SPARSE REPRESENTATION FOR FREQUENCY WARPING BASED VOICE CONVERSION

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

[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

This paper presents a sparse representation framework for weighted frequency warping based voice conversion. In this method, a frame-dependent warping function and the corresponding spectral residual vector are first calculated for each source-target spectrum pair. At runtime conversion, a source spectrum is factorised as a linear combination of a set of source spectra in the training data. The linear combination weight matrix, which is constrained to be sparse, is used to interpolate the frame-dependent warping functions and spectral residual vectors. In this way, the proposed method not only avoids the statistical averaging caused by GMM but also preserves the high-resolution spectral details for high-quality converted speech. Experiments are conducted on the VOICES database. Both objective and subjective results confirmed the effectiveness of the proposed method. In particular, the spectral distortion dropped from 5.55 dB of the conventional frequency warping approach to 5.0 dB of the proposed method. Compare to the state-of-the-art GMM-based conversion with global variance (GV) enhancement, our method achieved 68.5 % in an AB preference test.

Index Terms— Voice conversion, frequency warping, sparse representation, exemplar, residual compensation

1. INTRODUCTION

Voice conversion (VC) is a technique to transform the speech of one speaker (source) so that it sounds like it was uttered by another speaker (target) without changing the language context. The challenge is how to modify or transform the source speech parameters to match the target parameters while maintaining high speech quality. A number of statistical parametric approaches have been proposed, such as linear transformation implemented by Gaussian mixture model [1, 2, 3, 4] and partial least squares regression [5]; nonlinear transformation through neural network [6, 7, 8] and kernel partial least squares regression [9].

However, one of the major issues on the statistical parametric approaches is their attempts to minimise the difference between the converted and target features, or to maximise the joint likelihood of the source and target features. These optimisation criteria often introduce statistical averaging, which leads to over-smoothing in the converted speech. To address this problem, a number of research works appear recently. In [4], global variance enhancement was proposed to model the dynamics of natural speech, which improved the converted speech quality significantly. Other efforts include the non-parametric exemplar-based voice conversion, which directly use speech exemplars to synthesize the converted speech [10, 11, 12]. Operating on high-resolution spectra, exemplar-based methods are able to keep more spectral details.

An alternative way to avoid over-smoothing is to perform frequency warping (FW) based voice conversion, which shifts the frequency axis of the source spectra to that of the target. As the warping process does not remove any spectral details, FW methods facilitate the quality of converted speech. Several frequency warping based approaches have been proposed in the literature, such as vocal tract length normalization (VTLN) [13, 14], bilinear frequency warping (BLFW) [15] and correlation-based frequency warping (CFW) [16]. One of the successful methods is the weighted frequency warping (WFW) [17] with amplitude scaling (AS) [18]. WFW implements smooth warping through soft clustering based on GMM. On top of WFW, AS further shifts the amplitude of the warped spectra to match the target counterpart. Nevertheless, both WFW and AS heavily rely on GMM-based clustering. The statistical averaging nature of GMM inevitably reduces the variation in the converted speech, limiting the speech quality.

In this work, we propose a novel framework, named *sparse representation for weighted frequency warping*. In this framework, we first compute a warping function for each source-target frame pair. Then the corresponding amplitude difference between the warped and reference spectra, named *spectral residual*, is calculated. The warping functions and spectral residual vectors are treated as exemplars, similar to the spectrum exemplars in [12]. At runtime, each source spectrum is factorised as a linear combination of a set of source spectra in the training data. The linear combination weights, in the form of a sparse matrix, are used to interpolate the corresponding warping functions and spectral residual exemplars. The resultant weighted warping function is finally used to warp the source spectrum, then followed by the compensation by the weighted spectral residual vector. In practice, the linear combination weights are estimated by nonnegative matrix factorisation (NMF) technique similar to that in [12]. Note that this bypasses the computation of posterior probability associated to each Gaussian component in GMM-based approaches.

There are three advantages of our proposed method over the GMM-based weighted frequency warping:

a) High-resolution spectrum is directly used without any dimension reduction, for the estimation of linear combination weights, warping functions and spectral residuals.

b) Due to the sparsity constraint, only a small set of warping functions and spectral residual vectors are used to generate the target spectrum. Thus, the over-smoothing problem is avoided and converted speech will become lively.

This research is supported in part by Interactive and Digital Media Programme Office (IDMPO), National Research Foundation (NRF) hosted at Media Development Authority (MDA) of Singapore (Grant No.: MDA/IDM/2012/8/8-2 VOL 01).

c) It offers more flexibility than the GMM-based framework. For example, wide acoustic context is applicable to compute the weighted warping function and spectral residual.

2. PROBLEMS IN CONVENTIONAL WEIGHTED FREQUENCY WARPING

Weighted frequency warping (WFW) [17] is a popular implementation to produce smooth warping functions and avoid discontinuity across frames [19]. During WFW training, given N pairs of source $\mathbf{X}$ and target $\mathbf{Y}$ features, $\mathbf{Z} = [\mathbf{X}; \mathbf{Y}]$, where $\mathbf{z}_n = [\mathbf{x}_n; \mathbf{y}_n]$ denotes the n^{th} joint vector forming $\mathbf{Z}$. GMM with K mixtures is first employed to model the joint density. After that, a warping function $w_k(f)$ is computed between the source and target of the joint mean vector, $\mu_k^{(\text{Z})} = [\mu_k^{(\text{X})}; \mu_k^{(\text{Y})}]$ which is computed as

$$\mu_k^{(\text{Z})} = \sum_{n=1}^{N} \mathbf{z}_n \cdot \gamma_{n,k}, \tag{1}$$

where $\gamma_{n,k}$ is the occupation probability of the n^{th} vector belonging to the k^{th} Gaussian component.

In the conversion phrase, for each observation frame $\mathbf{x}$, $\mathbf{x}^{(\text{DFT})}$ denotes its spectrum, the warping function is calculated by interpolating pre-computed warping functions $\{w_1(f), \dots w_K(f)\}$ as:

$$w(\mathbf{x}, f) = \sum_{k=1}^{k} p_k(\mathbf{x}) \cdot w_k(f), \tag{2}$$

where $p_k(\mathbf{x})$ is the posterior probability of $\mathbf{x}$ belonging the k^{th} Gaussian component. Hence, the converted spectrum $\mathbf{y}^{(\text{DFT})'}$ is obtained by applying the warping function on the source spectrum $\mathbf{x}^{(\text{DFT})}$ as

$$\mathbf{y}^{(\text{DFT})'} = \mathbf{x}^{(\text{DFT})}(w^{-1}(\mathbf{x}, f)). \tag{3}$$

To further improve the performance, an amplitude scaling technique, proposed in [18], was used to compensate the amplitude difference between the warped and target spectra. As the difference is usually computed in log-amplitude scale, and similar to residual compensation (RC) in exemplar-based voice conversion [12], we denote it as RC throughout this paper. Given the frame $\mathbf{x}_n$, the residual spectrum $\mathbf{r}_n$ between the warped $\mathbf{y}_n^{(\text{DFT})'}$ and reference $\mathbf{y}_n^{(\text{DFT})}$ spectrum can be computed as

$$\mathbf{r}_n = \log \mathbf{y}_n^{(\text{DFT})} - \log \mathbf{x}_n^{(\text{DFT})}(w^{-1}(\mathbf{x}_n, f)). \tag{4}$$

Similar to Eq. (1) and (2), the residual compensation vector for each observation frame $\mathbf{x}$ could be computed as

$$\mathbf{r}' = \sum_{k=1}^{K} p_k(\mathbf{x}) \cdot \mu_k^{(\text{R})} = \sum_{k=1}^{K} p_k(\mathbf{x}) \cdot \sum_{n=1}^{N} \mathbf{r}_n \cdot \gamma_{n,k}, \tag{5}$$

where $\mu_k^{(\text{R})}$ and $p_k(\mathbf{x})$ are the spectral residual vector and corresponding posterior probability of k^{th} Gaussian component respectively.

The final converted spectrum is written as

$$\begin{aligned} \log \mathbf{y}^{(\text{Conv})} &= \log \mathbf{y}^{(\text{DFT})'} + \mathbf{r}' \\ &= \log \mathbf{x}^{(\text{DFT})}(\sum_{k=1}^{k} p_k(\mathbf{x}) \cdot w_k(f))^{-1} + \sum_{k=1}^{K} p_k(\mathbf{x}) \cdot \mu_k^{(\text{R})} \end{aligned} \tag{6}$$

Although current weighted frequency warping with residual compensation works well and produces higher quality speech than statistical parametric voice conversion, there are limitations in current implementation as shown in Eq. (6): a) rely on low-dimensional features to compute occupation probabilities $\gamma_{n,k}$ and posteriori probability $p_k(\mathbf{x}_n)$; b) problematic occupation probabilities will distort the fine structures the mean vectors $\mu_k^{(\text{X})}$ and $\mu_k^{(\text{Y})}$, which lead to a inaccurate warping function; c) small number of mixtures cannot introduce variation in the converted speech. These problems are all rooted in GMM. This work, is hence, aimed for a high-quality voice conversion framework for frequency warping without the use of GMM.

3. PROPOSED SPARSE REPRESENTATION BASED FREQUENCY WARPING

Motivated by the success of exemplar-based voice conversion [10, 11, 12], where each target spectrum is generated as a linear combination of a set of spectrum exemplars, an sparse representation based weighted frequency warping is proposed in the following. Without relying on GMM and low-resolution features, we preserve the spectral details in both the warped spectrum and residual spectrum.

The proposed framework is presented in Fig. 1, which consists of three stages: a) dictionary construction; b) frequency warping and c) residual compensation. We will explain the details of each stage in this Section.

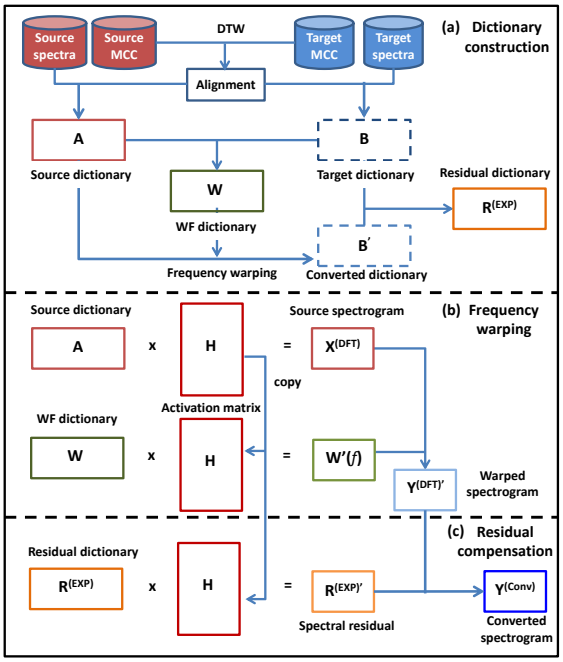

Fig. 1. Block digram of sparse representation based weighted frequency warping with residual compensation system. WF stands for warping function.

3.1. Dictionary construction

The dictionary construction stage is the only offline process in this framework. Alternative to the conventional GMM-based WFW or statistical parametric conversion approaches, there is no training stage in this framework, but dictionary construction instead.

Given the parallel training data, the source $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n, \cdots, \mathbf{a}_N]$ and target $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n, \cdots, \mathbf{b}_N]$ dictionaries can be obtained by dynamic time warping (DTW). We

note that $\mathbf{a}_n$ and $\mathbf{b}_n$, the exemplar version of $\mathbf{x}_n^{(\mathrm{DFT})}$ and $\mathbf{y}_n^{(\mathrm{DFT})}$, denote the spectra corresponding to $\mathbf{x}_n$, and $\mathbf{y}_n$, respectively. Note that, during the conversion phrase, only voiced frames will be transformed, thus the source and target dictionaries only contain voiced frames. For each paired spectrum $[\mathbf{a}_n; \mathbf{b}_n]$, a warping function $w_n(f)$ could be obtained. The warping function dictionary $\mathbf{W}$ can be express as

$$\mathbf{W} = [w_1, w_2, \cdots, w_n, \cdots, w_N] \tag{7}$$

Similarly, each source frame is associated with a spectral residual vector $\mathbf{r}_n$, which is computed from the warped spectrum and reference spectrum in log-scale, similar to Eq. (4). In order to use the matrix factorisation technique which is to be detailed in the next section, we exponentiate $\mathbf{r}_n$ as $\mathbf{r}_n^{(\mathrm{EXP})}$, denoted as

$$\mathbf{R}^{(\mathrm{EXP})} = [\mathbf{r}_1^{(\mathrm{EXP})}, \mathbf{r}_2^{(\mathrm{EXP})}, \cdots, \mathbf{r}_n^{(\mathrm{EXP})}, \cdots, \mathbf{r}_N^{(\mathrm{EXP})}] \tag{8}$$

Smoothed Warping Function: Because the spectrogram is smooth as long as source and target exemplars are continuous in time domain, the corresponding warping functions should be also smooth in such regions. However, the quality of warping function depends on other techniques as well, such as formant estimation. Thus, abrupt transition may appear even in continuous frames. In this work, a moving average filter is applied on the warping functions in $\mathbf{W}$ in continuous time domain (sentence by sentence) to prevent such discontinuity. The smoothing is applied before computing the spectral residual.

3.2. Frequency warping based conversion through exemplars

Three dictionaries $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{R}^{(\mathrm{EXP})}$ govern the runtime conversion process, that is the frequency warping conversion. As there is close correspondence between dictionaries, we assume that the same activation weights for interpolation are shared among the them.

At runtime, each source spectrum $\mathbf{x}^{(\mathrm{DFT})}$ can be factorised as linear combination of source spectrum exemplars in $\mathbf{A}$, express as

$$\mathbf{x}^{(\mathrm{DFT})} \approx \mathbf{A} \cdot \mathbf{h}, \tag{9}$$

where $\mathbf{h}$ is the activation vector, each element of which is a weight of the corresponding exemplar in source dictionary. As the constraint of the dictionaries $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{R}^{(\mathrm{EXP})}$ to be nonnegative, the activation vector can be estimated by the nonnegative matrix factorisation technique [20, 21] with criteria of minimising the following objective function

$$\mathbf{h} = \arg\min_{\mathbf{h}\geq 0} d(\mathbf{x}^{(\mathrm{DFT})}, \mathbf{A}\mathbf{h}) + \lambda \|\mathbf{h}\|_1, \tag{10}$$

where λ is the sparsity penalty factor, and $\|\mathbf{h}\|_1$ means L_1 norm on the activation vector. Technique details can be found in [21, 12].

The warping function of the source spectrum $\mathbf{x}^{(\mathrm{DFT})}$ can be obtained as

$$w^{'}(f) = \mathbf{W} \cdot \mathbf{h}, \tag{11}$$

where $\mathbf{W}$ is the warping function dictionary, $\mathbf{h}$ is the corresponding activation vector. After that, Eq. (3) can be applied directly to warp the source spectrum.

As each frame can be factorised and warped independently, the activation matrix $\mathbf{H}$ can be estimated directly by factorising the whole spectrogram $\mathbf{X}^{(\mathrm{DFT})}$, as illustrated in Fig. 1 (b). Similarly, the warping functions $\mathbf{W}^{'}(f)$ of the whole spectrogram can be calculated as

$$\mathbf{W}^{'}(f) = \mathbf{W} \cdot \mathbf{H}, \tag{12}$$

Then, the warped spectrogram $\mathbf{Y}^{(\mathrm{DFT})^{'}}$ could be obtained by applying the warping functions $\mathbf{W}^{'}(f)$ to the corresponding source spectrogram $\mathbf{X}^{(\mathrm{DFT})}$.

3.3. Residual compensation

Similar to the warping function calculation in Eq. (12), the same activation matrix $\mathbf{H}$ is used for the residual spectrogram calculation

$$\mathbf{R}^{(\mathrm{EXP})^{'}} = \mathbf{R}^{(\mathrm{EXP})} \cdot \mathbf{H}, \tag{13}$$

The residual spectrogram compensates the warped spectrogram $\mathbf{Y}^{(\mathrm{DFT})^{'}}$ to generate the final converted spectrogram $\mathbf{Y}^{(\mathrm{Conv})}$ in log-scale

$$\log \mathbf{Y}^{(\mathrm{Conv})} = \log \mathbf{Y}^{(\mathrm{DFT})^{'}} + \log \mathbf{R}^{(\mathrm{EXP})^{'}} \tag{14}$$

4. EVALUATION

We used the VOICES database [2] to assess the proposed method. Two male (*jal* and *jcs*) and two female (*leb* and *sas*) speakers were selected to conduct inter-gender and intra-gender conversions, including *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, while another non-overlapping 20 utterances for evaluation.

The speech signals were downsampled to 16 kHz. STRAIGHT [22] was used to extract 513-dimensional spectrum, 5 band aperiodicity measures and $\log F_0$. 25-dimensional Mel-Cepstral Coefficients (MCCs) and 15-dimensional linear spectrum frequencies (LSFs) were also used for the spectrum. In all the conversion methods, we used the same frame alignment, which was obtained by performing DTW on the MCC feature sequence.

In the experiments, we considered several state-of-the-art methods as our baselines, including WFW with residual compensation[1], and joint density GMM (JD-GMM) with global variance (GV) enhancement [4][2].

- **ML-GMM**: The JD-GMM with maximum likelihood parameter generation method as proposed in [4], was used as a reference baseline. In subjective test, post-filtering based GV enhancement as proposed in [23] was employed for better converted speech quality. MCC features were used to train the model, the optimal number of Gaussian mixtures was 64.
- **WAMF**: The classic weighted frequency warping (WFW) [17] with GMM-based residual compensation (or amplitude scaling) [18]. Automatic mapping of formants (AMF) method [17] was used to map the formants of source and target spectral pair with the constraint of minimum spectral distortion. Then, the warping function was defined by aligned formant pairs. LSFs feature was used for formant estimation, and the number of Gaussian components was set to 32, which is empirically found by the spectral distortion results.
- **WCFW**: The weighted correlation-based frequency warping [16] with GMM-based residual compensation. Spectral envelopes were used to find the warping function, based on formant segmentation. The segment boundary shift was constrained within 100 Hz.
- **NMF-AMF**: A variation of our **proposed** sparse representation based WFW with residual compensation, where the warping function is based on AMF.
- **NMF-CFW**: Another variation of our **proposed** method, where the warping function is based on our recently proposed CFW.

[1] As the difference is usually computed in log-amplitude scale, and similar to residual compensation (RC) in exemplar-based voice conversion [12], we denote it as RC rather than amplitude scaling, as in [18].

[2] Global variance (GV) enhancement was only applied in the listening test.

Only the performance with residual compensation will be reported for the frequency warping based methods. In order to achieve accurate activation weights as suggested in [12], in NMF-based approaches, NMF-AMF and NMF-CFW, a spectral compression factor was set to 0.4, which is based on our experimental results, and used in Eq. (9). In all the conversion methods, band aperiodicities (BAPs) were not converted, while F_0 was converted by a global linear transformation in log-scale.

4.1. Objective Evaluation

We conducted objective evaluation to assess the proposed method. The Mel-Cepstral Distortion (MCD) [4] was employed as the objective measure. The average MCD result over all evaluation pairs was reported. A lower MCD value indicates smaller distortion. As the frequency warping method was just applied on the voiced frame, only voiced frames were used in the MCD calculation.

Table 1. *Comparison of spectral distortions of different conversion methods.*

Conversion Method	Unsmoothed WF	Smoothed WF
WAMF	5.84 (dB)	N/A
WCFW	5.55 (dB)	N/A
NMF-AMF	5.85 (dB)	5.61 (dB)
NMF-CFW	5.14 (dB)	**5.0 (dB)**

Table 1 presents the MCD results for the frequency warping based methods involved in this work. In the GMM-based weighted frequency warping, as the single warping function was shared by each Gaussian, hence it is impossible to apply further smooth function. Comparing with WAMF, WCFW achieves a lower MCD, that is 5.55 dB over 5.84 dB of WAMF. It confirms the effectiveness of the CFW, and is consistent with our previous work [16].

Next, in comparison with WAMF, NMF-AMF achieves a similar MCD when the warping functions are not smoothed. However, a large improvement is observed then the warping functions are smoothed, that is from 5.85 dB to 5.61 dB. This implies the importance of smoothed warping functions to our sparse representation framework. Comparing with NMF-CFW and WCFW, the MCD drops from 5.55 dB of WCFW to 5.14 dB of NMF-CFW in the case of unsmoothed warping functions. Similarly, when the warping functions are smoothed, another 0.14 dB distortion reduction is observed. It confirms the effectiveness of the proposed sparse representation framework and the importance of smoothed warping functions, which can only be implemented in our framework.

Within the sparse representation framework, NMF-CFW has a 0.71 dB improvement in the case of unsmoothed warping functions over NMF-AMF, and a 0.61 dB improvement with smoothed warping functions. The improvement is much larger than that in the GMM-based framework. It confirms the effectiveness of the CFW, which is more efficient than AMF within the sparse representation framework.

Although the proposed method improves the MCD considerably in comparison to the other state-of-the-art frequency warping based approaches, the MCD of ML-GMM method achieves further 0.46 dB reduction with an MCD of 4.54 dB. However, objective results do not always correlate with subjective evaluation, especially between different categories of VC approaches, which was reported in previous work [15, 17, 18]. This phenomenon is also observed in our subjective evaluation, as detailed in next section.

4.2. Subjective Evaluation

We conducted listening tests to assess both speech quality and speaker similarity. 10 subjects participated in all the listening tests. The smoothed warping functions are used in both NMF-AMF and NMF-CFW methods. Here, GV enhancement [23] is employed in the ML-GMM method, which is proved slightly outperform than the NMF voice conversion method mentioned in [12].

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, similar to the AB preference test, 20 pairs were 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 and then A and B, after that, they should 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(%) (95% confidence interval)	
	Quality test	Similarity test
WAMF	23 ($\pm$ 4.43)	29 ($\pm$ 7.17)
WCFW	77 ($\pm$ 4.43)	71 ($\pm$ 7.17)
NMF-AMF	34 ($\pm$ 7.31)	33.5 ($\pm$ 7.47)
NMF-CFW	66 ($\pm$ 7.31)	66.5 ($\pm$ 7.47)
WCFW	21 ($\pm$ 7.56)	38.5 ($\pm$ 6.8)
NMF-CFW	79 ($\pm$ 7.56)	61.5 ($\pm$ 6.8)
ML-GMM (GV)	31.5 ($\pm$ 10.09)	46.4 ($\pm$ 10.75)
NMF-CFW	68.5 ($\pm$ 10.09)	53.6 ($\pm$ 10.75)

The subjective results are presented in Table 2. First, we compare the CFW and AMF approaches. It is clear that, in both quality and similarity tests, CFW approach achieves much higher preference score than AMF method in both frameworks. Then, we make a comparison between WCFW and NMF-CFW to examine the performance of sparse representation and GMM based framework. It is observed that NMF-CFW achieves significant improvement to WCFW in both quality and similarity. The above results confirm the effectiveness of the proposed method, and they were consistent with the spectral distortion results in Section 4.1.

Finally, the quality and similarity performance is compared between NMF-CFW and ML-GMM (GV), which is the state-of-the-art voice conversion method. *The results indicates that with comparable speaker similarity to the ML-GMM (GV) method, NMF-CFW achieves noticeable improvement in speech quality, and the improvement is significant*. This confirms the effectiveness of our proposed method. (Converted samples are available via: http://www.listeningtests.net/voiceconversion/xhtian2015icassp).

5. CONCLUSION

This paper proposed a sparse representation framework for frequency warping based voice conversion. By using exemplar-based framework, our proposed method not only preserves the speech quality but also bypasses the over-smoothing problem. The objective and subjective evaluation results indicate that, proposed method achieves lower spectral distortion and higher preference score in comparison with frequency warping methods. Moreover, compare to ML-GMM (GV) method, the proposed method produces significantly higher quality speech without decreasing the similarity.

With the flexibility of the sparse representation framework, we will include long-term contextual information for temporal constraint for frequency warping and residual compensation as a follow-up work.

6. REFERENCES

[1] Hadas Benisty and David Malah, "Voice conversion using GMM with enhanced global variance," in *Proc. INTERSPEECH*, 2011.

[2] Alexander Blouke Kain, *High resolution voice transformation*, Ph.D. thesis, OGI School of Science and Eng., Portland, Oregeon, USA, 2001.

[3] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] Tomoki Toda, Alan W Black, and Keiichi 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.

[5] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[6] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[7] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong 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.

[8] Fong-Long Xie, Yao Qian, Yuchen Fan, Frank K.Soong, and Haifeng Li, "Sequence error (SE) minimization training of neural network for voice conversion," in *Proc. INTERSPEECH*, 2014.

[9] Elina Helander, Hanna Silén, Tuomas Virtanen, and Moncef 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.

[10] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[11] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. Spoken Language Technology workshop (SLT)*, 2012.

[12] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou 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] David Sundermann and Hermann Ney, "VTLN-based voice conversion," in *Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2003.

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

[15] Daniel Erro, Eva Navas, and Inma 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.

[16] Xiaohai Tian, Zhizheng Wu, Siu Wa Lee, and Eng Siong Chng, "Correlation-based frequency warping for voice conversion," in *Proc. 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014.

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

[18] Elizabeth Godoy, Olivier Rosec, and Thierry 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.

[19] Hélène Valbret, Eric Moulines, and Jean-Pierre Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2, pp. 175–187, 1992.

[20] D Seung and L Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[21] Jort F. Gemmeke, Tuomas Virtanen, and Antti Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[22] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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.

[23] Tomoki Toda, Takashi Muramatsu, and Hideki Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, 2012.