Aihara *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:32
DOI 10.1186/s13636-015-0075-4

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization

Ryo Aihara[1*], Takao Fujii[1], Toru Nakashika[2], Tetsuya Takiguchi[1] and Yasuo Ariki[1]

Abstract

The need to have a large amount of parallel data is a large hurdle for the practical use of voice conversion (VC). This paper presents a novel framework of exemplar-based VC that only requires a small number of parallel exemplars. In our previous work, a VC technique using non-negative matrix factorization (NMF) for noisy environments was proposed. This method requires parallel exemplars (which consist of the source exemplars and target exemplars that have the same texts uttered by the source and target speakers) for dictionary construction. In the framework of conventional Gaussian mixture model (GMM)-based VC, some approaches that do not need parallel exemplars have been proposed. However, in the framework of exemplar-based VC for noisy environments, such a method has never been proposed. In this paper, an adaptation matrix in an NMF framework is introduced to adapt the source dictionary to the target dictionary. This adaptation matrix is estimated using only a small parallel speech corpus. We refer to this method as affine NMF, and the effectiveness of this method has been confirmed by comparing its effectiveness with that of a conventional NMF-based method and a GMM-based method in noisy environments.

Keywords: Voice conversion, Speech synthesis, Speaker adaptation, Noise robustness, Small parallel corpus

1 Introduction

Background noise is an unavoidable factor in speech processing. In the task of automatic speech recognition (ASR), one problem is that the recognition performance remarkably decreases under noisy environments, and this creates a significant problem in regard to the development of the practical use of ASR.

Non-negative matrix factorization (NMF) [1] is a popular approach for source separation or speech enhancement [2, 3]. In some approaches for NMF-based source separation, the exemplars, which are called "bases", are grouped for each source, and the mixed signals are expressed with a sparse representation of these bases. By using only the weights of the atoms related to the target signal, the target signal can be reconstructed. Gemmeke et al. [4] also propose an exemplar-based method for noise-robust speech recognition using NMF.

*Correspondence: aihara@me.cs.scitec.kobe-u.ac.jp
[1]Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan
Full list of author information is available at the end of the article

In our previous work, we proposed an exemplar-based method for noise-robust voice conversion (VC) using NMF. VC is a technique for converting a speaker's voice individuality while maintaining phonetic information in the utterance. In [5], we evaluated the conventional statistical VC method in a noisy environment and revealed that noise in the input signal is not only output with the converted signal but also tends to degrade the conversion performance itself due to the unexpected mapping of source features. In NMF-based VC, noise exemplars are extracted from before- and after-utterance sections and input noisy signals are decomposed into a linear combination of noise and speaker's clean exemplars. For this reason, no training processes related to noise signals are required. Only the weights related to the source exemplars are taken, and the target signal is constructed from the target exemplars and the weights. This method showed better performances than the conventional GMM-based method in speaker conversion experiments using noise-added speech data.

Moreover, we assume that our NMF-based VC creates a natural-sounding voice compared to statistical VC. In [6],

over-fitting and over-smoothing problems are reported in statistical VC. Because our NMF-based VC is not a statistical but an exemplar-based method, we assume that our approach can avoid the over-fitting problem and create a natural-sounding voice.

In spite of these efforts, VC is not used in practice. One reason for this is that conventional VC needs a large amount of parallel training data between the source and target speakers. In recent years, some statistical approaches that do not require parallel training data have been proposed [7–10]. In this paper, we propose noise-robust VC using a small parallel corpus based on an NMF-based speaker adaptation technique.

In [11], adaptation of speaker-specific bases in NMF for single-channel speech-music separation has been presented. In this framework, speaker-specific bases are adapted to the other speaker using an affine matrix. We call this method affine NMF (A-NMF) and apply it to VC. In VC, the source dictionary is constructed using sufficient source speaker data, and it is adapted using a small amount of parallel data (about ten words only) in order to obtain the target dictionary, where a linear regression transformation matrix (affine matrix) is trained based on NMF.

The contributions of this paper are summarized in two points. First, we have decreased the total amount of parallel training data required for NMF-based VC. Conventional NMF-based VC requires 216-word parallel data for dictionary construction. However, experimental results using our proposed approach, which requires only a small amount of parallel data, demonstrate a conversion quality that is almost the same as that of the conventional NMF-based VC. The second contribution is that there needs to be no concern about differences between parallel dictionaries. The parallel dictionary, as used in the conventional NMF-based VC, has a mismatched alignment, and this mismatch degrades the VC performance. In our proposed method, there is no mismatch because the target dictionary is estimated from the source dictionary using affine NMF. Details of this effect are given in Section 3.3.

The rest of this paper is organized as follows. Section 2 discusses related works, while Section 3 describes the conventional NMF-based VC method. An adaptation technique in an NMF framework is described in Section 4. Section 5 describes the results of the experiments, and the final section presents the conclusions.

2 Related works

A Gaussian mixture model (GMM)-based approach is widely used for VC because of its flexibility and good performance [12]. In this approach, the conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using a minimum mean-square error (MMSE) on a parallel training set. A number of improvements in this approach have been proposed. Toda et al. [13] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [6] proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. However, over-smoothing and over-fitting problems in these GMM-based approaches have been reported [6] because of statistical averages and a large number of parameters. These problems degrade the quality of synthesized speech.

The above statistical VC needs a large parallel corpus between the source and target speakers. In this paper, "parallel" means that the text of the corpus between the source and target speakers is the same. This constraint can be a difficult requirement to meet in practice. In GMM-based VC, there have been approaches that do not require parallel data. Lee et al. [7] used maximum a posteriori (MAP) in order to adapt training data. Mouchtaris et al. [8] proposed non-parallel training for GMM-based VC. Toda et al. [9] proposed eigen-voice GMM (EV-GMM) for many-to-many VC in which the source and target speech are represented by a super vector of the reference speakers. Saito et al. [10] proposed tensor representation for one-to-many GMM VC. However, these approaches do not work well in noisy environments because they are based on a statistical approach.

Our VC approach is exemplar-based, which is different from the conventional GMM-based VC. Exemplar-based VC using NMF has been proposed in [5]. In this framework, parallel training data is stored as source and target dictionaries. Input speech is decomposed into a linear combination of source exemplars from the source dictionary. Selected source exemplars are replaced with target exemplars, and the input speech is converted.

We assume that our approach using NMF has two advantages over conventional statistical VC. The first advantage is noise robustness and the second is the resultant natural-sounding converted voice. The noise robustness of this method was confirmed in [14]. In [15], we proposed multimodal NMF-based VC to enhance the noise robustness of our method. The natural-sounding converted voice in NMF-based VC was confirmed in [16]. Wu et al. [17] applied a spectrum compression factor to NMF-based VC and improved the conversion quality. The NMF-based VC has also been adapted for assistive technology for those with articulation disorders [18].

3 NMF-based voice conversion

3.1 Sparse representations for voice conversion

Figure 1 shows an exemplar-based voice conversion approach for a noisy environment. D, L, J, and K are the numbers of feature dimensions, frames, clean exemplars,

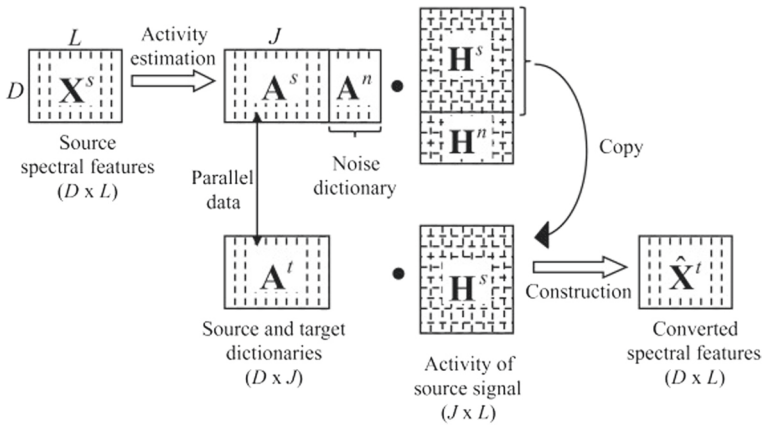

Fig. 1 Basic approach of exemplar-based voice conversion in a noisy environment

and noise exemplars, respectively. In approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of bases. We call the collection of these bases a "dictionary" and the collection of its weights "activities".

Figure 2 illustrates the process for constructing parallel dictionaries. First, we construct the parallel dictionaries of the source and target speakers. To do so, parallel spectra are extracted from the parallel words of the source and target speakers. Using dynamic time warping (DTW), these spectra are then aligned so that they have the same number of frames. Then, the source and target dictionaries are obtained by lining up the parallel spectra, which are used as parallel training data in GMM-based VC [13]. Therefore, the dictionary consists of short-time spectra obtained from all training speech data using short-time Fourier transform (STFT), where one spectrum corresponds to one basis of the dictionary. Using this method, unlike GMM-based VC, no dictionary training procedure is required.

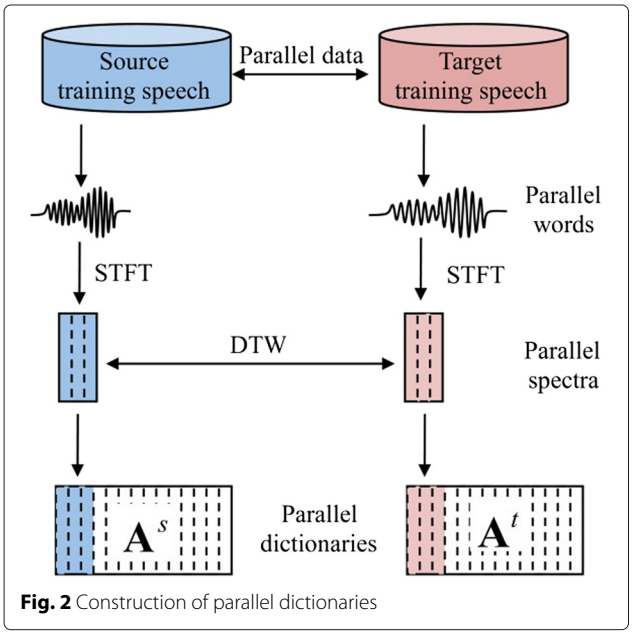

Fig. 2 Construction of parallel dictionaries

In the test stage, from the before- and after-utterance sections in the observed signal, the noise dictionary is extracted for each utterance. 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.

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

where $\mathbf{x}_l^s$ and $\mathbf{x}_l^n$ are the magnitude spectra of the source speaker and the noise, respectively. $\mathbf{A}^s$, $\mathbf{A}^n$, $\mathbf{h}_l^s$, and $\mathbf{h}_l^n$ are the source dictionary, noise dictionary, and their activities at frame l. Given the spectrogram, (1) can be written as follows:

$$\begin{aligned}\mathbf{X} &\approx [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0 \end{aligned} \tag{2}$$

where $\mathbf{H}^s$ and $\mathbf{H}^n$ denote the activity matrices of the source and noise dictionaries, respectively.

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

The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function [4]:

$$d(\mathbf{X}, \mathbf{AH}) + ||\left(\lambda \mathbf{1}^{(1\times L)}\right) .* \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{3}$$

$.*$ denotes element-wise multiplication. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}$ and $\mathbf{AH}$. The second term is the sparse constraint with a 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 $\lambda^{\mathbf{T}} = [\,\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$. In this study, the weights for source exemplars $[\,\lambda_1 \ldots \lambda_J]$ were set to 0.1, and those for noise exemplars $[\,\lambda_{J+1} \ldots \lambda_{J+K}]$ were set to 0 because the noise signal is less sparse compared to the speech signal. $\mathbf{H}$ minimizing (3) is estimated iteratively applying the following update rule:

$$\mathbf{H}_{n+1} = \mathbf{H}_n .* \left(\mathbf{A}^{\mathbf{T}}(\mathbf{X}./(\mathbf{AH}_n))\right) ./ \left(\mathbf{1}^{((J+K)\times L)} + \lambda \mathbf{1}^{(1\times L)}\right). \tag{4}$$

3.2 Target speech construction

$\mathbf{A}^t$ in Fig. 1 represents a target dictionary that consists of the target speaker's exemplars. $\mathbf{A}^s$ and $\mathbf{A}^t$ consisted of the same words and are aligned with DTW just as the conventional GMM-based VC is. 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. For this reason, we assume that when there are parallel dictionaries, the activity of the source features estimated with the source dictionary may be able to be substituted with that of the target features.

The target dictionary is also normalized for each frame in the same way the source dictionary was. 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 $\hat{\mathbf{X}}^t$ are constructed.

$$\hat{\mathbf{X}}^t = \mathbf{A}^t\mathbf{H}^s \tag{5}$$

The converted spectral features are de-normalized so that the sum of the magnitudes over frequency bins equals input spectral features.

3.3 Difference between parallel dictionaries

As mentioned in Section 3.2, this method assumes that if the parallel source and target spectra are decomposed into a parallel dictionary and its activities, the activity matrices will be approximately equivalent. In this framework, we assume that each basis in the dictionary represents a phoneme part and the activity matrix represents the phonetic information of the utterance, which is independent of the speaker.

Figure 3 shows an example of the activity matrices estimated from a single parallel Japanese word, where one is uttered by a male and the other by a female. These words are aligned by using DTW in advance, and the parallel dictionaries, which consist of 250 randomly chosen bases, are used in activity estimation. As shown in the figure, estimated activities are different although input features and dictionaries are parallel. We assume that there are two reasons for this. First, we assume that the alignment difference between the source and the target dictionaries causes this effect. Although the parallel dictionaries are aligned by DTW, there still seems to be a mismatch of alignment. These mismatch degrades the performance of the exemplar-based VC [16]. Second, we assume that the activity matrix contains not only phonetic information but also speaker information. In [19], we proposed a framework for solving this effect and improved the performance of the NMF-based VC; however, a large amount of parallel data is still needed when using this framework.

4 Exemplar-based voice conversion using a small-parallel corpus

In the framework of the conventional NMF-based VC which is described in Section 3, a large-parallel corpus of source and target speakers is needed for dictionary construction. In this section, we propose target dictionary estimation from a small-parallel corpus only.

Figure 4 shows the estimation procedure of our proposed method. $\mathbf{X}^s$ and $\mathbf{X}^t$ show a small amount of parallel data between the source and target speakers. In the activity estimation stage, a source spectral exemplar matrix $\mathbf{X}^s$ is decomposed into a linear combination of bases from the source dictionary $\mathbf{A}^s$. The source dictionary consists of the source speaker's exemplars. It is constructed the same way the dictionary is constructed when using the conventional NMF-based VC, as explained in Section 3. The indexes and weights of the bases are estimated using (4) as source activity $\mathbf{H}^s$.

In the dictionary adaptation stage, speaker adaptation is conducted in order to obtain a target dictionary from a source dictionary using a small amount of (parallel) target speech data. The adaptation is performed using a linear regression transformation matrix based on an NMF framework. Given the transformation matrix, $\mathbf{W}$, the target feature vector at the l-th frame is obtained as follows:

$$\mathbf{x}_l^t \approx \mathbf{W}\mathbf{A}^s\mathbf{h}_l^s \tag{6}$$

where $\mathbf{A}^s$ is the source dictionary and $\mathbf{h}_l^s$ is the activity vector of the source signal at the l-th frame.

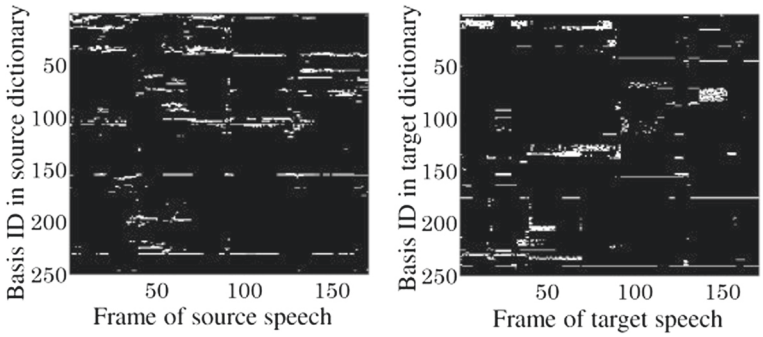

Fig. 3 Activity matrices for parallel utterances

In order to find the transformation matrix, an NMF framework which minimizes the KL divergence between $\mathbf{X}^t$ and $\mathbf{W}\mathbf{A}^s\mathbf{H}^s$ is used.

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}}\, d\left(\mathbf{X}^t, \mathbf{W}\mathbf{A}^s\mathbf{H}^s\right) \quad s.t. \quad \mathbf{W} \geq 0 \tag{7}$$

The transformation matrix, $\mathbf{W}$, is estimated using $\mathbf{A}^s$, $\mathbf{H}^s$, and a small amount of the parallel target speech data, $\mathbf{X}^t$, as follows:

$$\mathbf{W}_{n+1} = \mathbf{W}_n . * \left(\left(\mathbf{X}^t ./ \left(\mathbf{W}_n \left(\mathbf{A}^s\mathbf{H}^s\right)\right)\right)\left(\mathbf{A}^s\mathbf{H}^s\right)^{\mathbf{T}}\right) ./ \left(\mathbf{1}^{(D\times L)}\left(\mathbf{A}^s\mathbf{H}^s\right)^{\mathbf{T}}\right). \tag{8}$$

The new parallel target dictionary is given by $\hat{\mathbf{A}}^t = \mathbf{W}\mathbf{A}^s$.

In the test stage, the noisy input source speaker's spectra matrix $\mathbf{X}$ is decomposed into the multiplication of dictionary $\mathbf{A} = [\mathbf{A}^s\mathbf{A}^n]$ by its activity $\mathbf{H} = \left[\mathbf{H}^{s\mathbf{T}}\mathbf{H}^{n\mathbf{T}}\right]^{\mathbf{T}}$ as follows:

$$\mathbf{X} = \mathbf{A}\mathbf{H}. \tag{9}$$

The converted spectra matrix $\hat{\mathbf{X}}^t$ is constructed from the estimated target dictionary $\hat{\mathbf{A}}^t$, and the clean activity $\mathbf{H}^s$ as follows:

$$\hat{\mathbf{X}}^t = \hat{\mathbf{A}}^t\mathbf{H}^s. \tag{10}$$

In this method, we do not have to consider the difference between the parallel dictionaries in Section 3.3 because the parallel utterances are used as adaptation data, not as a dictionary. The activity matrix estimated

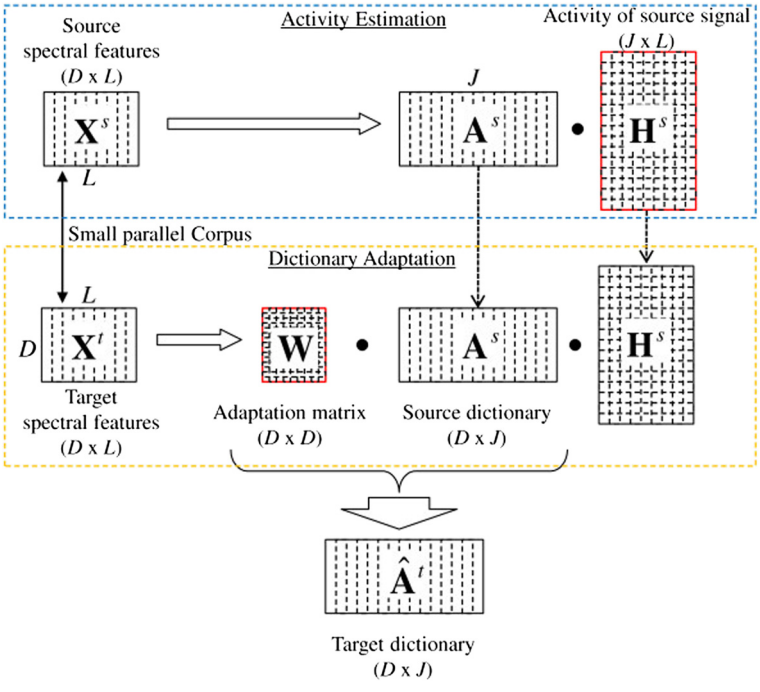

Fig. 4 Estimation of parallel dictionary using a speaker transformation matrix

from the source dictionary contains both the phoneme information and speaker information of the input utterance, as explained in Section 3.3. In this method, the adaptation matrix is estimated from the fixed source dictionary and source activity matrix, and target speaker information is extracted using the adaptation matrix in this procedure. In other words, the adaptation matrix is independent of the phoneme, and it is the conversion matrix from the source to the target speaker.

5 Experiments

5.1 Experimental conditions

The new VC technique was evaluated by comparing it with conventional techniques based on GMM [12] and NMF [5] in a speaker conversion task using noisy speech data. Speaker MMY, MAU, MNM, FTK, FYN, and FMS were selected from the ATR Japanese speech database [20], and we conducted male-to-female (MMY→FTK and MAU→FYN), male-to-male (MMY→MAU and MNM→MMY), and female-to-female (FTK→FYN and FMS→FTK) conversions. The sampling rate, frame shift, and window length are 8 kHz, 5 ms, and 25 ms, respectively.

We used 216 words of clean speech per each speaker to construct a source dictionary in NMF with a speaker adaptation and to train the GMM using the conventional method. Table 1 shows the average number of frames in each parallel dictionary. These training data were taken from the ATR Japanese speech database set A (all of task code B). The number of adaptation words was 10, 25, and 50 per each speaker. These adaptation words were randomly chosen from the ATR Japanese speech database set A (task code A). Fifty words, which are included in the ATR Japanese speech database set A (task code A) and are different from the training and adaptation data, were randomly chosen as test data.

The noisy speech signals were obtained by adding noise signals to clean speech data. We used three types of noise signal (restaurant, station, and exhibition), and these are randomly taken from the non-utterance section of CENSREC-1-C database [21]. The SNRs for each noise was set to 20 and 10 dB. They are added to a test word independently to each other. (A noisy speech signal includes one type of noise signal.) The average number of exemplars in the noise dictionary for each word was 104.

In the objective evaluation, all 50 test words with three types of noisy signal at two different types of SNRs were converted. Therefore, a of total 1800 words (6 pairs × 50 words × 3 noise types × 2 SNRs) were used for subjective evaluation. In subjective evaluation, the half of the test data with restaurant noise at 10 dB were used. Therefore, the total amount of test words was 150 (6 pairs × 25 words) in subjective evaluation.

Table 1 Number of frames in each parallel dictionary

No. of training data	216	50	25	10
No. of frames	61,168	13,839	6782	2826

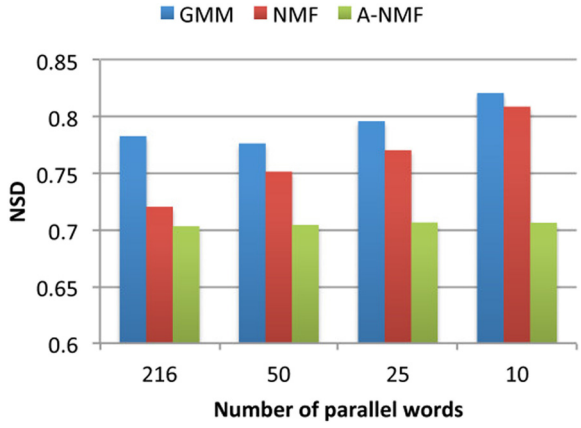

Fig. 5 NSD for male-to-female conversion at 10 dB

The spectrum, F0, and aperiodic components were extracted using STRAIGHT [22]. In the NMF-based method, a 513-dimensional spectrum extracted using STRAIGHT was used as the feature vector in the input signal and source dictionary. The number of iterations used to estimate the activity was 300 [16]. The activity and the transformation matrix were initialized with non-negative random values. In the GMM-based method, 40 linear-cepstral coefficients obtained from the STRAIGHT [22] spectrum were used as the feature vectors. The number of Gaussian mixtures was 64 which was chosen to obtain minimum distortion on test data. In this study, F0 information was converted using conventional linear regression in all VC methods based on the mean and standard deviation [13] as follows:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}\left(x_t - \mu^{(x)}\right) + \mu^{(y)}, \tag{11}$$

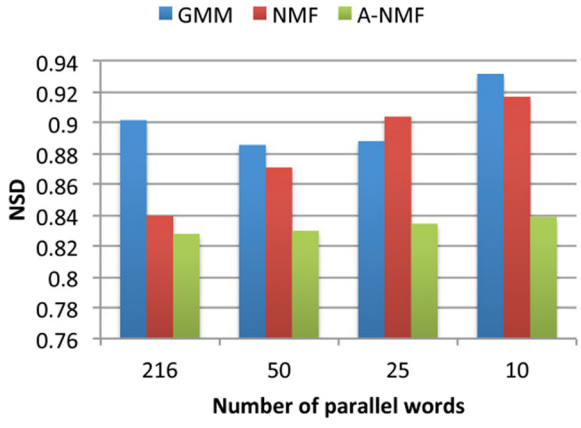

Fig. 6 NSD for male-to-male conversion at 10 dB

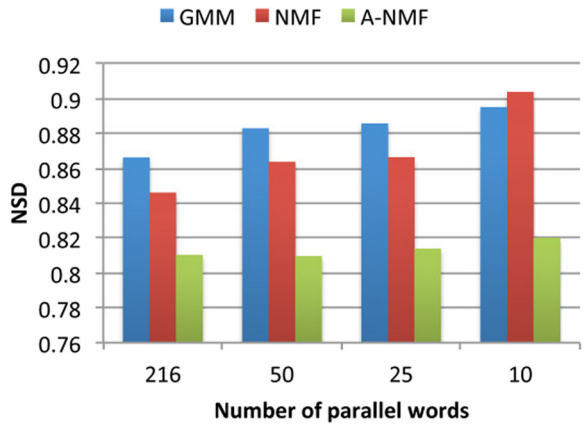

Fig. 7 NSD for female-to-female conversion at 10 dB

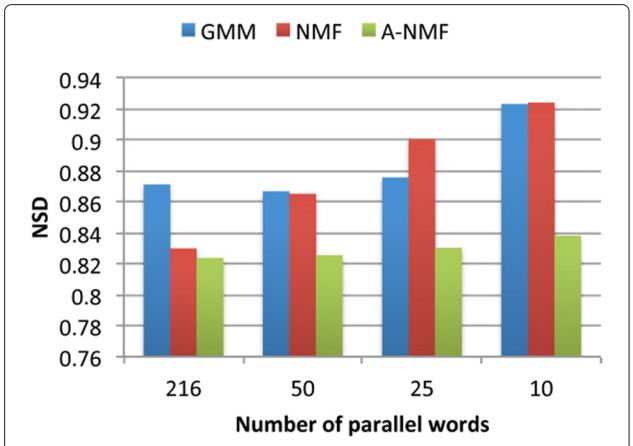

Fig. 9 NSD for male-to-male conversion at 20 dB

where $\mathbf{x}_t$ and $\hat{y}_t$ denote the log-scaled F0 of the source speaker and the converted word at frame t, respectively. $\mu^{(x)}$ and $\sigma^{(x)}$ denote the mean and standard deviation of the log-scaled F0, as calculated from the source speaker's training data. $\mu^{(y)}$ and $\sigma^{(y)}$ are the mean and standard deviation of the target speaker data. We made no conversions to the aperiodic components. With STRAIGHT, we used converted spectra, F0, and source aperiodic components for synthesizing the target voice.

5.2 Experimental results

Objective tests were carried out using the normalized spectrum distortion (NSD) [23]

$$NSD = \sqrt{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2 / ||\mathbf{X}^t - \mathbf{X}^s||^2}, \tag{12}$$

where $\mathbf{X}^s$, $\mathbf{X}^t$, and $\hat{\mathbf{X}}^t$ denote the source, target, and converted spectrum, respectively.

Figures 5, 6, and 7 show the NSD for each speaker at 10 dB. "NMF" shows the result using conventional NMF without speaker adaptation, and "A-NMF" shows the result using NMF with speaker adaptation. As shown in these figures, the performance of NMF without speaker adaptation decreases as the number of words used for the parallel dictionaries decreases. On the other hand, the performance of NMF with speaker adaptation does not decrease in comparison to the conventional NMF without speaker adaptation. Our A-NMF method obtained a better result than NMF when we used 216 parallel words for most speakers. We assume this to be due to the fact that the difference between parallel dictionaries degrades the performance of NMF.

Figures 8, 9, and 10 show the NSD for each speaker at 20 dB. Because of the low SNR conditions, the noise robustness of the NMF-based VC is lower, compared to Figs. 5, 6, and 7. However, the performance of NMF with speaker adaptation does not decrease as the number of words used for the parallel dictionaries decreases in comparison with the conventional NMF without speaker adaptation. Moreover, the performance of NMF with

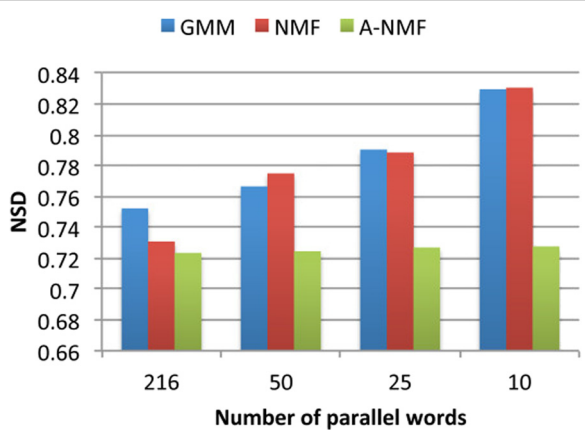

Fig. 8 NSD for male-to-female conversion at 20 dB

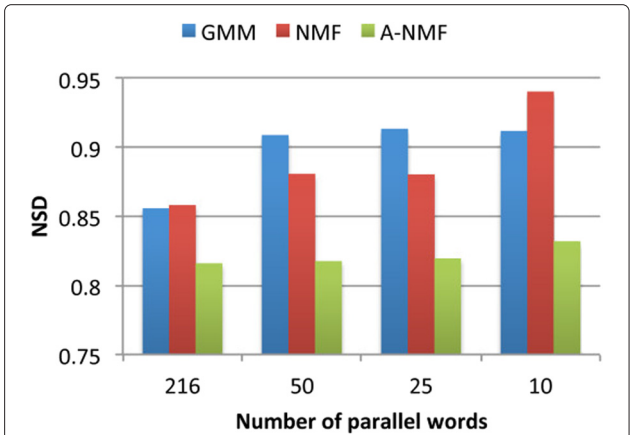

Fig. 10 NSD for female-to-female conversion at 20 dB

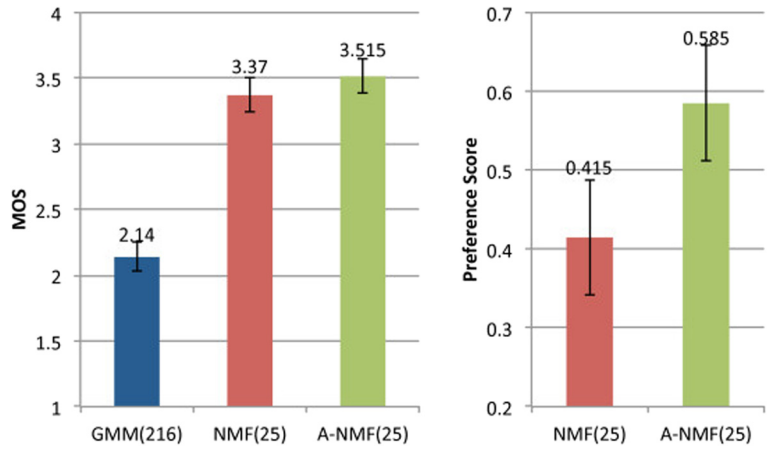

Fig. 11 Results of MOS test and XAB test

speaker adaptation is better than the conventional GMM-based VC. These results show the effectiveness of our NMF-based speaker adaptation technique.

For the speech quality evaluation, a mean opinion score (MOS) [24] test was performed. The opinion score was set to a five-point scale (5 excellent, 4 good, 3 fair, 2 poor, 1 bad). The number of participants was 8, and the SNR was 10 dB. Figure 11 shows the MOS test on the speech quality. As shown in Fig. 11, the NMF-based VC with speaker adaptation (25 adaptation words) obtained a better score than the conventional NMF-based VC (25 words). The result was confirmed by a p value test of 0.05.

For the evaluation of speaker individuality, an XAB test was carried out. In the XAB test, each participant listened to the target speech. The participant then listened to the speech converted by the two methods and selected the sample that sounded more similar to the target speech. Figure 11 shows that the NMF-based VC with speaker adaptation obtained a higher score than the conventional NMF-based VC without speaker adaptation. We confirmed this result by a 0.05 p value test.

6 Conclusions

In this paper, an exemplar-based VC technique using speaker adaptation was presented. This method requires only a small amount of parallel data, where a linear regression transformation matrix is used to adapt a source dictionary to a target dictionary and it is estimated in an NMF framework. In comparison experiments between GMM-based VC, NMF without speaker adaptation, and NMF with speaker adaptation, the NMF-based VC with speaker adaptation showed better performance.

Some problems remain with this method. The proposed method requires higher computation times than the GMM-based method. In [25], we proposed a framework that reduces computational time for NMF-based VC. In future work, we will investigate the optimal number of bases and evaluate performance under other noise conditions. In addition, this method is limited to only one-to-one voice conversion because it requires a small amount of parallel data. Hence, we will research a method for many-to-many VC within this framework and apply this method to other VC applications, such as assistive technology [18] or emotional VC [26].

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan. [2]Graduate School of Information Systems, University of Electro-Communications, 1-5-1, Chofugaoka, Chofu, Tokyo, Japan.

Received: 24 February 2015 Accepted: 12 October 2015
Published online: 25 November 2015

References

1. DD Lee, HS Seung, in *Proc. Neural. Inf. Process. Syst*. Algorithms for non-negative matrix factorization, vol. 13, (2001), pp. 556–562
2. T Virtanen, Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech and Lang. Process. **15**(3), 1066–1074 (2007)
3. MN Schmidt, RK Olsson, in *Proc. INTERSPEECH*. Single-channel speech separation using sparse non-negative matrix factorization, (2006), pp. 2614–2617
4. JF Gemmeke, T Viratnen, A Hurmalainen, Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio, Speech and Lang. Process. **19**(7), 2067–2080 (2011)
5. R Takashima, T Takiguchi, Y Ariki, in *Proc. SLT*. Exemplar-based voice conversion in noisy environment, (2012), pp. 313–317
6. E Helander, T Virtanen, J Nurminen, M Gabbouj, Voice conversion using partial least squares regression. IEEE Trans. On Audio, Speech, Lang. Process. **18**(5), 912–921 (2010)
7. CH Lee, CH Wu, in *Proc. INTERSPEECH*. MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training, (2006), pp. 2254–2257
8. A Mouchtaris, JV der Spiegel, P Mueller, Nonparallel training for voice conversion based on a parameter adaptation approach. IEEE Trans. Audio, Speech, and Lang. Processing. **14**(3), 952–963 (2006)
9. T Toda, Y Ohtani, K Shikano, in *Proc. INTERSPEECH*. Eigenvoice conversion based on Gaussian mixture model, (2006), pp. 2446–2449
10. D Saito, K Yamamoto, N Minematsu, K Hirose, in *Proc. INTERSPEECH*. One-to-many voice conversion based on tensor representation of speaker space, (2011), pp. 653–656

11. EM Grais, H Erdogan, in *Proc. INTERSPEECH*. Adaptation of speaker-specic bases in non-negative matrix factorization for single channel speech-music separation, (2011), pp. 569–572
12. Y Stylianou, O Cappe, E Moilines, Continuous probabilistic transform for voice conversion. IEEE. Trans. Speech and Audio Processing. **6**(2), 131–142 (1998)
13. T Toda, A Black, K Tokuda, Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech and Lang. Process. **15**(8), 2222–2235 (2007)
14. R Takashima, T Takiguchi, Y Ariki, Exemplar-based voice conversion using sparse representation in noisy environments. IEICE Trans. Fundam. Electron. Commun. Comp. Sci. **E96-A**(10), 1946–1953 (2013)
15. K Masaka, R Aihara, T Takiguchi, Y Ariki, in *Proc. INTERSPEECH*. Multimodal exemplar-based voice conversion using lip features in noisy environments, (2014), pp. 1159-1163
16. R Aihara, T Nakashika, T Takiguchi, Y Ariki, in *Proc. ICASSP*. Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary, (2014), pp. 7894–7898
17. Z Wu, T Virtanen, ES Chng, H Li, Exemplar-based sparse representation with residual compensation for voice conversion. IEEE Trans. Audio, Speech and Lang. Process. **22**(10), 1506–1521 (2014)
18. R Aihara, R Takashima, T Takiguchi, Y Ariki, A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary. EURASIP J. Audio, Speech, and Music Process. **2014**(5) (2014). doi:10.1186/1687-4722-2014-5
19. R Aihara, T Takiguchi, Y Ariki, in *Proc. ICASSP*. Activity-mapping non-negative matrix factorization for exemplar-based voice conversion, (2015), pp. 4899–4903
20. A Kurematsu, K Takeda, Y Sagisaka, S Katagiri, H Kuwabara, K Shikano, ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication. **9**, 357–363 (1990)
21. 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, S Nakamura, CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments. Acoustical Science and Technology. **30**(5), 363–371 (2009)
22. H Kawahara, H Matsui, in *Proc. ICASSP*. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation, vol. I, (2003), pp. 256–259
23. T En-Najjary, O Roec, T Chonavel, in *Proc. ICSLP*. A voice conversion method based on joint pitch and spectral envelope transformation, (2004), pp. 199–203
24. INTERNATIONAL TELECOMMUNICATION UNION, Methods for objective and subjective assessment of quality. ITU-T Recommendation, 800 (2003)
25. R Aihara, R Takashima, T Takiguchi, Y Ariki, Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization. IEICE Trans. Inf. Syst. **E97-D**(6), 1411–1418 (2014)
26. C Veaux, X Robet, in *Proc. INTERSPEECH*. Intonation conversion from neutral to expressive speech, (2011), pp. 2765–2768