EXEMPLAR-BASED VOICE CONVERSION IN NOISY ENVIRONMENT

Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University, Japan

ABSTRACT

This paper presents a voice conversion (VC) technique for noisy environments, where parallel exemplars are introduced to encode the source speech signal and synthesize the target speech signal. The parallel exemplars (dictionary) consist of the source exemplars and target exemplars, having 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, by using the weights of the source exemplars, the converted signal is constructed from the target exemplars. We carried out speaker conversion tasks using clean speech data and noise-added speech data. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method.

Index Terms— voice conversion, exemplar-based, sparse coding, non-negative matrix factorization, noise robustness

1. INTRODUCTION

Voice conversion (VC) is a technique for changing specific information in an input speech with holding 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 aid [3], and so on.

Many statistical approaches to VC have been studied [4, 5, 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. Helnder 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 does not require parallel data by using 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, the VC technique considering 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 exemplar and basis of NMF. In some approaches for source separation, the atoms are grouped for each source, and the mixed signal are expressed with a sparse representation of these atoms. By using only the weights of 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 Model for speech recognition.

In this paper, we propose an exemplar-based VC approach for noisy source signals. The parallel exemplars (called 'dictionaly' in this paper), which consist of a source exemplars and a 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, any training processes about noise signal are not 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. The effectiveness of this method has been confirmed by comparing it with a conventional method based on GMM in a speaker conversion task using clean speech data and noise-added speech data.

2. PROPOSED METHOD

2.1. Sparse Representations for Voice Conversion

In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small

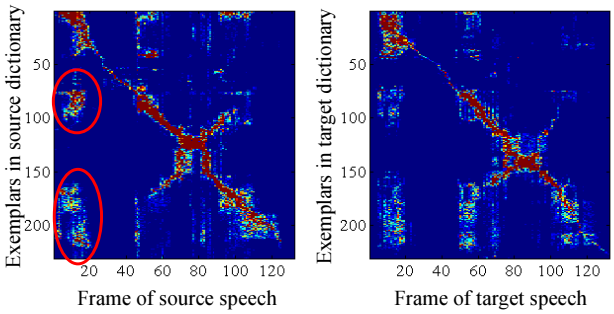

Fig. 1. Activity matrices of the source signal (left) and target signal (right)

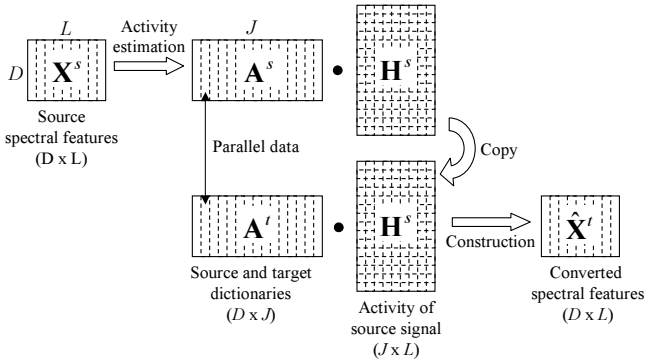

Fig. 2. Basic approach of exemplar-based voice conversion

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 of a small number of atoms that have non-zero weights. In this paper, each atom denotes the exemplar of speech or noise signal, and the collection of exemplar $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are called '*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. Figure 1 shows the activity matrices estimated from the source and target words uttered ('ikioi') and their dictionaries. The parallel dictionaries were structured from the same words aligned using dynamic programming (DP) matching. The source/target features and each atom in the dictionary are a spectral envelope extracted by STRAIGHT analysis [16]. When the source/target signal and its dictionary are the same word, the estimated activity will have high energies through the diagonal line. The reason some areas far from the diagonal line, such as the red-circled areas, also have high energies is that these areas correspond to the same utterance 'i'.

As shown in this figure, these activities have high energies at similar elements. For this reason, when there are parallel dictionaries, 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 as shown in Figure 2. D, L, J are the numbers of dimensions, frames and exemplars, respectively.

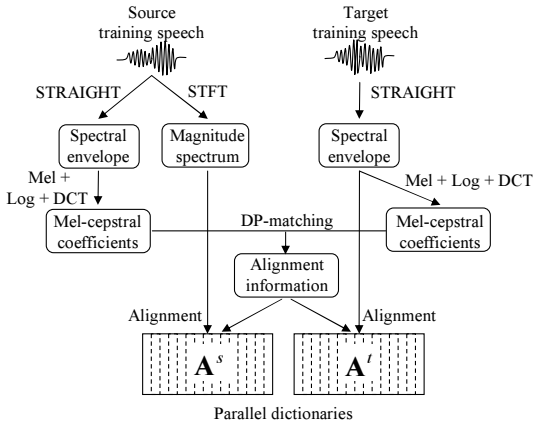

Fig. 3. Construction of source and target dictionaries

2.2. Dictionary Construction

In the preceding section, both dictionaries (source and target) consisted of the same spectral envelope features (STRAIGHT spectrum) for simplicity explaining the proposed method. Indeed, 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.

Figure 3 shows the process for constructing parallel dictionaries. For the target training speech, 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.

2.3. Estimation of Activity from Noisy Source Signal

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, h_l^n are the source dictionary, noise dictionary, and their activities at frame l. 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. 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 with 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 0.1, 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:

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

2.4. 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.

3. EXPERIMENTS

3.1. Experimental Conditions

The new VC technique was evaluated by comparing it with a conventional technique based on GMM [6] in a speaker conversion task using clean speech data and noise-added 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, respectively. The sampling rate was 8 kHz.

216 words of clean speech were used to construct parallel dictionaries in our proposed method and used to train the GMM in conventional method. The number of exemplars of source and target dictionaries was 57,033. 25 sentences of clean speech or noisy speech were used to evaluate. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database) to the clean speech sentences. The mean SNR was about 24 dB. The noise dictionary is extracted from the before and after utterance section in the evaluation sentence. The average number of noise dictionary for one sentence was 104.

In our proposed method, a 256-dimensional magnitude spectrum was used as the feature vectors for input signal, source dictionary and noise dictionary, and a 512-dimensional STRAIGHT spectrum was used for the target dictionary. The number of iterations used to estimate the activity was 500. In

Table 1. Mel-cepstral distortion [dB] for each method

	Original source	Conventional	Proposed
Clean speech	6.30	4.23	3.54
Noisy speech	6.70	4.74	3.97

the GMM-based method, the 1^{st} through 40^{th} linear-cepstral coefficients obtained from the STRAIGHT spectrum were used as the feature vectors.

3.2. Experimental Results

Table 1 shows the mel-cepstral distortion between the target mel-cepstra and that of a signal converted using each method. The mel-cepstral distortion is calculated as follows [7].

$$\text{Mel-CD[dB]} = 10/\ln 10 \sqrt{2 \sum_{d=1}^{24} (mc_d^t - \hat{mc}_d^t)^2} \qquad (9)$$

mc_d^t and $\hat{mc}_d^t$ are the d-th coefficients of the target and converted mel-cepstra, respectively. We calculated mel-cepstra from the converted STRAIGHT spectrum.

As shown in this table, our proposed method showed the lower distortion than the conventional method in both cases using clean speech and noisy speech for evaluation. The mel-cepstral distortion between source signal and target signal increased by 0.40 dB by adding noise signal to the source signal ($6.30 \rightarrow 6.70$ dB). On the other hand, the distortions in the conventional method and proposed method increased by 0.51 dB ($4.23 \rightarrow 4.74$ dB) and 0.43 dB ($3.54 \rightarrow 3.97$ dB) by adding noise signal to the source signal, respectively. The reason these increases were greater than that of original source is that the noise in the input signal is not only output with the converted signal, but also degrade the conversion performance itself due to unexpected mapping of source features. However, our proposed method could suppress the influence of the noise compared to the conventional method.

Next, we carried out the preference tests related to the naturalness and speaker individuality of the converted speech. The tests were carried out with 7 subjects. For the evaluation of naturalness, a paired comparison test was carried out, where each subject listened to pairs of speech converted by the two methods and selected which sample sounded more natural. For the evaluation of speaker individuality, the XAB test was carried out. In the XAB test, each subject listened to the target speech. Then the subject listened to the speech converted by the two methods and selected which sample sounded more similar to the target speech.

Figure 4 and Figure 5 show the preference scores of each method in the case of clean speech and noisy speech, respectively. The error bars show 95% confidence intervals. As shown in Figure 4, in both evaluation criteria, our proposed

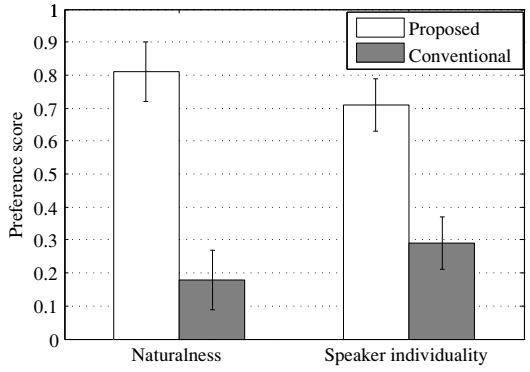

Fig. 4. Preference scores for the naturalness and the speaker individuality for each method in the case of clean speech

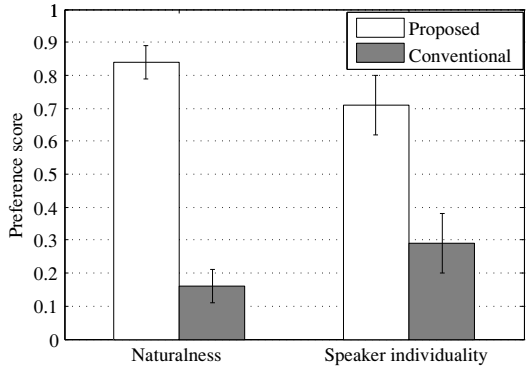

Fig. 5. Preference scores for the naturalness and the speaker individuality for each method in the case of noisy speech

method showed higher scores than the conventional method. As shown in Figure 5, the evaluation of speaker individuality in the case of noisy speech was rarely different from that in the case of clean speech. On the other hand, the preference scores of naturalness biased toward our proposed method greater than those in the case of clean speech since the mean preference score of our proposed method increased and the confidence interval narrowed. This might be because the noise caused unexpected mapping in the GMM-based method, and the speech was converted with a lack of naturalness.

4. 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 is 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 between a conventional GMM-based method and the proposed method, the proposed method showed better performances in both cases using clean speech and noisy speech for evaluation, especially in the naturalness in a noisy environment.

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. In addition, this method has a limitation 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 investigate a method not to use the parallel data. Future work will also include efforts to investigate other noise conditions, such as a low-SNR condition, and apply this method to other VC applications.

Acknowledgment

This research was supported in part by MIC SCOPE.

5. 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] 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.