A hybrid GMM and codebook mapping method for spectral conversion

[†]Yongguo Kang, [‡]Zhiwei Shuang, [†]Jianhua Tao, [‡]Wei Zhang, and [†]Bo Xu

[†]Institute of Automation, Chinese Academy of Science
[‡]China Research Lab, IBM
ygkang@nlpr.ia.ac.cn; shuangzw@cn.ibm.com; jhtao@nlpr.ia.ac.cn;
zhangwei@cn.ibm.com; xubo@hitic.ia.ac.cn

Abstract. This paper proposes a new mapping method combining GMM and codebook mapping methods to transform spectral envelope for voice conversion system. After analyzing overly smoothing problem of GMM mapping method in detail, we propose to convert the basic spectral envelope by GMM method and convert envelope-subtracted spectral details by GMM and phone-tied codebook mapping method. Objective evaluations based on performance indices show that the performance of proposed mapping method averagely improves 27.2017% than GMM mapping method, and listening tests prove that the proposed method can effectively reduce over smoothing problem of GMM method while it can avoid the discontinuity problem of codebook mapping method.

1 Introduction

Starting from speech signal uttered by a speaker, voice conversion aims at transforming the characteristics of the speech signal in such a way that a human naturally perceives the target speaker's own characteristics in the transformed speech[1]. An important task in voice conversion system is to map speech features which represent the speaker individuality between source and target speech. The underlying meaning of mapping is to find the relation between two sets of multi-dimension vectors.

There are a lot of mapping methods such as codebook mapping[2] [3], Linear Multivariate Regression (LMR)[8], , Neural Networks[9][10], Gaussian Mixture Model (GMM)[4][5] and Hidden Markov Model (HMM)[11]. Among these mapping methods, codebook mapping and GMM methods are two representative and popular mapping algorithms. Motivated by the fact that the disadvantages of two methods respectively are overly smoothing and discontinuity, a method combining GMM and codebook mapping is proposed. This method tries to grasp the basic spectral envelope using GMM and retain converted spectrum details using offset codebook mapping method. By this means, the problems of smoothing and discontinuity can be counteracted.

This paper is organized as follows. Section 2 describes the conventional GMM mapping methods and then investigates the reason of GMM's overly smoothing problem. Section 3 describes the proposed hybrid method combining GMM and codebook mapping to convert source spectrum. Evaluation and discussions are given in section 4 while the conclusions are drawn in section 5.

2 ANALYSIS OF GMM MAPPING ALGORITHMS

2.1 GAUSSIAN MIXTURE MODEL

The GMM assumes the probability distribution of the observed parameters takes the following form:

$$p(x) = \sum_{i=1}^{m} \alpha_i N(x; \mu_i, \Sigma_i), \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0 \tag{1}$$

where $N(x; \mu_i, \Sigma_i)$ denotes the m-dimensional normal distribution with mean vector μ and covariance matrix Σ, α_i is normalized positive scalar weight. The parameters (α, μ, Σ) can be estimated with the expectation-maximization algorithm.

The parameter of the conversion function is determined by the joint density of source and target features [5]. The combination of source and target vectors $z = [x^T y^T]^T$ is used to estimate GMM parameters (α, μ, Σ). The conversion function can be yielded using regression:

$$F(x) = \sum_{q=1}^{Q} p_q(x)[\mu_q^Y + \Sigma_q^{YX}(\Sigma_q^{XX})^{-1}(x - \mu_q^X)] \tag{2}$$

where $p_q(x)$ is the conditional probability of a GMM class q given x:

$$p_q(x) = \frac{\alpha_q N(x; \mu_q^X, \Sigma_q^X)}{\sum_{p=1}^{Q} \alpha_p N(x; \mu_p^X, \Sigma_p^X)} \tag{3}$$

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{XX} & \Sigma_q^{YX} \\ \\ \Sigma_q^{XY} & \Sigma_q^{YY} \end{bmatrix}; \mu_q = \begin{bmatrix} \mu_q^X \\ \\ \mu_q^Y \end{bmatrix}$$

and Q is the number of GMM components. The performance of GMM method has been proved that it can be as good as or better than other conversion function implementations. However, the conventional GMM based conversion tends to generate overly smoothed spectrum [13][7].

2.2 Analysis of overly smoothing problem

In the conventional GMM method, the transforming function includes two parts: the mean item and correlation item, as shown in Equation (4):

$$F(x) = \underbrace{\sum_{q=1}^{Q} p_q(x)\mu_q^Y}_{mean} + \underbrace{\sum_{q=1}^{Q} p_q(x) \underbrace{\Sigma_q^{YX}(\Sigma_q^{XX})^{-1}}_{covariance} \underbrace{(x - \mu_q^X)}_{offset}}_{correlation} \tag{4}$$

It can be concluded that the mean item grasps the basic shape of converted feature while the correlation item try to convert spectral details using the offset vector $((x - \mu_q^X))$. Toda [6] has pointed that the covariance matrices between a source feature and a target feature is critical to convert speech features continuously In GMM mapping method. However, the correlation is difficult to accurately estimate in particular using full covariance matrices. Chen [7] has point out the most items in the correlation matrix is nearly zero, thus the converted features are in fact close to the first item.

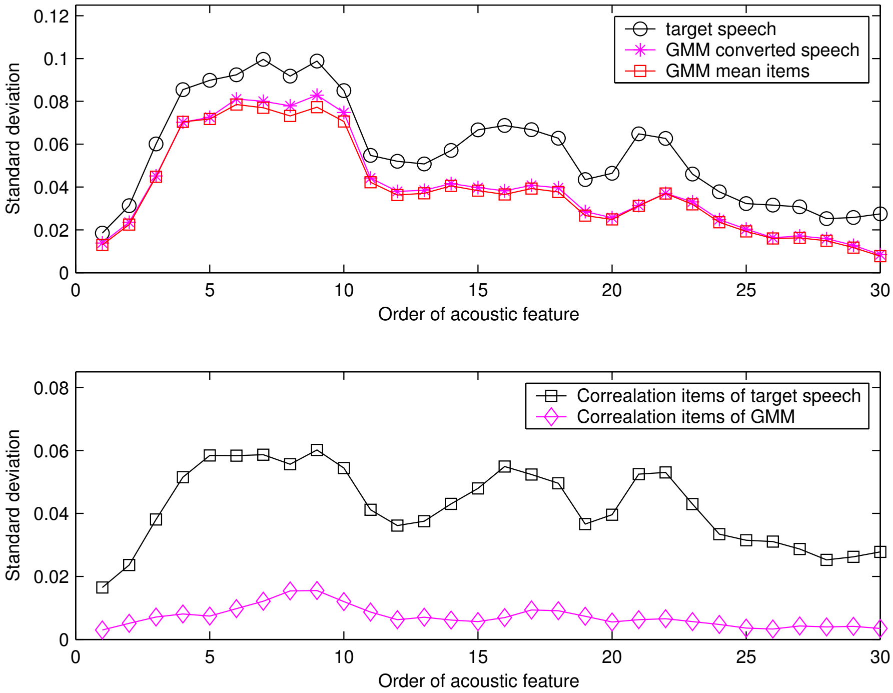

Fig. 1. The top are standard deviations of acoustic feature from target speech, GMM converted speech and hybrid method converted speech; The bottom are those from correlated items of target speech, GMM converted speech and hybrid method converted speech

In this study, standard deviation is employed to describe the smoothing level of acoustic feature. Figure 1 shows the standard deviations calculated from the following acoustic features: 1) target speech; 2) GMM converted features 3) GMM mean items; 4) correlation items of target speech, i.e. acoustic features from target speech subtracted GMM mean features; 5) GMM correlation items. It can be observed that standard deviations of GMM converted features are less than those of target speech, and the smaller standard deviations can indicate the overly smoothing problem. However, it is worth noting that the standard deviations of GMM correlation items are distinctly less than those of correlation items from target speech. Thus it can explain that in GMM methods, the basic envelope of ideally converted features is remained but the spectral details are lost.

3 A HYBRID MAPPING METHOD

Based on the analysis of previous section, the overly smmothing problem to be solved is how to reuse the offset vector $(x - \mu_q^X)$ and rebuild the missed spectral details. We will apply this phoneme-tied codebook mapping method to convert the offset vector in order to recover the spectral details. The procedures combining GMM and Mapping codebooks are described as:

- *Training procedure:*
  1. GMM training is first carried out.
  2. Generating offset codebook entries. For each joint vectors $z = [x^T y^T]^T$, the offset code words are defined as:

$$x^{\text{offset}} = \sum_{q=1}^{Q} p_q(x)(x - \mu_q^X) \tag{5}$$

$$y^{\text{offset}} = \sum_{q=1}^{Q} p_q(x)(y - \mu_q^Y) \tag{6}$$

- *Converting procedure:*
  1. Converting input source vector x only using corresponding target mean vectors:

$$\hat{y}^{\text{mean}} = \sum_{q=1}^{Q} p_q(x)\mu_q^Y \tag{7}$$

$$\hat{y}^{\text{corr}} = \sum_{q=1}^{Q} p_q(x)\Sigma_q^{YX}(\Sigma_q^{XX})^{-1}(x - \mu_q^X) \tag{8}$$

  2. Converting offset vector using codebook mapping method with phoneme-tied weighting :

$$\hat{y}^{\text{offset}} = F^{\text{offset}}(x^{\text{offset}}) \tag{9}$$

     where F^{offset} is the transforming function obtained by codebook mapping method, and x^{offset} is defined by equation (5);
  3. The final converted feature is:

$$\hat{y} = \hat{y}^{\text{mean}} + (1 - \lambda)\hat{y}^{\text{corr}} + \lambda\hat{y}^{\text{offset}} \tag{10}$$

The smooth converted spectrum is added spectral details using mapping offset codebook, and the new converted spectrum is shown in Fig.2. From Fig.2, it can be observed that the converted spectrum is much closer to the target spectrum than the smooth spectrum and the over smoothing problem is avoided.

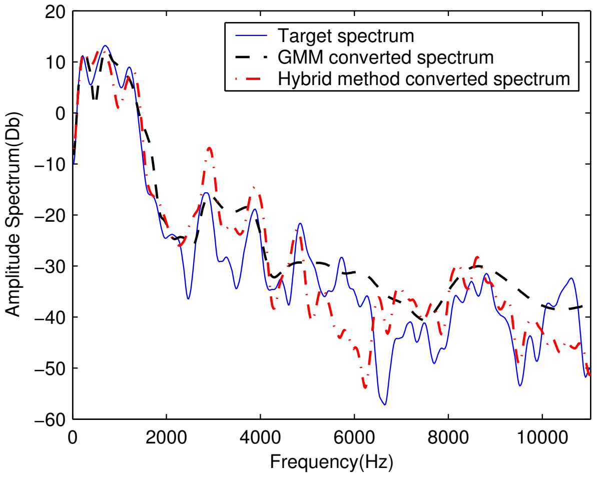

Fig. 2. The target spectrum, converted spectrum by GMM method and converted spectrum using hybrid mapping method

4 EXPERIMENT AND EVALUATION

4.1 EXPERIMENT

The corpus used for the experiments was recorded by two female speakers reading the same text. We use 22050 Hz's sampling rate and 16 bits per sample to store the speech data. The corpus has been segmented (manually supervised) into phonemes. There are 180 sentences which consist of 28037 vectors to train the transforming function and 12 sentences which consist of 2097 vectors to test voice conversion system.

Because STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectral envelope) algorithm can reproduce high quality speech from its coefficients and modify duration, F0 and spectral coefficients separately in large scale with little degradation to the quality, it is employed to analyze and synthesize speech signal. In this experiment we utilize the all-pole model to fit spectral envelope obtained by STRAIGHT, and employ linear spectral pair (LSP) parameter resulted from all-pole model as the mapping feature.

The pitch contour is converted using a linear transformation. The converted pitch $\hat{f}_t$ is obtained using:

$$\hat{f}_t = \mu_t + \frac{\sigma_t}{\sigma_s}(f_s - \mu_s) \tag{11}$$

where (μ_s, σ_s) and (μ_t, σ_t) are the mean and variance of source and target pitch contours respectively.

4.2 OBJECTIVE EVALUATION

The objective evaluation of mapping algorithms is based on spectral distance rather than acoustic feature distance. The spectral distance is measured using Kullback-Leibler (KL) distance:

$$D_{kl}(X,Y) = \int X(\omega)\log(\frac{X(\omega)}{Y(\omega)})d\omega \tag{12}$$

where $X(\omega)$ and $Y(\omega)$ are two spectral envelopes to be measured. Because the KL distance has the important property that it emphasizes differences in spectral regions with high energy more than differences in spectral regions with low energy, spectral peaks are emphasized more than valleys between the peaks and low frequencies are emphasized more than high frequencies [14].

To compare the performance of voice conversion system, the performance indices[5] is used in objective evaluation on converted spectral envelope, which is defined as:

$$P_{sd} = 1 - \frac{D_{sd(t(n),\hat{t}(n))}}{D_{sd(t(n),s(n))}} \tag{13}$$

Where $D_{sd(t(n),\hat{t}(n))}$ is the spectral distance between target speech and converted speech while $D_{sd(t(n),s(n))}$ is the spectral distance between target and source speech. It is noted that the objective evaluation is performed on spectral envelope rather than spectral feature.

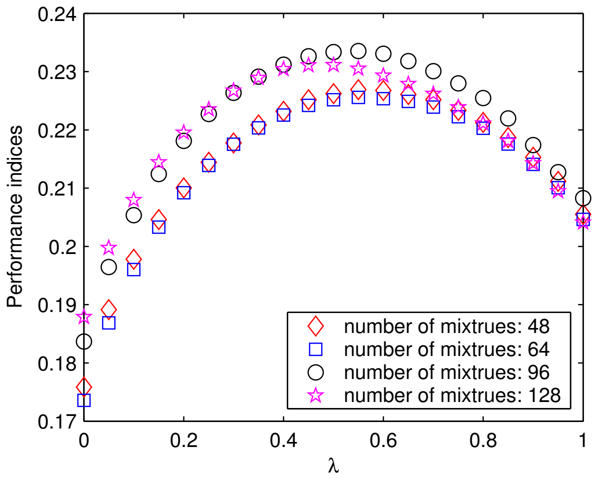

Fig. 3. Spectral distance with different λ using training method1

As shown in equation (10), the λ value adjusts the proportion between GMM method and codebook mapping method in final converted offset vector. The proposed method is traditional GMM method when $\lambda = 0$, while the offset vector is entirely converted by codebook mapping method when $\lambda = 1$.

The results with various λ using different training method are shown in Fig.3. From these figures, it is noted that the performance initially improves with the increase of λ, but then degrades after an optimal λ value, which is about 0.5. As well-known facts, the converted offset vector by GMM is overly smoothing while the results from codebook mapping method have so much spectral details that the converted spectrum is discontinuous. This evaluation indicates that the smoothing problem and the discontinuous problem can be properly counteracted with an optimal λ . The hybrid method with the optimal λ averagely improves 27.2017% than GMM mapping method.

5 Conclusion

In this paper, we propose a new voice conversion method combining the Gaussian Mixture Model (GMM) and codebook mapping method with phoneme-tied weighting. When transforming spectral features, the basic spectral envelopes are converted by GMM method and the envelope-subtracted spectral details are transformed by phone-tied codebook mapping method and GMM method. Evaluations are performed on speech quality and speaker individuality. All experiments shows that the converted speech using the proposed method both reduce the over smoothing problem of GMM method and avoid the discontinuity problem in spectrum of codebook mapping method.

Acknowledgements

This work is fulfilled as an intern in China Research Lab of IBM. I especially thank Liqin Shen for her giving the opportunity to study in CRL, and I am very grateful for studying together with IBM colleagues: Xijun Ma, Yi Liu, Qin Shi, Haiping Li, Yong Qin and Weibin. Zhu.

References

1. Moulines. E and Sagisaka.Y, .Voice conversion: State of the art and perspectives,. Speech Communication, vol. 16, no. 2, pp. 125-126, Feb 1995.
2. L. M. Arslan and D. Talkin, .Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum, in Proc: of the Eurospeech'97, Rhodes, Greece, 1997.
3. Zhi-Wei Shuang, Zi-Xiang Wang, Zhen-Hua Ling, and Ren-Hua Wang, .A novel voice conversion system based on codebook mapping with phoneme-tied weighting,. In Proc. ICSLP, Jeju, Oct. 2004.
4. Y. Stylianou and et al, .Continuous probabilistic transform for voice conversion,. IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131.142, March 1998.

5. Alexander Blouke Kain, High Resolution Voice Transformation, Ph.D. thesis, Oregon Health and Science University, October 2001.
6. T. Toda, H. Saruwatari, and K. Shikano, .Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum,. In Proc. Of ICASSP, 2001, pp. 841.944.
7. Yining Chen, Min Chu, and et al, .Voice conversion with smoothed gmm and map adaptation,. in Proc. Eurospeech, Geneva, Switzerland, Sept. 2003, pp. 2413-2416.
8. H Valbret, et al. Voice transformation using PSOLA technique [J]. Speech Communication. 1992, 11(2-3): 175-187.
9. M Narendranath, et al. Transformation of formants for voice conversion using artificial neural networks [J]. Speech Communi- cation. 1995, 16(2): 207-216.
10. T Watanabe, et al. Transformation of Spectral Envelope for Voice Conversion Based on Radial Basis Function Networks . Proc. ICSLP'2002 [C]. Denver, USA. Sept. 2002: 285-288.
11. E K Kim, et al. Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker [A]. Proc. Eurospeech [C]. Rhodes, Greece, 1997: 2519-2522.
12. M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. J. Acoust. Soc. Jpn. (E), Vol. 11, No. 2, pp.7176, 1990.
13. T. Toda, Alan W Black, Keiichi Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter, In Proc. Of ICASSP, 2005.
14. Esther Klabbers, Raymond Veldhuis. Reducing Audible Spectral Discontinuities, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 1, pp 39-51, JANUARY 2001.