INDIVIDUALITY-PRESERVING VOICE CONVERSION FOR ARTICULATION DISORDERS BASED ON NON-NEGATIVE MATRIX FACTORIZATION

Ryo AIHARA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI

Graduate School of System Informatics, Kobe University
1-1, Rokikodai, Nada, Kobe, 6578501, Japan

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 Non-negative Matrix Factorization (NMF) is applied to a voice with an articulation disorder. To preserve the speaker's individuality, we used a combined dictionary that is constructed from the source speaker's vowels and target speaker's consonants. Experimental results indicate that the performance of NMF-based VC is considerably better than conventional GMM-based VC.

Index Terms— Voice Conversion, NMF, Articulation Disorders, Voice Reconstruction, Assistive Technologies

1. INTRODUCTION

There are 34,000 people with speech impediments associated with an articulation disorder in Japan alone. One of the causes of speech impediments is cerebral palsy. Cerebral palsy results from 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 a mixture of these types [1].

In this study, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Athetoid symptoms develop in about 10-15% 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, in cases where movements are related to speaking, 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 great need for voice systems for them.

In this paper, we propose a voice conversion (VC) method for articulation disorders. The utterance of a person with an articulation disorder is difficult to understand for people who have not communicated with them. In recent years, people with an articulation disorder may use slideshows and a previously synthesized voice when they give a lecture. Veaux et al. used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders [2]. However, in case of people with an articulation disorder resulting from athetoid cerebral palsy, because their movement is restricted by their athetoid symptoms, to make slides or a synthesized 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 individuality. Many statistical approaches to VC have been studied and applied to various tasks, such as speaker conversion [3], emotion conversion [4, 5], speaking assistance [6], and so on. However, a speech conversion method for people with articulation disorders resulting from athetoid cerebral palsy has not been successfully developed.

A GMM-based approach is widely used for VC because of its flexibility and good performance [7]. 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. 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-disordered voice may be converted into a well-ordered voice. However, because the GMM-based approach has been developed for speaker conversion [3], the source speaker's voice individuality is also converted into the target speaker's individuality.

In the research discussed in this paper, we conducted VC for articulation disorders using Non-negative Matrix Factorization (NMF) [8]. NMF is a matrix decomposition method with non-negativity constraint. In the field of speech processing, NMF is a well-known approach for source separation and speech enhancement [9]. In these approaches, the observed vector is represented by a linear combination of a small number of elementary vectors, referred 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",

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

where $\mathbf{x}_l$ is the l-th frame of the observation, $\mathbf{a}_j$ and $h_{j,l}$ are the j-th 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. By 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 HMMs for speech recognition [10].

In our study, we adopt the supervised NMF approach [11], 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-disordered basis and its weights using NMF. By replacing an articulation-disordered basis 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.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Hence, by replacing the articulation-disordered basis of consonants only, a voice with an articulation disorder is converted into a clear voice that preserve the individuality of speaker's voice.

The rest of this paper is organized as follows: In Section 2, NMF-based VC is described, the experimental data is evaluated in Section 3, and the final section is devoted to our conclusions.

2. VOICE CONVERSION BASED ON NMF

2.1. Exemplar-based voice conversion

Fig. 1 shows the basic approach of our exemplar-based VC using NMF. D, d, L, and J represent the number of dimensions of source features, dimensions of target features, frames of dictionary, and basis of dictionary, respectively. 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 source features. Source features are constructed from an articulation-disordered spectrum and its segment features. The other dictionary is a target dictionary, which is constructed from target features. Target features are mainly constructed from a well-ordered spectrum. These two dictionaries consist of the same words and are aligned with dynamic time warping. Hence, these dictionaries have the same number of bases.

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

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

Fig. 2 shows an example of the activity matrices estimated from a word "ikioi" ("vigor" in English). One is uttered by a person with an articulation disorder, and the other is uttered by a physically unimpaired person. To show an intelligible example, each dictionary was structured from just the one word "ikioi" and aligned with dynamic time warping (DTW). As shown in Fig. 2, these activities have high energies at similar elements. For this reason, 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. Therefore, the target speech can be constructed using the target dictionary and the activity of the source signal as shown in Fig. 1.

Spectral envelopes extracted by STRAIGHT analysis [12] 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.

2.2. 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 Fig. 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 lining up each of the features extracted from parallel utterances.

Fig. 3 shows how to preserve a source speaker's voice individuality in our VC. Vowels of voice strongly imply a speaker's individuality. On the other hand, consonants of people with articulation disorders are often unstable. 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.

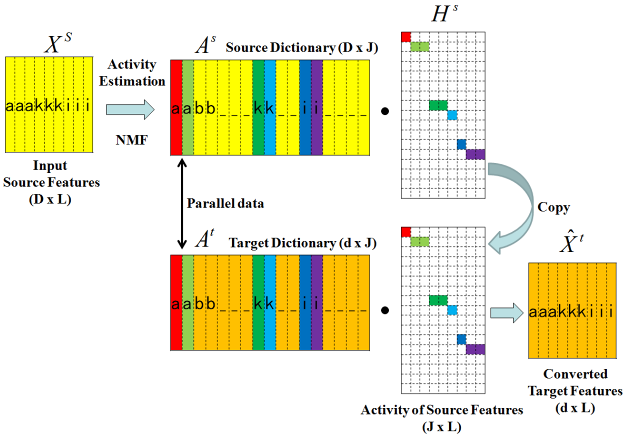

Fig. 1. Basic approach of NMF-based voice conversion

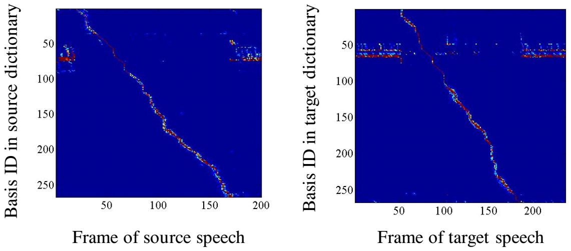

Fig. 2. Activity matrices for the articulation-disordered utterance (left) and well-ordered utterance (right)

2.3. Estimation of Activity

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

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

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

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

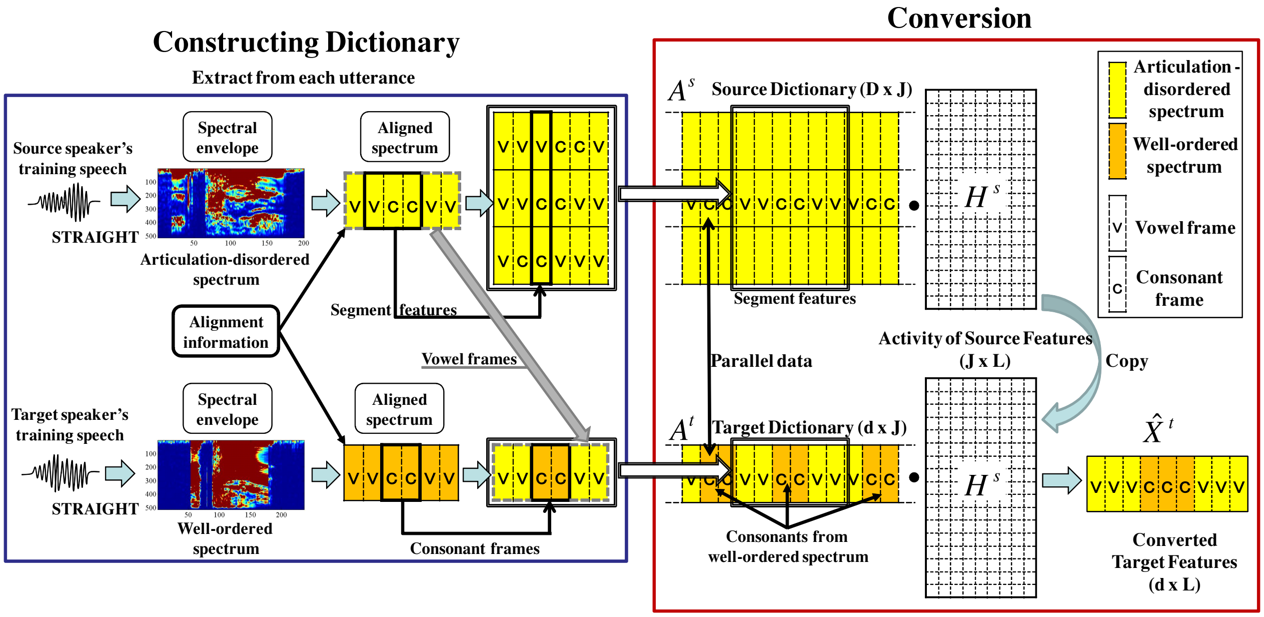

Fig. 3. Individuality-preserving voice conversion

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

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

$\mathbf{1}$ is an all-one matrix. 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 $\lambda^T = [\lambda_1 \dots \lambda_J]$. In this paper, all elements in λ were set to 0.1. $\mathbf{H}^s$ minimizing (4) is estimated iteratively applying the following update rule [8]:

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

with $.*$ and $./$ denoting element-wise multiplication and division, respectively. To increase the sparseness of $\mathbf{H}^s$, elements of $\mathbf{H}^s$, which are less than threshold, are rounded to zero.

By using the activity and the target dictionary, the converted spectral features are constructed.

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

3. EXPERIMENTAL RESULTS

3.1. 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. The speech signals were sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. A physically unimpaired Japanese male in the ATR Japanese speech database, was chosen as a target speaker. Two hundred sixteen utterances were used for training, and the other 216 utterances were used for the test. The number of dimensions of source and target features are, 2565 and 513. The Mel-cepstral coefficient, which is converted from the STRAIGHT spectrum, is used for DP-matching in order to align the temporal fluctuation.

We compared our NMF-based VC to conventional GMM-based VC. In GMM-based VC, the 1st through 24th cepstrum coefficients extracted by STRAIGHT are used as source and target features.

3.2. Subjective Evaluation

We performed a MOS (Mean Opinion Score) test on 4 subjective evaluations. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). On "listening intelligibility" and "clarity of consonants" evaluation, 50 words were converted using NMF-based VC and GMM-based VC, where the text was given and the subjects were asked about the listening intelligibility and the clarity of consonants in the articulation-disordered voice, the NMF-based converted voice, and the GMM-based converted voice. Each voice uttered by a physically unimpaired person was presented as a reference of 5 points on the MOS test.

On the "similarity" and "naturalness" evaluation, 23 words, which are difficult for a person with articulation disorder to utter, were evaluated, where the source text was given, and subjects were asked about the similarity to the source speaker and naturalness of the words. A total of 5 Japanese speakers performed the test using headphones.

3.3. Results and Discussion

Fig. 4 shows the results on the MOS test on the listening intelligibility and clarity of consonants. NMF-based VC can improve the listening intelligibility and clarity of consonants. On the other hand, GMM-based conversion can improve clarity of consonants but it deteriorates the listening intelligibility. This is because GMM-based conversion creates its conversion noise. NMF-based conversion also creates some conversion noise, but it is less than that created by GMM-based conversion.

Fig. 5 shows the results of the MOS test on the similarity to the source speaker and naturalness of the converted voice. NMF-based VC got a higher score than GMM-based conversion on similarity because NMF-based conversion used a combined dictionary. NMF-based VC also got a higher score than GMM-based conversion on naturalness although NMF-based conversion mixed the source speaker's vowels and target speaker's consonants.

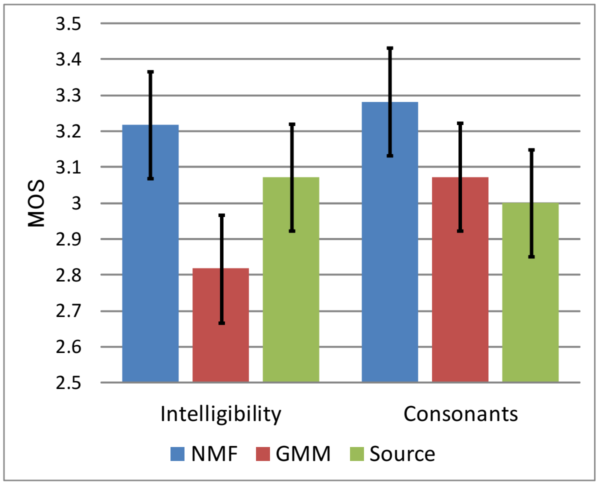

Fig. 4. Results of MOS test on listening intelligibility and clarity of consonants

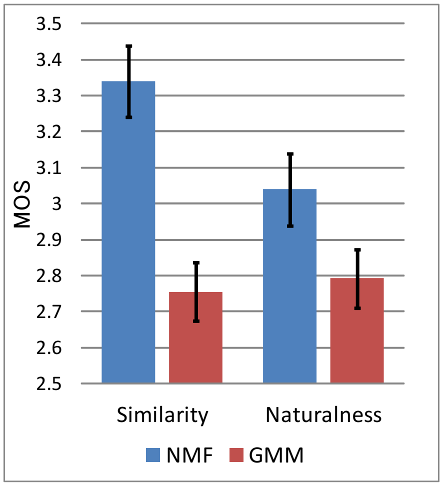

Fig. 5. Results of MOS test on the similarity to the source speaker and naturalness

4. CONCLUSIONS

We proposed a spectral conversion method based on NMF for a voice with an articulation disorder. Experimental results demonstrated that our VC method can improve the listening intelligibility and clarity of consonants of the words uttered by a person with an articulation disorder. Moreover, compared to conventional GMM-based VC, NMF-based VC can preserve the individuality of the source speaker's voice and the naturalness of the voice. 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.

5. ACKNOWLEDGMENT

This research was supported in part by MIC SCOPE.

6. REFERENCES

[1] S. T. Canale and W. C. Campbell, "Campbell's operative orthopaedics," Tech. Rep., Mosby-Year Book, 2002.

[2] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," *Proc. Interspeech*, 2012.

[3] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process., Vol. 15, No. 8, pp.2222-2235*, 2007.

[4] Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *IEEE Trans. Seech and Audio Proc., Vol. 7, pp. 2401-2404*, 1999.

[5] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing, Vol. 2 No. 5*, 2012.

[6] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication, Vol. 54, No. 1, pp. 134-146*, 2012.

[7] Y. Stylianou, O. Cappe, and E. Moilines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing, Vol. 6,No. 2, pp. 131-142*, 1998.

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

[9] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process., Vol. 15, No. 3, pp. 1066-1074*, 2007.

[10] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," *ICASSP, pp. 4546-4549*, 2010.

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

[12] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication, Vol. 27, No. 3-4*, 1999.