Text-independent voice conversion using speaker model alignment method from non-parallel speech

Peng Song\textsuperscript{1,2}, Yun Jin\textsuperscript{1}, Wenming Zheng\textsuperscript{3}, Li Zhao\textsuperscript{1}

\textsuperscript{1} School of Information Science and Engineering, Southeast University, Nanjing, P.R. China
\textsuperscript{2} School of Computer and Control Engineering, Yantai University, Yantai, P.R. China
\textsuperscript{3} Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing, P.R. China

\{pengsong, jinyun, wenming.zheng, zhaoli\}@seu.edu.cn

Abstract

In this paper, we propose a novel voice conversion method called speaker model alignment (SMA), which does not require parallel training speech. Firstly, the source and target speaker models, described by Gaussian mixture model (GMM), are trained, respectively. Then, the transformation function of spectral features is learned by aligning the components of source and target speaker models iteratively. Additionally, the transformation function is further combined with GMM, enabling the multiple local mappings, and a local consistent GMM (LCGMM) is also considered for model training to improve the conversion accuracy. Finally, we carry out experiments to evaluate the performance of the proposed method. Objective and subjective experimental results demonstrate that compared with the well-known INCA approach, the proposed method achieves lower spectral distortions and higher correlations, and obtains a significant improvement in perceptual quality and similarity.

Index Terms: voice conversion, speaker model alignment, INCA alignment, non-parallel speech

1. Introduction

The goal of voice conversion is to transform the characteristics of a source speaker to those of another speaker without changing the content information. In the past decades, this research has drawn more and more attention because of its potential applications, high-quality voice conversion may help to construct flexible text-to-speech (TTS) synthesis systems, and also can be applied to speech enhancement, speech-to-speech translation, etc.

To realize efficient voice conversion and obtain high-quality converted speech, many statistical approaches have been presented, such as Gaussian mixture model (GMM)\textsuperscript{1, 2}, artificial neural networks (ANN)\textsuperscript{3}, and codebook mapping methods\textsuperscript{4}. From the point of view of efficiency and practical applications, GMM is the most popular method for voice conversion. However, it has some main shortcomings, such as over-smoothing and overfitting. To tackle these problems, plenty of approaches have been proposed. E.g., Toda et al.\textsuperscript{5} introduce the global variance (GV) to solve the over-smoothing problem, Helander et al.\textsuperscript{6} propose a partial least squares regression (PLSR) approach to address the overfitting problem.

Most of current voice conversion approaches are conducted under the assumption that parallel data is available, however, it is not easily obtainable in practice. To address this problem, some research on voice conversion using non-parallel speech has been done over the last decade. The model adaptation algorithms, previously popular in speech recognition and TTS fields, are employed for non-parallel voice conversion in\textsuperscript{7, 8}, but these methods depend on the pre-trained reference speaker models. Ye et al.\textsuperscript{9} present a general maximum likelihood (ML) approach for the estimation of transformation function, but it greatly depends on the accuracy of target speaker model.

Different from traditional voice conversion approaches using parallel data, the dynamic time warping (DTW) algorithm\textsuperscript{1, 2, 3, 4} cannot be directly applied for feature alignment, and some methods have been presented for non-parallel data alignment, E.g., Tao et al.\textsuperscript{10} propose a supervisory data alignment algorithm, but it requires the prior phonetic information, which cannot be directly achieved in practice. Erro et al.\textsuperscript{11} propose an Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) method, which can be extended to combine with any of the existing voice conversion approaches. However, the INCA algorithm is conducted under the assumption that the spectral feature vectors close to each other in feature space belong to the same phoneme, which is not always accurate in practical situations.

In this paper, as most of current voice conversion approaches, we will focus on the transformation of spectral features, and propose a novel speaker model alignment (SMA) method for voice conversion using non-parallel speech. Different from the above-mentioned non-parallel based voice conversion approaches, the presented method requires neither any text-dependent information nor pre-trained reference speaker models. To satisfy these requirements and address the problem of INCA algorithm, in our approach, firstly, we train the source and target speaker models, respectively. Then, we can obtain a transformation function of spectral features by aligning the speaker models iteratively. Finally, we further combine it with GMM method, and also use a local consistent GMM (LCGMM) for model training to improve the conversion performance.

The remainder of this paper is structured as follows. In Section 2, we briefly introduce the baseline INCA approach. In Section 3, we present our proposed SMA approach, and extend it to the GMM based voice conversion. Section 4 gives and discusses the experimental results. Finally, we draw the conclusions in Section 5.

2. Baseline INCA algorithm

To solve the frame alignment problem for text-independent voice conversion using non-parallel speech, an iterative acoustic
feature alignment algorithm namely INCA is proposed by Erro [11]. Let \( X = \{x_j\}_{j=1,2,...,J} \) and \( Y = \{y_k\}_{k=1,2,...,K} \) be the spectral feature sequences of source and target speakers, respectively, the alignment procedure can be stated as follows:

1) **Initialization.** The transformation function \( F(x) \) is initialized as

\[
F(x) = x
\]

An auxiliary vector set \( X' = \{x'_j\}_{j=1,2,...,J} \) is created, and initialized by applying the transformation function to \( X \), which is given by

\[
x'_j = F(x_j)
\]

2) **nearest neighbor (NN) search.** For each vector \( x_j \), its closest index in \( Y \) is found and stored as \( p(j) \). Similarly, the index of \( q(k) \) of each \( y_k \) is found in \( X \).

\[
p(j) = \arg \min_k d(x_j, y_k), \quad q(k) = \arg \min_j d(x'_j, y_k)
\]

where \( d(\cdot) \) is the acoustic spectral distance between auxiliary and target vectors.

3) **Training.** The paired vectors \( \{x_j, y_{p(j)}\} \) and \( \{x'_j, y_{q(j)}\} \) are concatenated, and described by a joint vector set \( Z \). By employing the expectation maximization (EM) algorithm, the GMM of \( Z \) will be trained, and takes the form as

\[
p(Z) = \sum_{i=1}^{M} \alpha_i N(\mu_i, \Sigma_i)
\]

where \( \alpha_i \) is the weight satisfying \( \sum_{i=1}^{M} \alpha_i = 1 \), \( N(\cdot) \) denotes a Gaussian distribution, and \( \mu_i \) and \( \Sigma_i \) are the mean vectors and covariance matrices, respectively.

\[
\mu_i = \begin{bmatrix} \mu_{ix}^p \\ \mu_{iy}^p \end{bmatrix}, \quad \Sigma_i = \begin{bmatrix} \Sigma_{ix}^{xx} & \Sigma_{ix}^{xy} \\ \Sigma_{iy}^{xy} & \Sigma_{iy}^{yy} \end{bmatrix}
\]

By using the minimum mean square error (MMSE) algorithm, the auxiliary transformation function can be calculated

\[
F_{aux}(x) = \sum_{i=1}^{M} p(i|x) \left( \mu_{ix}^p + \Sigma_{ix}^{xy}(x - \mu_{ix}^p) \right)
\]

where \( p(i|x) \) is the posterior probability that \( x \) belongs to the \( i \)-th component

\[
p(i|x) = \frac{\alpha_i N(x, \mu_{ix}^p, \Sigma_{ix}^{xx})}{\sum_{r=1}^{M} \alpha_r N(x, \mu_{ir}^p, \Sigma_{ir}^{xx})}
\]

4) **Transformation.** By applying the transformation function to the spectral features of source speaker, the auxiliary vector set \( X' \) will be updated

\[
x'_j = F_{aux}(x_j)
\]

5) **Iteration.** Repeat steps 2) ~ 4) until the convergence is reached.

3. Speaker model alignment for voice conversion

3.1. Speaker model alignment

The INCA approach is carried out on the assumption that spectral features in feature space close to each other belong to the same phoneme, however, this assumption is not always accurate in reality. Meanwhile, the initialization of transformation function in INCA algorithm is too simple, which cannot ensure the accuracy of converted speech.

In this paper, different from INCA method, we choose the similarities between different distributions of spectral features for distance measurement, and propose a SMA method for feature alignment and estimation of the transformation function.

The SMA algorithm consists of the following steps:

1) **Model training.** By using the EM algorithm, we can train the source and target speaker models GMM\(_X\) and GMM\(_Y\), denoted as \( (x, \mu_x, \sigma_x) \) and \( (y, \mu_y, \sigma_y) \), respectively.

2) **Model initializing.** Given an auxiliary vector set \( X' = \{x'_j\}_{j=1,2,...,J} \), where \( x'_j \) is initialized as \( x_j \), it will be used as the intermediate set after each iteration. An intermediate speaker model GMM\(_{X'}\) is defined, and its model parameters are initialized as \( (x, \mu_x, \sigma_x) \).

3) **Model similarity calculation.** The similarities between different components of GMM\(_{X'}\) and GMM\(_Y\) are calculated, and a \( M \times M \) similarity matrix will be obtained.

\[
D = [d_{st}]_{M \times M}, \quad s, t = 1, 2, ..., M
\]

where \( d_{st} \) is the similarity between the \( s \)-th component of source speaker model and the \( t \)-th component of target speaker model. The Kullback-Leibler divergence (KLD) is adopted to measure the similarities, given two distributions \( P_s(c) \) and \( P_t(c) \), the KLD between them is computed as

\[
D(P_s(c)||P_t(c)) = \sum_c P_s(c) \log \frac{P_s(c)}{P_t(c)}
\]

The above function is not symmetric, and a symmetric function is adopted

\[
d_{st} = \frac{1}{2} \left[ D(P_s(c)||P_t(c)) + D(P_t(c)||P_s(c)) \right]
\]

4) **NN search.** For each component \( s \) of GMM\(_{X'}\), the index of the most similar component in GMM\(_Y\) is found, and marked as \( \sigma(s) \). Meanwhile, the index \( \Psi(t) \), the nearest component in GMM\(_{X'}\), for each component \( t \) of GMM\(_Y\) is found.

5) **Transformation.** An aligned mean vector sequence is established by concatenating \( \{\mu_{ix}^p, \mu_{iy}^p, \mu_{ix}^o, \mu_{iy}^o\} \) and \( \{\mu_{ix}^o, \mu_{iy}^o\} \). By employing the least square estimation (LSE) algorithm, a corresponding transformation function between \( \mu_{ix}^o \) and \( \mu_{iy}^o \) is computed, and given by

\[
f_{mn}(\mu_{ix}^o) = A \mu_{ix}^o + b
\]

where \( A = \mu_{iy}^o \mu_{ix}^o \), and \( b = \mu_{iy}^o - A \mu_{ix}^o \). \( f_{mn}(x') \) is directly utilized to transform the auxiliary spectral features, and the new intermediate vector set \( X' \) is set as

\[
x' = f_{mn}(x')
\]

We can also easily find the corresponding linear relationships between \( x \) and \( x' \), and denoted as \( x' = f_{mn}(x) \).
6) **Iteration.** When the similarities between two speaker models are reached to a certain threshold, the alignment procedure is stopped. Otherwise, the GMM is to be retrained, and the steps 3) ~ 5) will be repeated.

### 3.2. Combining SMA with GMM

The SMA approach can convert the source speech to the target one to some extent. However, it cannot ensure the accuracy of the converted speech, on one hand, a single linear transformation is unlikely effective to all data [6]. On the other hand, the training of transformation function cannot completely utilize the information of spectral features. So a GMM based SMA method is further presented.

After the last step of SMA, we will obtain the final values of the auxiliary vector set $X'$. As is shown in the steps 2) and 3) of INCA algorithm, a joint density GMM (JD-GMM) is trained using aligned auxiliary and target feature sequences, and the transformation function between auxiliary and target vector sets can be shown as

$$F_{au}(x') = \sum_{i=1}^{M} p(i|x') (\mu_i + \sum_{y} \sum_{x} \gamma_{xy} (x' - \mu_i))$$

(14)

where the definitions of the parameters can be referred to Eqs.(5) and (7).

So the transformation function between spectral features of source and target speakers is given as follows:

$$y = F(x) = F_{au}(f_{sa}(x))$$

(15)

### 3.3. Local consistent GMM training

In GMM based voice conversion method, a local linear transformation is estimated for each component, and the final transformation can be seen as a weighted combination of local linear transformations. These local transformations can perfectly fit the samples near to the center of each component, and cannot well deal with the samples far from the centers [12].

To alleviate this problem of conventional GMM and improve the conversion performance, the LCGMM algorithm [13], which considers the similarities of different components, is used for model training. In this method, a NN graph is established, and a KLD is employed for distance measurement. Given the spectral feature set $Z = \{z_n\}_{n=1}^{N}$, and a graph with M vertices, where each vertex refers to a vector, the edge of the graph $W$ is defined as following equations:

$$w_{mn} = \begin{cases} 1 & \text{if } z_m \in N_p(z_n) \text{ or } z_n \in N_p(z_m) \\ 0 & \text{otherwise} \end{cases}$$

(16)

where $N_p(z_m)$ are the $p$ NNs of $z_m$, and $N_p(z_n)$ are the $p$ NNs of $z_n$.

Let $P(i|z_m)$ and $P(i|z_n)$ denote the posterior probabilities of $z_m$ and $z_n$ belonging to the $i$-th component, respectively, as Eq.(11), the symmetric KLD is given as

$$d_{mn} = \frac{1}{2} \left( P(i|z_m)|P(i|z_n) + D(P(i|z_n)||P(i|z_m)) \right)$$

(17)

Incorporating Eqs.(16) and (17) as the smoothing term into GMM training, we will obtain the log likelihood as follows:

$$L = \sum_{n=1}^{N} \log \sum_{i=1}^{M} \alpha_{i} N(z_n|\mu_i, \Sigma_i) - \lambda \sum_{m,n=1}^{N} d_{mn} w_{mn}$$

(18)

where $\lambda$ is the regularization parameter, and the definitions of $\alpha_{i}$, $\mu_{i}$ and $\Sigma_{i}$ can be referred to Eqs.(4) and (5).

Just exactly like traditional GMM, an EM algorithm is also employed to maximize the regularized log likelihood $L$ [13], and all the model parameters ($\alpha_{i}$, $\mu_{i}$, $\Sigma_{i}$) will be computed.

### 4. Experiments

#### 4.1. Experimental setup

We carry out experiments on the CMU ARCTIC dataset to evaluate the performance of the proposed method. From this corpus, two English males (BDL and RMS) and two English females (SLT and CLB) are chosen, and 400 utterances of each speaker are used. To avoid overfitting, 5 independent tests are conducted repeatedly. In each test, 80 utterances for each speaker are randomly selected, among which, 50 utterances are chosen for training, while another 30 utterances, not including in the training dataset, are used for testing.

In our experiments, four types of inter-gender and intra-gender conversions including male-to-female (BDL to SLT), female-to-male (SLT to RMS), male-to-male (BDL to RMS) and female-to-female (SLT to CLB) are conducted, respectively, and four kinds of conversion methods, including the baseline INCA (INCA), the proposed SMA (SMA), the proposed GMM based SMA (GMM-SMA), and the proposed LCGMM based SMA (LCGMM-SMA) approaches are compared. The cepstral distortion and correlation coefficient are employed for objective evaluations. In addition, the mean opinion score (MOS) and ABX approaches are adopted for subjective measurements.

We employ the STRAIGHT analysis-synthesis model [14] to extract spectral features and F0s. The 24-order Mel-cepstral coefficients (MCCs) are used for spectral features, where the first item describes the energy. The 24-order MCC is transformed using the conversion methods mentioned above, while the first-order MCC and F0s are converted by transforming the means and covariances in logarithmic domain. The numbers of GMM and LCGMM are optimized as 256, and the numbers of NN and $\lambda$ for LCGMM are empirically set as 16 and 0.05, respectively.

#### 4.2. Objective evaluation

To evaluate the performance of converted speech objectively, we employ Mel-cepstral distortion (MCD) to describe the spectral distance between converted and target speech, which is defined as

$$\text{MCD} = \frac{10}{\ln 10} \left[ \frac{1}{2} \sum_{d=1}^{D} (m_{d} - m_{d}')^2 \right]$$

(19)

where $m_{d}$ and $m_{d}'$ are the $d$-th dimensional converted and target MCCs, respectively, and $D$ is the dimension of MCC. A lower MCD indicates a better conversion performance.

Figure 1 illustrates the results of MCDs using different conversion methods, and plots the MCDs as a function of the number of parallel training utterances. As is shown in the figure, for all the four methods, the MCDs become smaller with increase of the training data. We can also find that the SMA method always shows lower MCDs than the baseline INCA method, which demonstrates that the SMA alignment is more suitable for spectral feature alignment than INCA method. Meanwhile, we can easily observe that the GMM-SMA method

---

1http://www.festvox.org/cmu_arctic
outperforms the SMA or INCA methods greatly, which indicates that the multiple transformation is efficient to improve the conversion performance, and we can also see that the LCGMM-SMA approach shows the best performance, which means that the LCGMM algorithm is more appropriate for model training than traditional GMM method.

Another objective evaluation method namely Pearson product-moment correlation coefficient [15], which describes the correlations between converted and target MCCs, is also employed. The correlation of the \(d\)-th dimensional MCCs is calculated by

\[
 r_d = \frac{\sum_{n=1}^{N} (m_{nd}^c - \bar{m}_d^c)(m_{nd}^t - \bar{m}_d^t)}{\sqrt{\sum_{n=1}^{N} (m_{nd}^c - \bar{m}_d^c)^2} \sqrt{\sum_{n=1}^{N} (m_{nd}^t - \bar{m}_d^t)^2}} \tag{20}
\]

where \(m_{nd}^c\) and \(m_{nd}^t\) are the \(d\)-th dimensional MCCs of the \(n\)-th frame for converted and target speech, respectively, and \(\bar{m}_d^c\) and \(\bar{m}_d^t\) are the corresponding mean values of the \(d\)-dimensional MCCs. We use the average correlations of all dimensional MCCs for evaluation. The higher correlation, the better similarity.

The correlation results are presented in Figure 2. With the number of training utterances increasing, the correlations between converted and target speech increase accordingly. Meanwhile, the LCGMM-SMA always shows higher correlations than other three methods. Taking the spectral distortions and correlations into account, it can be stated that our proposed approaches are more efficient than the baseline INCA method.

4.3. Subjective evaluation

In order to further evaluate the performance of the proposed approach in terms of speech quality and identity, we conduct two kinds of methods, namely MOS and ABX, respectively. 50 randomly non-parallel utterances are chosen for building the voice conversion systems, and 10 subjects are employed for the listening tests.

To evaluate the quality of converted speech using different approaches, we adopt the MOS method, and the quality of the converted speech is graded in a 5-point range (5: excellent, 4: good, 3: fair, 2: poor and 1: bad).

We further perform an ABX test for similarity evaluations, where X is the converted speech, A and B are the source and target speech, respectively. In this test, the listeners are asked to choose whether X is closer to A or B, and a 5-point scale from '1' (totally different) to '5' (identical) is also introduced.

The average results of MOS and ABX evaluations are summarized in Figure 3, and the confidence interval is 95%. We can see that the SMA method shows higher scores than the traditional INCA method in terms of quality and similarity. Compared with SMA method, the GMM-SMA method shows significantly higher scores, and the LCGMMM-SMA method achieves the highest score among all the four approaches. These results are consistent with the results of objective MCDs and correlations.

5. Conclusion and future directions

In this paper, we have presented a SMA approach for voice conversion using non-parallel speech. Specifically, by aligning the source and target speaker models, the transformation function of spectral features is trained iteratively. In order to further improve the conversion performance, the SMA algorithm is extended to combine with GMM, and a LCGMM is used for model training. The experimental results clearly demonstrate that the proposed method significantly outperforms the baseline INCA algorithm.

As the proposed approach is designed for non-parallel speech, it is good for cross-linguistic voice conversion in theory. However, in our experiments, the voice conversion is carried out only for English utterances. So our further work will be done for cross-linguistic voice conversion.

6. Acknowledgements

The work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61375028, 61273266 and 61231002. The authors would like to thank all the volunteers who take part in the listening tests.
7. References


