INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-embedded Non-negative Matrix Factorization

Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki

Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan

aihara@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

This paper proposes a discriminative learning method for Non-negative Matrix Factorization (NMF)-based Voice Conversion (VC). NMF-based VC has been researched because of the natural-sounding voice it produces compared with conventional Gaussian Mixture Model (GMM)-based VC. In conventional NMF-based VC, parallel exemplars are used as the dictionary; therefore, dictionary learning is not adopted. In order to enhance the conversion quality of NMF-based VC, we propose Discriminative Graph-embedded Non-negative Matrix Factorization (DGNMF). Parallel dictionaries of the source and target speakers are discriminatively estimated by using DGNMF based on the phoneme labels of the training data. Experimental results show that our proposed method can not only improve the conversion quality but also reduce the computational times.

Index Terms: voice conversion, speech synthesis, NMF, spare representation

1. Introduction

Non-negative Matrix Factorization (NMF) [1] is one of the most popular sparse representation methods. The goal is to estimate the basis matrix $\mathbf{W}$ and its weight matrix $\mathbf{H}$ from the input observation $\mathbf{V}$ such that:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}. \tag{1}$$

In this paper, we refer to $\mathbf{W}$ as the "dictionary" and $\mathbf{H}$ as "activity". NMF has been applied to hyperspectral imaging [2], topic modeling [3], and the analysis of brain data [4].

The NMF-based method can be classified into two approaches: the dictionary-learning approach and exemplar-based approach. In the dictionary-learning approach, the dictionary and the activity are estimated simultaneously. This approach has been widely applied in the field of audio signal processing: for example single channel speech separation [5, 6] and music transcription [7]. By estimating the dictionary from the training data, reconstruction errors between $\mathbf{V}$ and $\mathbf{WH}$ tend to be small. However, because not only the activity but also the basis in the dictionary tend to be sparse, the formant structure of the spectral basis will suffer, and it degrades the performance. On the other hand, in the exemplar-based approach, only the activity becomes sparse because the dictionary is determined using exemplars and the activity is estimated using NMF. In the field of audio signal processing, Gemmeke *et al.* [8] proposed noise-robust automatic speech recognition using exemplar-based NMF. The disadvantage of this approach is the reconstruction error between $\mathbf{V}$ and $\mathbf{WH}$, which becomes larger than dictionary learning approach.

In recent years, exemplar-based NMF has been applied to Voice Conversion (VC) [9, 10]. VC is a technique for converting specific information in speech while maintaining the other information in the utterance. One of the most popular VC applications is speaker conversion [11]. In speaker conversion, a source speaker's voice individuality is changed to a specified target speaker's so that the input utterance sounds as though a specified target speaker had spoken it. VC is also being used for assistive technology [12], Text-To-Speech systems [13], spectrum restoring [14], and bandwidth extension for audio [15], etc.

The Gaussian Mixture Model (GMM)-based approach is widely used for VC because of its flexibility and good performance [11]. Toda *et al.* [16] introduced dynamic features and the Global Variance (GV) of the converted spectra over a time sequence. Helander *et al.* [17] proposed transforms based on Partial Least Squares (PLS), in order to prevent the over-fitting problem associated with standard multivariate regression.

The NMF-based approach has two advantages over conventional GMM-based VC methods. First, our approach results in a natural-sounding converted voice [18]. In statistical approaches, low-dimensional spectral features (for example mel-cepstrum) are used in order to avoid the "curse of dimensionality". In our NMF-based approach, high-dimensional spectra can be used because our approach is a non-statistical one. Second, our NMF-based VC method is noise robust [19]. The noise exemplars, which are extracted from the before- and after-utterance sections in the observed signal, are used as the noise dictionary, and the VC process is combined with NMF-based noise reduction.

However, because the conventional NMF-based approach employs exemplar-based NMF, the reconstruction error tends to be large, and we assume it results in "muffling effect". In order to enhance the conversion quality of NMF-based VC, we propose parallel dictionary learning using Discriminative Graph-embedded Non-negative Matrix Factorization (DGNMF). Source and target dictionaries are estimated under parallel constraint for activity between the source and target-training spectra. We introduce phoneme label-based discrimination and between-class constraint to Graph Regularized NMF (GRNMF) [20] in order to conduct discriminative learning and prevent over-fitting. Subjective evaluations show that our proposed method effectively enhanced NMF-based VC and alleviate the "muffling effect".

The rest of this paper is organized as follows: In Section 2, VC using exemplar-based NMF is described. In Section 3, our proposed method is described. In Section 4, the summary of our algorithm is described. In Section 5, the experimental data are evaluated, and the final section is devoted to our conclusions.

http://dx.doi.org/10.21437/Interspeech.2016-227

2. NMF-based Voice Conversion

2.1. Basic idea

Fig. 1 shows the basic approach of our exemplar-based VC, where I, J, and K represent the numbers of dimensions, frames, and bases, respectively. Our VC method needs two dictionaries that are phonemically parallel. $\mathbf{W}^s$ represents a source dictionary that consists of the source speaker's exemplars and $\mathbf{W}^t$ represents a target dictionary that consists of the target speaker's exemplars. These two dictionaries consist of the same words and are aligned with dynamic time warping (DTW) just as conventional GMM-based VC is. Hence, these dictionaries have the same number of bases. In this VC method, all the frames from the parallel training data are used as exemplars.

$\mathbf{W}^s$ and $\mathbf{W}^t$ are determined using parallel exemplars, and the source speaker's activity $\mathbf{H}^s$ is estimated using NMF. The cost function of NMF is defined as follows:

$$d_{KL}(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \;\; s.t. \;\; \mathbf{H}^s \geq 0 \tag{2}$$

In (2), the first term is the Kullback-Leibler (KL) divergence between $\mathbf{V}^s$ and $\mathbf{W}^s\mathbf{H}^s$ and the second term is the sparsity constraint with the L1-norm regularization term that causes the activity matrix to be sparse. λ represents the weight of the sparsity constraint. This function is minimized by iteratively updating the following equation.

$$\begin{aligned} \mathbf{H}^s \;\; \leftarrow \;\; & \mathbf{H}^s .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s./(\mathbf{W}^s\mathbf{H}^s))) \\ & ./(\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{(I\times J)} + \lambda\mathbf{1}^{(K\times J)}) \end{aligned} \tag{3}$$

$.*$, $./$ and $\mathbf{1}$ denote element-wise multiplication, division and all-one matrix, respectively. In this sense, the input spectra are represented by a linear combination of a small number of bases and the weights are estimated as activity.

This method assumes that when the source signal and the target signal (which are the same words but spoken by different speakers) are expressed with sparse representations of the source dictionary and the target dictionary, respectively, the obtained activity matrices are approximately equivalent. The estimated source activity $\mathbf{H}^s$ is multiplied to the target dictionary $\mathbf{W}^t$ and the target spectra $\hat{\mathbf{V}}^t$ are constructed.

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{4}$$

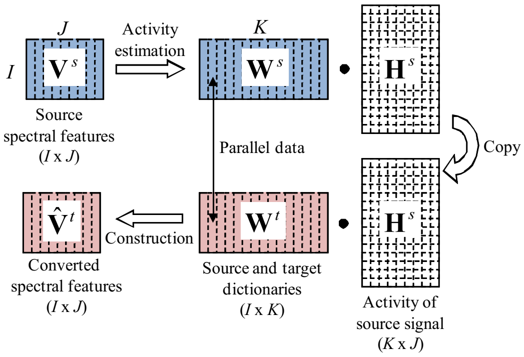

Figure 1: Basic approach of NMF-based voice conversion

2.2. Problems

In the case of exemplar-based NMF, the reconstruction error between $\mathbf{V}^s$ and $\mathbf{W}^s\mathbf{H}^s$ tends to be large (this point has been proven in experiments). In order to reduce the reconstruction error, a more compact dictionary needs to be estimated from the exemplars. Compact dictionary can reduce the computational times for NMF-based VC.

Moreover, in the NMF-based approach, input spectra are estimated from the source dictionary, and the converted spectra are constructed from the target dictionary. Fig. 2 shows an example of the activity matrices estimated from a single parallel Japanese word, which was uttered by a male and also by a female. These words are aligned using DTW in advance, and the parallel dictionaries, which consist of 250 bases, are used in activity estimation. As shown in the figure, the estimated activities are different, although the input features and dictionaries are parallel. We assume there are two reasons for this. First, we assume that the alignment difference between the source and the target dictionaries causes this effect. Although the parallel dictionaries are aligned by DTW, there still seems to be a mismatch of alignment. This mismatch degrades the performance of exemplar-based VC [18]. Second, we assume that the activity matrix contains not only phonetic information but also speaker information. In [21], we proposed a framework for dealing with this effect and improved the performance of NMF-based VC. However, a large amount of parallel adaptive data is needed when using this framework.

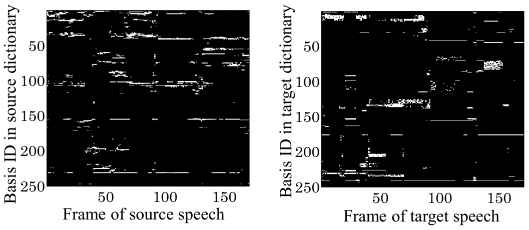

Figure 2: Activity matrices for parallel utterances

3. Discriminative Parallel Dictionary Learning for Voice Conversion

3.1. Discriminative Graph-embedded Non-negative Matrix Factorization

Because conventional dictionary learning using NMF is conducted on KL-divergence, it fails to discover the discriminative structure of the training data. We introduce discriminative constraint using graphs that are estimated from phoneme-labeled training data.

Adjacency matrices of a within-class scatter graph and a between-class scatter graph of training data are defined as follows:

$$\mathbf{A}^w_{ij} = \begin{cases} 1 & \begin{pmatrix} \mathbf{v}_i \in N_{k_w}(\mathbf{v}_i) \; or \; \mathbf{v}_j \in N_{k_w}(\mathbf{v}_j) \\ and \\ c_i = c_j \end{pmatrix} \\ 0 & (otherwise) \end{cases} \tag{5}$$

$$\mathbf{A}^b_{ij} = \begin{cases} 1 & \begin{pmatrix} \mathbf{v}_i \in N_{k_b}(\mathbf{v}_i) \; or \; \mathbf{v}_j \in N_{k_b}(\mathbf{v}_j) \\ and \\ c_i \neq c_j \end{pmatrix} \\ 0 & (otherwise) \end{cases} \tag{6}$$

where $N_{k_w}(\mathbf{v}_i)$ and $N_{k_b}(\mathbf{v}_i)$ denote the set of k_w nearest neighbors of $\mathbf{v}_i$ in the within-class scatter graph and k_b nearest neighbors of $\mathbf{v}_i$ in the between-class scatter graph. c_i and c_j denote the phoneme label of $\mathbf{v}_i$ and $\mathbf{v}_j$. Using adjacency matrices, graph Laplacians of within-class scatter and between-class scatter are defined as follows:

$$\mathbf{L}^w = \mathbf{D}^w - \mathbf{A}^w \quad (7)$$

$$\mathbf{L}^b = \mathbf{D}^b - \mathbf{A}^b \quad (8)$$

where $\mathbf{D}^w$ and $\mathbf{D}^b$ denote the diagonal column (or row) sum of $\mathbf{A}^w$ and $\mathbf{A}^b$.

Using graph Laplacians, the cost function of DGNMF is defined as follows:

$$d_{KL}(\mathbf{V}, \mathbf{WH}) + \frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{\mathsf{T}}\mathbf{L}^w\mathbf{H}) - \frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{\mathsf{T}}\mathbf{L}^b\mathbf{H}) \quad s.t.\ \ \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (9)$$

where $\mathbf{Tr}$, ϕ, and ψ denote the trace of matrix, weight of within-class scatter, and between-class scatter, respectively. In (9), the first term is the KL divergence between $\mathbf{V}$ and $\mathbf{WH}$, the second term is the within-class locality, and the third term is the between-class locality.

This function is minimized by iteratively updating the following equation.

$$\mathbf{W} \leftarrow \mathbf{W}.*((\mathbf{V}./(\mathbf{WH}))\mathbf{H}^{\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{\mathsf{T}}) \quad (10)$$

$$\mathbf{H} \leftarrow \frac{-\beta + \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha} \quad (11)$$

$$\alpha = ((\phi\mathbf{D}^w + \psi\mathbf{A}^b)\mathbf{H})./\mathbf{H} \quad (12)$$

$$\beta = \mathbf{W}^{\mathsf{T}}\mathbf{1}^{(I\times J)} - \mathbf{H}(\phi\mathbf{A}^w + \psi\mathbf{D}^b) \quad (13)$$

$$\gamma = \mathbf{H}.*(\mathbf{W}^{\mathsf{T}}(\mathbf{V}./(\mathbf{WH}))) \quad (14)$$

These equations are derived as the same manner as GNMF [20]. However, GNMF does not employ label-based discrimination and between-class locality. Therefore, DGNMF and GNMF are totally different approaches.

3.2. Parallel Dictionary Learning Using DGNMF

In order to construct a compact and discriminative dictionary, a parallel dictionary between the source and target speakers is estimated by parallel-constrained DGNMF. Fig. 3 shows the approach of our parallel dictionary learning. The objective function is represented as follows:

$$\begin{aligned} & d_{KL}(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_{KL}(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) \\ & + \frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{s\mathsf{T}}\mathbf{L}^{sw}\mathbf{H}^s) + \frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{t\mathsf{T}}\mathbf{L}^{tw}\mathbf{H}^t) \\ & - \frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{s\mathsf{T}}\mathbf{L}^{sb}\mathbf{H}^s) - \frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{t\mathsf{T}}\mathbf{L}^{tb}\mathbf{H}^t) \\ & + \lambda||\mathbf{H}^s||_1 + \lambda||\mathbf{H}^t||_1 + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 \\ & s.t.\ \ \mathbf{W}^s \geq 0, \mathbf{H}^s \geq 0, \mathbf{W}^t \geq 0, \mathbf{H}^t \geq 0 \quad (15) \end{aligned}$$

where $\mathbf{V}^s$, $\mathbf{V}^t$, $\mathbf{W}^s$, $\mathbf{W}^t$, $\mathbf{H}^s$, and $\mathbf{H}^t$ denote the source exemplars, the target exemplars, the source dictionary, the target dictionary, the source activity, and the target activity, respectively. $\mathbf{L}^{sw}$, $\mathbf{L}^{sb}$, $\mathbf{L}^{tw}$, and $\mathbf{L}^{tb}$ denote graph Laplacian of within-class scatter of source exemplars, between-class scatter graph of source exemplars, within-scatter graph of target exemplars, and between-class scatter graph of target exemplars, respectively. The source and target exemplars are aligned by DTW so that they have the same number of frames. ϵ and λ represent the parallel constraint weight and a the sparsity constraint weight. In (15), the first term to the sixth term are extended from (9). The seventh and eighth terms are sparsity constraint for $\mathbf{H}^s$ and $\mathbf{H}^t$, and the last term is the parallel constraint between $\mathbf{H}^s$ and $\mathbf{H}^t$.

This function is minimized by iteratively updating the following equation.

$$\mathbf{W}^s \leftarrow \mathbf{W}^s.*((\mathbf{V}^s./(\mathbf{W}^s\mathbf{H}^s))\mathbf{H}^{s\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{s\mathsf{T}}) \quad (16)$$

$$\mathbf{W}^t \leftarrow \mathbf{W}^t.*((\mathbf{V}^t./(\mathbf{W}^t\mathbf{H}^t))\mathbf{H}^{t\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{t\mathsf{T}}) \quad (17)$$

$$\mathbf{H}^s \leftarrow \frac{-\beta^s + \sqrt{\beta^{s2} + 4\alpha^s\gamma^s}}{2\alpha^s} \quad (18)$$

$$\alpha^s = ((\phi\mathbf{D}^{sw} + \psi\mathbf{A}^{sb})\mathbf{H}^s)./\mathbf{H}^s + \epsilon \quad (19)$$

$$\beta^s = \mathbf{W}^{s\mathsf{T}}\mathbf{1}^{(I\times J)} - \mathbf{H}^s(\phi\mathbf{A}^{sw} + \psi\mathbf{D}^{sb}) - \epsilon\mathbf{H}^t + \lambda \quad (20)$$

$$\gamma^s = \mathbf{H}^s.*(\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s./(\mathbf{W}^s\mathbf{H}^s))) \quad (21)$$

$$\mathbf{H}^t \leftarrow \frac{-\beta^t + \sqrt{\beta^{t2} + 4\alpha^t\gamma^t}}{2\alpha^t} \quad (22)$$

$$\alpha^t = ((\phi\mathbf{D}^{tw} + \psi\mathbf{A}^{tb})\mathbf{H}^t)./\mathbf{H}^t + \epsilon \quad (23)$$

$$\beta^t = \mathbf{W}^{t\mathsf{T}}\mathbf{1}^{(I\times J)} - \mathbf{H}^t(\phi\mathbf{A}^{tw} + \psi\mathbf{D}^{tb}) - \epsilon\mathbf{H}^s + \lambda \quad (24)$$

$$\gamma^t = \mathbf{H}^t.*(\mathbf{W}^{t\mathsf{T}}(\mathbf{V}^t./(\mathbf{W}^t\mathbf{H}^t))) \quad (25)$$

where $\mathbf{A}^{sw}$ and $\mathbf{A}^{sb}$ denote adjacency matrices of within the scatter graph and between the scatter of the source training data, and $\mathbf{D}^{sw}$ and $\mathbf{D}^{sb}$ denote their diagonal column (or row) sums. $\mathbf{A}^{tw}$, $\mathbf{A}^{tb}$, $\mathbf{D}^{tw}$, and $\mathbf{D}^{tb}$ are those for the target training data.

After discriminative parallel dictionaries are estimated, the input source speaker's spectra are converted using the dictionaries in the same manner described in Section 2.

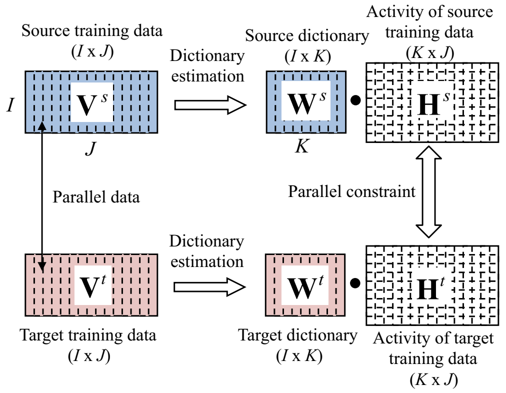

Figure 3: Parallel dictionary learning

4. Experimental Results

4.1. Experimental Conditions

The proposed VC technique was evaluated by comparing it with the conventional exemplar-based method [22] and the conventional GMM-based method in a speaker-conversion task using clean speech data. The source speaker and target speaker were one male and one female speaker, respectively, whose speech is stored in the ATR Japanese speech database [23]. The sampling rate was 12 kHz. Fifty sentences were used for training and another 50 sentences were used for testing. In our proposed method, $\epsilon, \phi, \psi, k_w, k_b, \lambda$ are set to be 40, 1, 10, 1024, 8192, and 0.1, respectively. The maximum number of NMF iterations is set to 10 for dictionary learning and 300 for conversion. Those parameters are chosen experimentally.

In the proposed and conventional GMM-based methods, mel-cepstrum + Δ is used as a spectral feature. Its number of dimensions is 48. In the NMF-based method (including our proposed method), the dimension number of the spectral feature is 1,539. It consists of a 513-dimensional STRAIGHT spectrum [24] and its consecutive frames (the frame coming before and the frame coming after). The number of Gaussian mixtures in the GMM-based method was set to 64, which is experimentally selected.

In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation [16]. The other information, such as aperiodic components, is synthesized without any conversion.

4.2. Results and Discussion

Objective tests were carried out using Mel-cepstrum distortion (MelCD) [dB] as follows: [16]

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \qquad (26)$$

where mc_d^{conv} and mc_d^{tar} denote the d-th dimension of the converted and target mel-cepstra.

Table 1 shows the MelCD and computational times for each method. Conv. denotes MelCD between the target and converted features. Recon. denotes MelCD between the source and reconstructed source features (mel-cepstra of $\mathbf{W}^s\mathbf{H}^s$ in (2)), which indicates the matrix factorization reconstruction eror. The number in brackets denotes the number of parallel dictionary bases. PDL denotes parallel dictionary learning (the same condition as $\phi = 0, \psi = 0$ in DGNMF).

In conventional NMF, the converted distortion using all bases and 5,000 bases is not significant although their reconstruction error is significantly small when we use all the bases. We assume that the activity mismatch problem, which we discussed in Section 2.2, degrades the conversion performance of NMF (all). The converted distortion between PDL (1,000) and NMF (1,000) is not significant. Moreover, the reconstruction error increased when we adapted PDL. On the other hand, our proposed method decreased not only the converted distortion but also the reconstruction error from NMF (1,000) and NMF (5,000). These results shows that our proposed method effectively enhanced the performance of the conventional NMF-based VC method.

The computational time is reduced as the number of bases becomes small. We assume this point to be the advantage of the dictionary-learning approach.

The subjective evaluation was conducted on "speech quality" and "similarity to the target speaker (individuality)". For the subjective evaluation, 25 sentences were evaluated by 10 Japanese speakers. For the evaluation on speech quality, we performed a Mean Opinion Score (MOS) test [25]. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For the similarity evaluation, a XAB test was carried out. In the XAB test, each subject listened to the voice of the target speaker. Then the subject listened to the voice converted by the two methods and selected which sample sounded most similar to the target speaker's voice. In NMF-based VC (including our proposed method), the number of bases is set at 5,000.

Fig. 4 shows the results of the subjective evaluation on speech quality. The MOS of our proposed method is better than that of the conventional NMF-based method and the GMM-based method. (The result is confirmed with the p-value test result of 0.05.) These result shows that our proposed method effectively alleviate the "muffling effect" in the NMF-based VC method. The difference in the results between the NMF-based method and the GMM-based method were not significant. We assume this is because our proposed method is significantly better than the other two methods. For this reason, the difference in MOS between NMF-based VC and GMM-based VC became relatively insignificant.

Fig. 5 shows the results of the subjective evaluation on individuality. The difference between our proposed method and the conventional NMF-based method is significant in the p-value test result of in 0.05. This result indicates that our proposed method enhanced the conversion quality of the NMF-based method. The difference between our proposed method and the conventional GMM-based method is not significant.

Table 1: MelCD and computational times of each method

	Conv. [dB]	Recon. [dB]	times [s]
GMM	2.96	-	2
NMF (all)	3.05	1.37	890
NMF (10,000)	3.11	1.59	680
NMF (5,000)	3.05	1.56	310
NMF (1,000)	3.11	1.59	71
PDL (1,000)	3.11	1.85	71
DGNMF (1,000)	3.08	1.09	71

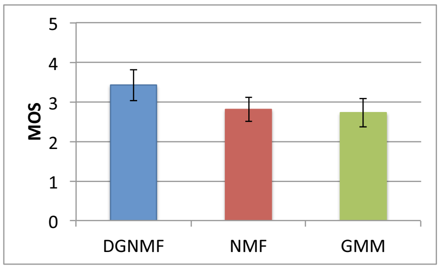

Figure 4: MOS test on speech quality

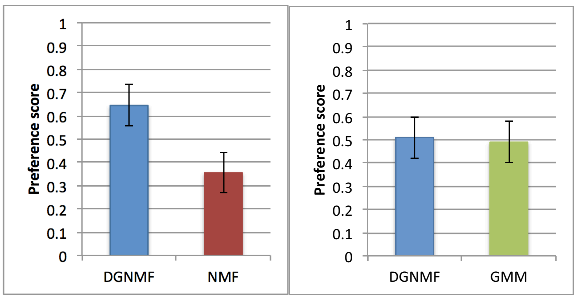

Figure 5: Preference score on individuality

5. Conclusions

In order to enhance the conversion performance of conventional exemplar-based VC using NMF, we proposed a dictionary-learning method using parallel constraint DGNMF. Our proposed DGNMF introduced phoneme-labeled discriminative learning on GNMF [20] and prevented over-fitting, which often occurs in dictionary-learning NMF. The parallel dictionary is estimated by using parallel-constraint DGNMF, and it enhanced the naturalness of the converted voice. As a result of dictionary learning, the computational times of the NMF-based VC method are also reduced. For our future work, we will adopt this method to NMF-based many-to-many VC [26]. We also assume our method can easily be adapted to hyperspectral imaging [2] or topic modeling [3].

6. References

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Neural Information Processing System*, pp. 556–562, 2001.

[2] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate non-negative matrix factorization," *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[3] T. Hofmann, "Probabilistic latent semantic indexing," *in Proc. SIGIR*, pp. 50–57, 1999.

[4] A. Cichocki, R. Zdnek, A. H. Phan, and S. Amari, *Non-negative Matrix and Tensor Factorization*. WILKEY, 2009.

[5] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," *in Proc. Interspeech*, 2006.

[6] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[7] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of itakura-saito nonnegative matrix factorization," *in Proc. ICASSP*, pp. 261–264, 2012.

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

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," *in Proc. SLT*, pp. 313–317, 2012.

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

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

[12] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[13] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *in Proc. ICASSP, vol. 1, pp. 285-288*, 1998.

[14] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "A mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech," *in Proc. Interspeech*, pp. 2494–2498, 2014.

[15] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "GMM-based bandwidth extension using sub-band basis spectrum model," *in Proc. Interspeech*, pp. 2489–2493, 2014.

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

[17] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5, pp. 912-921*, 2010.

[18] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," *in Proc. ICASSP*, pp. 7944–7948, 2014.

[19] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 6, pp. 1411–1418, 2014.

[20] D. Cai, X. He, and T. S. H. J. Han, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, 2010.

[21] R. Aihara, T. Takiguchi, and Y. Ariki, "Activity-mapping non-negative matrix factorization for exemplar-based voice conversion," *in Proc. ICASSP*, pp. 4899–4903, 2015.

[22] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E96-A, no. 10, pp. 1946–1953, 2013.

[23] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357–363, 1990.

[24] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[25] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, 2003.

[26] R. Aihara, T. Takiguchi, and Y. Ariki, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.