EXEMPLAR-BASED SPARSE REPRESENTATION OF TIMBRE AND PROSODY FOR VOICE CONVERSION

Huaiping Ming[1, 2]*, Dongyan Huang*[2]*, Lei Xie*[1]*, Shaofei Zhang*[1]*, Minghui Dong*[2]*and Haizhou Li*[2]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

Voice conversion (VC) aims to make one speaker (source) to sound like spoken by another speaker (target) without changing the language content. Most of the state-of-the-art voice conversion systems focus only on timbre conversion. However, the speaker identity is characterized by the source-related cues such as fundamental frequency and energy as well. In this work, we propose an exemplar-based sparse representation of timbre and prosody for voice conversion that does not necessitate separately timbre conversion and prosody conversions. The experiment results show that, in addition to the conversion of spectral features, the proper conversion of prosody features will improve the quality and speaker identity of the converted speech.

Index Terms— Voice conversion, exemplar, timbre, prosody, sparse representation

1. INTRODUCTION

Human voice is a powerful and fundamental aspect of speaker identity. Each of us has a unique voice that reflects our age, our size, even our lifestyle and personality. Voice conversion is a technique to modify the characteristics (timbre or/and prosody) of one speaker (source) to make it sounds like spoken by another speaker (target) without changing the language content. As voice timbre is characterized by spectral features, which plays a main role to identify the speaker individuality [1], most of the existing voice conversion systems focus only on spectral conversion [2, 3, 4, 5, 6, 7, 8]. However, the previous study has shown that the source-related cues play important roles in contribution in transmitting the speaker identity when the listener is familiar with the speaker [9]. The source-related cues include fundamental frequency, energy, duration of words, rhythm and so on.

There are many studies on spectral conversion. Frequency warping approaches apply a mapping function to shift source spectra to match those of the target, such as dynamic frequency warping [11], weighted frequency warping [10], and bilinear frequency warping [2]. The drawback of frequency warping approaches is that the speaker identity conversion quality is not satisfactory [2, 10]. Except for frequency warping approaches, a great number of statistical parametric approaches have been studied for spectral conversion, such as vector quantization approaches [12], GMM-based approaches [3, 4], partial least square regression approaches [13], and neural network based approaches [5, 6]. The statistical parametric approaches convert speaker identity better than frequency warping approaches, but the converted speech quality is unsatisfactory. The reason why converted speech quality degrades is mainly due to the statistical averaging problem and the usage of low-resolution features [14]. To address these problems, exemplar-based voice conversion is proposed recently [14, 15]. Exemplar-based voice conversion reconstructs a speech spectrogram by a weighted linear combination of high-resolution spectra. In order to avoid over-smoothing, the linear combination weights are constrained to be sparse.

While spectral conversion has been extensively studied, prosody conversion still remains to be a challenging research topic. As prosody is affected by both short term dependencies and long term dependencies, it is hard to model the variations of F0 in all temporal scales. Previous studies of prosody conversion concentrate on converting the pitch of the source speaker to that of target speaker's F0 [16]. The most common approach is to transform the mean and variance of F0 from source speaker to that of target speaker [17]. Some extensions of this approach are proposed such as GMM-based mapping [19, 20], higher-order polynomial [18] and piecewise linear transformation based on hand-labelled intonational target points [21]. Recently, continuous wavelets transform (CWT) is used for the modeling and conversion of F0 in multiple time levels to obtain promising results. In [22], F0 is decomposed into ten levels by CWT, and wavelet levels 3-8 were converted using dynamic kernel partial least square regression for voice conversion. In [23], a five-scale CWT representation of F0 is used for prosody conversion of emotional voice under the exemplar-based framework.

In this paper, instead of converting spectral and prosody features separately, we propose to convert the spectrum, energy contour and fundamental frequency (F0) simultaneously under a unique framework of exemplar-based voice conversion. In order to capture and convert the dynamics of F0 at different temporal levels, a five-scale CWT representation of F0 is used for pitch conversion. To convert spectral and prosody features simultaneously, we build a *joint exemplar* that consists of spectrum, aperiodicity component, energy and the five-scale CWT representation of F0. A collection of acoustically aligned source and target joint exemplars, called source and target *joint dictionary*, are constructed from the training data. In the conversion stage, spectral and prosody features of new speech data are approximated as a sparse linear combination (activation function) of the source dictionary elements. The features of the target voice are then constructed by applying the activation with the target dictionary. The residual error from the source is then mapped to the target using partial least square regression and added to the constructed target. The system is evaluated objectively and subjectively. The experiment results support the higher accuracy and the more efficiency of the proposed method comparing with the conventional exemplar-based voice conversion and the GMM-based voice conversion.

The rest of this paper is organized as follow: In section 2, we introduce the basic idea of exemplar-based voice conversion. The details of the proposed method are described in section 3. In sec-

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61175018, 61571363) and the Aeronautical Science Foundation of China (20155553038).

tion 4, the objective and subjective experiment results are presented. Conclusions are drawn in Section 5.

2. EXEMPLAR-BASED VOICE CONVERSION

The basic idea of exemplar-based voice conversion is to describe a magnitude spectrum as a linear combination of a set of basis spectra, called exemplars. Let $\mathbf{x}_i \in \mathbb{R}^{F\times 1}$ represent the high resolution spectrogram of a speech frame, where F is the dimension of the spectrum. Then it can be expressed as

$$\mathbf{x}_i \approx \sum_{n=1}^{N} \mathbf{a}_n \cdot h_{n,i} = \mathbf{A}\mathbf{h}_i \qquad s.t. \qquad \mathbf{h}_i \geq 0. \tag{1}$$

where $\mathbf{a}_n \in \mathbb{R}^{F\times 1}$ is the n-th exemplar, $h_{n,i}$ is the n-th nonnegative weight, $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N] \in \mathbb{R}^{F\times N}$ is the dictionary of exemplar built form the training data, $\mathbf{h}_i = [h_{1,i}, h_{2,i}, ..., h_{N,i}] \in \mathbb{R}^{N\times 1}$ is the activation vector. As each frame of spectrum is modeled independently, the spectrum of an utterance can be expressed as

$$\mathbf{X} \approx \mathbf{A}\mathbf{H}, \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{F\times M} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M]$ is the spectrum, M is the number of frames in the source spectrum and $\mathbf{H} \in \mathbb{R}^{N\times M} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_M]$ is the activation matrix.

In this method, source dictionary $\mathbf{A}^s$ and target dictionary $\mathbf{A}^t$ are firstly constructed. These two dictionaries consist of the same words and are aligned with dynamic time warping (DTW). When the source signal and the target signal are the same words spoken by different speakers, they can be expressed with sparse representations of the source dictionary and the target dictionary respectively, and the obtained activity matrices are approximately equivalent [15]. For this reason, the activation matrix $\mathbf{H}$ estimated from source spectrum $\mathbf{X}$ and source dictionary $\mathbf{A}^s$ can be applied to target dictionary to generate target spectrum:

$$\mathbf{Y} = \mathbf{A}^t\mathbf{H}. \tag{3}$$

As the source spectrum and dictionary are both non-negative, the non-negative matrix factorization (NMF) method [14, 15] is applied to estimate the activation matrix. Mathematically, the activation matrix is found by minimizing the following objective function:

$$\mathbf{H} = \arg\min_{\mathbf{H}\geq 0} d(\mathbf{X}, \mathbf{A}^s\mathbf{H}) + \lambda\|\mathbf{H}\|_1, \tag{4}$$

where λ is the sparsity penalty factor, and $d(\bullet)$ is the cost function. Applying Kullback-Leigler (KL) divergence as the cost function, the objective function in Eq. (4) can be iteratively minimized by the following update rule:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{(\mathbf{A}^s)^T \frac{\mathbf{X}}{\mathbf{A}^s\mathbf{H}}}{(\mathbf{A}^s)^T + \lambda}, \tag{5}$$

where $\otimes$ represents element-wise multiplication and divisions are also element-wise.

3. PROPOSED METHOD

This method can be applied to both spectrum and prosody conversion under the framework of exemplar conversion. The details about spectral and prosody features extraction, training and conversion are described in the following.

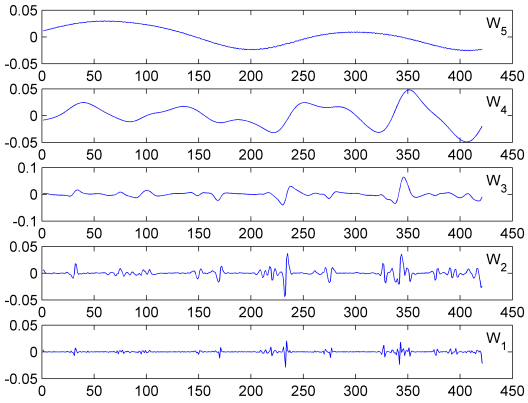

Fig. 1. An example of the five-scale representation of F0.

3.1. Spectral and Prosody Features

Given the parallel source and target data, spectrum, aperiodicity component and fundamental frequency are extracted from speech signals using STRAIGHT [24] analysis method. Denote the spectrum and aperiodicity component as $\mathbf{SP} \in \mathbb{R}^{F\times M}$, $\mathbf{AP} \in \mathbb{R}^{F\times M}$ respectively. To capture the energy contour of the speech signal, the energy of each frame is defined as

$$\mathbf{e}_m = \sqrt{\sum_{i=1}^{F} \mathbf{SP}_{i,m}^2} \quad , m = 1, ..., M. \tag{6}$$

By calculating the energy for each frame of a speech signal, we can obtain the energy contour vector $\mathbf{e} \in \mathbb{R}^{1\times M}$.

It is well known that fundamental frequency is influenced both at a supra-segmental level, by long-term dependencies, and at a segmental-level, by short-term dependencies. Inspired by the work in [25, 26], we adopt continuous wavelet transform to decompose the F0 contour into several temporal scales that model prosody at different temporal levels. The wavelet method is sensitive to the gaps in the F0 contour, so we only consider the voiced part. In order to explore the perceptual relevant information, the linear scale F0 contour is transformed to the logarithmic semitone scale. The log F0 contour is normalized to zero mean and unit variance as required by wavelet analysis, and we denote it as f_0.

The continuous wavelet transform of f_0 is defined by

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x)\psi\left(\frac{x-t}{\tau}\right) dx, \tag{7}$$

where $f_0(x)$ is the input signal and ψ is the Mexican hat mother wavelet. We fix the analysis at 10 discrete scales, each one octave apart. Then f_0 is represented by 10 separate components given by

$$W_i(f_0)(t) = W_i(f_0)(2^{i+1}\tau_0, t)(i + 2.5)^{-5/2}, \tag{8}$$

where $i = 1, ..., 10$ and $\tau_0 = 5$ ms. The original signal is approximately reconstructed by the following *ad hoc* reconstruction formula:

$$f_0(t) = \sum_{i=1}^{10} W_i(f_0)(t)(i + 2.5)^{-5/2}. \tag{9}$$

Attempting to relate the wavelet transform scales to levels of linguistic structure [25, 26], adjacent scales are combined, which result in a five-scale representation defined by

$$\mathbf{w}_i = W_{2i-1}(f_0)(t) + W_{2i}(f_0)(t), \tag{10}$$

where $i = 1, ..., 5$. An example of the five-scale representation of F0 is shown in Fig. 1. The lower scales (high frequencies) capture short-term variations and that higher scales (low frequencies) capture long-term variations. Thus, the five-scale representation can represent the dynamics of F0 in different time scales. As exemplar-based voice conversion requires non-negative features, the exponential value of the five-scale representation $\mathbf{w}$ is used in the proposed method:

$$\mathbf{F0}_{\text{cwt}} = \exp(\mathbf{w}). \tag{11}$$

3.2. Dictionary Construction

To align all the spectral and prosody features, we obtain Mel-cepstral coefficients (MCCs) by applying Mel-cepstral analysis technique [27] to the spectrum $\mathbf{SP}$. By applying dynamic time warping (DTW) to the source and target MCCs, the source-target frame align information is obtained. According to the frame level synchronization relationship between different features and the frame alignment information, we get the frame aligned source and target spectral and prosody feature as

$$\begin{bmatrix} \mathbf{sp}_1, \mathbf{sp}_2, ..., \mathbf{sp}_k, ..., \mathbf{sp}_K \\ \mathbf{ap}_1, \mathbf{ap}_2, ..., \mathbf{ap}_k, ..., \mathbf{ap}_K \\ e_1, e_2, ..., e_k, ..., e_K \\ \mathbf{f0}_{\text{cwt}1}, \mathbf{f0}_{\text{cwt}2}, ..., \mathbf{f0}_{\text{cwt}k}, ..., \mathbf{f0}_{\text{cwt}K} \end{bmatrix}, \tag{12}$$

where $\mathbf{sp} \in \mathbb{R}^{F\times 1}$, $\mathbf{ap} \in \mathbb{R}^{F\times 1}$, and $\mathbf{f0}_{\text{cwt}} \in \mathbb{R}^{5\times 1}$ are the column vectors from matrix $\mathbf{SP}$, $\mathbf{AP}$ and $\mathbf{F0}_{\text{cwt}}$ respectively, and $e \in \mathbb{R}^{+}$ is an energy value from energy vector $\mathbf{e}$.

With the acoustically aligned source and target features from training data, we can build the the *joint exemplars* thus construct the *joint dictionary* that consist of spectrum, aperiodicity component, energy and the five-representation of F0. Denote the frame aligned source and target spectrum, aperiodicity component, energy and five-representation of F0 as $\mathbf{SP}^s, \mathbf{AP}^s, \mathbf{e}^s, \mathbf{F0}^s_{\text{cwt}}, \mathbf{SP}^t, \mathbf{AP}^t, \mathbf{e}^t, \mathbf{F0}^t_{\text{cwt}}$ respectively, we propose to build a paired *joint exemplar* as

$$\mathbf{a}^s = \begin{bmatrix} \mathbf{sp}^s \\ \mathbf{ap}^s \\ e^s \\ \mathbf{f0}^s_{\text{cwt}} \end{bmatrix}, \quad \mathbf{a}^t = \begin{bmatrix} \mathbf{sp}^t \\ \mathbf{ap}^t \\ e^t \\ \mathbf{f0}^t_{\text{cwt}} \end{bmatrix}, \tag{13}$$

where $\mathbf{a}^s \in \mathbb{R}^{(2F+6)\times 1}$ is a super vector which represents a source exemplar, and $\mathbf{a}^t \in \mathbb{R}^{(2F+6)\times 1}$ is a super vector which represents a target exemplar. In the experiment, not all the exemplars are used to build the dictionary. We randomly select a subset of the paired *joint exemplars* to construct the coupled *joint dictionary* $\mathbf{A}^s \in \mathbb{R}^{(2F+6)\times(N)}$ and $\mathbf{A}^t \in \mathbb{R}^{(2F+6)\times(N)}$. The process of coupled *joint dictionary* construction is shown in Fig. 2.

3.3. Spectral and Prosody Feature Conversion

The aperiodicity component, energy and the five-scale representations of F0 are non-negative, and they can be represented as a linear combination of basis exemplars the same as spectrum. Moreover, the assumption that acoustically aligned source and target dictionary can share the same activation matrix still valid here. So we can convert spectrum, aperiodicity component, energy and fundamental frequency simultaneously in the framework of exemplar-based voice conversion.

Corresponding to the new *joint exemplar*, we need to redefine each column of matrix $\mathbf{X}$ and $\mathbf{Y}$ in Eq. (2) and Eq. (3) as a *joint super-vector*:

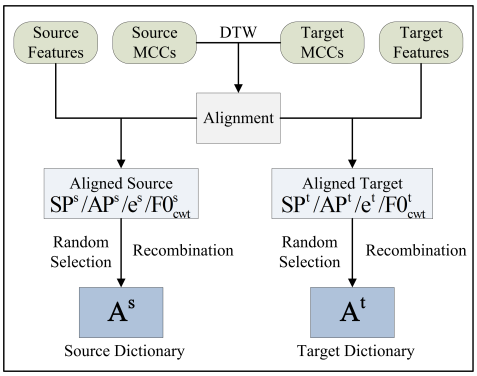

Fig. 2. The process of coupled *joint dictionary* construction.

$$\mathbf{x} = \begin{bmatrix} \mathbf{sp}^s \\ \mathbf{ap}^s \\ e^s \\ \mathbf{f0}^s_{\text{cwt}} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{sp}^t \\ \mathbf{ap}^t \\ e^t \\ \mathbf{f0}^t_{\text{cwt}} \end{bmatrix}, \tag{14}$$

where $\mathbf{sp}, \mathbf{ap}, e$ and $\mathbf{f0}_{\text{cwt}}$ are the features extracted from one frame of speech signal. In this way, Eq. (2) and Eq. (3) become:

$$\mathbf{X} \approx \mathbf{A}^s\mathbf{H}, \tag{15}$$

$$\mathbf{Y} = \mathbf{A}^t\mathbf{H}, \tag{16}$$

where $\mathbf{X} \in \mathbb{R}^{(2F+6)\times(M)}$ and $\mathbf{Y} \in \mathbb{R}^{(2F+6)\times(M)}$ are the new feature matrix with each column a *joint super-vector*. Note that, apart from feature dimension, Eq. (15) is the same as Eq. (3). Thus, we can use the same method for activation matrix estimation described in Eq. (5). Then the target joint spectral and prosody feature matrix $\mathbf{Y}$ is generated according to Eq. (16).

There is inevitably some modeling error between the source matrix $\mathbf{X}$ and the reconstructed $\mathbf{AH}$, called residual. A mapping implemented by partial least square regression (PLSR) [28] can be established between the source-target residual pairs. The predicted residual is compensated to $\mathbf{Y}$ as described in [14].

By reforming the components of $\mathbf{Y}$, we obtain the converted spectral and prosody features $\mathbf{SP}^c, \mathbf{AP}^c, \mathbf{e}^c$, and $\mathbf{F0}^c_{\text{cwt}}$. The logarithmic value of $\mathbf{F0}^c_{\text{cwt}}$ is calculated and the logarithmic scale converted F0 contour is reconstructed according to Eq. (9). Then the mean and variance of the converted logarithmic scale F0 contour are normalized to those of the target speaker. Finally, the exponential value of the logarithmic scale F0 is calculated to obtain the final converted F0 contour.

In order to make the energy contour of converted spectrum more close to that of the target, we take advantage of the information about converted energy $\mathbf{e}^c$. Firstly, we take the converted spectrum $\mathbf{SP}^c$ as input to calculate it's energy contour $\mathbf{e}^t$ according to Eq. (6). Then the energy ratio for each speech frame is calculated as

$$\mathbf{r} = \frac{\mathbf{e}^t}{\mathbf{e}^c}, \tag{17}$$

where the divisions are element-wise, and $\mathbf{r} \in \mathbb{R}^{1\times M}$. By replicating the energy ration vector, we get an energy ration matrix $\mathbf{R} \in \mathbb{R}^{F\times M}$. Finally, the energy contour improved spectrum is given by

$$\mathbf{SP}^c = \frac{\mathbf{SP}^c}{\mathbf{R}}, \tag{18}$$

where the divisions are also element-wise.

So far, we obtain the converted spectrum, aperiodicity component and fundamental frequency. These three features are passed to the STRAIGHT vocoder to reconstruct an audible speech signal. The process of spectral and prosody feature conversion is shown in Fig. 3.

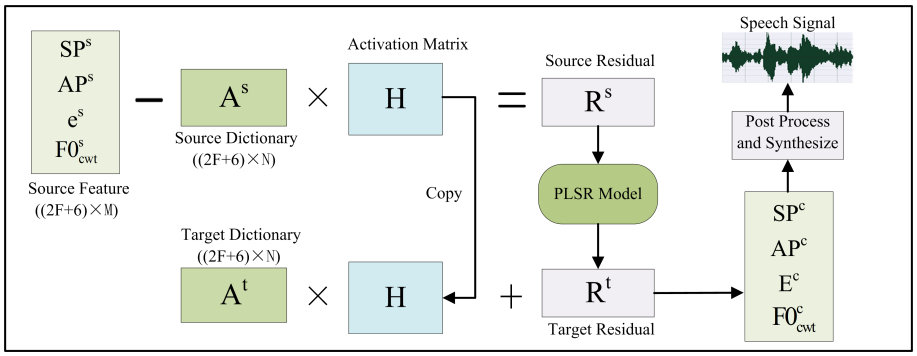

Fig. 3. The process of spectral and prosody feature conversion.

Table 1. The MCD of different voice conversion methods.

	Source	GMM	NMF-SP	Proposed Method
MCD (dB)	9.25	6.17	5.93	5.82

4. EXPERIMENTS

We conduct experiments using the CMU-ARCTIC [29] database to assess the performance of the proposed method. Speech data from two male speakers (bdl and rms) and two female speakers (clb and slt) are selected to conduct experiments. Voice conversion is conducted for all 12 speaker pairs. In each pair, 10 parallel utterances are selected as the training data, and another 10 utterances are selected as the development set. There are another 10 utterances selected as the evaluation set.

To validate our proposed method, we consider the state-of-the-art methods as our baselines, including the well established joint density GMM method with dynamic feature and global variance [4, 30], and exemplar-based voice conversion which only converts spectrum (NMF-SP) [14]. The fundamental frequency is linearly converted by transforming the mean and variance in the baseline methods.

4.1. Objective Evaluation

We use Mel-cepstral distortion (MCD) [4] of voiced part of speech samples as the objective measure to assess the proposed method. The average MCD result over all speaker pairs is reported. A lower MCD value means smaller spectral distortion.

The MCD results for different voice conversion methods are shown in Table 1. Comparing with GMM-based voice conversion, the proposed method achieves a lower MCD, that is 5.82dB over 6.17dB of GMM. This result confirms the effectiveness of the exemplar-based voice conversion, and it is consistent with the result in [14]. Comparing the proposed method with exemplar-based voice conversion which only converts spectrum, the MCD decreased 0.11dB. This confirms the effectiveness of the proposed method and implies the importance of prosody feature conversion for voice conversion.

4.2. Subjective Evaluation

We conduct listening tests to assess the performance of our proposed method and the baseline methods in terms of speech quality and speaker identity. In each test, 20 utterances are selected and 10 experienced listeners are involved.

We first conduct an AB preference test to assess speech quality. Speech sample A and B are obtained by different methods with the same input utterance. Speech A and B are presented to listeners in a random order. The listeners are required to choose the sample that has higher speech quality. If they are not able to perceive the difference of voice quality, then they can choose the option that claiming no preference.

Table 2. Comparing the proposed method with the GMM and the NMF-SP methods by quality and similarity preference score with 95% confidence intervals.

	Preference Score (%)	
	Quality Test	Similarity Test
GMM	33 (±15.1)	38 (±11.1)
Proposed Method	**55.5** (±14.3)	**48.5** (±8.4)
No Preference	11.5 (±3.7)	13.5 (±7.5)
NMF-SP	36 (±9.9)	34 (±9.0)
Proposed Method	**47** (±11.7)	**37.5** (±10.6)
No Preference	17 (±8.1)	28.5 (±7.5)

Then, an ABX test is conducted to assess the speaker similarity. Different from AB preference test, we have a reference target sample X. The listeners are asked to listen to the sample X first, then A and B. Then, they are required to choose a sample that is more closer to the target sample. If they are not able to decide which sample is closer to target, then they can choose the option claiming no preference.

The subject test results are presented in Table 2. Firstly, the proposed method is compared to GMM method. It is clear that the proposed method achieves a much higher preference score comparing to GMM method in both quality test and similarity test. The speech samples with no preference are less than 15% for both quality test and similarity test, which means there is a clear difference between the results of this two method. Then, the proposed method is compared to NMF-SP. We can see that the proposed method achieves a significant higher preference score in both quality test and similarity test. The above results confirm the effectiveness of the proposed method, and they are consistent with the objective evaluation results.

5. CONCLUSION

We propose a method to convert the spectrum, energy contour, aperiodicity component and fundamental frequency simultaneously in a sparse constrained exemplar-based voice conversion framework. The objective and subjective experiment results show that, the conversion of prosody features under exemplar-based voice conversion framework will lead to lower spectral distortion and higher preference score. The results suggest that in addition to spectral conversion, proper conversion of prosody is also critical for voice conversion.

6. REFERENCES

[1] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] 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.

[3] Hadas Benisty and David Malah, "Voice conversion using gmm with enhanced global variance," *INTERSPEECH*, pp. 669–672, 2011.

[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] 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/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.

[6] Feng-Long Xie, Yao Qian, Yuchen Fan, Frank K Soong, and Haifeng Li, "Sequence error (se) minimization training of neural network for voice conversion," *Proc. Interspeech*, pp. 2283–2287, 2014.

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

[8] Dong-Yan Huang, Ee Ping Ong, Susanto Rahardja, Minghui Dong, and Haizhou Li, "Transformation of vocal characteristics: A review of literature," *International Scholarly and Scientific Research and Innovation*, 2009.

[9] Rupal Patel, "Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 6, pp. 1401–1415, 2003.

[10] 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.

[11] 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.

[12] K Shikano, S Nakamura, and M Abe, "Speaker adaptation and voice conversion by codebook mapping," *IEEE International Sympoisum on Circuits and Systems*, pp. 594–597, 1991.

[13] 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.

[14] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 10, pp. 1506–1521, 2014.

[15] Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," pp. 7894–7898, 2014.

[16] Zhizheng Wu, "Spectral mapping for voice conversion," *PhD Thesis, St. Scho ol of Computer Engineering, Nanyang Technological University*, 2015.

[17] 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.

[18] David T Chappell and John HL Hansen, "Speaker-specific pitch contour modeling and modification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 885–888, 1998.

[19] Zeynep Inanoglu, "Transforming pitch in a voice conversion framework," *Master's Thesis, St. Edmunds College, University of Cambridge*, 2003.

[20] Zhizheng Wu, Tomi Kinnunen, Engsiong Chng, and Haizhou Li, "Text-independent f0 transformation with non-parallel data for voice conversion.," *INTERSPEECH*, pp. 1732–1735, 2010.

[21] Ben Gillett and Simon King, "Transforming f0 contours," *INTERSPEECH*, 2003.

[22] Gerard Sanchez, Hanna Silen, Jani Nurminen, and Moncef Gabbouj, "Hierarchical modeling of f0 contours for voice conversion," *INTERSPEECH*, 2014.

[23] Huaiping Ming, Dongyan Huang, Lei Xie, Shaofei Zhang, Minghui Dong, and Haizhou Li, "Fundamental frequency modeling using wavelets for emotional voice conversion," *Computing and Intelligent Interaction Workshop on Affective Social Multimedia Computing*, 2015.

[24] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3933–3936, 2008.

[25] Manuel Sam Ribeiro and Robert AJ Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[26] Antti Santeri Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, Martti Vainio, et al., "Wavelets for intonation modeling in hmm speech synthesis," *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[27] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation.," *International Conference on Spoken Language Processing (ICSLP)*, 1994.

[28] 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.

[29] John Kominek and Alan W Black, "The cmu arctic speech databases," *Fifth ISCA Workshop on Speech Synthesis*, 2004.

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