Individuality-Preserving Voice Conversion for Articulation Disorders Using Locality-Constrained NMF

Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI

Graduate School of System Informatics, Kobe University, Japan

aihara@me.cs.scitec.kobe-u.ac.jp

takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

We present in this paper a voice conversion (VC) method for a person with an articulation disorder resulting from athetoid cerebral palsy. The movements of such speakers are 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. In order to preserve the speaker's individuality, we use a combined dictionary that was constructed from the source speaker's vowels and target speaker's consonants. Also, in order to avoid an unclear converted voice, which is constructed using the combined dictionary, we used locality-constrained NMF. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method.

Index Terms: Voice Conversion, NMF, Articulation Disorders, Assistive Technologies

1. Introduction

In this study, we propose assistive technology for people with speech impediments. There are 34,000 people with speech impediments associated with an articulation disorder in Japan alone. Articulation disorders are classified into three types. Functional articulation disorders exist in the absence of any apparent cause and are related to deficiencies in the relatively peripheral motor processes. Organic articulation disorders are articulation problems that are associated with structural abnormalities and known impairments, such as cleft lip and palate, tongue tie, hearing impairment, etc. Motor speech disorders involve problems with strength and control of the speech musculature. We propose a voice conversion system, which converts an articulation-disordered voice into a non-disordered voice, for people with motor speech disorders.

Cerebral palsy is one of the typical causes of motor speech disorders. About two babies in 1,000 are born with cerebral palsy [1]. 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 [2].

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 10-20% of cerebral palsy sufferers [1]. In the case of a person with this type of articulation disorder, his/her movements are sometimes more unstable than usual. Because of this symptom, their utterances (especially their consonants) are often unstable or unclear. Most people suffering from athetoid cerebral palsy cannot communicate by sign language, writing or voice synthesizer [3, 4, 5] because athetoid symptoms also restrict the movement of the sufferer's arms and legs. For this reason, there is a great need for a voice conversion (VC) system for such people.

Automatic speech recognition system for people with articulation disorders resulting from athetoid cerebral palsy has been studied. Matsumasa et al. [6] proposed robust feature extraction based on PCA (Principal Component Analysis) with more stable utterance data instead of DCT. Miyamoto et al. [7] 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. The recognition rate for people with articulation disorders using a speaker-independent model trained by non-disordered speech is 3.5%. This result implies that the speech of a person with an articulation disorder is difficult to understand for people who have not communicated with them before.

A GMM-based approach is widely used for VC because of its flexibility and good performance [8]. This approach has been applied to various tasks, such as speaker conversion [9], emotion conversion [10, 11], and so on. In the field of assistive technology, Nakamura et al. [12] proposed a GMM-based speaking aid system for electrolaryngeal speech. In this approach, the conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using Minimum Mean-Square Error (MMSE) 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 non-disordered voice. However, because the GMM-based approach has been developed mainly for speaker conversion [9], the source speaker's voice individuality is also converted into the target speaker's individuality.

In this paper, we propose a VC method for articulation disorders. There are two main benefits to our VC method. 1) We convert the speaker's voice into a non-disordered voice, thus preserving their voice individuality. People with articulation disorders wish to communicate by their own voice if they can therefore, this is important for VC as assistive technology. 2) Our method outputs a natural-sounding voice. Because our VC is exemplar-based and there is no statistical model, we can create a natural sounding voice.

In the research discussed in this paper, we conducted VC for articulation disorders using Non-negative Matrix Factorization (NMF) [13]. NMF is a well-known approach for source

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 3–8,
Grenoble, France, 21–22 August, 2013.

separation and speech enhancement. In these approaches, the observed signal is represented by a linear combination of a small number of elementary vectors, referred to as the basis, and its weights. In some approaches for source separation, the bases are grouped for each source, and the mixed signals are expressed with a sparse representation of these bases. Gemmeke et al. proposes an exemplar-based method for noise robust 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 non-disordered articulation. The parallel exemplars (called the 'dictionary' in this paper), which consist of articulation-disordered exemplars and a non-disordered exemplars, are extracted from the parallel data. An input spectrum with an articulation disorder is represented by a linear combination of articulation-disordered exemplars using NMF. By replacing an articulation-disordered basis with a non-disordered basis, the original speech spectrum is replaced with a non-disordered spectrum.

In the voice of a person with an articulation disorder, their consonants are often unstable and that makes their voices unclear. Their vowels are relatively-stable compared to their consonants. Hence, by replacing the articulation-disordered basis of consonants only, a voice with an articulation disorder is converted into a non-disordered voice that preserves the individuality of the speaker's voice. In order to avoid a mixture of the source and target spectra in a converted phoneme which is constructed using the combined dictionary, we adopted locality-constraint to the supervised NMF.

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. Basic Approach of Exemplar-Based Voice Conversion

In the exemplar-based approach, the observed signal is represented by a linear combination of a small number of bases.

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

$\mathbf{x}_l$ is the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ are the j-th basis and the weight, respectively. $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ are the collection of the bases and the stack of weights. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination of a small number of bases that have non-zero weights. In this paper, each basis denotes the exemplar of the spectrum, and the collection of exemplar $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are called 'dictionary' and 'activity', respectively.

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 the dictionary, and basis of the dictionary, respectively. 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 (DTW). 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 bases from the source dictionary A^s by NMF. The weights of the bases are estimated as an activity 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 $\hat{X}^t$ which are represented by a linear combination of bases from the target dictionary. Because the source and target dictionary are parallel phonemically, the bases 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 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 [16] 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.

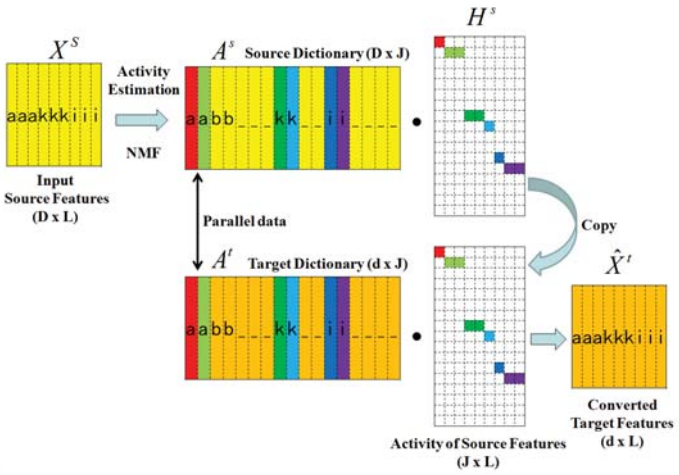


Figure 1: Basic approach of NMF-based voice conversion

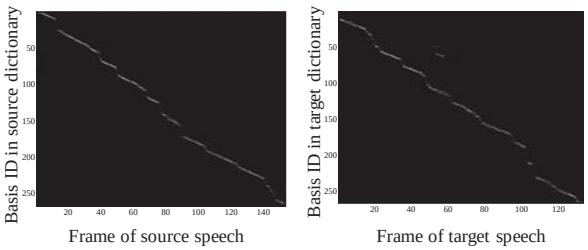


Figure 2: Activity matrices for the articulation-disordered utterance (left) and well-ordered utterance (right)

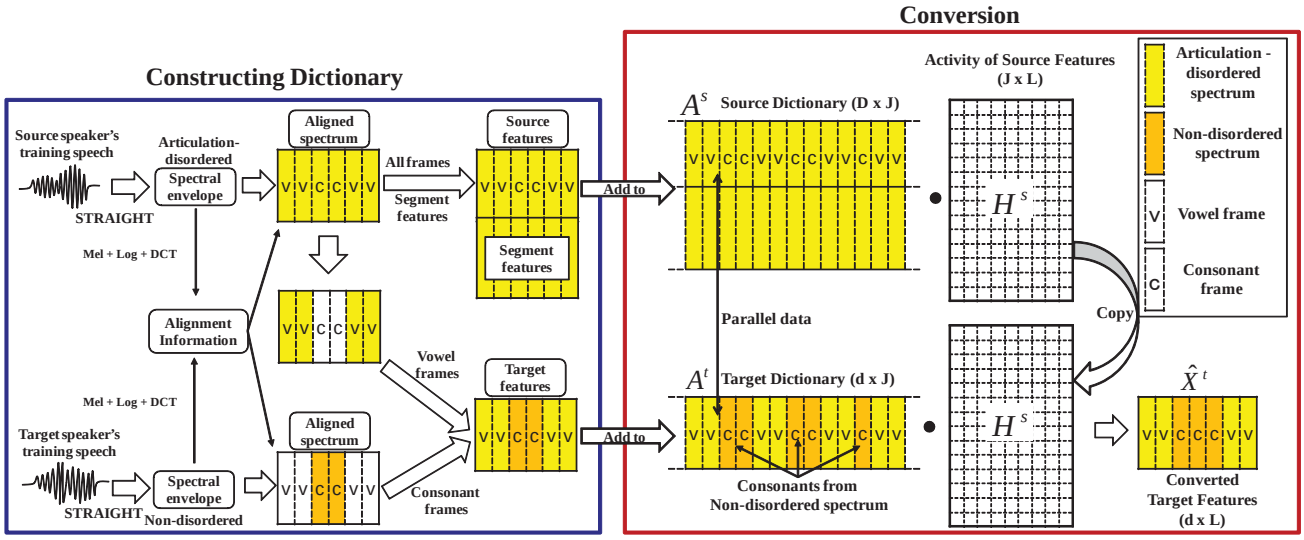


Figure 3: Individuality-preserving voice conversion

2.2. Constructing Dictionary to Preserve Individuality

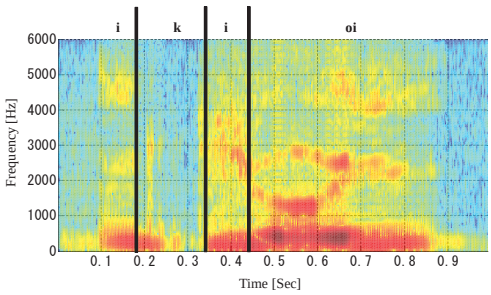
In order to make a parallel dictionary, some pairs of parallel utterances are needed, where each pair consists of the same text. One is spoken by a person with an articulation disorder (source speaker), and the other is spoken by a physically unimpaired person (target speaker). The left side of Fig. 3 shows the process for constructing a parallel dictionary. STRAIGHT spectrum is extracted from parallel utterances. The extracted STRAIGHT spectra are phonemically aligned with DTW. The Mel-cepstral coefficient, which is converted from the STRAIGHT spectrum, is used to align. In order to estimate the activities of the source features precisely, segment features of source 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.

The right side of Fig. 3 shows how to preserve a source speaker's voice individuality in our VC method. Fig. 4 shows examples of the spectrogram for the word "ikioi" ("vigor" in English) of a person with an articulation disorder and a physically unimpaired person. The vowels of a person's voice strongly imply a speaker's individuality. On the other hand, the consonants of people with articulation disorders are often unstable. In Fig. 4, the area labeled "k" in the articulation-disordered spectrum is not clear, compared to that of the same region spoken by a physically unimpaired person. Therefore, by combining the source speaker's vowels and target speaker's consonants in the target dictionary, the individuality of the source speaker's voice can be preserved.

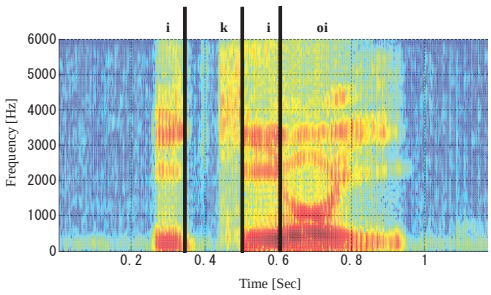
2.3. Estimation of Activity with Locality Constraint

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.

$$\begin{aligned} \mathbf{x}_l &= \mathbf{x}_l^s \\ &\approx \sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s \end{aligned} \tag{2}$$



(a) Spoken by a person with an articulation disorder



(b) Spoken by a physically unimpaired person

Figure 4: Examples of source and target spectrogram //i k i oi

$\mathbf{x}_l^s \in \mathbf{X}^s$ is the magnitude spectrum of the source signal.

Instead of using all bases, locality constraint is introduced.

$$\mathbf{\Delta}_{j,l} = \sqrt{(x_l^s - a_j^s)^2} \tag{3}$$

$\mathbf{\Delta}_l$ is a distance vector between $\mathbf{x}_l^s$ and $\mathbf{a}^s$. N nearest bases are chosen from all the bases.

$$\begin{aligned} \mathbf{S}_l^s &= \mathbf{nbest}_{\mathbf{\Delta}_l}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_J) \\ &= \mathbf{nbest}_{\mathbf{\Delta}_l}(\mathbf{A}) \end{aligned} \tag{4}$$

$\mathbf{S}_l^s$ is a set of nearest bases of $\mathbf{x}_l^s$. The number of basis is defined

by N. Eq. (2) can be written as follows:

$$\begin{aligned} \mathbf{x}_l &= \mathbf{x}_l^s \\ &\approx \sum_{j=1}^{N} \mathbf{S}_{l,j}^s h_{j,l}^s \\ &= \mathbf{S}_l \mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{5}$$

$$\mathbf{X}^s \approx \mathbf{S}^s \mathbf{H}^s \quad s.t. \quad \mathbf{H}^s \geq 0 \tag{6}$$

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

$$d(\mathbf{X}^s, \mathbf{S}^s \mathbf{H}^s) + ||(\lambda \mathbf{1}^{1\times L}) . * \mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geq 0 \tag{7}$$

$\mathbf{1}$ is an all-one matrix. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}^s$ and $\mathbf{S}^s\mathbf{H}^s$. 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 $\lambda^T = [\lambda_1 \ldots \lambda_J]$. In this paper, all elements in λ were set to 1. $\mathbf{H}^s$ minimizing Eq. (7) is estimated iteratively applying the following update rule [13]:

$$\begin{aligned} \mathbf{H}_{n+1}^s &= \mathbf{H}_n^s . * (\mathbf{S}^{sT}(\mathbf{X}^s ./ (\mathbf{S}^s \mathbf{H}_n^s))) \\ &\quad ./ (\mathbf{S}^{sT} \mathbf{1}^{\mathbf{D}\times L} + \lambda \mathbf{1}^{1\times L}) \end{aligned} \tag{8}$$

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 set of target basis which is parallel to $\mathbf{S}^s$, the converted spectral features are constructed.

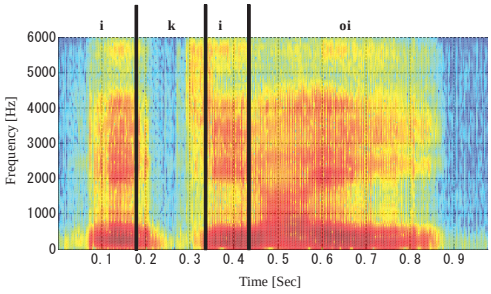
$$\hat{\mathbf{X}}^t = (\mathbf{S}^t \mathbf{H}^s) \tag{9}$$

3. Experimental Results

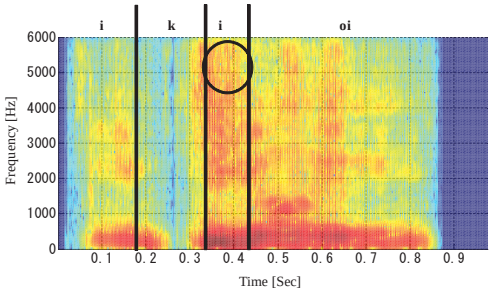
3.1. Experimental Conditions

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 [17]. The speech signals were sampled at 12 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 numbers of dimensions of source and target features are, 2,565 and 513. The number of bases of source and target dictionary is 64,467. We chose 10,000 nearest bases from dictionary by locality constraint.

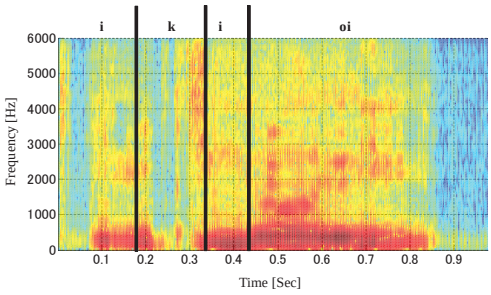
We compared our NMF-based VC to conventional GMM-based VC. In GMM-based VC, the 1st through 24th cepstrum coefficients extracted by STRAIGHT were used as source and target features.



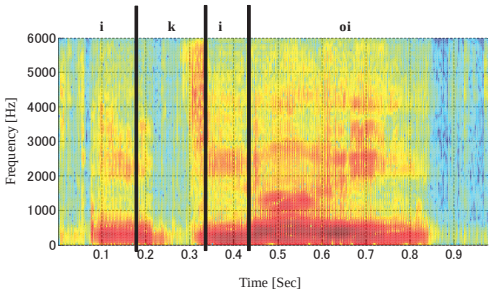
(a) Converted by GMM-based VC



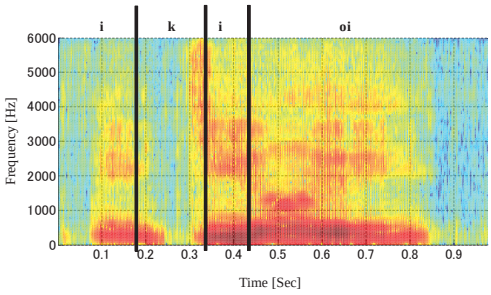
(b) Converted by NMF-based VC without locality



(c) Converted by NMF-based VC with 100 nearest bases



(d) Converted by NMF-based VC with 1,000 nearest bases



(e) Converted by NMF-based VC with 10,000 nearest bases

Figure 5: Examples of converted spectrograms for "i k i oi"

3.2. Subjective Evaluation

We conducted subjective evaluation on 3 topics. A total of 10 Japanese speakers took part in the test using headphones. For the "listening intelligibility" evaluation, we performed a MOS (Mean Opinion Score) test [18]. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Thirty-eight words, which are difficult for a person with an articulation disorder to utter, were evaluated. The subjects were asked about the listening intelligibility 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.

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.

3.3. Results and Discussion

Fig. 5 shows examples of converted spectrograms. Using GMM-based conversion, the area labeled "oi" becomes unclear compared to NMF-based conversion. This might be because unexpected mapping during the GMM-based VC degraded the conversion performance. Because NMF-based VC converts consonants only, the same area is relatively clear and similar to the labeled "oi" area in Fig. 4(a). In the spectrogram converted by NMF-based VC without locality, there are some misconversions in the black circled area. This is because there is some mixing of the vowel and consonant spectra. By using local constrained NMF, such misconversions are eliminated. Also, in comparison between (d) and (e) in Fig. 5, the converted voice using 10,000 nearest bases is more clear than that using 1,000 nearest bases, especially the areas labeled "i" and "oi". For this reason, locality-constraint is useful to the combined dictionary, however, using too few bases degrades conversion performance.

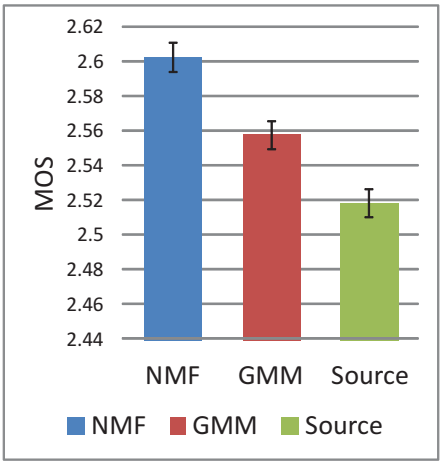


Figure 6: Results of MOS test on listening intelligibility

Fig. 6 shows the results of the MOS test for listening intelligibility. The error bars show a 95% confidence score. As shown in Fig. 6, NMF-based VC and GMM-based VC can improve listening intelligibility. NMF-based VC obtained a higher score than GMM-based VC. 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.

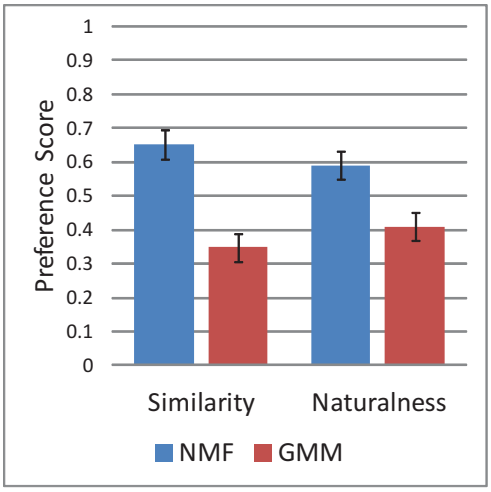


Figure 7: Preference scores for the similarity to the source speaker and naturalness

Fig. 7 shows the preference score on the similarity to the source speaker and naturalness of the converted voice. The error bars show a 95% confidence score. 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.

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 of 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. References

[1] M. V. Hollegaard, K. Skogstrand, P. Thorsen, B. Norgaard-Pedersen, D. M. Hougaard, and J. Grove, "Joint analysis of SNPs and proteins identifies regulatory IL18 gene variations decreasing the chance of spastic cerebral palsy," *Human Mutation, Vol. 34, pp. 143-148*, 2013.

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

[3] 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.

[4] A. Kain and M. Macon, "Personalizing a speech synthesizer by voice adaption," *Proceedings of the Third ESCA/COCOSDA International Speech Synthesis Workshop, pp.225-230.*, 1998.

[5] C. Jreige, R. Patel, and H. T. Bunnell, "VocaliD: Personalizing text-to-speech synthesis for individuals," *in Proceedings of ASSETS'09, pp.259-260*, 2009.

[6] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "Integration of metamodel and acoustic model for dysarthric speech recognition," *Journal of Multimedia, Volume 4, Issue 4, pp. 254-261*, 2009.

[7] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," *IEEE International Workshop on Multimedia Signal Processing (MMSP'10), pp. 517-520*, 2010.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

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

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

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

[17] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication, Vol. 9, pp. 357-363*, 1990.

[18] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, 2003.