.
4. The best paper award (short paper in regular session category), APSIPA 2012, Dec. 2012