Exemplar-Based Voice Conversion Using Sparse Representation in Noisy Environments

1. Introduction

Voice conversion (VC) is generally a technique for changing specific information in an input speech while maintaining the other information in the utterance, such as its linguistic information. One of the most popular applications using the VC technique is speaker conversion, where an utterance spoken by a source speaker is morphed so that it sounds as if it had been spoken by a specified target speaker. There have also been studies on various tasks, such as emotion conversion ([1], [2]), speaking assistance ([3], [4]), and so on, which make use of VC techniques.

Many statistical approaches to VC have been studied ([5]–[7]). Among these approaches, the GMM-based mapping approach [7] is widely used, and a number of improvements have been proposed. Toda et al. [8] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [9] 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 [10] or eigen-voice GMM (EV-GMM) ([11], [12]).

However, the effectiveness of these approaches was confirmed with clean speech data, and the utilization in noisy environments was not considered. Depending on the purpose of using VC system, the VC system will be utilized outside or in any noisy environments in the future. When the VC system is used in noisy environments, the noise in the input signal is output with the converted signal, and as a result, it may degrade the conversion performance. 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) [13] is a well-known approach for source separation and speech enhancement ([14], [15]). 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. [16] 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-utterance section 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. 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. Related Works

In recent statistical VC approaches ([5]–[7]), there are two types of VC methods: the codebook mapping-based method and transformation-based method. Abe et al. [5] proposed a
VC method based on a codebook mapping. In that approach, the source feature vector $x_l$ at frame $l$ is quantized to the nearest centroid vector $c^*_m$ in the source codebook. Then, the target feature vector $\hat{x}_l$ is determined by selecting the centroid vector $c^*_m$, which corresponds to $c^*_m$, in the target codebook.

$$\hat{x}_l = c^*_m$$  
(1)

$$m = \arg\min_j \text{dist}(x_l, c^*_j)$$  
(2)

Here, $\text{dist}(x, c)$ represents the distance between the vectors $x$ and $c$.

Nakamura et al. [17] proposed fuzzy vector quantization, which reduces the influence of the quantization error in the codebook mapping method. In that method, the converted feature vector is defined as a weighted sum of the centroid vectors in the mapping codebook:

$$\hat{x}_l = \sum_m w_{lm}^{f} x_m$$  
(3)

where $w_{lm}^{f}$ is the weight calculated from the distance between the source feature vector and the $m$-th centroid vector in the source codebook. However, codebook mapping-based methods do not represent the variations of the source speech minutely because those methods represent the converted speech by using only limited centroid vectors in the target codebook.

For more variable representations, transformation-based VC approaches have been studied and often show better conversion performances than codebook-mapping-based approaches. In those approaches, the source feature vector is transformed using the non-linear mapping function:

$$\hat{x}_l = f(x_l).$$  
(4)

This non-linear mapping function is often realized with piecewise linear mapping functions. Valbret et al. [6] proposed a VC method based on the piecewise linear multivariate regression with hard clustering:

$$\hat{x}_l = R_m x^f_l + u_m$$  
(5)

where $R_m$ and $u_m$ are the regression parameters related to the cluster to which the source vector belongs. VC methods using neural networks [18]–[20] have been proposed. Mapping based on the neural networks can also be considered as a piecewise linear transformation method.

Stylianou et al. [7] proposed a VC method based on a GMM. The GMM-based VC method is an approach that uses the piecewise linear mapping function with soft clustering:

$$\hat{x}_l = \sum_m w_{lm}^{s} (R_m x^f_l + u_m).$$  
(6)

The GMM-based approach is widely used in many studies related to VC, and a number of improvements have been proposed. Toda et al. [8] introduced dynamic features and the global variance of the converted spectra over a time sequence. Helander et al. [9] proposed transforms based on partial least squares and GMM in order to prevent the over-fitting problem of standard multivariate regression. There have also been approaches that do not require parallel data that use GMM adaptation techniques [10] or eigen-voice GMM ([11], [12]).

3. Proposed Method

3.1 Sparse Representations for Voice Conversion

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

$$x_l \approx \sum_{j=1}^{J} a_{j} h_{j,l} = A h_l$$  
(7)

$x_l$ is the $l$-th frame of the observation. $a_j$ and $h_{j,l}$ are the $j$-th atom and the weight, respectively. $A = [a_1 \ldots a_J]$ and $h_l = [h_{1,l} \ldots h_{J,l}]^T$ are the collection of the atoms 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 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 $A$ and the weight vector $h_l$ 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. Figure 1 shows the activity matrices estimated from the source and target words that were uttered (‘ikioi’) and their dictionaries. The parallel dictionaries were structured from the same words and aligned using dynamic programming (DP) matching. The source/target features and each atom in the dictionary are a spectral envelope extracted by STRAIGHT analysis [21]. 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’.

![Fig. 1 Activity matrices of the source signal (left) and target signal (right).](image-url)
1948

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 Fig. 2. $D, L, J$ are the numbers of dimensions, frames and exemplars, respectively.

3.2 Dictionary Construction

In Fig. 2, both dictionaries (source and target) consist of the spectral envelope features (STRAIGHT spectrum). 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 from the STRAIGHT spectrum by applying mel-filter bank, logarithmic transform and discrete cosine transform.

When an input source signal is converted, the source signal with a sparse representation of source and noise dictionaries. In this paper, mel-cepstral coefficients are calcu-

3.3 Estimation of Activity from Noisy Source Signals

From the before-utterance section 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 $i$ is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their activities at frame $i$.

$$x_i = x_i^s + x_i^n$$

$$\approx \sum_{j=1}^{J} a_{i,j} h_{i,j} + \sum_{k=1}^{K} a_{i,k} h_{i,k}$$

$$= [A^s A^n] \begin{bmatrix} h_i^n \\ h_i^s \end{bmatrix} \quad \text{s.t.} \quad h_i^n, h_i^s \geq 0$$

$$= A h_i \quad \text{s.t.} \quad h_i \geq 0 \quad (8)$$

$x_i^s$ and $x_i^n$ are the magnitude spectra of the source signal and the noise, respectively. $A^s$, $A^n$, $h_i^n$ and $h_i^s$ are the source dictionary, noise dictionary, and their activities at frame $i$, respectively. Given the spectrogram, Eq. (8) can be written as follows:

$$X \approx [A^s A^n] [H^n, H^s] \quad \text{s.t.} \quad H^n, H^s \geq 0$$

$$= AH \quad \text{s.t.} \quad H \geq 0 \quad (9)$$

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 [16].

$$m_i = \sum_{d=1}^{D} x_{i,d}, \quad x_{i,d} \leftarrow \frac{x_{i,d}}{m_i}$$

$$a_{d,i}^s \leftarrow \frac{a_{d,i}^s}{\sum_{d=1}^{D} a_{d,i}^s}, \quad a_{d,k}^n \leftarrow \frac{a_{d,k}^n}{\sum_{d=1}^{D} a_{d,k}^n} \quad (10)$$

d is the index of the dimension. $x_{i,j}$, $a_{i,j}^s$ and $a_{i,j}^n$ are $(i, j)$ elements of $X$, $A^s$ and $A^n$, respectively. The joint matrix $H$ is estimated based on NMF with the sparse constraint that minimizes the following cost function:

$$d(X, AH) + \sum_{j=1}^{J} \sum_{d=1}^{D} |a_{d,j}^s h_{d,j}| \quad \text{s.t.} \quad H \geq 0 \quad (11)$$
The first term is the Kullback-Leibler (KL) divergence between $X$ and $AH$. The second term is the sparse constraint with the L1-norm regularization term that causes $H$ to be sparse. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_J]$. In this paper, the weights for source exemplars $[\lambda_1 \ldots \lambda_J]$ were set to 0.1, and those for noise exemplars $[\lambda_{j1} \ldots \lambda_{jK}]$ were set to 0. $H$ minimizing Eq. (11) is estimated iteratively applying the following update rule:

$$
\hat{h}_{dl} = \frac{\sum_{d=1}^{D} a_{d,j} \cdot x_{d,l} / (AH)_{dl}}{1 + \lambda_j}
$$

(12)

$(AH)_{dl}$ is the $(d, l)$ element of $AH$.

3.4 Target Speech Construction

From the estimated joint matrix $H$, the activity of source signal $H^T$ 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.

$$
da'_{d,j} = \frac{a'_{d,j}}{\sum_{d=1}^{D} a'_{d,j}}
$$

(13)

Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in Eq. (10) are applied to the normalized target spectral feature.

$$
x'_{d,l} = m_l \sum_{j=1}^{J} a'_{d,j} h'_{j,l}
$$

(14)

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.

3.5 Differences from Related Works

As shown in Eqs. (3) and (14), our proposed method is similar to the codebook mapping-based methods in terms of the linear combination of feature vectors of the target speech. But the proposed method is different from these approaches in two main ways: the use of exemplars instead of centroids, and the formulation to calculate the weights.

In conventional codebook mapping-based methods, the training source and target speech data are quantized into the centroid vectors using hard or soft clustering. The converted speech synthesized only from limited centroid vectors does not represent the variations of the source speech minutely. On the other hand, our proposed method uses the exemplars that consist of all the training data without any quantizations, and the proposed method can represent the variation of the source speech more minutely than conventional codebook mapping-based methods.

In both the codebook mapping-based method (Eq. (3)) and transformation-based method (Eq. (6)) with soft clustering, the weight of each cluster is often calculated using the distance from the centroid or from Gaussian occupancy probability. This weighted sum is a kind of smoothing procedure, and it sometimes causes an unnatural conversion. As shown in Eq. (11), the weights in our proposed method are calculated so that the reconstructed source signal is similar to the input source signal. For this reason, our proposed method may be able to construct more natural target speech. Furthermore, this method can suppress the noise in the source signal by using noise exemplars like many NMF-based noise suppression techniques [14], [15].

4. Experiments

4.1 Experimental Conditions

The new VC technique was evaluated by comparing it with a conventional technique based on GMM [7] in a speaker conversion task using clean speech data and noise-added speech data. We conducted speaker conversion experiments for 3 pairs of source speaker and target speaker. The 3 speaker pairs consisted of “male (MMY) to female (FTK)”, “male (MMY) to male (MHT)” and “female (FKN) to female (FTK)”, whose speech is stored in the ATR Japanese speech database (B set) [22], respectively.

The noisy speech was created by adding a noise signal recorded in a restaurant and a street (taken from the CENSREC-1-C database [23]) to the clean speech sentences. The SNR was 20 dB and 10 dB. The sampling frequency of the noise signal in CENSREC-1-C database was 8 kHz. For this reason, we applied downsampling to the speech data in order to make the sampling frequency to be the same as that of the noise signal. This downsampling may degrade the quality of the synthesized speech. However, that may not make a serious problem for the purpose of comparing the performances of proposed method with that of the conventional method.

Fifty sentences of clean speech were used to construct parallel dictionaries in our proposed method and to train the GMM in the conventional method. For each pair of source and target speaker, the number of exemplars of the parallel dictionary was around 69,000. Twenty-five sentences of clean speech or noisy speech were used for the evaluation experiment. The noise dictionary is extracted from the before-utterance section in the evaluation sentence. The average number of exemplars for the noise dictionary for one sentence was 45.

The analysis window was a Hamming window having the length of 40 msec. In our proposed method, for each analysis window, 256-dimensional magnitude spectrum was computed and used as the feature vector for the input signal, source dictionary and noise dictionary. The number of dimension of the STRAIGHT spectrum was set to 512 in order to synthesize the output signal having high quality. The frame shift was 5 msec.
The number of iterations used to estimate the activity was 500. In the GMM-based method, the 1st through 40th linear-cepstral coefficients obtained from the STRAIGHT spectrum were used as the feature vectors. The number of mixtures was set to 64 empirically. We used diagonal covariances for training the GMM.

4.2 Experimental Results

First, the optimal hyper-parameters \([\lambda_1 \ldots \lambda_J]\) in Eq. (11) were determined by evaluating the mel-cepstral distortion between the target mel-cepstra and that of a signal converted using our proposed method. The mel-cepstral distortion is calculated as follows [8].

\[
\text{Mel-CD [dB]} = 10 \log_{10} \left( \frac{1}{J} \sum_{j=1}^{J} (c_j - c'_j)^2 \right) \tag{15}
\]

\(c_j\) and \(c'_j\) are the \(d\)-th coefficients of the target and converted mel-cepstra, respectively. We calculated mel-cepstra of the converted signal from the converted (absolute) STRAIGHT spectrum by applying mel-filter bank.

Figure 4 shows the mel-cepstral distortion as a function of the weights of the sparsity constraints under the restaurant noise environments with SNR of 20 dB. In this experiment, only the weights of the sparsity constraints for source exemplars \([\lambda_1 \ldots \lambda_J]\) were changed. The weights for noise exemplars \([\lambda_{J+1} \ldots \lambda_{J+K}]\) were fixed to 0 because the noise signal is not sparse. “\(\lambda = 0\)” means the sparsity constraints are not used. As shown in this figure, the sparsity constraints worked better. However, too strong sparsity constraints tended to distort the output signal and rather degraded the conversion performance. We also evaluated the weights of the sparsity constraints under the clean speech environment. In the clean speech environment, the performance of each parameter setting was almost the same even when the weights were 0. From the next experiments, the weights of the sparsity constraints for source exemplars were set to 0.1.

We compared the performance of our proposed method with that of GMM-based method. For the GMM-based method, we also evaluated the performance of the spectral subtraction (SS) as a front-end noise reduction processing (SS-GMM). In this method, the noise in the observed source signal was suppressed using the SS with noise observations at the before-utterance section in the evaluation sentence at the same condition of our proposed method. Then, the noise-suppressed source signal was converted into the target speech using the GMM-based method. Tables 1, 2 and 3 show the mel-cepstral distortion between the target mel-cepstra and that of a signal converted using each method.

As shown in Tables 1, 2 and 3, our proposed method showed a lower distortion than the conventional method under all experimental conditions. The difference in the mel-cepstral distortion between our proposed method and the GMM-based method is little in the clean speech environment. However, the difference became larger in noisy environments. This is because the noise in the input signal degraded the conversion performance. However, our proposed method could reduce the power of the noise signal, and as a result, our proposed method could suppress the influence of the noise to the conversion performance.

On the other hand, SS did not work well, and SS rather degrade the conversion performances under some conditions in comparison with the GMM without SS. It is difficult for the SS to reduce the unstable noise such as the restaurant noise because the SS uses only the mean value of noise spectra. The output signal also tends to be distorted by applying the SS. For these reasons, SS degrades the conversion performances in some experimental conditions.

Figure 5 shows the spectrograms of noisy source sig-
Fig. 5 Spectrograms of noise signal, noisy source signal, target signal and converted signals (restaurant noise, SNR: 20 dB).

Spectrograms of noise signal, noisy source signal, target signal and converted signals (restaurant noise, SNR: 20 dB). The noise signal had high energy in low frequency bins. As shown in the red-circled areas of the spectrograms of the noisy source signal and the signal converted using GMM, the energy of noise spread over the wide range of frequency in the spectrogram of the signal converted using GMM. This might be because the noise was identified as the speech of the source speaker and converted into the target speech by GMM. This noise was suppressed in the spectrogram of the signal converted using the proposed method compared with those of noisy source signal and signal converted using GMM.

Next, for clean speech and restaurant noise (SNR: 20 dB) environments, 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.

Figures 6 and 7 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 Fig. 6, in both evaluation criteria, our proposed method showed higher scores than the conventional method although mel-cepstral distortion is almost equal to each other in the case of clean speech. One of the possible reasons would be the effect of parameterization of spectral envelope. STRAIGHT spectrum is approximated with cepstrum in the GMM-based method. On the other hand, STRAIGHT spectrum is directly used in the proposed method. As shown in Fig. 7, the preference scores biased toward our proposed method greater than those in the case of clean speech in both evaluation criteria, especially naturalness. This is because the noise was output with the converted signal, and the noise may degrade the conversion performance in the GMM-based method. Our proposed method could suppress the energy of the noise, and as a result, the proposed method could prevent from the degradation of the conversion performances.

5. Conclusions

In this paper, we proposed an exemplar-based VC technique for noisy environments. 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 between a conventional GMM-based method and the proposed method, the proposed method showed better performance in both cases using clean speech and noisy speech for evaluation, and especially in regard to the naturalness in noisy environments.

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 in the sense that it can be applied to only one-to-one voice conversations because it requires parallel speech data having the same texts uttered by the source and target speakers. Hence, we will carry out research on a method that does not use the parallel data. Future work will also include efforts to study other noise conditions, such as a sudden noise. We will compare to the conventional method with other noise reduction techniques, such as the vector Taylor series. Furthermore, we will try to apply this method to other VC applications.

Acknowledgment

This research was supported in part by MIC SCOPE.

References


Ryoichi Takashima received his B.E. and M.E. degrees in computer science from Kobe University in 2008 and 2010, respectively. His current research interests include speech signal processing and pattern recognition. He is a member of ASJ.

Tetsuya Takiguchi received the B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and the M.E. and Dr. Eng. degrees in information science from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. He is currently an Associate Professor at Kobe University. His research interests include statistic signal processing and pattern recognition. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of the IEEE, the IPSJ, and the ASJ.
Yasuo Ariki received his B.E., M.E. and Ph.D. in information science from Kyoto University in 1974, 1976 and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IPSJ, JSAI, ITE and IIEEJ.