MAP-BASED ADAPTATION FOR SPEECH CONVERSION USING ADAPTATION DATA SELECTION AND NON-PARALLEL TRAINING

Chung-Han Lee and Chung-Hsien Wu

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan
Phone: +886-6-2089349, Fax: +886-6-2747076,
Email: {chlee,chwu} @csie.ncku.edu.tw

ABSTRACT

This study presents an approach to GMM-based speech conversion using maximum a posteriori probability (MAP) adaptation. First, a conversion function is trained using a parallel corpus containing the same utterances spoken by both the source and the reference speakers. Then a non-parallel corpus from a new target speaker is used for the adaptation of the conversion function which models the voice conversion between the source speaker and the new target speaker. The consistency among the adaptation data is estimated to select suitable data from the non-parallel corpus for MAP-based adaptation of the GMMs. In speech conversion evaluation, experimental results show that MAP adaptation using a small non-parallel corpus can reduce the conversion error and improve the speech quality for speaker identification compared to the method without adaptation. Objective and subjective tests also confirm the promising performance of the proposed approach.

Index Terms: Voice conversion, non-parallel training, data selection

1. INTRODUCTION

In recent years, voice conversion has been used in Text-to-speech (TTS) systems [1] to convert the synthesized speech to a certain target speech. These systems generate new voices of the target speaker without the need of collecting a large speech corpus of the target speaker. There are many approaches proposed for voice conversion, such as codebook mapping [2], artificial neural networks, Gaussian Mixture Models (GMM) [1][3][4], Hidden Markov Models (HMM) [5], and the combination of these methods. Most current voice conversion systems focused on spectral conversion and applied simple adjustment of the prosodic features.

In common, these approaches focused on short-term spectral property conversion of the speech signals. The parameters of these conversion functions are derived by minimizing the error measure (mean square error, cost function, etc.) during training. However, it is absolutely necessary that the training corpus should contain the parallel utterances (sentences) from both the source and the target speakers. In fact, it is obvious that collecting such a parallel corpus is difficult and even impossible. For example, if the source or the target speaker is a personage and the parallel corpus is not directly available. It is even impossible to collect such a corpus containing the desired sentences. The approach in [6] declared that it does not require a parallel corpus for training by considering the phones of the utterances from both the source and target speakers. However, speech recognition of high accuracy is required to correctly recognize the phones spoken by the source and target speakers.

In this study, a parallel corpus containing the speech utterances from the source and the reference speakers is collected. A general conversion function can be trained from the parallel corpus. For a new target speaker, a small non-parallel corpus is collected for conversion function adaptation. An adaptation data selection method is applied to remove the outliers in the adaptation corpus which present inconsistent property with other adaptation data. The adaptation procedure can be described as follows. We assume that there are three different speakers X, Y and Z; X is the source speaker, Y is the reference speaker, and Z is the target speaker. The parallel corpus with sufficient number of speech utterances is available for speakers X and Y. The purpose of the conversion is to convert the speech of speaker X to speaker Z. However, the corpus for speaker Z is difficult to collect and insufficient. The parallel corpus for speakers X and Y is used to train the source-to-reference conversion function. The source-to-reference conversion function between speakers X and Y is then adapted to that between speakers X and Z. The MAP adaptation method is proposed to adapt the conversion parameters in the source-to-reference conversion function.

2. CONVENTIONAL GMM-BASED VOICE CONVERSION

Recently, the Joint Density (JD) GMM method [1] is the mainstream method for voice conversion. The training parallel corpus is created from the source and the target speech utterances, and the parameters which model the short-term spectral envelope (line spectral frequencies -LSF) are extracted. The JD method is described below as the baseline of the proposed approach. This procedure results in two vector sequences; the source speech is represented by an n-frame time-series $X=[x_1,x_2, \cdots, x_n]$ and the target is represented by an m-frame time series $Y=[y_1,y_2, \cdots, y_m]$, where x_i and y_j are the k dimensional feature vectors for the ith frame and the jth frame respectively, i.e. $x_i=[x_1,x_2, \cdots, x_k]^T$ and $y_j=[y_1,y_2, \cdots, y_k]^T$. Assuming that x and y are joint Gaussian distribution for each class C_i, we can find the optimal choice for the function to minimize the mean square error using the EM algorithm [7]. The conversion function that converts the source feature x to the target feature y is given by the following equation

$$F(x) = E(y \mid x) = \sum_{i=1}^{M} P(C_i \mid x_t)[\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx^{-1}} (x_t - \mu_i^x)] \quad (1)$$

where $E(.)$ denotes the expectation , and

$$E[y \mid x = x_t] = \mu^y + \Sigma^{yx} \Sigma^{xx^{-1}} (x_t - \mu^x) \quad (2)$$

The dynamic time warping (DTW) algorithm is used to align the source features to their counterparts in the target series to create a new sequence $Z=[z_1,z_2,...,z_q]$ where $z=[x^T y^T]^T$. In order to estimate the GMM of z, it is required to correctly align vectors x_i and y_j during training. Then the parameters defined above can be calculated by estimating the GMM parameters of z, where

$$\Sigma_i^{zz} = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}, \mu_i^z = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \quad (3)$$

and

$$p(C_i \mid x) = \frac{p(C_i)N(x;\mu_i,\Sigma_i)}{\sum_{j=1}^{M} p(C_i)N(x;\mu_i,\Sigma_i)} \quad (4)$$

is the probability of x belonging to the *ith* component.

3. ADAPTATION DATA SELECTION AND MAP-BASED ADAPTATION

3.1 Adaptation Data Selection

In conversion function adaptation, some inconsistent sentence pairs are unsuitable for adaptation; it will disarrange the original conversion function. In the proposed adaptation data selection approach, the conversion function for each adaptation sentence pair is modeled by a GMM. A GMM, representing an adaptation sentence pair, with its mean and variance greatly different from that of other GMMs is regarded as an outlier and is unsuitable to serve as the adaptation data. The KL-divergence [10] is adopted as the distance measure between two GMMs f and g defined as follows:

$$KL(f \| g) = \int f \log \frac{f}{g} \approx \frac{1}{n}\sum_{t=1}^{n} \log \frac{f(x_t)}{g(x_t)} \quad (5)$$

such that $x_1,x_2, \cdots, x_n$ are sampled from $f(x)$. According to the KL-divergence is not symmetric, $KL(f\|g) \neq KL(g\|f)$, and does not satisfy the triangle inequality. The symmetric form is adopted as the distance between two GMMs f and g:

$$Dis(g,f) = \frac{1}{2}\left(KL(f \| g) + KL(g \| f)\right) \quad (6)$$

which is symmetric and nonnegative.

The distance between any two GMMs is estimated and used to generate the distance matrix containing the distance of every two GMMs. Accordingly, given the distance matrix, each GMM can be transformed into a corresponding vector in a high-dimensional space using MDS (Multi-Dimensional Scaling) which is a set of mathematical techniques used to uncover the "geometric structure" of datasets[11]. For example, given a set of objects with proximity values amongst themselves, MDS is able to create a 2D map of these objects.

Referring to Fig.1, we can transform GMMs into the corresponding vectors when the distance of any two GMMs is given. The mean of these vectors can be regarded as the centroid of these GMMs. The GMMs corresponding to the MDS-converted vectors of which the Euclidean distances to the centroid are greater than a threshold will be removed, where the threshold is empirically defined. Thereafter, the remaining sentence pairs are used as the adaptation data for further process.

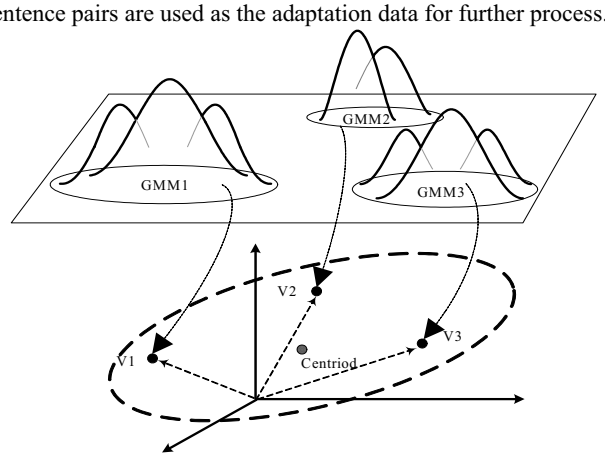

Figure 1: Block diagram outlining the concept of MDS converting the GMMs to the vectors in a high-dimensional space.

3.2 MAP-Based Adaptation

The GMM-based conversion method converts the acoustic features from a source speaker to a target speaker by minimizing the mean squared error (MMSE). However, some research [8] declared that the converted features are overly smoothed and this makes the reconstructed speech unclear. They thought that the correlation between the features of two speakers is not linear. With this assumption, the conversion function in Eq. (1) is modified to Eq. (7). In Eq. (7), the weighted distance between the mean vectors of the corresponding mixtures in the source GMM and the reference GMM represents the shift from the source features to the reference features.

$$\mathrm{y}=\mathrm{x}+\sum_{i=1}^{I} P_i(i \mid x)(\mu_i^y - \mu_i^x) \quad (7)$$

where μ_i^x and μ_i^y are the mean vectors of the *ith* component of the source and reference GMMs, and $P_i(i|x)$ is the probability of x belonging to the *ith* component.

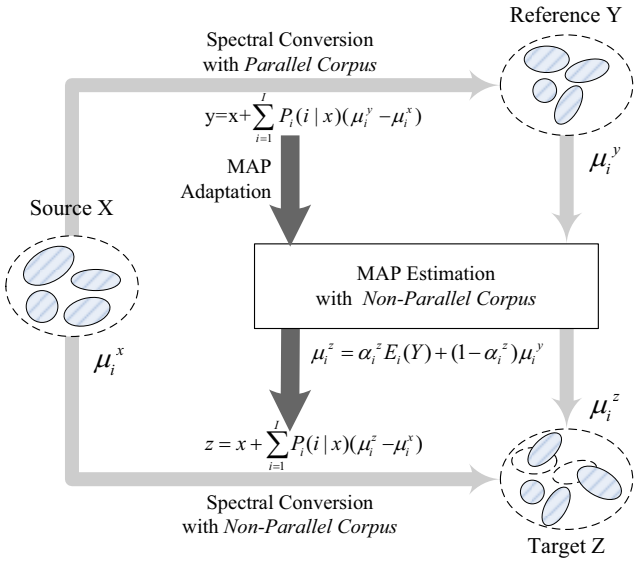

Figure 2: Block diagram outlining spectral conversion for a parallel and non-parallel corpus.

The conventional GMM-based method assumes that a parallel speech corpus is available for the source and the target speakers. In this approach, the source and target speakers do not necessarily utter the same sentences. We apply the maximum a posteriori probability (MAP) adaptation method [9] to the GMM conversion function. This approach only adapts the mean of the GMM conversion function because the importance of the mean is much higher than the covariance of the GMM. The MAP-based adaptation method is illustrated in Eq. (8).

$$\mu_i^z = \alpha_i^z E_i(Y) + (1-\alpha_i^z)\mu_i^y \tag{8}$$

where

$$\alpha_i^z = \frac{n_i}{n_i + r} \quad ,$$
$$n_i = \sum_{t=1}^{T} P(i \mid y_t), \tag{9}$$
$$E_i(Y) = \frac{1}{n_i}\sum_{t=1}^{T} P(i \mid y_t) y_t$$

and T is the total frame number of the selected adaptation data, r is a fixed factor [9] and x_t and y_t are the feature pairs aligned by the DTW algorithm.

In the adaptation process, first, we assume that a parallel speech corpus is available for the source and the reference speakers, referring to Fig.2. The spectral vectors which correspond to the source speaker are considered as the realization of the random vector x, while y corresponds to the reference speaker in the parallel corpus. The spectral vectors which correspond to the target speaker are considered as the realization of the random vector z, and a non-parallel corpus is available for the reference and the target speakers. In order to derive a conversion function for the non-parallel corpus, we relate the random vectors y and z by assuming that the target random vector z is related to the reference random vector y, and MAP adaptation is achieved using Eq. (10)

$$z = y + \sum_{i=1}^{I} P_i(i \mid y)(\mu_i^z - \mu_i^y) \tag{10}$$

These equations correspond to the GMM-based estimation that relates y with z and z with y in the block diagram of Fig.2. Then, the μ_i^z can be estimated from the non-parallel corpus using the MAP adaptation method.

$$\mu_i^z = \frac{\sum_{t=1}^{T} P(i \mid y_t) y_t}{r + \sum_{t=1}^{T} P(i \mid y_t)} + \frac{r\mu_i^y}{r + \sum_{t=1}^{T} P(i \mid y_t)} \tag{11}$$

Finally, the conversion function for the source speaker to the target speaker can be obtained by Eq. (12)

$$z = x + \sum_{i=1}^{I} P_i(i \mid x)(\mu_i^z - \mu_i^x) \tag{12}$$

According to the above equations, all the parameters of the conversion function between the source speaker x and the target speaker z are obtained using the non-parallel corpus and the MAP adaptation.

4. EXPERIMENTS AND DISCUSSION

There are two experiments conducted to evaluate the performance of the proposed method; one is the objective evaluation and the other is the subjective evaluation. The spectral vectors used in this study are the LSFs (22nd order). In the first experiment, we used a log-spectral distortion measure to provide an objective performance measure for the baseline system, and the performance for the parallel and the non-parallel corpora were compared. In the second experiment, speaker identification was investigated. In both experiments, the source-to-reference conversion function from the source to the reference speakers were trained using 20, 40 and 100 parallel utterances from the source and the reference speakers. The source-to-target conversion function is adapted from the source-to-reference conversion function using 5, 10, 15 and 20 utterances from the target speaker. The number of Gaussian mixtures was set to 16. All speakers are male.

The log-spectral distortion measure is defined as

$$d(S_1, S_2) = \sum_{k=1}^{120} (\log a_k^1 - \log a_k^2)^2 \tag{13}$$

where $\{a_k\}$ are the spectral amplitudes resampled from the spectral envelope S. Thus the overall conversion distortion can be estimated as the log distortion ratio of the converted-to-target conversion distortion and the source-to-target conversion distortion, which is defined as,

$$D = 10\log_{10}\frac{\sum_{t=1}^{N} d(S_T(t), S_C(t))}{\sum_{t1}^{N} d(S_T(t), S_S(t))} \tag{14}$$

where $S_T(t)$, $S_S(t)$ and $S_C(t)$ are the target spectral envelope, source spectral envelope and the converted spectral envelope at time t, respectively. N is the total number of vectors.

Table 1: Log-spectral distortion for three pairs of parameters derived from a parallel corpus, when applied to the target speaker with a non-parallel corpus.

Number of Non-parallel Sentences	Log-spectral Distortion			
	( ): Number of Parallel Sentences			
	None	Adapt(100)	Adapt(40)	Adapt(20)
5	-2.312	-2.531	-2.468	-2.409
10	-2.397	-2.596	-2.542	-2.491
15	-2.436	-2.712	-2.645	-2.573
20	-2.511	-2.943	-2.824	-2.641

In Table 1, the log-spectral distortion is given for four different numbers (5, 10, 15 and 20) of non-parallel utterances from the source and the target speakers in four different adaptation cases. The column denoted as "None" in this table corresponds to no adaptation, i.e. there is no non-parallel corpus available for the source and the target speakers. The parameters in the conversion function trained directly by the non-parallel corpus with the phones in the corpus having been segmented and aligned between the source and the target. The column denoted as "Adapt" in this table corresponds to the adaptation of the conversion function using Eq. (11). This table shows the performance of the proposed approach for four different choices of the training corpus. While the size of the parallel corpus increases, we can observe that the log-spectral distortion between the source and the target speakers will decrease. The performance improvement for increasing the number of non-parallel corpus only is less than that for increasing both the numbers of parallel and the non-parallel corpora. It is apparent from Table 1 that the adaptation methods result in error decrease compared to simply applying the conversion parameters of a given non-parallel corpus for the source and the target speakers.

In the second experiment, the ABX test is used to identify if the converted speech is close to the source speaker or the target speaker. In this experiment, three utterances were played sequentially to the subjects, who were expected to decide which the previous two utterances (A and B) was closer to the third one (X). In our result, X was the converted speech with MAP adaptation or the converted speech without MAP adaptation. There are two cases. The first is that there are 100 pairs of the parallel utterances which are available for the source and the reference speakers and 20 non-parallel utterances which are available for the target speakers. The conversion parameters were trained by the 100 pairs of parallel utterances to the target speaker by MAP adaptation method. The second case is that there are only 20 pairs of non-parallel utterances for the source and the target speakers, and the conversion parameters were trained directly without MAP adaptation. From Figure 3, it is obvious that voice quality of the converted speech generated by the MAP adaptation is better than those generated by the conventional GMM-based approach without MAP adaptation.

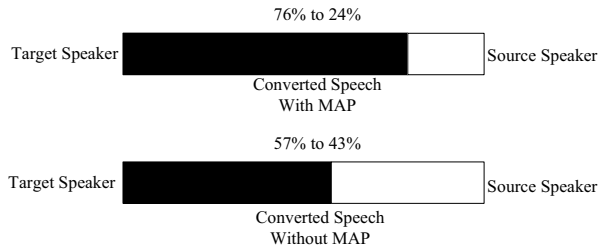

Figure 3: Results for the speaker identification evaluation

The results of the two experiments are given in Table 1 and Figure 3 respectively. Since the proposed method just adapts the mean vectors of GMM, the log-spectral distortion improvement is insignificant. We can see that significant error reduction was obtained and an improvement in speech quality for speaker identification was achieved.

5. CONCLUSIONS

In this study, a speech conversion approach that only requires a small non-parallel speech corpus for MAP-based adaptation has been presented. This approach uses a parallel corpus for the source and the reference speakers to derive a conversion function. A non-parallel corpus from the target speaker is then used for conversion function adaptation using the MAP adaptation method. An adaptation data selection method is proposed to select suitable data from the non-parallel corpus for MAP-based adaptation of the GMMs. Experiments show that the proposed method achieves a satisfactory result and decreases the conversion error even though the corpus of the target speakers is non-parallel to the source speaker corpus or insufficient.

6. REFERENCES

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (Seattle, WA),pp. 285 – 289, May 1998.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (New York, NY), pp. 655 – 658, April 1988.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131 – 142, March 1998.

[4] M. Mashimo, H. Toda, and K. Shikano, Campbell, N., "Evaluation of cross-language voice conversion based on GMM and STRAIGHT," in *Proc. of Eurospeech 2001,* pp. 361-364 , 2001.

[5] H. Duxans, A. Bonafonte, A. Kain, J. van Santen. "Including Dynamic and Phonetic Information in Voice Conversion Systems". *Proceedings of ICSLP*, Oct 2004.

[6] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 720–723, April 2003.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[8] Yining Chen, Min Chu, Eric Chang, Jia Liu and Runsheng Liu, "Voice Conversion with Smoothed GMM and MAP Adaptation," *proc. of the Eurospeech03*, pp. 2413–2416, 2003.

[9] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol.10, pp. 19-41, 2000.

[10] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transform and its application to speaker recognition", to appear in Proc. of Interspeech 2005.

[11] T. Cox and M. Cox, "Multidimensional Scaling", Chapman-Hall, 1994.