Consonant Enhancement for Articulation Disorders Based on Non-negative Matrix Factorization

Ryo AIHARA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI
Graduate School of System Informatics, Kobe University, Japan
E-mail: aihara@me.cs.scitec.kobe-u.ac.jp, takashima@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp
Tel/Fax: +81-78-803-6570

***Abstract*—We present consonant enhancement on a voice for a person with articulation disorders 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. Speech recognition for articulation disorders has been studied; however, its recognition rate is still lower than that of physically unimpaired persons. In this paper, an exemplar-based spectral conversion using Non-negative Matrix Factorization (NMF) is applied to consonant enhancement of a voice with articulation disorders. The source speaker's spectrum is easily converted into a well-ordered speaker's spectrum. Its effectiveness is examined for voice quality and clarity of consonants for a person with articulation disorders.**

I. Introduction

In recent years, information technology has been applied in welfare-related fields. For example, sign language recognition using image recognition technology [1], text reading systems from natural scene images [2], and the design of wearable speech synthesizers for those with voice disorders [3] have been studied.

There are 34,000 people with speech impediments associated with articulation disorders in Japan alone. One of the causes of speech impediments is cerebral palsy. About two babies in 1,000 are born with 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 type [4].

In this paper, 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 of them cannot communicate by sign language or writing, so voice systems for them are much needed.

In [5], we proposed robust feature extraction based on PCA (Principal Component Analysis) with more stable utterance data instead of DCT. In [6], 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.

In this paper, we investigate consonant enhancement on the voice for articulation disorders. In the voice of a person with an articulation disorder, his/her consonants are unstable, and that makes their voice unclear. We applied an exemplar-based voice conversion (VC) system and converted the speech of those with articulation disorders into speech with well-ordered articulation. Persons with an articulation disorder can communicate if their spoken consonants are enhanced. Many statistical approaches to VC have been studied and applied to various tasks, such as speaker conversion, emotion conversion, speaking assistance, and so on. However, a speech conversion method for people with articulation disorders resulting from athetoid cerebral palsy has not been successfully developed.

GMM-based approach is widely used for VC because of its flexibility and good performance [7]. The conversion function is interpreted as the expected value of the target spectral envelope and given the source spectral envelope. The conversion parameters are evaluated by the Minimum Mean-Square Error (MMSE) using a parallel training set. The well-known shortcoming of GMM-based approach is overfitting. Overfitting occurs when a model has too many degrees of freedom compared to the amount of training data available. Because recording the voices of people with articulation disorders is difficult, the amount of our training data is not enough for GMM-based VC.

In research discussed in paper, we conducted consonant enhancement for articulation disorders using Non-negative Matrix Factorization (NMF). In the field of speech processing, NMF is a well-known approach for source separation and speech enhancement. In these approaches, the observed signal is represented by a linear combination of a small number of atoms, such as exemplar and basis of NMF, and the collection of atoms is called a 'dictionary'.

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

where $\mathbf{x}_l$ is the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ are the j-th atom and the weight, respectively. $\mathbf{A}$ and $\mathbf{h}_l$ are the dictionary and the activity of frame l, respectively. In some approaches for source separation, the 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 atoms 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 [8].

In NMF, there are two approaches: an unsupervised approach [9] and a supervised approach [10]. In our study, we adopt the supervised approach, with focus on voice conversion from poorly articulated speech resulting from articulation disorders into well-ordered articulation. The parallel dictionary, which consists of a source dictionary and target dictionary, is extracted from the parallel data. The input source signal is expressed with a sparse representation of the source dictionary. Only the activity related to the source dictionary is picked up, and the target signal is constructed from the target dictionary with the activity. Hence, by replacing the source dictionary with the target dictionary, the original speech spectrum is replaced with the well-ordered spectrum. In NMF-based approach, overfitting is less likely to occur because there is no statistical model in this approach. Our consonant enhancement provides good performance on a small amount of training data.

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

II. Voice Conversion Based on NMF

A. Basic Idea

In our conversion method, the parallel dictionary is used to map the source signal to the target one. Fig. 1 shows the activity matrices estimated from the source and target words uttered “ikioi” (“vigor” in English) and their dictionaries. The parallel dictionary consists of source and target dictionaries that have the same size. It was structured from the same words aligned with dynamic time warping (DTW). Spectral envelope extracted by STRAIGHT analysis [11] is used as the source and target features.

As shown in Fig. 1, these activities have high energies at similar elements. For this reason, when there are parallel dictionary, the activity of source signal estimated with the source dictionary may be able to be substituted to that of target signal. Therefore, the target speech can be constructed by using the target dictionary and the activity of source signal as shown in Fig. 2, where D, L, and J represent the numbers of dimensions, frames and exemplars, respectively.

Fig. 3 shows the process for constructing the parallel dictionaries. Each dictionary is constructed using the STRAIGHT spectrum. The Mel-cepstral coefficient, which is converted from the STRAIGHT spectrum, is used for DP-matching in order to align the temporal fluctuation. The other features extracted by STRAIGHT analysis, such as F0 and the aperiodic components, are used to synthesize the converted signal.

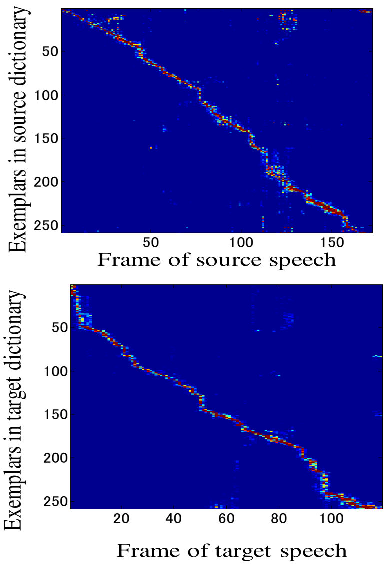

Fig. 1. Activity matrices of the source signal (top) and target signal (bottom)

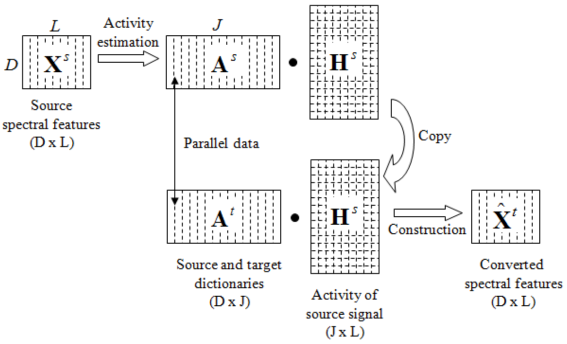
L
Activity estimation
J
D
X^s
A^s
•
H^s
Source spectral features (D x L)
Parallel data
Copy
A^t
•
H^s
Construction
$\hat{X}^t$
Source and target dictionaries (D x J)
Activity of source signal (J x L)
Converted spectral features (D x L)

Fig. 2. Basic approach of exemplar-based voice conversion

B. Estimation of Activity

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

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

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

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

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

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

$\mathbf{1}$ is an all-one matrix. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}$ and $\mathbf{A}^s\mathbf{H}^s$. The second term is the sparse constraint with L1-norm regularization term that couses $\mathbf{H}^s$ to be sparse. The weights of the sparsity constraints

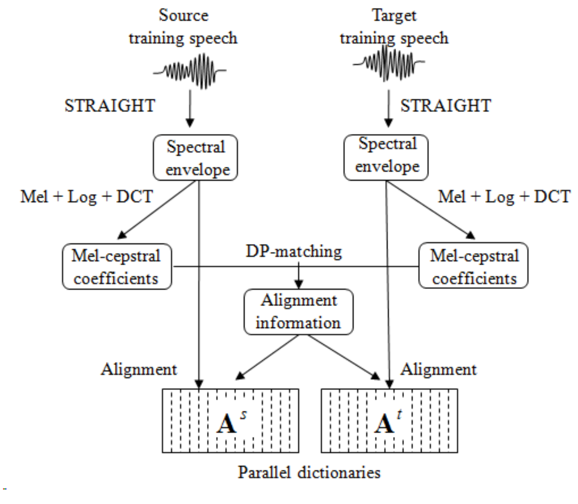

Fig. 3. Construction of source and target dictionaries

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. $\mathbf{H}$ minimizing (4) is estimated iteratively applying the following update rule:

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

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

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

III. Experimental Results

A. Experimental Conditions

We conducted on word-based conversion. In this experiment, we used 50 words from among 216 standard words in the ATR Japanese speech database, and recorded the same words uttered by a person with an articulation disorder. The speech signals were sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. Fig. 4 shows an example of a sound wave for the word “techou” (“notebook” in English) of a person with an articulation disorder. The spectrogram of a physically unimpaired person speaking the same word is shown in Fig. 5. We performed 2 types of conversions. One was a closed experiment in which all 50 words are used as training data, and we converted the same 50 words. The other was an open experiment (a one-leave-out cross-validation).

The pitch of the source speaker was converted as follows:

$$t_n = \frac{\sigma^{(t)}}{\sigma^{(s)}}(s_n - \mu^{(s)}) + \mu^{(t)} \tag{7}$$

where s_n and t_n are a log-scaled F_0 of the source and the converted speaker at frame n, respectively. Parameters $\mu^{(s)}$ and σ^s denote the mean and standard deviation of log-scaled F_0 calculated from features of the source speaker, respectively. Parameters $\mu^{(t)}$ and $\sigma^{(t)}$ are those of the target speaker.

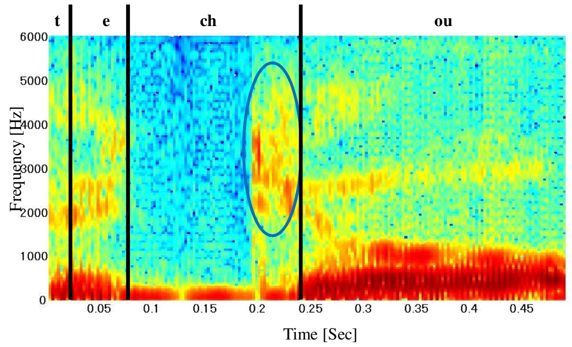

Fig. 4. Example of a spectrogram spoken by a person with an articulation disorder //t e ch ou

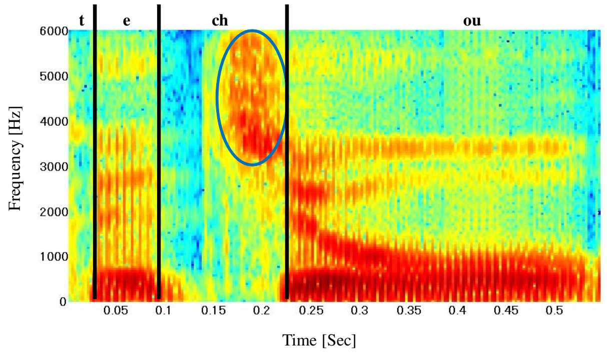

Fig. 5. Example of a spectrogram spoken by a physically unimpaired person //t e ch ou

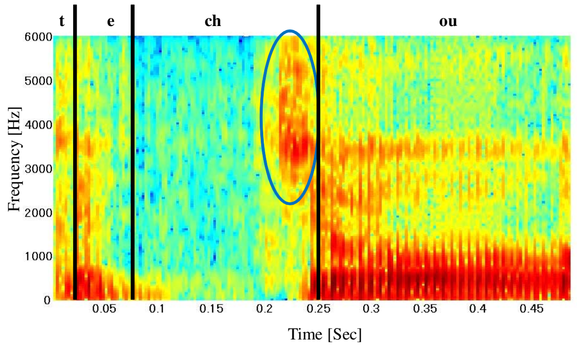

Fig. 6. Example of a spectrogram converted by close experiment //t e ch ou

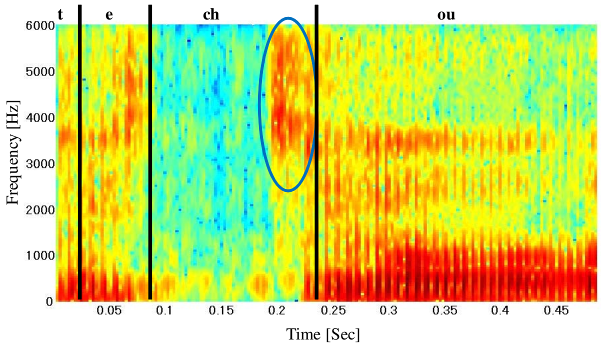

Fig. 7. Example of a spectrogram converted by open experiment //t e ch ou

The dictionary used in this experiment was constructed as follows. The source spectrum and target spectrum were first labeled by using phone HMMs and aligned with the dynamic programing matching. A spectrum will not be aligned well without a correct phone labeling because the spectrum with an articulation disorder is different from a well-ordered spectrum. Therefore, some labels were corrected manually. The Spectra extracted from each word are combined in order to construct a single dictionary.

B. Subjective Evaluation

We performed a MOS test on the speech quality and clarity of consonants. In the voice of a person with an articulation disorder, his/her consonants are often missing, which is seen in Fig. 4 in the region labeled "ch" (third region from left). In the MOS test, an opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). All the listeners were Japanese, and the number of the listeners was 5.

Fig. 6 shows an example of a converted spectrogram of "techou" in a closed experiment. The same word converted in an open experiment is also shown in Fig. 7. In Figs. 4 and 5, there is a big difference in the area labeled "ch" (blue circle in the figure). After the conversion, the "ch" area becomes similar to that of a physically unimpaired person.

Fig. 8 shows the results of the MOS test. Our consonant enhancement method can improve the speech quality and clarity of consonants. The voice of the person with an articulation disorder (source) has poor quality and many consonants are dropped, so it is difficult to understand what he/she said. By closed conversion, voice quality is greatly improved. Clarity of consonant is also improved and the speech with an articulation disorder becomes quite understandable. This fact can also be confirmed from Fig. 6. In open conversion, voice quality is a little worse than closed conversion, but better than the source voice. The same fact is confirmed in terms of the clarity of consonant, which is also shown in the area of "ch" in Fig. 7.

-

IV. Conclusions

Consonants of speech uttered by persons with speech disorders tend to become unstable due to strain on their speech-related muscles. We discussed a consonant enhancement based on NMF for a voice of articulation disorders. The parallel exemplars (dictionary) consist of the source exemplars (spectrum envelope of articulation disorders) and target exemplars (that of physically unimpaired person), having the same texts uttered by the source and target speakers. The input source spectrum is decomposed into the source exemplars (and their activities). Then, by using the activity of source exemplars, the converted spectrum is constructed from the target exemplars.

The amount of our training data is very small because recording the voice of a person with an articulation disorder is a difficult task. Even though there is not a large amount of training data, our consonant enhancement was quite successful because our approach does not use the statistics of the training data. Experimental results show that the quality of the voice and the clarity of constants can be drastically improved by using our method.

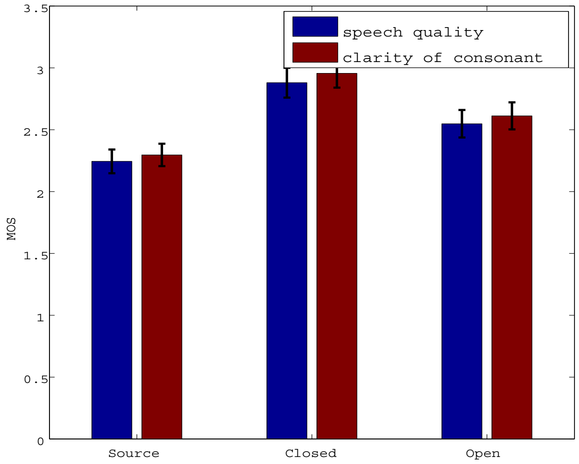

Fig. 8. Results of MOS test on speech quality and clearness of consonants

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. Further work can be done to compare our method with GMM-based VC or other consonant enhancement methods.

Acknowledgment

This research was supported in part by MIC SCOPE.

References

[1] J. Lin, W. Ying and T.S. Huang, "Capturing human hand motion in image sequences," IEEE Motion and Video Computing Workshop, pp. 99–104, 2002.

[2] M. K. Bashar, T. Matsumoto, Y. Takeuchi, H. Kudo and N. Ohnishi, "Unsupervised Texture Segmentation via Wavelet-based Locally Orderless Images (WLOIs) and SOM," 6th IASTED International Conference COMPUTER GRAPHICS AND IMAGING, 2003.

[3] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, "Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech," INTERSPEECH, pp. 1395–1398, 2006.

[4] S. T. Canale and W. C. Campbell, "Campbell's Operative Orthopaedics," Mosby-Year Book, 2002.

[5] H. Matsumasa, T. Takiguchi, Y. Ariki, I. LI and T. Nakabayashi, "PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders," INTERSPEECH, pp. 1150–1153, 2007.

[6] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal Speech Recognition of a Person with Articulation Disorders Using AAM and MAF," 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP'10), pp. 517-520, 2010.

[7] Y. Stylianou, O. Cappe and E. Moilines, "Statistical methods for voice quality transformation," Eurospeech, pp. 447-450, 1995.

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

[9] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 177–180, 2003.

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

[11] 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, pp. 187– 207, 1999.