Exemplar-based Emotional Voice Conversion Using Non-negative Matrix Factorization

Ryo AIHARA, Reina UEDA, Tetsuya TAKIGUCHI and Yasuo ARIKI
Graduate School of System Informatics, Kobe University, Japan
E-mail: aihara@me.cs.scitec.kobe-u.ac.jp, reina_1102@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp
Tel/Fax: +81-78-803-6570

***Abstract*—This paper presents an emotional voice conversion (VC) technology using non-negative matrix factorization, where parallel exemplars are introduced to encode the source speech signal and synthesize the target speech signal. The input source spectrum is decomposed into the source spectrum exemplars and their weights. By replacing source exemplars with target exemplars, the converted spectrum and F0 are constructed from the target exemplars and the target F0, which is paired with exemplars. In order to reduce the computational time, we adopted non-negative matrix factorization using active Newton set algorithms to our VC method. We carried out emotional voice conversion tasks, which convert an emotional voice into a neutral voice. The effectiveness of this method was confirmed with objective and subjective evaluations.**

I. 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. Particularly in telephone communication, emotional information in the human voice is important in understanding speaker's feelings or the various nuances of meaning. On the other hand, an emotional voice sometimes makes things stressful for the listener. For example, operators at call centers have to hear angry customers' claims all day long. If we can convert their angry voices into neutral voices, we can reduce the stress experienced by the operator. In this paper, we propose an exemplar-based emotional VC approach that converts an angry voice into a neutral voice.

In recent years, text-to-speech (TTS) techniques have been well developed. State-of-the-art TTS methods, such as unit selection [1] or Hidden Markov Model (HMM)-based speech synthesis [2], can produce high-quality speech in neutral reading styles. A concatenative approach, like unit selection, can create a natural-sounding voice; however, it requires a large speech corpus. A statistical parametric approach, such as HMM-based TTS, is more flexible compared to a concatenative approach. It requires less training data than unit selection and can transform the emotion or speech styles using speaker-adaptation techniques [3]. Emotional speech synthesis has been adapted for both unit selection and HMM-based TTS [4]; however, some problems remain regarding the naturalness of the synthesized speech.

Voice conversion (VC) techniques have been widely researched these days because of their flexibility. VC is a voice-to-voice technique that does not need text information in the input data, unlike TTS. One of the most popular VC applications is speaker conversion [5]. In speaker conversion, a source speaker's voice individuality is changed into a specified target speaker's so that the input utterance sounds as though a specified target speaker had spoken it. A statistical approach using Gaussian Mixture Model (GMM) is widely used in VC and a number of improvements in this approach have been proposed. In recent years, VC has been used for automatic speech recognition (ASR) or speaker adaptation in TTS systems [6].

Emotional VC is a technique for changing emotional information in input speech while maintaining linguistic information and speaker identity. Some researchers adopted GMM-based VC technique to emotional VC, but, because VC was mainly developed for spectrum conversion, it is hard to deal with prosody information in this framework. Because prosody information is more important than voice quality information in identifying emotion [7], it is a serious problem in emotional VC.

In [8], we proposed exemplar-based VC, which is based on the idea of sparse representation. This sparse representation-based approach has gained interest in a broad range of signal processing in recent years. In this approach, the observed signal is represented by a linear combination of a small number of atoms. In some approaches for source separation, the bases are grouped for each source, and the mixed signals are expressed with a sparse representation of these bases. By using only the weights of the bases related to the target signal, the target signal can be reconstructed. Gemmeke *et al.* [9] also proposed an exemplar-based method for noise-robust speech recognition. In that method, the observed speech is decomposed into speech bases, noise bases, and their weights. Then the weights of the speech bases are used as phonetic scores (instead of the likelihoods of hidden Markov models) for speech recognition.

In our exemplar-based VC [8], we use Non-negative Matrix Factorization (NMF) [10], which is a well-known approach for source separation and speech enhancement [11], [12]. 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

978-616-361-823-8

approach is not a statistical one, we assume that our approach can avoid the over-fitting problem and create a natural voice.

In this paper, we introduce a preliminary demonstration of emotional VC using NMF where input emotional spectra are converted to neutral spectra and prosody. In this paper, we used the active-set Newton algorithm for NMF [13], which decreases the computational time compared to conventional NMF. The effectiveness of this method was confirmed with objective and subjective evaluations.

The rest of this paper is organized as follows: In Section 2, related works are introduced. In Section 3, the NMF algorithm is explained. In Section 4, our proposed method is described. In Section 5, the experimental data are evaluated, and the final section is devoted to our conclusions.

II. Related Works

A. Conventional Voice Conversion

The statistical approaches to VC are most widely studied [5], [14], [15]. Among these approaches, the GMM-based mapping approach [5] 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.* [16] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander *et al.* [17] 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 [18] or eigen-voice GMM (EV-GMM) [19], [20].

B. Exemplar-based Voice Conversion

In [8], we proposed exemplar-based VC using NMF. In this approach, the input spectrum is converted by replacing the source speaker's exemplar with the target speaker's exemplar. In [21], we proposed advanced NMF-based VC using a phoneme-categorized dictionary. By using this method, we can improve the naturalness of the converted sound.

Our exemplar-based VC has noise robustness [8]. 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. In [22], we proposed multimodal VC using NMF. By using visual features which are combined with audio spectra, we can improve the noise robustness of our NMF-based VC.

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 [23], we proposed assistive technology for articulation disorders by using this function of our exemplar-based VC. From these applications, we assume that our exemplar-based VC using NMF is a flexible method that can be applied to many important tasks.

In addition, the computational time of NMF-based VC is one of its shortcomings compared to other VC methods. In this paper, we adapted the active-set Newton algorithm for NMF for our VC method, which has computational time that is 8 times faster than conventional NMF.

C. Emotional Voice Conversion

Mori *et al.* [24] proposed an F0 synthesis method for using subspace constraint in prosody. They assume that the combination of the number of syllables and accent type in Japanese determines the correlative dynamics of prosody. They employed principal component analysis and converted a word in each subspace, which is determined by its syllables and accent type. Wu *et al.*[25] proposed a hierarchical prosody conversion. The pitch contour of the source speech is decomposed into a hierarchical prosodic structure consisting of sentence, prosodic word, and sub-syllable levels.

Kawanami *et al.* [26] applied GMM-based spectrum conversion to emotional spectrum conversion. Veaux *et al.* [27] proposed an F0 conversion system based on GMM. However, because these methods convert only one feature in the human voice, some emotions were not converted well. In [28], we proposed GMM-based emotional VC that includes both voice quality and prosody. However, this method needs to deal with input speech F0 for each syllable unit. This shortcoming is not suited well to VC because, in most situations using VC, there is no text information in the input data.

In this paper, we introduce NMF-based emotional spectrum and prosody conversion. In this framework, an F0 syllable split is not necessary because F0 is converted based on the estimated exemplar of the target spectrum.

III. Non-negative Matrix Factorization

A. Formulation of NMF

NMF is one of the approach of sparse coding. In the idea of sparse coding, the observed signal is represented by a linear combination of a small number of bases.

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} \tag{1}$$

$\mathbf{v}_l$ represents the l-th frame of the observation (input). $\mathbf{w}_j$ and $h_{j,l}$ represent the j-th basis and the weight, respectively. In this paper, each basis denotes the exemplar of the spectrum. In NMF, the weight $h_{j,l}$ is constrained to non-negative. (1) can be rewritten by using matrix-vector product as follows,

$$\mathbf{v}_l \approx \mathbf{W}\mathbf{h}_l. \tag{2}$$

$\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^\mathsf{T}$ are the collection of the bases and the stack of weights. In this paper, the collection of exemplar $\mathbf{W}$ 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. (2) is expressed as the inner product of two matrices using the collection of the frames or bases.

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{3}$$

$$\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_L],\ \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{4}$$

L represents the number of the frames. $\mathbf{H}$ is a joint matrix of $\mathbf{h}_l$, which is called 'activities' in this paper.

$\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function,

$$f(\mathbf{H}) = d(\mathbf{V}, \mathbf{WH}) + \lambda||\mathbf{H}||_1 \ s.t. \ \mathbf{H} \geq 0. \quad (5)$$

The first term is the Kullback-Leibler (KL) divergence between $\mathbf{V}$ and $\mathbf{WH}$. The second term is the sparse constraint with the L1-norm regularization term that causes $\mathbf{H}$ to be sparse. λ represents the weight of sparsity constraint. The KL divergence is defined as

$$d(\mathbf{x}, \mathbf{y}) = \sum_d (x_d \log(y_d/x_d) - x_d + y_d). \quad (6)$$

B. Algorithm for NMF

When $\mathbf{W}$ or $\mathbf{H}$ is fixed, (6) is convex in $\mathbf{H}$ or $\mathbf{W}$. In this paper, $\mathbf{W}$ is fixed as the source dictionary and our object is to estimate $\mathbf{H}$. Many algorithms adopt an alternating minimization approach.

Lee *et al.* [10] proposed a standard method for minimizing (6). This method iteratively updates the following equation which alternately induces a descent in $\mathbf{H}$.

$$\begin{aligned} \mathbf{H}_{n+1} &= \mathbf{H}_n .* (\mathbf{W}^\mathsf{T}(\mathbf{V}./(\mathbf{WH}_n))) \\ &\quad ./(\mathbf{W}^\mathsf{T}\mathbf{1}^{D\times L} + \lambda \mathbf{1}^{D\times L}) \end{aligned} \quad (7)$$

$.*$ and $./$ denote element-wise multiplication and division, respectively. Because this update rule is derived from the expectation-maximization (EM) algorithm, this method is referred as EM-NMF.

The state-of-the-art-algorithm of NMF using KL-divergence is the active-set Newton algorithm (ASNA) proposed by Virtanen *et al.* [13]. This algorithm can reach a much lower divergence than EM-NMF, and is up to 8 times faster. In this paper, we applied this algorithm to exemplar-based VC. The EM-NMF algorithm requires more iteration to converge as the dictionary size increases. ASNA adds bases to an active set until the observation is adequately explained. In [13], experimental results indicate that ASNA is effective to estimate activities from a large-size dictionary. Since our source dictionary size is much larger than the other method using NMF [11], [12], ASNA is suited to our VC.

With active-set, it updates a set of active bases from a dictionary that has non-zero weights. For initialization, the basis that is most relative to the input observation vector is added to the active set. Newton's method is applied iteratively in order to estimate the weight of the active set. In each iteration, the most relative basis, that is not in the active set, is added to the active set. When the weight of a basis reaches to zero, the basis is removed. The detailed algorithm is explained in the following section.

IV. Emotional Voice Conversion Using NMF

A. Basic Idea

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

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.

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.

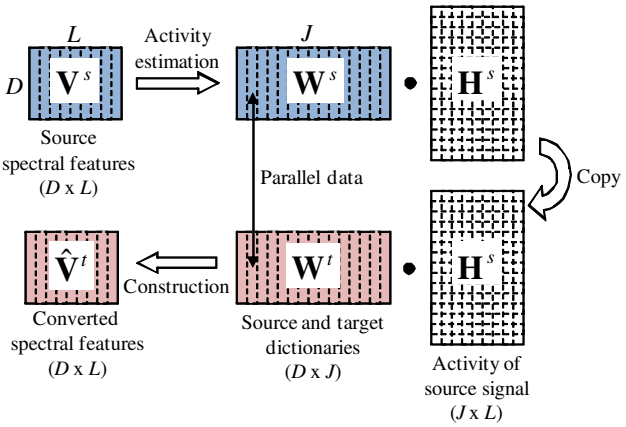

Fig. 1. Basic approach of NMF-based voice conversion

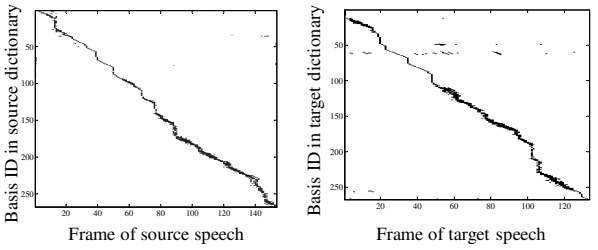

Fig. 2. Activity matrices for parallel utterances

B. Dictionary Construction

In order to make a parallel dictionary, some pairs of parallel utterances are needed, where each pair consists of the same text. Fig. 3 shows the process for constructing a parallel dictionary. $\mathbf{W}^s$, $\mathbf{W}^t_{spect}$, and $\mathbf{W}^t_{F0}$ represent the source spectrum dictionary, the target spectrum dictionary, the target F0 dictionary, respectively. Emotional utterances are set as source utterances and neutral utterances, which are the same text to emotional utterances, are set as target speeches. Spectrum envelopes and F0 are extracted from source and target utterances using STRAIGHT analysis [29]. Spectrum envelopes are extracted from source utterances. From the target utterances, spectrum envelopes and basic frequencies (F0) are extracted. Extracted features are aligned by using Dynamic Time Wrapping (DTW) so that each feature has the same number of frames. In order to estimate activities, a segment spectrum, which consists of some consecutive frames, is constructed. A segment F0, which consists of some consecutive frames, is constructed to convert F0 precisely. A source spectrum dictionary, target spectrum dictionary, and target F0 dictionary are constructed by lining up each feature extracted from parallel utterances.

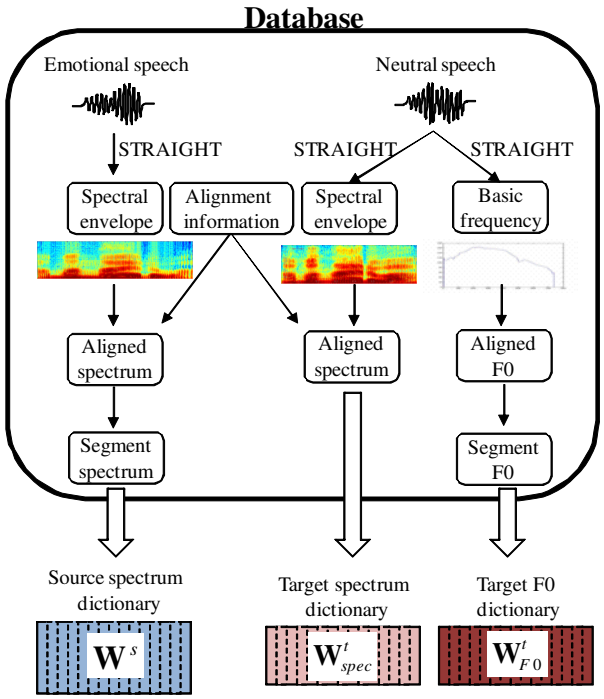

Fig. 3. Dictionary construction for emotional voice conversion

C. Estimation of Activity

Input emotional spectrum vector $\mathbf{v}^s$ is represented by a linear combination of basis in the activity set A using ASNA-NMF as follows.

$$\mathbf{v}^s \approx \sum_{j \in A} \mathbf{w}^s_j h^s_j \tag{8}$$

In order to estimate activity in the active set, a Newton algorithm is used. Before the activity estimation, the source dictionary $\mathbf{W}_s$ is normalized to unity norm, and input spectra $\mathbf{V}_s$ are normalized for each frame so that the sum of the magnitudes over frequency bins equals unity.

First, the active set is initialized with a single basis from the source dictionary minimizes (5) where the weight of each basis that minimizes the following equation,

$$h^s_j = \frac{\mathbf{1}^\mathsf{T}\mathbf{v}^s}{\mathbf{1}^\mathsf{T}\mathbf{w}^s_j + \lambda}. \tag{9}$$

The basis that gives the lowest weight (9) is added to the active set and the corresponding weight is used as its activity.

The active set is updated by adding the most promising basis not in the active set yet. In other words, it involves adding a basis that has the most negative partial derivative of (5) with respect to h_j given as

$$\frac{\partial f(\mathbf{h}^s)}{\partial h^s_j} = \mathbf{w}^{s\mathsf{T}}_j(\mathbf{1} - \frac{\mathbf{v}^s}{\mathbf{W}^s\mathbf{h}^s}) + \lambda. \tag{10}$$

The weight of the added basis is initialized with a small value 10^{-15}. If all partial derivatives are positive, no basis is added to the active set.

Next, the weights of bases in the active set are updated. The gradient of the cost function (5) with respect to the activities of the active bases is calculated as follows

$$\nabla_{\mathbf{h}^s_A} = \mathbf{W}^{s\mathsf{T}}_A(\mathbf{1} - \mathrm{diag}(\frac{\mathbf{v}^s}{\mathbf{W}^s_A\mathbf{h}^s_A})) + \lambda. \tag{11}$$

where $\nabla_{\mathbf{h}^s_A}$, $\mathbf{W}^s_A$, and $\mathbf{h}^s_A$ represent the gradient, bases in the active set, and the activities of bases in the active set, respectively. The Hessian matrix with respect to $\mathbf{h}^s_A$ is calculated as follows;

$$\mathbf{H}es_A = \mathbf{W}^{s\mathsf{T}}_A \mathrm{diag}(\frac{\mathbf{v}^s}{(\mathbf{W}^s_A\mathbf{h}^s_A)^2})\mathbf{W}^s_A. \tag{12}$$

The weights of bases in the active set are updated using the gradient and the Hessian as follows;

$$\mathbf{h}^s_{A,n+1} = \mathbf{h}^s_{A,n} - \alpha\mathbf{H}es^{-1}_A\nabla_{\mathbf{h}^s_A}. \tag{13}$$

α is a step size which is calculated as follows

$$\alpha = \min_{r_d>0} r_d \ where\ \mathbf{r} = \mathbf{h}^s_A/\mathbf{H}es^{-1}_A\nabla_{\mathbf{h}^s_A}. \tag{14}$$

According to the standard Newton algorithm, the step size is limited to 1. When the activity of the basis in the active set becomes, the basis is removed from the active set. In order to ensure the numerical stability in (12), an identity matrix multiplied by small positive constant 10^{-9} is added to the Hessian.

D. Spectrum and Prosody Conversion

The target spectrum dictionary $\mathbf{W}^t_{spect}$ is also normalized for each frame in the same way the source dictionary is. By using estimated activities $\mathbf{H}^s$ in (13), the target spectral feature $\hat{\mathbf{V}}^t_{spec}$ is constructed as follows;

$$\hat{\mathbf{V}}^t_{spec} = \mathbf{W}^t_{spec}\mathbf{H}^s. \tag{15}$$

Then, the magnitudes of the source signal are applied to the normalized target spectral feature.

For F0 conversion, one single frame is selected from the activity of each basis. Activities for F0 conversion $\hat{\mathbf{H}}^s$ are calculated as follows;

$$\hat{h}^s_{j,l} = \begin{cases} 1 & (h_{j,l} = \max \mathbf{h}^s_l) \\ 0 & (otherwise) \end{cases} \tag{16}$$

Target F0 is constructed as follows;

$$\hat{\mathbf{V}}^t_{F0} = \mathbf{W}^t_{F0}\hat{\mathbf{H}}^s. \tag{17}$$

Then, the segment F0 is converted to a single frame in order to construct the target emotional F0.

V. Experimental Results

A. Database

We used a database ofemotional Japanese speech constructed in [26]. From the database, we chose angry and neutral voices. A female professional narrator was asked to read the text set with emotion.

For the test data, 61 Japanese sentences uttered in anger were collected. Fifty sentences from the ATR Japanese phonetically balanced text set [30] were chosen as training data. These 50 sentences are designed to include a minimum phone set of Japanese. All the text are read with angry and neutral voices. In [26], a subjective evaluation shows that the recorded speech contains the target emotion correctly.

B. Experimental Conditions

The proposed method was evaluated on sentence-based VC for one person speaking with emotion. We set anger as the input emotion and neutral as the target emotion. The speech signals were sampled at 12 kHz and windowed with a 25-msec Hamming window every 10 msec. In the method based on NMF, the spectrum extracted by STRAIGHT was used. The number of spectrum dimensions was 513. For the source spectrum, before and after 2 frames were made up as segment spectrum and for target F0, before and after 20 frames were made up as segment F0. The Mel-cepstral coefficient, which was converted from the STRAIGHT spectrum, was used for DTW in order to align temporal fluctuations.

For this paper, objective and subjective evaluations were conducted. In both evaluations, 20 sentences were randomly selected from the test data. In the subjective evaluation, a total of 10 Japanese speakers took part in the test using headphones.

C. Objective Evaluation

Cepstrum distortion represented as the following equation was used for the objective evaluation of spectrum conversion.

$$CepD = (20/\log 10)\sqrt{2\sum_d (c_d^{conv} - c_d^{tar})^2} \tag{18}$$

where c_d^{conv} and c_d^{tar} denote the cepstrum coefficients of original/converted and target voice. In this paper, the number of dimensions of the cepstrum coefficients was 24. Fig. 13 shows the result of the cepstrum distortion test. As shown in the figure, both EM-NMF-based VC and the proposed VC method converted an emotional voice to a neutral voice effectively.

Root mean square error (RMSE) was used for the objective evaluation of F0 conversion. Fig. 5 shows RSME results. As shown in the figure, both EM-NMF-based VC and the proposed VC method converted emotional F0 to neutral F0 effectively.

Fig. 6 shows the computational times for estimating activities of one sentence for both methods. As shown in the figure, ASNA-NMF is 8 times faster than EM-NMF.

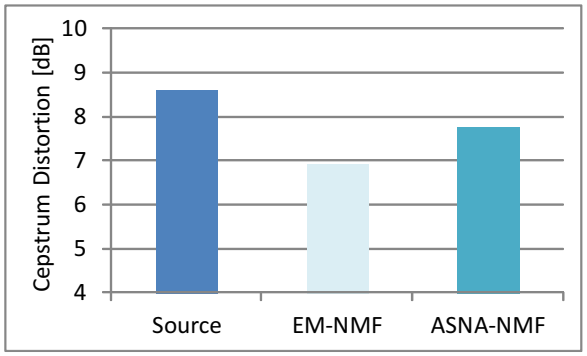

Fig. 4. Results of cepstrum distortion

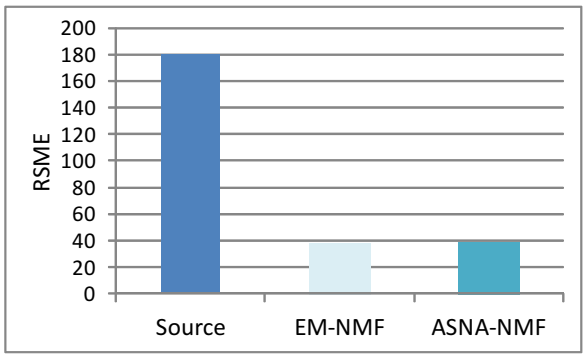

Fig. 5. Results of RSME

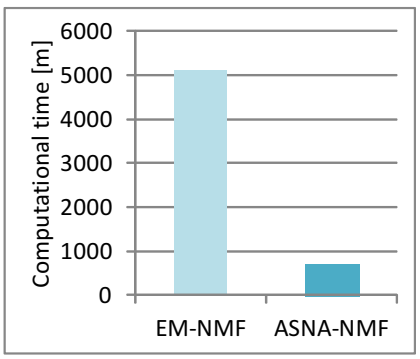

Fig. 6. Results of computational time

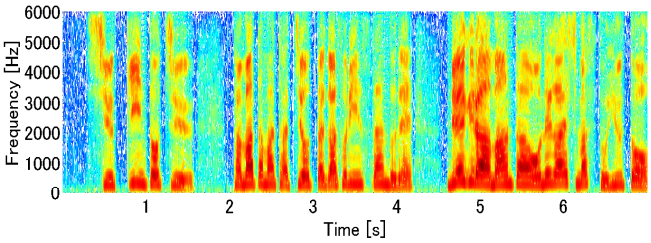

Fig. 7. Example of spectrogram spoken with anger

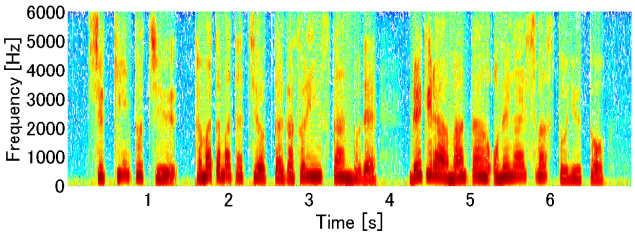

Fig. 8. Example of spectrogram spoken with neutrality

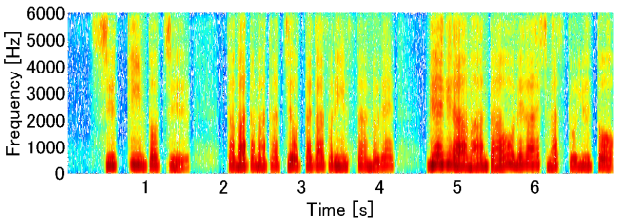

Fig. 9. Example of converted spectrogram

D. Subjective Evaluation

Figs. 7, 8 and 9 show examples of spectrogram spoken with angry, neutral and converted voices, respectively. Figs. 10, 11 and 12 show examples of F0 spoken with angry, neutral and converted voices, respectively.

We performed a MOS (Mean Opinion Score) test as a subjective evaluation. The opinion score was set to a 5-point scale (5:very neutral 4: neutral, 3: fair, 2: angry, 1: very angry). Fig. 13 shows the results of the MOS test. "Source" implies an input angry voice from the database. "Target" implies a target neutral voice from the database. In "Spectrum", only the spectrum is converted to neutral using our method. In "F0", only F0 is converted to neutral using our method. In "Proposed", both the spectrum and F0 are converted using the method proposed in this paper. The error bars show 95% confidence intervals.

From "Source" and "Target" in Fig. 13, it is confirmed that the input and target speech data contain the emotion of anger and neutrality. The MOS of "Spectrum" and "F0" is around 2: angry. This result shows that converting the spectrum or F0 only is not effective. The MOS of "Proposed" is significantly better than "Source" and "Target", and its MOS is around 3 : fair. This result shows that our proposed method effectively converted an angry voice to a neutral voice.

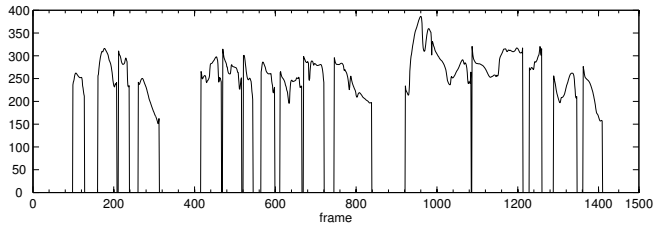

Fig. 10. Example of F0 spoken with anger

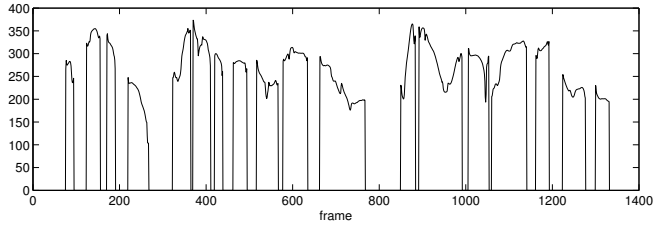

Fig. 11. Example of F0 spoken with neutrality

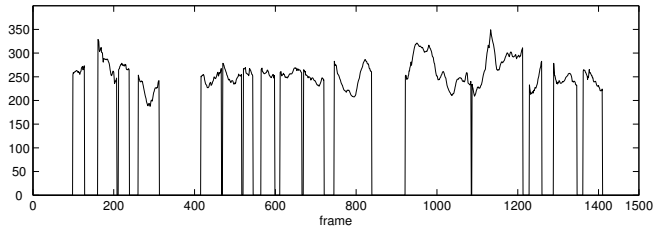

Fig. 12. Example of converted F0

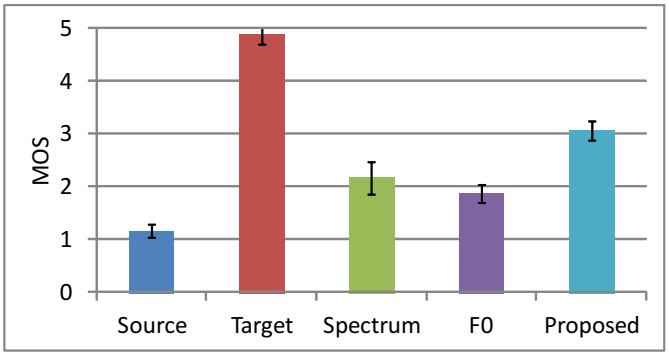

Fig. 13. Results of MOS test

E. Discussion

From the objective evaluation, NMF-based VC can convert an emotional voice to a neutral voice. Our proposed ASNA-NMF is a little worse in spectrum conversion than EM-NMF, but there is no significant difference in F0 conversion. The computational time of our proposed ASNA-NMF VC is significantly higher than EM-NMF VC.

From the subjective evaluations, the converting spectrum of F0 only is not effective in emotional VC. This result shows that our VC method can convert the angry voice into the neutral voice.

VI. Conclusions

We have introduced a preliminary demonstration of an exemplar-based emotional VC method using NMF. In our proposed method, the input emotional spectral feature can be represented by smaller numbers of exemplars compared to conventional EM-NMF-based VC. Objective and subjective evaluations show the effectiveness of our method. In particular, objective evaluation shows that our proposed VC method was about 8 times faster than that EM-NMF-based method.

There is still some problem in our proposed VC approach, however. In [31], we proposed a framework to train basis matrices of source and target exemplars in order to reduce computational cost. In future work, we will combine this method and the method proposed in this paper, and then we will investigate the optimal number of bases and evaluate the performances.

In addition, this method has a limitation in that it can be applied to only one-to-one voice conversation because it requires parallel speech data having the same texts uttered by the source and target speakers. Hence, we will research a method that does not use parallel data.

Comparing our VC approach to the other conventional VC methods in an emotional VC task will also our future work.

Acknowledgment

The authors are grateful to Prof. Tomoki Toda for permission to use the Japanese Emotional Database [26]

References

[1] A. W. Black and N. Cambpbell, "Optimising selection of units from speech database for concatenative synthesis," *in EUROSPEECH*, pp. 581–584, 1995.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *in ICASSP*, pp. 1315–1318, 2000.

[3] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech Audio Lang. Process.*, pp. 1208–1230, 2009.

[4] R. Barra-Chicote, J. Yamagichi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, pp. 394–404, 2010.

[5] 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.

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

[7] M. E. Ayadia, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, 2011.

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

[9] 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.

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

[11] 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.

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

[13] T. Virtanen, B. Raj, J. F. Gemmeke, and H. Van hamme, "Active-set newton algorithm for non-negative sparse coding of audio," *in ICASSP*, pp. 3116–3120, 2014.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, "Multimodal voice conversion using non-negative matrix factorization in noisy environments," *ICASSP2014*, pp. 1561–1565, 2014.

[23] 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.

[24] S. Mori, T. Moriyama, and S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," *IEEE Conference on Multimedia and Expo*, pp. 1093–1096, 2006.

[25] C. H. Wu, C. C. Hsia, and C. H. Lee, "Hierarchical prosody conversion using regression-based clustering for emotional synthesis," *IEEE Trans. Audio, Speech and Lang Proc.*, 2010.

[26] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *IEEE Trans. Speech and Audio Proc.*, 1999.

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

[28] 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, pp. 134–138, 2012.

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

[30] 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.

[31] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," *SSW8, 8th ISCA Speech Synthesis Workshop*, pp. 71–75, 2013.