Master Thesis

Eigenvoice-based character conversion and its evaluations

(Eigenvoice に基づくキャラクター変換とその評価)

48-106452 Teeraphon Pongkittiphan

(ポンキッティパン　ティーラポン)

Department of Information and Communication Engineering

Graduate School of Information Science and Technology

The University of Tokyo

16 August 2012

Supervisor: Prof. Nobuaki Minematsu

I would like to dedicate this thesis to my loving parents ...

Abstract

This thesis describes a new method of voice conversion, which aims at character conversion based on eigenvoice GMM (EV-GMM) approach. Using an eigenvoice space built from 273 speakers and speech samples of three different characters created by a single skilled voice actor/actress, the conversion can generate the voices of the three characters from an arbitrary speaker, while keeping the speaker identity. Listening tests were carried out by presenting two kinds of synthetic voices; before and after the character conversion. The results showed that listeners, both native and non-native speakers, can perceive well the character voice difference as what was intended by experimenters. It was also shown that this difference was perceived well even when F_0 difference between the two was very small, which indicates better performance of our method in character conversion compared to the general F_0-based conversion. Further, acoustic comparison between different characters in two cases of the voice actor and the proposed method was made. Results showed that the proposed method can realize acoustically valid modification between different characters.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

It's been a long journey since speech technology was born and mature enough for practical applications. Dating back to late 1870s [1], since the invention of telephone, the importance of speech technology was eventually widely concerned. In the middle of 1900s, when pulse code modulation (PCM) and linear predictive coding (LPC), and various complementary digital speech processing techniques were respectively invented, the combination of these techniques has realized a wide range of speech system, e.q. speech coding, speech synthesis, speech recognition and so on.

Among these systems, voice conversion is one of the most essential techniques that have been together developed. Voice conversion [2, 3] is a method to modify one speaker's speech so that it can be perceived as being spoken by another speaker without changing any linguistic contents. This technique is a potential tool for synthesizing speech with various kinds of speaker identity. Technically speaking, it can be generalized to be used not only for changing speaker identity but also for changing some other aspects of speech and for changing non-speech media. For example, it can be applied to cross-language voice conversion [4, 5] and hand-motion to speech conversion [6].

With the early limited knowledge in signal processing, the very first voice conversion method is normally done by using signal filtering. Later, the use of

statistical model has taken place and be a significant tool for the developing of mature voice conversion system. Several statistical techniques were proposed to estimate the conversion function. Among them, joint GMM approach, proposed by A. Kain et al. [3], is one of the most famous and efficient methods. This process requires a parallel set of utterances spoken by a source speaker and a target speaker. This set contains their utterances of the same sentences. The obtained conversion model provides a specific transformation between only that designated speaker pair.

Recently, T. Toda et al. [7] proposed eigenvoice conversion based on GMM (EV-GMM) that allows a flexible control of speaker characteristics. This eigenvoice technique is similar to eigenvoice based speech recognition [8] and eigenface [9] which originally developed for human face recognition. To train an eigenvoice GMM (EV-GMM), plenty of parallel utterance pairs between one reference speaker and many target speakers were used as prior knowledge. Using this data, the joint GMM between the one reference speaker and arbitrary target speakers can be estimated. In the eigenvoice speaker space, the identity of an arbitrary speaker can be represented as a unique weight vector and speaker identity can be controlled by adjusting its weight vector.

One of the remaining challenges is an attempt to expand the diversity of character or personality within a single speaker. Let us consider a scenario in voice acting industry, only one single voice actor can professionally perform voices for two or more animated characters.

To achieve this goal, I applied the EV-GMM approach not to speaker conversion, but to character conversion. Character conversion is the conversion from a source character to another character while keeping speaker identity. In other words, the conversion tries to generate various kinds of characters from a single speaker using training data of the various characters created by a single skilled voice actor/actress. This method is based on an eigenvoice space built from 273 speakers and uses the voices of some different characters given from another single voice actor/actress. The experimental results demonstrate that our proposed character conversion method can work well.

1.2 Objectives

The objectives of this research are to propose and develop a new method of voice conversion, called "character conversion", which is a conversion of speaker character or personality on the same individual, e.q. a conversion from a "childish" voice to an "elderly" voice of the same male speaker, and to investigate how the acoustic features change according to each character conversion.

1.3 Organization

The remainder of this thesis is organized as follows. Chapter 2 introduces the conventional voice conversion technique. Chapter 3 describes the basic eigenvoice conversion (EVC) technique. In Chapter 4, a framework of character conversion is described. Chapter 5 describes the experimental evaluation. Finally, the thesis is summarized in Chapter 6.

Chapter 2

Voice conversion

2.1 Introduction

Voice conversion [2, 3, 10] is a method to modify the speech of one speaker so that it can be perceived as if it was spoken by another speaker without changing any linguistic contents. This technique is a potential tool for synthesizing various kinds of speech. There are many systems that make use of VC, such as a speaker identity modification in Text-to-Speech (TTS) system [2], voice mail boxes that create desired voices without recording any further speakers [11], an enhancement of speech quality for telecommunications, hand-motion to speech conversion [6], cross language voice conversion [4, 5, 12] and a multi-lingual speech synthesizer [13].

To visualize the process flow of a general voice conversion, from Figure 2.1, a speech *"Let's take a walk"* uttered by a male source speaker X is converted via a conversion function F to the same sentence as being perceived as spoken by a female target speaker Y. This is called a "one-to-one voice conversion", which is a conversion from one source speaker's speech to that of one specific target speaker.

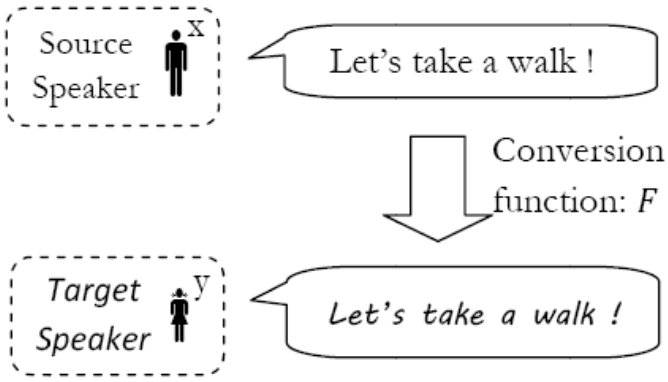

Figure 2.1: Process flow of voice conversion.

2.2 Acoustic Features

In the study of speech processing, there are several kinds of feature values being used to analyze the speech signal. The acoustic information extracted from this speech signal is called an "acoustic feature". In this section, the well-known and effective acoustic features, that usually mentioned and being used in the field of voice conversion, are introduced.

2.2.1 Fundamental frequency

The fundamental frequency (F_0) [1, 14], is the lowest frequency of a periodic waveform produced by the vibration of vocal cord. The human speech with higher number of F_0 will be perceived as a "high-pitch" voice, and as a "low-pitch" one in contrast. The typical voiced speech of a adult male will have a F_0 from 85 to 180 Hz, and that of a female from 165 to 255 Hz. Moreover, the F_0 of children voice is naturally higher than that of the mature one.

2.2.2 Cepstrum

The cepstrum [1, 14] represents the acoustic information of the vocal tract separately from the excitation. It is obtained by windowing the speech waveform to clip out the small pieces of frames at short time period, normally at millisecond scale. Then, the Discrete Fourier Transform (DFT) is used to extract its spec-

trum and then logarithmized to a logarithm scale. Finally, the logarithm power spectrum is transformed back to the time-domain by Inverse Discrete Fourier Transform (IDFT), as shown in Figure 2.2.

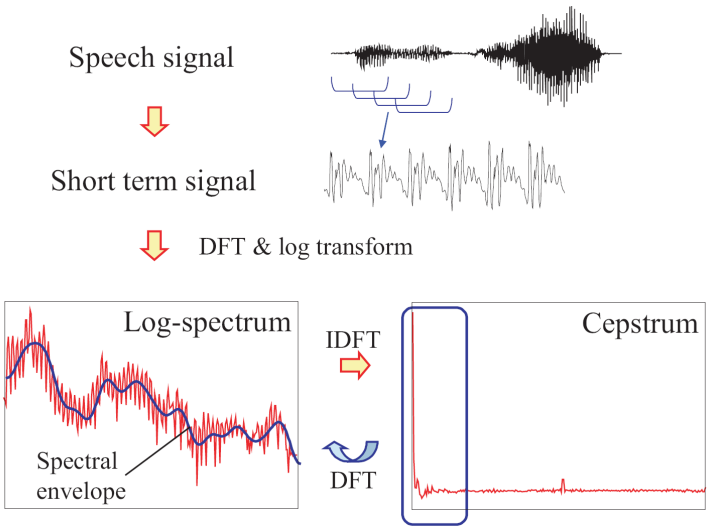

Figure 2.2: Extraction of cepstrum.

2.2.3 Mel scale cepstrum

Instead of using the hertz (Hz) scale, the mel scale is used to produce the equal linear pitch increments. This is because the perceptual scale of pitch judged by human is not linear in frequency domain, and the human ear is more sensitive to the lower frequency and can catch more information than that from higher frequency, as shown in Figure 2.3. The mel scale frequency can be calculated by the following equation,

$$m = 2595log_{10}(1 + \frac{f}{100}), \tag{2.1}$$

where m is the mel scale frequency and f is the hertz scale frequency.

In speech processing, the mel-frequency cepstral coeffiecient (MFCC) is commonly used as a feature value. It can be derived by taking the Fourier transform of a windowed signal and converting to the mel scale, according to Equation 2.1,

in order to get the mel scale spectrum. Then, the MFCC is calculated by taking the discrete cosine transform (DCT) of the log power of mel scale spectrum.

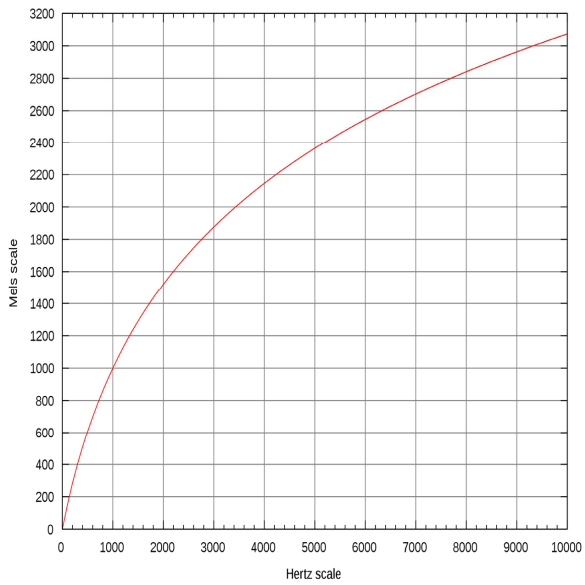

Figure 2.3: Plots of pitch mels versus hertz.

2.2.4 Delta cepstrum

The cepstrum itself only conveys the static feature of the speech at each time frame while omitting its detail of the change between each small frame of speech signal. The delta cepstrum was proposed as a acoustic feature that can represent further information of the time variance.

The delta cepstrum ($\triangle c_t$) and delta delta cepstrum ($\triangle^2 c_t$) are defined as first-order and second-order coefficients of the quadratic approximation for the time frame and its adjacent L frames, which can be obtained by the following equations,

$$\triangle c_t = \frac{\Sigma_{i=-L}^{L} i c_{t+i}}{\Sigma_{i=-L}^{L} i^2}, \tag{2.2}$$

$$\triangle^2 c_t = \frac{\Sigma_{i=-L}^{L} (a_0 i^2 - a_1) c_{t+i}}{\Sigma_{i=-L}^{L} (a_2 a_0 - a_1^2)}, \tag{2.3}$$

where $a_2 = \Sigma_{i=-L}^{L} i^4$, $a_1 = \Sigma_{i=-L}^{L} i^2$, $a_0 = \Sigma_{i=-L}^{L} 1$, and c_t is the cepstrum of t-th frame.

Many previous researches shown that using the concatenation $[c_t, \triangle c_t, \triangle^2 c_t]$, instead of c_t alone, improves the performance of speech conversion significantly [15, 16].

2.3 Conversion function

To realize the voice conversion, let $X = [x_1, x_2, ..., x_n]$ be the sequence of feature vectors representing a succession of speech spoken by the source speaker, and $Y = [y_1, y_2, ..., y_n]$ be those of the same sounds produced by target speaker. The goal is to find the conversion function F to convert the source vector x to the target vector y that minimizes the mean squared error as defined in the following equation [4],

$$E_{mse} = E[||Y - F(X)||^2], \tag{2.4}$$

where E denotes an expectation value.

Many researchers have proposed different solutions to find the effective conversion function. Recently, the voice conversion (VC) based on statistical approaches has been studied and eventually widespread. One of the earliest conventional methods was proposed by Abe et al. [17] that is a codebook mapping method based on hard clustering and discrete mapping, using vector quantization (VQ) technique. However, because of the parameter space of the converted envelop is limited to a discrete set of envelopes, this causes a drop in the quality of the converted speech. The use of fuzzy VC [18] was introduced to realize the soft clustering version of the Abe's one, but many authors commented that it is still lack of robustness [19].

There are still many methods proposed, e.q. conversion methods based on artificial neural networks [20], linear multivariate regression (LMR) [21] and speaker interpolation [22]. Among the probabilistic approaches, the VC based on maximum likelihood estimation (MLE) with a joint Gaussian mixture model (GMM) training becomes one of the most popular techniques [10].

2.4 Voice conversion based on maximum likelihood estimation (MLE)

In this section, the process flow of the VC based on MLE [10] is introduced. First, the feature vectors of source and target speech are extracted and used as the training data for the parallel joint GMM. In the case of VC, each parallel utterance pair is the same sentence read by the source and target speaker, respectively. Then, the dynamic time warping (DTW) is applied to these parallel feature vectors for preparing the time alignment to ensure that the small pieces of parallel speech frame share the same linguistic content or the same part of phoneme sound. Next, the statistical parameters of the joint GMM, e.q. the mean vectors and covariance matrix, are estimated. Finally, the conversion function is calculated with MLE. The details of each step is described in these following subsections.

2.4.1 Parallel joint GMM training

The $2D$-dimensional feature vectors $X_t = [x_t^\top, \Delta x_t^\top]^\top$ and $Y_t = [y_t^\top, \Delta y_t^\top]^\top$ are the static and dynamic features of source and target speakers, respectively, where $^\top$ denotes the transposition of a vector. The dynamic time warping (DTW) is applied to these feature vectors to ensure the correct time alignment. The GMM on joint probability density is trained with the joint vector of source and target speakers, and it is represented by using GMM parameter λ.

$$p(X_t, Y_t \mid \lambda) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(X_t^\top, Y_t^\top; \mu_m^{(X,Y)}, \Sigma_m^{(X,Y)}), \tag{2.5}$$

$$\mu_m^{(X,Y)} = \begin{bmatrix} \mu_m^X \\ \mu_m^Y \end{bmatrix}, \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}, \tag{2.6}$$

where M denotes the number of mixtures, $\mathcal{N}(x; \mu_m, \Sigma_m)$ denotes the normal distribution for m^{th}-mixture with mean vector μ_m and covariance matrix Σ_m, and α_m is the weight of mth-mixture.

2.4.2 Maximum likelihood estimation

The conversion function is derived based on the conditional probability density of Y_t, given X_t, according to the following equations [10],

$$p(Y_t \mid X_t, \lambda) = \sum_{m=1}^{M} p(m \mid X_t, \lambda)\, p(Y_t \mid X_t, m, \lambda), \tag{2.7}$$

where

$$p(m \mid X_t, \lambda) = \frac{\alpha_m \mathcal{N}(X_t; \mu_m^X, \Sigma_m^{(XX)})}{\sum_{m=1}^{M} \alpha_m \mathcal{N}(X_t; \mu_m^X, \Sigma_m^{(XX)})}, \tag{2.8}$$

$$p(Y_t \mid X_t, m, \lambda) = \mathcal{N}(Y_t; E_{m,t}^Y, D_m^Y). \tag{2.9}$$

The conditional mean vector $E_{m,t}^Y$, and the covariance matrix D_m^Y of the mth conditional probability distribution are written as

$$E_{m,t}^Y = \mu_m^Y + \Sigma_m^{(YX)} \Sigma_m^{(YX)^{-1}} (X_t - \mu_m^X), \tag{2.10}$$

$$D_m^Y = \Sigma_m^{(YY)} - \Sigma_m^{(YX)} \Sigma_m^{(XX)^{-1}} \Sigma_m^{(XY)}. \tag{2.11}$$

By using the EM algorithm, the converted feature vector $\hat{Y}$ can be determined as

$$\hat{Y} = = \sum_{m=1}^{M} p(m \mid X_t, \lambda) E_{m,t}^Y. \tag{2.12}$$

2.5 Summary

In this chapter, the conventional method of voice conversion and well-known acoustic features are introduced. In speech processing, the acoustic features are used as feature values for the speech analysis and synthesis. The fundamental frequency F_0, or the pitch of sound, represents the information from excitation source, while the cepstrum shows that from vocal tract. Instead of only using the cepstrum, the mel scale cepstrum and delta cepstrum are also included as

they provide more reliable acoustic information. In the case of voice conversion, among the statistical methods, the use of voice conversion based on maximum likelihood estimation (MLE), with joint Gaussian Mixture Model (GMM), is the one of the most popular approaches, which is effective and widely used.

Chapter 3

Eigenvoice conversion

3.1 Introduction

The voice conversion mentioned in Chapter 2, however, is still the method of the one-to-one voice conversion, which its conversion is kept to only the specific pair of speakers. The speech quality of the "one-to-one VC" is acceptable but there are many limitations, e.q. the less flexibility, the time consuming and the huge amount of parallel utterances required for the training.

Many researches proposed several effective approaches that using the voices of other speakers as prior knowledge to provide the flexibility when adopting the GMM for any desired speaker pair. For instance, Lee et al. [23] proposed the unsupervised training method based on maximum a posteriori (MAP) adaptation, and Mouchtaris et al. [24] proposed another unsupervised training method based on maximum likelihood constrained adaptation of the GMM trained with an existing parallel data set of a different speaker-pair.

Among the unsupervised training method using the prior knowledge, the eigenvoice-based conversion is one of the most successful approaches. There are different variation of the eigenvoice VC (EVC), i.e. the "one-to-many EVC" [7, 25] that is the conversion from a single source speaker to any arbitrary target speaker and the "many-to-one EVC" [25], which is the inverse version of one-to-many EVC, providing a conversion from any arbitrary source speaker to a specific target speaker. Moreover, when linking these two conversion by sharing the refer-

ence speaker, the most flexible "many-to-many EVC" [26] can be realized which is the conversion from any source speaker to any target speaker. In this chapter, the details of eigenvoice-based voice conversion (EVC) and its three variation, "one-to-many", "many-to-one" and "many-to-many" EVC will be explained.

3.2 Eigenvoice

Eigenvoice was first proposed by Kuhn [27] and was inspired by the eigenfaces technique [9]. Its main concept is to create a low-dimensional space describing speaker variability, then find a mapping between every point in this space and speaker dependent model (SD-model), and, finally, find a method of mapping new speech to a point in this space.

In this eigenvoice speaker space, the existance of any speakers can be ploted as a point in this space. Technically speaking, the speaker identity of any speakers is defined by its unique weight vector, and the modification of this weight vector will provide the flexibility of speaker characteristic conversion.

3.3 Principal component analysis (PCA)

In reduction techniques, Principal Component Analysis (PCA) is one method used to reduce the number of features used to represent data, by projecting data from a higher dimension to a lower dimensional space such that the error incurred by reconstructing the data in the higher dimension is minimized. The benefits of this dimensionality reduction include providing a simpler representation of the data, reduction in memory, and faster classification. Its methodolody is described as the following steps,

Suppose $x_1, x_2, ..., x_M$ are $N \times 1$ vectors

Step 1 : compute the mean vector,

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i \tag{3.1}$$

Step 2 : subtract the mean $\Phi_i = x_i - \bar{x}$

Step 3 : form the matrix $A = [\Phi_1, \Phi_2, ..., \Phi_M]$ ($N \times M$ matrix), then compute:

$$C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T \tag{3.2}$$

(sample covariance matrix, $N \times N$, charaterizes the scatter of data)

Step 4 : compute the eigenvalues of $C : \lambda_1 > \lambda_2 > ... > \lambda_N$

Step 5 : compute the eigenvectors of $C : u_1, u_2, ..., u_N$

Since C is symmetric, $u_1, u_2, ..., u_N$ form a basis, (i.e., any vector x or actually $(x - \bar{x})$, can be written as a linear combination of the eigenvectors):

$$x - \bar{x} = b_1 u_1 + b_2 u_2 + ...b_N u_N = \sum_{i=1}^{N} b_i u_i \tag{3.3}$$

Step 6 : (dimensionality reduction step) keep only the terms corresponding to the K largest eigenvalues:

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \;\; where \;\; K << N \tag{3.4}$$

The representation of $\hat{x} - \bar{x}$ into the basis $u_1, u_2, ..., u_K$ is thus

$$\begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_K \end{bmatrix}, \tag{3.5}$$

The linear transformation $R^N -> R^K$ that performs the dimensionality reduction is:

$$\begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \cdots \\ u_K^T \end{bmatrix} (x - \bar{x}) = U^T(x - \bar{x}). \tag{3.6}$$

3.4 Eigenvoice GMM (EV-GMM)

In this section, three variations of eigenvoice conversion, "one-to-many", "many-to-one" and "many-to-many" EVC, will be described, respectively.

3.4.1 "One-to-many" EVC

The one-to-many eigenvoice conversion (EVC) [7] is the most general version of EV. It gives the speech conversion from a single source speaker to any arbitrary traget speakers.

3.4.1.1 EV-GMM Training

The $2D$-dimensional feature vectors $X_t = [x_t^\top, \Delta x_t^\top]^\top$ and $Y_t = [y_t^\top, \Delta y_t^\top]^\top$ are the static and dynamic features of source and target speakers, respectively, where $^\top$ denotes the transposition of a vector. The EV-GMM on joint probability density is trained with the joint vector of source and target speakers, and it is represented by using GMM parameter λ and its weight vector w.

$$p(X_t, Y_t \mid \lambda, w) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(X_t^\top, Y_t^\top; \mu_m^{(X,Y)}(w), \Sigma_m^{(X,Y)}), \tag{3.7}$$

$$\mu_m^{(X,Y)}(w) = \begin{bmatrix} \mu_m^X \\ \mu_m^Y(w) \end{bmatrix}, \ \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}, \tag{3.8}$$

where M denotes the number of mixtures, $\mathcal{N}(x; \mu_m, \Sigma_m)$ denotes the normal distribution for m^{th}-mixture with mean vector μ_m and covariance matrix Σ_m, and α_m is the weight of mth-mixture.

To build the eigenvoice speaker space, as shown in Figure 3.1, first, we train

a target independent joint GMM (TI-GMM) using joint vectors of one reference source speaker and all the S target speakers. In other words, TI-GMM is a target independent but source dependent joint GMM. Then, target dependent GMMs (TD-GMM) for each targer speaker is separately trained by updating the target means of TI-GMM.

We prepare a $2DM \times S$ matrix, of which each column is a $2D \times M$ dimensional supervector that is the concatenation of target means of each Sth TD-GMM. To extract a small number of basis vectors, which can be used to represent any supervector, PCA is done. Finally, the target mean vector $\mu_m^Y(w)$ can be represented as the linear combination of bias vector $b_m^{(0)}$ and $B_m = [b_m^{(1)}, ..., b_m^{(K)}]$, which is a matrix of basis vectors, where $K < S$.

$$\mu_m^Y(w) = B_m w + b_m^{(0)}. \tag{3.9}$$

The flexibility control of speaker individuality can be done by adjusting the K-dimensional weight vector w.

3.4.1.2 Adaptation to any arbitrary target speakers

The adaptation to any given target speaker Y is to estimate his weight vector w done by applying the maximum likelihood eigen-decomposition (MLED) [8] so that the time-sequence marginal distribution likelihood of the target feature vector is maximized as follows:

$$\begin{aligned} \hat{w} &= \underset{w}{\operatorname{argmax}} \int p(X, Y \mid \lambda, w) dX, && (3.10) \\ &= \underset{w}{\operatorname{argmax}} \int p(Y \mid \lambda, w) p(X \mid Y, \lambda, w) dX, && (3.11) \\ &= \underset{w}{\operatorname{argmax}}\, p(Y \mid \lambda, w). && (3.12) \end{aligned}$$

We use EM-algorithm to find the weight vector for target speaker Y, which can be written as

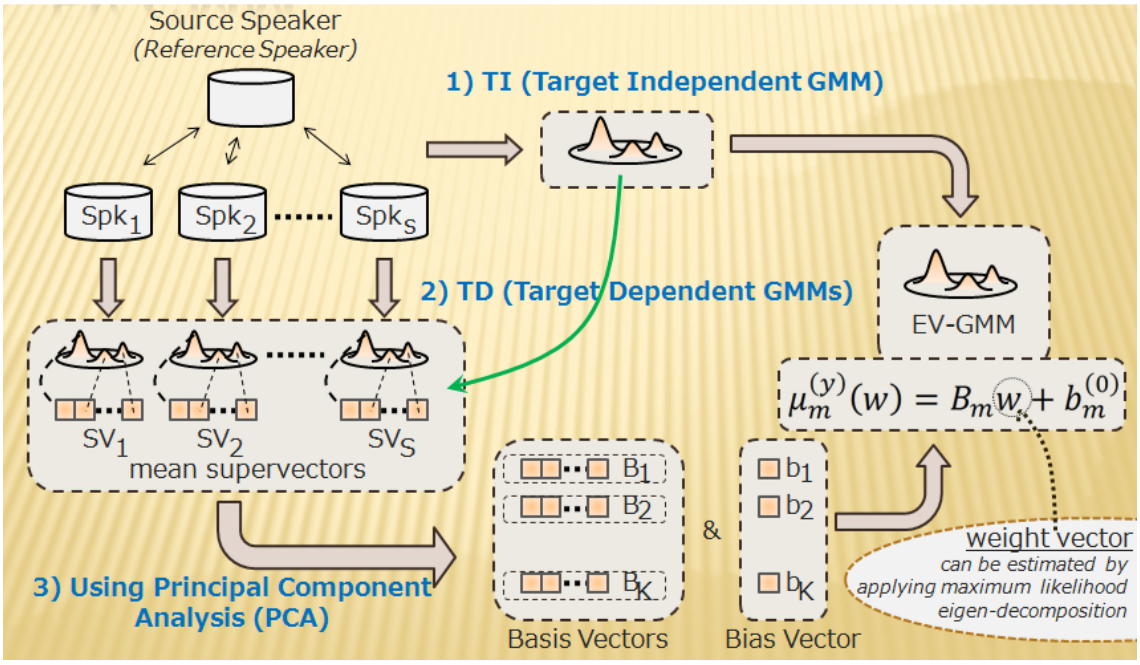

Figure 3.1: Training process of EV-GMM

$$\hat{w} = \{\sum_{m=1}^{M} \bar{\gamma}_m B_m^\top \Sigma_m^{(YY)^{-1}} B_m\}^{-1} \sum_{m=1}^{M} B_m^\top \Sigma_m^{(YY)^{-1}} \bar{Y}_m, \tag{3.13}$$

$$\bar{\gamma}_m = \sum_{t=1}^{T} \gamma_{m,t}, \;\; \bar{Y}_m = \sum_{t=1}^{T} \gamma_{m,t}(Y_t - b_m^{(0)}), \tag{3.14}$$

$$\gamma_{m,t} = p(m \mid Y_t, \lambda, w), \tag{3.15}$$

where Equation 3.11 approximates the weight vector $\hat{w}$ of the target speaker and TI-GMM is used as an initialization of Equation 3.13.

3.4.1.3 Conversion function estimation

Finally, the converted mean vector to target speaker Y from the source speaker X for the mth-mixture in EVC ($E_{t,m}$) can be estimated by MLE as the same

manner as mentioned in Chapter 2.4 as

$$E_{t,m} = [B_m w + b_m^{(0)}] + \Sigma_m^{(YX)} \Sigma_m^{(XX)^{-1}} (X_t - \mu_m^X). \tag{3.16}$$

3.4.2 "Many-to-one" EVC

The many-to-one eigenvoice conversion (EVC) [25] is a straightforward adaptation of the one-to-many EVC, described in section 3.4.1. It is the inverse version of one-to-many EVC by replacing the source and the target each other.

The $2D$-dimensional feature vectors $X_t = [x_t^\top, \Delta x_t^\top]^\top$ and $Y_t = [y_t^\top, \Delta y_t^\top]^\top$ are the static and dynamic features of source and target speakers, respectively, where $^\top$ denotes the transposition of a vector. The EV-GMM on joint probability density is trained with the joint vector of source and target speakers, and it is represented by using GMM parameter λ and its weight vector w.

$$p(X_t, Y_t \mid \lambda, w) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(X_t^\top, Y_t^\top; \mu_m^{(X,Y)}(w), \Sigma_m^{(X,Y)}), \tag{3.17}$$

$$\mu_m^{(X,Y)}(w) = \begin{bmatrix} \mu_m^X(w) \\ \mu_m^Y \end{bmatrix}, \ \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}, \tag{3.18}$$

$$\mu_m^X(w) = B_m w + b_m^{(0)}. \tag{3.19}$$

where M denotes the number of mixtures, $\mathcal{N}(x; \mu_m, \Sigma_m)$ denotes the normal distribution for m^{th}-mixture with mean vector μ_m and covariance matrix Σ_m, and α_m is the weight of mth-mixture.

In many-to-one VC, the source mean vector for the ith mixture is represented as a linear combination of a bias vector $b_i^X(0)$ and representative basis vectors $B_i^X = [b_i^X(1), ..., b_i^X(J)]$. The number of representative vectors is J. The source speaker individuality is controlled with only the J-dimnesional weight vector $w = [w(1), ..., w(J)]^\top$.

Instead of preparing target independent GMM (TI-GMM) as in one-to-many VC method, a source independent GMM (SI-GMM) is trained using all of the multiple parallel data sets, and the source dependent GMM (SD-GMM) of each

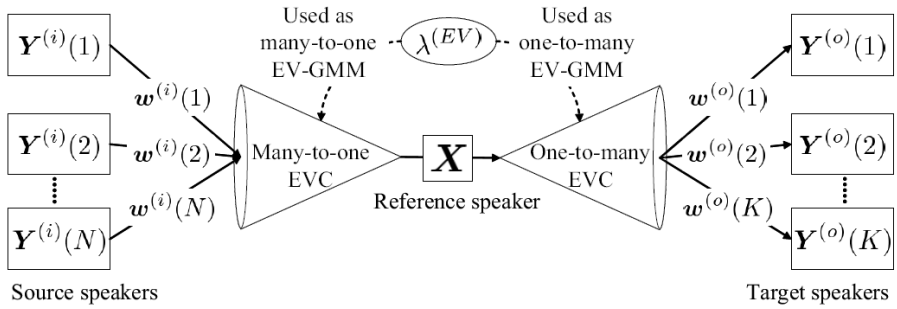

Figure 3.2: Overview of many-to-many EVC.

SI-GMM is trained by updating its source mean vector. Next, as the same manner as one-to-many VC, the PCA is applied to SD-GMM to extract the basis vectors B_i^X and a bias vector $b_i^X(0)$. Finally, the speaker space is constructed and the conversion method can be done based on the MLE voice conversion, as mentioned in Chapter 2.4.

3.4.3 "Many-to-many" EVC

The many-to-many eigenvoice conversion (EVC) [26] provides the most flexible conversion, which is the conversion from any arbitraty source speakesr to any arbitrary traget speakers. The implementation is done by the connection of the many-to-one and one-to-many EVC via a reference speaker at each single end-point.

The EV-GMM of the many-to-one and one-to-many EVC is previously trained using the same reference speaker at each single end-point. Actually, the one-to-many EV-GMM is first trained and the many-to-one EV-GMM can be later determined by just switching the source and the target features.

Figure 3.2 shows that the many-to-one and one-to-many EVC is linked via the reference speaker X at each own end-point. With a small amount of speech data from the source and target speaker, the adaptation of arbitary source speaker $y^{(i)}$ and target speaker $y^{(o)}$ and their weight vector $\hat{w}^{(i)}$ and $\hat{w}^{(o)}$ can be estimated. At last, the source speaker's voice is converted to the reference speaker's voice by many-to-one EVC, and then being converted to the target speaker's voice by the one-to-many EVC.

3.4.3.1 Conversion method based on multiple VC

In the first half, the source speech is converted to the reference speech using the many-to-one EVC. The maximum likelihood estimation of the feature vector of the reference speech $\bar{X}$ is calculated by maximizing the following likelihood fuction,

$$\bar{X} = \underset{x}{\operatorname{argmax}} \sum_{all\, m} p(m \mid Y^{(i)}, \lambda, \hat{w}^{(i)}) \times p(X \mid Y^{(i)}, m, \lambda, \hat{w}^{(i)}). \tag{3.20}$$

As the same manner as mentioned in section 3.4.2, the suboptimum mixture component sequence $\hat{m}^{(i)}$ is determined as the following equation,

$$\hat{m}^{(i)} = \underset{m}{\operatorname{argmax}}\, p(m \mid Y^{(i)}, \lambda, \hat{w}^{(i)}). \tag{3.21}$$

Next, the converted reference speech $\bar{X}$ is estimated as follows:

$$\bar{X} = \underset{x}{\operatorname{argmax}}\, p(X \mid Y^{(i)}, \hat{m}^{(i)}, \lambda, \hat{w}^{(i)}). \tag{3.22}$$

In the second half, the reference speech is later converted to the target speech using the one-to-many EVC. The feature vector of the target speech $\hat{y}^{(o)}$ is estimated by maximizing the following likelihood fuction,

$$\hat{y}^{(o)} = \underset{y}{\operatorname{argmax}} \sum_{all\, m} p(m \mid \bar{X}, \lambda) \times p(Y^{(o)} \mid \hat{X}, m, \lambda, \hat{w}^{(o)}). \tag{3.23}$$

The suboptimum mixture component sequence $\hat{m}^{(o)}$ is determined as follows:

$$\hat{m}^{(o)} = \underset{m}{\operatorname{argmax}}\, p(m \mid \hat{X}, \lambda). \tag{3.24}$$

Then, the converted feature vectors of target speech $\hat{y}^{(o)}$ is estimated as follows:

$$\hat{y}^{(o)} = \underset{y^{(o)}}{\operatorname{argmax}}\, p(Y^{(o)} \mid \hat{X}, \hat{m}^{(o)}, \lambda, \hat{w}^{(o)}),\ subject\, to\, Y^{(o)} = W y^{(o)}. \tag{3.25}$$

Note that the alignment of the mixture component index in many-to-one EVC, shown in Equation 3.21, is not always corresponding to that in one-to-many EVC shown in Equation 3.24. The inconsistency of the mixture component index may causes the wrong conversion across different phonemic spaces.

3.4.3.2 Conversion method with shared mixture components

To eliminate the inconsistency problem of the mixture component index, a sequential conversion approach, while sharing the same mixture component index in both many-to-one and one-to-many EVC, is proposed.

The converted feature vectors of target speech $\hat{y}^{(o)}$ is determined by maximizing the following likelihood function,

$$\hat{y}^{(o)} = \underset{y}{\operatorname{argmax}} \sum_{all\, m} p(m \mid Y^{(i)}, \lambda, \hat{w}^{(i)}) \times p(Y^{(o)} \mid Y^{(i)}, \lambda, \hat{w}^{(i)}, \hat{w}^{(o)}), \tag{3.26}$$

where,

$$\begin{aligned} & p(Y^{(o)} \mid Y^{(i)}, \lambda, \hat{w}^{(i)}, \hat{w}^{(o)}) && (3.27) \\ = & \int p(Y^{(o)} \mid X, m, \lambda, \hat{w}^{(o)}) \times p(X \mid Y^{(i)}, m, \lambda, \hat{w}^{(i)}) dX, && (3.28) \\ = & \prod_{t=1}^{T} \mathcal{N}(Y_t^{(o)}; \bar{E}_{m,t}^{(Y)}, \bar{D}_m^{(Y)}), && (3.29) \end{aligned}$$

$$\begin{aligned} \bar{E}_{m,t}^{(Y)} &= B_m \hat{w}^{(o)} + b_m^{(o)} + A_m \Sigma_m^{(YY)^{-1}} (Y_t^{(i)} - B_m \hat{w}^{(i)} - b_m^{(o)}), && (3.30) \\ \bar{D}_m^{(Y)} &= \Sigma_m^{(YY)} - A_m^T \Sigma_m^{(YY)^{-1}} A_m, && (3.31) \\ A_m &= \Sigma_m^{(XX)} \Sigma_m^{(XX)^{-1}} \Sigma_m^{(XY)}. && (3.32) \end{aligned}$$

The feature vectors of the reference speaker X is considered as a hidden variable. A graphical visualization of the relationship between individual variables in the conversion process is shown in Figure 3.3.

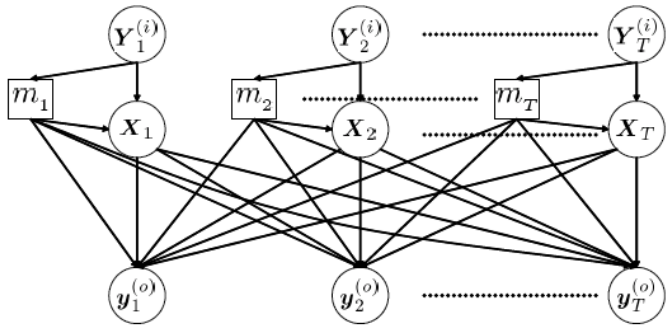

Figure 3.3: Graphical representation of relationship between individual variables in many-to-many EVC with reference speech.

As the same manner as described in section 3.4.2, the suboptimum mixture component sequence $\hat{m}$ is determined by

$$\hat{m} = \underset{m}{\operatorname{argmax}}\, p(m \mid Y^{(i)}, \lambda, \hat{w}^{(i)}). \tag{3.33}$$

Finally, the converted feature vectors of target speech $y^{(o)}$ is determined as follows:

$$\hat{y}^{(o)} = \underset{y^{(o)}}{\operatorname{argmax}}\, p(Y^{(o)} \mid Y^{(i)}, \hat{m}, \lambda, \hat{w}^{(i)}, \hat{w}^{(o)}),\ subject\, to\, Y^{(o)} = Wy^{(o)}. \tag{3.34}$$

3.5 Summary

This chapter describes another technique of voice conversion by using the concept of eigenvoice. Eigenvoice-based voice conversion (EVC) can solve the problems and limitations found in the conventional one-to-one VC, mentioned in Chapter 2. The use of eigenvoice speaker space provides more flexible modification of speaker identity. The principle component analysis (PCA), which is a reduction technique used for the eigenvoice speaker space estimation, is described. Moreover, the details of three variations of EVC, which are “one-to-many”, “many-to-one” and “many-to-many” EVC, are introduced.

Chapter 4

Character Conversion

4.1 Introduction

The concept of character conversion is quite different from conventional voice conversion since its objective is to realize the character speech of the same speaker not to change the speaker identity of a source speaker to that of different target speakers. In other words, an imaginary character speech sample of the same speaker can be realized even if it cannot be naturally uttered by that speaker in a real situation.

In the perspective of vector space of speakers, as shown in Figure 4.1, in the conventional voice conversion (left panel), the bold and the dash conversion vectors are different because either of them is depending on the specific pair of source and target speakers. However, in character conversion (right panel), both of the dash vectors are the same, which is originally extracted from the target character voices and the natural voices of a single voice artist. This delta vector is applied as it is to the natural voices of a source speaker to obtain the target character voice of that source speaker. In other words, the character conversion's vector is once calculated and still practical to any arbitrary speakers.

In this thesis, when mentioned about character conversion, we call the natural speech of a voice artist as "source character" voice and his/her different character speech as "target character" one. Meanwhile, we call the "source speaker" as another general speaker that his/her speech will be converted into the "target-

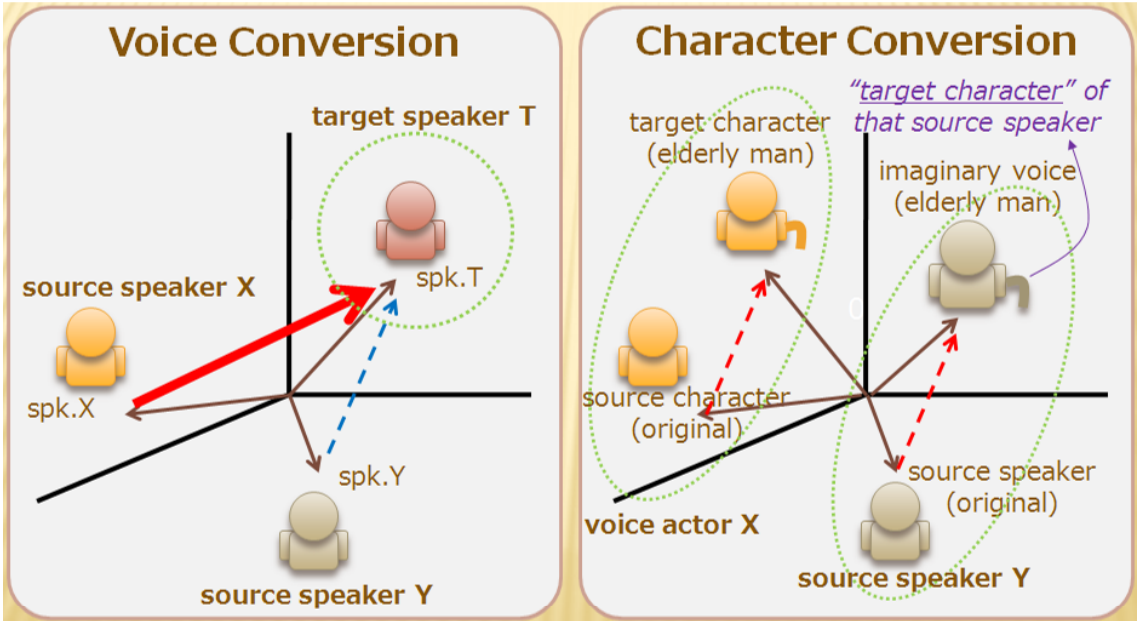

Figure 4.1: Illustration of character conversion.

character-like" voice.

4.2 Delta-weight vector calculation

To realize a conversion from *character-A* to *character-B* on a given source speaker, a parallel set of *character-A* and *B* utterances performed by a single voice actor were used to estimate both weight vectors of *character-A* and *B*. Instead of choosing two characters from two different voice actors, I collected those from the same speaker to avoid an error that may be caused by an articulation variation between different speakers.

Note that, in this thesis, *character-A* often indicates the original and natural voice of that voice actor.

Then, delta-weight vector A→B, henceforth $\Delta W_{A\to B}$, which indicates a change from *character-A* to *character-B* on the same voice actor, is calculated by the following function,

$$\Delta W_{A\to B} = W_{(B)} - W_{(A)}, \tag{4.1}$$

where $W_{(A)}$ and $W_{(B)}$ are the weight vector of *character-A* and that of *character-B* of the same voice actor, respectively.

The $\Delta W_{A \to B}$ is later added to another source speaker's weight vector.

$$W_{(A \to B\, on\, src)} = W_{(src)} + \alpha \Delta W_{A \to B}, \tag{4.2}$$

where $W_{(src)}$ is a weight vector of another source speaker, and $W_{(A \to B\, on\, src)}$ is an estimated weight vector of the expected *character-B-like* voice of that source speaker.

With this modified weight vector, the expected "*character-B-like*" voice of that source speaker can be realized. Moreover, a coefficient α is included to control the strength of delta-weight vector (ΔW). The larger α is, the larger change to target *character-B* will be applied on the source speaker.

4.3 Fundamental frequency conversion

The spectral conversion can be done directly by our proposed method, mentioned in 4.2. In the case of F_0 conversion, I consider the change of F_0 from source character (A) to target character (B) on the same voice actor and use a linear pitch scaling to project the F_0 of the source speaker to that of the same speaker with target character by the following equation,

$$y = \mu_x + \Delta\mu_{A \to B} + \frac{\sigma_B}{\sigma_x}(x - \mu_x), \tag{4.3}$$

where y is a converted F_0, x is an input source speech F_0, μ_x is the F_0 mean of source speaker, $\Delta\mu_{A \to B}$ is the change of F_0 mean from source character (A) to target character (B) on the same voice actor and σ_x, σ_B are the standard deviation of F_0 of source speaker and that of target character (B) of the voice actor, respectively.

4.4 Character speech corpus

However, only the concept of the character conversion itself cannot realize the practical conversion. The well-prepared speech corpus is one of the most important issues as well. The collection and corpus recording are necessary because the current provided resources of the speech corpus, that mainly focusing on the change of speaker character, is not available.

The group of semi-professional voice artists, 2 voice actors and 6 voice actresses, from the annoucer and voice training school are asked to join the character speech corpus recording. Each person has to read the same set of sentences with the natural speech and 2-3 kinds of his/her skilled character speech, e.q a "childish", "elderly", "autie", "teen student", "shy girl" voice and so on. For the reading style, they also has to try keeping the reading speed and the pause between phrase as the same as those read in every character to avoid the inconsistency of time variant. The reading material is the standard content consisting of 503 Japanese sentences, which can be divided into 10 sets of phoneme-balanced sentences. The examples of sentences are provided in the appendix part.

4.5 Summary

In this chapter, the framework of my proposed method, called "character conversion", is described. Character conversion is a new concept in the field of voice conversion. Instead of changing the voice of one speaker into another speaker, this method is to convert the speaker character or personality within a single speaker, e.q. a conversion from a "childish" voice to an "elderly" voice of the same male speaker.

In the eigenvoice speaker space, any arbitrary speaker identity can be represented as a unique weight vector, the conversion of speech from one character to another character can be realized by pre-calculating the "delta-weight vector" of two different characters performed by a single voice artist, and then adding it to the original weight vector estimated from the natural voice of any speaker.

Chapter 5

Experiments and evaluations

In this chapter, two kinds of experiment and evaluation methods are described. To evaluate the performance of the character conversion, the subjective subjective experimental evaluation and the acoustic evaluation were conducted.

The first experiment is the listening test of the synthetic samples of before and after character conversion. The native Japanese and native Thai subjects have joined the experiments. The aims of this experiment are to evaluate whether the chatacter conversion is practical enough, to investigate how well the subjects can perceive the change in each character and also to compare the cross-cultural factor that may affect the perception of character.

The other one is the acoustic evaluation based on the analyzing of the spectrogram of the speech signal. Its objective is to analyze what kind of acoustic change can be found in each character conversion.

5.1 Subjective experimental evaluation

5.1.1 Speaker space construction

In order to train EV-GMM, we used one male speaker as a reference speaker; MHT speaker in ATR database [28]. And, the 273 speakers of JNAS database [29], consisting of 136 male and 136 female speakers, are used as the target speakers. All speakers read the same Japanese phonetic-balanced 50 sentences. The DTW was preformed to automatically prepare the time-aligned parallel data sets.

Speech signals are sampled at 16 kHz sampling rate, and STRAIGHT [30, 31] analysis is used to extract the spectral envelop which later be converted to mel-cepstral coefficients using a recursion formula. The feature vector is 48 dimensional including 24 mel-cepstral coefficients and their delta values. Then, EV-GMM is finally trained with 128 mixtures. As a result, the speaker space is built as a weight vector space using a reference speaker and many target speakers, as described in Chapter 3.

5.1.2 Design of evaluation

We ran evaluation experiments by conducting 3 kinds of comparative conversions, named 'F0', 'X1' and 'X2'.

'F0' means the conventional F_0-based voice conversion where only F_0 of a source speaker is projected to that of the target character with linear transformation.

'X1' and 'X2' are the character conversions done by our proposed method. They both apply F_0 conversion, and spectral conversion with coefficient α =1 and α =2, respectively. We have observed that the value of α coefficient should be constantly fixed from 1 to 2, so that the converted speech sounded natural. The large value of α will cause over-conversion problem, and the converted speech might be unnatural and creaky.

Namely, all samples done by 3 conversion methods have the same F_0 patterns, while there are only differences in spectral features.

5.1.3 Preparation of eigenvoice-based resynthesized speech

We selected source characters from a well-performed voice actor and actress in his/her age of thirty, who individually gives us 3 distinct character voices, shown in Table 5.1. We selected 2 males and 2 females as source speakers on whose voices character conversion will be applied for in closed-gender experiment as shown in Table 5.2. All of voice artists and source speakers are Japanese native speakers and the recording was done in Japanese.

We estimated a weight vector using 24 utterances of each source speaker. Using this weight vector and the EV-GMM we built, any speech sample of the

Table 5.1: *F_0 mean and standard deviation (s.d.) of three characters of two voice artists.*

speaker	character	mean	s.d.
voice actor	A: original	159.31	39.95
	B: elderly man	117.88	27.20
	C: cheerful boy	200.90	48.79
voice actress	A: original	238.89	58.75
	B: young girl	296.01	70.08
	C: elderly woman	230.69	44.86

Table 5.2: *F_0 mean and standard deviation (s.d.) of source speakers.*

source speaker	mean	s.d.
male#1	144.66	48.01
male#2	141.97	37.82
female#1	237.16	46.20
female#2	231.80	54.62

reference speaker, MHT, can be converted to that source speaker. This process can be interpreted as generation process of utterances of the source speaker through EV-GMM using MHT as reference speaker. We call these utterances as "resynthesized" utterances through EV-GMM with reference speaker MHT.

The aim of the listening test to investigate whether our proposed method can do "character" conversion well by comparing resynthesized speech samples of a source speaker and those of the same speaker with target character, created through EV-GMM character conversion.

Since we have 3 different characters (A, B and C), there are 6 possible character conversion patterns; A→B, A→C, B→C and vice versa. For each conversion pattern, 3 different conversion methods (F0, X1 and X2) were applied to synthesize the sample pairs for subjective evaluation. To conclude, there are 18 (6 combinations x 3 conversion methods) sample pairs of a source character and a target character of the same speaker, and 72 sample pairs in total for 4 source speakers.

Table 5.3: *Correctness of subjects' judgment in the three conversion methods (percentage unit).*

1) Male source speakers conversion

	Japanese			Thai		
method	M	P	H	M	P	H
F0	13.9	8.3	77.8	20.8	13.7	65.5
X1	17.4	7.6	75.0	13.7	7.7	78.6
X2	10.4	6.3	83.3	17.9	10.7	71.4

2) Female source speakers conversion

	Japanese			Thai		
method	M	P	H	M	P	H
F0	11.1	27.8	61.1	25.0	29.8	45.2
X1	5.6	12.5	81.9	14.3	27.4	58.3
X2	4.2	11.1	84.7	19.0	26.2	54.8

Note: M='miss', P='partially miss', H='totally hit'

5.1.4 Listening test

The subjective evaluation is divided into two parts; the first 36 sample pairs of the two male source speakers and the remaining 36 samples of the two female source speakers. The subjects are 6 native Japanese, and 7 native Thai subjects who have no any linguistic knowledge of Japanese. We added the non-native speakers as subjects to observe whether either linguistic or non-linguistic factors might affect the discrimination among distinct characters.

To evaluate the conversion performance, firstly, a subject had to listen to resynthesized speech samples of the three characters (A,B and C) of the voice actor and paid very careful attention to what kind of character voice differences can be perceived in each of the 6 combinations.

Then, the first 36 sample pairs were randomly presented to a subject, and he/she had to judge what kind of conversion was done on each sample pair. There are six possible answers according to six conversion combinations. If a sample pair was a conversion result from character A to B, the correct answer will be "A →B" choice. Similarly, for the second half, each subject now compared

another 3 characters of resynthesized samples of the voice actress and judged the similarity of each of the remaining 36 sample pairs. It should be noted that we had told the subjects that the voice artist's resynthesized speech and the stimuli for listening test had different speaker identity.

5.1.5 Experimental results

The subjective evaluation results are shown in Table 5.3 as correctness of subjects' judgment. Since the number of candidates in each selection is 6, the chance level is 16.7%. Considering the correct judgment shown in the totally hit (H) column, we can say that the subjects can perceive well the change of character voices intended by experimenters. As shown in Table 5.1, each character has its own F_0 range and only F_0 conversion may be able to change the character of a speaker. Table 5.3 shows, however, that spectral conversion in addition to F_0 conversion improves the conversion performance both in the cases of male and female speakers.

In the case of conversion among the voice actress characters, the F_0-based approach has much lower performance because F_0 means of character A (original) and character C (elderly woman) are considerably similar to each other, which is thought to make it difficult for the subjects to distinguish between these two characters only by F_0 range. Especially for Thai subjects, in Table 5.4, the confusion matrix shows that most of the partially incorrect answers, the underlined numbers in diagonal cells, are found due to the misjudgment between character A and C. For instance, the presented sample pair is an A→B conversion, but the subjects judged it as a C→B one. That is, the subjects perceived character C as it was character A.

This means our proposed method is still effective with a very small F_0 change observed between different characters. This implies that only converting spectral features may increase the diversity of speaker character.

The sentences used for comparison were the same among different character pairs. So, we can assume that judgment of the subjects is mainly affected by extra-linguistic factors in the stimuli. In the case of Thai subjects, they cannot understand the linguistic content of the stimuli at all. However, when we look at the results from a cross-cultural viewpoint, it is quite obvious that the native

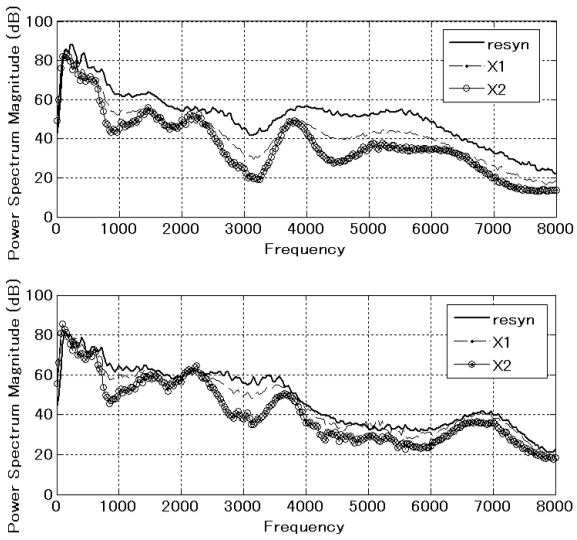

Figure 5.1: Examples of power spectral density of resynthesized and converted stimulus by "elderly man" character conversion. The upper one is that of source speaker male#1 and the lower one is that of source speaker male#2.

subjects have better performance and more precise discrimination among distinct character voices. This implies that the "acoustic" image of characters has both culture-universal and culture-specific aspects.

5.2 Acoustic evaluation and discussion

To notice how the spectral structure changes in each character conversion, we estimate the power spectral density (PSD) of the resynthesized stimuli of source speakers (male#1, male#2) and those of the character converted samples (A→B, A→C) with X1 and X2 conversion methods on the same source speaker. Note that, all of them have the same linguistic context and PSD can be considered as the long-time average of the spectrum along the time axis.

From the study of voice imitation and its perception, Elisabeth Z. [32] mentions that the human voice changes during an individual's lifetime. The voice of children has the highest F_0 , as well as formant frequency, but the aging one undergoes some changes in F_0 and be more breathy voice as a result of the loss

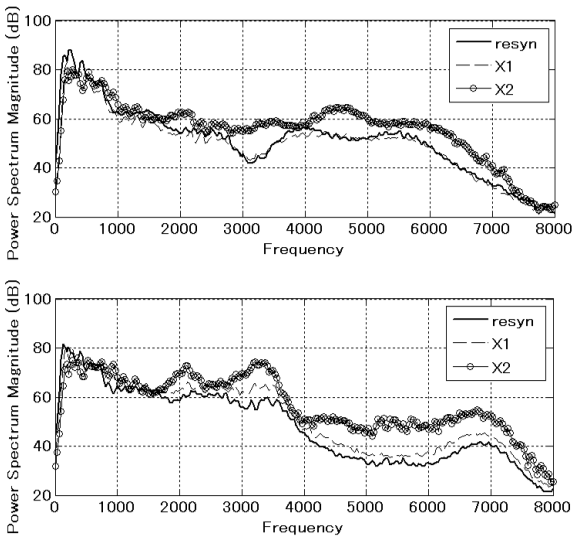

Figure 5.2: Examples of power spectral density of resynthesized and converted stimulus by "cheerful boy" character conversion. The upper one is that of source speaker male#1 and the lower one is that of source speaker male#2.

of control of laryngeal muscles.

In the case of character A to B conversion, namely "elderly man" character conversion, Figure 5.1 shows that the first, second and third formant (F1, F2, F3) stay constant or slightly change, and their amplitudes also decrease only slightly. However, the amplitudes at spectral valleys are considerably decreased by 5-20 dB corresponding to the strength of α coefficient. In other words, the formant frequencies of the converted samples, as well as their peak level, are somewhat unchanged, but the decrease of the amplitudes of the surrounding valleys might cause "elderly man" likeness, and make the subjects to perceive these stimulus as spoken by an elderly man.

Meanwhile, in the case of character A to C conversion, namely "cheerful boy" character conversion, , contrary to our expectation, Figure 5.2 shows that the formant frequencies are somewhat unchanged but their amplitudes increase by 5-10 dB. The peak levels of each surrounding valley are also going higher as the α coefficient increases.

Figure 5.3 shows another acoustic analysis result, which was done on the PSD of speech samples that were uttered by the same voice actor in the three characters

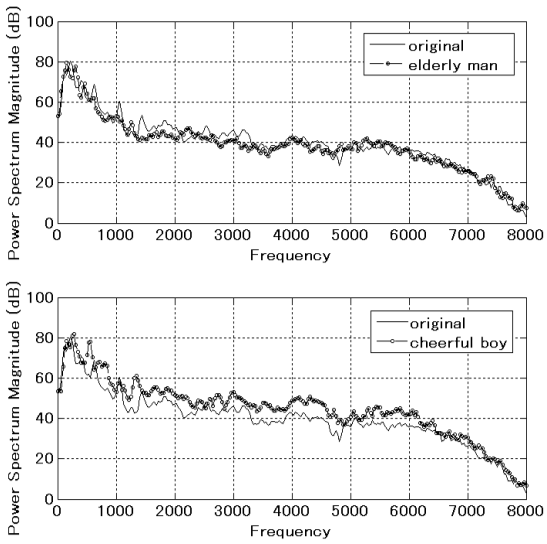

Figure 5.3: Examples of power spectral density of speech samples uttered by voice actor in three characters (original voice, elderly man and cheerful boy).

(original voice, elderly man and cheerful boy). It affirms that the modification patterns of spectral amplitudes found in synthetic speech are also found in that naturally created "elderly man" voices.

In spite of main focus on the analysis of formant frequency, the secondary spectral properties, that are the amplitude of spectral peaks and valleys, cannot be overlooked. Many studies [33, 34] claim that the vowel perception can be changed when increasing/decreasing the amplitudes of F1, F2 and F3, though the formant frequency is keeping unchanged. To increase the local spectral contrast that makes the formant peaks more prominent to other spectral peaks, without changing either formant frequency, is maybe another possible factor that affects the speaker character's perception of the listeners.

Table 5.4: *Confusion matrix of Thai subjects' answers on character conversion of female speakers (percentage unit).*

presented sample	subjects' answers (F0 method)					
	A→B	A→C	B→A	B→C	C→A	C→B
A→B	**50.0**	0	0	0	7.1	42.9
A→C	7.1	**50.0**	0	0	42.9	0
B→A	0	0	**42.9**	50.0	0	7.1
B→C	0	0	42.9	**50.0**	7.1	0
C→A	14.3	50.0	0	0	**35.7**	0
C→B	35.7	14.3	7.1	0	0	**42.9**

presented sample	subjects' answers (X1 method)					
	A→B	A→C	B→A	B→C	C→A	C→B
A→B	**57.1**	0	14.3	0	0	28.6
A→C	7.1	**71.4**	0	0	21.4	0
B→A	0	21.4	**57.1**	14.3	7.1	0
B→C	0	0	35.7	**64.3**	0	0
C→A	0	14.3	14.3	7.1	**64.3**	0
C→B	50.0	0	7.1	0	7.1	**35.7**

presented sample	subjects' answers (X2 method)					
	A→B	A→C	B→A	B→C	C→A	C→B
A→B	**50.0**	0	0	0	0	50.0
A→C	0	**50.0**	7.1	0	14.3	28.6
B→A	0	21.4	**50.0**	21.4	7.1	0
B→C	0	7.1	21.4	**71.4**	0	0
C→A	0	28.6	21.4	7.1	**42.9**	0
C→B	21.4	7.1	0	0	7.1	**64.3**

Chapter 6

Conclusions and future works

I propose a new method of character conversion by applying EV-GMM technique. Using training data of the various characters created by a single skilled voice actor/actress, the conversion can generate various kinds of characters from a single speaker. From the results of the listening test, we can say that human subjects can perceive well the change of character voices intended by experimenters even there is very small change of F_0 between two characters. Moreover, the spectral structure of converted samples has clearer local spectral contrast that make the formant peaks more prominent to their surrounding peaks, while formant frequencies are somewhat unchanged. The secondary spectral properties, e.q. formant amplitude, seem to work as important factors that can control the speaker character.

In this research, EV-GMM was used but it provided us with frame-wise conversion only. Long-span or dynamic speech features such as speaking style and temporal structure were not converted and those of the reference speaker (MHT) were found somewhat intact in converted speech. I plan to improve the naturalness of the converted speech by introducing the conversion with respect to these aspects of speech.

Acknowledgements

I would like to express my greatest gratitude to all the people who have helped and supported me throughout my research. I am heartily thankful to my supervisor, Prof. Nobuaki MINEMATSU, whose encouragement, guidance and support from the initial to the final level enabled me to develop more understanding of the speech technology. A speacial thank of mine goes to my colleagues and seniors who helped me in discussing their interesting ideas and thoughts. I would also like to extend my thanks to the technician of the laboratory, Mr. Noboru TAKAHASHI, for his support and advices during the corpus recording phase, and Mrs. Megumi IKEGAMI, the secretary of the laboratory, who always taking care of me about student affairs and helping me organize all the Japanese documents. I am indebted to my best mentor and Thai friends, Ms. Oraphan Krityakien and her term, who always being by my side sharing experiences, advices and encouragement. I could not enjoy and pass a two-years living in Tokyo without a warm support and care from my beloved Japanese mother, Mrs. Sumie NOUCHI. At last, I wish to thank my parents from Thailand for their love, smile and encouragement throughout my living and studing here in Japan.

References

[1] Sadaoki Furui. *Digital Speech Processing: Synthesis and Recognition*. 2001. 1, 5

[2] O. Cappe Y. Stylianou and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process*, 6(2):131–142, Mar. 1998. 1, 4

[3] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. ICASSP*, 1:285–288, 1998. 1, 2, 4

[4] A. Bonafonte H. Ney D. Sunderman, H. Hoge and J. Hirschbreg. Text-independent cross-language voice conversion. *Proc. INTERSPEECH*, 2006. 1, 4, 8

[5] T. Toda A. Moinet M. Charlier, Y. Ohtani and T. Dutoit. Cross-language voice conversion based on eigenvoices. *Proc. INTERSPEECH*, 2009. 1, 4

[6] N. Minematsu A. Kunikoshi, Y. Qiao and K. Hirose. Speech generation from hand gestures based on space mapping. *Proc. INTERSPEECH*, pages 308–311, 2009. 1, 4

[7] K. Shikano T. Toda, Y. Ohtani. Eigenvoice conversion based on gaussian mixture model. *Proc. INTERSPEECH*, 2006. 2

[8] P. Nguyen R. Kuhn, J-C. Junqua and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. Speech and Audio Processing*, 8(6):695–707, 2000. 2, 16

[9] Turk Matthew A. and Alex P. Pentland. Face recognition using eigenfaces. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1991. 2, 13

[10] Alan W. Black T. Toda and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(8), 2007. 4, 8, 9, 10

[11] E. Moulines and Eds. Y. Sagisaka. Voice conversion: State of the art and perspectives (special issue of speech communication). *Amsterdam, The Netherlands: Elsevier*, 16, 1995. 4

[12] H. Kawanami K. Shikano M. Mashimo, T. Toda and N. Campbell. Cross-language voice conversion evaluation using bilingual databases. *IPSJ*, 43(7):2177–2185, 2002. 4

[13] Alan W. Black and Kevin A. Lenzo. Multilingual text-to-speech synthesis. *ICASSP*, 2004. 4

[14] Raymond D. Kent. *The Acoustic Analysis of Speech.* Singular Publishing Group, Inc., 1992. 5

[15] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech and Signal Processing, IEEE Trans*, 34(1):52–59, 1986. 8

[16] S. Sagayama and F. Itakura. On individuality in a dynamic measure of speech. *Proc. ASJ Spring Conferrence*, 1979. 8

[17] K. Shikano M. Abe, S. Nakamura and H. Kuwabara. Voice conversion through vector quantization. . *Acoust. Soc. Jpn. (E)*, 11(2):71–76, 1990. 8

[18] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: Control and conversion. *Acoustics, Speech and Signal Processing, IEEE Trans*, 16. 8

[19] Voice conversion algorithm based on piecewise linear conversion rule of formant frequency and spectrum tilt. *Speech Commun.*, 16:153–164, 1995. 8

[20] B. Yegenanarayana A. W. Black S.Desai, E. V. Raghavendra and K. Prahallad. Voice conversion using artificial neural networks. *Proc. ICASSP*, pages 3893–3896, 2009. 8

[21] E. Moulines H. Valbret and J. P. Tubach. Voice transformation using psola technique. *Speech Commun*, 11(2-3):175–187, 1992. 8

[22] N. Iwahashi and Y. Sagisaka. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. *Speech Commun.*, 16(2):139–151, 1995. 8

[23] C.-H. Lee and C.-H Wu. Map-based adaptation for speech conversion using adaptation data selection and non-parallel training. *Proc. Interspeech2006-ICSLP*, pages 2254–2257, 2006. 12

[24] J. V. der Spiegel A. Mouchtaris and P. Mueller. Non-parallel training for voice conversion based on a parameter adaptation approach. *IEEE Trans. Audio, Speech and Language Processing*, 14(3):952–963, 2006. 12

[25] Y. Ohatani T. Toda and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoice. *IEEE Trans. ICASSP*, (4):1249–1252, 2007. 12, 18

[26] H. Saruwatari Y. Ohtani, T. Toda and K. Shikano. Many-to-many eigenvoice conversion with reference voice. *INTERSPEECH*, pages 1623–1626, 2009. 13, 19

[27] R. Kuhn et al. Eigenvoices for speaker adaptation. *Fifth International Conference on Spoken Language Processing*, 1998. 13

[28] et al A. Kurematsu. Atr japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9(4):357–363, 1990. 27

[29] Jnas: Japanese newspaper article sentences. *http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/instruct.html*. 27

[30] I. Masuda-Katsuse H. Kawahara and A. de 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*, 27(3-4):187–207, 1999. 28

[31] H. Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorpohic decomposition of speech sounds. *Acoustic Sci. and Tech*, 27, 2006. 28

[32] Elisabeth Z. *Voice imitation: A phonetic study of perceptual illusions and acoustic success*. Sweden, Studentlitteratur Lund. 32

[33] A. ito M. Ito, K. Ohara and M. Yano. An effect of formant amplitude in vowel perception. *Proc. INTERSPEECH*, 2010. 34

[34] T. Enright M. Kiefte and L. Marshall. The role of formant amplitude in the perception of /i/ and /u/. *J. Acoust. Soc. Amer.*, 127(4):2611–2621, April 2010. 34

Publications

Domestic Conferences

1. T. Pongkittiphan, N. Minematsu, D. Saito and K. Hirose, "Character conversion based on eigenvoice technique", Proc. Spring Meeting of the Acoustic Society of Japan, 2012-3.
2. T. Pongkittiphan, N. Minematsu, D. Saito and K. Hirose, "Eigenvoice-based character conversion and its evaluation", IEICE technical report. Speech, Vol. 112, No. 81, pp 7-12, 2012.

International Conference

1. T. Pongkittiphan, N. Minematsu, D. Saito and K. Hirose, "Eigenvoice-based character conversion", IEEE Workshop on Spoken Language Technology (SLT), 2012. (submitted)

Appendix A

Appendix

Figure A.1 shows the examples of Japanese sentences used for the reading material of character speech corpus recording.

a01: あらゆる現実を、すべて自分のほうへねじ曲げたのだ。
a02: 一週間ばかり、ニューヨークを取材した。
a03: テレビゲームやパソコンで、ゲームをして遊ぶ。
a04: 物価の変動を考慮して、給付水準を決める必要がある。
a05: 救急車が十分に動けず、救助作業が遅れている。
a06: 言論の自由は、一歩譲れば、百歩も千歩も攻め込まれる。
a07: 会場の周辺には、原宿駅や、代々木駅もあるし、ちょっと歩けば、新宿御苑駅もある。
a08: 老人ホームの場合は、健康器具や、ひざ掛けだ。
a09: ちょっと遅い昼食をとるため、ファミリーレストランに入ったのです。
a10: 嬉しいはずが、ゆっくり寝てもいられない。
---
a11: 自然の研究者は、自然をねじ伏せようとしてはいけない。
a12: おごりを捨て、謙虚な姿勢を取り戻さねば、冬は過ごせない。
a13: 先だって、ごく短期間だが、久方ぶりに、ヨーロッパへ行った。
a14: しかし、このプロ野球ブームも、永久に続くとは限らぬ。
a15: お客さんは、それじゃあ練習さえすれば、誰にでも出来るんじゃないかなっ、て考え始める
a16: アフリカ人は、実に巧みに、ぴゅんとつばを吐く。
a17: 前者を、普遍文化と呼び、後者を、個別文化と呼ぶことにする。
a18: 叔父さんは、岬の一軒家に独りぼっちで住んでいた。
a19: 立春が過ぎても、厳しい寒さの日々が続く。
a20: 大昔のフィリピンには、豊かな土地があった。

Figure A.1: The examples of sentences from the reading material.