System Fusion for High-Performance Voice Conversion

Xiaohai Tian[1,2], Zhizheng Wu[3], Siu Wa Lee[4], Nguyen Quy Hy[1,2], Minghui Dong[4], and Eng Siong Chng[1,2]

[1]School of Computer Engineering, Nanyang Technological University (NTU),Singapore
[2]Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly, NTU, Singapore
[3]Center for Speech Technology Research, University of Edinburgh, United Kingdom
[4]Human Language Technology Department, Institute for Infocomm Research, Singapore

Abstract

Recently, a number of voice conversion methods have been developed. These methods attempt to improve conversion performance by using diverse mapping techniques in various acoustic domains, e.g. high-resolution spectra and low-resolution Mel-cepstral coefficients. Each individual method has its own pros and cons. In this paper, we introduce a system fusion framework, which leverages and synergizes the merits of these state-of-the-art and even potential future conversion methods. For instance, methods delivering high speech quality are fused with methods capturing speaker characteristics, bringing another level of performance gain. To examine the feasibility of the proposed framework, we select two state-of-the-art methods, Gaussian mixture model and frequency warping based systems, as a case study. Experimental results reveal that the fusion system outperforms each individual method in both objective and subjective evaluation, and demonstrate the effectiveness of the proposed fusion framework.

Index Terms: Voice conversion, system fusion, high-performance, frequency warping, GMM

1. Introduction

Voice conversion (VC) is a technology to modify the speech uttered by a source speaker to make it as if it was spoken by another speaker (target) without changing the language content. Typically, VC can operate with three different types of feature, i.e. spectrum, prosody and duration. As compared to the prosodic and the duration, the spectrum feature can more significantly affect the conversion quality as it contains a greater amount of speaker identity information. Hence, learning a robust spectral mapping in the spectrum domain is an essential topic in VC.

To achieve this goal, several types of VC approaches have been proposed. Statistical parametric voice conversion is one of the effective techniques, which offers both linear and nonlinear feature mapping. To construct a linear mapping, Gaussian mixture model (GMM)-based approach [1, 2] and partial least squares regression [3] are proposed. Alternatively, the nonlinear methods, such as neural network [4, 5, 6] and kernel partial least squares regression [7] have also been proposed. These approaches are usually applied to low-dimensional features, which model the shape of spectral envelop. However, the converted speech was degraded due to over-smoothing. To address this problem, global variance (GV) enhancement was proposed in [8, 9], which improves the converted speech quality significantly.

The exemplar-based voice conversion is a non-parametric approach which directly uses the target speech exemplars to synthesize the converted speech [10, 11, 12]. As high-resolution spectra are usually employed as the basis exemplars, exemplar-based methods can maintain more spectral details and achieve better speaker similarity. However, as this approach operates in spectrum domain, the spectral variation at the temporal domain might not be effective enhanced.

Unlike statistical parametric and exemplar-based methods, frequency warping (FW) based voice conversion shifts the frequency axis of the source spectra to match that of the target. Several frequency warping based approaches have been proposed in the literature, such as vocal tract length normalization (VTLN) [13, 14], weighted frequency warping (WFW) [15], bilinear frequency warping (BLFW) [16] and correlation-based frequency warping (CFW) [17]. High naturalness of this kind of methods has been reported in these studies. As frequency warping itself only shift the frequency axis and cannot match the slope of the target spectrum, residual compensation [18] also called amplitude scaling in [19] will be useful to improve the speaker similarity performance.

As we discussed above, each voice conversion method has its own pros and cons. One voice conversion system might be able to address the problems that arise in other voice conversion systems. Inspired by the system combination ideas in speech recognition [20], speaker recognition [21] and speech synthesis [22], we propose a system fusion framework to combine different types of VC systems. As High-resolution feature maintains the spectral details, spectrum is preferred in this framework. In this paper we consider fusing two types of VC system, namely Gaussian mixture model (GMM) and frequency warping (FW) based systems, for a case study. The reason to choose the two systems is that GMM-based systems can capture the general shape of spectral envelope, while frequency warping systems are good at preserving spectral details for higher naturalness performance. However, in a more general case, different types of all possible systems can be combined.

2. State-of-the-art voice conversion approaches

The objective of most voice conversion systems is to learn the transformation functions from the source to the target based on a set of aligned feature vector pairs. In conversion phrase, a conversion function maps the source feature vector $\mathbf{x}_k$ into the target feature vector $\hat{\mathbf{y}}_k$ for k-th frame, expressed as:

$$\hat{\mathbf{y}}_k = \mathcal{F}(\mathbf{x}_k). \quad (1)$$

The conversion function $\mathcal{F}(\cdot)$ is optimized by minimizing the prediction error between converted frame $\hat{\mathbf{y}}_k$ and target frame $\mathbf{y}_k$.

In this section, we review two types of state-of-the-art voice conversion approaches.

2.1. Statistical parametric based method

The statistical approach applies statistical models to estimate the mapping relationship between the spectral features of the source and target speakers. During training phrase, the transformation, $\mathcal{F}(\cdot)$, is defined by a set of parameters, which are found with the criterion of minimizing the difference or maximizing the joint likelihood of the converted and target features. During runtime conversion, the source spectral features are converted by Eq. (1).

In practice, $\mathcal{F}(\cdot)$ can be either linear transform, such as GMM [1, 2] and partial least squares regression [3], or nonlinear transform, such as neural network [4, 5, 6] and kernel partial least squares regression [7]. Low-resolution feature, e.g. Mel-cepstral coefficients (MCCs), is usually used in these methods, which can be used to construct mapping functions that convert speaker identity successfully. However, the spectral details are eliminated due to the low feature dimension. This degrades the quality of converted speech.

To improve the converted speech quality of GMM-based voice conversion, the global variance (GV) was proposed in [8]. The statistics of the GV, trained from the speech of target speaker, are used for post-filter the spectral features generated by above methods. As the variance of converted features tend to be smaller than that of target speech, the speech quality will be improved by this GV compensation.

2.2. Frequency warping based method

Frequency warping (FW) is an alternative voice conversion approach, which moves the frequency axis of source spectra to that of the target. Given a source spectral envelope $\mathbf{x}_k^{(\mathrm{DFT})}$ and its warping function $w_k(f)$, the Eq. (1) could be written as:

$$\hat{\mathbf{y}}_k^{(\mathrm{DFT})} = \mathcal{F}(\mathbf{x}_k^{(\mathrm{DFT})}) = \mathbf{x}_k^{(\mathrm{DFT})}(w_k^{-1}(f)). \qquad (2)$$

$w_k(f)$ can be found by either minimizing the spectral distance between $\hat{\mathbf{y}}_k^{(\mathrm{DFT})}$ and $\mathbf{y}_k^{(\mathrm{DFT})}$ [23, 15] or maximizing the correlation between them [17].

Similar to GMM-based methods [2] and exemplar-based methods [12], FW relies on a subset of aligned training spectral pairs, so as to estimate the warping function. Hence, FW can be easily combined with the above two type of methods, as reported in [15] and [18], respectively.

FW-based approach operates directly on the high-resolution spectral feature, which does not remove the details of source spectra and hence leads to good naturalness in the converted speech. Moreover, the residual compensation (or amplitude scaling) function [19, 18] is also used to further enhance the speech quality.

3. Proposed system fusion

3.1. Framework for system fusion

Studies shown that existing approaches often achieve either good similarity voices or high quality speech. Now a system fusion framework is proposed in the following to leverage any state-of-the-art voice conversion methods, and even the methods invented in the future.

Given a set of source spectral features $\mathbf{X}$, it is first transformed by candidate VC methods to obtain the converted features $\hat{\mathbf{Y}}$. Theoretically, $\hat{\mathbf{Y}}_l$ of l-th VC system could be any spectral feature, such as MCCs and spectrum. As different features will be used in candidate VC methods, $\hat{\mathbf{Y}}_l$ should be transformed to the same feature type for fusion. High-resolution feature maintains the spectral details, hence spectrum is preferred in this framework.

Finally, the fused spectrogram can be obtained as:

$$\hat{\mathbf{Y}}^{(\mathrm{DFT})} := \sum_{l=1}^{L} \alpha_l \cdot \hat{\mathbf{Y}}_l^{(\mathrm{DFT})}, \qquad \sum_{l=1}^{L} \alpha_l = 1, \qquad (3)$$

where, $\hat{\mathbf{Y}}_l^{(\mathrm{DFT})}$ is the converted spectrogram of l-th VC system. The fusion ratio $\boldsymbol{\alpha} = [\alpha_1, \ldots \alpha_l, \ldots, \alpha_L]$ could be obtained by minimizing the error on training or development data as following,

$$\boldsymbol{\alpha} = \arg \min_{s.t.\ \sum \alpha_l = 1} d(\mathbf{Y}^{(\mathrm{DFT})}, \hat{\mathbf{Y}}^{(\mathrm{DFT})}), \qquad (4)$$

where, $d(\cdot)$ is the spectral distortion.

3.2. GMM-based and FW-based system fusion

Recall that, GMM-based approach is good at capturing the general shape of spectral envelope, while FW-based approach generates high quality speech [15, 18]. In this work, we apply the fusion to these two approaches as an example to demonstrate the merits of the fusion framework. Three state-of-the-art methods are chosen as the candidate systems, including JD-GMM [2] and GV enhancement [8] as the GMM-based approaches, and sparse representation based FW [18] as the FW-based approaches.

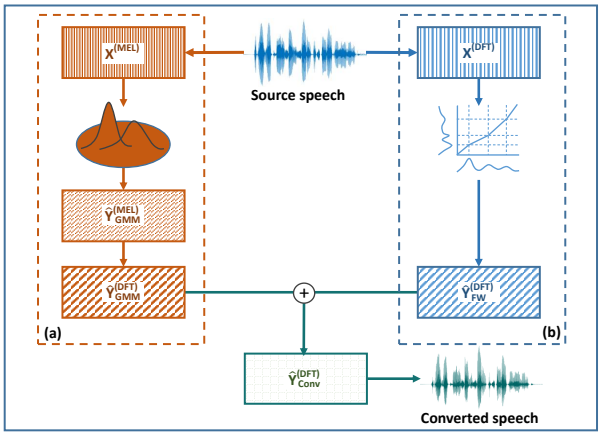

Figure 1: Block diagram of voice conversion system fusion. (a) is the conversion process of GMM-based VC system, (b) is the conversion process of FW-based VC system.

As different features will be used in FW-based and GMM-based approaches, spectrum and MCCs features will be extracted. The aligned source and target frames are obtained by applying dynamic time warping (DTW) to the MCCs feature sequence. The aligned MCCs and spectrum are used for the model training of GMM-based VC approaches and dictionary construction of FW-based VC approaches respectively. As only voiced frames will be transformed in FW-based method, while the unvoiced frames are not modified, the aligned spectra contain voiced frames only.

The proposed framework, as shown in Figure 1, contains following steps:

a) Extract the MCCs, $\mathbf{X}^{(\mathrm{Mel})}$, and spectrogram, $\mathbf{X}^{(\mathrm{DFT})}$, features of source speech.

b) Each frame of $\mathbf{X}^{(\mathrm{Mel})}$ and $\mathbf{X}^{(\mathrm{DFT})}$ will be converted by Eq. (1), GMM-based method, and Eq. (2), FW-based method, respectively.

c) The converted MCCs, $\hat{\mathbf{Y}}^{(\mathrm{Mel})}_{\mathrm{GMM}}$, of GMM-based system will be transformed to spectrogram, $\hat{\mathbf{Y}}^{(\mathrm{DFT})}_{\mathrm{GMM}}$.

d) Then the system fusion will be applied to the converted spectrogram of voiced frames from two methods, $\hat{\mathbf{Y}}^{(\mathrm{DFT})}_{\mathrm{GMM}}$ and $\hat{\mathbf{Y}}^{(\mathrm{DFT})}_{\mathrm{FW}}$. Eq. (3) could be written as:

$$\hat{\mathbf{Y}}^{(\mathrm{DFT})}_{\mathrm{Conv}} = \alpha \cdot \hat{\mathbf{Y}}^{(\mathrm{DFT})}_{\mathrm{GMM}} + (1-\alpha) \cdot \hat{\mathbf{Y}}^{(\mathrm{DFT})}_{\mathrm{FW}}, \quad (5)$$

Based on human perception, the system is fused in a bandwise manner. We uniformly divide the frequency range into a number of frequency bands in bark scale [24].

In each critical band, the converted spectrograms from the two systems will be merged by linear combination. As the speech signals are sampled as 16kHz, the first 21 bark bands, up to 7700 Hz, are used in this work. The fusion ratio of each frequency band will be set by grid search on development data to minimize the spectral distortion.

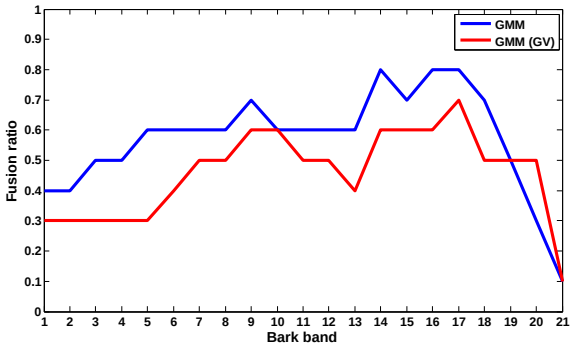

Figure 2: The fusion ratio of GMM and GMM(GV) for each bark band.

As shown in Figure 2, both the fusion ratio of GMM and GMM(GV) changes over bark bands, which indicates the performances of individual VC methods vary over frequency. Our preliminary experimental results showed that when using a single fusion ratio for all frequency bins, the fusion system does not outperform the best candidate system and the spectral distortion is higher than the best candidate system. Fusing system in a bandwise manner results in a spectral distortion even lower than any of the candidate systems. Note that, this fusion is only applied to voiced frames, while unvoiced frames are copied from GMM-system directly.

4. Experimental evaluations

4.1. Experimental setup

The VOICES database [25] was used to assess the proposed method. Four speakers were selected: two male speakers, *jal* and *jcs*, and two female speakers, *leb* and *sas*. Inter-gender and intra-gender conversions were conducted between following pairs: *jal* to *jcs* (M2M), *jal* to *sas* (M2F), *leb* to *jcs* (F2M) and *leb* to *sas* (F2F). 20 parallel utterances of each speaker were used as training data, another non-overlapping 20 utterances for evaluation and the rest 10 utterances for development data.

The speech signals were downsampled to 16 kHz. STRAIGHT [26] was used to extract 513-dimensional spectrum, aperiodicity coefficients and $\log F_0$. 25-dimensional MCCs and 15-dimensional linear spectrum frequencies (LSFs) were also calculated from the spectrum. In all the conversion methods, the same frame alignment was used.

- **GMM** (baseline): The JD-GMM with maximum likelihood parameter generation method as proposed in [2]. The number of Gaussian mixtures was set to 64.
- **GMM(GV)** (baseline): We use the same setting as GMM, and the converted MCC features were revised by GV enhancement as proposed in [27].
- **FW** (baseline): The sparse representation based CFW [18] with residual compensation. We use the same setting as [18].
- **FW+GMM (proposed)**: Fusion of the FW and GMM methods, mentioned in Section 3.2.
- **FW+GMM(GV) (proposed)**: Fusion of the FW and GMM(GV) methods, mentioned in Section 3.2.

In all the conversion methods, aperiodicity coefficients were not converted, while F_0 was converted by a global linear transformation in log-scale.

4.2. Objective evaluation

We conducted objective evaluation to assess the proposed method. The log spectral distortion (LSD) [28] was employed. The distortion of k-th order of log spectrum is calculated as:

$$d(\mathrm{x}^{(\mathrm{DFT})}_k, \mathrm{y}^{(\mathrm{DFT})}_k) = \sum_{i=1}^{M} (\log \mathrm{x}^{(\mathrm{DFT})}_{k,i} - \log \mathrm{y}^{(\mathrm{DFT})}_{k,i})^2, \quad (6)$$

where, M is the total number of the frequency bins. A distortion ratio between converted-to-target distortion and the source-to-target distortion could be defined as:

$$\mathrm{LSD} = \frac{\sum_{k=1}^{K} d(\hat{\mathrm{y}}^{(\mathrm{DFT})}_k, \mathrm{y}^{(\mathrm{DFT})}_k)}{\sum_{k=1}^{K} d(\mathrm{x}^{(\mathrm{DFT})}_k, \mathrm{y}^{(\mathrm{DFT})}_k)} \times 100\%, \quad (7)$$

where, $\mathrm{x}^{(\mathrm{DFT})}_k$ and $\mathrm{y}^{(\mathrm{DFT})}_k$ denote the source and target spectra respectively. $\hat{\mathrm{y}}^{(\mathrm{DFT})}_k$ is the converted spectrum. The average LSD result over all evaluation pairs was reported. A lower LSD value indicates smaller distortion.

Table 1: *Comparison of log spectral distortion (LSD) ratio of different conversion methods.*

Conversion Method	Voiced frames (%)	All frames (%)
GMM	76.0	82.3
GMM(GV)	75.8	83.1
FW	62.3	77.0
FW+GMM	**59.8**	**72.5**
FW+GMM(GV)	60.0	73.5

Table 1 presents the LSD results for the baseline methods and our proposed methods. In FW method, as the unvoiced frames are not involved in the conversion procedure, the LSD of all frames are calculated with converted voiced frames and original unvoiced frames.

We first analyse the LSD of different methods on voiced frames. We observe that two GMM-based methods, GMM and

GMM(GV), got similar LSD on voiced frames, that is 76.0% to 75.8%. Comparing with two GMM-based methods, FW achieves a lower LSD (62.3%), which is around 13% lower than GMM-based methods. It confirms the effectiveness of the FW, and is consistent with our previous finding in [18].

In comparison with GMM, FW+GMM achieves a much lower LSD, that is from 76.0% to 59.8%. Improvement is also observed by comparison FW with FW+GMM, the LSD drops from 62.3% to 59.8%. This indicates the two VC methods complement each other. Similarly complementary effect is found by combining FW and GMM(GV). Comparing to GMM(GV) and FW, the LSD of FW+GMM(GV) drops 15.8% and 2.3% respectively. It confirms the effectiveness of the proposed system combination framework.

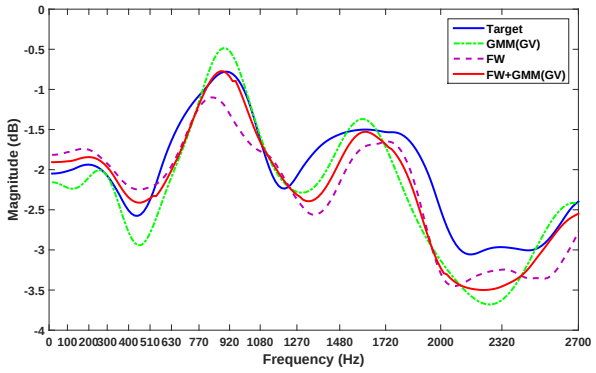

Figure 3: The converted spectral envelopes of GMM(GV), FW and fusion system.

Figure 3 shows an example of converted spectral envelope from GMM(GV), FW and fusion system. Comparing to GMM(GV) and FW, the spectral envelope converted by FW+GMM(GV) is the nearest to the target.

We now examine the LSD of different methods for all frames. Comparing to GMM-based methods, the LSD of proposed methods on all frames are consistent with the results on voiced frames only. This is because that, in FW+GMM and FW+GMM(GV), the unvoiced frames are copied from the results of GMM-based methods directly and the change comes from the voiced part only.

In comparison with FW, the LSD of FW+GMM and FW+GMM(GV) drop 4.5% and 3.5% respectively. These gaps are larger than that of voiced frames, which are 2.5% and 2.3%.

Note that, the FW+GMM and FW+GMM(GV) obtain very similar LSD. In the following, we will examine the performance in subjective listening test.

4.3. Subjective evaluation

We conducted listening tests to assess both speech quality and speaker similarity. 10 subjects participated in all the listening tests. As proved in [8], the converted speech of GMM(GV) outperform that of GMM. In the following, GMM(GV), FW, FW+GMM and FW+GMM(GV), are chosen for this evaluation.

We first performed AB preference tests to assess speech quality. 20 pairs were randomly selected from the 80 paired samples. In each pair, A and B were the samples from the proposed method and one of the baseline methods, respectively, in a random order. Each listener was asked to listen to both samples and then decide which sample is better in term of quality.

We then conducted an XAB test to assess the speaker similarity. In the test, similarly to the AB preference test, 20 pairs were randomly selected from the 80 paired samples. In each pair, X was the reference target sample, A and B were the converted samples of comparison methods listed in the first column of Table 2, in a random order. We note that X, A and B have the same language content. The listeners were asked to listen to the sample X first, then A and B, and then decide which sample is closer to the reference target sample.

Table 2: *Results of average quality and similarity preference tests with 95% confidence intervals for different methods.*

Conversion method	Preference score(%)	
	Quality test	Similarity test
FW+GMM	26 ($\pm$ 10.81)	29 ($\pm$ 7.69)
FW+GMM(GV)	74 ($\pm$ 10.81)	71 ($\pm$ 7.69)
GMM(GV)	32 ($\pm$ 8.34)	33 ($\pm$ 5.22)
FW+GMM(GV)	68 ($\pm$ 8.34)	67 ($\pm$ 5.22)
FW	46 ($\pm$ 8.29)	43 ($\pm$ 5.4)
FW+GMM(GV)	54 ($\pm$ 8.29)	57 ($\pm$ 5.4)

The subjective results are presented in Table 2. First, we evaluate the two proposed approaches, FW+GMM and FW+GMM(GV). It is clearly shown, in both quality and similarity tests, FW+GMM(GV) approach achieves much higher preference score than FW+GMM method.

We take two set of evaluations, comparing GMM(GV) to FW+GMM(GV), and FW to FW+GMM(GV), to examine the performance of the fused system and each separate system. In the comparison between GMM(GV) and FW+GMM(GV), FW+GMM(GV) achieves significant improvement to GMM(GV) in both quality and similarity. While comparing to FW, FW+GMM(GV) achieves noticeable improvement in speaker identity, and comparable speech quality. The above results confirm the effectiveness of the proposed method, and are consistent with the log spectral distortion results in Section 4.2. They are also consistent with the previous results reported in [18]. [1]

5. Conclusions

This paper proposed a framework to fuse the GMM-based and FW-based voice conversion methods. By tuning the band-wise fusion ratio, the fused system leverages each single method and improve conversion performance in various aspects, e.g. quality and similarity. The objective results indicate that, proposed method achieves lower log spectral distortion ratio. The subjective results show that, comparing to GMM(GV) method, proposed method achieves higher score in both quality and similarity. Moreover, comparing to FW, the proposed method improve the speaker similarity and preserve the speech quality.

6. Acknowledgements

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

[1] Converted samples are available via: http://www.listeningtests.net/voiceconversion/xhtian2015interspeech.

7. References

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

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 285–288.

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

[4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3893–3896.

[5] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE Transactions on Speech and Audio Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[6] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (SE) minimization training of neural network for voice conversion," in *INTERSPEECH*, 2014.

[7] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[8] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[9] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *INTERSPEECH*, 2011, pp. 669–672.

[10] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *8th ISCA Speech Synthesis Workshop*, 2013.

[11] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Spoken Language Technology workshop (SLT)*, 2012, pp. 313–317.

[12] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[13] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2003, pp. 556–559.

[14] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 676–681.

[15] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[16] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[17] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 211–215.

[18] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. S. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) to appear*, 2015.

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

[20] M. J. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[21] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. Van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[22] H. Zen, M. J. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 794–805, 2012.

[23] H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2, pp. 175–187, 1992.

[24] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[25] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, Rockford College, 2001.

[26] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[27] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion." in *INTERSPEECH*, 2012.

[28] H. Ye and S. Young, "High quality voice morphing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2004, pp. 1–9.