VOICE CONVERSION BASED ON NON-NEGATIVE MATRIX FACTORIZATION USING PHONEME-CATEGORIZED DICTIONARY

Ryo AIHARA, Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI

Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 6578501, Japan

ABSTRACT

We present in this paper an exemplar-based voice conversion (VC) method using a phoneme-categorized dictionary. Sparse representation-based VC using Non-negative matrix factorization (NMF) is employed for spectral conversion between different speakers. In our previous NMF-based VC method, source exemplars and target exemplars are extracted from parallel training data, having the same texts uttered by the source and target speakers. The input source signal is represented using the source exemplars and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. However, this exemplar-based approach needs to hold all the training exemplars (frames), and it may cause mismatching of phonemes between input signals and selected exemplars. In this paper, in order to reduce the mismatching of phoneme alignment, we propose a phoneme-categorized sub-dictionary and a dictionary selection method using NMF. By using the sub-dictionary, the performance of VC is improved compared to a conventional NMF-based VC. 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, sub-dictionary

1. INTRODUCTION

The human voice is rich in information. A listener perceives not only linguistic information from a speaker's voice but also speaker identity, emotional information, etc. 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.

There have also been studies on several tasks that make use of VC. Emotion conversion is a technique for changing emotional information in input speech while maintaining linguistic information and speaker individuality [2, 3]. VC is also being adopted as assistive technology that reconstructs a speaker's individuality in electrolaryngeal speech [4], disordered speech [5] or speech recorded by NAM microphones [6]. In recent years, VC has been used for automatic speech recognition (ASR) or speaker adaptation in text-to-speech (TTS) systems [7]. These studies show the varied uses of VC.

Many statistical approaches to VC have been studied [1, 8, 9]. 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. [10] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [11] 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 [12] or eigen-voice GMM (EV-GMM) [13, 14].

In recent years, approaches based on sparse representations have gained interest in a broad range of signal processing. In [15], we proposed exemplar-based VC, which is based on the idea of sparse representation. In approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of atoms. In some approaches for source separation, the atoms are grouped for each source, and the mixed signals are expressed with a sparse representation of these atoms. By using only the weights of the atoms related to the target signal, the target signal can be reconstructed. Gemmeke et al. [16] also propose an exemplar-based method for noise-robust speech recognition. In that method, the observed speech is decomposed into speech atoms, noise atoms, and their weights. Then the weights of the speech atoms are used as phonetic scores (instead of the likelihoods of hidden Markov models) for speech recognition.

In our exemplar-based VC [15], we use Non-negative Matrix Factorization (NMF) [17], which is a well-known approach for source separation and speech enhancement [18, 19]. 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 has noise robustness [15]. 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 [5], we proposed assistive technology for articulation disorders by using this function of our exemplar-based VC. From these two applications, we assume that our exemplar-based VC using NMF is a flexible method that can be

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

applied to many important tasks.

In this paper, we propose advanced exemplar-based VC using NMF. In order to improve the performance of speaker conversion of exemplar-based VC, we applied a phoneme-categorized dictionary and a dictionary selection method to our VC using NMF. In conventional NMF-based VC, the number of dictionary frames becomes large because the dictionary holds all the training exemplar frames. Therefore, it may cause a phoneme mismatching between input signals and selected exemplars. In this paper, a training exemplar is divided into a phoneme-categorized sub-dictionary, and an input signal is converted by using the selected sub-dictionary. 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 experimental data are evaluated, and the final section is devoted to our conclusions.

2. VOICE CONVERSION USING NON-NEGATIVE MATRIX FACTORIZATION

2.1. Basic Approach

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

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \quad (1)$$

$\mathbf{x}_l$ represents the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ represent the j-th basis and the weight, respectively. $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ are the collection of the bases and the stack of weights. In this paper, each basis denotes the exemplar of the spectrum, and the collection of exemplar $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are called the 'dictionary' and 'activity', respectively. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination of a small number of bases that have non-zero weights. Eq. (1) is expressed as the inner product of two matrices using the collection of the frames or bases.

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \quad (2)$$

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \quad (3)$$

L represents the number of the frames.

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. $\mathbf{A}^s$ represents a source dictionary that consists of the source speaker's exemplars and $\mathbf{A}^t$ 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.

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 an example of the activity matrices estimated from a Japanese word "ikioi" ("vigor" in English), where one is uttered by a male, the other is uttered by a female, and each dictionary is structured from just one word "ikioi" as the simple example.

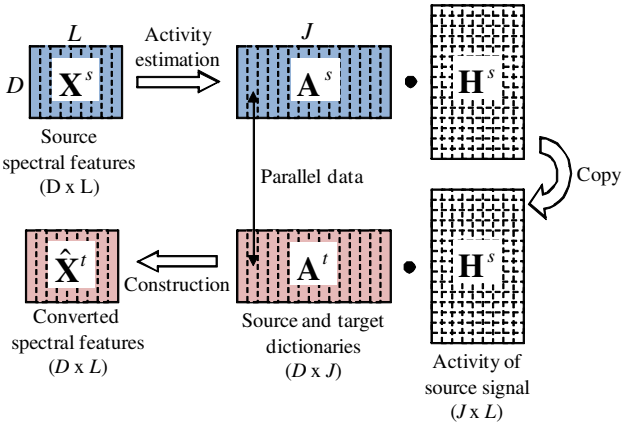

Fig. 1. Basic approach of NMF-based voice conversion

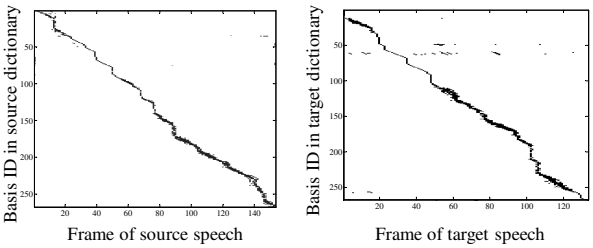

Fig. 2. Activity matrices for parallel utterances

As shown in Fig. 2, these activities have high energies at similar elements. 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. Therefore, the target speech can be constructed using the target dictionary and the activity of the source signal as shown in Fig. 1. In this paper, we use Non-negative Matrix Factorization (NMF), which is a sparse coding method in order to estimate the activity matrix.

2.2. Estimation of Activity

The joint matrix $\mathbf{H}^s$ in Fig. 1 is estimated based on NMF with the sparse constraint that minimizes the following cost function,

$$d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + ||(\lambda \mathbf{1}^{1\times L}) .* \mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geq 0 \quad (4)$$

$\mathbf{1}$ is an all-one matrix. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}^s$ and $\mathbf{A}^s\mathbf{H}^s$. The second term is the sparse constraint with the L1-norm regularization term that causes $\mathbf{H}^s$ to be sparse. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \dots \lambda_J]$. In this paper, all elements in λ were set to 0.1. $\mathbf{H}^s$ minimizing Eq. (4) is estimated iteratively applying the following update rule [17]:

$$\mathbf{H}^s_{n+1} = \mathbf{H}^s_n .* (\mathbf{A}^{sT}(\mathbf{X}^s ./ (\mathbf{A}^s\mathbf{H}^s_n))) ./ (\mathbf{A}^{sT}\mathbf{1}^{\mathbf{D}\times L} + \lambda \mathbf{1}^{1\times L}) \quad (5)$$

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

By using the activity and the target dictionary, the converted spectral features are constructed, as follows:

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) \quad (6)$$

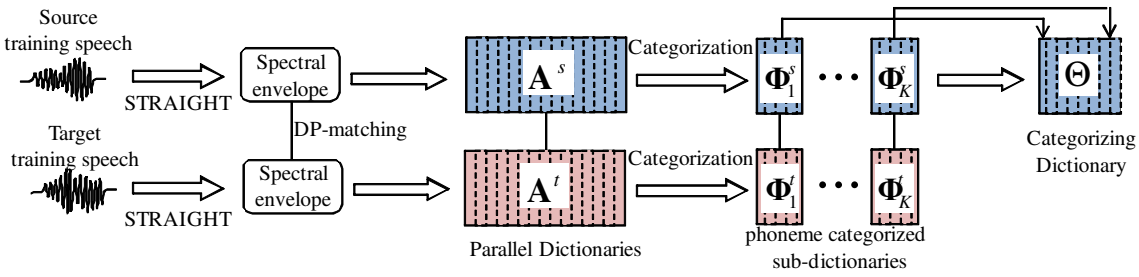

Fig. 3. Construction of categorizing dictionary that consists of representative vectors from each sub-dictionary.

3. NON-NEGATIVE MATRIX FACTORIZATION USING PHONEME-CATEGORIZED DICTIONARY

In the NMF-based approach described in Sec. 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 becomes to be represented by a linear combination of a LARGE number of bases rather than 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 mismatch between the input signal and the selected bases. We assume that this problem degrades the performance of the exemplar-based VC. Hence, we use a phoneme-categorized sub-dictionary in place of the large dictionary in order to reduce the number of the bases that represent the input signal.

3.1. Phoneme-categorized Dictionary

Fig. 3 shows how to construct the sub-dictionary. $\mathbf{A}^s$ and $\mathbf{A}^t$ imply the source and target dictionary which hold all the bases from training data. These dictionaries are divided into K dictionaries. In this paper, the dictionaries are divided into 10 categories according to Japanese phoneme categories shown in Table 1.

In order to select the sub-dictionary, a "categorizing-dictionary" which consists of the representative vector from each sub-dictionary is constructed. The representative vectors for each phoneme category consist of the mean vectors of the Gaussian Mixture Model (GMM).

$$p(\mathbf{x}_n^{(k)}) = \sum_{m=1}^{M_k} \alpha_m^{(k)} N(\mathbf{x}_n^{(k)}, \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}) \tag{7}$$

M_k, $\alpha_m^{(k)}$, $\boldsymbol{\mu}_m^{(k)}$ and $\boldsymbol{\Sigma}_m^{(k)}$ represent the number of the Gaussian mixture, the weights of mixture, mean and variance of the m-th mixture of the Gaussian, in the k-th sub-dictionary, respectively. Each parameter is estimated by using an EM algorithm.

The basis of the categorizing-dictionary which corresponds to the sub-dictionary is represented using the estimated phoneme GMM as follows:

$$\boldsymbol{\Phi}_k^s = [\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{N_k}^{(k)}] \tag{8}$$

$$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_1^{(k)}, \ldots, \boldsymbol{\mu}_{M_k}^{(k)}] \tag{9}$$

$$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \tag{10}$$

$\boldsymbol{\Phi}_k^s$, N_k and $\boldsymbol{\Theta}$ represent the k-th sub-dictionary, and the number of frames of the k-th sub-dictionary, the categorizing dictionary, respectively.

Table 1. Sub-dictionary categories

Category	Phoneme
a	a
e	e
i	i
o	o
u	u
plosives	p, t, k, b, d, g, s
fricatives	s, h, z
nasals	m, n, N
semi-vowel	j, w
liquid	r

3.2. Dictionary Selection and Voice Conversion

Fig. 4 shows the flow of the dictionary selection and VC. The input spectral features $\mathbf{X}^s$ are represented by a linear combination of bases from the categorizing-dictionary $\boldsymbol{\Theta}$. The weights of the bases are represented as activities $\mathbf{H}_{\boldsymbol{\Theta}}^s$.

$$\mathbf{X}^s \approx \boldsymbol{\Theta}\mathbf{H}_{\boldsymbol{\Theta}}^s \quad s.t. \quad \mathbf{H}_{\boldsymbol{\Theta}}^s \geq 0 \tag{11}$$

$$\mathbf{X}^s = [\mathbf{x}_1^s, \ldots, \mathbf{x}_L^s] \tag{12}$$

$$\mathbf{H}_{\boldsymbol{\Theta}}^s = [\mathbf{h}_{\boldsymbol{\Theta}1}^s, \ldots, \mathbf{h}_{\boldsymbol{\Theta}L}^s] \tag{13}$$

$$\mathbf{h}_{\boldsymbol{\Theta}l}^s = [\mathbf{h}_{\boldsymbol{\theta}_1 l}^s, \ldots, \mathbf{h}_{\boldsymbol{\theta}_K l}^s]^T \tag{14}$$

$$\mathbf{h}_{\boldsymbol{\theta}_k l}^s = [h_{\boldsymbol{\theta}_1 l}^s, \ldots, h_{\boldsymbol{\theta}_{Mk} l}^s]^T \tag{15}$$

The activities $\mathbf{H}_{\boldsymbol{\Theta}}^s$ are estimated by Eq. (5)

Then, the l-th frame of input feature $\mathbf{x}_l^s$ is represented by a linear combination of bases from the source speaker sub-dictionary. The sub-dictionary $\boldsymbol{\Phi}_{\hat{k}}^s$, which corresponds to $\mathbf{x}_l$, is selected as follows:

$$\hat{k} = \arg\max_k \mathbf{1}^{1\times M_k} \mathbf{h}_{\boldsymbol{\theta}_k l}^s = \arg\max_k \sum_{m=1}^{M_k} h_{\boldsymbol{\theta}_m l}^s \tag{16}$$

$$\mathbf{x}_l = \boldsymbol{\Phi}_{\hat{k}}^s \mathbf{h}_{\hat{k},l} \tag{17}$$

The activity $\mathbf{h}_{l,\hat{k}}$ in Eq. (17) is estimated by Eq. (5) from the selected source speaker sub-dictionary.

By using the activity and the sub-dictionary of the target speaker $\boldsymbol{\Phi}_{\hat{k}}^t$, the l-th frame of the converted spectral feature $\hat{\mathbf{y}}_l$ is constructed as follows:

$$\hat{\mathbf{y}}_l = \boldsymbol{\Phi}_{\hat{k}}^t \mathbf{h}_{\hat{k},l} \tag{18}$$

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

The proposed VC technique was evaluated by comparing it with the conventional NMF-based method [15] (referred as the "sample-based method" in this paper) and the conventional GMM-based

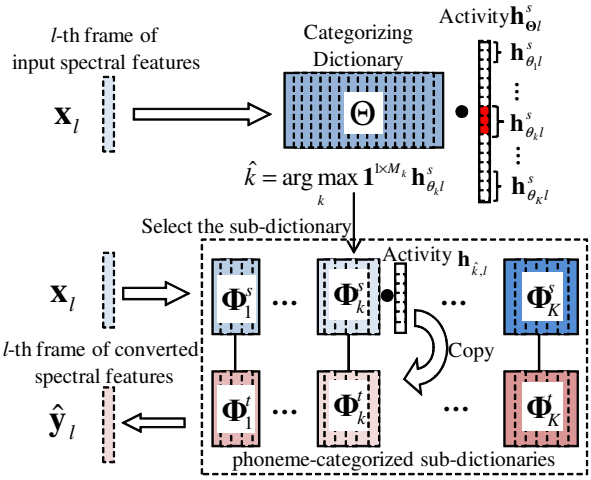

Fig. 4. NMF-based VC using the categorized dictionary

method [1] in a speaker-conversion task using clean speech data. The source speaker and target speaker were one male and one female speaker, whose speech is stored in the ATR Japanese speech database [20], respectively. The sampling rate was 8 kHz. A total of 216 words were used to construct parallel dictionaries in the sample-based VC, proposed VC, and used to train the GMM in the GMM-based method. The other 100 words were used for the test.

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 [21] and its consecutive frames (the 2 frames coming before and the 2 frames coming after). The Gaussian mixture, which is used to construct the categorizing-dictionary, is 1/500 of the number of bases of each sub-dictionary. The number of iterations for estimating the activity in the proposed and sample-based methods was 300. In the conventional GMM-based method, mfcc+Δmfcc+$\Delta\Delta$mfcc is used as a spectral feature. Its number of dimensions is 64. The number of Gaussian mixture was set to 64, which is experimentally selected.

In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation [10]. 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. NSD (Normalized Spectrum Distortion) [22] represented as the following equation was used for objective evaluation.

$$NSD = \sqrt{\frac{||S^Y - \hat{S}^X||^2}{||S^Y - S^X||^2}} \tag{19}$$

S^X, S^Y and $\hat{S}^X$ represent 513-dimensional STRAIGHT spectra of source, target and converted utterance, respectively. The subjective evaluation was conducted on "naturalness" and "similarity to the target speaker". For the evaluation on naturalness, we performed a Mean Opinion Score (MOS) test [23]. 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 an XAB test, each subject listened to the voice of the target speaker. Then the subject listened to the voice converted by the two methods and selected which sample sounded most similar to the target speaker's voice. In these two subjective evaluations, a total of 10 Japanese speakers took part in the test using headphones.

4.2. Results and Discussion

The left side of Fig. 5 shows the NSD for each method. As shown in the figure, the distribution of the sample-based method (NMF) is higher than that of GMM-based method (GMM). However, our proposed method obtained a lower distribution than the other two methods. This result shows the effectiveness of a phoneme-categorized sub-dictionary in NMF-based VC.

The right side of Fig. 5 shows the MOS test on naturalness. The error bars show 95% confidence intervals. There are no significant differences between the sample-based method (NMF) and the GMM-based method (GMM). However, our proposed method obtained a significantly higher score than the other two methods.

Fig. 6 shows the result of the XAB test on similarity. Similar to the right side of Fig. 5, there are no significant differences between the sample-based method (NMF) and the GMM-based method (GMM). However, our proposed method obtained a higher score than the sample-based method (NMF).

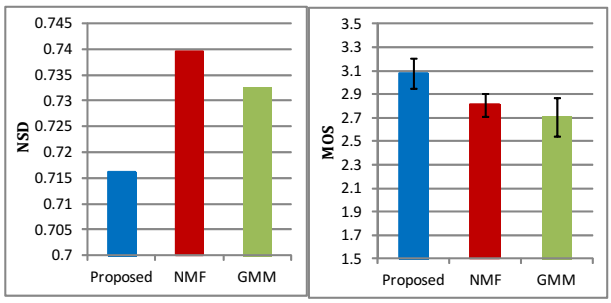

Fig. 5. Normalized spectrum distortion calculated from converted speech using each method (left), and results of MOS test (right)

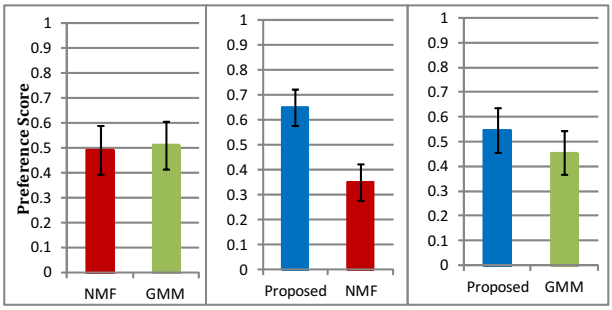

Fig. 6. Preference scores for the similarity

5. CONCLUSIONS

We have proposed an advanced exemplar-based VC using NMF that implements phoneme-categorized dictionary selection. In our proposed method, the input spectral feature can be represented by smaller numbers of exemplars, which are closer to the original phonemes compared to conventional NMF-based VC. Objective and subjective evaluations show the effectiveness of our method compared to conventional NMF and GMM-based VC. Especially, subjective evaluation shows that our proposed VC creates a natural voice compared to the other two VC methods. In future work, we will apply our method to noisy environments and an assistive technology for people with articulation disorders.

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] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," *in Interspeech*, pp. 2765–2768, 2011.

[3] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, 2012.

[4] 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.

[5] 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.

[6] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," *in Interspeech*, pp. 148–151, 2006.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

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

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