MANY-TO-ONE VOICE CONVERSION USING EXEMPLAR-BASED SPARSE REPRESENTATION

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 a many-to-one VC method using exemplar-based sparse representation, which is different from conventional statistical VC. In our previous exemplar-based VC method, input speech was 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. This method requires parallel exemplars (which consist of the source exemplars and target exemplars that have the same texts uttered by the source and target speakers) for dictionary construction. In this paper, we propose a many-to-one VC method in an exemplar-based framework which does not need training data of the source speaker. Some statistical approaches for many-to-one VC have been proposed; however, in the framework of exemplar-based VC, such a method has never been proposed. The effectiveness of our many-to-one VC has been confirmed by comparing its effectiveness with that of a conventional one-to-one NMF-based method and one-to-one GMM-based method.

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

1. INTRODUCTION

In recent years, approaches based on sparse representations have gained interest in a broad range of signal processing. In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of basis vectors,

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

where $\mathbf{v}_j$ represents the j-th frame of the observation. $\mathbf{w}_k$ and $h_{k,j}$ represent the k-th basis vector 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 basis vectors and the stack of weights. When the weight vector 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.

In the field of speech processing, Non-negative Matrix Factorization (NMF) [1] is a well-known approach for source separation and speech enhancement based-on sparse representation [2, 3]. In some source separation approaches, a dictionary is constructed for each source, and the mixed signals are expressed with a sparse representation of these dictionaries. By using only the weights (called "activity" in this paper) of basis in the target dictionary, the target signal can be reconstructed. Gemmeke *et al.* also proposed an exemplar-based method using NMF for noise-robust speech recognition [4].

Inspired by these sparse representation-based approaches, we proposed exemplar-based Voice Conversion (VC) in [5, 6]. 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 [7]. 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. In our exemplar-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 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.

The most popular conventional approach to VC is a statistical one [7, 8, 9]. Among these approaches, the Gaussian Mixture Model (GMM)-based mapping approach [7] 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; however, over-smoothing and over-fitting problems have been reported [10] because of statistical averages and the large number of parameters.

The benefits to an exemplar-based approach can be summarized in two points. First, an exemplar-based approach can convert speech into a natural-sounding voice. 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 [11]. 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.

A high hurdle for the practical use of VC has been the fact 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; however, an exemplar-based approach without parallel data has never been proposed. This paper proposes many-to-one VC using an exemplar-based sparse representation, which does not need any training data from a source speaker. We introduce Multiple Non-negative Matrix Factorization (Multi-NMF) and the parallel dictionaries that 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 easily applied to many-to-many VC.

The rest of this paper is organized as follows: In Section 2, related works are introduced. In Section 3, conventional one-to-one NMF-based VC is described. In Section 4, our proposed method

978-1-4799-7450-4/15/$31.00 ©2015 IEEE

is described. In Section 5, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. RELATED WORKS

The GMM-based approach is widely used for VC because of its flexibility and good performance [7]. Toda *et al.* [13] introduced dynamic features and the Global Variance (GV) of the converted spectra over a time sequence. Helander *et al.* [10] proposed transforms based on Partial Least Squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. GMM-based VC is also being used for assistive technology [14], Text-to-Speech (TTS) systems [15], spectrum restoring [16], and audio bandwidth extension [17].

The statistical VC mentioned above needs a large-volume parallel corpus between the source and target speakers. In this paper, "parallel" means that the text of the corpus between the source and target speakers is the same. This constraint can be a difficult requirement to meet in practice. In GMM-based VC, there have been approaches that do not require parallel data. Lee *et al.* [18] used Maximum A Posteriori (MAP) in order to adapt training data. Mouchtaris *et al.* [19] proposed non-parallel training for GMM-based VC. Toda *et al.* [20] proposed eigen-voice GMM (EV-GMM) for many-to-many VC in which the source and target speech are represented by a super vector of the reference speakers. Saito *et al.* [21] proposed tensor representation for one-to-many GMM VC.

Our VC approach is exemplar-based, which is different from conventional GMM-based VC. Exemplar-based VC using NMF was first proposed in [5]. The noise robustness of this exemplar-based approach is revealed in [6]. In [22], we proposed multimodal NMF-based VC to enhance the noise robustness of our method. The natural sound of the converted voice produced using NMF-based VC has been confirmed in [11]. Wu *et al.* [23] applied a spectrum compression factor to NMF-based VC and improved the conversion quality. NMF-based VC is being also adapted to assistive technology for those with speech articulation disorders [24].

3. EXEMPLAR-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 exemplars $\mathbf{W}$ 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.

$$\mathbf{V} \approx \mathbf{WH} \tag{2}$$

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_J], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_J]. \tag{3}$$

J represents the number of the frames. In this paper, we use NMF [1], 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, the numbers of frames, and the numbers of 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.

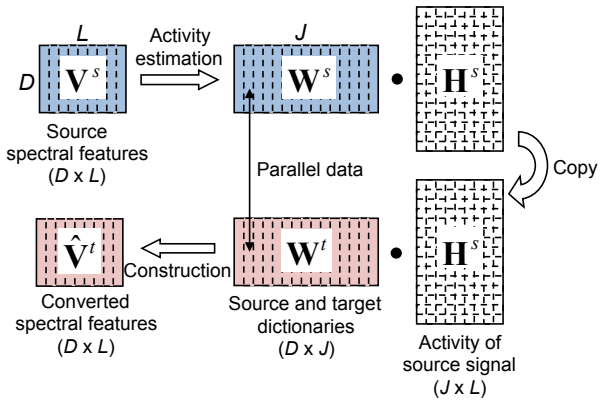

Figure 1: One-to-one VC using NMF

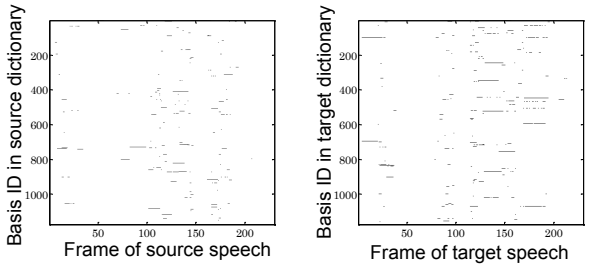

Figure 2: Activity matrices for parallel utterances.

4. MANY-TO-ONE VOICE CONVERSION USING MULTI-NMF

4.1. Flow of the proposed method

Our proposed method is based on the following assumptions:

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

Fig. 3 shows the flow of the proposed method. $\mathbf{V}^s$, $\hat{\mathbf{V}}^s$, $\mathbf{a}^s$, and $\mathbf{H}^s$ denote the matrix of input source spectra, the matrix of converted spectra, the source speaker's weight vector, and the activity matrix of the source speaker, respectively. D, L, and J denote the number of dimensions for a spectrum, the frame of the source spectra, and the frame of the dictionary, respectively.

$\mathbf{W}^M \in \mathbb{R}^{(D\times J\times K)}$ denotes the source 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 $\mathbf{W}_k^M \in \mathbb{R}^{(D\times J)}$. $\mathbf{W}^t \in \mathbb{R}^{(D\times J)}$ denotes the target dictionary matrix, which consists of the parallel exemplars of the target speaker.

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

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

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 (4). The summation in (4) can be represented as $\mathbf{W}^s \approx \sum_{k=1}^{K} a_k \mathbf{W}_k^M$. In this framework, the source dictionary $\mathbf{W}^s$, which is used in one-to-one VC using NMF, is represented by a linear combination of the dictionaries in $\mathbf{W}^M$. We assume that the activity matrix represents the phoneme information and the speaker weight vector represents the speaker identities. Therefore, Multi-NMF can extract the phoneme information and the speaker information from the input speech in the matrix representation.

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

$$\hat{\mathbf{V}}^s = \mathbf{W}^t\mathbf{H}^s \tag{5}$$

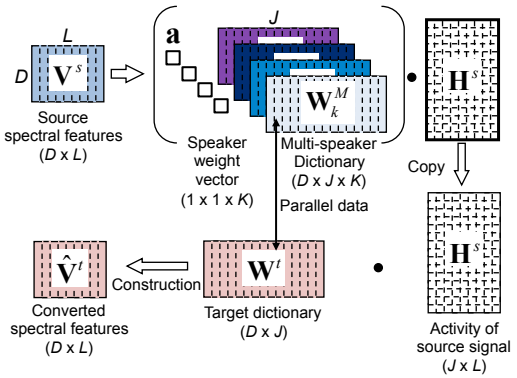

Figure 3: Flowchart of Many-to-one VC using Multi-NMF

4.2. 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 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,

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

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].

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

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

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

5. EXPERIMENTS

5.1. Experimental conditions

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. Six males and one female from the ATR Japanese speech database [25] were used in this experiment. In our proposed method, the source speaker is chosen from the six males and the source dictionary is constructed from the parallel utterances of the other males. The female speaker is set as the target speaker. In each method, 50 parallel sentences of each speaker were used for dictionary construction or GMM training.

In the proposed and conventional NMF-based methods, the number of dimensions for 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, 75 sentences that are not included in the training data were evaluated. We used Mel-cepstrum distortion (MelCD) [dB] [13] as a measurement of objective evaluations, which is defined as follows,

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

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, 36 sentences were evaluated by 10 Japanese speakers. For the

[1]The derivation of (7) and (8) is uploaded to `http://www.me.cs.scitec.kobe-u.ac.jp/aihara/WASPAA2015.pdf`

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.

5.2. Results and discussions

Fig. 4 shows the Mel-CD of source speech and converted speech. 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 figure, the difference between conventional one-to-one NMF and one-to-one GMM is not statistically significant. Although our proposed method does not include the source speaker's spectra in the dictionary, the difference between one-to-one VC methods and our proposed many-to-many VC method is quite small. For speaker C, our proposed method is slightly better than one-to-one NMF. This result shows that our proposed method has the potential to outperform conventional one-to-one VC.

Fig. 5 shows the MOS test on speech quality. The error bar shows the 95% confidence interval. The difference between our proposed method and one-to-one NMF-based VC is not statistically significant. However, our proposed method obtained the better score compared to one-to-one GMM-based VC.

Fig. 6 shows the results of the XAB test on speaker similarity between the proposed method and one-to-one NMF-based VC. For speakers A and B, the difference between these methods are not statistically significant. However, our proposed method obtained a slightly better score than one-to-one NMF-based VC in the case of speaker C. This result supports the objective evaluation of speaker C.

Fig. 7 shows the results of the XAB test on speaker similarity between the proposed method and one-to-one GMM-based VC. Our proposed method obtained a significantly higher score than one-to-one GMM.

6. CONCLUSIONS

This paper proposed exemplar-based many-to-one VC using sparse representation. 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 introduced Multi-NMF estimates the source speaker weight vector and its activities from input spectra and a dictionary. Therefore, the source speaker's utterance is converted to the target speaker's utterance without source speaker's training data. We assume that Multi-NMF makes it possible to decompose input speech into phonetic information, which is estimated as activities, and the 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 that requires source speaker's training data. We assume this method can be easily applied to many-to-many VC.

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-one VC methods will also be a part of our future work.

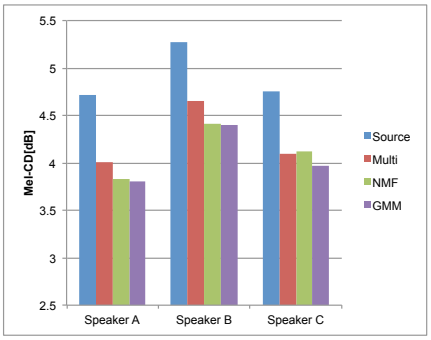

Figure 4: MelCD calculated from source speech and converted speech using each method

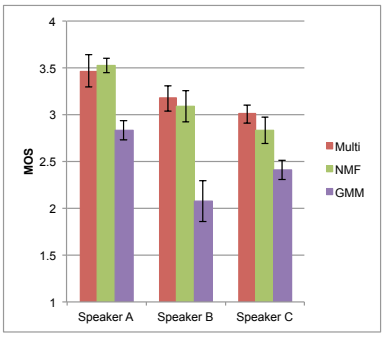

Figure 5: MOS of speech quality

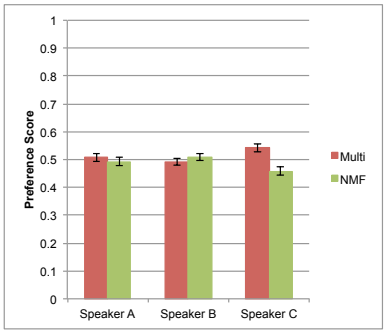

Figure 6: XAB test between our proposed method and one-to-one NMF VC

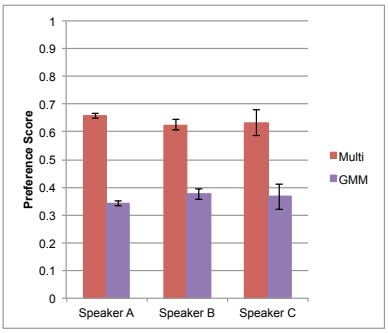

Figure 7: XAB test between our proposed method and one-to-one GMM VC

7. REFERENCES

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

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

[3] 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.

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

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

[6] ——, "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.

[7] 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.

[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] 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.

[11] 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.

[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] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[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.