NOISE-ROBUST VOICE CONVERSION USING A SMALL PARALLEL DATA BASED ON NON-NEGATIVE MATRIX FACTORIZATION

Ryo Aihara[*], *Takao Fujii*[*], *Toru Nakashika*[†], *Tetsuya Takiguchi*[*], *Yasuo Ariki*[*]

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

[†] Graduate School of Information Systems
University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo, Japan

ABSTRACT

This paper presents a novel framework of voice conversion (VC) based on non-negative matrix factorization (NMF) using a small parallel corpus. In our previous work, a VC technique using NMF for noisy environments has been proposed, and it requires parallel exemplars (dictionary), which consist of the source exemplars and target exemplars, having the same texts uttered by the source and target speakers. The large parallel corpus is used to construct a conversion function in NMF-based VC (in the same way as common GMM-based VC). In this paper, an adaptation matrix in an NMF framework is introduced to adapt the source dictionary to the target dictionary. This adaptation matrix is estimated using a small parallel speech corpus only. The effectiveness of this method is confirmed by comparing its effectiveness with that of a conventional NMF-based method and a GMM-based method in a noisy environment.

Index Terms— voice conversion, speaker adaptation, noisy environments, small parallel corpus

1. INTRODUCTION

Voice conversion (VC) is a technique for converting speaker's voice individuality while maintaining phonetic information in the utterance. The VC techniques have been applied not only to speaker conversion but also emotion conversion [1], speaking aid [2], and so on.

Many statistical approaches to VC have been studied [3–5]. Among these approaches, the GMM-based mapping approach [5] is widely used, and a number of improvements have been proposed. Toda *et al.* [6] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander *et al.* [7] proposed transforms based on Partial Least Squares (PLS) in order to prevent the over-fitting problem of standard multivariate regression.

In our previous work [8], an exemplar-based VC approach for noisy source signals has been presented using Non-negative Matrix Factorization (NMF) [9]. We assume that our approach using NMF has two advantages compared to conventional statistical VC. The first advantage is the noise robustness. In statistical VC, the noise in the input signal is not only output with the converted signal, but may also degrade the conversion performance itself due to unexpected mapping of source features [8]. Because NMF can separate back-ground noise and speech spectra [10], our VC approach can avoid performance degradation in noisy environments. The second advantage is naturalness of the converted voice. In statistical VC, over-smoothing in converted voice is reported [7]. Because our approach is not a statistical one, we assume that our approach can avoid the over-fitting problem and create a natural voice.

However, a conventional VC approach needs parallel training data between source and target speakers and this constraint can be a difficult requirement to meet in practice. In statistical VC, some approaches that do not require parallel data have been proposed [11–14]. In this paper, we propose noise robust VC using small parallel corpus based on an NMF-based speaker adaptation technique.

In [16], adaptation of speaker-specific bases in NMF for single channel speech-music separation has been presented. The adaptation technique is applied to voice conversion in this paper. In VC, the source dictionary is constructed using sufficient data of the source speaker, and it is adapted using a small amount of parallel data only (about ten words only) in order to obtain the target dictionary, where a linear regression transformation matrix is trained based on NMF.

This rest of this paper is organized as follows. In section 2 a voice conversion technique based on NMF is described, and an adaptation technique in an NMF framework is described in section 3. Section 4 describes the results of experiments.

2. VOICE CONVERSION BASED ON NMF

2.1. Sparse Representations for Voice Conversion

Fig. 1 shows an approach of exemplar-based voice conversion in a noisy environment. D, L, and J are the numbers of dimensions, frames and exemplars, respectively. In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of bases. We call the collection of the bases as dictionary and

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

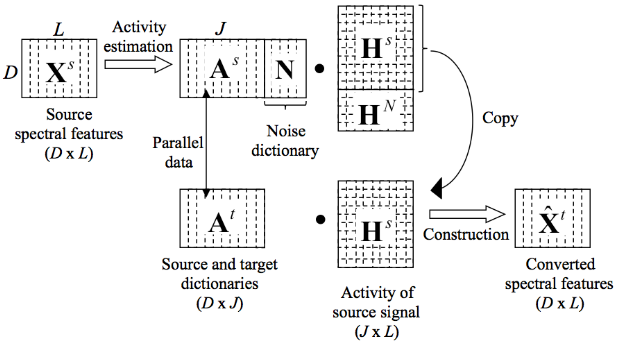

Fig. 1. Basic approach of exemplar-based voice conversion in a noisy environment

the stack of its weights as activities. From the before and after utterance sections in the observed signal, the noise dictionary is extracted for each utterance. The spectrum of the noisy source signal at frame l is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their activities.

$$\begin{aligned}
\mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\
&\approx \textstyle\sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s + \sum_{k=1}^{K} \mathbf{a}_k^n h_{k,l}^n \\
&= [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\
&= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \qquad (1)
\end{aligned}$$

$\mathbf{x}_l^s$ and $\mathbf{x}_l^n$ are the magnitude spectra of the source speaker's and the noise. $\mathbf{A}^s$, $\mathbf{A}^n$, $\mathbf{h}_l^s$, and $\mathbf{h}_l^n$ are the source dictionary, noise dictionary, and their activities at frame l. Given the spectrogram, (1) can be written as follows:

$$\begin{aligned}
\mathbf{X} &\approx [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\
&= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \qquad (2)
\end{aligned}$$

In order to consider only the shape of the spectrum, $\mathbf{X}$, $\mathbf{A}^s$ and $\mathbf{A}^n$ are first normalized for each frame or exemplar so that the sum of the magnitudes over frequency bins equals unity.

The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function [10]:

$$d(\mathbf{X}, \mathbf{A}\mathbf{H}) + ||(\lambda \mathbf{1}^{(1\times L)}) .* \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \qquad (3)$$

$.*$ denotes element-wise multiplication. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}$ and $\mathbf{A}\mathbf{H}$. The second term is the sparse constraint with L1-norm regularization term that causes $\mathbf{H}$ to be sparse. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$. In this paper, the weights for source exemplars $[\lambda_1 \ldots \lambda_J]$ were set to 0.1, and those for noise exemplars $[\lambda_{J+1} \ldots \lambda_{J+K}]$ were set to 0. $\mathbf{H}$ minimizing (3) is estimated iteratively applying the following update rule:

$$\begin{aligned}
\mathbf{H}_{n+1} &= \mathbf{H}_n .* (\mathbf{A}^{\mathsf{T}}(\mathbf{X}./(\mathbf{A}\mathbf{H}))) \\
&\quad ./(\mathbf{1}^{((J+K)\times L)} + \lambda \mathbf{1}^{(1\times L)}). \qquad (4)
\end{aligned}$$

2.2. Target speech construction

$\mathbf{A}^t$ in Fig. 1 represents a target dictionary that consists of the target speaker's exemplars. $\mathbf{A}^s$ and $\mathbf{A}^t$ consisted of the same words and are aligned with dynamic time warping (DTW) just as conventional GMM-based VC is. 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. For this reason, we assume that when there are parallel dictionaries, the activity of the source features estimated with the source dictionary may be able to be substituted with that of the target features.

The target dictionary is also normalized for each frame in the same way the source dictionary was. From the estimated joint matrix $\mathbf{H}$, the activity of source signal $\mathbf{H}^s$ is extracted, and by using the activity and the target dictionary, the converted spectral features $\hat{\mathbf{X}}^t$ are constructed.

$$\hat{\mathbf{X}}^t = \mathbf{A}^t \mathbf{H}^s \qquad (5)$$

The converted spectral features are de-normalized so that the sum of the magnitudes over frequency bins equals input spectral features.

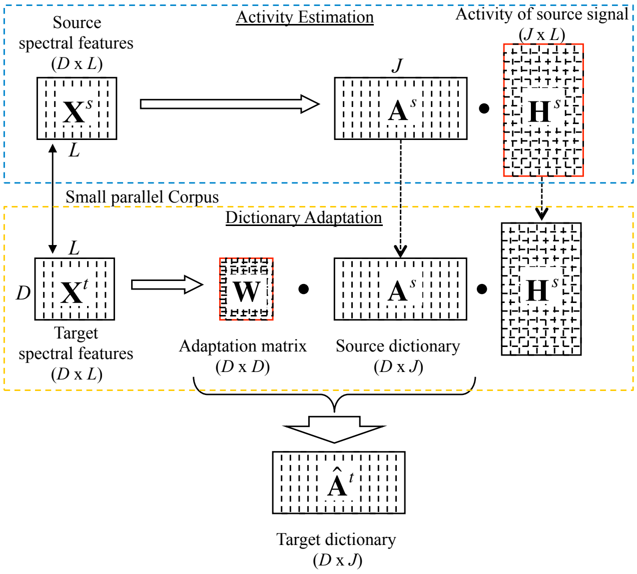

Fig. 2. Estimation of parallel dictionary using a speaker transformation matrix

3. ESTIMATION OF TARGET DICTIONARY USING A SMALL PARALLEL CORPUS ONLY

In the framework of conventional NMF-based VC which is described in section 2, large parallel corpus of source and target speaker is needed for dictionary construction. In this section, we propose a target dictionary estimation from small parallel corpus only.

Fig. 2 shows the estimation procedure of our proposed method. $\mathbf{X}^s$ and $\mathbf{X}^t$ show the small parallel data between source and target speakers. In the Activity Estimation stage, a source spectral exemplar matrix $\mathbf{X}^s$ is decomposed into a linear combination of bases from the source dictionary $\mathbf{A}^s$. The source dictionary is consist of source speaker's exemplars. The indexes and weights of the bases are estimated using (4) as source activity $\mathbf{H}^s$.

In the Dictionary Adaptation stage, speaker adaptation is conducted in order to obtain a target dictionary from a source dictionary using a small amount of (parallel) target speech signals. The adaptation is performed by using a linear regression transformation matrix based on an NMF framework. Given the transformation matrix, $\mathbf{W}$, the target feature vector at the l-th frame,

$$\mathbf{x}_l^t \approx \mathbf{W}\mathbf{A}^s\mathbf{h}_l^s \tag{6}$$

where $\mathbf{A}^s$ is the source dictionary and $\mathbf{h}_l^s$ is the activity vector of source signal at the l-th frame.

In order to find the transformation matrix, an NMF framework which minimizes the KL divergence between $\mathbf{X}^t$ and $\mathbf{W}\mathbf{A}^s\mathbf{H}^s$ is used.

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}}\, d(\mathbf{X}^t, \mathbf{W}\mathbf{A}^s\mathbf{H}^s) \tag{7}$$

The transformation matrix, $\mathbf{W}$, is estimated using $\mathbf{A}^s$, $\mathbf{H}^s$, and a small amount of the parallel target speech signals, $\mathbf{X}^t$, as follows:

$$\begin{aligned}\mathbf{W} \;\leftarrow\;& \mathbf{W}.*((\mathbf{X}^t./(\mathbf{W}(\mathbf{A}^s\mathbf{H}^s)))(\mathbf{A}^s\mathbf{H}^s)^\mathsf{T})\\ & ./(\mathbf{1}^{(D\times L)}(\mathbf{A}^s\mathbf{H}^s)^\mathsf{T}).\end{aligned} \tag{8}$$

The new parallel target dictionary is given by $\hat{\mathbf{A}}^t = \mathbf{W}\mathbf{A}^s$ and the source features are converted into the target features according to section 2.

4. EXPERIMENTS

4.1. Experimental Conditions

The new VC technique was evaluated by comparing it with conventional techniques based on GMM [5] and NMF [8] in a speaker conversion task using noisy speech data. Two males and two females are selected as source or target speakers from the ATR Japanese speech database and we conducted male-to-female, male-to-make and female-to-female conversion. The sampling rate was 8 kHz.

Two-hundred sixteen words of clean speech were used to construct a source dictionary in NMF with speaker adaptation and used to train the GMM in the conventional method. The number of adaptation words was 10, 25, and 50, respectively. 50 words for test are different from those used in training and adaptation.

The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database [17]) to the clean speech data. The SNR was 20 and 10 dB. The average number of exemplars in the noise dictionary for each utterance was 104.

In the NMF-based method, a 512-dimensional spectrum was used as the feature vectors for input signal and source dictionary. The number of iterations used to estimate the activity was 300. In the GMM-based method, 40 linear-cepstral coefficients obtained from the STRAIGHT [15] spectrum were used as the feature vectors. The number of Gaussian mixtures was 64. In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation [6]. The other information, such as aperiodic components, is synthesized without any conversion.

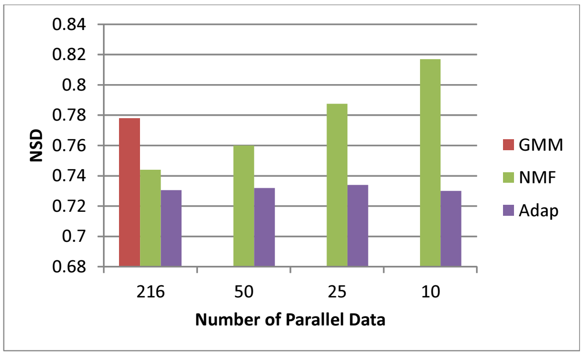

Fig. 3. NSD for male-to-female voices at 10dB converted by each method

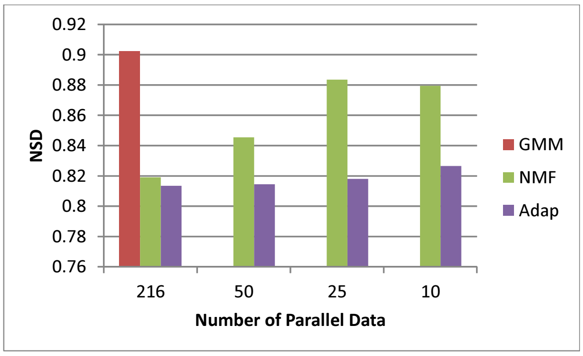

Fig. 4. NSD for male-to-male voices at 10dB converted by each method

4.2. Experimental Results

Objective tests are carried out using the normalized spectrum distortion.

$$NSD = \sqrt{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2/||\mathbf{X}^t - \mathbf{X}^s||^2} \tag{9}$$

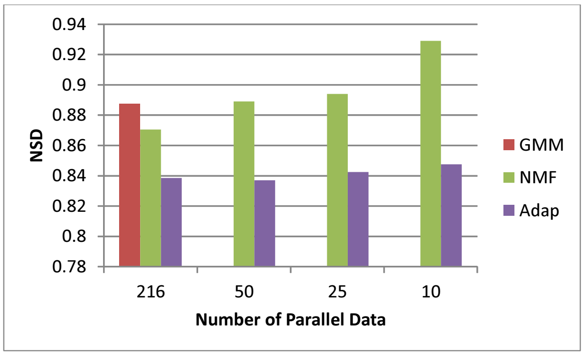

Fig. 5. NSD for female-to-female voices at 10dB converted by each method

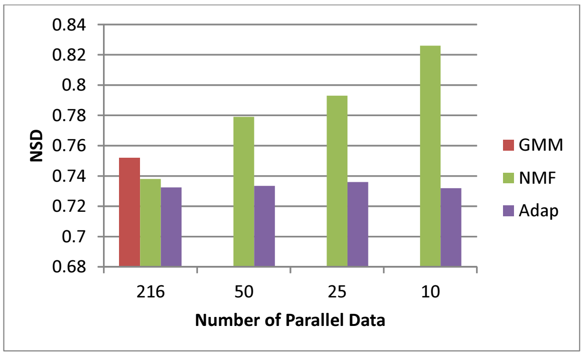

Fig. 6. NSD for male-to-female voices at 20dB converted by each method

Figs. 3, 4, and 5 show the NSD for a male-to-female conversion, a male-to-male conversion, and a male-to-female conversion in 10 dB, respectively. "NMF" shows the result using the conventional NMF without speaker adaptation and "Adap" shows the result using NMF with speaker adaptation. As shown in these figures, the performance of NMF without speaker adaptation decreases as the number of words used for the parallel dictionaries decreases. On the other hand, the performance of NMF with speaker adaptation does not decrease in comparison with the conventional NMF without speaker adaptation.

Figs. 6, 7, and 8 show the NSD for a male-to-female conversion, a male-to-male conversion, and a female-to-female conversion in 20 dB, respectively. Because of low SNR condition, the effectiveness of noise robustness of NMF-based VC is lower, compared to Figs. 3, 4, and 5. However, the performance of NMF with speaker adaptation does not decrease as the number of words used for the parallel dictionaries decreases in comparison with the conventional NMF without speaker adaptation. Moreover, the performance of NMF with speaker adaptation is better than conventional GMM-based VC. These results shows the effectiveness of our NMF-based speaker adaptation technique.

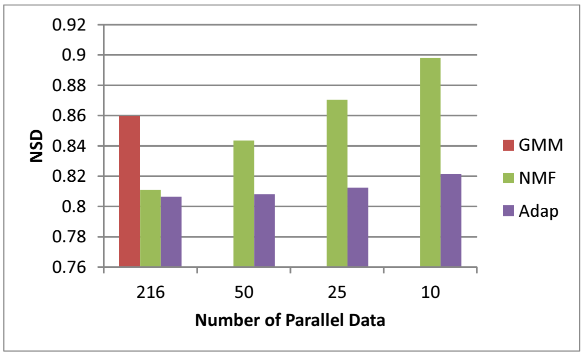

Fig. 7. NSD for male-to-male voices at 20dB converted by each method

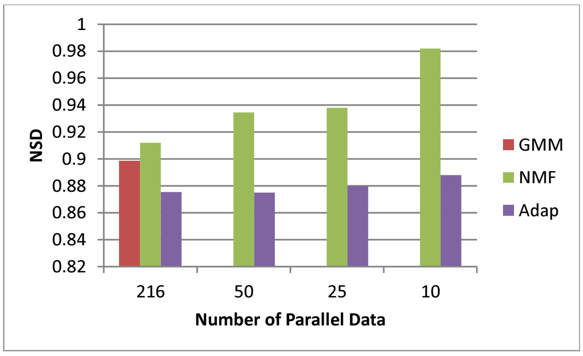

Fig. 8. NSD for female-to-female voices at 20dB converted by each method

For the speech quality evaluation, a mean opinion score (MOS) test was performed. The opinion score was set to a five-point scale (5 excellent, 4 good, 3 fair, 2 poor, 1 bad). The number of subjects was 8 and the SNR was 10 dB. Fig. 9 (left side) shows the MOS test on the speech quality. As shown in Fig. 9, NMF-based VC with speaker adaptation (25 adaptation words) obtained a better score than the conventional NMF-based VC (25 words). The result was confirmed by a p-value test of 0.05.

For the evaluation of speaker individuality, the XAB test was carried out. In the XAB test, each subject listened to the target speech. Then the subject listened to the speech converted by the two methods and selected which sample sounded more similar to the target speech. Fig. 9 (right side) shows the NMF-based VC with speaker adaptation obtained a higher score than the conventional NMF-based VC without speaker adaptation. The result was confirmed by a p-value test of 0.05.

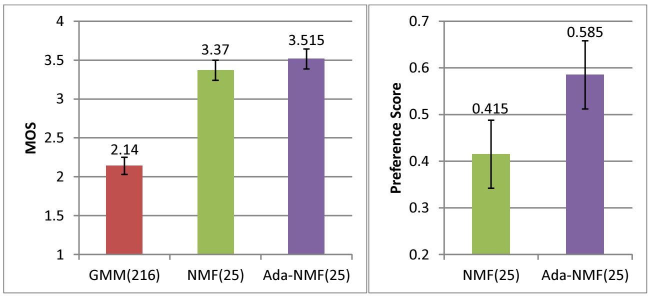

Fig. 9. Results of MOS test and XAB test

5. CONCLUSIONS

In this paper, an exemplar-based VC technique using speaker adaptation was presented. This method uses a small amount of parallel data only, where a linear regression transformation matrix is used to adapt a source dictionary to a target dictionary and it is estimated in an NMF framework.

In comparison experiments between GMM-based VC, NMF without speaker adaptation and NMF with speaker adaptation, the NMF-based VC with speaker adaptation showed better performance. Future work will include efforts to apply this method to other exemplar-based VC applications [18–20].

REFERENCES

[1] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in Proc. INTERSPEECH, pp. 2765–2768, 2011.

[2] 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.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Vice conversion through vector quantization," in Proc. ICASSP, pp. 655–658, 1988.

[4] H. Valbret, E. Moulines and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, Vol. 11, No. 2-3, pp. 175–187, 1992.

[5] 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.

[6] 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.

[7] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," IEEE Trans. Audo, Speech, Lang. Process., Vol. 18, No. 5, pp. 912–921, 2010.

[8] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-Based Voice Conversion in Noisy Environment," in Proc. IEEE Workshop on Spoken Language Technology, pp. 313-317, 2012.

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

[10] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," IEEE Trans. Audio, Speech, Lang. Process., Vol. 19, Issue 7, pp. 2067–2080, 2011.

[11] C. H. Lee and C. H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in Proc. INTERSPEECH, pp. 2254–2257, 2006.

[12] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in Proc. INTERSPEECH, pp. 2446–2449, 2006.

[13] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in Proc. INTERSPEECH, pp. 653–656, 2011.

[14] A. Mouchtaris, J. Van der Spiegel and P. Mueller "Non-parallel training for voice conversion based on a parameter adaptation approach", IEEE Trans. Audio, Speech, Lang. Process., Vol. 14, Issue 3, 952-963, 2006.

[15] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, "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, pp. 187–207, 1999.

[16] E. M. Grais and H. Erdogan, "Adaptation of speaker-specic bases in non-negative matrix factorization for single channel speech-music separation," in Proc. INTERSPEECH, pp. 569-572, 2011.

[17] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," Acoustical Science and Technology, Vol. 30, No. 5, pp. 363–371, 2009.

[18] 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, 2014:5, doi:10.1186/1687-4722-2014-5, 10 pages, 2014.

[19] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, "Multi-modal voice conversion using non-negative matrix factorization in noisy environments," in Proc. ICASSP, pp. 1561–1565, 2014.

[20] 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.