PAN ARCHIVES OF ACOUSTICS
Vol. **38**, No. 1, pp. 39–45 (2013)

DOI: 10.2478/aoa-2013-0005

Spectral Mapping Using Kernel Principal Components Regression for Voice Conversion

Peng SONG[(1)], Li ZHAO[(1)], Yongqiang BAO[(2)]

[(1)] *Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education*
Southeast University
Nanjing, 210096, P.R. China; e-mail: pengsongseu@gmail.com

[(2)] *School of Communication Engineering, Nanjing Institute of Technology*
Nanjing 211167, P.R. China

(*received July 2, 2012; accepted December 7, 2012*)

The Gaussian mixture model (GMM) method is popular and efficient for voice conversion (VC), but it is often subject to overfitting. In this paper, the principal component regression (PCR) method is adopted for the spectral mapping between source speech and target speech, and the numbers of principal components are adjusted properly to prevent the overfitting. Then, in order to better model the nonlinear relationships between the source speech and target speech, the kernel principal component regression (KPCR) method is also proposed. Moreover, a KPCR combined with GMM method is further proposed to improve the accuracy of conversion. In addition, the discontinuity and oversmoothing problems of the traditional GMM method are also addressed. On the one hand, in order to solve the discontinuity problem, the adaptive median filter is adopted to smooth the posterior probabilities. On the other hand, the two mixture components with higher posterior probabilities for each frame are chosen for VC to reduce the oversmoothing problem. Finally, the objective and subjective experiments are carried out, and the results demonstrate that the proposed approach shows greatly better performance than the GMM method. In the objective tests, the proposed method shows lower cepstral distances and higher identification rates than the GMM method. While in the subjective tests, the proposed method obtains higher scores of preference and perceptual quality.

Keywords: spectral mapping, overfitting, oversmoothing, discontinuity, kernel principal component regression.

1. Introduction

In text-to-speech (TTS) synthesis system, the personalized speech generation is one of the biggest challenges. Voice conversion (VC) can be seen as an efficient technique to solve this problem, which refers to modifying the speech spoken by a source speaker to be perceived as that spoken by a target speaker. It also has many other realistic applications, such as identity disguise in secure communications, single channel speech enhancement, and speech-to-speech translation system, etc.

In the last two decades, many spectral mapping approaches have been proposed for VC. The vector quantization (VQ) method is first introduced for spectral transformation (Abe *et al.*, 1988), it can efficiently map the space of the source speaker to that of the target speaker. However, the transformation is performed in discrete spaces, it will lead to poor perceptual speech quality. The linear multivariate regression method is also proposed to improve the VC performance (Valbret *et al.*, 1992), although it can obtain better performance than VQ method, it still can not achieve satisfactory results. The Gaussian mixture model (GMM) method is further proposed for VC (Stylianou *et al.*, 1998; Kain, Macon, 1998), the results show that it can significantly outperform the prior methods. There are also some other methods, such as the artificial neural network (ANN) method (Desai *et al.*, 2010) and the support vector regression (SVR) method (Song *et al.*, 2011). The experimental results show that compared to the GMM method, these methods can get comparable or even better performance.

From the state-of-the-art references, the GMM method is proven to be most prevalent and efficient, and is chosen as the baseline method in the paper. However, the GMM method has some main shortcomings. Firstly, if the spectral features are sparse and the model is complex, the GMM always tends to overfitting, it has been proven that the GMM with a full covariance matrix is difficult to estimate and will lead to overfitting (Mesbahi *et al.*, 2007), and the partial least squares regression (PLSR) approach (Helander *et al.*, 2010) is proposed to avoid this problem efficiently. Secondly, it inevitably introduces the oversmoothing problem, which is caused by the statistical averaging of the model. A hybrid GMM and maximum a posterior (MAP) method (Chen *et al.*, 2003) has been proposed to solve this problem, and the global variance (GV) is considered to reduce this issue (Toda *et al.*, 2001). Thirdly, the traditional GMM method deals with every frame independently regardless of the adjacent frames, this will lead to discontinuities in the converted spectral features and degrade the perceptual quality. Similar to the hidden Markov model (HMM) based speech synthesis, the relationships between the static and dynamic features are considered to obtain the optimal spectral trajectory (Toda *et al.*, 2005).

In this paper, we first propose a VC method based on the principal component regression (PCR). The numbers of principal components are adjusted to overcome the overfitting problem. In order to describe the nonlinear mapping between the source speech and target speech, the kernel PCR (KPCR) method is proposed, and the combined KPCR and GMM method is further presented to improve the VC performance. Then, in order to solve the discontinuity problem, an adaptive median filtering strategy, which is prevalent in image processing, is adopted to smooth the posterior probabilities. Meanwhile, only the two mixture components with higher posterior probabilities for each frame are chosen to reduce the oversmoothing problem. Finally, the objective and subjective experiments are carried out. Compared to the baseline GMM method, the proposed approach can efficiently avoid the overfitting problem when the number of training utterances is limited. Meanwhile, the converted speech using the proposed method shows better perceptual quality, and is much closer to the target one.

The paper is organized as follows, Sec. 2 describes the baseline GMM based VC method. Section 3 gives the PCR and KPCR based VC methods, respectively, and also presents the combined KPCR and GMM method. The novel post-processing approaches using posterior probability information are proposed to solve the discontinuity and oversmoothing problems in Sec. 4. The experimental results are given and discussed in Sec. 5. Finally, Sec. 6 draws the conclusions of this paper.

2. GMM based spectral mapping

In the GMM based VC method, the GMM is employed to model the augmented source and target spectral features. Let $\mathbf{x}$ and $\mathbf{y}$ denote the spectral sequences of source and target speakers, respectively, where $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ and $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$ aligned by the dynamic feature warping (DTW) algorithm. The distribution of the augmented features $\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$ can be seen as a mixture of M Gaussian components, and the probability density function can be written as

$$\begin{aligned} p\begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix} &= \sum_{m=1}^{M} \alpha_m N\left[\begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right], \\ \sum_{m=1}^{M} \alpha_m &= 1, \end{aligned} \tag{1}$$

where $N\left[\begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right]$ denotes a Gaussian distribution, α_m is a prior probability of each frame belonging to the m-th component, and $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and covariance matrix of the m-th Gaussian component, respectively, which can be represented as

$$\boldsymbol{\mu}_m = \begin{bmatrix}\boldsymbol{\mu}_m^x\\ \boldsymbol{\mu}_m^y\end{bmatrix}, \qquad \boldsymbol{\Sigma}_m = \begin{bmatrix}\boldsymbol{\Sigma}_m^{xx} & \boldsymbol{\Sigma}_m^{xy}\\ \boldsymbol{\Sigma}_m^{yx} & \boldsymbol{\Sigma}_m^{yy}\end{bmatrix}, \tag{2}$$

where $\boldsymbol{\mu}_m^x$ and $\boldsymbol{\mu}_m^y$ denote the mean vectors of the m-th component for source and target speakers, respectively. $\boldsymbol{\Sigma}_m^{xx}$, $\boldsymbol{\Sigma}_m^{yy}$, $\boldsymbol{\Sigma}_m^{xy}$ and $\boldsymbol{\Sigma}_m^{yx}$ are the blocks of covariance matrix of the m-th component. The unknown parameters α_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ can be estimated by the expectation maximization (EM) algorithm.

The spectral mapping can be seen as a regression problem. Let $F(\,)$ denotes the conversion function, the total squares error can be shown as

$$\varepsilon = \sum_{n=1}^{N} |\mathbf{y}_n - F(\mathbf{x}_n)|^2 . \tag{3}$$

Introducing the least squares estimation (LSE) algorithm, the unknown parameters can be computed, and the conversion function can be represented as

$$\begin{aligned} F(\mathbf{x}_n) &= E(\mathbf{y}_n|\mathbf{x}_n) \\ &= \sum_{m=1}^{M} p_m(\mathbf{x}_n)\left[\boldsymbol{\mu}_m^y + \frac{\boldsymbol{\Sigma}_m^{yx}}{\boldsymbol{\Sigma}_m^{xx}}(\mathbf{x}_n - \boldsymbol{\mu}_m^x)\right], \end{aligned} \tag{4}$$

where $p_m(\mathbf{x}_n)$ denotes the posterior probability of $\mathbf{x}_n$ belonging to the m-th component, and is given by

$$p_m(\mathbf{x}_n) = \frac{\alpha_m N(\mathbf{x}_n, \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^{xx})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}_n, \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})} . \tag{5}$$

3. Proposed spectral mapping using KPCR

3.1. PCR based Spectral mapping

When the training utterances are sparse and the number of GMM components is large, the GMM based VC always tends to overfitting. In the paper, a PCR based VC method is proposed to solve this problem. The PCR method is a linear regression, and the principal component analysis (PCA) (JOLLIFFE, 1982) is embedded to work for regression. It can efficiently overcome the multicollinearity and the overfitting problems.

Mathematically, given the spectral sequences of the source and target speakers, $\mathbf{x}$ and $\mathbf{y}$, respectively. By subtracting the mean vectors $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$, we shall obtain the zero-mean matrices $\widetilde{\mathbf{x}} = [\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, ..., \widetilde{\mathbf{x}}_N]^{\mathrm{T}}$ and $\widetilde{\mathbf{y}} = [\widetilde{\mathbf{y}}_1, \widetilde{\mathbf{y}}_2, ..., \widetilde{\mathbf{y}}_N]^{\mathrm{T}}$, respectively. The n-th frames of the spectral features take the forms as

$$\widetilde{\mathbf{x}}_n = \mathbf{x}_n - \boldsymbol{\mu}_x\,, \qquad \widetilde{\mathbf{y}}_n = \mathbf{y}_n - \boldsymbol{\mu}_y\,; \tag{6}$$

$$\boldsymbol{\mu}_x = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,, \qquad \boldsymbol{\mu}_y = \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\,. \tag{7}$$

In order to perform PCA, the covariance matrix of the source speech is calculated as given by

$$\boldsymbol{\Sigma} = \frac{1}{N}\widetilde{\mathbf{x}}^{\mathrm{T}}\,\widetilde{\mathbf{x}}. \tag{8}$$

Performing the eigen-decomposition of the covariance matrix $\boldsymbol{\Sigma}$, we shall obtain

$$\boldsymbol{\Lambda} = \mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{Q}, \tag{9}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_K)$ and $\mathbf{Q}_K = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_K]$ with the K sorted eigenvalues as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$, and the associated eigenvectors $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_K$ of the matrix $\boldsymbol{\Sigma}$. So the ordinary PCA projection is given by

$$\mathbf{S}_K = \widetilde{\mathbf{x}}\mathbf{Q}_K\,. \tag{10}$$

After choosing a suitable number of principal components, the important features of $\widetilde{\mathbf{x}}$ are retained by $\mathbf{S}_K$. Then, we perform a linear regression between $\mathbf{S}_K$ and the target spectral feature sequence $\widetilde{\mathbf{y}}$,

$$\widetilde{\mathbf{y}} = \mathbf{S}_K\mathbf{C} + \mathbf{E}, \tag{11}$$

where $\mathbf{C}$ is the regressor, and $\mathbf{E}$ is the residual error. Employing the LSE algorithm, the unknown parameter $\mathbf{C}$ is given by

$$\mathbf{C} = \left(\mathbf{S}_K\mathbf{S}_K^{\mathrm{T}}\right)^{-1}\mathbf{S}_K^{\mathrm{T}}\widetilde{\mathbf{y}}, \tag{12}$$

and it is necessary to turn to $\mathbf{x}$ for prediction, so the conversion function can be written as

$$F(\mathbf{x}_n) = (\mathbf{x}_n - \boldsymbol{\mu}_x)\mathbf{Q}_K\mathbf{C} + \boldsymbol{\mu}_y\,. \tag{13}$$

3.2. KPCR based spectral mapping

In fact, the relationships between the spectral features of source speech and target speech are nonlinear, so the KPCR algorithm is further proposed for the spectral mapping. It have many advantages, such as it can make a perfect nonlinear regression, shows better mapping performance than the PCR method, and can avoid the overfitting problem efficiently (SCHOLKOPF *et al.*, 1997). Let $\phi(\widetilde{\mathbf{x}})$ be a nonlinear mapping function from the lower feature space to the higher feature space, so that $\widetilde{\mathbf{x}}_n$ is thereby projected to be $\phi(\widetilde{\mathbf{x}}_n)$. Assuming $\sum_{n=1}^{N}\phi(\widetilde{\mathbf{x}}_n) = 0$, the covariance matrix in the feature space is given by

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\phi(\widetilde{\mathbf{x}})\phi(\widetilde{\mathbf{x}})^{\mathrm{T}}. \tag{14}$$

The eigenvector $\mathbf{q}_l$ can be obtained by a linear combination of $\phi(\widetilde{\mathbf{x}}_n)$, and given by

$$\mathbf{q}_l = \sum_{n=1}^{N}\beta_{l,n}\phi(\widetilde{\mathbf{x}}_n). \tag{15}$$

The projection onto eigenvector $\mathbf{q}_l$ is given by

$$\begin{aligned} B_l &= \langle \mathbf{q}_l,\, \phi(\widetilde{\mathbf{x}}_j)\rangle = \sum_{n=1}^{N}\beta_{l,n}\phi(\widetilde{\mathbf{x}}_j)^{\mathrm{T}}\phi(\widetilde{\mathbf{x}}_n) \\ &= \sum_{n=1}^{N}\beta_{l,n}k(\widetilde{\mathbf{x}}_n, \widetilde{\mathbf{x}}_j), \end{aligned} \tag{16}$$

where $k(\widetilde{\mathbf{x}}_n, \widetilde{\mathbf{x}}_j)$ is the kernel function, and $l = 1, 2, ..., L$ are the numbers of principal components. However, in general, the projected data given by $\phi(\widetilde{\mathbf{x}})$ does not have zero mean, and it is too difficult to calculate the mean values of $\phi(\widetilde{\mathbf{x}})$. Employing the kernel function can avoid directly working in the feature space. Denote the kernel matrix $\mathbf{K}$ and the centralized kernel matrix $\widetilde{\mathbf{K}}$, respectively, and represented as

$$[\mathbf{K}]_{i,j} = [k(x_i, x_j)]\,, \qquad [\widetilde{\mathbf{K}}]_{i,j} = \left[\widetilde{k}(x_i, x_j)\right]. \tag{17}$$

The $\widetilde{\mathbf{K}}$ is computed after centralizing the kernel function, and given by

$$\widetilde{\mathbf{K}} = \mathbf{K} - \mathbf{I}_N\mathbf{K} - \mathbf{K}\mathbf{I}_N + \mathbf{I}_N\mathbf{K}\mathbf{I}_N\,, \tag{18}$$

where $\mathbf{I}_N$ is a $N \times N$ matrix with every element taking the value of $1/N$. So the formula (16) can be modified as

$$\mathbf{B}_l = \sum_{n=1}^{N}\beta_{l,n}\widetilde{k}(\widetilde{\mathbf{x}}_n, \widetilde{\mathbf{x}}_j). \tag{19}$$

The KPCR is based on the analysis of KPCA, and the regression function takes the form as follows

$$F(\mathbf{x}_n) = \mathbf{B}_L\mathbf{w} + \boldsymbol{\mu}_y\,, \tag{20}$$

where

$$\mathbf{B}_L = \sum_{i=1}^{N} \beta_{l,i} K(\widetilde{\mathbf{x}}_i, \mathbf{x}_n - \boldsymbol{\mu}_x)$$

and

$$\mathbf{w} = (\mathbf{B}_L \mathbf{B}_L^{\mathrm{T}})^{-1} \mathbf{B}_L^{\mathrm{T}} \widetilde{\mathbf{y}}.$$

The optimal number of principal components is critical for the KPCR method, and is chosen by a ten-fold cross-validation in the paper.

3.3. Combining with GMM

From the state-of-the-art references, it is not possible to perfectly describe the relationships between source and target speech using single regression function (**Helander** *et al.*, 2010; Song *et al.*, 2012). In this paper, the KPCR method is further extended to combine with GMM. Similar to the forms of GMM based VC function mentioned in Sec. 2, the regression function can be seen as a mixture of local regressions. In the m-th mixture, the regression function is given by

$$F_m(\mathbf{x}_n) = \mathbf{B}_{L,m} \mathbf{w}_m + \boldsymbol{\mu}_m^y \,, \tag{21}$$

where $\mathbf{B}_{L,m} = \sum_{i=1}^{N} \beta_{l,i} k(\mathbf{x}_i - \boldsymbol{\mu}_x,\, \mathbf{x}_n - \boldsymbol{\mu}_m^x)$, and $\mathbf{w}_m$ is the regressor. So the global conversion function can be calculated as shown

$$F(\mathbf{x}_n) = \sum_{m=1}^{M} p_m(\mathbf{x}_n) F_m(\mathbf{x}_n), \tag{22}$$

where $p_m(\mathbf{x}_n)$ is the posterior probability as shown in Eq. (5).

The kernel selection is also essential to the accuracy of spectral transformation. There are many kinds of kernel functions, such as linear kernel, polynomial kernel, Gaussian kernel, and wave kernel, etc. Among which, the Gaussian kernel is simple and efficient, and is chosen for spectral mapping, which takes the form as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right), \tag{23}$$

where σ is the width of the Parzen window, and the decent range is used to find its best value. In the paper, the σ is optimized as 0.6 also by a ten-fold cross-validation.

4. Post-processing using posterior probability

The proposed method can be seen as an extension of GMM method, and can efficiently avoid the overfitting problem. But it is well known that except the overfitting problem, GMM method has another two main shortcomings, one is the discontinuity problem because of the frame-by-frame based transformation, and the other is the oversmoothing problem caused by the statistical averaging of the model. In this section, the two problems are analyzed and efficiently overcome using the posterior probability information.

From the prior studies, it is well known that each frame often obviously dominates one component of the GMM (Helander *et al.*, 2010). Table 1 gives the results obtained from the 1500 frames based on a GMM with 8 components, we can easily find that most of the frames have the maximum posterior probabilities higher than 90%, and only a very few with the maximum posterior probabilities lower than 50%. So it can be assumed that each frame belongs to single component of GMM, and it also has been proven that the changes of posterior probabilities from one component to another are rapid (Helander *et al.*, 2010), which will lead to the discontinuities in the converted speech, and degrade the perceptual quality.

Table 1. Number of frames in different maximum posterior probabilities.

Maximum posterior probability [%]	Number of frames
90–100	1052
80–90	213
70–80	78
60–70	52
50–60	38
0–50	67

In this paper, in order to reduce the discontinuity problem, an adaptive median filtering approach is adopted to smooth the posterior probabilities. The adaptive median filtering is a prevalent and efficient algorithm in image processing, and can efficiently smooth the images while preserving the details (Hwang, Haddad, 1995). Let W be the rectangle moving window, and the initial length and maximum length are set as 3 and 7, respectively. In each component of GMM, p_{cur} is the posterior probability of current frame, $p_{\min}$, $p_{\max}$ and p_{med} are the minimum, maximum and median values of the posterior probability in the window, respectively. The adaptive filtering can be seen as a two level structure:

Level A:

$$A_1 = p_{\mathrm{med}} - p_{\min} \,,$$
$$A_2 = p_{\mathrm{med}} - p_{\max} \,.$$

In Level A, if $A_1 > 0$ and $A_2 < 0$, then go to level B. Or increase the length of W and repeat level A. If the length exceeds the maximum, then p_{cur} is the output.

Level B:

$$B_1 = p_{\mathrm{cur}} - p_{\min} \,,$$
$$B_2 = p_{\mathrm{cur}} - p_{\max} \,.$$

In level B, if $B_1 > 0$ and $B_2 < 0$, then p_{cur} is the output, or p_{med} is chosen as the output.

Meanwhile, the statistical averaging of GMM will introduce the oversmoothing problem, and smoothing the posterior probabilities by adaptive median filtering will also aggravate this issue to some extent, although it can efficiently decrease the discontinuities. As mentioned above, in most cases, each frame dominates one component of GMM. So a new approach is proposed, to each frame, the mixture component with highest posterior probability is chosen, and the local regression in this component is adopted to replace the global transformation. Theoretically, it can efficiently solve the oversmoothing problem. Meanwhile, it can be also observed that after adaptive smoothing, there are always two higher posterior probabilities to each frame in many cases, especially when the highest posterior probabilities of the adjacent frames belong to the different mixture components. So a top-two selection strategy is further proposed, in which, after adaptive smoothing, the mixture components with two higher posterior probabilities are chosen, and the two probabilities are normalized so that they sum to a unity. The transformation function can be modified as

$$F(x_n) = \sum_{m=1}^{2} p_m F_m(x_n), \tag{24}$$

where p_m is the normalized posterior probability, which satisfies $\sum_{m=1}^{2} p_m = 1$.

5. Experiments

We perform the experiments on CMU ARCTIC corpus. Two male and two female speakers are chosen. 25 parallel utterances of each speaker are prepared for the experiments, in which, 15 utterances (about 5 minutes) are used for training, while the others are used for testing. The 30-order Mel-cepstral coefficients (MCEPs) and their Δ and Δ^2 features (totally 90th order) are chosen to represent the spectral features (Toda *et al.*, 2005). Four kinds of VC methods are compared, they are the baseline GMM method (GMM), the PCR method (PCR), the KPCR method (KPCR), and the combined KPCR and GMM method with post-processing (GKPCR). Meanwhile, four types of VC strategies are adopted, including male-to-male transformation (M-M), male-to-female transformation (M-F), female-to-female transformation (F-F), and female-to-male transformation (F-M). The objective and subjective experiments are carried out to evaluate the performance of the proposed method, respectively. The Mel-cepstral distortion (MCD) and speaker identification system are chosen for the objective evaluation, while the ABX test and MOS test are used for the subjective evaluation. The number of GMM is set as 16. The number of principal components is optimized as 40, and 8 experienced people are employed for the listening tests.

5.1. Objective evaluation

The MCD is a common method to evaluate the objective performance of VC. It measures the cepstral distance between the converted speech and target speech, and the formula can be computed as follows

$$MCD = 10/\log 10 \sqrt{2 \sum_{j=1}^{D} (\mathbf{mc}_j^c - \mathbf{mc}_j^t)^2}, \tag{25}$$

where $\mathbf{mc}_j^c$ and $\mathbf{mc}_j^t$ are the i-th coefficients of converted and target MCEPs, respectively, and D is the dimension of MCEPs.

Figures 1 to 4 give the average MCD results of different methods, it should be noted that a lower MCD value means a better VC performance. We can observe that when the numbers of principal components are adjusted properly, the proposed method can show greatly better performance than the baseline GMM method.

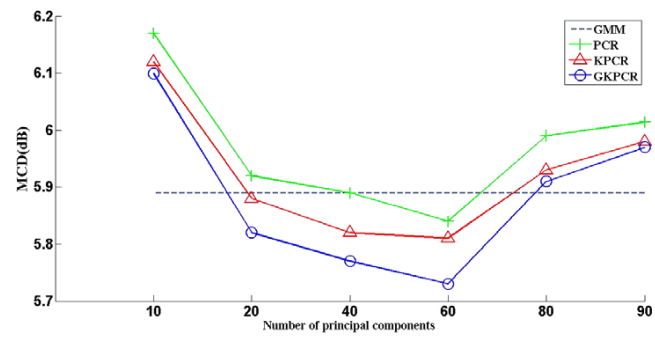

Fig. 1. MCD of M-M.

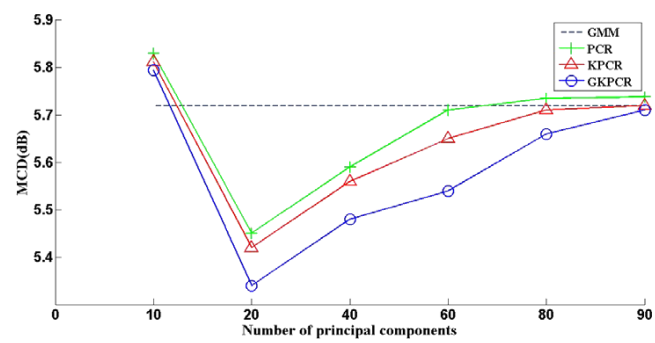

Fig. 2. MCD of M-F.

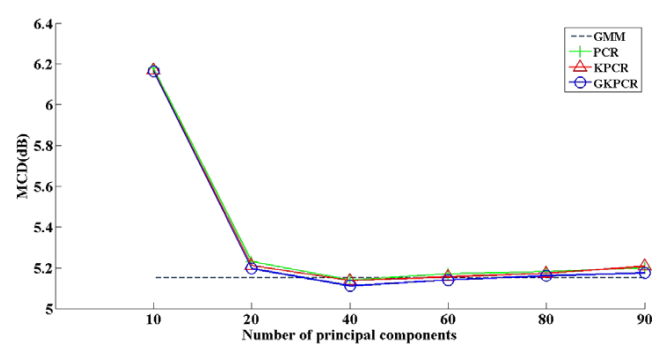

Fig. 3. MCD of F-F.

In the F-F VC, when the numbers of principal components are 40 and 60, respectively, the proposed method shows a little better performance. While in other cases, the proposed method is superior to the other methods.

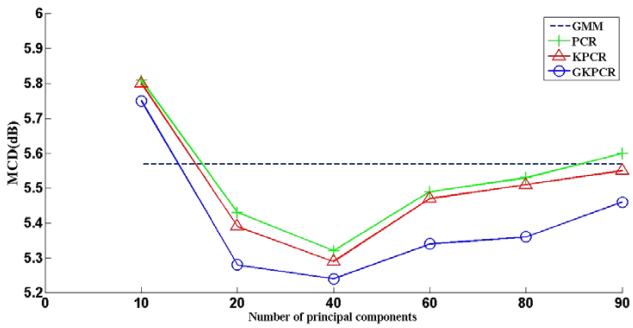

Fig. 4. MCD of F-M.

The GMM based speaker identification system (Reynolds *et al.*, 2000) is also adopted for the objective evaluation of VC. Let λ_S and λ_T be the trained speaker models of source and target speakers, respectively, the performance measure θ_{ST} is given by

$$\theta_{ST} = \log P(\mathbf{O}|\lambda_T) - \log P(\mathbf{O}|\lambda_S), \tag{26}$$

where $\mathbf{O}$ is the observed sequence of spectral features, the average recognition results are summarized in Table 2. It can be found that before conversion all the values of θ_{ST} are minus, which demonstrates that the speech is recognized as the source speech, and after conversion the values will become plus, which indicates that the converted speech is recognized as the target one. It can be also observed that in all cases, the converted speech is more likely to be recognized as the target one. Compared to other methods, the proposed GKPCR method can efficiently increase the values of θ_{ST}.

Table 2. Results of the speaker identification.

Strategy	θ_{ST}				
	Before conversion	After conversion			
		GMM	PCR	KPCR	GKPCR
M-M	−4.53	+3.25	+3.31	+3.34	+3.36
M-F	−5.16	+3.01	+3.08	+3.09	+3.12
F-F	−4.02	+3.51	+3.55	+3.58	+3.63
F-M	−5.41	+3.14	+3.26	+3.27	+3.34

5.2. Subjective evaluation

The ABX test is carried out for the similarity between converted and target speech. In this method, A and B are the converted speech using GKPCR method and other methods, respectively. X is the target speech. The listeners are asked to choose whether A or B is closer to X. Figure 5 summarizes the overall performance of different VC methods, and the results confirms the conclusions of objective tests. It can be found that the proposed GKPCR method shows significantly better preference than the baseline GMM method (With probability of 90%). Meanwhile, compared to the PCR method and KPCR method, the proposed method also shows obvious superiority.

Fig. 5. Results of ABX test.

The MOS test is also performed to evaluate the perceptual quality of the converted speech. The converted utterances using different methods are shown to the listeners, who are asked to rate the converted speech using a 5-point score from 1 "bad" to 5 "excellent". The average scores of different methods are summarized in Table 3. The mean score and standard deviation (SD) are given, respectively, and the confidence interval is set as 95%. We can easily find that the proposed GKPCR method greatly outperforms the baseline GMM method and other methods, with higher mean scores and lower SD.

Table 3. Results of perceptual quality evaluation.

	GMM	PCR	KPCR	GKPCR
Mean score	3.72	3.79	3.83	4.09
SD	0.51	0.48	0.47	0.42

6. Conclusions

In this paper, a novel spectral mapping method using KPCR is proposed for VC. The proposed KPCR method can efficiently solve the overfitting problem, and also perfectly describe the nonlinear relationships between the source speech and target speech. To further improve the VC performance, the combined KPCR and GMM method is proposed. The discontinuity and oversmoothing problems of traditional GMM method are also analyzed, and can be efficiently overcome by adopting the novel post-processing strategies using posterior probability information. The adaptive median filter is employed to reduce the discontinuities, while a top-two selection strategy is proposed to solve the oversmoothing. Experimental results demonstrate that the proposed method greatly outperforms the traditional GMM method.

Acknowledgment

The authors acknowledge the work is supported by the National Natural Science Foundation of China (Grant Nos. 60975017 and 51075068), the Natural Science Foundation of Higher Education Institutions of Jiangsu Province (Grant No. 10KJB510005), and the Natural Science Foundation of Guangdong Province (Grant No. 10252800001000001).

References

1. Abe M., Nakamura S., Shikano K., Kuwabara H. (1998), *Voice conversion through vector quantization*, Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing, pp. 655–658, New York.
2. Chen Y., Chu M., Chang E., Liu J., Liu R. (2003), *Voice conversion with smoothed GMM and MAP adaptation*, Proceedings of Eurospeech 2003, pp. 2413–2416, Geneva.
3. Desai S., Black A. W., Yegnanarayana B., Prahallad K. (2010), *Spectral mapping using artificial neural networks for voice conversion*, IEEE Transactions on Audio, Speech, and Language Processing, **18**, 5, 954–964.
4. Helander E., Virtanen T., Nurminen J., Gabbouj M. (2010), *Voice conversion using partial least squares regression*, IEEE Transactions on Audio, Speech, and Language Processing, **18**, 5, 912–921.
5. Hwang H., Haddad R. A. (1995), *Adaptive median filters: new algorithms and results*, IEEE Transactions on Image Processing, **4**, 4, 499–502.
6. Jolliffe I. T. (1982), *A note on the use of principal components in regression*, Applied Statistics, **31**, 3, 300–303.
7. Kain A., Macon M. W. (1998), *Spectral voice conversion for text-to-speech synthesis*, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 285–288, Seattle.
8. Mesbahi L., Barreaud V., Boeffard O. (2007), *GMM-based speech transformation systems under data reduction*, Proceedings of the 6th ISCA workshop on Speech Synthesis, pp. 119–124, Bonn.
9. Reynolds D. A., Quatieri T. F., Dunn R. B. (2000), *Speaker verification using adapted Gaussian mixture models*, Digital Signal Processing, **10**, 1, 19–41.
10. Song P., Bao Y. Q., Zhao L., Zou C. R. (2011), *Voice conversion using support vector regression*, Electronics Letters, **47**, 18, 1045–1046.
11. Song P., Jin Y., Zhao L., Zou C. R. (2012), *Voice conversion based on hybrid SVR and GMM*, Archives of Acoustics, **37**, 2, 143–149.
12. Scholkopf B., Smola A., Muller K. R. (1997), *Kernel principal component analysis*, Proceedings of the 7th International Conference on Artificial Neural Networks, pp. 583–588, Berlin.
13. Stylianou Y., Cappe O., Moulines E. (1998), *Continuous probabilistic transform for voice conversion*, IEEE Transactions Speech and Audio Processing, **6**, 2, 131–142.
14. Toda T., Saruwatari H., Shikano K. (2001), *Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum*, Proceedings of the 2001 International Conference on Acoustics, Speech, and Signal Processing, pp. 841–944, Salt Lake City.
15. Toda T., Black A. W., Tokuda K. (2005), *Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter*, Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing, pp. 9–12, Philadelphia.
16. Valbret H., Mulines E., Tubach J. (1992), *Voice transformation using PSOLA techniques*, Speech Communication, **11**, 2-3, 175–187.