NON-PARALLEL TRAINING FOR VOICE CONVERSION BASED ON ADAPTATION METHOD

Peng Song[1], Wenming Zheng[2], Li Zhao[1]

[1]School of Information Science and Engineering,
[2]Research Center for Learning Science,
Southeast University, Nanjing 210096, China
{pengsong, wenming_zheng, zhaoli}@seu.edu.cn

ABSTRACT

In this paper, we propose a simple and efficient non-parallel training scheme for voice conversion (VC). First, the speaker models are adapted from the background model using maximum a posteriori (MAP) technique. Then, by utilizing the parameters of adapted speaker models, the Gaussian normalization and mean transformation methods are proposed for VC, respectively. In addition, to improve the conversion performance of the proposed methods, a combination approach is further presented. Finally, objective and subjective experiments are carried out to evaluate the performance of the proposed scheme, the results demonstrate that our scheme can obtain comparable performance with the traditional GMM method based on parallel corpus.

***Index Terms*—** Voice conversion, non-parallel training, MAP, Gaussian normalization, mean transformation

1. INTRODUCTION

The goal of voice conversion (VC) is to convert the speech spoken by a source speaker to sound like that spoken by a target speaker. The applications include personalized text-to-speech synthesis, spoofing attacks to speaker recognition systems, and providing speaker individuality in ultra low bit-rate communication systems, etc.

Many approaches have been proposed for VC, such as mapping codebooks [1], Gaussian mixture model (GMM) [2, 3], artificial neural networks (ANN) [4], partial least squares regression (PLSR) [5], dynamic frequency warping (DFW) [6], and some combinations of them [7]. All these methods can obtain satisfactory results to some degree. However, they are conducted on a large parallel corpus, which is usually unavailable in practical situations.

In the past decade, several non-parallel training methods have been proposed. Ye et al. present a maximum likelihood (ML) approach to solve the non-parallel training problem [8]. In this approach, first, the hidden Markov model (HMM) is used to train the statistical model of target speaker, then the conversion function is estimated by maximizing the likelihood of the source spectral vectors with respect to the target model. Mouchtaris et al. employ a speaker adaptation technique to adapt the speech pairs of source and target speakers, and the conversion function is derived from a different pair of reference speakers [9]. Lee et al. propose a GMM based VC method using MAP adaptation, which needs the parallel utterances of source and reference speakers [10]. Erro et al. propose an iterative alignment method to improve the versatility of current VC systems, which allows non-parallel or even cross-lingual conditions [11]. All these methods can obtain comparable performance to the current parallel training method. However, they have some disadvantages, such as depending on large-scale corpus, or prior parallel reference conversion functions.

Different from the above-mentioned methods, in the paper, we propose a novel non-parallel training scheme for VC using small training corpus. The source and target speaker models are adapted from the background model, and two VC approaches are presented according to the parameters of adapted speaker models, one is the Gaussian normalization method, the other is the mean transformation method. In addition, the combination of these two methods is further proposed to improve the conversion performance.

The paper is organized as follows. Section 2 describes the baseline GMM method. Section 3 first gives the steps of the proposed scheme, and then presents the non-parallel training methods. The experimental results are reported and discussed in section 4. Finally, the main conclusions of the paper are provided in section 5.

2. CONVENTIONAL GMM BASED VOICE CONVERSION METHOD

GMM is the most popular approach for VC [2], and is chosen as the baseline of our method. Let $X = \{x_1, x_2, ..., x_T\}$ and $Y = \{y_1, y_2, ..., y_T\}$ represent the aligned parallel spectral feature sequences of source and target speakers, respectively,

The work is supported by NSFC (61231002, 60975017), and Ph.D. Programs Foundation of Ministry of Education of China (20120092110054).

and $Z = \{z_1, z_2, ..., z_T\}$ be the spectral feature pair series, where $z_t = [x_t^T, y_t^T]^T$ (the superscript ^{T} denotes transposition). Z is modeled by a GMM, which takes the form as follows:

$$p(Z) = \sum_{i=1}^{M} \alpha_i N(z, \mu_i, \Sigma_i) \tag{1}$$

Where α_i is the prior probability of Z, M is the number of mixture components, $N(z, \mu_i, \Sigma_i)$ denotes the Gaussian distribution of the i-th component, and μ_i and Σ_i are the mean vector and covariance matrix, respectively. The conversion function between source feature x and target feature y is given as follows:

$$F(x) = E(y|x) = \sum_{i=1}^{M} p(i|x)\Big(\mu_i^y + \frac{\Sigma_i^{xy}}{\Sigma_i^{xx}}(x - \mu_i^x)\Big) \tag{2}$$

Where $\mu_\mathbf{i} = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$, $\mathbf{\Sigma_i} = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}$, and $p(i|x)$ is the posterior probability of x belonging to the i-th component, which is given by

$$p(i|x) = \frac{\alpha_i N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{M} \alpha_j N(x, \mu_j^x, \Sigma_j^{xx})} \tag{3}$$

3. PROPOSED SCHEME

The GMM method is efficient and robust for VC. However, it is performed on a parallel corpus, which is often not feasible in practice. In order to address this problem, all kinds of approaches have been proposed in [8–11]. However, these methods still have some limitations, they need large-scale corpus or pre-defined parallel reference conversion functions. In this section, we propose a novel VC scheme based on a small non-parallel corpus.

The training process of the proposed scheme can be divided into the following steps:

1) First, like the universal background model (UBM) in speaker recognition systems [12], a background model is learned from the utterances of reference speakers.
2) Then, the models of source and target speakers are adapted from the background model by using the training utterances, respectively.
3) Finally, the conversion functions are estimated by utilizing the means and variances of speaker models.

3.1. Model adaptation

The MAP approach [13] is the popular adaptation strategy for GMM-UBM speaker recognition system, and is chosen for model adaptation of the proposed scheme. As small adaptation data cannot exactly describe each parameter of Gaussian components, only means and variances are considered in this paper. Let the observed spectral feature sequences $o = \{o_1, o_2, ..., o_T\}$, and ω_i, μ_B^i, and σ_B^i be the weight, mean and variance of the i-th component of the background model, respectively. The updated formulas of mean and variance are written as

$$\hat{\mu}_B^i = \gamma_i E_i(o) + (1 - \gamma_i)\mu_B^i \tag{4}$$

$$\hat{\sigma}_B^{i\,2} = \gamma_i E_i(o^2) + (1 - \gamma_i)(\mu_B^{i\,2} + \sigma_B^{i\,2}) - \hat{\mu}_B^{i\,2} \tag{5}$$

Where $E_i(o)$ and $E_i(o^2)$ are the statistics of mean and variance of the i-th component, respectively, and γ_i is the adaptation factor [12], and is given by

$$\gamma_i = \frac{n_i}{n_i + \rho} \tag{6}$$

Where n_i is the statistic of weight, and ρ is the coefficient describing the correlations between adapted and background models, and is optimized as 16 for our experiments. After adaptation, the parameters of source and target speaker models, $\{\omega_i, \mu_x^i, \sigma_x^i\}$ and $\{\omega_i, \mu_y^i, \sigma_y^i\}$ will be computed, respectively.

3.2. Gaussian normalization

The Gaussian normalization method is proposed for non-parallel spectral transformation. Different from the above-mentioned parallel or non-parallel VC methods, it can efficiently simplify the VC process, which can even avoid the training process of conversion function. The flowchart of this approach is shown in Fig.1.

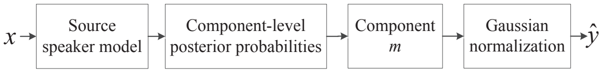


Fig. 1. Flowchart of Gaussian normalization method.

In the conversion phase, given a frame of spectral feature of source speaker x_t and the adapted source speaker model, the posterior probabilities of x_t belonging to the Gaussian components are computed. In [5], the experimental results have demonstrated that single Gaussian component often dominates each frame, so the component with maximum posterior probability is chosen, which is given by

$$m = \underset{i}{arg} \max p(i|x_t) \qquad i = 1, 2, ..., M \tag{7}$$

Where $p(i|x_t)$ takes the same form as Eq.(3). To the m-th component, we assume that the spectral features of source and target speakers belong to a Gaussian distribution with specific mean and variance, and will obtain the following equation:

$$\frac{x - \mu_x^m}{\sigma_x^m} = \frac{\hat{y} - \mu_y^m}{\sigma_y^m} \tag{8}$$

The converted spectral features $\hat{y}$ will be computed, and the conversion function can be defined as

$$F(x) = \frac{\sigma_y^m}{\sigma_x^m} x + \mu_y^m - \frac{\sigma_y^m}{\sigma_x^m} \mu_x^m \tag{9}$$

3.3. Mean transformation

The mean transformation approach is also proposed for non-parallel training. In this method, the mapping abilities of mean vectors of GMMs from source speaker to target speaker are investigated. Given the mean spectral feature sequences of source and target speaker models, $\mu_x = \{\mu_x^1, \mu_x^2, \ldots, \mu_x^M\}$ and $\mu_y = \{\mu_y^1, \mu_y^2, \ldots, \mu_y^M\}$, respectively, the mean transformation function between μ_x and μ_y is written as follows:

$$F(\mu_x) = A\mu_x + b \tag{10}$$

Where A and b are the transformation parameters. Assume $\bar{\mu}_x = \frac{1}{M}\sum_{i=1}^{M}\mu_x^i$ and $\bar{\mu}_y = \frac{1}{M}\sum_{i=1}^{M}\mu_y^i$, by employing the least squared algorithm, the unknown transformation parameters A and b will be computed, and given by

$$A = \hat{\mu}_y\hat{\mu}_x^T(\hat{\mu}_x\hat{\mu}_x^T)^{-1}, \qquad b = \bar{\mu}_y - A\bar{\mu}_x \tag{11}$$

Where $\hat{\mu}_x = \mu_x - \bar{\mu}_x$ and $\hat{\mu}_y = \mu_y - \bar{\mu}_y$.

It can be easily found that different from the alignment procedures in parallel VC, there exists a one-to-one correspondence between the mean vectors of source and target speaker models, it can avoid the forced alignment errors by means of dynamic time warping (DTW) algorithm in traditional GMM methods [2, 3]. The mean transformation function can roughly describe the mapping relationships between the spectral features of source and target speakers, and is directly adopted for spectral transformation. The conversion function is given as below:

$$F(x) = Ax + b \tag{12}$$

3.4. Combination algorithm

The proposed two approaches can convert source speech to target speech to a certain degree. On one hand, the Gaussian normalization method can be regarded as local linear regressions in discrete spaces. On the other hand, the mean transformation method is a global round prediction method. In order to further improve the conversion performance, an algorithm is proposed to combine the local and global regression methods, and the combination function is defined as

$$F(x) = \theta F_g(x) + (1-\theta)F_m(x) \tag{13}$$

Where $F_g(x)$ and $F_m(x)$ are the conversion functions of Gaussian normalization and mean transformation approaches, respectively, and θ is the coefficient and satisfies $0 \leq \theta \leq 1$. The selection of θ is key to the performance of combination algorithm. In this paper, it is chosen by using the iteration algorithm with step size 0.01. In this paper, when the number of adaptation utterances is lower than 8, θ is optimized as 0.19 for the experiments, or it will be set as 0.73.

3.5. Model optimization using K-L divergence

It is worth noting that the speaker models are trained by limited adaptation utterances, which cannot ensure that the parameters of each mixture component are updated, and will affect the precision of conversion function. In this paper, an algorithm by employing Kullback-Leibler (K-L) divergence is proposed to address this problem [14]. The divergence between two distributions $f(c)$ and $g(c)$ is defined as

$$D\big(f(c)||g(c)\big) = \sum_c f(c)\log\frac{f(c)}{g(c)} \tag{14}$$

Note that Eq.(14) is not symmetric, we can adopt a symmetric divergence to measure the distance between $f(c)$ and $g(c)$, which is given as follows:

$$D_{fg} = \frac{1}{2}\Big(D\big(f(c)||g(c)\big) + D\big(g(c)||f(c)\big)\Big) \tag{15}$$

For simplicity, only means are considered. The proposed algorithm consists of the following steps:

1) First, the similarities between different components of each speaker model are computed by symmetric divergence.

2) Then, if μ_x^i equals μ_B^i or μ_y^i equals μ_B^i, update μ_x^i with μ_x^j or μ_y^i with μ_y^l, respectively, where j and l are the most similar components of the i-th components for source and target speakers, respectively.

3) Repeat step **2**), until it satisfies the conditions: $\mu_x^i \neq \mu_B^i$ and $\mu_y^i \neq \mu_B^i$.

4. EXPERIMENTS

We carry out experiments on the CMU ARCTIC corpus. The speakers BDL and CLB (each has 500 utterances) are chosen for the background model training, while speakers RMS and SLT are chosen for voice conversion. Two cases: the transformation of RMS-to-SLT (male to female, M-F) and transformation of SLT-to-RMS (female to male, F-M) are used for evaluations. Four kinds of VC methods are compared, they are the baseline GMM method with parallel training data (GMM), and the proposed three kinds of non-parallel training approaches, including the Gaussian normalization method (GN), the mean transformation method (MT), and the combination method (GNMT).

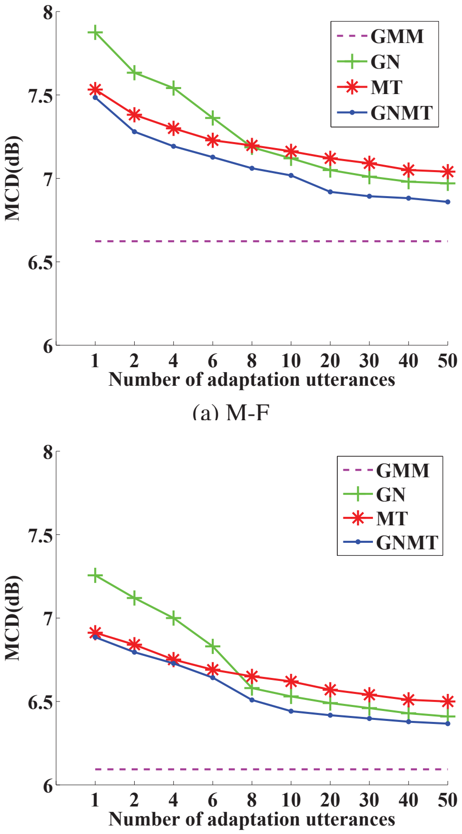
Each subset (RMS or SLT) consists of 120 utterances, from which, the same 50 utterances are prepared for the baseline parallel training, and different 50 utterances are prepared for the proposed approaches, while another 20 utterances are used for testing. The 24-order Mel-cepstral coefficients (MCEPs) are extracted to represent the spectral features, and the mixture number of background model is set as 256, while the mixture number of baseline GMM method is optimized as 16. The F0s are converted using logarithm Gaussian normalization approach [4], and 6 subjects participate in the listening tests.

4.1. Objective evaluation

Mel-cepstral distance (MCD) is a known error measurement between converted speech and target speech, and has been widely employed for objective evaluations of VC [4]. It is defined as follows:

$$MCD = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}(m_d^t - m_d^c)^2} \qquad (16)$$

Where m_d^t and m_d^c are the d-th coefficient of MCEPs of target and converted speech, respectively.



(a) M-F

(b) F-M

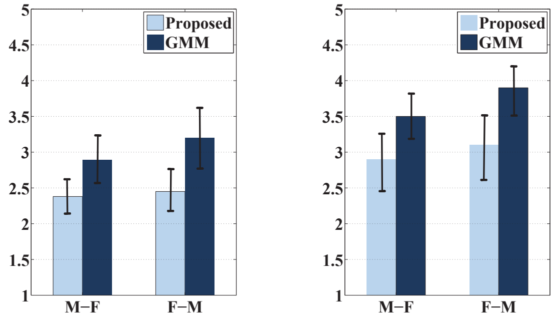
Fig. 2. Average MCDs using different number of adaptation utterances. The dashed line corresponds to the GMM based VC using parallel corpus.

Fig.2 gives the results of MCDs of M-F and F-M. It can be easily found that as expected, with increase of the number of adaptation utterances, the MCDs of the proposed approaches show the same trend directions: they all become nearer to those of the baseline method. This is due to the fact that with increase of adaptation data, the adaptation models become closer to the true speaker models. When the number of adaptation utterances is smaller than about 8, the MT method shows superior results to the GN method. Meanwhile, when the adaptation corpus becomes larger, the GN method performs better. It can be also found that the MCDs of combination algorithm show always lower values than those of GN or MT approaches, which indicates that the combination strategy can be an efficient way to boost the GN and MT methods.

4.2. Subjective listening evaluation

In subjective tests, 5 utterances of each speaker are used for model adaptation. The mean opinion score (MOS) and similarity tests are conducted, respectively. In MOS test, the listeners are asked to rate the perceptual quality of the converted speech in a 5-point score: (1:bad, 2:poor, 3:fair, 4:good, 5:excellent). While in similarity test, the listeners evaluate the similarities between the converted speech and target speech, also in a 5-point range from 1 ”different” to 5 ”identical”.

The proposed GNMT and baseline GMM approaches are compared for evaluations. In Fig.3(a) and Fig.3(b), the overall results of quality and similarity tests are presented, respectively. We can observe that the proposed method can yield comparable results to the baseline GMM method, which confirms the results of objective evaluations to some extent.



(a) Quality

(b) Similarity

Fig. 3. Subjective test results (95% confidence interval).

5. CONCLUSIONS

This paper proposes a novel non-parallel VC scheme using small training corpus. The MAP adaptation technique is adopted to train the models of source and target speakers, and the Gaussian normalization, mean transformation, and combination approaches using parameters of adaptation models are presented, respectively. Experimental results on CMU ARCTIC corpus indicate that compared to the traditional GMM method based on parallel corpus, our scheme has a comparable performance in terms of cepstral distortion, and also obtains satisfactory speech perceptibility and similarity.

6. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*. IEEE, april 1988, vol. 1, pp. 655 –658.

[2] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*. IEEE, may 1998, vol. 1, pp. 285 –288.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131 –142, march 1998.

[4] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 954 –964, july 2010.

[5] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 912 –921, july 2010.

[6] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1313 –1323, may 2012.

[7] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum," in *Proc. ICASSP*. IEEE, 2001, vol. 2, pp. 841 –844.

[8] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1301 –1312, july 2006.

[9] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 952 – 963, may 2006.

[10] C.H. Lee and C.H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Proc. Interspeech*. ISCA, 2006, pp. 2254 –2257.

[11] D. Erro, A. Moreno, and A. Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 944 –953, july 2010.

[12] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1, pp. 19 – 41, 2000.

[13] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291 –298, april 1994.

[14] T.A. Myrvoll and F.K. Soong, "On divergence based clustering of normal distributions and its application to hmm adaptation.," in *Proc. Interspeech*. ISCA, 2003, pp. 1517 –1520.