Proceedings of the 2013 IEEE/SICE International Symposium on System Integration, Kobe International Conference Center, Kobe, Japan, December 15-17,

Voice Conversion based on Non-negative Matrix Factorization in Noisy Environments

Takao Fujii, Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

Abstract— **This paper presents a voice conversion (VC) technique for noisy environments. We prepared parallel exemplars (dictionary) that consist of the source and target exemplars, which have the same texts uttered by the source and target speakers. The input source signal is decomposed into the source exemplars, noise exemplars obtained from the input signal, and their weights (activities). Then, the converted signal is obtained by calculating the linear combination of the target exemplars and the weights which are calculated using the source exemplars. In the proposed method, a Gaussian Mixture Model (GMM) -based conversion method is also applied to the feature vectors generated by the sparse coding in order to compensate a mismatch between the weights of source and target exemplars. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional method.**

I. INTRODUCTION

Voice conversion (VC) is a technique for changing specific information in an input speech while maintaining the other information in the utterance, such as its linguistic information. The VC techniques have been applied to various tasks, such as speaker conversion, emotion conversion [1], [2], speaking assistance [3], and so on.

Many statistical approaches to VC have been studied [4]-[6]. Among these approaches, the GMM-based mapping approach [6] is widely used, and a number of improvements have been proposed. Toda et al. [7] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [8] proposed transforms based on Partial Least Squares (PLS) in order to prevent the over-fitting problem of standard multivariate regression. There have also been approaches that do not require parallel data that make use of GMM adaptation techniques [9] or eigen-voice GMM (EV-GMM) [10], [11].

However, the effectiveness of these approaches was confirmed with clean speech data, and the utilization in noisy environments was not considered. The noise in the input signal is not only output with the converted signal, but may also degrade the conversion performance itself due to unexpected mapping of source features. Hence, a VC technique that takes into consideration the effect of noise is of interest.

Recently, approaches based on sparse representations have gained interest in a broad range of signal processing. In the field of speech processing, Non-negative Matrix Factorization (NMF) [12] is a well-known approach for source separation and speech enhancement [13], [14]. In these approaches, the observed signal is represented by a linear combination of a small number of atoms, such as the exemplar and basis of NMF. 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. [15] also proposes an exemplar-based method for noise robust speech recognition. In that method, the observed speech is decomposed into the 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 this paper, we propose an exemplar-based VC approach for noisy source signals. The parallel exemplars (called 'dictionary' in this paper), which consist of source exemplars and target exemplars, are extracted from the parallel data that were used as training data in conventional GMM-based approaches. Also, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. For this reason, no training processes for the noise signal are required. The input source signal is expressed with a sparse representation of the source exemplars and noise exemplars. Only the weights (called 'activity' in this paper) related to the source exemplars are picked up, and the target signal is constructed from the target exemplars and the picked-up weights. In addition, a GMM-based conversion method is also applied to the feature vectors generated by the sparse coding in order to approximate to target features. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional method.

II. PROPOSED METHOD

A. Voice Conversion based on Sparse Coding

In this section, we give expression to voice conversion based on sparse coding [16]. In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of atoms.

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

$\mathbf{x}_l$ is the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ are the j-th atom 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 atoms and the stack of weights. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination

T. Fujii, R. Aihara, R. Takashima, T. Takiguchi and Y. Ariki are with Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada, Kobe, Hyogo 657-8501, Japan
[fujii, aihara]@me.cs.scitec.kobe-u.ac.jp,
[ariki, takigu]@kobe-u.ac.jp

978-1-4799-2625-1/13/$31.00 ©2013 IEEE

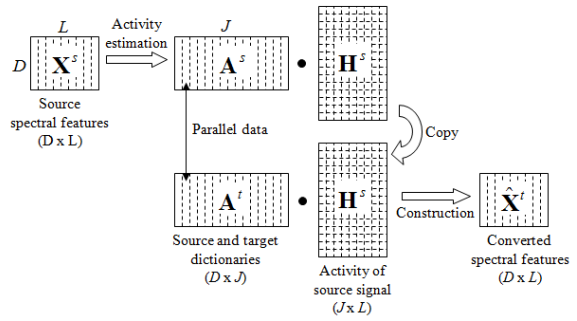

Fig. 1. Basic approach of exemplar-based voice conversion

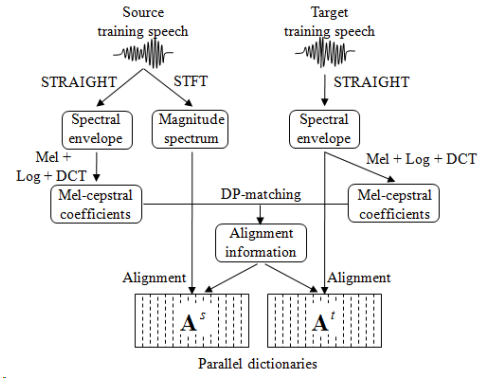

Fig. 2. Construction of source and target dictionaries

of a small number of atoms that have non-zero weights. In this paper, each atom denotes the exemplar of speech or of a noise signal, and the collection of exemplar $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are referred to as the '*dictionary*' and '*activity*', respectively.

In our proposed method, the parallel exemplars (dictionaries) are used to map the source signal to the target one. The parallel dictionaries consist of source and target dictionaries that have the same size.

Fig. 1 shows the basic approach of voice conversion based on sparse coding. The activity matrices are estimated from source words and the source dictionary. When there are parallel dictionaries constructed from source and target speech features, the activity of the source signal estimated with the source dictionary may be able to be substituted for that of the target signal. Therefore, the target speech can be constructed by using the target dictionary and the activity of the source signal.

B. Dictionary Construction

The parallel dictionaries are constructed from source and target spectral envelopes extracted by STRAIGHT analysis [17]. The use of these features worked without any problems in a preliminary experiment using clean speech data. However, when it came to constructing a noise dictionary, STRAIGHT analysis could not express the noise spectrum well since STRAIGHT itself is an analysis and synthesis method for speech data. In order to express the noisy source speech with a sparse representation of source and noise dictionaries, a simple magnitude spectrum calculated by short-time Fourier transform (STFT) is used to construct the source and noise dictionaries.

Fig. 2 shows the process for constructing parallel dictionaries. For the target training speech, the STRAIGHT spectrum is used to extract its dictionary. For the source training speech, on the other hand, the STRAIGHT spectrum is converted into mel-cepstral coefficients and only used for DP-matching in order to align the temporal fluctuation, and the magnitude spectrum is used to extract its dictionary. When an input source signal is converted, the source signal is also applied to STFT and STRAIGHT analysis. The magnitude spectrum is used to extract the noise dictionary and used to estimate the activity. The STRAIGHT spectrum is not used in the conversion process, but the other features extracted by STRAIGHT analysis, such as F0 and aperiodic components, are used to synthesize the converted signal.

C. Estimation of Activity from Noisy Source Signals

From the before- and after-utterance sections in the observed (noisy) signal, the noise dictionary is extracted for each utterance. In the exemplar-based approach, the spectrum of the noisy source signal at frame l is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their activities.

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\ &\approx \sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s + \sum_{k=1}^{K} \mathbf{a}_k^n h_{k,l}^n \\ &= [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\ &= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{2}$$

$\mathbf{x}_l^s$ and $\mathbf{x}_l^n$ are the magnitude spectra of the source signal and the noise. A^s, A^n, h_l^s, and h_l^n are the source dictionary, noise dictionary, and their activities at frame l, respectively. Given the spectrogram, (2) can be written as follows:

$$\begin{aligned}\mathbf{X} &\approx [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \end{aligned} \tag{3}$$

In order to consider only the shape of the spectrum, X, A^s and A^n are first normalized for each frame or exemplar so that the sum of the magnitudes over frequency bins equals unity.

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X} \\ \mathbf{X} &\leftarrow \mathbf{X}./\mathbf{M} \\ \mathbf{A} &\leftarrow \mathbf{A}./(\mathbf{1}^{(D\times D)}\mathbf{A}) \end{aligned} \tag{4}$$

$\mathbf{1}$ is an all-one matrix. ./ denotes element-wise division. The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function [15]:

$$d(\mathbf{X}, \mathbf{A}\mathbf{H}) + ||(\lambda \mathbf{1}^{(1\times L)}) .* \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{5}$$

The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}$ and $\mathbf{A}\mathbf{H}$. The second term is the sparse constraint

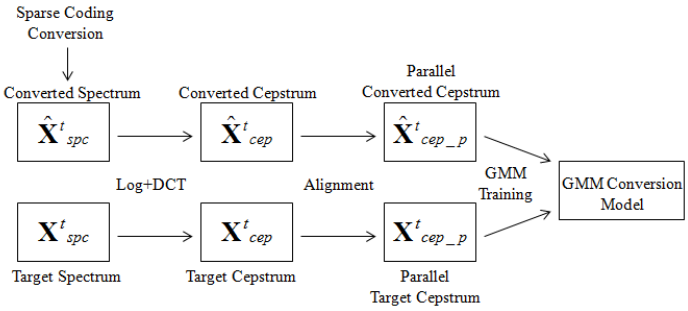

Fig. 3. Training of GMM conversion model after Sparse Coding conversion

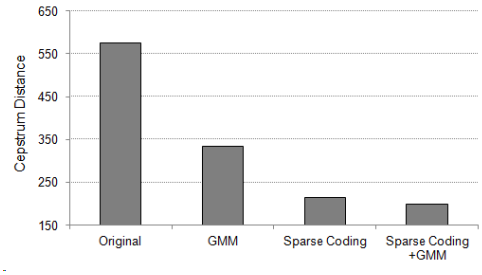

Fig. 4. Cepstrum distance for each method in the case of using 50 words (10 dB)

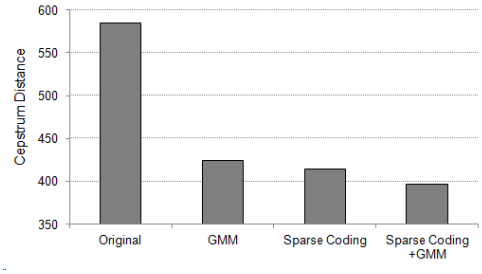

Fig. 5. Cepstrum distance for each method in the case of using 25 sentences (10 dB)

with the L1-norm regularization term that causes $\mathbf{H}$ to be sparse. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \dots \lambda_J \dots \lambda_{J+K}]$. In this paper, the weights for source exemplars $[\lambda_1 \dots \lambda_J]$ were set to 0.2, and those for noise exemplars $[\lambda_{J+1} \dots \lambda_{J+K}]$ were set to 0. $\mathbf{H}$ minimizing (5) is estimated iteratively applying the following update rule:

$$\begin{aligned}\mathbf{H}_{n+1} &= \mathbf{H}_n .* (\mathbf{A}^T(\mathbf{X}./(\mathbf{AH}))) \\ &\quad ./(\mathbf{1}^{((J+K)\times L)} + \lambda \mathbf{1}^{(1\times L)}).\end{aligned} \tag{6}$$

D. Target Speech Construction

From the estimated joint matrix $\mathbf{H}$, the activity of source signal $\mathbf{H}^s$ is extracted, and by using the activity and the target dictionary, the converted spectral features are constructed. Then, the target dictionary is also normalized for each frame in the same way the source dictionary was.

$$\mathbf{A}^t \leftarrow \mathbf{A}^t./(\mathbf{1}^{(D\times D)}\mathbf{A}^t) \tag{7}$$

Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in (4) are applied to the normalized target spectral feature.

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) .* \mathbf{M} \tag{8}$$

The input source feature is the magnitude spectrum calculated by STFT, but the converted spectral feature is expressed as a STRAIGHT spectrum. Hence, the target speech is synthesized using a STRAIGHT synthesizer. Then, F0 information is converted using a conventional linear regression based on the mean and standard deviation.

E. Compensation of a Mismatch between the Weights of Source and Target Exemplars Based on GMM

Although source and target dictionaries are consisted of parallel data, a mismatch may occur in the estimated weights of source and target exemplars. In order to compensate the mismatch, in this paper, GMM-based conversion is applied to the feature vectors generated by the sparse coding. In the compensation process, as shown in Fig. 3, the spectra converted by the sparse coding and the target speech data are converted to the cepstrum domain, and they are used as training data of GMM. Training of GMM is performed by the same procedure as the conventional VC based on GMM [6].

III. EXPERIMENTS

In the experiments, noise-added speech data was used as an input signal. The proposed VC technique was evaluated by comparing it with a conventional technique based on GMM [6] and the VC technique based on sparse coding without GMM-based conversion [16]. The source speaker and target speaker were one male and one female speaker, whose speech is stored in the ATR Japanese speech database, respectively. The sampling rate was 8 kHz. Two-hundred sixteen words of clean speech were used to construct parallel dictionaries in our proposed method and used to train the GMM in the conventional method. The number of exemplars of source and target dictionaries was 57,033. Fifty words or twenty-five sentences of noisy speech were used as the test dataset. Fifty words were included in parallel dictionaries, and twenty-five sentences were not. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database) to the clean words or sentences. The mean SNR was about 10 dB and 15 dB. The noise dictionary is extracted from the before- and after-utterance sections in the evaluation sentence. The average number of exemplars for the noise dictionary for one sentence was 104. In our proposed method, a 256-dimensional magnitude spectrum was used as the feature vector for the input signal, source dictionary and noise dictionary, and a 512-dimensional STRAIGHT spectrum was used for the

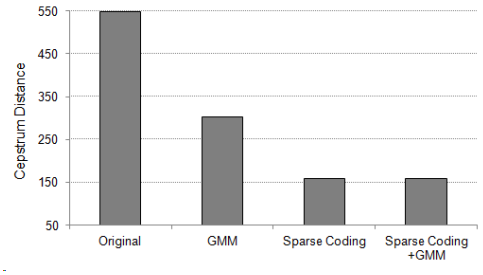

Fig. 6. Cepstrum distance for each method in the case of using 50 words (15 dB)

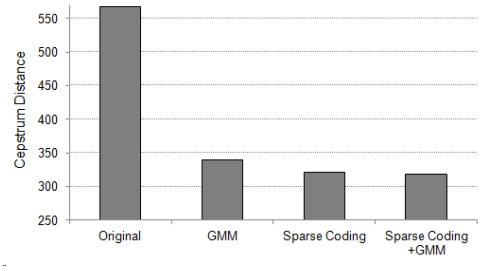

Fig. 7. Cepstrum distance for each method in the case of using 25 sentences (15 dB)

target dictionary. The number of iterations used to estimate the activity was 500.

IV. EXPERIMENTAL RESULTS

Fig. 4 - 7 show the cepstrum distance between target cepstrum and that of a signal converted using each method. As shown in these figures, our proposed method showed a better performance than the conventional method in both the use of fifty words and twenty-five sentences as the test data. In the case of fifty words (test data), a largest cepstrum distance is shown in Fig. 4 when a GMM-based VC is used. This might be because the noise caused unexpected mapping in the GMM-based method. When twenty-five sentences that were not included in parallel dictionaries were used as test data, there was a slight improvement as compared with the conventional method based on GMM.

V. CONCLUSIONS

In this paper, we proposed an exemplar-based VC technique for a noisy environment. This method uses parallel exemplars (dictionaries) that consist of the source and target dictionaries. By using the source dictionary and noise dictionary, only the weights (activity) corresponding to the source dictionary are extracted from the noisy source. The converted speech is constructed from the target dictionary and the activity of the source dictionary. In a comparison experiment, the proposed method showed better performance than the conventional method.

However, this method requires the estimation of activity of each atom in the dictionary, and it requires high computation times. Therefore, we will research ways to reduce the atoms in the dictionary efficiently, and we will try to introduce dynamic information, such as segment features. Also, in our proposed method, a GMM-based conversion method is applied to the feature vectors generated by the sparse coding in order to compensate a mismatch between the weights of source and target exemplars. Future work will also include further efforts to adapt the weights of source exemplars to that of target exemplars.

REFERENCES

[1] Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based Voice Conversion Applied to Emotional Speech Synthesis," IEEE Trans. Seech and Audio Proc., Vol. 7, pp. 2401–2404, 1999.
[2] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in Proc. INTERSPEECH, pp. 2765–2768, 2011.
[3] 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.
[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Vice conversion through vector quantization," in Proc. ICASSP, pp. 655–658, 1988.
[5] H. Valbret, E. Moulines and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, Vol. 11, No. 2-3, pp. 175–187, 1992.
[6] 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.
[7] 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.
[8] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," IEEE Trans. Audo, Speech, Lang. Process., Vol. 18, No. 5, pp. 912–921, 2010.
[9] 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.
[10] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in Proc. INTERSPEECH, pp. 2446–2449, 2006.
[11] 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.
[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. Neural Information Processing System, pp. 556–562, 2001.
[13] 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.
[14] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in Proc. INTERSPEECH, pp. 2614-2617, 2006.
[15] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," IEEE Trans. Audio, Speech, Lang. Process., Vol. 19, Issue 7, pp. 2067–2080, 2011.
[16] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-Based Voice Conversion in Noisy Environment," IEEE Workshop on Spoken Language Technology (SLT2012), pp. 313-317, 2012.
[17] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, Vol.27, pp. 187–207, 1999.