Individuality-preserving Voice Conversion for Articulation Disorders Using Dictionary Selective Non-negative Matrix Factorization

Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki
Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 6578501, 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. However, this exemplar-based approach needs to hold all the training exemplars (frames), and it may cause mismatching of phonemes between input signals and selected exemplars. In this paper, in order to reduce the mismatching of phoneme alignment, we propose a phoneme-categorized sub-dictionary and a dictionary selection method using NMF. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based and conventional NMF-based method.

1 Introduction

In this study, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. About two babies in 1,000 are born with cerebral palsy (Hollegaard et al., 2013). Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Cerebral palsy is classified into the following types: 1)spastic, 2)athetoid, 3)ataxic, 4)atonic, 5)rigid, and a mixture of these types (Canale and Campbell, 2002).

Athetoid symptoms develop in about 10-15% of cerebral palsy sufferers (Hollegaard et al., 2013). 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 great need for voice systems for them.

In this paper, we propose a voice conversion (VC) method for articulation disorders. Regarding speech recognition for articulation disorders, the recognition rate using a speaker-independent model which is trained by well-ordered speech, is 3.5% (Matsumasa et al., 2009). 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 slideshows 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. Rudzicz et al. (Rudzicz, 2011; Rudzicz, 2014) proposed speech adjustment method for people with articulation disorders based on the observations from the database.

In (Aihara et al., 2014), we proposed individuality-preserving VC for articulation disorders. In our VC, source exemplars and target exemplars are extracted from the parallel

Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pages 29–37,
Baltimore, Maryland USA, August 26 2014.

training data, having the same texts uttered by the source and target speakers. The input source signal is expressed with a sparse representation of the source exemplars using Non-negative Matrix Factorization (NMF). By replacing a source speaker's exemplar with a target speaker's exemplar, the original speech spectrum is replaced with the target speaker's spectrum. People with articulation disorders wish to communicate by their own voice if they can; therefore, we proposed a combined-dictionary, which consists of a source speaker's vowels and target speaker's well-ordered consonants. 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 this paper, we propose advanced individuality-preserving VC using NMF. In order to avoid a mixture of the source and target spectra in a converted phoneme, we applied a phoneme-categorized dictionary and a dictionary selection method to our VC using NMF. In conventional NMF-based VC, the number of dictionary frames becomes large because the dictionary holds all the training exemplar frames. Therefore, it may cause phoneme mismatching between input signals and selected exemplars and some frames of converted spectra might be mixed with the source and target spectra. In this paper, a training exemplar is divided into a phoneme-categorized sub-dictionary, and an input signal is converted by using the selected sub-dictionary. The effectiveness of this method was confirmed by comparing it with a conventional NMF-based method and a conventional Gaussian Mixture Model (GMM)-based method.

The rest of this paper is organized as follows: In Section 2, related works are introduced. In Section 3, the basic idea of NMF-based VC is described. In Section 4, our proposed method is described. In Section 5, the experimental data are evaluated, and the final section is devoted to our conclusions.

2 Related Works

Voice conversion (VC) is a technique for converting specific information in speech while maintaining the other information in the utterance. One of the most popular VC applications is speaker conversion (Stylianou et al., 1998). In speaker conversion, a source speaker's voice individuality is changed to a specified target speaker's so that the input utterance sounds as though a specified target speaker had spoken it.

There have also been studies on several tasks that make use of VC. Emotion conversion is a technique for changing emotional information in input speech while maintaining linguistic information and speaker individuality (Veaux and Robet, 2011). In recent years, VC has been used for automatic speech recognition (ASR) or speaker adaptation in text-to-speech (TTS) systems (Kain and Macon, 1998). These studies show the varied uses of VC.

Many statistical approaches to VC have been studied (Valbret et al., 1992). Among these approaches, the Gaussian mixture model (GMM)-based mapping approach (Stylianou et al., 1998) is widely used. 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) on a parallel training set. A number of improvements in this approach have been proposed. Toda et al. (Toda et al., 2007) introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. (Helander et al., 2010) proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem associated with standard multivariate regression. There have also been approaches that do not require parallel data that make use of GMM adaptation techniques (Lee and Wu, 2006) or eigen-voice GMM (EV-GMM) (Toda et al., 2006).

In the field of assistive technology, Nakamura et al. (Nakamura et al., 2012; Nakamura et al., 2006) 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 however, because these statistical approaches are mainly proposed for speaker conversion, the target speaker's individuality will be

changed to the source speaker's individuality. People with articulation disorders wish to communicate by their own voice if they can and there is a needs for individuality-preserving VC.

Text-to-speech synthesis (TTS) is a famous voice application that is widely researched. Veaux et al. (Veaux et al., 2012) used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from Amyotrophic Lateral Sclerosis (ALS). Yamagishi et al. (Yamagishi et al., 2013) proposed a project named "Voice Banking and Reconstruction". In that project, various types of voices are collected and they proposed TTS for ALS using that database. The difference between TTS and VC is that TTS needs text input to synthesize speech, whereas VC does not need text input. In the case of people with articulation disorders resulting from athetoid cerebral palsy, it is difficult for them to input text because of their athetoid symptoms.

Our proposed NMF-based VC (Takashima et al., 2012) is an exemplar-based method using sparse representation, which is different from the conventional statistical method. In recent years, approaches based on sparse representations have gained interest in a broad range of signal processing. In approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of bases. In some approaches for source separation, the atoms are grouped for each source, and the mixed signals are expressed with a sparse representation of these atoms. By using only the weights of the atoms related to the target signal, the target signal can be reconstructed. Gemmeke et al. (Gemmeke et al., 2011) also propose an exemplar-based method for noise-robust speech recognition. In that method, the observed speech is decomposed into the 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.

In (Takashima et al., 2012), we proposed noise-robust VC using NMF. The noise exemplars, which are extracted from the before- and after-utterance sections in an observed signal, are used as the noise-dictionary, and the VC process is combined with an NMF-based noise-reduction method. On the other hand, NMF is one of the clustering methods. In our exemplar-based VC, if the phoneme label of the source exemplar is given, we can discriminate the phoneme of the input signal by using NMF. In this paper, we proposed a dictionary selection method using this property of NMF.

3 Voice Conversion Based on Non-negative Matrix Factorization

3.1 Basic Idea

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

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

$\mathbf{x}_l$ represents the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ represent 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. 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 the 'dictionary' and 'activity', 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 that have non-zero weights. Eq. (1) is expressed as the inner product of two matrices using the collection of the frames or bases.

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \quad (2)$$

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \quad (3)$$

L represents the number of the frames.

Fig. 1 shows the basic approach of our exemplar-based VC, where D, L, and J represent the numbers of dimensions, frames, and bases, respectively. Our VC method needs two dictionaries that are phonemically parallel. $\mathbf{A}^s$ represents a source dictionary that consists of the source speaker's exemplars and $\mathbf{A}^t$ represents a target dictionary that consists of the target speaker's exemplars. These two dictionaries consist of the same words and are aligned with dynamic time warping (DTW) just as conventional GMM-based VC is. Hence, these dictionaries have the same number of bases.

This method assumes that when the source signal and the target signal (which are the same words but spoken by different speakers) are expressed with sparse representations of the source dictionary and the target dictionary, respectively, the ob-

tained activity matrices are approximately equivalent. Fig. 2 shows an example of the activity matrices estimated from a Japanese word "ikioi" ("vigor" in English), where one is uttered by a male, the other is uttered by a female, and each dictionary is structured from just one word "ikioi" as the simple example.

As shown in Fig. 2, these activities have high energies at similar elements. For this reason, we assume that when there are parallel dictionaries, the activity of the source features estimated with the source dictionary may be able to be substituted with that of the target features. Therefore, the target speech can be constructed using the target dictionary and the activity of the source signal as shown in Fig. 1. In this paper, we use Non-negative Matrix Factorization (NMF), which is a sparse coding method in order to estimate the activity matrix.

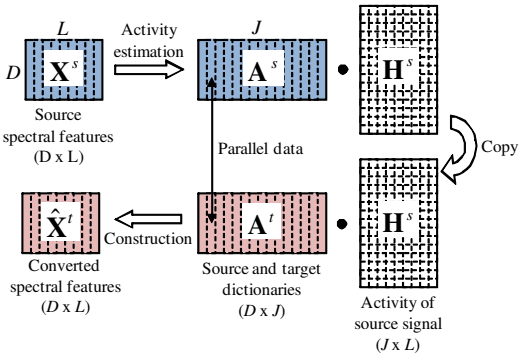


Figure 1: Basic approach of NMF-based voice conversion

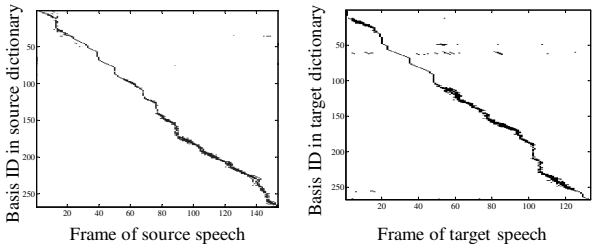


Figure 2: Activity matrices for parallel utterances

3.2 Individuality-preserving Voice Conversion Using Combined Dictionary

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). Spectrum envelopes, which are extracted from parallel utterances, are phonemically aligned by using DTW. 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.

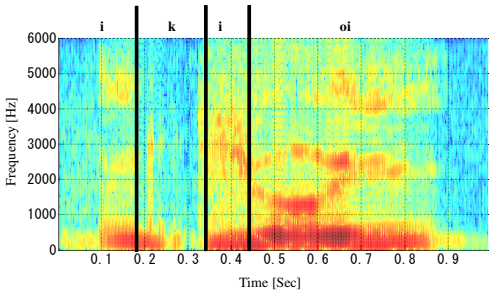
The vowels voiced by a speaker strongly indicate the speaker's individuality. On the other hand, consonants of people with articulation disorders are often unstable. Fig. 3(a) shows an example of the spectrogram for the word "ikioi" ("vigor" in English) of a person with an articulation disorder. The spectrogram of a physically unimpaired person speaking the same word is shown in Fig. 3(b). In Fig. 3(a), the area labeled "k" is not clear, compared to the same region in to Fig. 3(b). These figures indicate that consonants of people with articulation disorders are often unstable and this deteriorates their voice intelligibility. In order to preserve their voice individuality, we use a "combined-dictionary" that consists of a source speaker's vowels and target speaker's consonants.

We replace the target dictionary $\mathbf{A}^s$ in Fig. 1 with the "combined-dictionary". 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 bases 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 combined-dictionary in order to obtain converted spectral features $\hat{\mathbf{X}}^t$, which are represented by a linear combination of bases from the source speaker's vowels and target speaker's consonants. Because the source and target are parallel phonemically, the bases used in the converted features are phonemically the same as that of the source features.

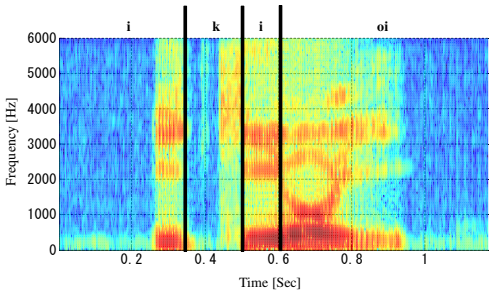
3.3 Problems

In the NMF-based approach described in Sec. 3.2, the parallel dictionary consists of the parallel training data themselves. Therefore, as the number of the bases in the dictionary increases, the input

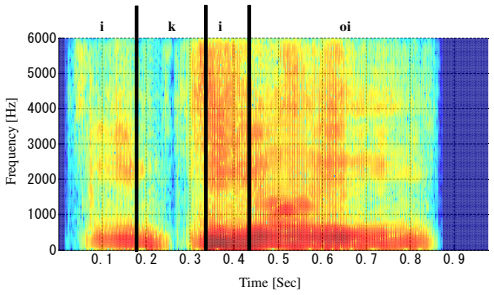
signal comes to be represented by a linear combination of a large number of bases rather than a small number of bases. When the number of bases that represent the input signal becomes large, the assumption of similarity between source and target activities may be weak due to the influence of the mismatch between the input signal and the selected bases. Moreover, in the case of a combined-dictionary, the input articulation-disordered spectrum may come to be represented by a combination of vowels and consonants. We assume that this problem degrades the performance of our exemplar-based VC. Hence, we use a phoneme-categorized sub-dictionary in place of the large dictionary in order to reduce the number of the bases that represent the input signal and avoid the mixture of vowels and consonants.



(a) Spoken by a person with an articulation disorder



(b) Spoken by a physically unimpaired person



(c) Converted by NMF-based VC

Figure 3: Examples of spectrogram //i k i oi

4 Non-negative Matrix Factorization Using a Phoneme-categorized Dictionary

4.1 Phoneme-categorized Dictionary

Fig. 4 shows how to construct the sub-dictionary. $\mathbf{A}^s$ and $\mathbf{A}^t$ imply the source and target dictionary which hold all the bases from training data. These dictionaries are divided into K dictionaries. In this paper, the dictionaries are divided into 10 categories according to the Japanese phoneme categories shown in Table 1.

In order to select the sub-dictionary, a "categorizing-dictionary", which consists of the representative vector from each sub-dictionary, is constructed. The representative vectors for each phoneme category consist of the mean vectors of the Gaussian Mixture Model (GMM).

$$p(\mathbf{x}_n^{(k)}) = \sum_{m=1}^{M_k} \alpha_m^{(k)} N(\mathbf{x}_n^{(k)}, \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}) \quad (4)$$

M_k, $\alpha_m^{(k)}$, $\boldsymbol{\mu}_m^{(k)}$ and $\boldsymbol{\Sigma}_m^{(k)}$ represent the number of the Gaussian mixture, the weights of mixture, mean and variance of the m-th mixture of the Gaussian, in the k-th sub-dictionary, respectively. Each parameter is estimated by using an EM algorithm.

The basis of the categorizing-dictionary, which corresponds to the k-th sub-dictionary $\boldsymbol{\Phi}_k^s$, is represented using the estimated phoneme GMM as follows:

$$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_1^{(k)}, \ldots, \boldsymbol{\mu}_{M_k}^{(k)}] \quad (5)$$

$$\boldsymbol{\Phi}_k^s = [\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{N_k}^{(k)}] \quad (6)$$

N_k represents the number of frames of the k-th sub-dictionary. The categorizing-dictionary $\boldsymbol{\Theta}$ is given as follows:

$$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K] \quad (7)$$

4.2 Dictionary Selection and Voice Conversion

Fig. 5 shows the flow of the dictionary selection and VC. The input spectral features $\mathbf{X}^s$ are represented by a linear combination of bases from the categorizing-dictionary $\boldsymbol{\Theta}$. The weights of the

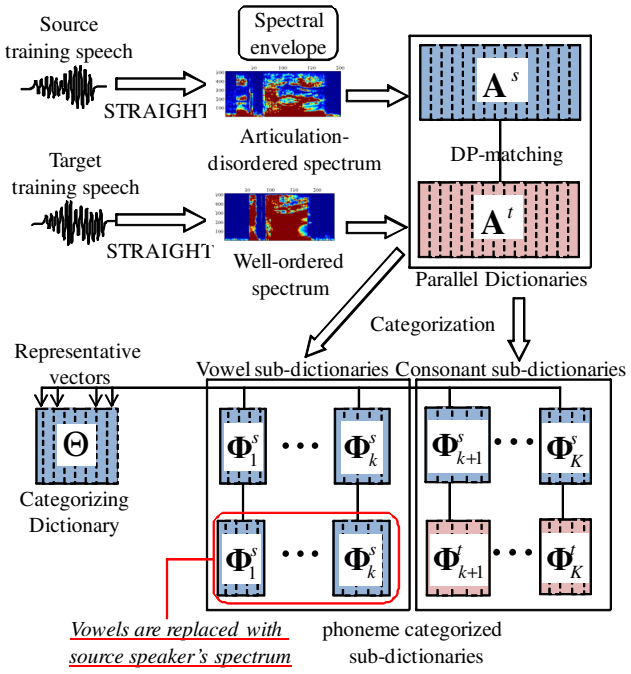


Figure 4: Making a sub-dictionary

bases are represented as activities $\mathbf{H}^s_{\mathbf{\Theta}}$.

$$\mathbf{X}^s \approx \mathbf{\Theta}\mathbf{H}^s_{\mathbf{\Theta}} \quad s.t. \quad \mathbf{H}^s_{\mathbf{\Theta}} \geq 0 \tag{8}$$

$$\mathbf{X}^s = [\mathbf{x}^s_1, \ldots, \mathbf{x}^s_L] \tag{9}$$

$$\mathbf{H}^s_{\mathbf{\Theta}} = [\mathbf{h}^s_{\mathbf{\Theta}1}, \ldots, \mathbf{h}^s_{\mathbf{\Theta}L}] \tag{10}$$

$$\mathbf{h}^s_{\mathbf{\Theta}l} = [\mathbf{h}^s_{\boldsymbol{\theta}_1 l}, \ldots, \mathbf{h}^s_{\boldsymbol{\theta}_K l}]^T \tag{11}$$

$$\mathbf{h}^s_{\boldsymbol{\theta}_k l} = [h^s_{\boldsymbol{\theta}_1 l}, \ldots, h^s_{\boldsymbol{\theta}_{Mk} l}]^T \tag{12}$$

Then, the l-th frame of input feature $\mathbf{x}^s_l$ is represented by a linear combination of bases from the sub-dictionary of the source speaker. The sub-dictionary $\mathbf{\Phi}^s_{\hat{k}}$, which corresponds to $\mathbf{x}_l$, is selected as follows:

$$\hat{k} = \arg\max_k \mathbf{1}^{1\times M_k} \mathbf{h}^s_{\boldsymbol{\theta}_k l} = \arg\max_k \sum_{m=1}^{M_k} h^s_{\boldsymbol{\theta}_m l} \tag{13}$$

$$\mathbf{x}_l = \mathbf{\Phi}^s_{\hat{k}} \mathbf{h}_{\hat{k},l} \tag{14}$$

The activity $\mathbf{h}_{l,\hat{k}}$ in Eq. (14) is estimated from the selected source speaker sub-dictionary.

If the selected sub-dictionary $\mathbf{\Phi}^s_{\hat{k}}$ is related to consonants, the l-th frame of the converted spectral feature $\hat{\mathbf{y}}_l$ is constructed by using the activity and the sub-dictionary of the target speaker $\mathbf{\Phi}^t_{\hat{k}}$.

$$\hat{\mathbf{y}}_l = \mathbf{\Phi}^t_{\hat{k}} \mathbf{h}_{\hat{k},l} \tag{15}$$

On the other hand, if the selected sub-dictionary $\mathbf{\Phi}^s_{\hat{k}}$ is related to vowels, the l-th frame of the converted spectral feature $\hat{\mathbf{y}}_l$ is constructed by using the activity and the sub-dictionary of the source speaker $\mathbf{\Phi}^s_{\hat{k}}$.

$$\hat{\mathbf{y}}_l = \mathbf{\Phi}^s_{\hat{k}} \mathbf{h}_{\hat{k},l} \tag{16}$$

Table 1: Japanese phoneme categories

	Category	phoneme
vowels	a	a
	e	e
	i	i
	o	o
	u	u
consonants	plosives	Q, b, d, dy, g, gy, k, ky, p, t
	fricatives	ch, f, h, hy, j, s, sh, ts, z
	nasals	m, my ny, N
	semivowels	w,y
	liquid	r, ry

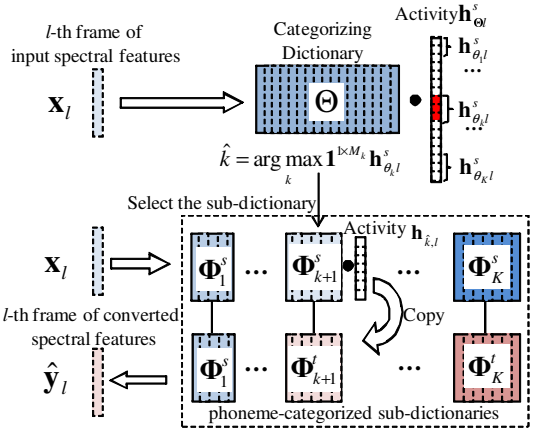


Figure 5: NMF-based voice conversion using categorized dictionary

5 Experimental Results

5.1 Experimental Conditions

The proposed VC technique was evaluated by comparing it with the conventional NMF-based method (Aihara et al., 2014) (referred to as the "sample-based method" in this paper) and the conventional GMM-based method (Stylianou et al., 1998) using clean speech data. We recorded 432 utterances (216 words, each repeated two times) included in the ATR Japanese speech

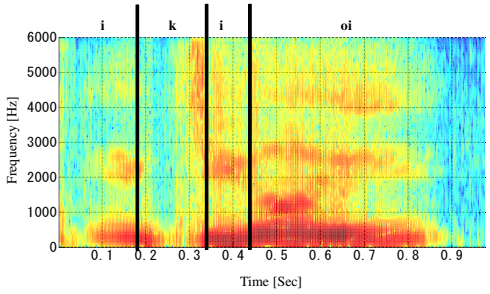
database (Kurematsu et al., 1990). The speech signals were sampled at 12 kHz and windowed with a 25-msec Hamming window every 10 msec. A physically unimpaired Japanese male in the ATR Japanese speech database, was chosen as a target speaker.

In the proposed and sample-based methods, the number of dimensions of the spectral feature is 2,565. It consists of a 513-dimensional STRAIGHT spectrum (Kawahara et al., 1999) and its consecutive frames (the 2 frames coming before and the 2 frames coming after). The Gaussian mixture, which is used to construct a categorizing-dictionary, is 1/500 of the number of bases of each sub-dictionary. The number of iterations for estimating the activity in the proposed and sample-based methods was 300. In the conventional GMM-based method, MFCC+ΔMFCC+$\Delta\Delta$MFCC is used as a spectral feature. Its number of dimensions is 74. The number of Gaussian mixtures is set to 64, which is experimentally selected.

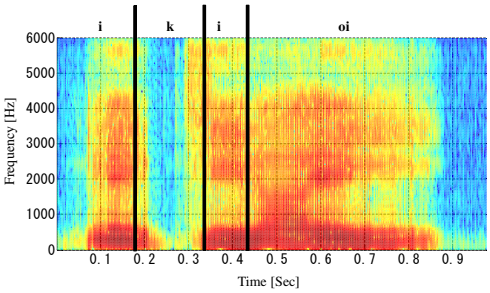
In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation (Toda et al., 2007). The other information such as aperiodic components, is synthesized without any conversion.

We conducted a subjective evaluation of 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 ("INTERNATIONAL TELECOMMUNICATION UNION", 2003). The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Twenty-two words that 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 voice converted by our proposed method, and the GMM-based converted voice.

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.



(a) Converted by proposed method



(b) Converted by GMM-based VC

Figure 6: Examples of converted spectrogram //i k i oi

5.2 Results and Discussion

Fig. 6(a) and 6(b) show examples of converted spectrograms using our proposed method and the conventional GMM-based method, respectively. In Fig. 6(a), there are fewer misconversions in the vowel part compared to Fig. 3(c). Moreover, by using GMM-based conversion, the area labeled "oi" becomes unclear compared to NMF-based conversion.

Fig. 7 shows the results of the MOS test for listening intelligibility. The error bars show a 95% confidence score; thus, our proposed VC method is shown to be able to improve the listening intelligibility and clarity of consonants. On the other hand, GMM-based conversion can improve the clarity of consonants, but it deteriorates the listening intelligibility. This is because GMM-based conversion has the effect of noise resulting from measurement error. Our proposed VC method also has the effect of noise, but it is less than that created by GMM-based conversion.

Fig. 8 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. Our proposed VC method obtained a higher score than Sample-based and GMM-based conversion on similarity. Fig. 9

shows the preference score on the naturalness. The error bars show a 95% confidence score. Our proposed VC also method obtained a higher score than Sample-based and GMM-based conversion methods in regard to naturalness.

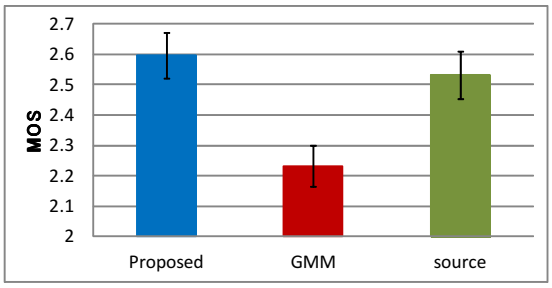


Figure 7: Results of MOS test on listening intelligibility

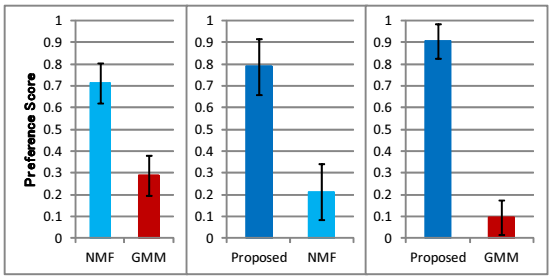


Figure 8: Preference scores for the individuality

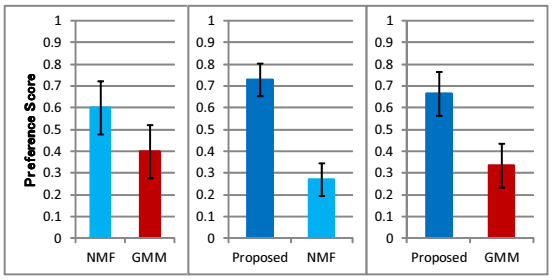


Figure 9: Preference scores for the naturalness

6 Conclusion

We proposed a spectral conversion method based on NMF for a voice with an articulation disorder. Our proposed method introduced a dictionary-selection method for conventional NMF-based VC. 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 and conventional NMF-based VC, our proposed VC method 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.

References

R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki. 2014. A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary. *EURASIP Journal on Audio, Speech, and Music Processing, 2014:5, doi:10.1186/1687-4722-2014-5*.

S. T. Canale and W. C. Campbell. 2002. Campbell's operative orthopaedics. Technical report, Mosby-Year Book.

J. F. Gemmeke, T. Viratnen, and A. Hurmalainen. 2011. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 7*, pages 2067–2080.

E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj. 2010. Voice conversion using partial least squares regression. *IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5*, pages 912–921.

Mads Vilhelm Hollegaard, Kristin Skogstrand, Poul Thorsen, Bent Norgaard-Pedersen, David Michael Hougaard, and Jakob Grove. 2013. Joint analysis of SNPs and proteins identifies regulatory IL18 gene variations decreasing the chance of spastic cerebral palsy. *Human Mutation, Vol. 34*, pages 143–148.

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

A. Kain and M. W. Macon. 1998. Spectral voice conversion for text-to-speech synthesis. *in ICASSP, vol. 1*, pages 285–288.

H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne. 1999. 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*, 27(3-4):187–207.

A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9:357–363.

C. H. Lee and C. H. Wu. 2006. Map-based adaptation for speech conversion using adaptation data selection and non-parallel training. *in Interspeech*, pages 2254–2257.

H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayachi. 2009. Integration of metamodel and acoustic model for dysarthric speech recognition. *Journal of Multimedia, Volume 4, Issue 4*, pages 254–261.

K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. 2006. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. *in Interspeech*, pages 148–151.

K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. 2012. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146.

F. Rudzicz. 2011. Acoustic transformations to improve the intelligibility of dysarthric speech. *in proc. the Second Workshop on Speech and Language Processing for Assistive Technologies*.

F. Rudzicz. 2014. Adjusting dysarthric speech signals to be more intelligible. *in Computer Speech and Language, 27(6), September*, pages 1163–1177.

Y. Stylianou, O. Cappe, and E. Moilines. 1998. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, 6(2):131–142.

R. Takashima, T. Takiguchi, and Y. Ariki. 2012. Exemplar-based voice conversion in noisy environment. *in SLT*, pages 313–317.

T. Toda, Y. Ohtani, and K. Shikano. 2006. Eigenvoice conversion based on Gaussian mixture model. *in Interspeech*, pages 2446–2449.

T. Toda, A. Black, and K. Tokuda. 2007. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(8):2222–2235.

H. Valbret, E. Moulines, and J. P. Tubach. 1992. Voice transformation using PSOLA technique. *Speech Communication, vol. 11, no. 2-3, pp. 175-187*.

C. Veaux and X. Robet. 2011. Intonation conversion from neutral to expressive speech. *in Interspeech*, pages 2765–2768.

C. Veaux, J. Yamagishi, and S. King. 2012. Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders. *in Interspeech*, pages 1–4.

J. Yamagishi, Christophe Veaux, Simon King, and Steve Renals. 2013. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology, Vol. 33 (2012) No. 1*, pages 1–5.