Many-to-many Voice Conversion Based on Multiple Non-negative Matrix Factorization

Ryo Aihara, Testuya 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

10.21437/Interspeech.2015-579

Abstract

We present in this paper an exemplar-based Voice Conversion (VC) method using Non-negative Matrix Factorization (NMF), which is different from conventional statistical VC. NMF-based VC has advantages of noise robustness and naturalness of converted voice compared to Gaussian Mixture Model (GMM)-based VC. However, because NMF-based VC is based on parallel training data of source and target speakers, we cannot convert the voice of arbitrary speakers in this framework. In this paper, we propose a many-to-many VC method that makes use of Multiple Non-negative Matrix Factorization (Multi-NMF). By using Multi-NMF, an arbitrary speaker's voice is converted to another arbitrary speaker's voice without the need for any input or output speaker training data. We assume that this method is flexible because we can adopt it to voice quality control or noise robust VC.

Index Terms: voice conversion, speech synthesis, many-to-many, exemplar-based, NMF

1. Introduction

Voice Conversion (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 [1]. 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 [2], Text-To-Speech (TTS) systems [3], spectrum restoring [4], bandwidth extension for audio [5], and more.

Many statistical approaches to VC have been studied [1, 6, 7]. Among these approaches, the Gaussian Mixture Model (GMM)-based mapping approach [1] is widely used. In this approach, the conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using Minimum Mean-Square Error (MMSE) on a parallel training set. A number of improvements in this approach have been proposed. Toda *et al.* [8] introduced dynamic features and the Global Variance (GV) of the converted spectra over a time sequence. Helander *et al.* [9] proposed transforms based on Partial Least Squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. However, over-smoothing and over-fitting problems have been reported [9] in regard to these GMM-based approaches because of statistical averages and the large number of parameters. These problems degrade the quality of synthesized speech.

In recent years, exemplar-based VC has been researched [10, 11] because of its flexibility and the naturalness of converted voice. In [10, 12], we proposed exemplar-based VC based on Non-negative Matrix Factorization (NMF) [13]. NMF is a well-known approach for source separation and speech enhancement [14, 15, 16]. In our VC method, source exemplars and target exemplars are extracted from the parallel training data, having the same texts uttered by the source and target speakers. The input source signal is expressed with a sparse representation of the source exemplars using NMF. By replacing a source speaker's exemplar with a target speaker's exemplar, the original speech spectrum is replaced with the target speaker's spectrum. Because our approach is not a statistical one, we assume that our approach can avoid the over-fitting problem and create a more natural voice [17].

Moreover, our exemplar-based VC method has noise robustness [12]. The noise exemplars, which are extracted from the before- and after-utterance sections in an observed signal, are used as the noise dictionary, and the VC process is combined with an NMF-based noise reduction method. On the other hand, NMF is one of the clustering methods. In our exemplar-based VC, if the phoneme label of a source exemplar is given, we can discriminate the phoneme of the input signal by using NMF. In [18], we proposed assistive technology for people who have articulation disorders by using this function of our exemplar-based VC. NMF-based VC is also applied to multimodal VC [19]. Wu *et al.* applied a spectrum compression factor to NMF-based VC and improved the conversion quality [11].

In spite of these efforts, VC has still not been put into practical use. One reason for this is that conventional VC needs a large amount of parallel training data between the source and target speakers. In GMM-based VC, there have been approaches that do not require parallel data. Lee *et al.* [20] used Maximum A Posteriori (MAP) in order to adapt training data. Mouchtaris *et al.* [21] proposed non-parallel training for GMM-based VC. Toda *et al.* [22] proposed eigen-voice GMM (EV-GMM) for one-to-many VC and many-to-one VC in which the source and target utterances are represented by a super vector of the reference speakers. Ohtani *et al.* [23] expand EV-GMM to many-to-many VC using a reference voice. Saito *et al.* [24] proposed tensor representation for one-to-many GMM-based VC. However, exemplar-based many-to-many VC has never been proposed.

This paper proposes a many-to-many exemplar-based VC approach using Multiple Non-negative Matrix Factorization (Multi-NMF). Parallel dictionaries, which are needed in conventional NMF-based VC, are replaced with dictionaries that are represented by the dictionaries of many speakers. We assume this method can be applied to voice quality control, noise-robust VC, and assistive technology.

The rest of this paper is organized as follows: In Section 2,

conventional one-to-one NMF-based VC is described. In Section 3, our proposed method is described. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. NMF-based Voice Conversion

In the exemplar-based approach, the observed signal is represented by a linear combination of a small number of bases. In this VC method, each basis denotes the exemplar of the spectrum, and the collection of exemplar $\mathbf{W}$ and the weight vector $\mathbf{h}_j$ are called the 'dictionary' and 'activity', respectively. When the weight vector $\mathbf{h}_j$ is sparse, the observed signal can be represented by a linear combination of a small number of bases that have non-zero weights.

$$\mathbf{V} \approx \mathbf{WH} \quad (1)$$

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_J], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_J]. \quad (2)$$

J represents the number of the frames. In this paper, we use NMF [13], which is a sparse coding method, in order to estimate the activity matrix.

Fig. 1 shows the basic approach of our exemplar-based VC, where D, L, and J 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.

A matrix of input source spectra $\mathbf{V}^s$ is decomposed into the source dictionary $\mathbf{W}^s$ and the activity matrix $\mathbf{H}^s$ by using NMF. 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. Fig. 2 shows the activity matrices estimated from parallel dictionaries. As shown in the figure, these activities have high energies at similar elements. Therefore, a matrix of target spectra $\hat{\mathbf{V}}^t$ can be constructed using the target dictionary $\mathbf{W}^t$ and the activity matrix of the source signal $\mathbf{H}^s$ as shown in Fig. 1.

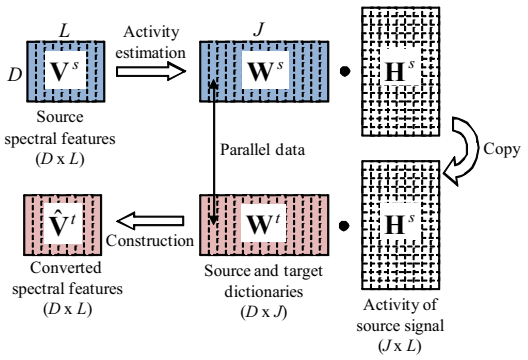

Figure 1: One-to-one VC using NMF

3. Many-to-many Voice Conversion Using Multi-NMF

3.1. Flow of the proposed method

Our proposed method is based on the following assumptions:

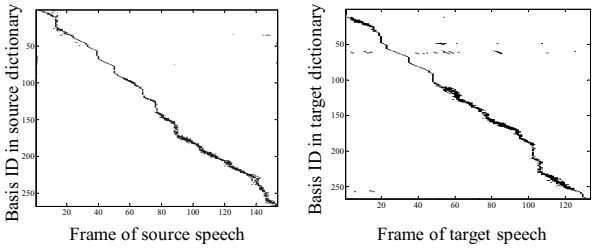

Figure 2: Activity matrices for parallel utterances.

1. Spectra of arbitrary speaker are represented by a linear combination of the basis of many speakers.
2. An activity matrix represents phoneme information which is speaker independent.

Fig. 3 shows the flow of the proposed method. $\mathbf{V}^s$, $\mathbf{V}^t$, $\hat{\mathbf{V}}^s$, $\mathbf{a}^s$, $\mathbf{a}^t$, $\mathbf{H}^s$, $\mathbf{H}^t$, denote the matrix of input source spectra, the matrix of the adaptation target speaker's spectra, the matrix of converted spectra, the source speaker's weight vector, the target speaker's weight vector, the activity matrix of the source speaker, the activity matrix of the target speaker, respectively. D, L, L', J denote the number of the dimension of a spectrum, the frame of the source spectra, the frame of the adaptive spectra, the frame of the dictionary, respectively. $\mathbf{W}^M \in \mathbb{R}^{(D \times J \times K)}$ denotes the dictionary matrix, which consists of the parallel exemplars of many speakers and K is the number of speakers who are included in it. The superscript of $\mathbf{W}^M$ means that it consists of the dictionaries of many speakers. The k-th speaker's dictionary is denoted by $\mathbf{W}_k^M \in \mathbb{R}^{(D \times J)}$.

First, the matrix of input source spectra $\mathbf{V}^s$ is represented as follows based on the assumption 1,

$$\mathbf{V}^s \approx \left(\sum_{k=1}^{K} a_k^s \mathbf{W}_k^M \right) \mathbf{H}^s \quad (3)$$

where a_k^s denotes the k-th element of $\mathbf{a}^s$. We emphasize that each speaker's dictionary is multiplied by the same activity matrix element of $\mathbf{H}^s$ in 3.

Next, some frames of the target speaker spectra $\mathbf{V}^t$ are used as adaptive spectra, and the target speaker's weight vector and adaptive data activity matrix are estimated as $\mathbf{a}^t$ and $\mathbf{H}^t$, respectively.

$$\mathbf{V}^t \approx \left(\sum_{k=1}^{K} a_k^t \mathbf{W}_k^M \right) \mathbf{H}^t \quad (4)$$

Finally, the converted spectra $\hat{\mathbf{V}}^t$ are constructed from the estimated target speaker's weight vector $\mathbf{a}^t$ and the source speaker's activity matrix $\mathbf{H}^s$ based on the assumption 2.

$$\hat{\mathbf{V}}^t = \left(\sum_{k=1}^{K} a_k^t \mathbf{W}_k^M \right) \mathbf{H}^s \quad (5)$$

In this method, the dictionary consists of either male only or female only spectra. Therefore, in cross-gender conversion, $\mathbf{W}_k^M$ in (3) is replaced with $\mathbf{W}_k^{sM}$, and $\mathbf{W}_k^M$ in (4) and (5) is replaced with $\mathbf{W}_k^{tM}$, where $\mathbf{W}_k^{sM}$ and $\mathbf{W}_k^{tM}$ denote the dictionaries of the source gender and the target gender, respectively.

3.2. Multi-NMF

We are proposing Multi-NMF, which estimates a speaker vector $\mathbf{a} \in \mathbb{R}^{(1 \times 1 \times K)}$ and an activity matrix $\mathbf{H} \in \mathbb{R}^{(J \times L)}$ from input

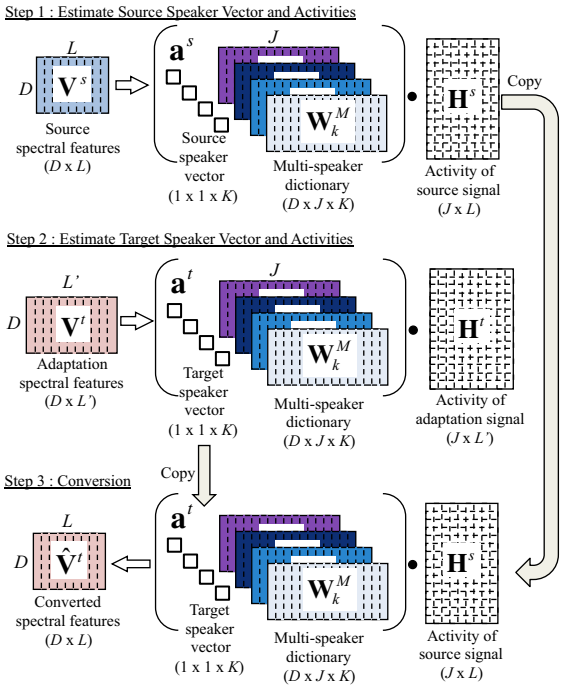

Figure 3: Many-to-many VC using Multi-NMF

spectra $\mathbf{V} \in \mathbb{R}^{(D \times L)}$ and given dictionary $\mathbf{W}^M \in \mathbb{R}^{(D \times J \times K)}$. The cost function of Multi-NMF is defined as follows,

$$d(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}) + \lambda ||\mathbf{H}||_1 \quad (6)$$

where the first term is the Kullback-Leibler (KL)-divergence between $\mathbf{V}$ and $\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}$, and the second term is the L1-norm regularization term that causes the activity matrix to be sparse. λ represents the weight of the sparse constraint.

$\mathbf{H}$ and $\mathbf{a}$ are estimated by minimizing (6). The updating rule is determined by adapting Jensen's inequality [1].

$$a_k \leftarrow \frac{a_k}{\sum_{d,l}(\mathbf{W}_k^M \mathbf{H})_{dl}} \sum_{d,l} \left(\frac{v_{dl}(\mathbf{W}_k^M \mathbf{H})_{dl}}{\sum_k a_k (\mathbf{W}_k^M \mathbf{H})_{dl}} \right) \quad (7)$$

$$\mathbf{H} \leftarrow \mathbf{H}. * ((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^{\mathsf{T}} (\mathbf{V}./(\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}))) ./((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^{\mathsf{T}} \mathbf{1}^{D \times L} + \lambda \mathbf{1}^{J \times L}) \quad (8)$$

where v_{dl} denotes the element of $\mathbf{V}$ and $.*$ and $./$ denote element-wise multiplication and division, respectively.

4. Experiments

4.1. Experimental conditions

We used the ATR Japanese speech database set C [25], which contains of the speech of 10 males and 10 females. The utterance of half of the males and the half of the females are stored as training data and the rest are stored as test data. The sampling rate was 12 kHz. We compared our method with conventional one-to-one NMF-based VC and one-to-one GMM-based VC, which use parallel data between the source and the target speakers as training data. In each method, 50 parallel sentences of each speaker were used for dictionary construction or training of GMM. In the proposed method, 2 sentences uttered by the target speaker, which were not included in the test or training data were used as adaptation data.

In the proposed and conventional NMF-based methods, the dimension number of the spectral feature was 2,565. It consisted of a 513-dimensional STRAIGHT [26] spectrum and its consecutive frames (the 2 frames coming before and the 2 frames coming after). The number of iterations of NMF and Multi-NMF was 300 and λ in (6) was set to 0.1.

In the conventional GMM-based method, MFCC+ΔMFCC+ΔΔMFCC is used as a spectral feature. Its number of dimensions is 60. The number of Gaussian mixtures was set to 64, which is experimentally selected. In this paper, in order to focus on the spectra conversion, F0 information was converted using parallel training data. It was converted using conventional linear regression based on the mean and standard deviation. The other information, such as aperiodic components, was synthesized without any conversion.

In order to evaluate our proposed method, we conducted objective and subjective evaluations. For the objective evaluation, 50 sentences that are not included in the training data were evaluated. We used Mel-cepstrum distortion (MelCD) [dB] [8] as a measurement of objective evaluations, which is defined as follows,

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (9)$$

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

The subjective evaluation was conducted on "speech quality" and "similarity to the target speaker". 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 [27]. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). On the similarity evaluation, the 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.

4.2. Results and discussions

Tables 1 to 4 show Mel-CD of male-to-male conversion, female-to-female conversion, male-to-female conversion, and female-to-male conversion, respectively. Source, Multi, NMF and GMM denote Mel-CD between the target and the source speech, converted by the proposed method, converted by one-to-one NMF, and converted by one-to-one GMM, respectively. As shown in the tables, although our proposed method includes neither source nor target speaker's spectra in the dictionaries, the distortion between one-to-one VC methods and our proposed many-to-many VC method is quite small. Moreover, in some pairs of speakers, the distortion of the proposed method is almost the same as that of one-to-one conversions (for example, F5→F10 and F2→M2).

Fig. 4 shows the results of the MOS test on speech quality. M-to-M, F-to-F, M-to-F and F-to-M denote male-to-male

[1] The derivation of (7) and (8) is uploaded to `http://www.me.cs.scitec.kobe-u.ac.jp/aihara/Interspeech2015.pdf`

conversion, female-to-female conversion, male-to-female conversion, and female-to-male conversion, respectively. In inter-gender conversion, our proposed method obtained a better score than conventional one-to-one NMF and GMM-based VC. These results were confirmed by a p-value test of 0.05. In cross-gender conversion, the difference between our proposed method and one-to-one NMF-based VC is not significant. However, our proposed method obtained the better score compared to one-to-one GMM-based VC. These results were confirmed by a p-value test of 0.05. We assume that the performance difference between inter-gender VC and cross-gender VC is caused by the difference between dictionaries. In our cross-gender conversion, the dictionary, which we used in the activity estimation of input speech, is different from that which we used in the target weight vector estimation, and this difference may impact the assumption 2 in Section 3.

Fig. 5 shows the results of the XAB test on speaker similarity between the proposed method and one-to-one NMF-based VC. The score of our proposed method is slightly lower than that of the one-to-one NMF-based method except for female-to-female conversion. We assume that is because the dictionary, which we used in our proposed method contains neither the source speaker's spectra nor the target speaker's spectra. Fig. 6 shows the results of the XAB test on speaker similarity between the proposed method and one-to-one GMM-based VC. The difference between our proposed method and the one-to-one GMM-based method is not significant. This speaker similarity test shows that our proposed many-to-many VC approach effectively converts the individuality of the source speaker's voice to the target speaker's voice.

Table 1: Mel-CD of male-to-male conversion [dB]

	Source	Multi	NMF	GMM
M1→M6	4.76	4.16	4.06	3.93
M2→M7	5.29	4.92	4.71	4.74
M3→M8	4.68	4.47	4.15	4.23
M4→M9	4.59	4.18	3.92	3.92
M5→M10	4.29	4.02	3.69	3.62
Mean	4.72	4.35	4.11	4.09

Table 2: Mel-CD of female-to-female conversion [dB]

	Source	Multi	NMF	GMM
F1→F6	4.74	4.38	4.19	4.20
F2→F7	4.88	4.52	4.51	4.51
F3→F8	4.77	4.25	4.07	3.99
F4→F9	4.78	4.40	4.18	4.10
F5→F10	4.50	4.07	4.06	4.01
Mean	4.73	4.32	4.20	4.16

Table 3: Mel-CD of male-to-female conversion [dB]

	Source	Multi	NMF	GMM
M1→F1	5.46	4.59	4.32	4.59
M2→F2	5.05	4.59	4.32	4.37
M3→F3	5.22	4.44	4.24	4.27
M4→F4	5.89	4.95	4.83	4.73
M5→F5	5.05	4.39	4.04	4.06
Mean	5.34	4.57	4.35	4.41

5. Conclusions

This paper introduced exemplar-based many-to-many VC using multi-NMF. In this framework, the input speaker's spectra are represented by linear combinations of spectra from a dictionary that contains the spectra of many speakers. Our proposed Multi-NMF estimates the source speaker weight vector and its activities from input spectra and a dictionary. A target speaker weight vector is estimated from adaptation data and the target

Table 4: Mel-CD of female-to-male conversion [dB]

	Source	Multi	NMF	GMM
F1→M1	5.46	4.69	4.48	4.67
F2→M2	5.05	4.42	4.24	4.42
F3→M3	5.22	4.37	4.11	4.24
F4→M4	5.89	4.99	4.75	4.75
F5→M5	5.05	4.34	4.07	4.10
Mean	5.34	4.56	4.33	4.43

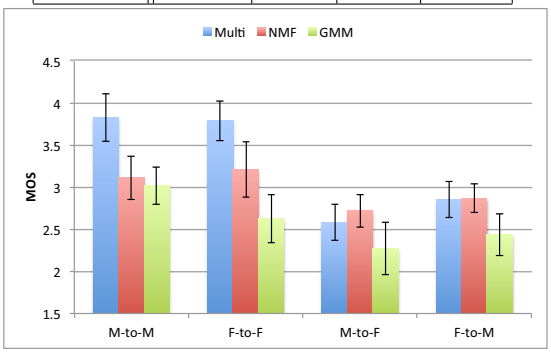

Figure 4: MOS of speech quality

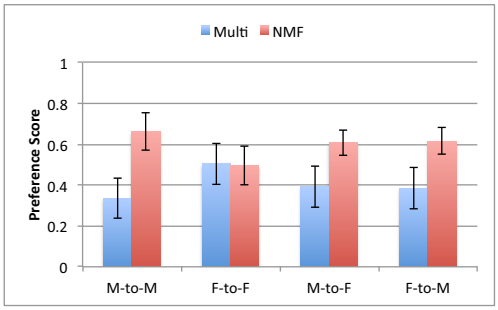

Figure 5: XAB test between proposed method and NMF

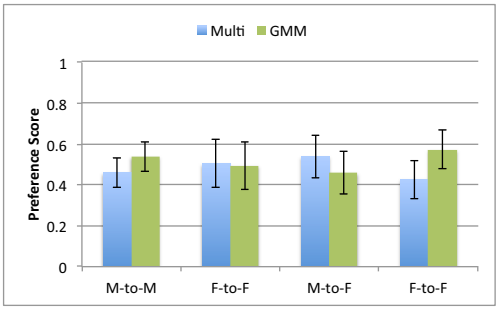

Figure 6: XAB test between proposed method and GMM

speech is synthesized from the target speaker weight vector and activities of input speech. We assume that Multi-NMF makes it possible to decompose input speech into phonetic information, which is estimated as activities and speaker information, which is estimated as the speaker weight vector. Experimental results revealed that the conversion quality of the proposed method is almost the same as that of conventional one-to-one VC although our proposed method includes neither the source speaker's spectra nor the target speaker's spectra.

In future work, we will apply our method to noisy environments and an assistive technology for people with articulation disorders. Comparison between our method and other many-to-many VC methods will also be a part of our future work. We assume that the proposed method can be easily applied to voice quality control by using regression of speaker weight vectors and voice expression words. We also plan to research speaker identification using the speaker weight vector.

6. References

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

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

[4] 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.

[5] 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.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models," *in Proc. ICASSP*, pp. 655–658, 1988.

[7] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication, vol. 11, no. 2-3, pp. 175-187*, 1992.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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, no. 7, pp. 2067–2080, 2011.

[17] 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.

[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*, vol. 2014:5, doi:10.1186/1687-4722-2014-5, 2014.

[19] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, "Multimodal exemplar-based voice conversion using lip features in noisy environments," *in Proc. INTERSPEECH*, vol. 1159-1163, 2014.

[20] 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.

[21] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *Audio, Speech, and Language Processing, IEEE Transactions on 14 (3)*, pp. 952–963, 2006.

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

[23] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," *in Proc. Interspeech*, pp. 1623–1626, 2009.

[24] 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.

[25] 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.

[26] H. Kawahara, "STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, pp. 349–353, 2006.

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