www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **37**, No. 2, pp. 143–149 (2012)

DOI: 10.2478/v10168-012-0020-9

Voice Conversion Based on Hybrid SVR and GMM

Peng SONG[(1)], Yun JIN[(2),(3)], Li ZHAO[(1)], Cairong ZOU[(1)]

[(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 Physics and Electronic Engineering, Xuzhou Normal University*
Xuzhou, 221116, P.R. China

[(3)] *Key Laboratory of Child Development and Learning Science of Ministry of Education*
Southeast University
Nanjing, 210096, P.R. China

(*received October 20, 2011; accepted March 13, 2012*)

A novel VC (voice conversion) method based on hybrid SVR (support vector regression) and GMM (Gaussian mixture model) is presented in the paper, the mapping abilities of SVR and GMM are exploited to map the spectral features of the source speaker to those of target ones. A new strategy of F0 transformation is also presented, the F0s are modeled with spectral features in a joint GMM and predicted from the converted spectral features using the SVR method. Subjective and objective tests are carried out to evaluate the VC performance; experimental results show that the converted speech using the proposed method can obtain a better quality than that using the state-of-the-art GMM method. Meanwhile, a VC method based on non-parallel data is also proposed, the speaker-specific information is investigated using the SVR method and preliminary subjective experiments demonstrate that the proposed method is feasible when a parallel corpus is not available.

Keywords: voice conversion, support vector regression, Gaussian mixture model, F0 prediction, speaker-specific information.

1. Introduction

VC (voice conversion) is a technique which refers to transforming the characteristics of a source speaker to those of a target speaker. A wide variety of applications are available, ranging from expressive text-to-speech synthesis and preserving speaker individuality in an ultra low bit communication system, to aiding the speech-impaired people.

Several VC methods have been proposed over the past decades, such as the mapping codebook (Abe *et al.*, 1988), the discrete transformation function (Mizuno, Abe, 1995), GMM (Stylianou *et al.*, 1998; Kain, Macon, 1998), and the ANN (artificial neural network) (Desai *et al.*, 2010). In the mapping codebook method, the VQ (vector quantization) a clustering approach is applied to the spectral parameters of the source and target speakers and the mapping function is obtained from the two resulting codebooks. One main shortcoming of this technique is that the converted parameters are limited in a discrete space, which will cause severe degradation of the speech quality. The discrete transformation using a piecewise linear function has been then proposed to replace the mapping codebook method, however it results in discontinuities in the converted speech. In the GMM based VC method, the conversion is established on the basis of continuous probabilistic functions and the experimental results show that better results can be obtained compared to the other prior transformation methods. The ANN method as a continuous and nonlinear function has also been ved investigated and it has been proved that results comparable to those of the GMM method can be achieved, but there are several main shortcomings, such as a greater computing burden, multiple local minima depending on empirical risk minimization and always involving the overfitting problem. From the state-of-the-art references,

the GMM is the most popular and well-established VC method. Many improved GMM based methods have been proposed, such as the GMM and DFW (dynamic frequency warping) method (Toda *et al.*, 2001), the GMM and MAP (maximum a posteriori) method (Chen *et al.*, 2003), and the GMM using ML (maximum likelihood) parameter generation method (Toda *et al.*, 2005). These methods can avoid more or less the over-smooth phenomenon or discontinuity problem. In fact, the relationship between the source and target speakers is non-linear. Different from traditional GMM or ANN methods, the SVR approach can perfectly map the non-linear relationship between the source and target speakers, it needs less training data and is less prone to local minima. So a hybrid SVR and GMM VC approach is proposed, in which the SVR mapping is carried out instead of the linear regression in each component of GMM.

The above mentioned methods focus on the conversion of the spectral envelope, but the prosodic features, particularly the F0s, are also very important to the speaker individuality. Most of the current VC systems often employ simple F0 transformations, such as transforming the mean F0s from the source speaker to the target one (Kain, Macon, 1998), and shifting the means and variances of the F0 distribution to map the source and target speakers (Inanoglu, 2003). The strategy adopted in this paper is different from these methods, the F0s and spectral features are modeled in a joint model and the F0s are predicted from the spectral parameters using SVR.

The common VC methods are carried out basing on a parallel corpus, which contains the same utterances of the source and target speakers. It is evident that collecting such a corpus is difficult and even impossible in many cases. Several approaches have been proposed to resolve this issue, similar to the adaptation methods in speech recognition, the adaptation technique is employed to fit spectral features of different speakers based on a previously trained conversion function (Mouchtaris *et al.*, 2004), the conversion function is extended to a ML formulation which requires non-parallel data of the source and target speakers (Ye, Young, 2006), and an iterative method based on the acoustic distance is proposed that proves to be suitable for the text-independent and cross-language VC (Erro, Moreno, 2007). In the paper, the SVR method is exploited to capture the specific information of the target speaker, which can efficiently decrease the need for parallel data.

The remainder of the paper is organized as follows. Section 2 describes the conversion methods of spectral parameters in details and also gives a new F0 transformation approach. Section 3 introduces the VC based on a non-parallel corpus. The experimental results are given and discussed in Sec. 4. The conclusions are finally drawn in Sec. 5.

2. Hybrid SVR and GMM based on VC

The flowcharts of VC using the proposed method are shown in Fig. 1 and Fig. 2, respectively, the STRAIGHT (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram) analysis and synthesis method (Kawahara *et al.*, 1999) is adopted to extract the spectral features and F0s, and DTW (dynamic time warping) technique is used to align the spectral features.

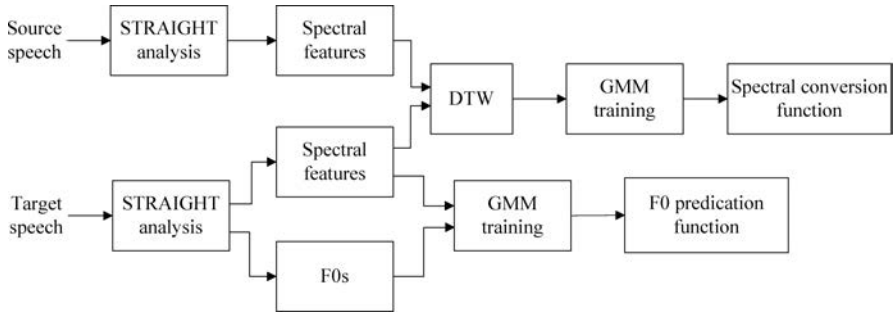

Fig. 1. Training structure.

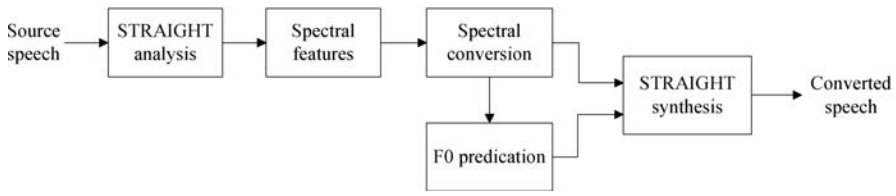

Fig. 2. Conversion structure.

2.1. Baseline GMM based on spectral conversion

There are two mainstream VC approaches based on GMM, the LSE (least squared estimation) method (Stylianou *et al.*, 1998) and JDE (joint density estimation) method (Kain, Macon, 1998) respectively. They show equivalent performance and the latter one is chosen as the baseline of the presented method.

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ be the sequences of the spectral parameters of source and target speakers, respectively, where $\mathbf{x}_i = \{x_{i1}, ..., x_{iJ}\}$ and $\mathbf{y}_i = \{y_{i1}, ..., y_{iJ}\}$. $\mathbf{x}$ is aligned to the counterpart $\mathbf{y}$ to get a parallel sequence pair $\mathbf{z} = (\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}$ (where the superscript T denotes transposition), which is used to train the joint GMM parameters $(\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The GMM can be written as a sum of M Gaussian components, which takes the form

$$p(\mathbf{z}) = \sum_{i=1}^{M} \alpha_i N(\mathbf{z}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where α_i denotes the prior probability of the i-th component and satisfies $\sum_{i=1}^{M} \alpha_i = 1$, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance matrices of the i-th component. Minimizing the mean squared errors between the converted and target speech, the conversion function can be written as

$$F(\mathbf{x}) = E(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{M} p_i(\mathbf{x}) \left[\boldsymbol{\mu}_i^y + \frac{\boldsymbol{\Sigma}_i^{yx}}{\boldsymbol{\Sigma}_i^{xx}}(\mathbf{x} - \boldsymbol{\mu}_i^x)\right], \tag{2}$$

$$p_i(\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{k=1}^{M} \alpha_k N(\mathbf{x}, \boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx})}, \tag{3}$$

where

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

and $p_i(\mathbf{x})$ is the probability of $\mathbf{x}$ belonging to the i-th component.

2.2. Proposed hybrid SVR and GMM methods

A novel hybrid SVR and GMM VC approach is proposed in this paper, the SVR is adopted in each GMM component, which is radically different with traditional GMM or ANN methods. It performs a perfect non-linear mapping between the source and target speakers and can efficiently avoid the over-fitting problem, and finds always the global minima. Different from the traditional one-dimension output SVR, a multi-dimensional SVR based VC is proposed; the conversion function as a regression in the m-th component is given by

$$f_m(\mathbf{x}) = \langle \mathbf{W}, \varphi(\mathbf{x}) \rangle + \mathbf{b}, \tag{4}$$

where $\varphi(\mathbf{x})$ is a non-linear mapping function from a low dimensional space to a higher one, $\mathbf{W} = \{w_1, ..., w_J\}^{\mathrm{T}}$ and $\mathbf{b} = \{b_1, ..., b_J\}^{\mathrm{T}}$ define two J-dimensional regressors in the higher dimensional space, respectively. The regression function can be resolved by the optimization problem:

$$\begin{aligned} &\min \frac{1}{2} \sum_{j=1}^{J} \|w_j\|^2 + C \sum_{i=1}^{N} L(\xi_i), \\ &\text{s.t. } \|\mathbf{y}_i - \langle \mathbf{W}, \varphi(\mathbf{x}_i) \rangle - \mathbf{b}\| \le \varepsilon + \xi_i, \\ &\xi_i \ge 0, \qquad i = 1, ..., N. \end{aligned} \tag{5}$$

Here C is a penalty factor, ε and ξ_i are variables to cope with the cost errors on the training points, and $L(\xi_i)$ denotes the cost function. Instead of the hypercubic intensive zone used in the ε-based SVR, a hyperspherical insensitive zone is adopted to cope with the multi-dimensional output, and an IRWLS (iterative reweighted least squares) method (Perez–Cruz *et al.*, 2000; 2002) is employed to resolve the Lagrangian as follows

$$\begin{aligned} L(\mathbf{W}, \mathbf{b}) &= \frac{1}{2} \sum_{j=1}^{J} \|w_j\|^2 + C \sum_{i=1}^{N} L(\xi_i) \\ &- \sum_{i=1}^{N} \alpha_i [(\varepsilon + \xi_i)^2 - \|\mathbf{y}_i - \langle \mathbf{W}, \varphi(\mathbf{x}_i) \rangle - \mathbf{b}\|^2] - \sum_{i=1}^{N} \mu_i \xi_i, \end{aligned} \tag{6}$$

where α_i and μ_i are Lagrange multipliers. Introducing the kernel function and iteration procedures, the unknown w_j and b_j parameters will be computed in each dimension. So the GMM based VC function can be modified as

$$F(\mathbf{x}) = \sum_{i=1}^{M} p_i(\mathbf{x}) f_i(\mathbf{x}). \tag{7}$$

As is well known, the selection of kernel is the key to the SVR performance; The RBF (radial basis function) and the polynomial function are two typical kernel methods. As shown in the literature (Smits, Jordan, 2002), the RBF K_{rbf} has a better interpolation performance, while the polynomial function K_p shows a better extrapolation ability, a mixed kernel is introduced to improve the conversion performance.

$$K_{\mathrm{mix}} = \lambda K_{\mathrm{rbf}} + (1 - \lambda) K_p, \qquad 0 \le \lambda \le 1. \tag{8}$$

The weight λ varies between 0 and 1 at a 0.05 step size, and was finally optimized as 0.85 in the paper.

2.3. F0 transformation

A typical F0 transformation method is based on GMM (Inanoglu, 2003), which takes a form similar to formula (2). Many studies have indicated the relationships existing between spectral parameters and F0s. The joint GMM is used to model F0s and spectral features (En-Najjary *et al.*, 2003). The F0 prediction

from the MFCC (Mel-frequency cepstral coefficient) vectors using GMM and HMM (hidden Markov model) methods (Shao, Milner, 2004) indicates that they can achieve satisfactory results as predicted, but there still exist some shortcomings, such as the non-linearity of the relationships between the spectral parameters and F0s, and need of a large corpus in the training phase. So the SVR method, which makes non-linear mapping with less training data, is adopted and the F0s are predicted from the spectral parameters. Differently from the traditional F0 transformation methods, only the target features are necessary for the training process. The F0 modification is carried out as the following steps:

Step 1. In the training phase, the sequences of spectral parameters $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ and F0s $\mathbf{f} = (f_1, ...f_N)$ of the target speech are calculated using the STRAIGHT method (Kawahara *et al.*, 1999), respectively.

Step 2. Then the EM (expectation maximization) algorithm is employed to model $\mathbf{y}$ and $\mathbf{f}$ in a joint GMM. In each component of GMM, the conversion function is trained between $\mathbf{y}$ and $\mathbf{f}$ using the ε-based SVR with the mixed kernel.

Step 3. In the conversion phase, the F0s are estimated from the converted spectral parameters using the trained SVR conversion functions.

3. VC based on the non-parallel corpus

The previously discussed VC methods are based mainly on parallel training data, which requires the same utterances of the source and target speakers. Recent approaches using a non-parallel corpus have been investigated (Mouchtaris *et al.*, 2004; Ye, Young, 2006); they can get satisfactory results, but still need some prior information from the mapping function between the source and target speakers, which is not always feasible in real applications. In this paper, a SVR method is adopted to capture speaker-specific information, which doesn't need any prior information from the source speaker and makes it possible to do VC from an arbitrary source speaker to the target one.

The idea is stimulated by the speaker-specific mapping for speaker recognition (Misra *et al.*, 2003). Let L denote the linguistic information, and LS correspond to the linguistic and speaker information; a mapping function $\Omega(L)$ is calculated to get the relationships between L and LS and computed using the LSE method on the training data so as to minimize the squared errors,

$$\varepsilon_{SE} = \sum_{i=1}^{N} \|LS_i - \Omega(L_i)\|^2. \tag{9}$$

Assuming m and n are the orders of L and LS, respectively, which are difficult to determine. According to the literature (Misra *et al.*, 2003), a low order of the LP (linear predictive) analysis (m: 4∼8) can grossly capture the linguistic information of the speaker, while a higher LP order (n: >12) can capture both the linguistic and speaker-specific information. Figure 3 shows a flowchart of the training process of SVR based on VC using speaker-specific information; a VTLN (vocal tract length normalization) technique as the pre-processing module is adopted to extract the linguistic information.

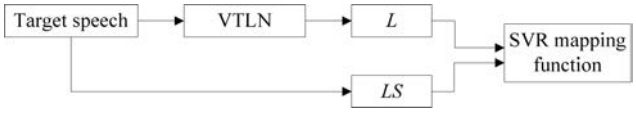

Fig. 3. Training structure of VC using speaker-specific information.

4. Experiments and discussions

The work was based on the ARCTIC database (Kominek, Black, 2004). Two U.S. females and two U.S. males were chosen to evaluate the performance of the proposed VC method. The tests were performed for four speaker pairs: male-to-male (M-M), male-to-female (M-F), female-to-female (F-F), and female-to-male (F-M). 200 phonetically balanced utterances with the same linguistic contents of each speaker were selected, 100 sentences were chosen as the training data, while another 100 sentences were used for testing. The 16-order LSFs (line spectral frequencies) were extracted as spectral parameters, and F0s were derived in a log-scaled domain. Three types of VC methods using parallel corpus were compared: the JDE GMM method (Kain, Macon, 1998), the proposed hybrid SVR and GMM methods, and the proposed VC with the F0 prediction accordingly. The GMM based F0 conversions were performed for the first two methods, and finally they all were evaluated by objective and subjective tests. The VC using non-parallel data was also assessed by subjective tests, the orders of L and LS were optimized as 4 and 16 respectively, and the number of the GMM components M was set as 64.

4.1. Objective evaluation

The normalized mean squared error was employed to evaluate the distance between the converted and target speech, which takes the form

$$\varepsilon_{NE} = \frac{\frac{1}{T}\sum_{i=1}^{N} \|\mathbf{y}_i - F(\mathbf{x}_i)\|^2}{\frac{1}{T}\sum_{i=1}^{N} \|\mathbf{y}_i - \boldsymbol{\mu}^y\|^2}. \tag{10}$$

The figures below summarize the results of the converted speech using the above mentioned three meth-

ods based on parallel data. It is obvious that the normalized errors of the proposed method are significantly lower than those of the GMM baseline method. Compared to the GMM based F0 transformation method, the proposed F0 prediction method can efficiently decrease the normalized error. When the numbers of GMM components increase above 64, the trends of normalized errors become nearly constant, which means that the number of training data is enough to train the model.

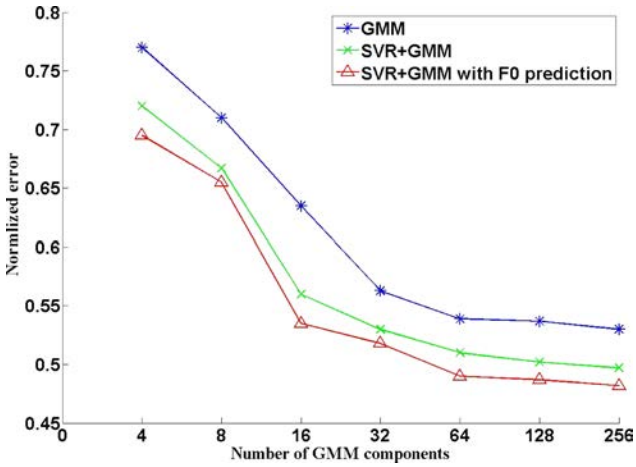

Fig. 4. Normalized error (M-M).

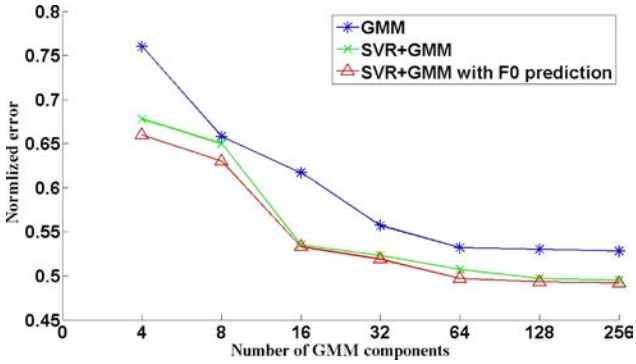

Fig. 5. Normalized error (M-F).

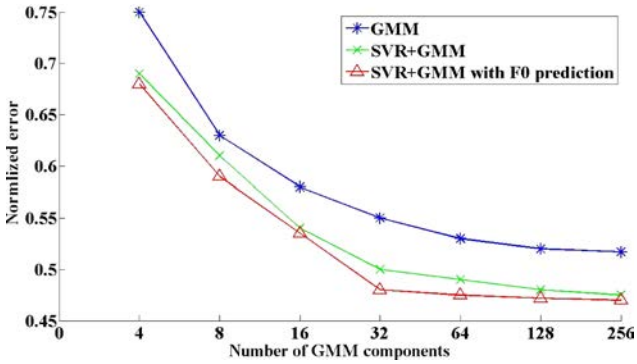

Fig. 6. Normalized error (F-F).

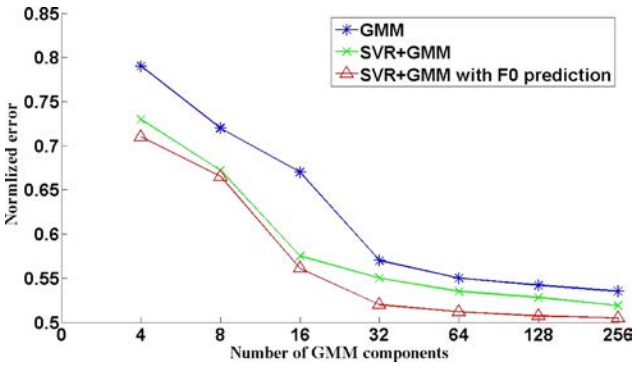

Fig. 7. Normalized error (F-M).

4.2. Subjective evaluation

Subjective tests were carried out to evaluate the identity and quality of the converted speech. 12 experienced listeners participated in all the tests. An ABX test was carried out to evaluate the similarity between the converted and target speech, in which three utterances were shown to the listeners, who were asked to judge whether A (source speech) or B (target speech) was closer to X (converted speech). Experimental results are shown in the Table 1. It can be found that the proposed hybrid SVR and GMM method achieves the best performance with an 81.09% average correct response. Adopting the F0 prediction instead of the GMM based F0 conversion can enhance the VC performance with an about 0.18% improvement of average correct response.

Table 1. Results of the ABX test.

Methods	Correct response (%)				
	M-M	M-F	F-F	F-M	Average
GMM	80.23	81.27	80.68	79.04	80.31
GMM+SVR	80.72	81.69	81.23	80.01	80.91
GMM+SVR with F0 prediction	81.06	81.87	81.28	80.13	81.09

A MOS (mean opinion score) experiment was also conducted to evaluate the overall performance of the converted speech. Each pair of the utterances including the converted speech using different VC methods and the target speech were shown to the listeners, who were asked to rate the similarity using a 10-point score from 0 for "totally different" to 9 for "identical". The pairs of speech were grouped with source speech, target speech, the converted speech using GMM, the converted speech using hybrid SVR and GMM, and the converted speech using hybrid SVR and GMM with F0 prediction. Different utterances were chosen to make these pairs, so the listeners could judge the speaker individuality instead of the sentence level similarity.

Table 2 compares various VC methods with a criterion of MOS and its standard deviation (SD), and the confidence interval of MOS is 95%. The results clearly reveal the efficiency of the proposed method, the score of the hybrid SVR and GMM method greatly outperforms that of the traditional GMM method, while the F0 prediction method can enhance the quality of the converted speech.

Table 2. Results of the MOS test.

	GMM	GMM+SVR	GMM+SVR with F0 prediction
MOS	5.68	6.03	6.12
SD	0.65	0.59	0.62

Complementary experiments were also performed to evaluate the performance of VC using speaker-specific information. Two mentioned subjective tests including ABX and MOS were adopted and the training utterances were all from the target speaker. Table 3 depicts the average results of the ABX and MOS tests, respectively. As seen from the table, with the increasing number of the training data, the probability of correct responses and the opinion scores show an upward trend. Although the quality and identity of the converted speech are not too satisfactory, the results indicate that VC using the target speaker-specific information is feasible and can be used in the many-to-one or even in the cross-language VC.

Table 3. Subjective tests of VC using speaker-specific information.

Number of training sentences	ABX test	MOS test	
	Correct response [%]	MOS	SD
1	50.23	1.53	0.89
5	52.98	2.03	0.87
10	55.16	2.98	0.79
20	56.92	3.62	0.81
50	58.28	4.18	0.78
100	59.13	4.36	0.75

5. Conclusions

A novel VC method based on hybrid SVR and GMM is proposed in the paper, which shows better performance than the GMM baseline method. A new F0 conversion approach is also presented to enhance the VC performance, finally the VC based on non-parallel data which needs only the specific information of the target speaker has been also investigated. The objective and subjective experimental results confirm the efficiency of the proposed methods, but an ideal VC should also take account of other aspects, such as durations, speaking rates, and speaking styles. Further research will focus on the conversion of these features.

Acknowledgment

The authors acknowledge the support of this work by the Natural Science Foundation of China (Grant Nos. 60872073, 51075068 and 60975017), the Open Research Foundation of Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education of Southeast University (Grant No. UASP1003), and the Autonomous Fund of Science and Technology on Acoustic Antagonizing Laboratory in 2009 (Grant No. 09ZD.2).

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. En–Najjary T., Rosec O., Chonavel T. (2003), *A new method for pitch prediction from spectral envelope and its application in voice conversion*, Proceedings of Eurospeech 2003, pp. 1753–1756, Geneva.

5. Erro D., Moreno A. (2007), *Frame Alignment Method for Cross-lingual Voice Conversion*, Proceedings of Interspeech 2007, pp. 1969–1972, Antwerp.

6. Inanoglu Z. (2003), *Transforming pitch in a voice conversion framework*, Master Thesis, St. Edmund's College, University of Cambridge.

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. Kominek J., Black A.W. (2004), *The CMU Arctic speech databases*, Proceedings of the 5th ISCA Speech Synthesis Workshop, pp. 223–224, Pittsburgh.

9. Kawahara H., Masuda–Katsuse T., Cheveigne A. (1999), *Restructuring speech representation using pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of*

a repetitive structure in sounds, Speech Communication, **27**, 3, 187–207.

10. Misra H., Ikbal S., Yegnanarayana B. (2003), *Speaker-specific mapping for text-independent speaker recognition*, Speech Communication, **39**, 3–4, 301–310.

11. Mizuno H., Abe M. (2005), *Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt*, Speech Communication, **16**, 2, 153–164.

12. Mouchtaris A., Spiegel J.V., Mueller P. (2004), *Non-parallel training for voice conversion by maximum likelihood constrained adaptation*, Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1–4, Montreal.

13. Perez–Cruz F., Camps–Valls G., Soria–Olivas E., Perez–Ruixo J.J., Figueiras–Vidal A.R., Artes–Rodriguez A. (2002), *Multi-dimensional function approximation and regression estimation*, Proceedings of the International Conference on Artificial Neural Networks, pp. 757–762, Madrid.

14. Perez–Cruz F., Navia–Vazquez A., Alarcon–Diana P., Artes–Rodriguez A. (2000), *An IRWLS procedure for SVR*, Proceedings of the 10th European Signal Processing Conference, pp. 725–728, Tampere.

15. Shao X., Milner B. (2004), *Pitch prediction from MFCC vectors for speech reconstruction*, Proceedings of the 2004 International Conference on Acoustics, Speech, and Signal Processing, pp. 97–100, Montreal.

16. Smits G.F., Jordan E.M. (2002), *Improved SVM regression using mixtures of kernels*, Proceeding of the 2002 International Joint Conference on Neural Networks, pp. 2785–2790, Honolulu.

17. Stylianou Y., Cappe O., Moulines E. (1998), *Continuous probabilistic transform for voice conversion*, IEEE Transactions Speech and Audio Processing, **6**, 2, 131–142.

18. 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.

19. 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.

20. Ye H., Young S. (2006), *Quality-enhanced voice morphing using maximum likelihood transformations*, IEEE Transactions on Audio, Speech and Language Processing, **14**, 4, 1301–1312.