EXEMPLAR-BASED VOICE CONVERSION USING NON-NEGATIVE SPECTROGRAM DECONVOLUTION

Zhizheng Wu[1,2], Tuomas Virtanen[3], Tomi Kinnunen[4], Eng Siong Chng[1,2], Haizhou Li[1,2,5]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[4]School of Computing, University of Eastern Finland, Joensuu, Finland
[5]Human Language Technology Department, Institute for Infocomm Research, Singapore
wuzz@ntu.edu.sg

ABSTRACT

In the traditional voice conversion, converted speech is generated using statistical parametric models (for example Gaussian mixture model) whose parameters are estimated from parallel training utterances. A well-known problem of the statistical parametric methods is that statistical average in parameter estimation results in the over-smoothing of the speech parameter trajectories, and thus leads to low conversion quality. Inspired by recent success of so-called exemplar-based methods in robust speech recognition, we propose a voice conversion system based on non-negative spectrogram deconvolution with similar ideas. Exemplars, which are able to capture temporal context, are employed to generate converted speech spectrogram convolutely. The exemplar-based approach is seen as a data-driven, non-parametric approach as an alternative to the traditional parametric approaches to voice conversion. Experiments on VOICES database indicate that the proposed method outperforms the conventional joint density Gaussian mixture model by a wide margin in terms of both objective and subjective evaluations.

Index Terms— Voice conversion, exemplar, non-negative matrix factorization, non-negative matrix deconvolution, temporal information

1. INTRODUCTION

Voice conversion is a process of modifying source speaker's voice to sound like it was spoken by another speaker (target). It can be applied to speaker identity conversion in speech synthesis systems when only a few recording samples from a specific target speaker are available.

In general, voice conversion techniques operate on several different speech features, such as spectral envelope [1, 2], formants [3], fundamental frequency [4, 5] and duration [6]. Spectral envelope contains most of the speaker identity information and is the focus in most of the voice conversion studies, including this one. Spectral conversion involves two phases, training and run-time conversion. During training, a transformation function is estimated from frame-aligned source-target feature vectors. The trained conversion model is then applied to unseen utterances at system run-time. Implementation of the conversion function is the most important part of a voice conversion system.

To implement a robust spectral conversion function, a number of data-driven statistical parametric methods have been proposed in the past two decades. A straightforward way to model the relationship between source and target speech is to employ vector quantization (VQ) to learn a codebook from the paired source-target frame vectors, and apply this codebook during conversion phase [7].

To alleviate the frame-to-frame discontinuity problem caused by VQ, joint density Gaussian mixture model (JD-GMM) was proposed [8, 9, 1]. It implements a smoothed local linear transformation function for each frame. Other local linear transformation methods, such as partial least square regression [10], trajectory GMM/hidden Markov model (HMM) [11], mixture of factor analyzers [12], local linear transformation [13], noisy channel model [2] and so on, have been proposed to reduce the over-smoothing and over-fitting problems of JD-GMM. In addition to the linear transformation functions, which assume the source and target speech features to be linearly correlated, nonlinear methods, such as artificial neural network [3, 14], support vector regression [15], kernel partial least square regression [16], and conditional restricted Boltzmann machine [17], have been studied to implement nonlinear conversion.

Due to inherent statistical averaging in parametric methods, over-smoothed speech samples are generated from the averaged parameters, which leads to unnatural speech quality. Inspired by the success of so-called *exemplar*-based noise robust speech recognition [18, 19, 20], we propose a non-parametric exemplar-based voice conversion method as an alternative to statistical parametric methods. We define an exemplar to be a segment of speech spectrogram spanning multiple frames. Utilizing multiple frames, as opposed to single frame in the conventional methods, allows contextual modelling which helps increasing the resulting speech quality.

We study two examplar-based voice conversion variants: *non-negative spectrogram factorization (NMF)* and *non-negative spectrogram deconvolution (NMD)*. In the former variant, each spectrogram frame is represented as a convex combination of several basis spectra (atoms) forming a dictionary. In the deconvolution variant, a converted spectrogram is generated as a convolution of exemplars and activations. Comparing with the most related work in [21], our work has the following novel contributions:

a) We utilize multiple-frame exemplar rather than single-frame spectrum as the basis in the dictionary;

b) We employ low-dimensional filter-bank energies instead of the original magnitude spectrum to represent source spectrogram and source dictionary for efficient computation;

c) We employ a convolutive model to include temporal context information in the converted spectrogram.

2. BASELINE JOINT DENSITY GAUSSIAN MIXTURE MODEL METHOD

Among the statistical parametric methods, joint density Gaussian mixture model (JD-GMM) method [8, 1] is one of the most successful methods, due to the probabilistic treatment and flexible implementation. Therefore, we employ the JD-GMM method as our baseline method in this study.

The JD-GMM method involves two phases: off-line training and run-time conversion phases. During the training phase, given parallel training data from a source speaker $\mathbf{X}$ and a target speaker $\mathbf{Y}$, dynamic time warping (DTW) algorithm is used to align the source speech vectors and target speech vectors to obtain the paired speech feature vector $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t, \ldots, \mathbf{z}_T]$, where $\mathbf{z}_t = [\mathbf{x}_n^\top, \mathbf{y}_m^\top]^\top \in \mathcal{R}^{2d}$, and $\mathbf{x}_n \in \mathcal{R}^d$ and $\mathbf{y}_m \in \mathcal{R}^d$ are source and target speech feature vectors, respectively.

Gaussian mixture model (GMM) is adopted to model the distribution of the paired feature vector sequence $\mathbf{Z}$, which represents the joint distribution of source speech $\mathbf{X}$ and target speech $\mathbf{Y}$. The joint probability density is given as follows:

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{Z}) = \sum_{k=1}^{K} w_k^{(z)} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k^{(z)}, \boldsymbol{\Sigma}_k^{(z)}), \quad (1)$$

$$\boldsymbol{\mu}_k^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_k^{(x)} \\ \boldsymbol{\mu}_k^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_k^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(xx)} & \boldsymbol{\Sigma}_k^{(xy)} \\ \boldsymbol{\Sigma}_k^{(yx)} & \boldsymbol{\Sigma}_k^{(yy)} \end{bmatrix},$$

where K is the number of Gaussian components, $\boldsymbol{\mu}_k^{(z)}$ and $\boldsymbol{\Sigma}_k^{(z)}$ are the mean vector and the covariance matrix of the k^{th} Gaussian component $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k^{(z)}, \boldsymbol{\Sigma}_k^{(z)})$, respectively. The prior probability $w_k^{(z)}$ of the k^{th} Gaussian component is constrained by $\sum_{k=1}^{K} w_k^{(z)} = 1$. To estimate the model parameters of the joint density Gaussian mixture model $\lambda^{(z)} = \{w_k^{(z)}, \mu_k^{(z)}, \boldsymbol{\Sigma}_k^{(z)} | k = 1, 2, \ldots, K\}$, the well-known expectation-maximization (EM) algorithm is adopted to maximize likelihood of the training data.

In the run-time conversion phase, JD-GMM model parameters are employed to implement the conversion function. To be more specific, for each input source speech feature vector $\mathbf{x}$, the conversion function $F(\mathbf{x})$ implemented with minimum mean square error is used to predict the target's feature vector $\hat{\mathbf{y}}$ is given as:

$$\hat{\mathbf{y}} = F(\mathbf{x}) = \sum_{k=1}^{K} p_k(\mathbf{x})(\boldsymbol{\mu}_k^{(y)} + \boldsymbol{\Sigma}_k^{(yx)}(\boldsymbol{\Sigma}_k^{(xx)})^{-1}(\mathbf{x} - \boldsymbol{\mu}_k^{(x)})), \quad (2)$$

$$p_k(\mathbf{x}) = \frac{w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx})}{\sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx})},$$

where $p_k(\mathbf{x})$ is the posterior probability of the source vector $\mathbf{x}$ generated from the k^{th} Gaussian component.

We note that during the JD-GMM model parameter estimation process, the mean vector of each Gaussian component is updated as:

$$\boldsymbol{\mu}_k^{(z)} = \frac{\sum_{t=1}^{T} \mathbf{z}_t p_k(\mathbf{z}_t, \lambda^{(z)})}{\sum_{t=1}^{T} p_k(\mathbf{z}_t, \lambda^{(z)})}. \quad (3)$$

Similarly, the covariance matrix of each Gaussian component is updated as:

$$\boldsymbol{\Sigma}_k^{(z)} = \frac{\sum_{t=1}^{T} p_k(\mathbf{z}_t, \lambda^{(z)})(\mathbf{z}_t - \boldsymbol{\mu}_k^{(z)})(\mathbf{z}_t - \boldsymbol{\mu}_k^{(z)})^\top}{\sum_{t=1}^{T} p_k(\mathbf{z}_t, \lambda^{(z)})} \quad (4)$$

From (3) and (4), we observe that when calculating mean and covariance for each Gaussian component, all the training samples are used, which is the so-called statistical average. The statistical average results in over-smoothing of the converted speech. We also note that if the correlation between the paired source and target feature vectors is low, the value of the covariance matrix $\Sigma_k^{(yx)}$ will be very small, therefore, only $\boldsymbol{\mu}_k^{(y)}$ contributes to the converted speech as observed and reported in [22].

3. PROPOSED EXEMPLAR-BASED VOICE CONVERSION METHOD

To tackle the over-smoothing problem, we propose an exemplar-based method to generate the converted speech from the spectrogram segments (exemplar). We employ two matrix factorization techniques to implement the exemplar-based method: non-negative spectrogram factorization and non-negative spectrogram deconvolution. Both implementations have the same procedures as follows:

1 Training: construct parallel source and target dictionaries;

2 Conversion:

- 2.a Extract source spectrogram;
- 2.b Given source spectrogram and source dictionary, estimate activation matrix;
- 2.c Utilize the activation matrix estimated in step 2.b and the target dictionary to generate the converted spectrogram;

The two implementations using matrix factorization techniques are briefly introduced in this section.

3.1. Non-negative spectrogram factorization (NMF)

The first exemplar-based method is based on *non-negative spectrogram factorization*. The basic idea of this method is to represent a magnitude spectrum as a linear combination of a set of basis spectra (*speech atoms*). It is formulated as follows:

$$\mathbf{x} = \sum_{t=1}^{T} \mathbf{a}_t^{(\mathrm{X})} \cdot h_t = \mathbf{A}^{(\mathrm{X})} \cdot \mathbf{h}, \quad (5)$$

where $\mathbf{x} \in \mathcal{R}^{p \times 1}$ represents the spectrum of one frame, T is the total number of speech atoms, $\mathbf{A}^{(\mathrm{X})} = [\mathbf{a}_1^{(\mathrm{X})}, \mathbf{a}_2^{(\mathrm{X})}, \cdots, \mathbf{a}_T^{(\mathrm{X})}] \in \mathcal{R}^{p \times T}$ is the dictionary of speech atoms built from training source speech, $\mathbf{a}_t^{(\mathrm{X})}$ is the t^{th} speech atom which has the same dimension as $\mathbf{x}$, $\mathbf{h} = [h_1, h_2, \cdots, h_T] \in \mathcal{R}^{T \times 1}$ is the non-negative weight or activation vector and h_t is the activation of the t^{th} speech atom.

Therefore, the spectrogram of each source utterance can be represented as:

$$\mathbf{X} = \mathbf{A}^{(\mathrm{X})} \cdot \mathbf{H}, \quad (6)$$

where $\mathbf{X} \in \mathcal{R}^{p \times M}$ is the source spectrogram, and $\mathbf{H} \in \mathcal{R}^{T \times M}$ is the activation matrix, the column vector of which is the activation vector in Eq. (5).

In order to generate converted speech spectrogram, we assume that the aligned source and target dictionaries share the same activation matrix. To this end, we represent the converted spectrogram as:

$$\mathbf{Y} = \mathbf{A}^{(\mathrm{Y})} \cdot \mathbf{H}, \quad (7)$$

where $\mathbf{Y} \in \mathcal{R}^{q \times M}$ is the converted spectrogram, and $\mathbf{A}^{(\mathrm{Y})} \in \mathcal{R}^{q \times T}$ is the dictionary of the target speech atoms from target training data.

The illustration of Eq. (6) and (7) is presented in Fig. 1. The source and target dictionaries $\mathbf{A}^{(\mathrm{X})}$ and $\mathbf{A}^{(\mathrm{Y})}$ are constructed from parallel training data and they remain the same during the conversion phase. During the conversion phase, the source spectrogram is given and the activation matrix is obtained as a solution of non-negative matrix factorization as in [18]. Then, the activation matrix estimated from Eq. (6) is then directly employed in Eq. (7) to generate the converted spectrogram.

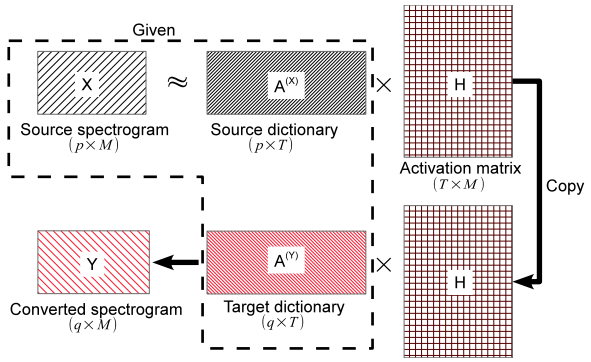

Fig. 1. Illustration of non-negative spectrogram factorization for exemplar-based voice conversion

3.2. Non-negative spectrogram deconvolution (NMD)

Although temporal constraints can be included in the estimation of activation matrix by using multiple-frame exemplars as source speech atoms, the converted speech spectrogram is still generated frame-by-frame. In order to utilize temporal context in the generation process of the converted spectrogram, we propose *non-negative spectrogram deconvolution (NMD)* method for exemplar-based voice conversion. In the NMD method, a spectrogram is represented as a convolution of exemplars and activations. The idea is formulated as follows:

$$\mathbf{X} = \sum_{l=1}^{L} \mathbf{A}_l^{(\mathrm{X})} \cdot \overset{\rightarrow(l-1)}{\mathbf{H}}, \tag{8}$$

$$\mathbf{Y} = \sum_{l=1}^{L} \mathbf{A}_l^{(\mathrm{Y})} \cdot \overset{\rightarrow(l-1)}{\mathbf{H}}, \tag{9}$$

where $\mathbf{A}_l^{(\mathrm{X})} \in \mathcal{R}^{p\times T}$ and $\mathbf{A}_l^{(\mathrm{Y})} \in \mathcal{R}^{q\times T}$ are the matrices consisting of the l^{th} frame of the source and target atoms, respectively, L is the number of adjacent frames within an examplar and $\mathbf{H}$ is the activation (weights) matrix as that in Eq. (6). $\overset{\rightarrow(l-1)}{(\cdot)}$ operator shifts the matrix entries (columns) to the right by $(l-1)$ units. In practice, several consecutive frames of an exact frame can be stacked into one supervector to represent the exact frame for constructing the source dictionary $\mathbf{A}_l^{(\mathrm{X})} \in \mathcal{R}^{p\times T}$. Therefore, $p = L \times d$ other than $p = d$, where d is the dimension the spectrum. During conversion, a source spectrogram $\mathbf{X}$ is first decomposed to estimate the activation matrix, and then the converted speech spectrogram $\mathbf{Y}$ is generated as a convolution of the target speech atoms and the corresponding activation matrix. The activation matrix is obtained by minimizing the generalized Kullback-Leibler divergence as explained in [19].

3.3. Dictionary construction

As discussed above, dictionary is important for both estimating the activation matrix and generating the converted speech signal. Before introducing how to construct dictionary, we first introduce the related features used to represent spectrum. In this work, the STRAIGHT [23] system is employed to extract spectral envelope and fundamental frequency (F0). The following three features are involved in this study:

a) *Magnitude spectra (MSP)*: Magnitude spectra consist a sequence of spectral envelopes extracted by STRAIGHT. We use 513 dimensional spectra. Magnitude spectra can be passed to STRAIGHT for reconstructing speech signal. In this work, target dictionary and converted spectrogram are always represented by MSP.

b) *Mel-scale magnitude spectra (MMSP)*: MMSP is obtained by passing the magnitude spectrogram to a 23-channel Mel-scale filter-bank. The minimum frequency is set to be 133.33 Hz, and the maximum frequency is set to be 6,855.5 Hz. In this work, MMSP is only used in the source dictionary to estimate the activation matrix but not for synthesizing speech.

c) *Mel-cepstral coefficient* (MCC): MCC is obtained by employing mel-cepstral analysis technique on the magnitude spectrogram and keeping 24 coefficients as the feature. During synthesis, MCCs are converted back to magnitude spectrogram, which is then passed to the STRAIGHT synthesis filter to reconstruct speech signal. In this work, MCC is only used in the JD-GMM method and in the dynamic time warping to align two parallel utterances.

Given one pair of parallel utterances from source and target, the following process is employed to construct the dictionary.

1) Extract magnitude spectrogram (spectral envelopes) from both source and target speech signal using STRAIGHT;

2) Apply mel-cepstral analysis [24] on the spectrograms to obtain mel-cepstral coefficients (MCCs);

3) Apply 23-channel Mel-scale filter-bank to obtain 23-dimensional MMSP;

4) Perform dynamic time warping on the source and target MCC sequence to align the speech to obtain source-target frame pairs;

5) Apply the alignment information to the source and target spectrograms. The resulting spectrum pairs are stored in the source and target dictionaries (column vectors), respectively.

The above five steps are applied for all the parallel training utterances. All the spectrum pairs (column vectors in source and target dictionaries) are used as speech atoms. In order to include multiple frames, consecutive frames are stacked into a super-vector to represent one frame. We note that for simple explanation, same features (both spectral envelopes) are used in step 5. As the size of the activation matrix is independent of the dimensionality of the features (column dimensionality), therefore, 23-dimensional MMSP can be used to replace 513-dimensional MSP in the source dictionary. While *513-dimensional MSP is always used in the target dictionary for synthesizing speech purpose*. More details will be discussed in Section 4.

4. EXPERIMENTS

To evaluate the proposed methods, we conduct experiments using the VOICES database [25]. Male-to-female and female-to-male conversions are conducted. For each conversion, 10 utterances from each speaker are selected as training data and 20 utterances, which are not included in the training data, are used as testing data.

In the experiments, three methods are compared. They are summarized as follows:

a) *JD-GMM*: The joint density Gaussian mixture model method (Section 2). The number of Gaussian components is set to be 32.

b) *NMF*: The proposed non-negative spectrogram factorization method (Section 3.1).

c) *NMD*: The proposed non-negative spectrogram deconvolution method (Section 3.2).

In the JD-GMM method, 24-dimensional MCC features are used to represent spectral envelope and to synthesize speech signal, while in NMF and NMD method, 513-dimensional MSP is used in the target dictionary and to synthesize speech signal. Log-scale F0 is converted by equalizing the mean and variance of the source and target speech.

4.1. Objective evaluation

Two objective measures are employed to evaluate the proposed method objectively. The first objective measure is spectral distortion: *mel-cepstral distortion* (MCD), which is calculated between a converted frame and the corresponding original target frame. We note that the frame alignment is obtained by performing dynamic time warping between parallel source and target sentences. The MCD for the m^{th} frame is calculated as:

$$\text{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}(c_{m,d} - c_{m,d}^{\text{conv}})}, \qquad (10)$$

where, M is frame number in one utterance, $c_{m,d}$ and $c_{m,d}^{\text{conv}}$ are the d^{th} dimension of the original target and converted MCCs of the m^{th} frame, respectively. We report the average MCD value over all the frames. A lower MCD value indicates smaller distortion. The second objective measure is the *correlation coefficient*, which is calculated between the original target and the converted MCC parameter trajectories dimension-by-dimension. The correlation coefficient γ_d of the d^{th} MCC trajectory is computed as follows:

$$\gamma_d = \frac{\sum_{m=1}^{M}(c_{m,d} - \overline{c_d})(c_{m,d}^{\text{conv}} - \overline{c_d^{\text{conv}}})}{\sqrt{\sum_{m=1}^{M}(c_{m,d} - \overline{c_d})^2}\sqrt{\sum_{m=1}^{M}(c_{m,d}^{\text{conv}} - \overline{c_d^{\text{conv}}})^2}}, \qquad (11)$$

where $\overline{c_d}$ and $\overline{c_d^{\text{conv}}}$ are the mean values of the original target and converted MCCs of the d^{th} dimension, respectively. We note that correlation coefficient is calculated sentence-by-sentence and we report the average correlation coefficient. Different from MCD, correlation coefficient focuses on the trajectory-level similarity, which is not affected by the mean and variance of the MCC trajectory, and has been used to measure the fundamental frequency trajectory similarity [5, 26]. Bigger correlation coefficient indicates Higher similarity between the original target and the converted MCC trajectories. We report the average correlation coefficient over all dimension.

In order to obtain comparable MCD and correlation results, in the NMF and NMD method, mel-cepstral analysis is applied to the converted spectrogram to get the 24-dimensional MCCs for computing MCD and correlation coefficient. Both MCD and correlation coefficient results reported in this work are averaged over the conversion pairs.

As shown in Eq. (6) and (7), and Fig. 1, the dimensionality of the activation matrix is independent of the dimensionality of the exemplars in both the source and the target dictionaries. Therefore, we first evaluate the performance of NMF using different features in source dictionary for estimating the activation matrix. *We note that for all the experiments, target dictionary always use the 513-dimensional magnitude spectra*, as the target dictionary does not affect the activation matrix and also is used to synthesize speech signal.

As discussed above, the dimensionality of the spectral envelope from STRAIGHT is 513 (1024-point FFT). If the original magnitude spectra (MSP) are used to estimate the activation matrix, as illustrated in Fig. 1, the dimensionality of the source dictionary $\mathbf{A}^{(\text{X})}$ will be $513 \times T$, assuming that each exemplar spans only one frame. If each exemplar spans 11 frame, the dimensionality of the source dictionary $\mathbf{A}^{(\text{X})}$ will be $5,643 \times T$, where T is the number of atoms. The huge dimensionality of the source dictionary will increase the computation and memory usage considerably. To reduce computation and memory usage, low-dimensional features will be a better choice. In this study, we propose to use 23-dimensional MMSP instead of the 513-dimensional original MSP to make the source dictionary for estimating the activation matrix. While the target dictionary reminds same as discussed above.

Table 1 presents the spectral distortions and correlations of NMF using 513-dimensional MSP and 23-dimensional MMSP in the source dictionary. Here, an exemplar spans only one frame. The results show that, even the dimensionality is reduced from 513 to 23, the distortion only increases 0.06 dB, and the correlation decreases by 0.003. The benefit of using 23-dimensional MMSP instead of 513-dimensional MSP in source dictionary to represent speech signal is that more consecutive frames can be included in the exemplar to estimate the activation matrix without increasing the computation cost and memory usage too much.

Table 1. Comparison of NMF results using 513-dimensional magnitude spectra (MSP) and 23-dimensional Mel-scale magnitude spectra (MMSP) in the source dictionary $\mathbf{A}^{(\text{X})}$. 513-dimensional MSP is always used in the target dictionary $\mathbf{A}^{(\text{Y})}$.

Features in source dictionary $\mathbf{A}^{(\text{X})}$	MCD (dB)	Correlation
MSP (513 dimensions)	5.47	0.439
MMSP (23 dimensions)	5.53	0.436

We then evaluate the performance of NMF using multiple frames in an exemplar for source dictionary. The spectral distortion results as a function of the window size (number of consecutive frames) of an exemplar is presented in Fig. 2. For Mel-scale magnitude spectra, the window size of exemplar is varied. While for 513-dimensional magnitude spectra, only one frame spectrum is employed in the exemplar due to computation restrictions discussed above. The results show that when the window size is larger than 3, 23-dimensional MMSP yields lower MCD and higher correlation coefficient than 513-dimensional MSP. NMF with exemplar using MMSP and spanning 9 frames gives the lowest spectral distortion. We note that the dimensionality of exemplar using MMSP and spanning 9 frames is $23 \times 9 = 207$, which is still much smaller than 513. The correlation results in Fig. 3 agree well with the spectral distortion results.

Next, we evaluate the proposed non-negative deconvolution (NMD) method using 23-dimensional MMSP. As shown above, 9

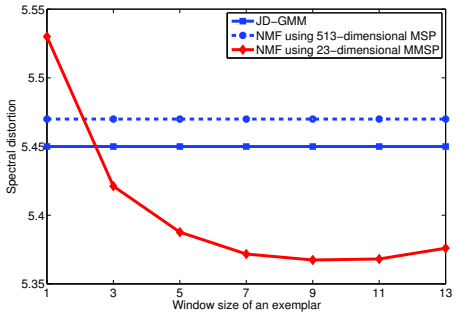

Fig. 2. The spectral distortion results of NMF method using different features with the baseline JD-GMM method as a reference

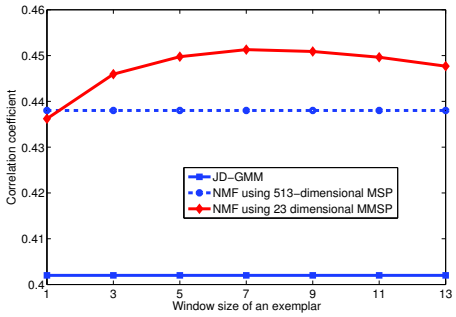

Fig. 3. The correlation coefficient results of NMF method using different features with the baseline JD-GMM method as a reference

frame exemplars give lowest distortion in the NMF method, therefore, in the NMD method, we stack 9 consecutive frames of an exact frame to represent the exact frame. Therefore, In Eq. (1), $p = 9 \times 23 = 207$. The spectral distortion results are presented in Fig. 4, as a function of the window size of an exemplar. Comparing with JD-GMM method, we observe that NMD method always obtains lower spectral distortion. NMD and NMF methods have similar performance in terms of spectral distortion when the window size is 5 or 7. The correlation coefficient results are shown in Fig. 5. It clearly shows that NMD has the highest correlation coefficients in all the cases. We note that different from NMF, NMD method utilizes multiples target frames (an exemplar) not only to estimate the activation matrix but also to generate the converted spectrogram.

4.2. Subjective evaluation

To assess the similarity of the converted speech to the target speech, a similarity preference listening test was conducted. The JD-GMM, and the two proposed methods: NMF and NMD are compared. 10 converted utterances from each method were randomly selected, including 5 utterances from the male-to-female conversion and the other 5 utterances from the female-to-male conversion. 11 subjects were asked to listen to a reference target speech and then the three converted speech samples representing the three methods. After that they were asked to decide which speech sample is more closer to the reference target speech sample. The preference scores with 95% confidence interval are presented in Fig. 6. We can clearly observe that the proposed NMF and NMD methods are both able to generate speech samples which are more similar to the target speaker than the

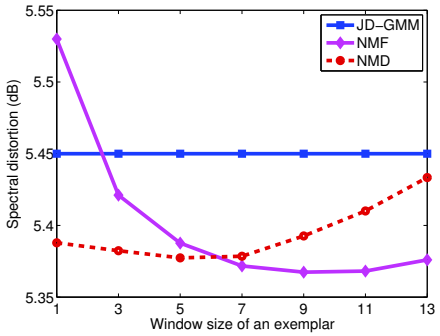

Fig. 4. Comparison of the spectral distortion results of JD-GMM, NMF and NMD methods as a function of the window size of an exemplar

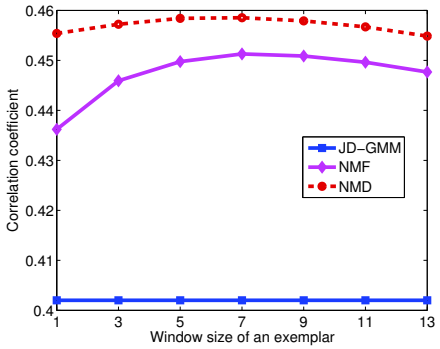

Fig. 5. Comparison of the correlation results of JD-GMM, NMF and NMD methods as a function of the window size of an exemplar

baseline JD-GMM method. We note that during the listening test, when the subjects are not able to distinguish the similarity across speech samples, they prefer to choose the one which gives better quality. Therefore, the similarity can reflect the speech quality of the three methods to some degree.

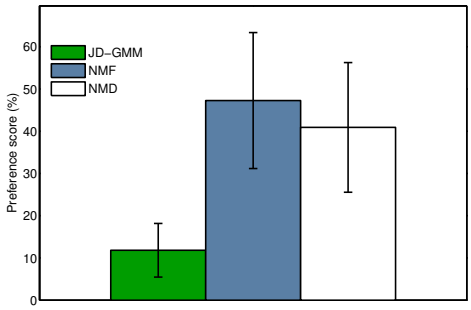

Fig. 6. Similarity results of the preference score with 95% confidence interval

5. CONCLUSIONS

In this paper, we proposed an exemplar-based voice conversion method utilizing the matrix/spectrogram factorization techniques. Two implementations, non-negative spectrogram factorization and non-negative spectrogram deconvolution, are proposed to use original target spectrogram directly without any dimension reduction to synthesize the converted speech. The experiment results show the proposed method outperforms the conventional joint density Gaussian mixture model considerably.

6. ACKNOWLEDGEMENT

The work of Tuomas Virtanen (projects no. 258708) and Tomi Kinnunen was supported by Academy of Finland (projects no. 253120). The authors would like to thank all the listeners who take part in the subjective evaluation test.

7. REFERENCES

[1] 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.

[2] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1784–1794, 2012.

[3] M. Narendranath, H.A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[4] B. Gillet and S. King, "Transforming F0 contours," in *Proceedings of Eurospeech*, 2003, pp. 101–104.

[5] Z.Z. Wu, T. Kinnunen, E.S. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[6] C. H. Wu, C. C. Hsia, T. H. Liu, and J. F. Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1109–1116, 2006.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP 1998*.

[8] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP 1998*.

[9] 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.

[10] 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.

[11] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory hmms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.

[12] Z. Wu, T. Kinnunen, E. Chng, and H. Li, "Mixture of factor analyzers using priors from non-parallel speech for voice conversion," *Signal Processing Letters, IEEE*, 2012.

[13] V. Popa, H. Silen, J. Nurminen, and M. Gabbouj, "Local linear transformation for voice conversion," in *ICASSP 2012*.

[14] S. Desai, E.V. Raghavendra, .B Yegnanarayana, A.W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *ICASSP 2009*.

[15] P. Song, Y.Q. Bao, L. Zhao, and C.R. Zou, "Voice conversion using support vector regression," *Electronics letters*, vol. 47, no. 18, pp. 1045–1046, 2011.

[16] 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.

[17] Z. Wu, E.S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *the first IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013.

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

[19] A. Hurmalainen, J. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *ICASSP 2011*.

[20] T.N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J.F. Gemmeke, J.R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98 –113, nov. 2012.

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

[22] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Eurospeech-2003*.

[23] 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.

[24] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP 1992*.

[25] A. KAIN, "High resolution voice transformation," *Ph. D. Thesis, OGI School of Science and Engineering, Oregon Health and Science University*, 2001.

[26] Y. Qian, Z. Wu, B. Gao, and F.K. Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1702–1710, 2011.