PAPER *Special Section on Advances in Modeling for Real-world Speech Information Processing and its Application*

# Noise-Robust Voice Conversion Based on Sparse Spectral Mapping Using Non-negative Matrix Factorization

**Ryo AIHARA**[†a)], **Ryoichi TAKASHIMA**[†], ***Nonmembers***, **Tetsuya TAKIGUCHI**[††], ***and*** **Yasuo ARIKI**[††], ***Members***

**SUMMARY** This paper presents a voice conversion (VC) technique for noisy environments based on a sparse representation of speech. Sparse representation-based VC using Non-negative matrix factorization (NMF) is employed for noise-added spectral conversion between different speakers. In our previous exemplar-based VC method, source exemplars and target exemplars are extracted from parallel training data, having the same texts uttered by the source and target speakers. The input source signal is represented using the source exemplars and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. However, this exemplar-based approach needs to hold all training exemplars (frames), and it requires high computation times to obtain the weights of the source exemplars. In this paper, we propose a framework to train the basis matrices of the source and target exemplars so that they have a common weight matrix. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. The effectiveness of this method was confirmed by comparing its effectiveness (in speaker conversion experiments using noise-added speech data) with that of an exemplar-based method and a conventional Gaussian mixture model (GMM)-based method.

***key words:*** *voice conversion, sparse representation, non-negative matrix factorization, noise robustness*

## 1. Introduction

The human voice contains a variety of information, such as linguistic information, speaker individuality, emotional information, and so on. Voice conversion (VC) is a technique for converting specific information in an input speech while maintaining the other information in the utterance. One of the most popular VC applications is speaker conversion [1]. In speaker conversion, a source speaker's voice individuality is changed to a specified target speaker's so that the input utterance sounds as if it had been spoken by a specified target speaker. In recent years, VC has been used for speaker adaptation in text-to-speech (TTS) systems or for automatic speech recognition (ASR) [2].

There have also been studies on several tasks that make use of VC. Emotion conversion is a technique for changing emotional information in input speech while maintaining linguistic information and speaker individuality [3], [4]. VC is also being adopted as assistive technology that reconstructs a speaker's individuality in electrolaryngeal speech or disordered speech [5], [6]. These studies show the varied uses of VC.

Many statistical approaches to VC have been studied [1], [7], [8]. Among these approaches, the GMM-based mapping approach [1] 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. [9] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [10] proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. There have also been approaches that do not require parallel data that make use of GMM adaptation techniques [11] or eigen-voice GMM (EV-GMM) [12], [13].

However, the effectiveness of these approaches was confirmed with clean speech data, and their 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. In this paper, we propose noise-robust VC based on sparse representation.

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) [14] is a well-known approach for source separation and speech enhancement [15], [16]. 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. [17] also propose 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 [18], we discussed an exemplar-based VC technique for noisy environments. In that report, source exemplars

Manuscript received September 27, 2013.

Manuscript revised January 17, 2014.

[†] The authors are with the Graduate School of System Informatics, Kobe University, Kobe-shi, 657–8501 Japan.

[††] The authors are with Organization of Advanced Science and Technology, Kobe University, Kobe-shi, 657–8501 Japan.

a) E-mail: aihara@me.cs.scitec.kobe-u.ac.jp

DOI: 10.1587/transinf.E97.D.1411

and target exemplars are extracted from the parallel training data, having the same texts uttered by the source and target speakers. Also, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. For this reason, no training processes related to noise signals are required. The input source signal is expressed with a sparse representation of the source exemplars and noise exemplars. Only the weights related to the source exemplars are picked up, and the target signal is constructed from the target exemplars and the picked-up weights. This method showed better performances than the conventional GMM-based method in speaker conversion experiments using noise-added speech data. However, this exemplar-based approach needs to hold all training exemplars (frames), and it requires high computation times to obtain the weights of the source exemplars.

In this paper, we propose a framework to train the basis matrices of source and target exemplars so that they have a common weight matrix. The basis matrix of the source exemplars is trained using NMF, and then the weight matrix of the source exemplars is obtained. Next, the basis matrix of the target exemplars is trained using NMF, where the weight matrix is fixed to that obtained from the source exemplars. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. The effectiveness of this method was confirmed by comparing its effectiveness (in speaker conversion experiments using clean speech data and noise-added speech data) with that of an exemplar-based method and the conventional Gaussian mixture model (GMM)-based method.

The rest of this paper is organized as follows: In Sect. 2, baseline GMM-based VC is introduced. In Sect. 3, basic idea of NMF-based VC is described. In Sect. 4, our noise-robust VC is proposed. In Sect. 5, the experimental data is evaluated, and the final section is devoted to our conclusions.

## 2. Baseline Gaussian Mixture Model-Based Voice Conversion

A joint density Gaussian mixture model (JD-GMM) method is one of the most successful VC methods because of its flexibility and good performance [1]. This section describes a VC method based on JD-GMM.

### 2.1 Probability Density Function

JD-GMM VC is divided into two phases: training and conversion phases. In the training phase, a mapping function between the source and target spectrum is estimated. Let $\mathbf{x}_t$ and $\mathbf{y}_t$ be the source and target $D$-dimensional feature vectors at the $t$-th frame, respectively. A dynamic time warping (DTW) algorithm is used to align these vectors. Defining the paired feature $\mathbf{z}_t = [\mathbf{x}_t{}^T, \mathbf{y}_t{}^T]^T$, its joint probability density is set as

$$p(\mathbf{z}_t|\boldsymbol{\theta}^{(\mathbf{z})}) = \sum_{m=1}^{M} \alpha_m N(\mathbf{z}_t; \boldsymbol{\mu}_m^{(\mathbf{z})}, \boldsymbol{\Sigma}_m^{(\mathbf{z})}) \tag{1}$$

where $\boldsymbol{\theta}^{(\mathbf{z})}$ is a parameter set of the weight $\alpha_m$, source mean vector $\boldsymbol{\mu}_m^{(\mathbf{x})}$, target mean vector $\boldsymbol{\mu}_m^{(\mathbf{y})}$, and covariance matrix $\boldsymbol{\Sigma}^{(\mathbf{z})}$, and they are given by

$$\boldsymbol{\Sigma}_m^{(\mathbf{z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\mathbf{xx})} & \boldsymbol{\Sigma}_m^{(\mathbf{xy})} \\ \boldsymbol{\Sigma}_m^{(\mathbf{yx})} & \boldsymbol{\Sigma}_m^{(\mathbf{yy})} \end{bmatrix}, \quad \boldsymbol{\mu}_m^{(\mathbf{z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\mathbf{x})} \\ \boldsymbol{\mu}_m^{(\mathbf{y})} \end{bmatrix}.$$

The normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted as $N(\mu, \boldsymbol{\Sigma})$. The prior probability $\alpha_m$ of the $m$-th mixture is constrained by $\Sigma_{m=1}^{M} \alpha_m = 1$. The matrices $\boldsymbol{\Sigma}_m^{\mathbf{xx}}$ and $\boldsymbol{\Sigma}_m^{\mathbf{yy}}$ are the covariance matrix of the $m$-th mixture component for the source and that for the target, respectively. The matrices $\boldsymbol{\Sigma}_m^{\mathbf{xy}}$ and $\boldsymbol{\Sigma}_m^{\mathbf{yx}}$ are the cross covariance matrices of the $m$-th mixture component for the source and the target, respectively. In this study, these covariance matrices $\boldsymbol{\Sigma}_m^{\mathbf{xx}}$, $\boldsymbol{\Sigma}_m^{\mathbf{yy}}$, $\boldsymbol{\Sigma}_m^{\mathbf{xy}}$ and $\boldsymbol{\Sigma}_m^{\mathbf{yx}}$ are diagonal. These parameters are estimated by using the expectation-maximization (EM) algorithm.

### 2.2 Mapping Function

In the conversion phase, estimated parameters are used to implement the conversion function. Given $\mathbf{x}_t$, the likelihood function of $\mathbf{y}_t$ is given by

$$P(\mathbf{y}; \mathbf{x}_t, \boldsymbol{\theta}^{\mathbf{z}}) = \sum_{m=1}^{M} P(m; \mathbf{x}_t, \boldsymbol{\theta}^{z}) P(\mathbf{y}; \mathbf{x}_t, m, \boldsymbol{\theta}^{z}). \tag{2}$$

The $m$-th conditional probability distribution is given by

$$P(m; \mathbf{x}_t, \boldsymbol{\theta}^{(\mathbf{z})}) = \frac{\alpha_m N(\mathbf{x}_t; \boldsymbol{\mu}_m^{(\mathbf{x})}, \boldsymbol{\Sigma}_m^{(\mathbf{xx})})}{\sum_{m=1}^{M} N(\mathbf{x}_t; \boldsymbol{\mu}_m^{(\mathbf{x})}, \boldsymbol{\Sigma}_m^{(\mathbf{xx})})} \tag{3}$$

$$P(\mathbf{y}; \mathbf{x}_t, m, \boldsymbol{\theta}^{\mathbf{z}}) = N(\mathbf{y}_t; \mathbf{E}_{m,t}^{(\mathbf{y})}, \mathbf{D}_m^{(\mathbf{y})}) \tag{4}$$

$$\mathbf{E}_{m,t}^{(\mathbf{y}_t)} = \boldsymbol{\mu}_m^{(\mathbf{y})} + \boldsymbol{\Sigma}_m^{(\mathbf{yx})} (\boldsymbol{\Sigma}_m^{(\mathbf{xx})})^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_m^{(\mathbf{x})}) \tag{5}$$

$$\mathbf{D}_m^{(\mathbf{y})} = \boldsymbol{\Sigma}_m^{(\mathbf{yy})} - \boldsymbol{\Sigma}_m^{(\mathbf{yx})} (\boldsymbol{\Sigma}_m^{(\mathbf{xx})})^{-1} \boldsymbol{\Sigma}_m^{(\mathbf{xy})}. \tag{6}$$

The conversion function $F(\mathbf{x})$, which is implemented with mean square error, is used to get the target feature $\hat{\mathbf{y}}$ as follows:

$$\hat{\mathbf{y}} = F(\mathbf{x}) \tag{7}$$

$$= \int P(\mathbf{y}_t; \mathbf{x}_t, \boldsymbol{\theta}^{(\mathbf{z})}) d\mathbf{y}_t \tag{8}$$

$$= \int \sum_{m=1}^{M} P(m; \mathbf{x}_t, \boldsymbol{\theta}^{(\mathbf{z})}) P(\mathbf{y}_t; \mathbf{x}_t, \boldsymbol{\theta}^{(\mathbf{z})}) \mathbf{y}_t d\mathbf{y}_t \tag{9}$$

$$= \sum_{m=1}^{M} P(m; \mathbf{x}_t, \boldsymbol{\theta}^{(\mathbf{z})}) \mathbf{E}_{m,t}^{(\mathbf{y})}. \tag{10}$$

In each mixture component, the conditional target mean vector for the given source feature vector is calculated using a simple linear conversion as shown in (5). The converted feature vector is defined as the weighted sum of the conditional mean vectors in (10).

Although the GMM-based mapping function works well and is widely used, the effectiveness of this method has only evaluated using clean speech data. In a real environment, background noise is inevitable and it may deteriorate the performance of conversion. In order to maintain the quality of the performance of VC in a noisy environment, some noise reduction method must be added.

## 3. Voice Conversion Using Non-negative Matrix Factorization (NMF)

NMF is a well-known algorithm for noise suppression based on sparse representation. This section describes a basic VC method using NMF [18].

### 3.1 Basic Idea of NMF-Based Voice Conversion

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

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \tag{11}$$

$\mathbf{x}_l$ is the $l$-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ are the $j$-th atom and the weight, respectively. $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_J]$ and $\mathbf{h}_l = [h_{1,l} \ldots h_{J,l}]^T$ are the collection of the atoms and the stack of weights. When the weight vector $\mathbf{h}_l$ 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, the collection of atoms $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are called '*dictionary*' and '*activity*', respectively. For the frame sequence data $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_L]$, (11) is expressed as the inner product of two matrices.

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \tag{12}$$

$$\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1 \ldots \mathbf{h}_L] \tag{13}$$

$L$ is the number of the frames.

Figure 1 shows the schema of the VC method based on the sparse representation. $D$, $L$, and $J$ are the numbers of dimensions, frames and atoms, respectively. In this method, the parallel dictionaries, which consist of source and target dictionaries of the same size, are used to map the source signal to the target one. The parallel dictionaries are structured from the parallel training data, which have the same texts uttered by the source and target speakers. First, they are labeled by using forced-alignment from phoneme-HMMs recognition. Then, each labeled area is stretched so that they have same number of frames by using dynamic time wrapping (DTW). In this paper, DTW alignment allows duplicated frames.

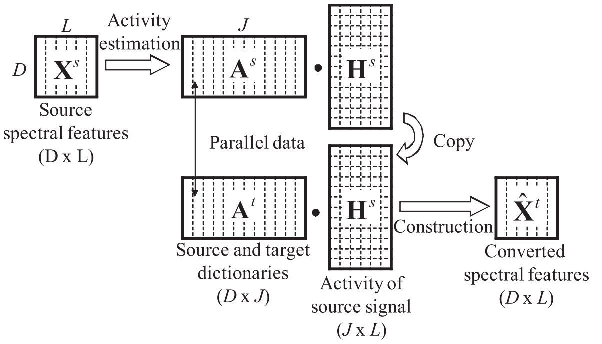
L
Activity estimation
J
D
$\mathbf{X}^s$
$\mathbf{A}^s$
$\mathbf{H}^s$
Source spectral features (D x L)
Parallel data
Copy
$\mathbf{A}^t$
$\mathbf{H}^s$
Construction
$\hat{\mathbf{X}}^t$
Source and target dictionaries ($D$ x $J$)
Activity of source signal ($J$ x $L$)
Converted spectral features ($D$ x $L$)

**Fig. 1** Voice conversion based on sparse representation.

This VC method can be combined with an NMF-based noise reduction method. Then, the noise dictionary is extracted from the before- and after-utterance sections in an observed signal, and the noise dictionary is concatenated with the source dictionary. The noisy source signal is expressed with a sparse representation of the source dictionary and noise dictionary. Only the weights related to the source dictionary are picked up, and the target signal is constructed from the target dictionary and the picked-up weights.

### 3.2 Problem

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 as shown in Fig. 2. Based on this assumption, the activity of the source signal estimated with the source dictionary can be substituted for that of the target signal.

Figure 3 shows the activity matrices estimated from the source and target words uttered ('ikioi') and their dictionaries. To show an intelligible example, each dictionary was structured from just the one word "ikioi" and aligned with DTW. The source/target features and each atom in the dictionary are a spectral envelope extracted by STRAIGHT analysis [19]. When the source/target signals 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'.

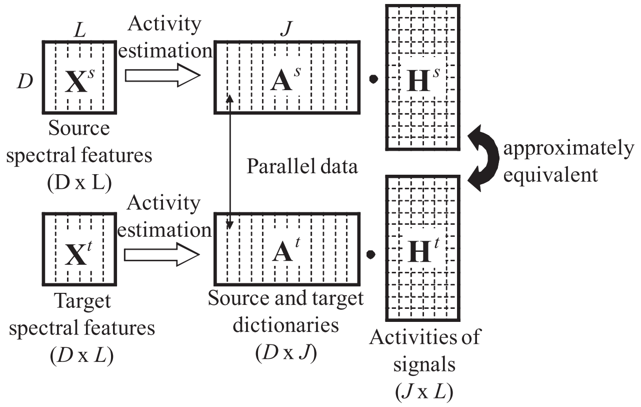

**Fig. 2** Assumption of the parallelism of source and target dictionaries.

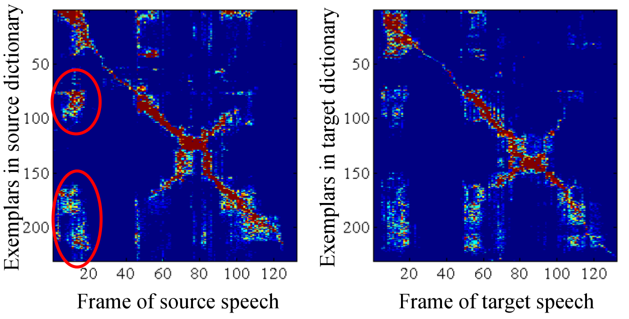

**Fig. 3** Activity matrices of the source signal (left) and target signal (right).

Therefore, as shown in Fig. 1, the input source signal is represented using the source dictionary and the activity. Then, the converted speech is constructed from the target dictionary and the activity related to the source dictionary.

However, this exemplar-based approach defines the parallel dictionary using the parallel training data. Hence, this method needs to hold all training exemplars (frames), and it requires high computation times to obtain the weights of the source exemplars. In conventional NMF-based noise reduction methods, dictionary **A** is defined with much fewer bases. In [17], 4,000 bases are chosen from real speech features and used as **A**. In [20], a small number of bases of **A** are trained using NMF. However, when the basis matrices of the source exemplars and target exemplars are trained using NMF independently, the parallelism of the source and target dictionaries shown in Fig. 2 is lost.

Therefore, in this paper, we propose a framework to train the basis matrices of source and target exemplars so that they have a common weight matrix. By using the basis matrices instead of the exemplars, VC can be performed with lower computation times than with the exemplar-based method.

## 4. Proposed Method

This section describes the proposed framework for training the basis matrices of the source and target exemplars.

### 4.1 Dictionary Construction for a Noisy Environment

In the preceding section, both dictionaries (source and target) consisted of the same spectral envelope features (STRAIGHT spectrum) for simplicity in 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 using short-time Fourier transform (STFT) is used to construct the source and noise dictionaries.

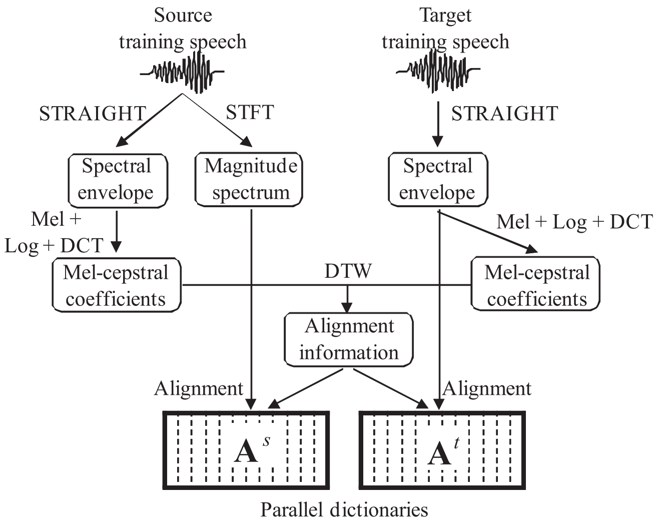

**Fig. 4** Construction of source and target dictionaries.

Figure 4 shows the process for constructing parallel dictionaries. Parallel dictionaries are constructed from clean speech data. For the target training speech, a STRAIGHT spectrum is used to extract its dictionary. Mel-cepstral coefficients are estimated from the STRAIGHT spectrum and used for DTW. For the source training speech, on the other hand, the STRAIGHT spectrum is converted into mel-cepstral coefficients and only used for DTW 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 to estimate the activity. The STRAIGHT spectrum, F0 and aperiodic components are used to synthesize the converted signal.

### 4.2 Training of the Parallel Basis Matrices

We optimize the source basis matrix $\mathbf{A}^s$ and target basis matrix $\mathbf{A}^t$ so that when the source signal and target signal are expressed with the sparse representations of $\mathbf{A}^s$ and $\mathbf{A}^t$, respectively, the obtained activity matrices are equivalent, as shown in Fig. 2.

Table 1 shows the algorithm of the training of the parallel basis matrices. At first, for the training source data (exemplars) $\mathbf{X}^s$, the basis matrix $\mathbf{A}^s$ and the activity matrix $\mathbf{H}^s$ are optimized using NMF with the sparse constraint [17]. In the framework of NMF with the sparse constraint, it minimizes the following cost function:

$$d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + \|(\lambda_{train}\mathbf{1}^{(1\times L)}) .* \mathbf{H}^s\|_1 \qquad s.t.\ \ \mathbf{A}^s,\ \mathbf{H}^s \geq 0. \tag{14}$$

Here, $.*$ and $\mathbf{1}$ are an element-wise multiplication and an all-one vector, respectively. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}^s$ and $\mathbf{A}^s\mathbf{H}^s$. The second term is the sparse constraint with the L1-norm regularization term that causes $\mathbf{H}^s$ to be sparse. $\lambda_{train}$ is the weight of the

**Table 1** Algorithm of the training of the parallel basis matrices.

| |
|---|
| **Training of source basis matrix $\mathbf{A}^s$** |
| • Set source training exemplars to $\mathbf{X}^s$ |
| • Optimize $\mathbf{A}^s$ and $\mathbf{H}^s$ by (15) and (16) |
| **Training of target basis matrix $\mathbf{A}^t$** |
| • Set target training exemplars to $\mathbf{X}^t$ |
| • Fix the activity matrix to $\mathbf{H}^s$, and optimize $\mathbf{A}^t$ by (18) |

sparse constraint. $\mathbf{A}^s$ and $\mathbf{H}^s$ minimizing (14) are estimated iteratively applying the following update rules:

$$\mathbf{A}^s_{n+1} = \mathbf{A}^s_n . * (\mathbf{H}^s_n(\mathbf{X}^s./\mathbf{A}^s_n\mathbf{H}^s_n)^{\mathsf{T}}./(\mathbf{H}^s_n\mathbf{1}^{(L\times D)}))^{\mathsf{T}} \quad (15)$$

$$\begin{aligned}\mathbf{H}^s_{n+1} &= \mathbf{H}^s_n . * (\mathbf{A}^{s^{\mathsf{T}}}_n(\mathbf{X}^s./(\mathbf{A}^s_n\mathbf{H}^s_n))) \\ &\quad ./(\mathbf{A}^{s^{\mathsf{T}}}_n\mathbf{1}^{(J\times L)} + \lambda_{train}\mathbf{1}^{(1\times L)}) \quad (16)\end{aligned}$$

where ./ is an element-wise division.

Next, using the activity matrix $\mathbf{H}^s$ obtained by (16), the target basis matrix $\mathbf{A}^t$ of the training target exemplars $\mathbf{X}^t$ is optimized. Then, $\mathbf{A}^t$ is optimized so that the activity matrix is equivalent to $\mathbf{H}^s$, i.e. $\mathbf{A}^t$ is optimized to minimize the following cost function:

$$d(\mathbf{X}^t, \mathbf{A}^t\mathbf{H}^s) \quad s.t. \quad \mathbf{A}^t \geq 0. \quad (17)$$

In this optimization, the activity matrix is fixed to $\mathbf{H}^s$, and only $\mathbf{A}^t$ is updated by the following update rule:

$$\mathbf{A}^t_{n+1} = \mathbf{A}^t_n . * (\mathbf{H}^s(\mathbf{X}^t./\mathbf{A}^t_n\mathbf{H}^s)^{\mathsf{T}}./(\mathbf{H}^s\mathbf{1}^{(L\times D)}))^{\mathsf{T}}. \quad (18)$$

$\mathbf{A}^t_n$ and $\mathbf{A}^t_{n+1}$ represent the target basis matrices of the $n$-th and the $(n+1)$-th iteration, respectively.

## 4.3 Voice Conversion of Noisy Source Signal

### 4.3.1 Estimation of Activity from Noisy Source Signal

Figure 5 shows the conversion framework of our method. The exemplars (frames) of the noise are extracted from the before- and after-utterance sections in the observed (noisy) signal, and the noise dictionary is structured from the noise exemplars for each utterance. For this reason, no training processes related to noise signals are required. In the approach based on the sparse representation, the spectrum of the noisy source signal at frame $l$ is approximately expressed by a non-negative linear combination of the clean source dictionary, noise dictionary, and their activities.

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}^s_l + \mathbf{x}^n_l \\ &\approx \sum_{j=1}^{J}\mathbf{a}^s_j h^s_{j,l} + \sum_{k=1}^{K}\mathbf{a}^n_k h^n_{k,l} \\ &= [\hat{\mathbf{A}}^s\mathbf{A}^n]\begin{bmatrix}\mathbf{h}^s_l \\ \mathbf{h}^n_l\end{bmatrix} \quad s.t. \quad \mathbf{h}^s_l, \mathbf{h}^n_l \geq 0 \\ &= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \quad (19)\end{aligned}$$

$\mathbf{x}^s_l$ and $\mathbf{x}^n_l$ are the magnitude spectra of the source signal and the noise, respectively. $\hat{\mathbf{A}}^s$, $\mathbf{A}^n$, $\mathbf{h}^s_l$ and $\mathbf{h}^n_l$ are the source dictionary (basis matrix) trained by (15), noise dictionary (exemplars), and their activities at frame $l$, respectively. Given the spectrogram, (19) can be written as follows:

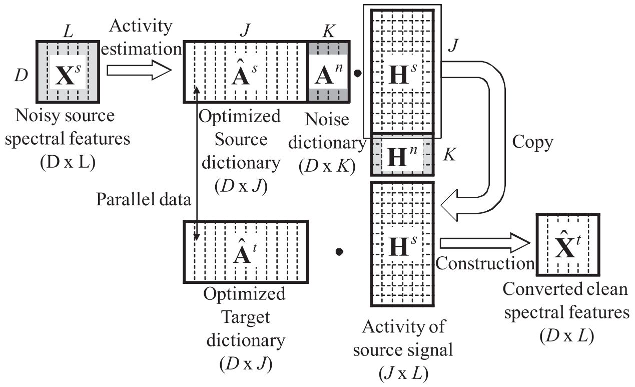
L
Activity estimation
D
$\mathbf{X}^s$
Noisy source spectral features (D x L)
J
K
$\hat{\mathbf{A}}^s$
$\mathbf{A}^n$
Optimized Source dictionary ($D$ x $J$)
Noise dictionary ($D$ x $K$)
$\mathbf{H}^s$
J
$\mathbf{H}^n$
K
Copy
Parallel data
$\hat{\mathbf{A}}^t$
$\mathbf{H}^s$
Construction
$\hat{\mathbf{X}}^t$
Optimized Target dictionary ($D$ x $J$)
Activity of source signal ($J$ x $L$)
Converted clean spectral features ($D$ x $L$)

**Fig. 5** Proposed noise-robust voice conversion.

$$\begin{aligned}\mathbf{X} &\approx [\hat{\mathbf{A}}^s\mathbf{A}^n]\begin{bmatrix}\mathbf{H}^s \\ \mathbf{H}^n\end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \quad (20)\end{aligned}$$

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

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X} \\ \bar{\mathbf{X}} &\leftarrow \mathbf{X}./\mathbf{M} \\ \bar{\mathbf{A}} &\leftarrow \mathbf{A}./(\mathbf{1}^{(D\times D)}\mathbf{A}) \quad (21)\end{aligned}$$

$\bar{\mathbf{X}}$ and $\bar{\mathbf{A}}$ are normalized $\mathbf{A}$ and $\mathbf{X}$, respectively. The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function:

$$d(\bar{\mathbf{X}}, \bar{\mathbf{A}}\mathbf{H}) + ||(\lambda_{conv}\mathbf{1}^{(1\times L)}). * \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \quad (22)$$

The weights of the sparsity constraints can be defined for each basis and exemplar by defining $\lambda_{conv}{}^{\mathsf{T}} = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$. In this paper, the weights for source bases $[\lambda_1 \ldots \lambda_J]$ were set to 0.15, and those for noise exemplars $[\lambda_{J+1} \ldots \lambda_{J+K}]$ were set to 0. $\mathbf{H}$ minimizing (22) is estimated iteratively applying the following update rule:

$$\begin{aligned}\mathbf{H}_{n+1} &= \mathbf{H}_n . * (\bar{\mathbf{A}}^{\mathsf{T}}(\bar{\mathbf{X}}./(\bar{\mathbf{A}}\mathbf{H}))) \\ &\quad ./(\mathbf{1}^{((J+K)\times L)} + \lambda_{conv}\mathbf{1}^{(1\times L)}). \quad (23)\end{aligned}$$

$\mathbf{H}_n$ and $\mathbf{H}_{n+1}$ represent the activity matrices of the $n$-th and the $(n+1)$-th iteration, respectively.

### 4.3.2 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 basis in the same way the source dictionary was.

$$\overline{\hat{\mathbf{A}}^t} \leftarrow \hat{\mathbf{A}}^t./(\mathbf{1}^{(D\times D)}\hat{\mathbf{A}}^t) \quad (24)$$

$\hat{\mathbf{A}}^t$ is the target dictionary (basis matrix) trained by (18) and

$\overline{\hat{\mathbf{A}}^t}$ is the normalized target dictionary of $\hat{\mathbf{A}}^t$. Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in (21) are applied to the normalized target spectral feature.

$$\hat{\mathbf{X}}^t = (\overline{\hat{\mathbf{A}}^t}\mathbf{H}^s). * \mathbf{M} \tag{25}$$

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 as follows:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)} \tag{26}$$

where $x_t$, $\hat{y}_t$, $\mu^{(x)}$, $\mu^{(y)}$, $\sigma^{(x)}$, and $\sigma^{(y)}$ are a log F0 of the source speaker and the converted F0 at frame $t$, mean of the source and target speaker's log F0, standard deviation of the source and target speaker's log F0, respectively. Mean and standard deviation are calculated from training data of the source and target speaker.

## 5. Experiments

### 5.1 Experimental Conditions

The proposed VC technique was evaluated by comparing it with an exemplar-based method [18] and a conventional GMM-based method [1] 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 [21], respectively. The sampling rate was 8 kHz.

A total of 216 words of clean speech were used to construct parallel dictionaries in the methods based on the sparse representation and used to train the GMM in the GMM-based method. In the exemplar-based method, the number of exemplars of the source and target dictionaries was 58,426. Then, in our proposed method, several bases were trained from the exemplars for each dictionary. Twenty-five sentences of clean speech or noisy speech were used in the evaluation. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database [22]) to the clean speech sentences. The SNR was 15 dB. The noise dictionary is extracted from the before- and after-utterance sections in the evaluation sentence. The average number of exemplars in the noise dictionary for one sentence was 110.

In the methods based on sparse representation, a 257-dimensional magnitude spectrum was used as the feature vectors for the input signal, source dictionary and noise dictionary, and a 513-dimensional STRAIGHT spectrum was used for the target dictionary. 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 64.

**Table 2** Spectral distortion improvement ratio (SDIR) [dB] for noisy speech.

| | SDIR [dB] | time [s] |
|---|---|---|
| Proposed (1,000 bases) | 5.14 | 30 |
| Proposed (2500 bases) | 4.68 | 75 |
| Proposed (5,000 bases) | 4.38 | 137 |
| Exemplar-based (58,426 exemplars) | 5.23 | 910 |
| Exemplar-based (1,000 exemplars) | 4.91 | 30 |
| GMM-based (64 mixtures) | 4.11 | 1 |

### 5.2 Experimental Results

Table 2 shows the spectral distortion improvement ratio (SDIR) [dB] and the computation time of the conversion method (1 sentence on Intel Core i7 2.80 GHz personal computer) for noisy input source signal. In our proposed method, 1,000, 2,500 and 5,000 bases were trained from the exemplars for each dictionary. In the exemplar-based method, all 58,426 exemplars and 1,000 exemplars (which are chosen randomly) are used. The SDIR is defined as follows.

$$\text{SDIR[dB]} = 10\log_{10}\frac{\sum_d |\mathbf{X}^t(d) - \mathbf{X}^s(d)|^2}{\sum_d |\mathbf{X}^t(d) - \hat{\mathbf{X}}^t(d)|^2} \tag{27}$$

Here, $\mathbf{X}^s$, $\mathbf{X}^t$ and $\hat{\mathbf{X}}^t$ are normalized so that the sum of the magnitudes over frequency bins equals unity. As shown in this table, the distortion improvement for the methods based on the sparse representation was higher than the GMM-based method regardless of the number of the trained bases. In our proposed method, the case of 1,000 bases shows the best distortion improvement. The distortion improvement of the proposed method was slightly lower than that of the exemplar-based method which uses all 58,426 exemplars. However, compared to the exemplar-based method (which uses 1,000 exemplars) our proposed method obtained higher distortion improvement. Moreover, for obtaining the activity matrix, the computation time of the proposed method (which uses 1,000 exemplars) was about 30 times faster than that of the exemplar-based method, which uses all 58,426 exemplars. The computation time is reduced as the number of the bases is reduced.

We performed a mean opinion score (MOS) test [23] on the naturalness and speaker individuality of the converted speech. In the opinion test, the opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The tests were carried out with 7 subjects. For the evaluation of naturalness, each subject listened to the converted speech and evaluated how natural the sample sounded. For the evaluation of speaker individuality, each subject listened to the target speech. Then the subject listened to the converted speech and evaluated how similar the converted speech and the target one were.

Figure 6 shows the mean opinion scores (MOS) for each method. The error bars show 95% confidence intervals. As shown in this figure, when clean speech data was used, the performances of the three methods were not so

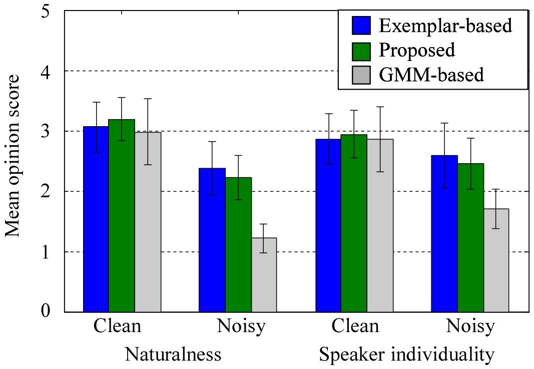

**Fig. 6** Mean opinion scores (MOS) for each method.

different under both evaluation criteria. However, when noisy speech data was used, the performances of the GMM-based method degraded considerably, especially in naturalness. This might be because the noise caused unexpected mapping in the GMM-based method, and the speech was converted with a lack of naturalness. On the other hand, the degradations of the performances of the VC methods based on the sparse representation were less than those of the GMM-based method. The performances of the proposed method were slightly lower than those of the exemplar-based method when noisy speech data was used.

## 6. Conclusions

In this paper, we discussed a noise-robust VC technique based on sparse representation. We proposed a framework to train the basis matrices of source and target exemplars so that they have a common activity matrix. The basis matrix of the source exemplars is trained using NMF. Then, the basis matrix of the target exemplars is trained using NMF, where the weight matrix is fixed to that obtained from the source exemplars. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. When a noisy input signal is converted to the target signal, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. The noisy signal is expressed with a sparse representation of the source basis matrix and noise exemplars. The target signal is constructed from the target basis matrix and the activity matrix related to the source basis matrix.

In comparison experiments between the proposed method, an exemplar-based method and a conventional GMM-based method, the proposed method showed better performances than the GMM-based method when evaluating noisy speech. The performances of the proposed method were slightly lower than those of the exemplar-based method when noisy speech data was used. But for obtaining the activity matrix, the computation time of the proposed method was about 30 times faster than that of the exemplar-based method.

However, the proposed method still requires higher computation times than that of the GMM-based method. While our proposed method took about 30 seconds to convert the speech features for 1 sentence, the GMM-based method took about 1 second to do this. In future work, we will investigate the optimal number of bases and evaluate the performances under other noise conditions.

In this paper, the source and target dictionaries are estimated separately. We can estimate the source and target dictionaries simultaneously by using the joint vector of the source and the target features just as is done in the conventional GMM-based VC. However, the performance of that method, which was evaluated experimentally, is worse than our proposed method. We will also try to improve the performance of that method.

In [24], exemplar-based VC using temporal information is proposed. We will also try to introduce dynamic information, such as segment features. 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 investigate a method that does not use parallel data. Future work will also include efforts to study other noise conditions, such as a low-SNR condition, and apply this method to other VC applications.

### References

[1] Y. Stylianou, O. Cappe, and E. Moilines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech Audio Process., vol.6, no.2, pp.131–142, 1998.

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

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

[4] 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, 2012.

[5] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," Speech Commun., vol.54, no.1, pp.134–146, 2012.

[6] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization," ICASSP, pp.8037–8040, 2013.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models," ICASSP, pp.655–658, 1988.

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

[9] 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.

[10] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," IEEE Trans. Audio Speech Lang. Process., vol.18, Issue:5, pp.912–921, 2010.

[11] C.H. Lee and C.H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," Interspeech, pp.2254–2257, 2006.

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

[13] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-

many voice conversion based on tensor representation of speaker space," Interspeech, pp.653–656, 2011.

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

[15] 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.

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

[17] 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, no.7, pp.2067–2080, 2011.

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

[19] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," Speech Commun., vol.27, no.3-4, pp.187–207, 1999.

[20] B. Schuller, F. Weninger, M. Wollmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," ICASSP, 2010.

[21] 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 Commun., vol.9, pp.357–363, 1990.

[22] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," Acoustical Science and Technology, vol.30 (2009), no.5, pp.363–371, 2009.

[23] International Telecommunication Union, "Methods for objective and subjective assessment of quality," ITU-T Recommendation P.800, 2003.

[24] Z. Wu, T. Virtanen, T. Kinnunen, E.S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," SSW8, 2013.

**Ryo Aihara** received his B.E. degrees in computer science from Kobe University in 2012. His research interest is voice conversion and statistic signal processing. He is a member of ASJ.

**Ryoichi Takashima** received his B.E., M.E., and Dr. Eng. degrees in computer science from Kobe University in 2008, 2010, and 2013 respectively. His current research interests include speech signal processing and pattern recognition. He is a member of ASJ.

**Tetsuya Takiguchi** received his B.S. degrees in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and his M.E. and Dr. of Engineering 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 acoustic and image 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.