An Evaluation of Many-to-One Voice Conversion Algorithms with Pre-Stored Speaker Data Sets

Daisuke Tani, Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{daisuke-t, yamato-o, tomoki, sawatari, shikano}@is.naist.jp

Abstract

This paper describes an evaluation of many-to-one voice conversion (VC) algorithms converting an arbitrary speaker’s voice into a particular target speaker’s voice. These algorithms effectively generate a conversion model for a new source speaker using multiple parallel data sets of many pre-stored source speakers and the single target speaker. We conducted experimental evaluations for demonstrating the conversion performance of each of the many-to-one VC algorithms, including not only the conventional algorithms based on a speaker independent GMM and on eigenvoice conversion (EVC), but also new algorithms based on speaker selection and on EVC with speaker adaptive training (SAT). As a result, it is shown that an adaptation process of the conversion model improves significantly conversion performance, and the algorithm based on speaker selection works well even when using a very limited amount of adaptation data.

Index Terms: voice conversion, many-to-one VC, EVC, SAT, speaker selection

1. Introduction

Voice conversion (VC) is a technique for converting an input speaker’s voice into another speaker’s voice while keeping the linguistic information [1]. VC is applied to various applications such as modification of the synthetic speech of Text-to-Speech [2], bandwidth extension of cellular speech [3], and body-transmitted speech enhancement [4]. One of the most useful VC applications is cross-language VC [5, 6], which converts speaker individuality across different two languages. This technique realizes a speech translation system or a CALL (Computer Assisted Language Learning) system synthesizing non-native language with user’s own voice.

Statistical approaches are often employed in VC, the most popular being the conversion method based on the Gaussian Mixture Model (GMM) [7]. A GMM representing joint probability density of the source and the target speech parameters is trained in advance using parallel data consisting of utterance pairs of the source and the target speakers. The trained GMM allows the determination of the target speech parameters for the given source speech parameters based on minimum mean square error (MMSE) estimation [7] or maximum likelihood estimation (MLE) [8], without any linguistic restrictions. Thus an arbitrary sentence uttered by the source speaker is rendered as a sentence uttered by the target speaker. One essential problem of this approach is the need for parallel data for model training. Moreover, several tens of phoneme-balanced sentences whose total duration is around 3 to 5 minutes are generally required to train the GMM sufficiently for conversion performance. These constraints clearly limit VC applications.

In order to alleviate the problem of parallel training, two main approaches have been proposed. One is parallel data generation from non-parallel data; the other is model adaptation with non-parallel data. The first approach conducts frame alignment between the source and the target voices based on HMM state alignment [9] or unit selection [10]. In that case, conventional model training is employed with the resulting parallel data. This approach requires only the source and the target speakers’ voices for the model training. Therefore, the same amount of training data is basically necessary. On the other hand, the latter approach uses other speaker’s voices as a prior knowledge for training the model for the desired speaker pair. Mouchtaris et al. [11] proposed a non-parallel training method based on maximum likelihood constrained adaptation. The GMM trained with an existing parallel data set of a certain source and target speaker pair is adapted for the desired source and target speakers separately. Lee et al. [12] proposed the adaptation method based on maximum a posteriori (MAP). In order to use a more reliable prior knowledge and reduce the amount of adaptation data, Toda et al. [13] proposed the eigen-voice conversion (EVC). EVC trains the eigenvoice GMM (EV-GMM) in advance using multiple parallel data sets consisting of utterance pairs of a single speaker and many pre-stored speakers. Effectively using the feature correlation between those speakers extracted from pre-stored parallel data sets enables unsupervised adaptation of EV-GMM for the desired speaker using only a few arbitrary sentences.

There are two novel VC frameworks to which EVC has been applied, i.e., one-to-many VC and many-to-one VC [14]. One-to-many VC converts the particular source speaker’s voice into arbitrary speaker’s one. On the other hand, many-to-one VC converts arbitrary speakers’ voices into the particular speaker’s voice. This paper focuses on many-to-one VC. It enables the conversion of any language uttered by an arbitrary speaker as utterances of the specific target speaker. It has been reported that not only the EV-GMM but also the source-independent GMM (SI-GMM) works in many-to-one VC [14] without any adaptation processes, but simply trained simultaneously using multiple parallel data sets of many pre-stored source speakers and the single target speaker. In this paper, we propose another many-to-one VC method based on speaker selection [15]. And, we introduce speaker adaptive training (SAT) [16] into EVC to further improve conversion performance. These many-to-one VC methods are compared with each other in both objective and subjective evaluations.

This paper is organized as follows. In Section 2, we describe the Many-to-One VC algorithms. In Section 3, we describe an experimental evaluation. Finally, we summarize this paper in Section 4.
2. Many-to-One VC algorithms
We describe four many-to-one VC algorithms based on 1) SI-GMM, 2) speaker selection, 3) EVC, 4) EVC with SAT. The difference among these algorithms is only in the way of constructing the conversion model for a given new source speaker. Every algorithm employs the MLE-based conversion method [8] using the resulting conversion model.

2.1. Many-to-One VC based on SI-GMM [14]
We use 2D-dimensional acoustic features, $X_i^{(s)} = [x_i^{(s)}; \Delta x_i^{(s)}]_T$ (the $s$th source speaker’s), and $Y_i = [y_i; \Delta y_i]_T$ (target speaker’s), consisting of $D$-dimensional static and dynamic features, where $T$ denotes transposition of the vector. Joint probability density of $Z_i^{(s)} = [X_i^{(s)}; Y_i]_T$ consisting of time-aligned source and target features determined by DTW is modeled with a GMM as follows:

$$P(Z_i^{(s)} | \lambda) = \sum_{t=1}^{M} \alpha_t N(Z_i^{(s)}; \mu_i^{(Z)}, \Sigma_i^{(ZZ)})$$

$$\mu_i^{(Z)} = \begin{bmatrix} \mu_i^{(XX)} \\ \mu_i^{(YY)} \end{bmatrix}, \Sigma_i^{(ZZ)} = \begin{bmatrix} \Sigma_i^{(XX)} & \Sigma_i^{(XY)} \\ \Sigma_i^{(YX)} & \Sigma_i^{(YY)} \end{bmatrix}$$

where $N(x; \mu, \Sigma)$ shows the normal distribution with a mean vector $\mu$ and a covariance matrix $\Sigma$. The $i^{th}$ mixture weight is $\alpha_i$. The total number of mixtures is $M$. This paper employs diagonal covariance matrices for individual blocks, $\Sigma_i^{(XX)}, \Sigma_i^{(YY)}$, and $\Sigma_i^{(XY)}$.

The SI-GMM is trained with all multiple parallel data sets consisting of utterance-pairs of multiple pre-stored source speakers and one target speaker, as follows:

$$\lambda^{(0)} = \arg \max_{\lambda} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(Z_i^{(s)} | \lambda)$$

where $S$ is the number of pre-stored source speakers. Figure 1 shows the previous training process. In the conversion, the SI-GMM is directly used without any adaptation processes.

2.2. Many-to-One VC based on Speaker Selection
It is well known that phonemic spaces of a certain speaker often overlap with those of another speaker. Therefore, the SI-GMM might cause a conversion error especially for source speakers whose voice characteristics are quite different from those of the average voice among the pre-stored source speakers. Therefore, a model adaptation process for each source speaker is useful for alleviating this problem.

Speaker selection is one of the model adaptation techniques. Figure 2 shows the previous training and the adaptation processes in many-to-one VC based on speaker selection. The conversion model is trained not with all parallel data sets but with only those consisting of the pre-stored source speakers whose voice characteristics are similar to those of the given source speaker. In order to reduce considerably the computational cost for this adaptation, we employ the single EM update of the SI-GMM, using pre-calculated sets of sufficient statistics for individual speaker-pairs. Note that this process allows unsupervised adaptation.

2.2.1. Previous training
A set of sufficient statistics for each speaker pair is computed with each parallel data set and the SI-GMM as follows:

$$\bar{\gamma}_i^{(s)} = \sum_{t=1}^{T_s} P(m_i | Z_i^{(s)}; \lambda^{(0)})$$

$$\bar{Z}_i^{(s)} = \sum_{t=1}^{T_s} P(m_i | Z_i^{(s)}; \lambda^{(0)}) Z_i^{(s)}$$

Moreover, individual source dependent GMMs (SD-GMMs) are trained using each of the calculated sets of sufficient statistics for individual speaker pairs.
2.2.2. Unsupervised Adaptation

Firstly, a likelihood of each SD-GMM for given adaptation data of the new source speaker $X^{(org)}$ is calculated as follows:

$$ L^{(s)} = \int P(X^{(org)}, Y|\lambda^{(s)}) dY. \quad (7) $$

Then, the likelihoods of individual SD-GMMs are sorted and the top $N$ speaker-pairs are selected. The conversion model is generated using the N-best sets of sufficient statistics for the selected speaker pairs as follows:

$$ \hat{\alpha}_i = \sum_{s \in S_N} \gamma^{(s)}_i, \quad (8) $$

$$ \hat{\mu}_i = \sum_{s \in S_N} Z^{(s)}_i \sum_{s \in S_N} \gamma^{(s)}_i, \quad (9) $$

$$ \hat{\Sigma}_i = \sum_{s \in S_N} V^{(s)}_i \sum_{s \in S_N} \gamma^{(s)}_i, \quad (10) $$

where $\hat{\alpha}_i$, $\hat{\mu}_i$, and $\hat{\Sigma}_i$ are the updated mixture weight, mean vector and covariance matrix, and $S_N$ is a set of the N-best speakers.

2.3. Many-to-One VC based on EVC [14]

2.3.1. Eigenvoice Gaussian Mixture Model (EV-GMM)

The mean vector of the EV-GMM for many-to-one VC is given by

$$ \mu_i^{(Z)} = \left[ B_i^{(X)} w + b_i^{(X)}(0) \right], \quad (11) $$

The source mean vector for the $i^{th}$ source is given by

$$ \mu_{org}^{(Z)} = \left[ b_i^{(X)}(1), b_i^{(X)}(2), \ldots, b_i^{(X)}(J) \right]. $$

The number of the representative vectors $J$ is fixed. The speaker source individuality is controlled with only the $J$-dimensional weight vector $w = [w(1), w(2), \ldots, w(J)]^\top$.

2.3.2. Training of EV-GMM

Figure 3 shows the previous training and the adaptation processes of many-to-one VC based on the EVC. Each SD-GMM is trained by updating only source mean vectors of the SI-GMM using each of the multiple parallel data sets. As a source dependent parameter, a supervector for each pre-stored source speaker is constructed by concatenating the source mean vectors of each of the SD-GMMs. The bias and representative vectors, i.e., eigenvectors, are determined with principal component analysis (PCA) for all source speakers’ supervectors. Finally, the EV-GMM is constructed from the resulting bias and representative vectors and parameters of the SI-GMM.

2.3.3. Unsupervised Adaptation of EV-GMM

The EV-GMM is adapted for arbitrary speakers by estimating the optimum weight vector for given speech samples without any linguistic information. The weight vector is estimated so that the likelihood of the marginal distribution for a time sequence of the given source features $X^{(org)}$ is maximized as follows:

$$ \hat{w} = \arg \max_w \int P(X^{(org)}, Y|\lambda^{(EV)}) dY. \quad (12) $$

This estimation is performed with EM algorithm.

2.4. Many-to-One VC based on EVC with SAT

The tied parameters of the PCA-based EV-GMM are from the SI-GMM. They are affected by acoustic variations of many pre-stored source speakers. Especially, source covariance values are much larger than those of the SD-GMM. They would cause performance degradation of the adapted EV-GMM.

In order to train an appropriate canonical EV-GMM, we apply speaker adaptive training (SAT) to the EV-GMM training. Figure 4 shows the previous training and the adaptation processes of many-to-one VC based on the EVC with SAT. The canonical EV-GMM is trained by maximizing the following likelihood of the adapted models for individual pre-stored source speakers,

$$ \lambda^{(EV)}(\hat{w}_s) = \arg \max_{\lambda} \prod_{s=1}^{S} \prod_{i=1}^{T_s} P(Z^{(s)}_i|\lambda^{(EV)}(w_s)), \quad (13) $$

where $\lambda^{(EV)}(w_s)$ denotes the adapted model for the $s$-th pre-stored source speaker with the weight vector $w_s$. SAT estimates both canonical EV-GMM parameters $\lambda^{(EV)}$ and a set of weight vectors for pre-stored source speakers $\hat{w}_s = [w_{s1}, w_{s2}, \ldots, w_{sj}]$. The estimation is performed with EM algorithm by maximizing the following auxiliary function:

$$ Q(\hat{\lambda}^{(EV)}(\hat{w}_s), \hat{\lambda}^{(EV)}(\hat{w}_s)) = \sum_{s=1}^{S} \sum_{i=1}^{M} \gamma^{(s)}_i \log P(m_s, Z^{(s)}_i|\hat{\lambda}^{(EV)}(w_s)), $$

(14)
where

$$z_i^{(s)} = \sum_{t=1}^{T_s} P(m_t|Z_i^{(s)}, \lambda^{(EV)}(w_s)).$$

It is difficult to update all parameters simultaneously because some depend on others. Therefore, each parameter of EV-GMM is updated as follows:

$$Q(\lambda^{(EV)}(w_1^n), \lambda^{(EV)}(w_1^n))$$
$$\leq Q(\lambda^{(EV)}(w_1^n), (\hat{\omega}_i^n, B_i^n, b_i^n(0), \mu_i^n, \gamma_i^n, \alpha_i, \Sigma_i^{(zz)}))$$
$$\leq Q(\lambda^{(EV)}(w_1^n), (\hat{\omega}_i^n, B_i^n, b_i^n(0), \mu_i^n, \gamma_i^n, \alpha_i, \Sigma_i^{(zz)})).$$

The ML estimate of the weight vector for the s-th pre-stored source speaker is written as

$$\hat{w}_s = \left( \sum_{i=1}^{M} \gamma_i^{(s)} B_i^{(X)} Y_i^{(s)} \right)^{-1} \sum_{i=1}^{M} \gamma_i^{(s)} B_i^{(X)} Y_i^{(s)} - \gamma_i^{(s)} B_i^{(X)}$$
$$+ B_i^{(X)} P_i^{(XX)} (X_i^{(s)} - \gamma_i^{(s)} b_i^{(X)}(0))).$$

(15)

where

$$Z_i^{(s)} = \begin{bmatrix} X_i^{(s)} \\ Y_i^{(s)} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T_s} P(m_t|Z_i^{(s)}, \lambda^{(EV)}(w_s)) X_i^{(s)} \\ \sum_{t=1}^{T_s} P(m_t|Z_i^{(s)}, \lambda^{(EV)}(w_s)) Y_i^{(s)} \end{bmatrix}.$$
3.2. Objective evaluations

We conducted objective evaluations using the mel-cepstral distortion between the converted and the target mel-cepstra as an evaluation measure. The averaged distortion over all test speakers was 8.11 [dB] before the conversion.

Figure 5 shows the mel-cepstral distortion as a function of the number of selected pre-stored speakers in many-to-one VC based on speaker selection. Note that results when selecting 160 pre-stored speakers are the same as those in many-to-one VC based on SI-GMM. The adaptation method based on speaker selection improves conversion accuracy, compared with the method based on the SI-GMM, because the adapted GMM models an acoustic space of the given source speaker more properly than does the SI-GMM. We can see that the best conversion accuracy is achieved when selecting around 20 to 40 pre-stored speakers. Rapid degradation of conversion accuracy is observed when setting the number of the selected speakers much smaller. Using directly the conversion model for another speaker does not obviously work even if, among the pre-stored speakers, his voice characteristics are the most similar to those of the given source speaker. Therefore, it is important to construct the conversion model properly, covering the acoustic space of the given source speaker by mixing multiple speakers’ data sets. The number of selected speakers is set to 27 in many-to-one VC based on speaker selection in the following evaluations.

Figure 6 shows mel-cepstral distortion as a function of the number of mixtures. Each many-to-one VC algorithm with the adaptation process outperforms that based on the SI-GMM. Many-to-one VC based on speaker selection allows the adaptation of every parameter of the conversion model. However, its adaptation mechanism is rougher than that of EVC, in terms of using a constant rate of mixing data sets of the selected pre-stored speakers. On the other hand, EVC estimates the best mixing rate, i.e., weights for eigenvectors, in the sense of ML, although it allows only the adaptation of source mean vectors. Consequently the conversion performance in speaker selection is comparable to that in EVC. SAT optimizes tied parameters of the EV-GMM considering the adaptation process. Therefore, the performance of EVC is obviously improved by applying SAT into the EV-GMM training.

Figure 7 shows the mel-cepstral distortion as a function of the number of adaptation sentences. The number of mixtures is set to 128.

In the 1/32 sentence, only one of 32 blocks into which one sentence is divided is used as adaptation data. It includes only around 32 frames, whose total duration is around 0.16 seconds. EVC obviously improves performance compared with the SI-GMM, when using one or more adaptation sentences. Introducing SAT into EVC improves further performance. However, when decreasing the adaptation data less than one sentence, conversion accuracy starts to degrade rapidly due to over-adaptation. Although reducing the number of representative vectors alleviates the over-adaptation problem, similar degradation tendencies were still observed in another experiment which is not shown here. On the other hand, speaker selection still keeps the performance improvements compared with the SI-GMM even when decreasing the adaptation data less than one sentence. Compared with EVC, speaker selection is more robust against the amount of adaptation data. These results show that the best adaptation method differs according to the amount of adaptation data.
Many-to-one VC algorithms

![Graph showing preference score on speech quality for different methods.](image)

Figure 8: Result of subjective evaluation in many-to-one VC.

### 3.3. Subjective evaluation

We conducted subjective evaluation of speech quality of the converted voices. Four many-to-one VC methods were evaluated in the preference test. The number of mixtures was set to 128. The number of adaptation sentences was set to 2. The number of subjects was 6. We randomly presented a pair of the converted voices from two different methods. The subjects were asked which sample sounded more natural. Each subject evaluated 120 sample-pairs including every pair-combination of the four methods.

Figure 8 shows the result of the preference test. Three adaptation methods significantly improve the converted speech quality compared with the method based on the SI-GMM. Each adaptation process alleviates unstable sounds of the converted speech sometimes caused by the SI-GMM. This result is very similar as shown in the objective evaluations.

### 4. Conclusions

This paper described many-to-one voice conversion (VC) algorithms that convert an arbitrary speaker’s voice into a particular target speaker’s voice. We conducted an experimental evaluation of many-to-one VC algorithms, using not only the conventional methods based on the source independent GMM (SI-GMM) and on EVC, but also two new methods based on speaker selection and EVC with speaker adaptive training (SAT). Results of objective and subjective evaluations showed that, in many-to-one VC, the adaptation process results in a better conversion model than the SI-GMM. Moreover, an algorithm based on speaker selection worked well with very little amount of adaptation data.

### 5. Acknowledgements

This research was supported in part by MEXT’s (the Japanese Ministry of Education, Culture, Sports, Science and Technology) e-Society project.

### 6. References


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