SEMI-NON-NEGATIVE MATRIX FACTORIZATION USING ALTERNATING DIRECTION METHOD OF MULTIPLIERS FOR VOICE CONVERSION

Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI

Graduate School of System Informatics, Kobe University, Japan

ABSTRACT

Voice conversion (VC) is being widely researched in the field of speech processing because of increased interest in using such processing in applications such as personalized Text-To-Speech systems. A VC method using Non-negative Matrix Factorization (NMF) has been researched because of its natural sounding voice, however, huge memory usage and high computational times have been reported as problems. We present in this paper a new VC method using Semi-Non-negative Matrix Factorization (Semi-NMF) using the Alternating Direction Method of Multipliers (ADMM) in order to tackle the problems associated with NMF-based VC. Dictionary learning using Semi-NMF can create a compact dictionary, and ADMM enables faster convergence than conventional Semi-NMF. Experimental results show that our proposed method is 76 times faster than conventional NMF, and its conversion quality is almost the same as that of the conventional method.

***Index Terms*—** NMF, ADMM, Voice Conversion, Speech Synthesis, Sparse Representation

1. INTRODUCTION

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

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

NMF has been applied to hyperspectral imaging [2], topic modeling [3], and the analysis of brain data [4].

In the field of audio signal processing, NMF has been applied to single channel speech separation [5, 6] and music transcription [7]. Some approaches using NMF employ an exemplar-based sparse representation method, which determines the dictionary using exemplars and only estimates the activity. Gemmeke *et al.* [8] proposed noise robust automatic speech recognition using exemplar-based NMF.

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]. Over-smoothing and over-fitting problems have been reported [17] in statistical approaches because of statistical averaging and the large number of parameters. Because our approach is a non-statistical one, it should avoid the over-fitting problem. Second, our exemplar-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, NMF-based VC also suffers from a problem with high computational times. There are three major reasons:

1. **Using high-dimensional spectra:** Because of the non-negativity constraint, the NMF-based VC method cannot use mel-cepstrum or other low-dimensional features, which include a negative value.
2. **Large number of bases:** The NMF-based VC method uses all the spectra of the training data; therefore, the number of bases in the dictionary is huge.
3. **Poor optimization method:** NMF-based VC method employs multiplicative updates using the Majorization Minimization (MM) algorithm.

This paper tackles the problems using the following tactics:

1. **Using Semi-Non-negative Matrix Factorization (Semi-NMF),** which relaxes the non-negativity constraint in the dictionary and makes the use of mel-cepstrum possible.
2. **Dictionary learning:** Parallel dictionaries are estimated by using Semi-NMF, resulting in a smaller number of the bases in the dictionary.
3. **Using the Alternating Direction Method of Multipliers (ADMM) [20]**, which provides fast convergence to Semi-NMF.

Semi-NMF using the MM algorithm has been proposed in [21], but it has never been applied to VC as far as we know. Also, NMF using ADMM has been proposed in [22, 23], and we have expanded it to Semi-NMF in this paper.

The rest of this paper is organized as follows: In Section 2, two major prior works and their problems are 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.

2. PRIOR WORKS

2.1. NMF-based VC

2.1.1. 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 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 by using NMF. The cost function of NMF is defined as follows:

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

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.

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

$.*$, $./$ 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 $\hat{\mathbf{V}}^t$ is constructed.

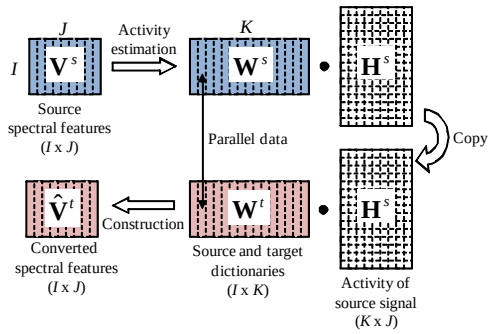

Fig. 1. Basic approach of NMF-based voice conversion

2.1.2. Problems

Because of the non-negative constraint in NMF, the usable features are restricted to the linear-spectrum, and we cannot use Δ or $\Delta\Delta$ features. In [18], we also used 513-demensional spectra and the consecutive frames. For this reason, the memory usage of this method is huge.

Also, in the NMF-based approach, the parallel dictionary consists of the parallel training data themselves. Activity is estimated from a dictionary that consists of a large number of bases and requires long running times. From these points of view, the NMF-based method is not applicable for practical use.

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, where one is uttered by a male and the other by a female. These words are aligned by using DTW in advance, and the parallel dictionaries, which consist of 250 bases, are used in activity estimation. As shown in the figure, 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 [24], we proposed a framework for solving this effect and improved the performance of NMF-based VC. However, a large amount of parallel adaptive data is needed when using this framework.

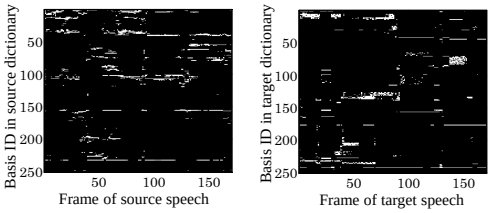

Fig. 2. Activity matrices for parallel utterances

2.2. Semi-NMF Using the Majorization Minimization Algorithm

2.2.1. Formulation

The cost function of Semi-NMF is defined as follows:

$$d_F(\mathbf{V}, \mathbf{WH}) + \lambda||\mathbf{H}||_1 \;\; s.t. \;\; \mathbf{H} \geq 0 \tag{4}$$

In (4), the first term is the Frobenius norm between $\mathbf{V}$ and $\mathbf{WH}$, and the second term is the sparsity constraint with the L1-norm regularization term that causes the activity matrix to be sparse. λ represents a weight of sparsity constraint. Because we relaxed the non-negativity constraint in $\mathbf{W}$, the cost function of Semi-NMF is restricted to the Frobenius norm or Euclidian distance. This function is minimized by iteratively updating the following equation;

$$\begin{aligned} \mathbf{H} \quad \leftarrow \quad & (-\lambda\mathbf{H}^{\mathsf{T}} + (\mathbf{H}^{\mathsf{T}} .* \sqrt{\lambda^2 + 16(\mathbf{A} .* \mathbf{B})})) \\ & ./(4\mathbf{A}) \end{aligned} \tag{5}$$

$$\mathbf{A}^{\mathsf{T}} \quad = \quad (\mathbf{V}^{\mathsf{T}}\mathbf{W})^{-} + (\mathbf{H}^{\mathsf{T}}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{+}) \tag{6}$$

$$\mathbf{B}^{\mathsf{T}} \quad = \quad (\mathbf{V}^{\mathsf{T}}\mathbf{W})^{+} + (\mathbf{H}^{\mathsf{T}}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-}) \tag{7}$$

where we separate the positive and negative parts of matrix $\mathbf{X}$ as $\mathbf{X}^+ = (|\mathbf{X}| + \mathbf{X})/2, \mathbf{X}^- = (|\mathbf{X}| - \mathbf{X})/2$.

2.2.2. The problem

Semi-NMF can decompose negative values. Therefore, we can use mel-cepstrum or Δ parameters. This function will decrease the memory usage compared to NMF. However, in the updating equation (7), there is a term that includes $\mathbf{W}^\mathsf{T}\mathbf{W}$. If we use a parallel dictionary, which consists of the parallel training data themselves, it requires a huge memory and a large number of computations. Therefore, we need a dictionary learning scheme for a Semi-NMF based VC method. Moreover, this Semi-NMF using the MM algorithm has slow convergence.

3. VOICE CONVERSION USING SEMI-NON-NEGATIVE MATRIX FACTORIZATION

3.1. Basic Idea

In order to solve the problems mentioned in the above section, we propose a new VC method using Semi-NMF based on ADMM. First, the source and target dictionaries are estimated by using parallel-constrained Semi-NMF. This parallel-constraint solves the activity gap problem that occurred in the conventional NMF-based VC method. Moreover, estimating a compact dictionary with this scheme can reduce the memory usage and computational times.

Input source spectra are decomposed into a linear combination from the basis from the estimated dictionary by using ADMM-based Semi-NMF. ADMM-based Semi-NMF enables fast convergence and estimation of sparse activity compared to Semi-NMF using the MM algorithm.

3.2. Dictionary Learning

In order to construct a compact dictionary, a parallel dictionary between the source and target speakers is estimated by parallel-constrained Semi-NMF using ADMM. The objective function is represented as follows:

$$\begin{aligned} min \quad & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_F(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) \\ & + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 + \lambda||\mathbf{H}^s||_1 + \lambda||\mathbf{H}^t||_1 \\ sub\ to \quad & \mathbf{H}^s = \mathbf{H}^s_+, \mathbf{H}^s_+ \geq 0, \mathbf{H}^t = \mathbf{H}^t_+, \mathbf{H}^t_+ \geq 0 \end{aligned} \tag{8}$$

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. The source and target exemplars are aligned by DTW so that they have the same number of frames. ϵ and λ represent a weight of parallel constraint and a weight of sparsity constraint. The augmented Lagrangian corresponding to (8) is as follows:

$$\begin{aligned} & L_\rho(\mathbf{W}^s, \mathbf{H}^s, \mathbf{W}^t, \mathbf{H}^t, \mathbf{H}^s_+, \mathbf{H}^t_+) = \\ & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_F(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 \\ & + \lambda||\mathbf{H}^s||_1 + \langle \alpha_{\mathbf{H}^s}, \mathbf{H}^s - \mathbf{H}^s_+ \rangle + \frac{\rho}{2}||\mathbf{H}^s - \mathbf{H}^s_+||_F^2 \\ & + \lambda||\mathbf{H}^t||_1 + \langle \alpha_{\mathbf{H}^t}, \mathbf{H}^t - \mathbf{H}^t_+ \rangle + \frac{\rho}{2}||\mathbf{H}^t - \mathbf{H}^t_+||_F^2 \end{aligned} \tag{9}$$

where ρ denotes the tuning parameter that controls the convergence rate. The updates alternately optimize (9) with respect to each of the four primal variables, followed by gradient ascent in each of the two dual variables. This is summarized below.

Table 1. Algorithm of Dictionary Learning

Input $\mathbf{V}^s, \mathbf{V}^t$
Initialize $\mathbf{W}^s, \mathbf{H}^s, \mathbf{W}^t, \mathbf{H}^t, \mathbf{H}^s_+, \mathbf{H}^t_+, \alpha_{\mathbf{H}^s}, \alpha_{\mathbf{H}^t}$
Repeat
$\mathbf{W}^s \leftarrow (\mathbf{V}^s(\mathbf{H}^{s\mathsf{T}}))/(\mathbf{H}^s(\mathbf{H}^{s\mathsf{T}}))$
$\mathbf{W}^t \leftarrow (\mathbf{V}^t(\mathbf{H}^{t\mathsf{T}}))/(\mathbf{H}^t(\mathbf{H}^{t\mathsf{T}}))$
$\mathbf{H}^s \leftarrow (2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s + (\rho+\epsilon)\mathbf{I})$
$\quad \backslash(2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s - \alpha_{\mathbf{H}_s} + \rho\mathbf{H}^s_+ + \epsilon\mathbf{H}^t - \lambda)$
$\mathbf{H}^t \leftarrow (2\mathbf{W}^{t\mathsf{T}}\mathbf{W}^t + (\rho+\epsilon)\mathbf{I})$
$\quad \backslash(2\mathbf{W}^{t\mathsf{T}}\mathbf{W}^t - \alpha_{\mathbf{H}_t} + \rho\mathbf{H}^t_+ + \epsilon\mathbf{H}^s - \lambda)$
$\mathbf{H}^s_+ \leftarrow \max(\mathbf{H}^s + \frac{1}{\rho}\alpha_{\mathbf{H}^s}, 0)$
$\mathbf{H}^t_+ \leftarrow \max(\mathbf{H}^t + \frac{1}{\rho}\alpha_{\mathbf{H}^t}, 0)$
$\alpha_{\mathbf{H}^s} \leftarrow \alpha_{\mathbf{H}_s} + \rho(\mathbf{H}^s - \mathbf{H}^s_+)$
$\alpha_{\mathbf{H}^t} \leftarrow \alpha_{\mathbf{H}_t} + \rho(\mathbf{H}^t - \mathbf{H}^t_+)$
Until convergence **return** $\mathbf{W}^s, \mathbf{H}^s_+, \mathbf{W}^t, \mathbf{H}^t_+$

3.3. Conversion

After a pair of parallel dictionaries, $\mathbf{W}^s$ and $\mathbf{W}^t$ is estimated, input source spectra $\mathbf{V}^s$ is converted to $\hat{\mathbf{V}}^t$ by using Semi-NMF based on ADMM. The objective is represented as follows:

$$\begin{aligned} min \quad & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \\ sub\ to \quad & \mathbf{H}^s = \mathbf{H}^s_+, \mathbf{H}^s_+ \geq 0. \end{aligned} \tag{10}$$

The augmented Lagrangian corresponding to (10) is as follows:

$$\begin{aligned} & L_\rho(\mathbf{W}^s, \mathbf{H}^s, \mathbf{H}^s_+) = \\ & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \\ & + \langle \alpha_{\mathbf{H}^s}, \mathbf{H}^s - \mathbf{H}^s_+ \rangle + \frac{\rho}{2}||\mathbf{H}^s - \mathbf{H}^s_+||_F^2 \end{aligned} \tag{11}$$

$\mathbf{W}^s$ is determined and $\mathbf{H}^s$ is estimated by using the following algorithm.

Table 2. Algorithm of Conversion

Input $\mathbf{V}^s, \mathbf{W}^s$
Initialize $\mathbf{H}^s, \mathbf{H}^s_+, \alpha_{\mathbf{H}^s}$
Repeat
$\mathbf{H}^s \leftarrow (2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s + \rho\mathbf{I})$
$\quad \backslash(2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s - \alpha_{\mathbf{H}_s} + \rho\mathbf{H}^s_+ - \lambda)$
$\mathbf{H}^s_+ \leftarrow \max(\mathbf{H}^s + \frac{1}{\rho}\alpha_{\mathbf{H}^s}, 0)$
$\alpha_{\mathbf{H}^s} \leftarrow \alpha_{\mathbf{H}_s} + \rho(\mathbf{H}^s - \mathbf{H}^s_+)$
Until convergence **return** $\mathbf{H}^s_+$

By using the estimated activity $\mathbf{H}^s$ and the target dictionary $\mathbf{W}^t$, the converted spectra $\hat{\mathbf{V}}^t$ is constructed as follows:

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{12}$$

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

The proposed VC technique was evaluated by comparing it with the conventional NMF-based method [25] 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 [26]. The sampling rate was 12 kHz. Two-hundred sixteen words were used for training and 50 sentences were used for testing. In our proposed method, ρ, ϵ, λ are set to be 1, 1, 0.1, respectively. The maximum number of iterations of Semi-NMF is set to 50 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, the dimension number of the spectral feature is 1,539. It consists of a 513-dimensional STRAIGHT spectrum [27] 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 128, 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

First, we tested the convergence speed between the proposed Semi-NMF method using ADMM and the Semi-NMF method using MM. The number of bases in the dictionary was set to 1,000. The convergence of the different algorithms is show in Fig. 3, where the x-axis shows the iteration and y-axis shows the convergence on a logarithmic scale. This figure shows that ADMM can produce faster convergence than MM when ρ is small.

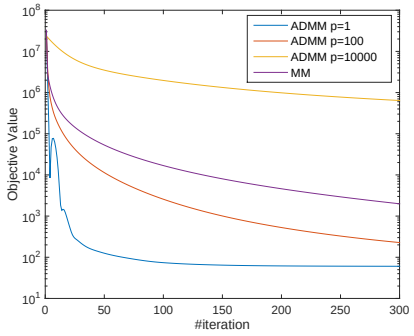

Fig. 3. Objective value as a function of iteration

Objective tests were carried out using the Normalized Spectrum Distortion (NSD) [28].

$$NSD = \sqrt{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2 / ||\mathbf{X}^t - \mathbf{X}^s||^2}, \quad (13)$$

where $\mathbf{X}^s$, $\mathbf{X}^t$, and $\hat{\mathbf{X}}^t$ denote the source, target, and converted spectrum, respectively. Table 3 shows the NSD and computational times for each method. In our proposed method, 1,000 or 5,000 bases are estimated. The distortion between our proposed method using 1,000 bases and 5,000 bases is not significant. Our proposed method is slightly worse than the NMF-based method, but the proposed method can reduce the computational times. The distortion between the proposed method and GMM-based method is not significant.

For the subjective evaluation, a total of 15 Japanese speakers took part in the test using headphones. We compared our proposed method (Proposed(1,000) in Table 3), an NMF-based VC method, and a GMM-based VC method. The left side of Fig. 4 shows the results of a MOS test on speech quality. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The difference between Semi-NMF and GMM, and NMF and GMM is marginally significant in the p-value test.

Table 3. NSD and computational times of each method

	NSD	times [s]
GMM	1.66	2
NMF	1.54	916
Proposed(1,000)	1.69	12
Proposed(5,000)	1.70	310

The right side of Fig. 4 shows the results of a DMOS test on similarity to the target speaker. The opinion score was set to a 5-point scale (5: very similar, 4: similar, 3: fair, 2: different, 1: very different). These differences are not significant.

Based on from these evaluations, the conversion quality of our proposed VC method was almost the same as the conventional NMF-based VC method, and we were able to effectively reduce the computational times and memory usage compared to the NMF-based VC method.

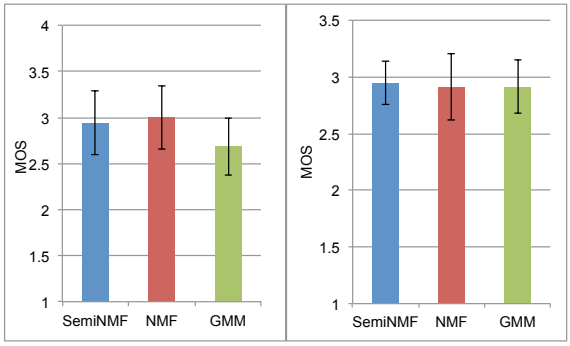

Fig. 4. MOS test on speech quality (left) and similarity (right)

5. CONCLUSIONS

This paper proposed a Semi-NMF-based VC method using ADMM. In order to reduce the computational times and memory usage, in the NMF-based VC method, NMF is replaced with Semi-NMF so that we can use compact spectral features. The convergence of conventional Semi-NMF using the MM algorithm is slow, and we proposed Semi-NMF using ADMM, which enables faster convergence and estimation of sparse activity. Also, we proposed a dictionary-learning scheme to estimate parallel compact dictionaries. We assume our method can easily adapt to hyperspectral imaging [2] or topic modeling [3].

Some problems still remain with our method. The proposed method requires longer running time than the GMM-based method. Wu *et al.* proposed a method for NMF-based VC to reduces the computational cost [10]. In future work, we will combine these methods and investigate the optimal number of bases for better performance.

In [29], we proposed a phoneme-categorized dictionary that enhances the performance of exemplar-based VC. We assume our proposed method can achieve better performance by combining it with this method. Also, we will apply our method to noisy environments and an assistive technology for people with articulation disorders.

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 nonnegative 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, *Nonnegative 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 nonnegative 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] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[21] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorization," *IEEE Trans. Pettern Analysis And Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.

[22] D. L. Sun and C. Févotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," *in Proc. ICASSP*, pp. 6242–6246, 2014.

[23] Y. Xu, W.Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Frontiers of Mathematics in China*, vol. 7, no. 2, pp. 365–384, 2012.

[24] 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.

[25] 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.

[26] 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.

[27] H. Kawahara, I. Masuda-Katsuse, and A.de 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.

[28] T. En-Najjary, O. Roec, and T. Chonavel, "A voice conversion method based on joint pitch and spectral envelope transformation," *in Proc. ICSLP*, pp. 199–203, 2004.

[29] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014:5, doi:10.1186/1687-4722-2014-5, 2014.