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

RESEARCH Open Access

A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary

Ryo Aihara[1*], Ryoichi Takashima[1], Tetsuya Takiguchi[2] and Yasuo Ariki[2]

Abstract

We present in this paper a voice conversion (VC) method for a person with an articulation disorder resulting from athetoid cerebral palsy. The movement of such speakers is limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In this paper, exemplar-based spectral conversion using nonnegative matrix factorization (NMF) is applied to a voice with an articulation disorder. To preserve the speaker's individuality, we used an individuality-preserving dictionary that is constructed from the source speaker's vowels and target speaker's consonants. Using this dictionary, we can create a natural and clear voice preserving their voice's individuality. Experimental results indicate that the performance of NMF-based VC is considerably better than conventional GMM-based VC.

Keywords: Voice conversion; NMF; Articulation disorders; Voice reconstruction; Assistive technologies

1 Introduction

In recent years, a number of assistive technologies using information processing have been proposed, for example, sign language recognition using image recognition technology [1-3], text reading systems from natural scene images [4-6], and the design of wearable speech synthesizers [7]. In this study, we focused on a person with an articulation disorder resulting from athetoid cerebral palsy. There are about 34,000 people with speech impediments associated with an articulation disorder in Japan alone, and one of the causes of speech impediments is cerebral palsy.

Cerebral palsy is a result of damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder: before birth, at the time of delivery, and after birth. Cerebral palsy is classified into the following types: (1) spastic, (2) athetoid, (3) ataxic, (4) atonic, (5) rigid, and (6) a mixture of these types [8].

Athetoid symptoms develop in about 10% to 15% of cerebral palsy sufferers. In the case of a person with this type of articulation disorder, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to the athetoid symptoms. Athetoid symptoms also restrict the movement of their arms and legs. Most people suffering from athetoid cerebral palsy cannot communicate by sign language or writing, so there is a great need for voice systems for them.

In [9], we proposed robust feature extraction based on principal component analysis (PCA) with more stable utterance data instead of DCT. In [10], we used multiple acoustic frames (MAF) as an acoustic dynamic feature to improve the recognition rate of a person with an articulation disorder, especially in speech recognition using dynamic features only. In spite of these efforts, the recognition rate for articulation disorders is still lower than that of physically unimpaired persons. Maier et. al. [11] proposed automatic speech recognition systems for the evaluation of speech disorders that resulted from head and neck cancer.

In this paper, we propose a voice conversion (VC) method for articulation disorders. Regarding speech recognition for articulation disorders, the recognition rate

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

using a speaker-independent model, which is trained by well-ordered speech, is 3.5% [9]. This result implies that the utterance of a person with an articulation disorder is difficult to understand for people who have not communicated with them before. In recent years, people with an articulation disorder may use slide shows and a previously synthesized voice when they give a lecture. However, because their movement is restricted by their athetoid symptoms, to make slides or synthesize their voice in advance is hard for them. People with articulation disorders desire a VC system that converts their voice into a clear voice that preserves their voice's individuality. However, a speech conversion method for people with articulation disorders resulting from athetoid cerebral palsy has not been successfully developed.

In the research discussed in this paper, we conducted VC for articulation disorders using nonnegative matrix factorization (NMF) [12]. NMF is a matrix decomposition method with nonnegativity constraint. In the field of speech processing, NMF is a well-known approach for source separation and speech enhancement [13]. In these approaches, the observed vector is represented by a linear combination of a small number of elementary vectors, referred to as the basis, and its weights. The collection of the basis is called a 'dictionary', and the joint matrix of weights is called an 'activity':

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

where $\mathbf{x}_l$ is the lth frame of the observation, and $\mathbf{a}_j$ and $h_{j,l}$ are the jth basis and the weight, respectively. $\mathbf{A}$ and $\mathbf{h}_l$ are the dictionary and the activity of frame l, respectively. In some separation approaches, a dictionary is constructed for each source, and the mixed signals are expressed with a sparse representation of these dictionaries. Using only the weights (called activity in this paper) of basis in the target dictionary, the target signal can be reconstructed. Gemmeke et al. also used the activity of the speech dictionary as phonetic scores instead of likelihoods of hidden Markov models (HMMs) for speech recognition [14].

In our study, we adopt the supervised NMF approach [15], with a focus on VC from poorly articulated speech resulting from articulation disorders into well-ordered articulation. An input spectrum with an articulation disorder is represented by a linear combination of an articulation disorder basis and its weights using NMF. By replacing the basis produced by someone with an articulation disorder with a well-ordered basis, the original speech spectrum is replaced with a well-ordered spectrum. In the voice of a person with an articulation disorder, their consonants are often unstable and that makes their voices unclear. Hence, by replacing the articulation disorder basis of consonants only, a voice with an articulation disorder is converted into a clear voice that preserves the individuality of the speaker's voice.

The rest of this paper is organized as follows. Section 2 discusses related works, while Section 3 describes the NMF-based VC method. Section 4 presents the experimental results, and the final section presents the conclusions.

2 Related works

VC is a technique for changing specific information in an input speech while retaining the other information in the utterance such as its linguistic information. Unlike speech synthesis, VC application does not need text input. Many statistical approaches to VC have been studied and applied to various tasks. One of the most popular VC applications is speaker conversion [16]. 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. The other information, such as its linguistic information or emotional information, is retained. Emotion conversion is a technique for changing emotional information in input speech while maintaining linguistic information and speaker individuality [17,18]. With those approaches, a mapping function is trained in advance using a small amount of training data consisting of utterance pairs consisting of source and target voices.

A Gaussian mixture model (GMM)-based approach is widely used for VC because of its flexibility and good performance [19]. 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) using a parallel training set. A number of improvements to this approach have been proposed. Toda et al. [16] introduced the global variance (GV) of the converted spectra over time sequence. Helander et al. [20] proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem of standard multivariate regression. However, over-smoothing and over-fitting problems in these GMM-based approaches have been reported [20] because of statistical averages and large number of parameters. These problems degrade the quality of synthesized speech. There have also been approaches that do not require parallel data using GMM adaptation techniques [21] or eigen-voice GMM (EV-GMM) [22].

However, these approaches have been developed for speaker conversion. If the person with an articulation disorder is set as a source speaker and a physically unimpaired person is set as a target speaker, an articulation disorder voice may be converted into a well-ordered voice, but the source speaker's voice individuality is also converted into the target speaker's individuality.

In the field of assistive technology, Nakamura et al. [23,24] proposed GMM-based VC systems that reconstruct a speaker's individuality in electrolaryngeal speech and speech recorded by NAM microphones. These systems are effective for electrolaryngeal speech and speech recorded by NAM microphones, but the target speaker's individuality will be changed to the source speaker's individuality. Veaux et al. [25] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders. HMM-based speech synthesis needs text input to synthesize speech. In the case of people with an articulation disorder resulting from athetoid cerebral palsy, it is difficult for them to input text because of their athetoid symptoms.

The goals of this study can be divided into in following three points: (1) convert the voice uttered by a person with an articulation disorder so that everyone can understand what he/she said, (2) preserve the individuality of the speaker's voice, and (3) output a natural-sounding voice. Our proposed exemplar-based VC can create a natural-sounding voice because there is no statistical model in our approach, and the source speaker's individuality can be preserved using our individuality-preserving dictionary.

3 Voice conversion based on NMF

3.1 Exemplar-based voice conversion

Figure 1 shows the basic approach of our exemplar-based VC using NMF. L and J represent frames of dictionary and basis of dictionary, respectively. D represents the number of dimensions of the source feature, which consists of a segment feature of the source speaker's spectrum, and d represents the number of dimensions of a target feature, which consists a single feature of the target speaker's spectrum.

A dictionary is a collection of source or target basis. Our VC method needs two dictionaries that are phonemically parallel. One dictionary is a source dictionary, which is constructed from a source feature. The source feature is constructed from an articulation-disordered spectrum and its segment feature which consists of some consecutive frames. We use a segment feature in order to consider temporal information in activity estimation. The other dictionary is a target dictionary, which is constructed from a target feature. The target feature is mainly constructed from a well-ordered spectrum. These two dictionaries consist of the same words and are aligned with dynamic time warping (DTW). Hence, these dictionaries have the same number of bases.

Input source features, $\mathbf{X}^s$, which consist of an articulation-disordered spectrum and its segment features, are decomposed into a linear combination of bases from the source dictionary $\mathbf{A}^s$ by NMF. The weights of the basis are estimated as an activity $\mathbf{H}^s$. Therefore, the activity includes the weight information of input features for each basis.

Then, the activity is multiplied by a target dictionary in order to obtain converted spectral features $\hat{\mathbf{X}}^t$ which are represented by a linear combination of bases from the target dictionary. Because the source and target dictionaries are parallel phonemically, the basis used in the converted

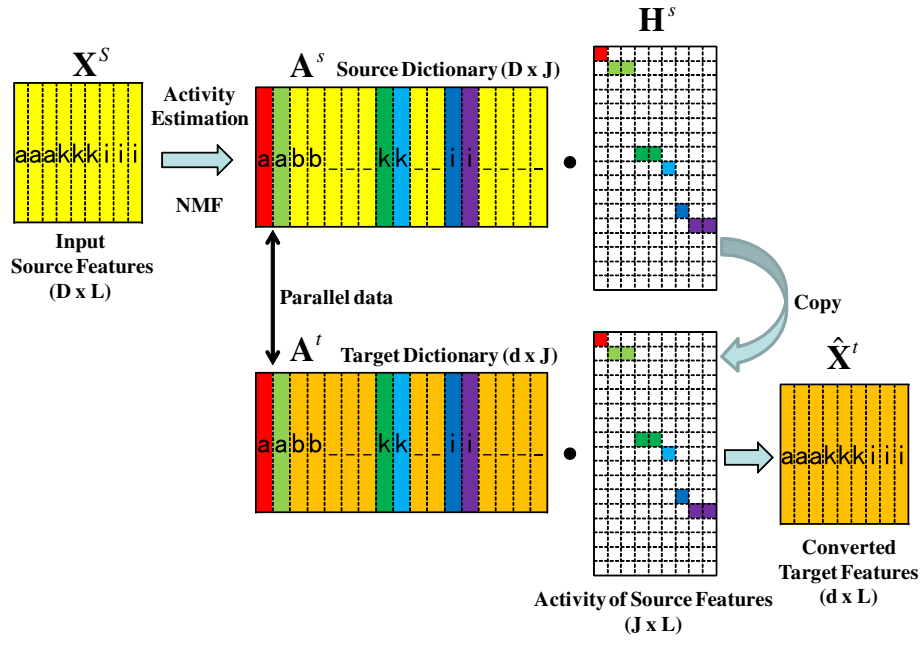

Figure 1 Basic approach of NMF-based voice conversion.

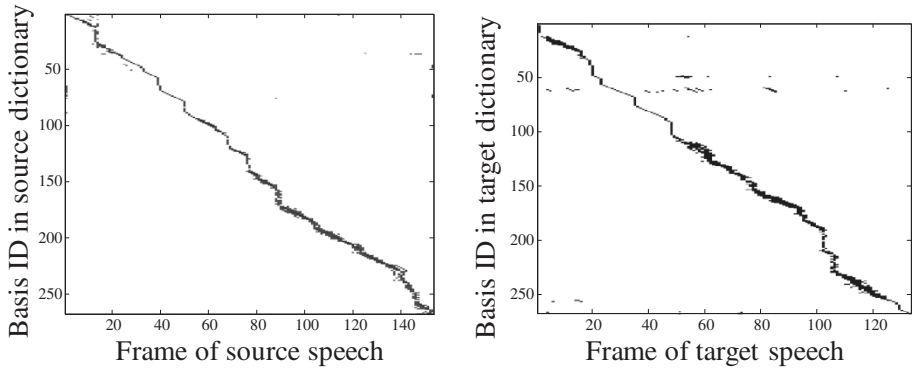

Figure 2 Activity matrices for the articulation-disordered utterance (left) and well-ordered utterance (right).

features is phonemically the same as that of the source features.

Figure 2 shows an example of the activity matrices estimated from the word *ikioi* ('vigor' in English). One was uttered by a person with an articulation disorder and the other by a physically unimpaired person. To show an intelligible example, each dictionary was structured from just the one word *ikioi* and aligned with DTW. As shown in Figure 2, these activities have high energies at similar elements. For this reason, when there are parallel dictionaries, it may be possible to substitute the activity of the source features estimated with the source dictionary with that of the target features. Therefore, the target speech can be constructed using the target dictionary and the activity of the source signal as shown in Figure 1.

Spectral envelopes extracted by STRAIGHT analysis [26] are used in the source and target features. The other features extracted by STRAIGHT analysis, such as F0 and the aperiodic components, are used to synthesize the converted signal without any conversion.

3.2 Preserving the individuality of the speaker's voice

In order to make a parallel dictionary, some pairs of parallel utterances are needed, where each pair consists of the same text. One is spoken by a person with an articulation disorder (source speaker), and the other is spoken by a physically unimpaired person (target speaker).

The left side of Figure 3 shows the process for constructing a parallel dictionary. Spectrum envelopes, which are extracted from parallel utterances, are phonemically aligned. In order to estimate activities of source features precisely, segment features, which consist of some consecutive frames, are constructed. Target features are constructed from consonant frames of the target's aligned spectrum and vowel frames of the source's aligned spectrum. Source and target dictionaries are constructed by

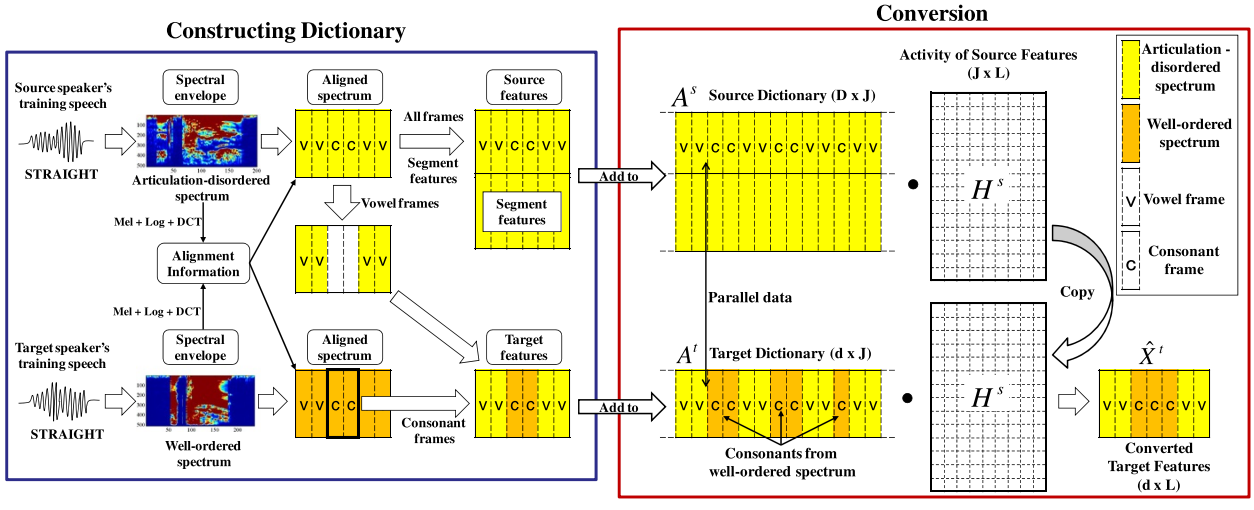

Figure 3 Individuality-preserving voice conversion.

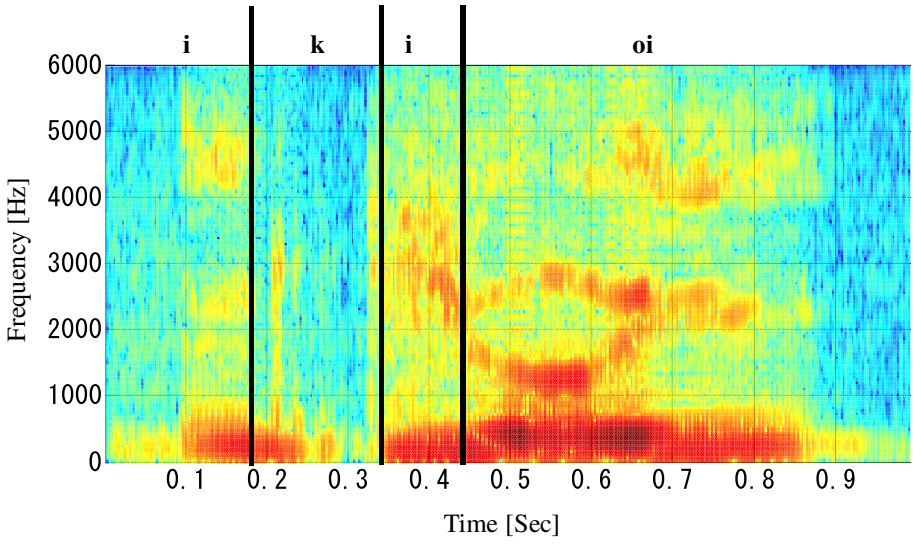

Figure 4 Example of a spectrogram spoken by a person with an articulation disorder i k i oi.

lining up each of the features extracted from parallel utterances.

Figure 3 shows how to preserve a source speaker's voice individuality in our VC. The vowels voiced by the speaker strongly indicate the speaker's individuality. On the other hand, consonants of people with articulation disorders are often unstable. Figure 4 shows an example of the spectrogram for the word *ikioi* (vigor in English) of a person with an articulation disorder. The spectrogram of a physically unimpaired person speaking the same word is shown in Figure 5. In Figure 4, the area labeled 'k' is not clear, compared to the same region in Figure 5. By combining a source speaker's vowels and target speaker's consonants in the target dictionary, the individuality of the source speaker's voice can be preserved.

3.3 Estimation of activity

In the NMF-based approach, the spectrum source signal at frame l is approximately expressed by a nonnegative linear combination of the source dictionary and the activities:

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}_l^s \\ &\approx \sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s \\ &= \mathbf{A}\mathbf{h}_l \quad \text{s.t.} \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{2}$$

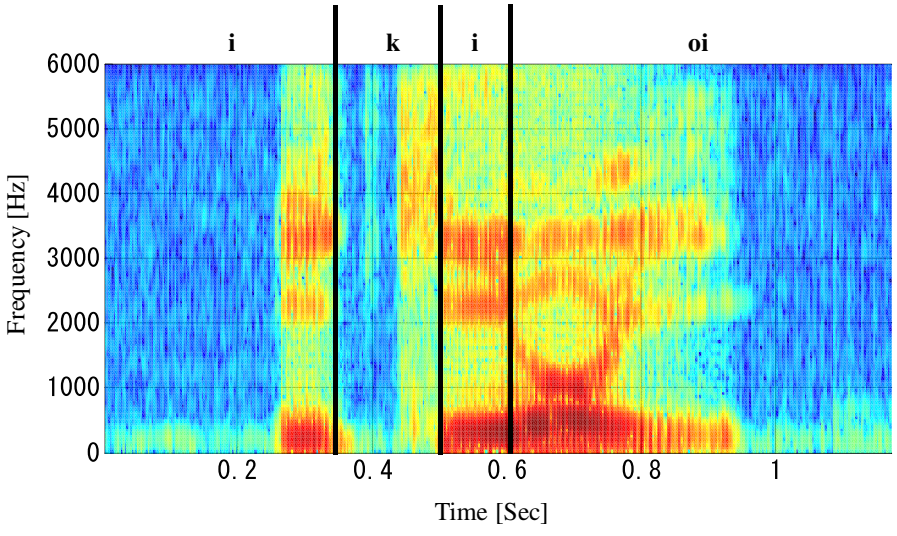

Figure 5 Example of a spectrogram spoken by a physically unimpaired person i k i oi.

where $\mathbf{x}_l^s \in \mathbf{X}^s$ is the magnitude spectra of the source signal. Given the spectrogram, Equation 2 can be written as follows:

$$\mathbf{X}^s \approx \mathbf{A}^s\mathbf{H}^s \quad \text{s.t.} \quad \mathbf{H}^s \geq 0. \tag{3}$$

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

$$d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + ||(\lambda \mathbf{1}^{1\times L}) .* \mathbf{H}^s||_1 \quad \text{s.t.} \quad \mathbf{H}^s \geq 0 \tag{4}$$

where $\mathbf{1}$ is an all-one matrix and $.*$ denotes element-wise multiplication, 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. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \ldots \lambda_J]$.

$\mathbf{H}^s$ minimizing Equation 4 is estimated iteratively applying the following update rule [12,27]:

$$\mathbf{H}^s_{n+1} = \mathbf{H}^s_n .* (\mathbf{A}^{sT}(\mathbf{X}^s ./ (\mathbf{A}^s\mathbf{H}^s))) \\ ./(\mathbf{A}^{sT}\mathbf{1}^{D\times L} + \lambda \mathbf{1}^{1\times L}). \tag{5}$$

To increase the sparseness of $\mathbf{H}^s$, elements of $\mathbf{H}^s$, which are less than threshold, are rounded to zero.

Using the activity and the target dictionary, the converted spectral features are constructed:

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s). \tag{6}$$

4 Experiments

4.1 Experimental conditions

The proposed method was evaluated on word-based VC for one person with an articulation disorder. We recorded 432 utterances (216 words, repeating each two times) included in the ATR Japanese speech database [28]. The speech signals were sampled at 16 kHz and windowed with a 25-ms Hamming window every 10 ms. A physically unimpaired Japanese male in the ATR Japanese speech database was chosen as a target speaker.

The source feature is a 2,565-dimensional segment spectrum, and the target feature is a 513-dimensional single spectrum. Those spectra are extracted by STRAIGHT analysis. The mel-cepstral coefficient, which is converted from the STRAIGHT spectrum, is used for DTW in order to align the temporal fluctuation.

We compared our NMF-based VC to conventional GMM-based VC. In GMM-based VC, the first through 24th cepstral coefficients extracted by STRAIGHT are used as source and target features.

4.2 Objective evaluation of exemplar-based VC

Mel-cepstral distortion between the target and converted mel-cepstra given by the following equation [29] was used in order to evaluate the size of training words and the weights of the sparsity constraint:

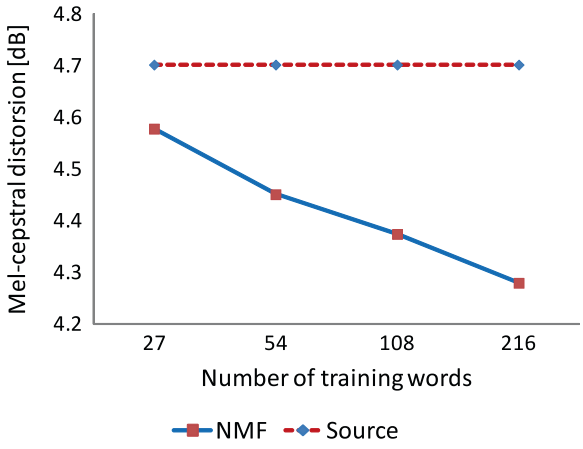

Figure 6 Mel-cepstral distortion as a function of the number of training words.

$$\mathrm{MCD}(v^{\mathrm{conv}}, v^{\mathrm{ref}}) = (\alpha/T)\sum_{t=0}^{T-1}\sqrt{\sum_{d=1}^{24}(v_d^{\mathrm{conv}}(t) - v_d^{\mathrm{ref}}(t))^2} \tag{7}$$

$$\alpha = 10\sqrt{2}/\ln 10 \tag{8}$$

where $v_d^{\mathrm{conv}}(t)$ and $v_d^{\mathrm{ref}}(t)$ are the dth coefficients on frame t of the converted and target mel-cepstra, T is a total number of frames, and α is a scaling factor, respectively.

We selected 50 words at random as an evaluation set and used the target speaker's utterance as a reference. An individuality-preserving dictionary is not suitable for mel-cepstral distortion because vowel frames in converted voices are different from the reference. Therefore, in this

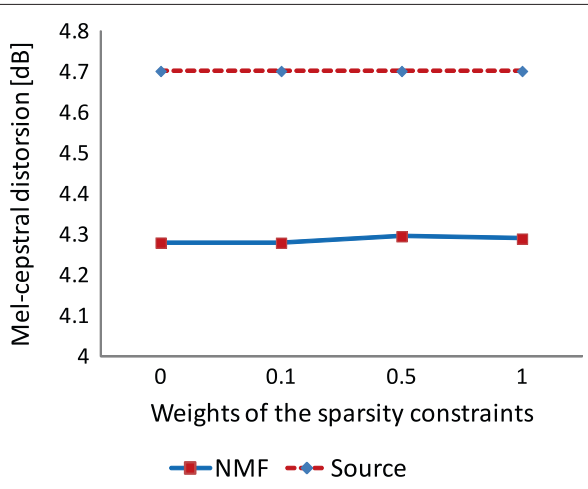

Figure 7 Mel-cepstral distortion as a function of the weights of the sparsity constraints.

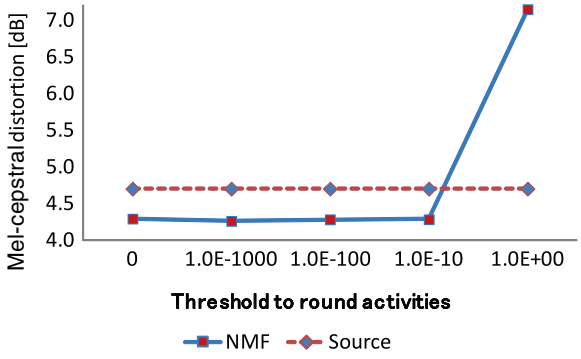

Figure 8 Mel-cepstral distortion as a function of the threshold to round activities.

subsection, we used a target dictionary that is constructed from the target speaker's vowels and consonants.

Figure 6 shows the mel-cepstral distortions in the evaluation set as a function of the number of training words. The distortion before conversion (shown by the dotted line) is 4.70 dB. With NMF-based VC, there is less distortion as the number of training words increases.

Figure 7 shows the mel-cepstral distortions in the evaluation set as a function of the weights of the sparsity constraint (λ in Equation 4). The distortion between the converted voice and the reference is almost the same despite the increase in the weights of the sparsity constraint.

Figure 8 shows the mel-cepstral distortions in the evaluation set as a function of the threshold to round activities to zero. From Figure 8, the threshold was set to 0.1.

4.3 Subjective evaluation

We used 216 utterances for training and used the other 216 utterances for the test. We conducted subjective evaluation on three topics. A total of ten Japanese speakers

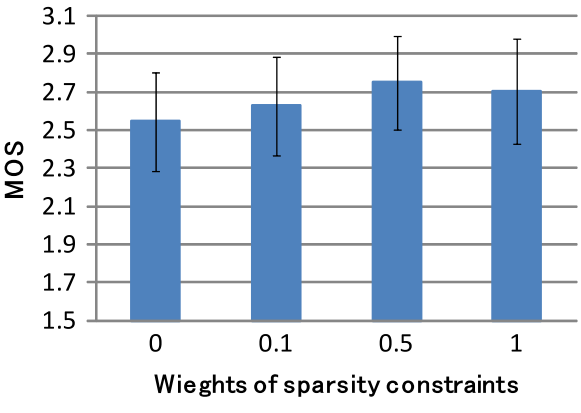

Figure 9 Results of MOS test on speech quality as a function of the weights of the sparsity constraints.

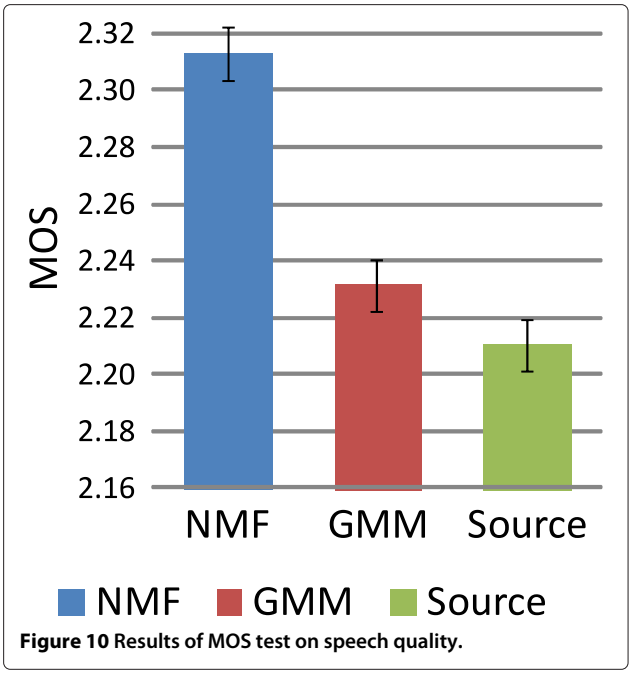

Figure 10 Results of MOS test on speech quality.

took part in the test using headphones. For the 'speech quality' evaluation, we performed a mean opinion score (MOS) test [30]. The opinion score was set to a five-point scale (5 excellent, 4 good, 3 fair, 2 poor, 1 bad). Thirty-two words, which are difficult for a person with an articulation disorder to utter, were evaluated. The subjects were asked about the speech quality in the articulation-disordered voice, the NMF-based converted voice, and the GMM-based converted voice. Each voice uttered by a physically

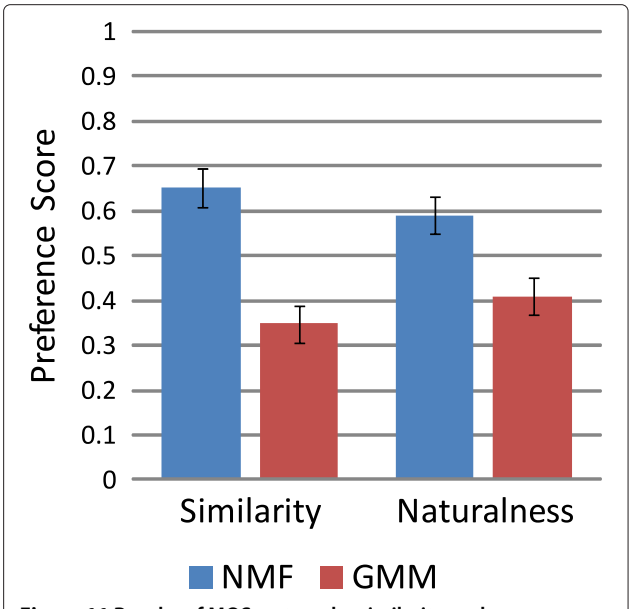

Figure 11 Results of MOS test on the similarity to the source speaker and naturalness.

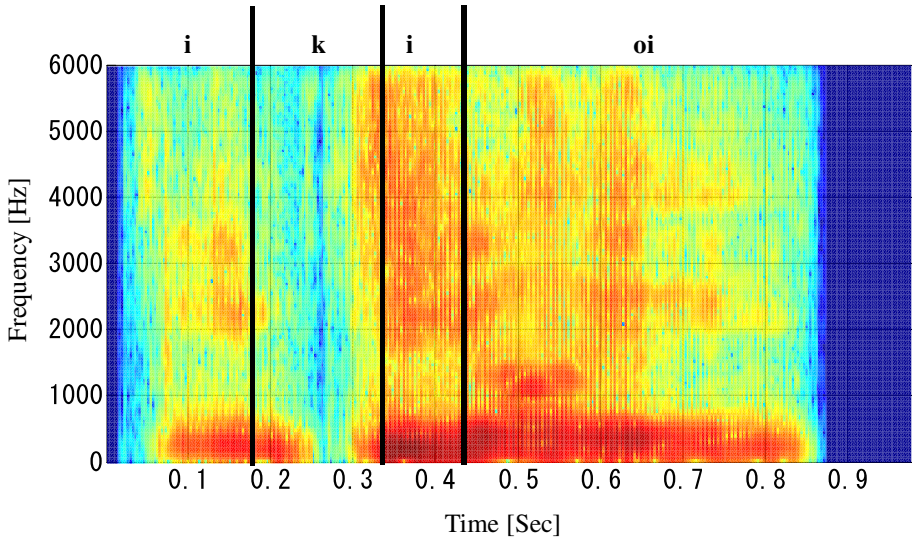

Figure 12 Example of a spectrogram converted by NMF-based VC i k i oi.

unimpaired person was presented as a reference of five points on the MOS test.

Fifty words were converted using NMF-based VC and GMM-based VC for the following evaluations. On the 'similarity' evaluation, the XAB test was carried out. In the XAB test, each subject listened to the articulation-disordered voice. Then the subject listened to the voice converted by the two methods and selected which sample sounded most similar to the articulation-disordered voice. On the 'naturalness' evaluation, a paired comparison test was carried out, where each subject listened to pairs of speech converted by the two methods and selected which sample sounded more natural.

4.4 Results and discussion

Figure 9 shows the results of the MOS test on the speech quality of the words converted by NMF-based VC, where the weights of the sparsity constraints are marked on the horizontal axis. The error bars show a 95% confidence score. The p value test of 0.05 showed that there are no significant differences in these scores. In this paper, all elements for λ in Equation 5 were set to 0.5 from Figure 9.

Figure 10 shows the results on the MOS test on speech quality. The error bars show a 95% confidence score. The difference of MOS between NMF-based VC and GMM-based VC was confirmed by a p value test of 0.05. Moreover, the p value test of 0.05 showed that there are

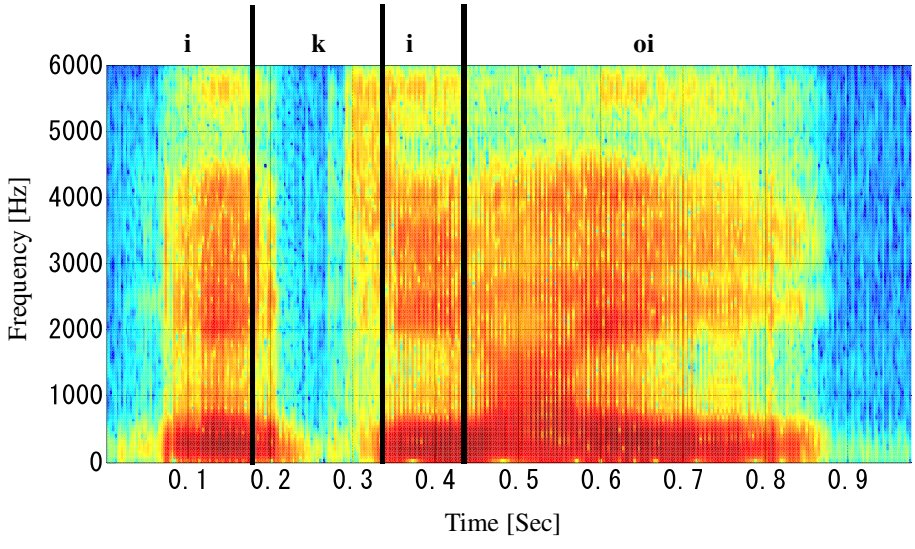

Figure 13 Example of a spectrogram converted by GMM-based VC i k i oi.

no significant differences between GMM-based VC and utterances of a person with an articulation disorder. The difference between NMF-based VC and an articulation-disordered utterance was significant at the 0.1 level. As shown in Figure 10, NMF-based VC obtained a higher score than GMM-based VC on the speech quality test. This is because GMM-based VC creates conversion noise. NMF-based VC also creates some conversion noise, but it is less than that created by GMM-based VC. Using the individuality-preserving dictionary, we can reduce the effect of noise in the vowel portion.

Figure 11 shows the results of the XAB test on the similarity to the source speaker and naturalness of the converted voice. The error bars show a 95% confidence score. The NMF-based VC obtained a higher score than the GMM-based conversion on similarity because NMF-based conversion uses an individuality-preserving dictionary. The NMF-based VC also obtained a higher score than the GMM-based conversion on naturalness although NMF-based conversion mixed the source speaker's vowels and target speaker's consonants. The results of this test were confirmed by a p value test of 0.05.

Figure 12 shows an example of an NMF-based converted spectrogram of *ikioi*. The same word converted by GMM-based VC is also shown in Figure 13. In Figure 12, the 'oi' area is similar to the same area in Figure 4, as compared to Figure 13. Using the individuality-preserving dictionary, NMF-based VC can keep naturalness in the vowel portion because our method converts consonants only.

5 Conclusions

We proposed a spectral conversion method based on NMF for a voice with an articulation disorder. By combining articulation-disordered vowels and well-ordered consonants, we constructed an individuality-preserving dictionary. Experimental results demonstrated that our VC method can improve the listening speech quality of the words uttered by a person with an articulation disorder.

Our proposed method has the following benefits compared to conventional GMM-based VC: (1) NMF-based VC can preserve the individuality of the source speaker's voice using an individuality-preserving dictionary, and (2) our NMF-based VC can create a natural voice because the individuality-preserving dictionary keeps the source speaker's vowels.

Some problems remain with this method. Articulation-disordered speech has a co-articulation effect between phonemes. As shown in Figure 4, the mispronunciation of consonants also affects the vowels that follow. Solving this problem will be the focus of our future work. In this study, there was only one subject person, so in future experiments, we will increase the number of subjects and further examine the effectiveness of our method.

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

Acknowledgements
This research was supported in part by MIC SCOPE.

Author details
[1]Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan. [2]Organization of Advanced Science and Technology, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan.

Received: 5 April 2013 Accepted: 25 January 2014
Published: 1 February 2014

References

1. J Lin, W Ying, TS Huang, Capturing human hand motion in image sequences, in *IEEE Workshop on Motion and Video Computing,* Orlando, 5-6 Dec 2002 (IEEE, Piscataway, 2002), pp. 99–104
2. T Starner, J Weaver, A Pentland, Real-time American sign language recognition using desk and wearable computer based video. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12), 1371–1375 (1998)
3. G Fang, W Gao, D Zhao, Large vocabulary sign language recognition based on hierarchical decision trees. 5th International Conference on Multimodal Interfaces. **34**(3), 125–131 (2004)
4. N Ezaki, M Bulacu, L Schomaker, Text detection from natural scene images: towards a system for visually impaired persons. Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004). **2**, 683–686 (2004)
5. MK Bashar, T Matsumoto, Y Takeuchi, H Kudo, N Ohnishi, Unsupervised texture segmentation via wavelet-based locally orderless images (WLOIs) and SOM, in *Computer Graphics and Imaging* (IASTED/ACTA Press, Calgary, 2003), pp. 279–284
6. V Wu, R Manmatha, EM Riseman, Textfinder: an automatic system to detect and recognize text in images. IEEE Trans. Pattern Anal. Mach. Intell. **21**(11), 1224–1229 (1999)
7. K Yabu, T Ifukube, S Aomura, A basic design of wearable speech synthesizer for voice disorders [Japanese]. EIC Technical report (Institute Electron. Inf. Commun. Eng). **105**(686), 59–64 (2006)
8. ST Canale, WC Campbell, *Campbell's Operative Orthopaedics*, vol. 4 (Mosby-Year Book, St. Louis, 2002)
9. H Matsumasa, T Takiguchi, Y Ariki, I Li, T Nakabayachi, Integration of metamodel and acoustic model for dysarthric speech recognition. J. Multimedia. **4**(4), 254–261 (2009)
10. C Miyamoto, Y Komai, T Takiguchi, Y Ariki, I Li, Multimodal speech recognition of a person with articulation disorders using AAM and MAF, in *IEEE International Workshop on Multimedia Signal Processing (MMSP'10),* St. Malo, 4–6 Oct 2010 (IEEE, Piscataway, 2010), pp. 517–520
11. A Maier, T Haderlein, F Stelzle, E Noth, E Nkenke, F Rosanowski, A Schutzenberger, M Schuster, Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. EURASIP J. Audio Speech Music Process. **2010**, 926951 (2010)
12. D Lee, HS Seung, Algorithms for non-negative matrix factorization, in *Advances in Neural Information Processing 13 (NIPS 2000)* (MIT Press, Massachusetts, 2001), pp. 556–562
13. 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)
14. JF Gemmeke, T Virtanen, Noise robust exemplar-based connected digit recognition, in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP),* Dallas, 14–19 March 2010 (IEEE, Piscataway, 2010), pp. 4546–4549
15. MN Schmidt, RK Olsson, Single-channel speech separation using sparse non-negative matrix factorization, in *Interspeech 2006 - ICSLP , Ninth International Conference on Spoken Language Processing,* Pittsburgh, 17–21 Sept 2006 (Curran Associates, Inc., New York, 2006), pp. 2614–2617
16. 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)
17. Y Iwami, T Toda, H Saruwatari, K Shikano, GMM-based voice conversion applied to emotional speech synthesis. IEEE Trans. Speech Audio Process. **7**, 2401–2404 (1999)

18. R Aihara, R Takashima, T Takiguchi, Y Ariki, GMM-Based emotional voice conversion using spectrum and prosody features. Am. J. Signal Process. **2**(5), 135–138 (2012)
19. Y Stylianou, O Cappe, E Moulines, Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. **6**(2), 131–142 (1998)
20. E Helander, T Virtanen, J Nurminen, M Gabbouj, Voice conversion using partial least squares regression. IEEE Trans. Audio Speech Lang. Process. **18**(5), 912–921 (2010)
21. CH Lee, CH Wu, Map-based adaptation for speech conversion using adaptation data selection and non-parallel training, in *Interspeech 2006 - ICSLP , Ninth International Conference on Spoken Language Processing,* Pittsburgh, 17–21 Sept 2006 (Curran Associates, Inc., New York, 2006), pp. 2254–2257
22. T Toda, Y Ohtani, K Shikano, Eigenvoice conversion based on Gaussian mixture model, in *Interspeech 2006 - ICSLP , Ninth International Conference on Spoken Language Processing,* Pittsburgh, 17–21 Sept 2006 (Curran Associates, Inc., New York, 2006), pp. 2446–2449
23. K Nakamura, T Toda, H Saruwatari, K Shikano, Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun. **54**(1), 134–146 (2012)
24. K Nakamura, T Toda, H Saruwatari, K Shikano, Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech, in *Interspeech 2006 - ICSLP, Ninth International Conference on Spoken Language Processing,* Pittsburgh, 17–21 Sept 2006 (Curran Associates, Inc., New York, 2006), pp. 1395–1398
25. C Veaux, J Yamagishi, S King, Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders, in *13th Annual Conference of the International Speech Communication Association 2012 (INTERSPEECH 2012),* Portland, 9–13 September 2012 (Curran Associates, Inc. New York, 2012), pp. 966–969
26. H Kawahara, I Masuda-Katsuse, A Cheveigne, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. **27**(3–4), 187–207 (1999)
27. JF Gemmeke, T Viratnen, A Hurmalainen, Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2067–2080 (2011)
28. 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 Commun. **9**, 357–363 (1990)
29. J Kominek, T Schultz, AW Black, Synthesizer voice quality of nwe languages calibrated with mean mel cepstral distortion, in *The International Workshop on Spoken Language Technology for Under-Resourced Languages (SLTU)* (Hanoi University of Technology, Hanoi, 5–7 May 2008)
30. International Telecommunication Union, *ITU-T Recommendation P.800-P.899: Methods for Objective and Subjective Assessment of Quality* (ITU, Geneva, 2003)

doi:10.1186/1687-4722-2014-5
Cite this article as: Aihara *et al.*: **A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:5.