ACTIVITY-MAPPING NON-NEGATIVE MATRIX FACTORIZATION FOR EXEMPLAR-BASED 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. We present in this paper an exemplar-based VC method using Non-negative Matrix Factorization (NMF), which is different from conventional statistical VC. In our previous exemplar-based VC method, input speech is represented by the source dictionary and its sparse coefficients. The source and the target dictionaries are fully coupled and the converted voice is constructed from the source coefficients and the target dictionary. In this paper, we propose an Activity-mapping NMF approach and introduce mapping matrices between source and target sparse coefficients. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method and a conventional NMF-based method.

Index Terms— voice conversion, sparse representation, non-negative matrix factorization, 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 systems [3], spectrum restoring [4], bandwidth extension for audio [5], etc.

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. There have also been approaches that do not require parallel data that make use of GMM adaptation techniques [10] or eigen-voice GMM (EV-GMM) [11, 12].

In recent years, approaches based on sparse representations have gained interest in a broad range of signal processing. In [13, 14], we proposed exemplar-based VC, which is based on the idea of sparse representation. In our exemplar-based VC, we use Non-negative Matrix Factorization (NMF) [15], which is a well-known approach for source separation and speech enhancement [16, 17, 18]. In our VC, 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 natural voice.

Moreover, our exemplar-based VC method has noise robustness [14]. 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 [19], we proposed assistive technology for articulation disorders by using this function of our exemplar-based VC. NMF-based VC is also applied to multimodal VC [20]. Wu *et. al* applied a spectrum compression factor to NMF-based VC and improved the conversion quality [21].

In this paper, we propose advanced exemplar-based VC using Activity-mapping NMF. In those conventional NMF-based VC methods, input speech is represented by the source dictionary and its sparse coefficients. The source and the target dictionaries are fully coupled and the converted voice is constructed from the source coefficients and the target dictionary. However, there is mismatching between the source coefficients and the target coefficients. In this paper, we propose Activity-mapping NMF for exemplar-based VC in order to solve this problem. By using Activity-mapping NMF, a mapping function, which shows the relationship between two speakers, is learned so that mismatching between the source and the target sparse coefficients are compensated for. The effectiveness of this method was confirmed by comparing it with the conventional NMF-based method and the conventional GMM-based method.

The rest of this paper is organized as follows: In Section 2, the basic idea of NMF-based VC 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.

2. EXEMPLAR-BASED VOICE CONVERSION USING NON-NEGATIVE MATRIX FACTORIZATION

In the exemplar-based approach, the observed signal is represented by a linear combination of a small number of bases.

$$\mathbf{v}_j \approx \sum_{k=1}^{K} \mathbf{w}_k h_{k,j} = \mathbf{W}\mathbf{h}_j \quad (1)$$

$\mathbf{v}_j$ represents the j-th frame of the observation. $\mathbf{w}_k$ and $h_{k,j}$ represent the k-th basis and the weight, respectively. $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_K]$

and $\mathbf{h}_j = [h_{1,j} \dots h_{K,j}]^\mathsf{T}$ are the collection of the bases and the stack of weights. 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. (1) is expressed as the inner product of two matrices using the collection of the frames or bases.

$$\mathbf{V} \approx \mathbf{WH} \quad (2)$$

$$\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_J], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_J]. \quad (3)$$

J represents the number of the frames. In this paper, we use NMF, 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 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 frames from parallel training data are used as exemplars.

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

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \quad (4)$$

In (4), the first term is Kullback-Leibler (KL) divergence between $\mathbf{V}^s$ and $\mathbf{W}^s\mathbf{H}^s$ and the second term is the sparse constraint with the L1-norm regularization term that causes activity matrix to be sparse. λ represents a weight of sparse constraint. This function is minimized by iteratively updating the following equation.

$$\mathbf{H}^s_{n+1} = \mathbf{H}^s_n .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \; ./ (\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{I\times J} + \lambda \mathbf{1}^{K\times J}) \quad (5)$$

$.*$ and $./$ denote element-wise multiplication and division, respectively.

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. Estimated source activity $\mathbf{H}^s$ is multiplied to target dictionary $\mathbf{W}^t$ and target spectra $\hat{\mathbf{V}}^t$ is constructed.

3. VOICE CONVERSION BASED ON ACTIVITY-MAPPING NON-NEGATIVE MATRIX FACTORIZATION

3.1. Basic Idea

In the NMF-based approach described in Section 2, the parallel dictionary consists of the parallel training data themselves. Therefore, as the number of the bases in the dictionary increases, the input signal comes to be represented by a linear combination of a large number of bases rather than as a small number of bases. When the number of bases that represent the input signal becomes large, the assumption of similarity between source and target activities may be weak due to the influence of the mismatching between the input signal and the selected bases.

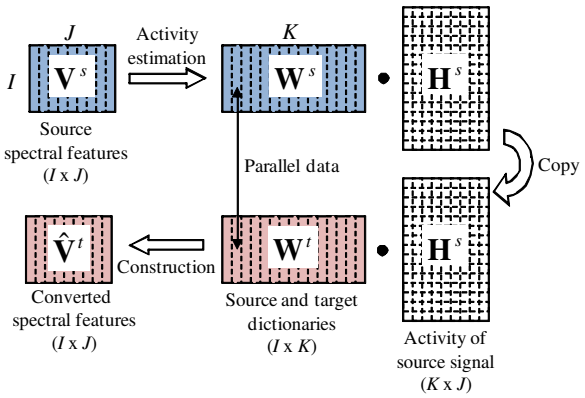

Fig. 1. Basic approach of NMF-based voice conversion

Fig. 2 shows an example of the activity matrices estimated from a Japanese word "akogareru" ("adore" in English), where one is uttered by a male and the other by a female. These words are aligned by using DTW in advance, and the source and target speaker's dictionaries, which consist of 250 bases, are used in activity estimation. The estimated activities are different although input features and dictionaries are parallel. We assume that this problem degrades the performance of the exemplar-based VC. Hence, in this paper, we introduce a mapping function between the source and target speaker's activities.

Fig. 3 shows the conversion procedure of our VC method. In the Activity Estimation stage, a source spectral exemplar matrix $\mathbf{V}^s$ is decomposed into a linear combination of bases from the source dictionary $\mathbf{W}^s$. The indexes and weights of the bases are estimated using NMF as source activity $\mathbf{H}^s$. In the next stage, the Activity Transformation stage, estimated source activity $\mathbf{H}^s$ is transformed into target activity $\mathbf{H}^t$ using a pre-learned mapping matrix $\mathbf{A}$. This mapping matrix spans a hidden space between the source and target speakers. Finally, in the Target Construction stage, transformed activity $\mathbf{H}^t$ is multiplied by the target dictionary which consists of exemplars of the target speaker's spectra, and then the converted speech $\hat{\mathbf{V}}^t$ is constructed.

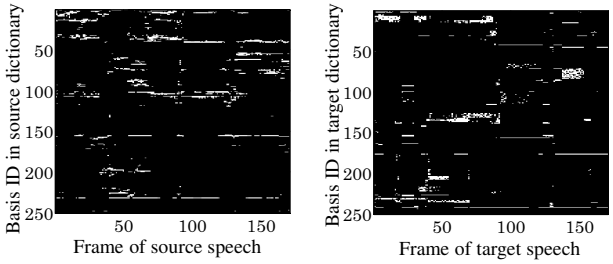

Fig. 2. Activity matrices for parallel utterances

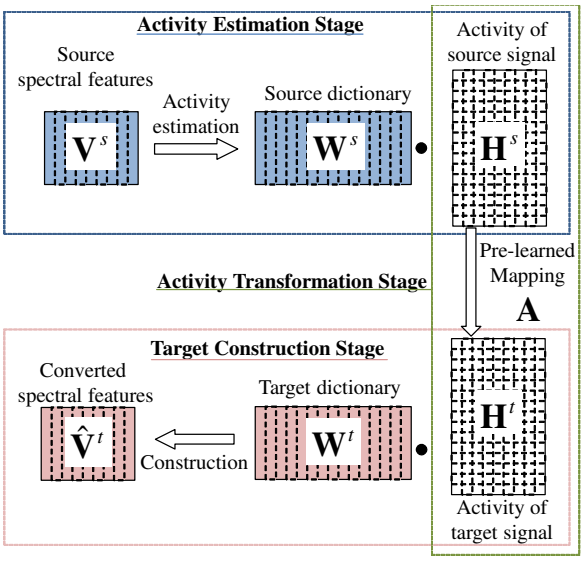

Fig. 3. Flow chart of Voice Conversion Using Activity-mapping NMF

3.2. Activity-mapping Non-negative Matrix Factorization

We propose the following cost function in order to estimate the desired mapping:

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 + d(\mathbf{V}^t, \mathbf{W}^t\mathbf{A}\mathbf{H}^s) + \lambda||\mathbf{A}\mathbf{H}^s||_1 \quad (6)$$

The first and the second terms are the same as (4). The third term is the KL-divergence between $\mathbf{V}^t$ and $\mathbf{W}^t\mathbf{A}\mathbf{H}^s$ and the fourth term is the sparse constraint of $\mathbf{A}\mathbf{H}^s$.

In order to estimate $\mathbf{H}^s$ precisely in the Activity Estimation stage, $\mathbf{H}^s$ is estimated using the first term and the second term of (6). The activity mapping matrix $\mathbf{A}$ is estimated by minimizing (6). The updating rule is determined by adapting Jensen's inequality [1].

$$\mathbf{A}_{n+1} = \mathbf{A}_n./((\mathbf{W}^{t\mathsf{T}}\mathbf{1}^{(I\times J)} + \lambda\mathbf{1}^{(K\times K)}).*(\mathbf{1}^{(J\times K)}\mathbf{H}^{s\mathsf{T}})) \\ .*(\mathbf{W}^{t\mathsf{T}}(\mathbf{V}^t./(\mathbf{W}^t\mathbf{A}_n\mathbf{H}^s))\mathbf{H}^{s\mathsf{T}}) \quad (7)$$

In the Activity Estimation stage, the source activity $\mathbf{H}^s$ is estimated using (5). In the next Activity Transformation stage, the target activity $\mathbf{H}^t$ is obtained using pre-trained $\mathbf{A}$ and the converted spectra $\hat{\mathbf{V}}^t$ is constructed as follows in the final Target Construction stage:

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^t = \mathbf{W}^t\mathbf{A}\mathbf{H}^s \quad (8)$$

4. DICTIONARY CLUSTERING AND SELECTING

Our conventional VC in Section 2 holds all the training data as one dictionary pair. In the case of activity-mapping NMF, a mapping matrix obtained from only one dictionary pair is not enough to cover all the variations of speech spectra because speech spectra vary widely. Therefore, we divide the parallel data into a numbers of clusters and adopt activity-mapping NMF for each cluster.

[1]The derivation of (7) is uploaded to `http://www.me.cs.scitec.kobe-u.ac.jp/aihara/ICASSP2015.pdf`

Our clustering algorithm is similar to the k-means algorithm but uses KL-divergence instead of squared Euclidean distances. Input data $\mathbf{z}_j = [\mathbf{v}_j^{s\mathsf{T}}, \mathbf{v}_j^{t\mathsf{T}}]^\mathsf{T}$ is clustered by using the following cost function:

$$D = \sum_{j=1}^{J} d(\mathbf{z}_j, \mathbf{m}_{c_j}) \quad (9)$$

$\mathbf{z}_j$, $\mathbf{m}_{c_j}$ and J represent the j-th joint feature of the source and target spectra, the c_j-th cluster and number of frames, respectively. c_j represents a cluster index of the j-th frame, which is decided as follows:

$$c_j = \arg\min_l d(\mathbf{z}_j, \mathbf{m}_l) \quad (10)$$

l represents the index of cluster.

In the target construction stage, a cluster is selected for an input spectrum. The cluster of the input spectrum is decided using NMF.

$$c_j = \arg\min_l d(\mathbf{v}_j^s, \mathbf{W}_l^s\mathbf{h}_{l\,j}^s) + \lambda||\mathbf{h}_{l\,j}^s||_1 \\ s.t. \quad \mathbf{h}_{l\,j}^s \geq 0 \quad (11)$$

$\mathbf{W}_l^s$ represents the dictionary of l-th cluster. $\mathbf{h}_{l\,j}^s$ minimizing (11) is estimated iteratively, applying (5).

Our proposed algorithm is summarized in Table 1.

Table 1. Algorithm of Activity-Mapping VC

Initializing for estimation of activity-mapping
Set source and target exemplars to $\mathbf{V}^s$ and $\mathbf{V}^t$.
Set source and target dictionaries to $\mathbf{W}^s$ and $\mathbf{W}^t$.
$\mathbf{A}$ and $\mathbf{H}^s$ are initialized with a random matrix.
Clustering
Jointed $\mathbf{V}^s$ and $\mathbf{V}^t$ are clustered using (9).
For each iteration
For each cluster
• Optimize $\mathbf{H}^s$ by (5)
• Optimize $\mathbf{A}$ by (7)

Initializing for conversion
Set input spectra $\mathbf{V}^s$, mapping matrix $\mathbf{A}$,
source dictionary $\mathbf{W}^s$, target dictionary $\mathbf{W}^t$.
Clustering
Cluster the input spectrum $\mathbf{v}_j^s$ using (11).
For each iteration
For each cluster
• Optimize $\mathbf{H}^s$ by (5)
• Construct $\hat{\mathbf{V}}^t$ by (8)

5. EXPERIMENTAL RESULTS

5.1. Experimental Conditions

The proposed VC technique was evaluated by comparing it with the conventional NMF-based method [14] (referred to as the "sample-based method" in this paper) 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 [22]. The sampling rate was 12 kHz. In our proposed method, λ is set to be 0.1. The number of dictionary clusters

is set to be 64. Fifty sentences were used for training the dictionary and the mapping matrix for our proposed VC. The same 50 sentences were used as training data for the GMM-based VC and sample-based method.

In the proposed and sample-based methods, the dimension number of the spectral feature is 2,565. It consists of a 513-dimensional STRAIGHT spectrum [23] and its consecutive frames (the 2 frames coming before and the 2 frames coming after). The number of iterations for estimating the activity in the proposed and sample-based methods was 500. 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 96, which is experimentally selected.

In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation [8]. The other information, such as aperiodic components, is 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. The spectral distortion improvement ratio (SDIR) [dB] represented as the following equation, was used for objective evaluation.

$$SDIR[dB] = 10\log_{10}\frac{\sum_i^I |\mathbf{V}^t(i) - \mathbf{V}^s(i)|^2}{\sum_i^I |\mathbf{V}^t(i) - \hat{\mathbf{V}}^t(i)|^2} \quad (12)$$

Here, the source spectrum matrix $\mathbf{V}^s$, target spectrum matrix $\mathbf{V}^t$ and converted spectrum matrix $\hat{\mathbf{V}}^t$ are 513-dimensional STRAIGHT spectra, which are normalized so that the sum of the magnitudes over frequency bins equals unity. For the subjective evaluation, the XAB test was carried out. Twenty-five sentences that are not included in the training data were evaluated. In an XAB test, the target utterance is shown as reference X. The subject listened to a voice converted by our proposed method and by conventional methods. They selected the samples which they felt have better quality. A total of 10 Japanese speakers took part in the test using headphones.

5.2. Results and Discussion

Fig. 4 shows the SDIR for each method. "GMM" shows the SDIR of the conventional statistical VC method. "NMF" shows the SDIR of the conventional NMF-based method explained in Sec. 2. "cls-NMF" shows the SDIR of the NMF-based method with dictionary clustering but without activity mapping. "Act-NMF" shows the SDIR of our proposed method which is combined dictionary clustering and activity mapping. As shown in the figure, the improvement ratio of "NMF" is lower than that of the conventional GMM-based method. However, the improvement ratio of our proposed "Act-NMF" is higher than the other method. This result shows the effectiveness of activity mapping in NMF-based VC. The improvement ratio of "cls-NMF" is lower than the other methods. These results imply that there is still room for improvement in the dictionary clustering method. The improvement of dictionary clustering will enhance the effectiveness of our proposed VC method.

Fig. 5 shows the results of the XAB test. The left side shows the preference score between GMM-based VC and the proposed method. The right side shows the preference score between sample-based VC and the proposed method. The error bars show a 95% confidence score. The results of these tests were confirmed by a p-value test of 0.05. Our proposed VC method obtained a higher score than GMM-based conversion. We assume that our proposed method can create a natural-sounding voice because our method is exemplar-based. It also obtained a higher score than the sample-based VC method. This result shows that Activity-mapping NMF can capture the relationship between source and target exemplars and enhance the conversion quality of exemplar-based VC.

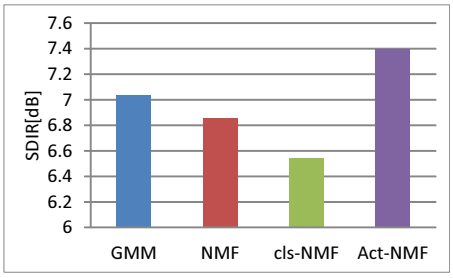

Fig. 4. SDIR calculated from converted speech using each method

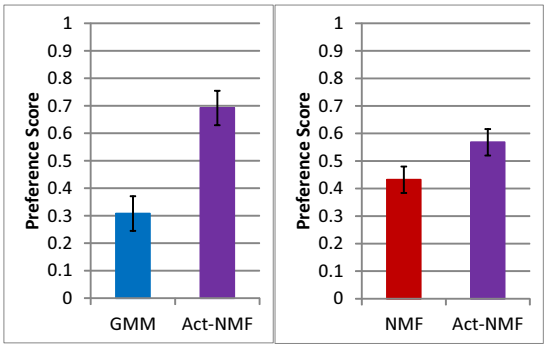

Fig. 5. Results of XAB test

6. CONCLUSIONS

We have proposed a novel activity mapping method for exemplar-based VC using NMF. Our proposed method optimizes the mapping matrix in the sparse space between the source and target dictionaries. The learned mapping matrix spans a hidden space between the source and target speakers. Objective and subjective evaluations show the effectiveness of our method compared to conventional NMF and GMM-based VC.

Some problems still remain with our method. We employed a k-means based dictionary clustering method to our VC because our activity mapping method degrade the performance of VC when it is adopted for a large-size dictionary. However, the objective evaluation shows that the dictionary clustering without activity mapping degrades the performance of exemplar-based VC. In [24], we proposed a phoneme-categorized dictionary that enhances the performances of exemplar-based VC. A phoneme-categorized dictionary does not work well with activity-mapping NMF because the number of exemplars in each sub-dictionary is too large. (There are 10 sub-dictionaries in [24]. We used 64 sub-dictionaries in this paper.) In future work, we will research a novel dictionary clustering method that is better matched with our exemplar-based VC using activity mapping.

The proposed method requires higher computation times than the GMM-based method. In [25], we proposed NMF-based VC that reduce the computational cost for conversion. Wu *et. al* also proposed method for NMF-based VC to reduces the computational cost [21]. In future work, we will combine these methods and investigate the optimal number of bases for better performance.

Also, we will apply our method to noisy environments and an assistive technology for people with articulation disorders.

7. 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 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 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 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 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] C. H. Lee and C. H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," *in Interspeech*, pp. 2254–2257, 2006.

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

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

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

[14] 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.

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

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

[17] 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.

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

[19] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization," *in ICASSP*, pp. 8037–8040, 2013.

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

[21] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 22, no. 10*, pp. 1506–1521, 2014.

[22] 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.

[23] 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.

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

[25] 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.