Masaka *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:24
DOI 10.1186/s13636-015-0067-4

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

RESEARCH Open Access

Multimodal voice conversion based on non-negative matrix factorization

Kenta Masaka[1*], Ryo Aihara[1], Tetsuya Takiguchi[2] and Yasuo Ariki[2]

Abstract

A multimodal voice conversion (VC) method for noisy environments is proposed. In our previous non-negative matrix factorization (NMF)-based VC method, source and target exemplars are extracted from parallel training data, in which the same texts are uttered by the source and target speakers. The input source signal is then decomposed into source exemplars, noise exemplars, and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. In this study, we propose multimodal VC that improves the noise robustness of our NMF-based VC method. Furthermore, we introduce the combination weight between audio and visual features and formulate a new cost function to estimate audio-visual exemplars. Using the joint audio-visual features as source features, VC performance is improved compared with that of a previous audio-input exemplar-based VC method. The effectiveness of the proposed method is confirmed by comparing its effectiveness with that of a conventional audio-input NMF-based method and a Gaussian mixture model-based method.

Keywords: Voice conversion; Multimodal; Image features; Non-negative matrix factorization; Noise robustness

1 Introduction

Background noise is an unavoidable factor in speech processing. In automatic speech recognition (ASR) tasks, one problem is that recognition performance decreases significantly in noisy environments, which impedes the development of practical ASR applications.

The same problem occurs in VC, which modifies non-linguistic information such as voice characteristics, while maintaining linguistic information in its original state. The noise in the input signal is output with the converted signal and degrades conversion performance because of unexpected mapping of source features. To address this problem, we propose a noise-robust VC method that is based on sparse representations.

In recent years, sparse representation-based approaches have gained interest in a broad range of signal processing techniques. NMF [1], which is based on the idea of sparse representations, is a well-known approach for source separation and speech enhancement [2, 3]. In such 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 source separation approaches, atoms are grouped for each source, and the mixed signals are expressed with a sparse representation of these atoms. The target signal can then be reconstructed using only the weights of the atoms related to the target signal. Gemmeke et al. [4] proposed an exemplar-based method for noise-robust speech recognition using NMF. In their method, the observed speech is decomposed into 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.

Previously, we have discussed a noise-robust VC technique using NMF [5]. In that method, source and target exemplars are extracted from parallel training data, in which the same texts are uttered by the source and target speakers. In addition, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. Consequently, no training processes related to noise signals are required. The input source signal is expressed with a sparse representation of the source exemplars and (non-sparse) noise exemplars. Only the weights related to the source exemplars are picked up, and the target signal is constructed from the target

*Correspondence: makka@me.cs.scitec.kobe-u.ac.jp
[1] Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada-ku, 657-8501 Kobe, Japan
Full list of author information is available at the end of the article

exemplars and the picked-up weights. This method has demonstrated better performance than the conventional Gaussian mixture model (GMM)-based method [6] in VC experiments using noise-added speech data. However, the performance of the method was not sufficient for practical use.

Audio-visual speech recognition which uses dynamic visual lip and audio information has been studied as a technique for robust speech recognition in noisy environments. In audio-visual speech recognition, there are three priminary integration methods: early integration [7], which connects the audio feature vector with the visual feature vector; late integration [8], which weighs the likelihood of the result obtained by a separate process for audio and visual signals, and synthetic integration [9], which calculates the product of the output probability in each state. A discrete cosine transform (DCT) is widely used as a visual feature in audio-speech recognition. Previously, we have proposed audio-visual speech recognition using a visual feature extracted from an active appearance model [10, 11]. The feature contains shape information that expresses lip movement and texture information that expresses intensity changes, such as tooth.

In this study, we propose a multimodal VC technique using NMF with a combination weight between audio and visual features. The visual information is extracted from videos that capture the lip movement of utterances. The extracted visual features are connected to the audio features and used as source exemplars. The input noisy audio-visual feature is represented by a linear combination of source and noise exemplars. Then, the source exemplars are replaced with related parallel target exemplars extracted from clean audio features. The effectiveness of the proposed method has been confirmed by comparing it with that of the conventional audio-input NMF-based method and the conventional GMM-based method.

The remainder of this paper is organized as follows. In Section 2, related works are introduced. In Section 3, the proposed method is described. Experimental data are evaluated in Section 4 and conclusions are presented in Section 5.

2 Related works

VC is a technique for converting specific information to speech while maintaining other information within the utterance. One of the most popular VC applications is speaker conversion [6] where a source speaker's voice individuality is changed to that of a specified target speaker such that the input utterance sounds as though the specified target speaker has spoken the utterance.

Other studies have examined several tasks that use VC. Emotion conversion is a technique that changes emotional information in input speech while maintaining linguistic information and speaker individuality [12, 13]. VC has also been adopted as assistive technology that reconstructs a speaker's individuality in electrolaryngeal speech [14], disordered speech [15] or speech recorded by non-audible murmur microphones [16]. Recently, VC has been used for ASR and speaker adaptation in text-to-speech (TTS) systems [17].

Statistical approaches to VC are the most widely studied [6, 18, 19]. Among these approaches, a GMM-based mapping approach [6] is the most common. In this approach, the conversion function is interpreted as the expectation value of the target spectral envelope, and the conversion parameters are evaluated using minimum mean-square error (MMSE) on a parallel training set. A number of improvements to this approach have been proposed. Toda et al. [20] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [21] proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. In addition, other approaches that make use of GMM adaptation techniques [22] or eigenvoice GMM (EV-GMM) [23, 24] do not require parallel data.

Our VC approach is exemplar-based, which differs from conventional GMM-based VC. Exemplar-based VC using NMF has been proposed previously [5]. We assume that our NMF approach is advantageous in that it results in a more natural-sounding converted voice compared to conventional statistical VC. The natural sounding converted voice in NMF-based VC has been confirmed [25]. Wu et al. [26] applied a spectrum compression factor to NMF-based VC to improve conversion quality. However, the effectiveness of these approaches was confirmed using clean speech data; thus, their utilization in noisy environments was not considered. Noise in the input signal may degrade conversion performance due to unexpected mapping of source features.

The contributions of this paper are summarized as follows. First, we propose multimodal exemplar-based VC for noisy environments. The effectiveness of conventional VC approaches has been confirmed with clean speech data; however, their utilization in noisy environments was not considered. Therefore, noise-robust VC is required for real environments because noise in the input signal may degrade conversion performance due to unexpected mapping of source features. The main framework for exemplar-based multimodal VC has been proposed previously [27, 28]. In this paper, we evaluate our multimodal VC using continuous digital utterances which have been used in most studies related to audio-visual signal processing. Second, we have conducted detailed objective evaluation and confirmed the effectiveness of

visual data in VC. Note that we analyzed the performance of the proposed method for vowel and consonant parts separately. The evaluation revealed that visual input data improved the conversion quality of the consonant part. Third, we conducted subjective evaluations and confirmed that the proposed multimodal VC reduces noise effectively compared to conventional VC.

3 Multimodal voice conversion

3.1 Basic approach

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

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

Here, $\mathbf{x}_l$ represents the l-th frame of the observation, and $\mathbf{w}_j$ and $h_{j,l}$ represent the j-th basis and the weight, respectively. $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ represent the collection of bases and collection of weights, respectively. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination of a small number of bases with non-zero weights. In this paper, each basis represents the exemplar of the speech or noise signal, and the collection of exemplar $\mathbf{W}$ and the weight vector $\mathbf{h}_l$ are referred to as "dictionary" and "activity", respectively.

Figure 1 shows the basic approach of the proposed exemplar-based VC using NMF. Here D, d, L, and J represent the number of dimensions of source features,the dimensions of target features, the frames of the dictionary, and the basis of the dictionary, respectively.

The proposed VC method requires two phonemically parallel dictionaries, where one dictionary (i.e. the source dictionary) is constructed from source features and the other dictionary (i.e., the target dictionary) is constructed from target features. These dictionaries consist of the same words and are aligned with dynamic time warping (DTW); thus, they have the same number of bases.

An input source feature matrix $\mathbf{X}^s$ is decomposed into a linear combination of bases from the source dictionary $\mathbf{W}^s$ using NMF. The weights of the bases are estimated as an activity $\mathbf{H}^s$. Therefore, the activity includes the weight information of input features for each basis. The activity is then multiplied by a target dictionary to obtain the converted spectral feature matrix $\hat{\mathbf{X}}^t$, which is represented by a linear combination of bases from the target dictionary. The source and target dictionaries are parallel phonemically; therefore, the bases used in the converted features are phonemically the same as those of the source features.

3.2 Multimodal dictionary construction

Figure 2 shows the process for constructing a parallel dictionary. To construct a parallel dictionary, some pairs of parallel utterances are required, with each pair consisting of the same text. The source dictionary $\mathbf{W}^s$ consists of jointed audio-visual features, while the target dictionary $\mathbf{W}^t$ consists of only audio features.

For audio features, a simple magnitude spectrum calculated by short-time Fourier transform (STFT) is extracted from clean parallel utterances. Mel-frequency cepstral coefficients (MFCCs) are calculated from the STRAIGHT spectrum to obtain alignment information in DTW.

For visual features, DCT of lip motion images of the source speaker's utterance is used. We have adopted 2D-DCT for lip images and perform a zigzag scan to obtain the 1D-DCT coefficient vector. Note that DCT

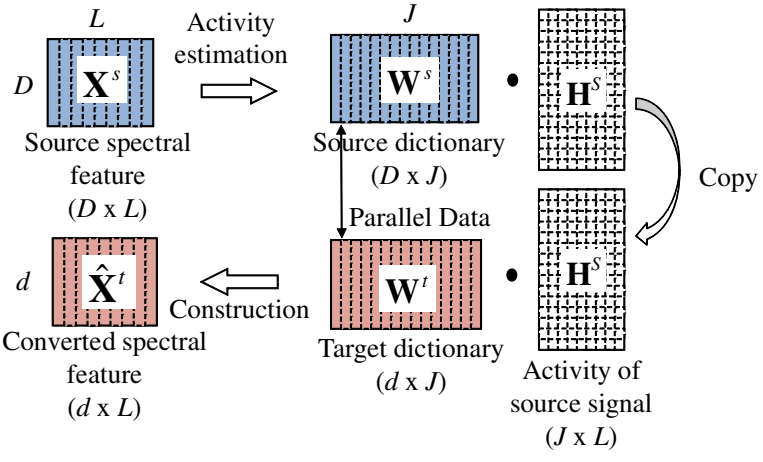

Fig. 1 Basic approach of NMF-based voice conversion

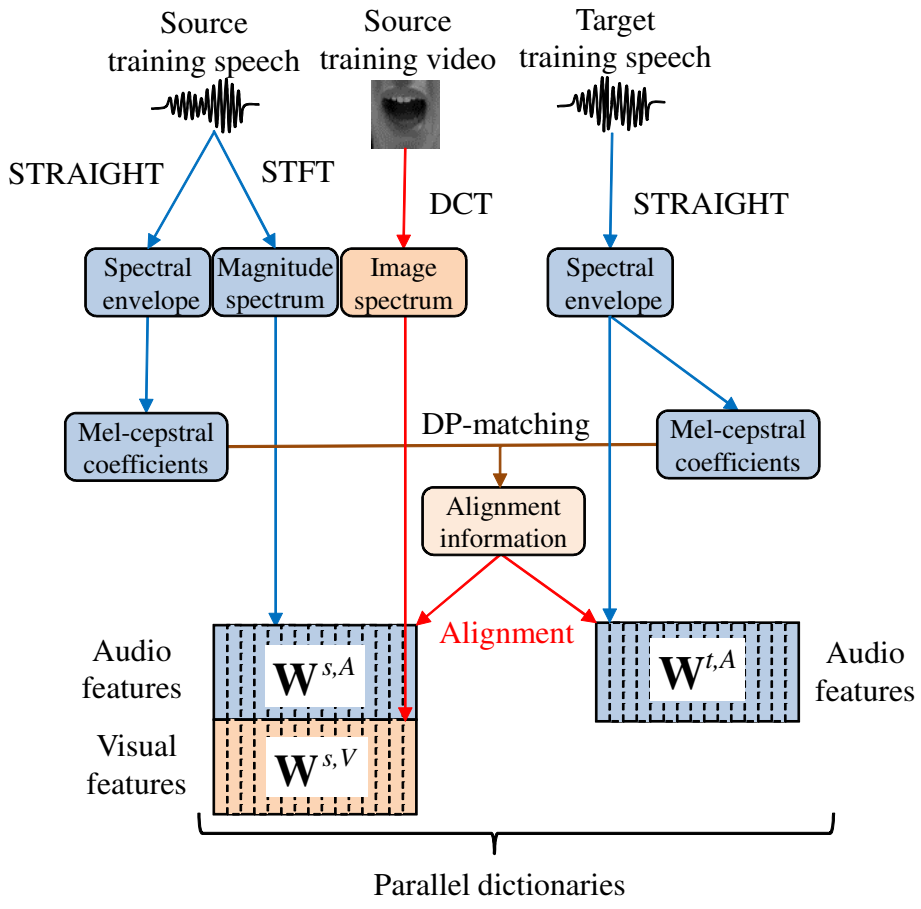

Fig. 2 Multimodal dictionary construction

coefficients contain negative data; therefore, we added constant value to satisfy the non-negativity constraint of NMF such that the scale of the frame data is not changed. The visual features extracted from images taken by commonly-used cameras are interpolated by spline interpolation to fill the sampling rate gap between audio features. Aligned audio and visual features of the source speaker are joined and used as a source feature. The source and target dictionaries are constructed by lining up each of the features extracted from parallel utterances.

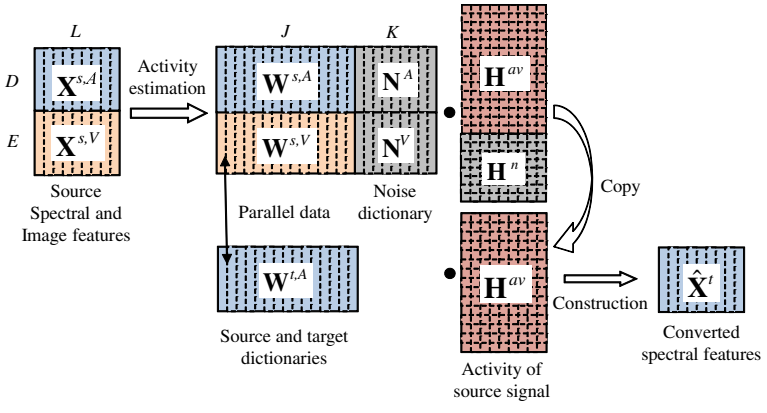

Fig. 3 Flow chart of multimodal conversion

Table 1 Experimental data

Number of digits	Total number of utterances	Dictionary and test data
Total	50	{40, 10}
1	10	{10, 0}
2	9	{6, 3}
3	7	{6, 1}
4	10	{8, 2}
5	8	{6, 2}
7	6	{4, 2}

The audio feature of the noise dictionary is extracted from the before- and after-utterance sections in the input-noisy audio signal. Note that the visual feature of the noise dictionary is extracted in the same way.

3.3 Estimation of activity from noisy source signals using NMF with a combination weight

In the exemplar-based approach, the spectrum of the noisy source signal at a frame is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their activities.

$$\begin{aligned}\mathbf{x} &= \mathbf{x}^s + \mathbf{x}^n \\ &\approx \sum_{j=1}^{J} \mathbf{w}_j^s h_j^{av} + \sum_{k=1}^{K} \mathbf{w}_k^n h_k^n \\ &= [\,\mathbf{W}^s\ \mathbf{N}] \begin{bmatrix} \mathbf{h}^{av} \\ \mathbf{h}^n \end{bmatrix} \quad s.t. \quad \mathbf{h}^{av}, \mathbf{h}^n \geq 0 \\ &= \mathbf{W}\mathbf{h} \quad s.t. \quad \mathbf{h} \geq 0 \end{aligned} \tag{2}$$

Here, $\mathbf{x}^s$ and $\mathbf{x}^n$ represent the spectrum of the source signal and the noise, respectively. $\mathbf{W}^s$, $\mathbf{N}$, and $\mathbf{h}^{av}$ and $\mathbf{h}^n$ represent the source dictionary, the noise dictionary and their activities at a frame, respectively. Note that all spectra are normalized for each frame.

Figure 3 shows a flow chart describing the proposed method. Here, $\mathbf{X}^{s,A}$ (D × L), $\mathbf{X}^{s,V}$(E × L), $\mathbf{W}^{s,A}$(D × J), $\mathbf{W}^{s,V}$(E × J), $\mathbf{N}^A$(D × K), $\mathbf{N}^V$(E × K), $\mathbf{H}^{av}$(J × L), $\mathbf{H}^n$(K × L), and $\mathbf{W}^{t,A}$(D' × J) represent the source audio signal, source visual signal, source audio dictionary, source visual dictionary, audio noise dictionary, visual noise dictionary, audio-visual activity, noise activity and target audio dictionary, respectively.

The joint matrix $\mathbf{h}$ is estimated on the basis of NMF with the sparse constraint that minimizes a cost function [4]. Previously, we used simple NMF without considering the weights of audio and visual parameters when estimating the activity [27]. Thus, we introduce audio-visual weights α and β because we must adjust the weight depending on the signal-to-noise ratio (SNR) and the new cost function as follows.

$$\begin{aligned}\alpha d\left(\mathbf{x}^{s,A}, \left[\mathbf{W}^{s,A}\ \mathbf{N}^A\right]\mathbf{h}\right) + \beta d\left(\mathbf{x}^{s,V}, \left[\mathbf{W}^{s,V}\ \mathbf{N}^V\right]\mathbf{h}\right) \\ + ||\lambda. * \mathbf{h}||_1 \quad s.t. \quad \mathbf{h} \geq 0\end{aligned} \tag{3}$$

Here, the first and second terms are the Kullback-Leibler (KL) divergence of audio data and visual data, respectively. The third term is the sparsity constraint with the L1-norm regularization term that causes $\mathbf{h}$ to be sparse. The symbol .∗ denotes element-wise multiplication. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\,\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$. Here, the weights for source exemplars $[\,\lambda_1 \ldots \lambda_J]$ are set to 0.1, and those for noise exemplars $[\,\lambda_{J+1} \ldots \lambda_{J+K}]$ are set to 0.

Note that the human voice has sparseness; however, noise signals do not. Therefore, we separate the human voice and noisy signals using NMF sparseness constraint λ. We experimentally set the value of λ of speech to 0.1 and of noise to 0 [5, 29] because the spectrum of the source signal should be expressed with a sparse representation of the source exemplars and the noise spectrum should not be expressed with a sparse representation.

$\mathbf{h}$ minimizing (3) is iteratively estimated by applying the following update rule.

$$h_j \leftarrow h_j \frac{\sum_d^D \mathbf{f}_d + \sum_e^E \mathbf{g}_e}{\alpha + \beta + \lambda_j} \tag{4}$$

$$\mathbf{f}_d = \alpha W_{d,j}^{s,A} \alpha x_d^a / (\alpha \mathbf{W}^{s,A} \mathbf{h}^{av})_d \tag{5}$$

$$\mathbf{g}_e = \beta W_{e,j}^{s,V} \beta x_e^v / (\beta \mathbf{W}^{s,V} \mathbf{h}^{av})_e \tag{6}$$

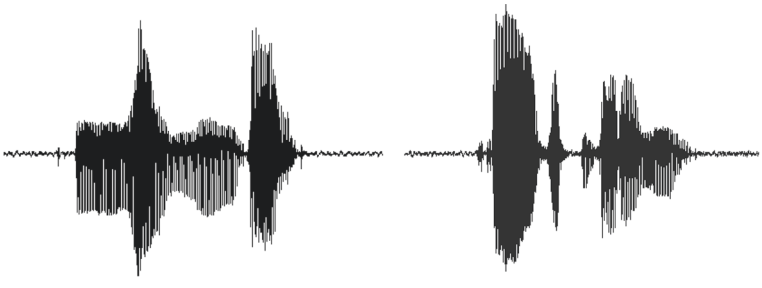

Fig. 4 Audio waves

Fig. 5 Lip images

Here, D and E represent the dimensions of the audio and visual dictionaries, respectively.

3.4 Target speech construction

From the estimated joint matrix $\mathbf{h}$, the activity of the source signal $\mathbf{h}^{av}$ is extracted, and using the activity and the target dictionary, the converted spectral features are constructed.

$$\hat{\mathbf{x}}^t = \mathbf{W}^{t,A}\mathbf{h}^{av} \tag{7}$$

The input source and converted spectral features are expressed as a STRAIGHT spectrum. Thus, the target speech is synthesized using a STRAIGHT synthesizer. The other features extracted by STRAIGHT analysis, such as F0 and the aperiodic components, are used to synthesize the converted signal without conversion.

4 Experimental results

4.1 Experimental conditions

The proposed multimodal VC technique was evaluated by comparing it with an exemplar-based audio-input method [5] and a conventional GMM-based method [6] in a speaker-conversion task using clean speech data and noise-added speech data.

The source speaker was a Japanese male, and the target speaker was a Japanese female. The target female audio data was taken from the CENSREC-1-AV [30] database. We recorded the source male audio-visual data with the same text as the target female utterances. Table 1 shows the content of the audio data taken from the CENSREC-1-AV database. We used a video camera (HDR-CX590, SONY) and a pin microphone (ECM-66B, SONY) for recording. We recorded audio and visual data simultaneously in a dark anechoic room. The camera was positioned 65 cm from the speaker and 130 cm from the floor.

Figures 4 and 5 show the recorded audio waves and lip images, respectively. We labelled the recorded audio data manually and used the labelled data in a subsequent experiment. The sampling rate of the audio data in each database was 8 kHz, and the frame shift was 5 ms.

A total of 40 utterances of clean continuous digital speech were used to construct parallel dictionaries in the NMF-based methods. These utterances were also used to train the GMM in the GMM-based method. Ten randomly selected utterances of clean and noisy continuous digital speech were used in the evaluation. Table 1 shows the content of the database, dictionary, and test data. We used the noise data from the CENSREC-C-1 [31] database. The noisy speech was created by adding white noise or a noise signal recorded in a car, airport, restaurant, or subway to the clean speech data. The SNRs were 0, 10, and 20 dB. The noise dictionary was extracted from the before- and after-utterance sections in the evaluation sentence. The average number of noisy frames was 223.

In the NMF-based methods, a 257-dimensional magnitude spectrum was used for the source and noise dictionaries and a 512-dimensional STRAIGHT spectrum was used for the target dictionary. STRAIGHT analysis

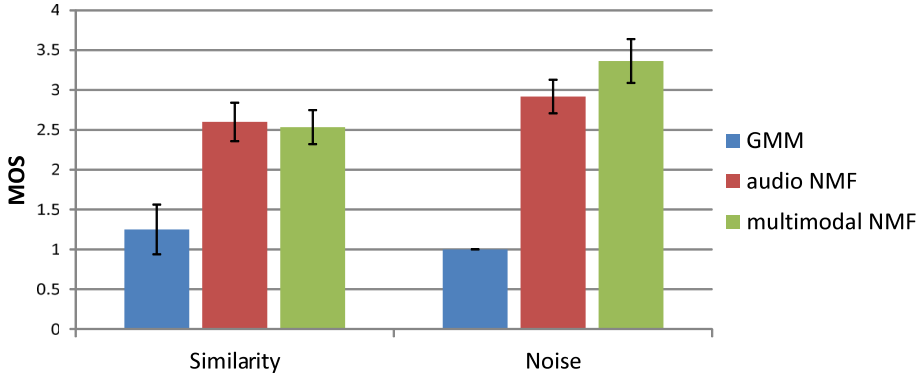

Fig. 6 Mel-cepstrum distortion in white noise environments (β is the weight of image feature)

Table 2 MCD in each noisy environment [dB]

Noise	Audio NMF	Multimodal NMF
White	3.113	2.788
Car	3.041	2.896
Airport	2.978	2.869
Restaurant	3.021	2.893
Subway	3.064	3.006

could not express the noise spectrum well. To express the noisy source speech with a sparse representation of source and noise dictionaries, a simple magnitude spectrum was used to construct the source and noise dictionaries. The DFT length was 40 ms. The number of iterations used to estimate the activity was 300. In the GMM-based method, the 1^{st} through 24^{th} MFCC obtained from the STRAIGHT spectrum were used as feature vectors. The number of mixtures was 32.

The frame rate of the visual data was 30 fps. The lip image size was 130 × 80 pixels. For visual features, 50-dimensions of the DCT coefficient of the lip motion images of the source speaker's utterance were used. We introduced segment features for the DCT coefficient that consist of consecutive frames (two frames before and two frames after). Therefore, the total dimension of the visual feature was 250. For the weights of the audio-visual feature, α was 1, and β was changed to 10 from 1.

4.2 Results and discussion

Figure 6 shows the mel-cepstral distortion (MCD) between the target signal and converted signals. The MCD is defined as follows.

$$\mathrm{MCD\ [dB]} = 10/\ln 10 \sqrt{2\sum_{d=1}^{24}\left(mc_d^t - \hat{mc}_d^t\right)^2} \qquad (8)$$

Here, mc_d^t and $\hat{mc}_d^t$ represent the d^{th} coefficient of the target and the converted mel-cepstra, respectively. We calculated the mel-cepstra from the converted STRAIGHT spectrum. Figure 6 shows that the distortion of the proposed method is lower than that of the conventional NMF and GMM-based methods in noisy environments. The proposed method obtained lower distortion than the non-weighted method ($\beta = 1$) by selecting an optimal image feature weight. These results indicate the effectiveness of the proposed method.

For an SNR of 0 dB, the best performance of the proposed method was 2.976 dB ($\beta = 5$), which is 0.338 dB less than that of audio-input NMF. In addition, for an SNR of 20 dB, the best performance was 2.674 dB ($\beta = 7$), which is 0.271 lower than audio-input NMF. Therefore, the performance difference between the conventional NMF method and the proposed method was greater in a low SNR environment.

In a clean environment, the proposed method did not show significant difference when compared with audio-input NMF.

We also investigated the effectiveness of the proposed method in various noisy environments. Table 2 shows the MCD in each noisy environment, where the SNR was 10 dB and the image feature weight β was set to 5. Table 2 shows that the MCD of the proposed method is less than that of audio-input NMF in each noisy environment. Thus, we confirm the effectiveness of the proposed method in various noisy environments.

In addition, we calculated the improvement ratio for vowels and consonants using the labelled data. We used the mel-cepstral distortion ratio (MCDR) to evaluate improvement ratio. These values are shown in Table 3. The MCDR is defined as follows.

$$\mathrm{MCDR} = \frac{\sqrt{\sum_{d=1}^{24}(mc_d^t - mc_d^s)^2}}{\sqrt{\sum_{d=1}^{24}(mc_d^t - \hat{mc}_d^t)^2}} \qquad (9)$$

Table 3 Improvement ratio of MCD

Phoneme	Audio NMF	Multimodal NMF
Vowels	1.508	1.620
Consonants	1.477	1.676

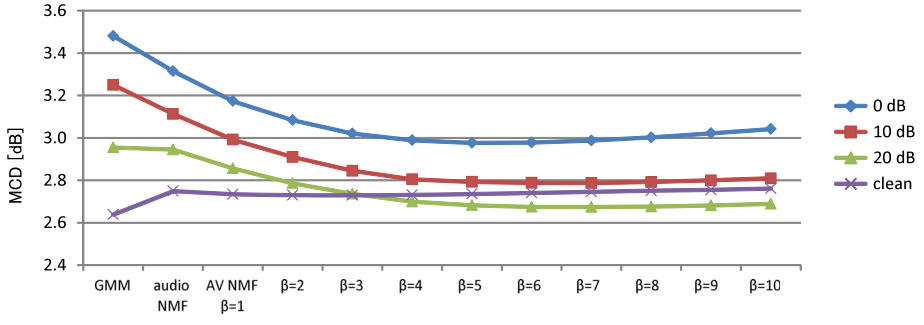

Fig. 7 Mean opinion score for subjective evaluations

Here, mc_d^s represents the d^{th} coefficient of the source signal. The SNR was 10 dB and the image feature weight β was 5 in this experiment. Table 3 shows that the performance of the proposed method outperforms audio-input NMF in the evaluation of both vowels and consonants. In audio-input conversion, the improvement ratio of vowels is better than that of consonants; however, in audio-visual conversion, the improvement ratio of consonants is better because image features compensate for the conversion of the consonant part, which is degraded by the noisy signal in audio-input conversion.

We also performed a mean opinion score (MOS) test [32] on the similarity and noise suppression of the converted speech. The opinion score was set to a 5-point scale (5; excellent, 4; good, 3; fair, 2; poor, 1; bad). The tests were performed with 11 subjects. For the evaluation of similarity, each subject listened to the converted speech and evaluated how similar the sample was to the target speech. For the evaluation of noise suppression, each subject listened to the converted speech and evaluated the degree of noise suppression in the sample.

Figure 7 shows the results of the MOS test. The error bars show 95 % confidence intervals. As can be seen, the performance of the GMM-based method degraded considerably. This may be because the noise caused unexpected mapping in the GMM-based method. Conversely, the performance degradations of the VC methods based on our proposed multimodal NMF, and audio NMF were less than that of the GMM-based method. In addition, in the noise suppression test, the proposed method obtained a higher score than the other two methods. This result demonstrates the noise robustness of the proposed multimodal VC method.

5 Conclusions

We have proposed multimodal VC using NMF based on the idea of sparse representation and introduced the weight of audio-visual features. In the proposed method, the joint audio-visual feature is used as the source feature. Noisy audio-visual features are then decomposed into a linear combination of the clean audio-visual feature and the noise feature. By replacing the source speaker's audio-visual feature with the target speaker's audio feature, the voice individuality of the source speaker is converted to the target speaker. Furthermore, we have introduced audio-visual weights and formulated a new cost function. By selecting an optimal weight of the image feature, we achieve good transformation accuracy. Our experimental results demonstrate the superior effectiveness of the proposed VC technique compared with conventional audio-input NMF and GMM-based VC. In addition, we have shown that the proposed method is effective in several noisy environments.

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-ku, 657-8501 Kobe, Japan. [2]Organization of Advanced Science and Technology, Kobe University, 1-1 Rokkodai, Nada-ku, 657-8501 Kobe, Japan.

Received: 26 December 2014 Accepted: 8 June 2015
Published online: 04 September 2015

References

1. DD Lee, HS Seung, Algorithms for non-negative matrix factorization. Advances in Neural Information Processing System 13, 556–562 (2000)
2. T Virtanen, Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech, Lang. Process. **15**(3), 1066–1074 (2007)
3. MN Schmidt, RK Olsson, in *Interspeech*. Single-channel speech separation using sparse non-negative matrix factorization (Pittsburgh, Pennsylvania, USA, 2006)
4. JF Gemmeke, T Viratnen, A Hurmalainen, Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio, Speech and Language Processing. **19**(7), 2067–2080 (2011)
5. R Takashima, T Takiguchi, Y Ariki, in *SLT*. Exemplar-based voice conversion in noisy environment (Miami, Florida, USA, 2012), pp. 313–317
6. Y Stylianou, O Cappe, E Moilines, Continuous probabilistic transform for voice conversion. IEEE Trans. Speech and Audio Processing. **6**(2), 131–142 (1998)
7. G Potamianos, HP Graf, in *ICASSP*. Discriminative training of HMM stream exponents for audio-visual speech recognition (Seattle, Washingron, USA, 1998), pp. 3733–3736
8. A Verma, T Faruquie, C Neti, S Basu, A Senior, in *ASRU*. Late integration in audio-visual continuous speech recognition (Keystone, Colorado, USA, 1999)
9. MJ Tomlinson, MJ Russell, NM Brooke, in *ICASSP*. Integrating audio and visual information to provide highly robust speech recognition (Atlanta, Georgia, USA, 1996), pp. 821–824
10. Y Komai, N Yang, T Takiguchi, Y Ariki, in *ACM Multimedia*. Robust aam-based audio-visual speech recognition against face direction changes (Nara, Japan, 2012), pp. 1161–1164
11. Timothy F, GJE Cootes, CJ Taylor, Active appearance models. IEEE Trans. Pattern. Anal. Mach. Intell. **23**, 681–685 (2001)
12. C Veaux, X Robet, in *Interspeech*. Intonation conversion from neutral to expressive speech (Florence, Italy, 2011), pp. 2765–2768
13. R Aihara, R Takashima, T Takiguchi, Y Ariki, GMM-based emotional voice conversion using spectrum and prosody features. Am. J. Signal Process. **2**(5), 134–138 (2012)
14. K Nakamura, T Toda, H Saruwatari, K Shikano, Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Comm. **54**(1), 134–146 (2012)
15. R Aihara, R Takashima, T Takiguchi, Y Ariki, in *ICASSP*. Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization (Lyon, France, 2013), pp. 8037–8040
16. K Nakamura, T Toda, H Saruwatari, K Shikano, in *Interspeech*. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech (Pittsburgh, Pennsylvania, USA, 2006), pp. 148–151
17. A Kain, MW Macon, in *ICASSP*. Spectral voice conversion for text-to-speech synthesis (Las Vegas, Nevada, USA, 1998), pp. 285–288
18. M Abe, S Nakamura, K Shikano, H Kuwabara, in *ICASSP*. Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models (New York, USA, 1988), pp. 655–658
19. H Valbret, E Moulines, JP Tubach, Voice transformation using PSOLA technique. Speech Comm. **11**, 175–187 (1992)
20. T Toda, A Black, K Tokuda, Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech, Lang. Process. **15**(8), 2222–2235 (2007)
21. E Helander, T Virtanen, J Nurminen, M Gabbouj, Voice conversion using partial least squares regression. IEEE Trans. Audio, Speech, Lang. Process. **18**, 912–921 (2010)

22. CH Lee, CH Wu, in *Interspeech*. Map-based adaptation for speech conversion using adaptation data selection and non-parallel training (Pittsburgh, Pennsylvania, USA, 2006), pp. 2254–2257
23. T Toda, Y Ohtani, K Shikano, in *Interspeech*. Eigenvoice conversion based on Gaussian mixture model (Pittsburgh, Pennsylvania, USA, 2006), pp. 2446–2449
24. D Saito, K Yamamoto, N Minematsu, K Hirose, in *Interspeech*. One-to-many voice conversion based on tensor representation of speaker space (Florence, Italy, 2011), pp. 653–656
25. R Aihara, R Takashima, T Takiguchi, Y Ariki, in *ICASSP*. Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary (Florence, Italy, 2014), pp. 7944–7948
26. Z Wu, T Virtanen, ES Chng, H Li, Exemplar-based sparse representation with residual compensation for voice conversion. IEEE/ACM Transactions on Audio, Speech, and Language. **22**, 1506–1521 (2014)
27. K Masaka, R Aihara, T Takiguchi, Y Ariki, in *ICASSP*. Mutimodal voice conversion using non-negative matrix factorization in noisy environments (Florence, Italy, 2014), pp. 1561–1565
28. K Masaka, R Aihara, T Takiguchi, Y Ariki, in *Interspeech*. Multimodal exemplar-based voice conversion using lip features in noisy (Singapore, 2014), pp. 1159–1163
29. T Virtanen, BT Raj, JF Gemmeke, HV Hamme, in *ICASSP*. Active-set newton algorithm for non-negative sparse coding of audio (Florence, Italy, 2014), pp. 3092–3096
30. T Satoshi, M Chiyomi, Censrec-1-av an evaluation framework for multimodal speech recognition (japanese). Technical report. **SLP-82**(7), 1–6 (2010)
31. 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)
32. INTERNATIONAL TELECOMMUNICATION UNION, Methods for objective and subjective assessment of quality. ITU-T Recommendation, 800–899 (2003)