Eigenvoice-based Approach to Voice Conversion and Voice Quality Control

Tomoki Toda
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan

Abstract

This paper reviews our proposed approach to voice conversion (VC) and voice quality control based on an eigenvoice technique. VC is a technique to modify nonlinguistic information such as speaker individuality while keeping linguistic information unchanged. In the traditional VC framework, a conversion model for a source and target speaker-pair needs to be trained in advance using a parallel data set consisting of utterance-pairs of these two speakers. To make VC technologies more practical, we have developed a new VC paradigm for flexibly building the conversion model for an arbitrary speaker-pair by effectively using speech samples of many other speakers. In this paper, we give an overview of eigenvoice conversion (EVC) as one of our proposed VC techniques.

1. Introduction

Speech enables us to convey both linguistic and nonlinguistic information simultaneously. Various phonemic sounds are generated by controlling our articulators to convey linguistic information. On the other hand, it is essentially difficult to intentionally control some nonlinguistic information such as speaker individuality by generating various speech sounds because voice quality is constrained by physical factors such as vocal tract length.

Voice conversion (VC), which is a technique for converting voice characteristics while keeping linguistic information unchanged, allows us to generate the desired voice quality beyond our physical constraints. One approach to VC is based on simple rules, e.g., time-invariant filtering, to modify speech parameters. Although this approach generally yields very high-quality converted speech, the converted voice quality strongly depends on the original one, and thus, it is indeed difficult to attain specific target voice quality. To overcome this problem, an alternative approach based on statistical process has been studied for around two decades [1].

In a statistical approach to VC, a conversion model for modifying speech parameters is trained in advance using a parallel data consisting of utterance-pairs (e.g., 50 utterance-pairs) of the source speaker and the target speaker. The trained model is capable of converting any utterance of the source speaker as if uttered by the target speaker [2]. One of the most popular VC methods is a probabilistic conversion method based on a Gaussian mixture model (GMM) [3]. Using a GMM of joint probability density of the source and the target speech parameters as the conversion model [4], the source speech parameters are converted into the target ones on the basis of the minimum mean-square error (MMSE) [3] or maximum likelihood estimation (MLE) [5]. Conversion performance of the GMM-based method has progressively been improved by various approaches, e.g., considering dynamic features and higher-order moment of speech parameters [5].

There have been many attempts at applying a statistical VC framework to various applications beyond speaker conversion, which has been originally studied as a VC application. To improve telephone speech quality, a VC technique has been used for bandwidth extension from narrow band speech into wide band speech [6]. To improve noise robustness, a probabilistic noise suppression technique has been successfully integrated into VC-based bandwidth extension [7]. As an attempt at eliminating external noise, a VC approach has been proposed for recovering clean speech sounds using speech signals recorded by a bone-conductive microphone [8]. It is also very effective for enhancing various types of body conductive speech recorded with another type of body conductive microphone [9], which is called Non-Audible Murmur (NAM) microphone developed for detecting not only normal speech but also very small speech signals useful in private talk [10]. A VC technique has further been applied to speaking aid, e.g., for making voices of total laryngectomees sound more natural [11] or improving intelligibility of dysarthric speech [12]. Moreover, there have been studied applications other than mentioned above, such as cross language VC [13], speech production [14], acoustic-to-articulatory inversion mapping [14], and so on. As shown in these examples, VC is a promising technique to make our speech communication more flexible and convenient.
However, there still remain several problems to be solved towards the practical use of VC applications. One of the biggest problems is the model training process using parallel data. To address this problem, two typical approaches to accepting non-parallel data in training have been proposed: one is based on the generation of (pseudo) parallel data from nonparallel data; the other is based on model adaptation techniques widely used in the speech recognition area.

In the nonparallel training approach based on the parallel data generation, frame alignment is performed between the source speech samples and the target speech samples in nonparallel data. This process is based on the state alignment of hidden Markov model (HMM) [15] or frame selection [16] with a unit selection algorithm [17]. The conversion model is trained using the generated parallel data in the same manner as the traditional framework. Hence this approach needs an equal or larger amount of training data compared with the conventional approach although the constraint of parallel data is alleviated.

In the other nonparallel training approach based on the model adaptation, one conversion model is trained in advance using an existing parallel data of a certain speaker-pair, and then, it is adapted to the desired source and target speakers. Mouchtaris et al. [18] proposed an adaptation method of the GMM of joint probability density based on maximum-likelihood stochastic-transformation [19]. Linear transforms for the source feature space and those for the target feature space are independently estimated using nonparallel data, and then they are applied to the GMM. As another approach, Lee and Wu [20] applied a training method of a constrained GMM [21] based on maximum a posteriori (MAP) estimation [22] to the nonparallel training. These methods suggest that the acoustic feature correlation between a certain speaker-pair is effectively used as prior knowledge for building the conversion model for another speaker-pair. Consequently, the model adaptation approach enables the use of a smaller amount of nonparallel data compared with the data generation approach.

Inspired by the model adaptation approach, we have developed a new paradigm for flexibly building the conversion model for arbitrary speakers using a large amount of speech data of many pre-stored speakers as prior knowledge. Two VC frameworks have hitherto been proposed: one is many-to-one VC capable of converting an arbitrary source speaker's voice into the target speaker's voice; the other is one-to-many VC capable of converting the source speaker's voice into an arbitrary target speaker's voice [23]. As one of methods for achieving these goals, we have proposed eigenvoice conversion (EVC) [24] by integrating an adaptation technique based on eigenvoices [25] into the GMM-based VC method. EVC also allows voice quality control of converted speech by manipulating small amount of parameters.

In this paper, we review EVC techniques and show their potential to make VC applications more practical.

2. Traditional Voice Conversion

As one of the state-of-the-art VC methods, we describe a conversion method based on MLE of speech parameter trajectories considering a global variance (GV) [5].

2.1. Training

Let us assume a source static feature vector

\[ x_t = [x_t(1), \ldots, x_t(D)]^T \]

and a target static feature vector

\[ y_t = [y_t(1), \ldots, y_t(D)]^T \]

at frame \( t \), respectively, where \( T \) denotes transposition of the vector. As a speech parameter vector, we use joint static and dynamic feature vectors

\[ X_t = [x_t^T, Ax_t^T]^T \]

for the source and

\[ Y_t = [y_t^T, Ay_t^T]^T \]

for the target, respectively.

Using parallel training data consisting of time-aligned source and target parameter vectors \([X_1^T, AX_1^T]^T, \ldots, [X_T^T, AX_T^T]^T\), the joint probability density of the source and target parameter vectors is modeled by a GMM [4] as follows:

\[
\begin{align*}
P(X_t, Y_t | \lambda) & = \\
= & \sum_{m=1}^{M} \alpha_m N\left( [X_t^T, Y_t^T]^T ; \mu_m^{(X,Y)}, \Sigma_m^{(X,Y)} \right) \\
\mu_m^{(X,Y)} & = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}, \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(XY)} & \Sigma_m^{(YY)} \end{bmatrix} \\
\end{align*}
\]

where \( N(\cdot; \mu, \Sigma) \) denotes the Gaussian distribution with a mean vector \( \mu \) and a covariance matrix \( \Sigma \). The mixture component index is \( m \). The total number of mixture components is \( M \). A GMM parameter set \( \lambda \) consists of weights \( \alpha_m \), mean vectors \( \mu_m^{(X,Y)} \), and covariance matrices \( \Sigma_m^{(X,Y)} \) for individual mixture components.

Furthermore, the probability density of the GV of the target static feature vectors over an utterance is also modeled by a Gaussian distribution,

\[
P(v_y | \lambda_y) = N(v_y; \mu_v, \Sigma_v)
\]

where the GV vector \( v_y = [v_y(1), \ldots, v_y(D)]^T \) is calculated by

\[
v_y(d) = \frac{1}{T} \sum_{t=1}^{T} (y_t(d) - \frac{1}{T} \sum_{t=1}^{T} y_t(d))^2.
\]

A parameter set \( \lambda_y \) consists of a mean vector \( \mu_v \) and a covariance matrix \( \Sigma_v \).
2.2. Conversion

Let us assume a time sequence of the source parameter vectors \(X = [X_1^T, \cdots, X_T^T]^T\) and that of the target parameter vectors \(Y = [Y_1^T, \cdots, Y_T^T]^T\). The converted static feature sequence \(\hat{y} = [\hat{y}_1^T, \cdots, \hat{y}_T^T]^T\) is determined by maximizing a product of the conditional probability density function of \(Y\) given \(X\) and the GV probability density function as follows:

\[
\hat{y} = \arg \max_y P(Y | X, \lambda)P(v_y | \lambda) \omega \tag{5}
\]

subject to \(Y = Wy\). \(\omega\) \(\in [0, 1]\) is fixed over \(T\), \(\alpha\) is determined by

\[
\hat{m} = \arg \max_m P(m | X, \lambda) \tag{8}
\]

And then, under the constraint shown by Eq. (6) the converted static feature sequence is determined by

\[
\hat{y} = \arg \max_y P(Wy | X, \hat{m}, \lambda)P(v_y | \lambda) \omega \tag{9}
\]

We iteratively update the converted sequence \(\hat{y}\) using the gradient method [5].

3. Many-to-One Voice Conversion

Many-to-One VC is a framework for converting an arbitrary speaker's voice into a pre-defined target speaker's voice. First, a conversion model capturing prior knowledge for building the conversion models for various source speakers is trained in advance using the multiple parallel data sets consisting of many pre-stored source speakers and the target speaker. And then, the trained conversion model is adapted to arbitrary source speakers using only a small amount of their speech data. Finally, the conversion process is performed using the adapted model.

We have hitherto proposed several many-to-one VC algorithms such as a method using a source speaker independent GMM [23], a method based on speaker selection [26], and EVC [23]. EVC is the most sophisticated method among them, and it is feasible to further apply various well-known techniques to EVC. To make it possible to instantaneously convert an input speech sample of a new source speaker, we have integrated the MAP algorithm into EVC adaptation process for rapidly adapting the conversion model to the source speaker using a very limited amount of speech data, e.g., less than one second [27].

In this section, we describe the many-to-one EVC algorithm with MAP adaptation.

3.1. Eigenvoice GMM (EV-GMM)

The joint probability density of the source and target feature vectors is modeled by an EV-GMM as follows:

\[
P(X, Y | \lambda^{(EV)}, w) = \sum_{m=1}^{M} \alpha_m N \left( [X_T^T, Y_T^T]^T ; \mu_m^{(X,Y)}(w), \Sigma_m^{(X,Y)} \right) \tag{10}
\]

where the mean vector \(\mu_m^{(X,Y)}(w)\) is given by

\[
\mu_m^{(X,Y)}(w) = \begin{bmatrix} \mu_m^{(X)}(w) \\ \mu_m^{(Y)}(w) \end{bmatrix} = \begin{bmatrix} B_m^{(X)}w + b_m^{(X)}(0) \\ \mu_m^{(Y)} \end{bmatrix} \tag{11}
\]

The source mean vector of the \(m\)th mixture component is represented as a linear combination of a bias vector \(b_m^{(X)}(0)\) and representative vectors \(B_m^{(X)} = [b_m^{(X)}(1), \cdots, b_m^{(X)}(J)]\), where the number of representative vectors is \(J\). The EV-GMM is adapted to individual source speakers by adjusting the \(J\)-dimensional weight vector \(w = [w(1), \cdots, w(J)]^T\) while the parameter set of the EV-GMM \(\lambda^{(EV)}\) is fixed over different source speakers.

Dominant voice characteristics extracted from many pre-stored source speakers (i.e., eigenvoices) are modeled by the representative vectors, and voice characteristics averaged over these speakers (i.e., average voice) are modeled by the bias vector. Seeing the EV-GMM from a different angle, several GMMs for converting each eigenvoice or the average voice into the target voice are interpolated based on the various weight values \(w(1), \cdots, w(J)\) for generating new GMMs for converting various source speakers' voices into the target voice. This idea is closely related to the speaker selection [26] and speaker interpolation [28].

3.2. Training

The EV-GMM is trained in advance using all of the pre-stored multiple parallel data sets. To build a canonical model yielding significant improvements of the model adaptation performance, we train the EV-GMM in Speaker Adaptive
Training (SAT) paradigm [29]. A set of the weight vectors \(\mathbf{w}^{(s)}_1 = \{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(5)}\}\) adapted to individual pre-stored source speakers as well as the tied parameter set \(\lambda^{(EV)}\) are estimated as follows:

\[
\hat{\lambda}^{(EV)} \hat{\mathbf{w}}^{(s)}_1 = \arg \max_{\lambda^{(EV)} \mathbf{w}^{(s)}_1} \prod_{i=1}^{S} \prod_{t=1}^{T} P(X^{(s)}_t, Y_t | \lambda^{(EV)}, \mathbf{w}^{(s)}_1) \quad (12)
\]

where \(X^{(s)}_t\) and \(Y_t\) are the parameter vector of the \(s^{th}\) pre-stored source speaker and that of the target speaker at frame \(t\) in their parallel data. More details are in [30].

It is noted that the correspondence of each mixture component into a phonemic space should be consistent for all pre-stored source speakers and the target speaker to keep linguistic information unchanged through the conversion. This consistency would be kept due to two constraints. One constraint is to fix the probability density function on the target space over all pre-stored source speakers in order to keep the target phonemic spaces modeled by the individual mixture components of the EV-GMM unchanged when adapting the EV-GMM to individual pre-stored source speakers. The other constraint is to optimize model parameters using parallel data sets in order to align each mixture component of the adapted EV-GMM to the same phonemic space on the individual source speakers as that on the target speaker. Consequently, a subspace representing speaker dependent features while not touching phonemic features is constructed by the representative vectors. This is regarded as a data-driven approach to modeling phonemic features and speaker dependent features separately without any phoneme transcriptions.

To consider prior information on the weight vector in the adaptation process, the probability density of the weight vector is modeled by a Gaussian distribution as follows:

\[
P(\mathbf{w} | \lambda_w) = \mathcal{N}(\mathbf{w}; \mu_w, \Sigma_w). \quad (13)
\]

A model parameter set \(\hat{\lambda}_w\) consisting of the mean vector \(\mu_w\) and the covariance matrix \(\Sigma_w\) is trained in advance using the set of the weight vectors for individual pre-stored source speakers \(\hat{\mathbf{w}}^{(s)}_1\) estimated in Eq. (12).

3.3. Unsupervised adaptation

The EV-GMM is adapted for an arbitrary source speaker by estimating the optimum weight vector for given speech samples of the source speaker in a completely unsupervised manner, i.e., using neither parallel data nor linguistic information. To estimate the weight vector, MAP adaptation is performed for a time sequence of the given source parameter vectors \(X' = [X_1^T, \ldots, X_T^T]^T\) as follows:

\[
\hat{w} = \arg \max_{w} P(w | X', \tau, \lambda)
\]

\[
= \arg \max_{w} P(w | \hat{\lambda}_w)^* P(X' | \lambda^{(EV)}_w, \mathbf{w})
\]

\[
P(X' | \lambda^{(EV)}_w, \mathbf{w}) = \int P(X', Y_t | \lambda^{(EV)}_w, \mathbf{w})dY_t \quad (15)
\]

where the hyper-parameter \(\tau\) is empirically determined. The MAP estimate is basically calculated by interpolation between the prior mean vector \(\mu_w\) and the ML estimate given by maximizing \(P(X' | \lambda^{(EV)}_w, \mathbf{w})\) where an interpolation rate dynamically changes according to the amount of adaptation data. For instance, if the adaptation data is not given, the MAP estimate of the weight vector is equal to the prior mean vector: i.e., the conversion model for average voice characteristics of various speakers is initially used. As the amount of adaptation data increases, the MAP estimate asymptotically approaches the ML estimate; i.e., the conversion model adapted to the given source speaker becomes available. In fact, the EV-GMM is capable of rapidly adapting to a new speaker (see also [27]).

Using both the adapted EV-GMM and the probability density function of the target GV also trained in advance, the conversion process is performed in a manner as in Section 2.2.

4. One-to-Many Voice Conversion and Voice Quality Control

One-to-Many VC is a framework for converting a pre-defined source speaker’s voice into an arbitrary speaker’s voice. The basic process is the same as many-to-one VC. An essential difference compared with many-to-one VC is that model parameters related to the target voice are adapted in one-to-many VC, and thus, converted speech quality is severely affected by model adaptation performance. In fact, the conversion methods based on speaker independent model and speaker selection are not effective in one-to-many VC although they work reasonably well in many-to-one VC. It is indispensable to use the conversion method capable of accurately estimating probability density of an arbitrary target speaker, such as EVC.

One of inherent properties of the eigenvoice technique is a capability to modify the probability density function by manipulating a small number of weighting parameters. In HMM-based speech synthesis, voice quality control based on eigenvoices has been developed [31]. This method has been further extended to the use of multiple regression HMMs for intuitively controlling voice quality of synthetic speech [32]. Inspired by these methods, we have developed voice quality control with a multiple regression GMM (MR-GMM) in the statistical VC framework [33].

In this section, we describe one-to-many EVC [23] and voice quality control based on the MR-GMM [33].
4.1. One-to-Many EVC

The joint probability density of the source and target parameter vectors is modeled by the EV-GMM of which mean vectors are given by

\[
\begin{bmatrix}
\mu_m^{(X)}(w) \\
\mu_m^{(Y)}(w)
\end{bmatrix}
=\begin{bmatrix}
\mu_m^{(X)} \\
\mu_m^{(Y)}(0)
\end{bmatrix} + \begin{bmatrix}
B_m^{(Y)} w \\
B_m^{(Y)} + b_m^{(Y)}(0)
\end{bmatrix}. \tag{16}
\]

Unlike the EV-GMM for many-to-one EVC, the target mean vectors are represented in the subspace for modeling speaker individuality. The model parameters are trained using all of multiple parallel data sets consisting of a single source speaker and many pre-stored target speakers in the SAT paradigm as described in Section 3.2.

The eigenvoice technique is also applied to the probability density of the GV [34] as follows:

\[
P(v_y | \lambda^{(EV)}_v, w_v) = N(v_y; \mu_v(w_v), \Sigma_v) \tag{17}
\]

\[
\mu_v(w_v) = B_v w_v + b_v(0). \tag{18}
\]

It is straightforward to apply SAT to training of a model parameter set \( \lambda^{(EV)}_v \) including \( B_v, b_v(0), \) and \( \Sigma_v. \)

For given speech samples of an arbitrary target speaker \( Y' = [Y'_1^T, \ldots, Y'_T^T]^T \), the unsupervised adaptation of the EV-GMM is performed on the basis of MLE of the weight vector as follows:

\[
\hat{w} = \arg \max_{w} \prod_{t=1}^{T} P(X_t, Y'_t | \lambda^{(EV)}, w). \tag{19}
\]

The weight vector for the GV is also determined by maximizing the model likelihood for the GV vectors calculated from the given target speech samples utterance by utterance as follows:

\[
\hat{w}_v = \arg \max_{w_v} \prod_{t=1}^{T} P(v_{y(t)} | \lambda^{(EV)}_v, w_v) \tag{20}
\]

where \( v_{y(t)} \) is the GV of the \( k \)-th utterance of the adaptation data and the number of utterances is \( K \). We may also employ the MAP adaptation as described in Section 3.3. The adapted EV-GMM works reasonably well even if using only a few utterances of the target speaker (see also [24, 34]).

4.2. Voice quality control based on multiple regression GMM (MR-GMM)

The one-to-many EV-GMM allows us to efficiently generate converted speech with various voice characteristics by manipulating the weight vector since the representative vectors are determined so that the constructed subspace widely covers voice characteristics of pre-stored target speakers. However, performance of voice quality control is not high enough in terms of the controllability because each representative vector doesn't usually represent voice characteristics specific to physical meaning.

In the MR-GMM, the weight vector is regarded as an explanatory variable and the target mean vectors are regarded as an objective variable. Its model structure is the same as the EV-GMM but the representative vectors are intentionally optimized so that each of them models voice characteristics described by each of voice quality expression words, e.g., gender (male - female) or age (younger - elder), which are designed so as to effectively describe voice quality of various speakers.

First, we manually assign perceptual scores on individual voice quality expression words to each pre-stored target speaker, e.g., using a 5-point scale such as gender (from -2: male to 2: female) and age (from -2: younger to 2: elder). And then, we develop the voice quality control vector consisting of the assigned scores \( w_c = [w_{c,1}(1), \ldots, w_{c,K}(K)]^T \) for each pre-stored target speaker, where the number of voice quality expression words is \( K \). Using all of pre-stored multiple parallel data sets and the developed voice quality control vectors, a parameter set of MR-GMM \( \lambda^{(MR)} \) is optimized as follows:

\[
\arg \max_{\lambda^{(MR)}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(X_{i,s}, Y_{i,s} | \lambda^{(MR)}, w_{c(s)}). \tag{21}
\]

where \( w_{c(s)} \) is the voice quality control vector for the \( s \)-th pre-stored target speaker. Unlike the EV-GMM training, the speaker-dependent parameters are fixed through training. The trained MR-GMM enables us to intuitively control converted voice quality by manipulating the voice quality control vector.

5. Conclusions

This paper reviews our proposed voice conversion (VC) techniques based on eigenvoices to give new VC frameworks; many-to-one VC converting an arbitrary source speaker's voice into the target speaker's voice and one-to-many VC converting vice versa. The proposed techniques also enable us to intuitively control converted voice quality by manipulating a small number of parameters. Although the proposed techniques make a statistical VC framework quite flexible, we still require speech samples of a specific speaker, i.e., the target speaker in many-to-one VC and the source speaker in one-to-many VC. Our next step is to develop many-to-many VC techniques.

Acknowledgement

This research was supported in part by MIC SCOPE. The author is grateful to Mr. Yamato Ohtani, Ms. Kumi Ohta, and Mr. Daisuke Tani of Nara Institute of Science and Technology, Japan, for their support in this research.

---

1. K. Ohta is currently with Brother Industries, Ltd., Japan.
2. D. Tani is currently with SHARP Corporation, Japan.
References


