Robust Feature Extraction to Utterance Fluctuations Due to Articulation Disorders Based on Sparse Expression

Toshiya Yoshioka, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki
Graduate School of System Informatics, Kobe University, Japan
E-mail: yoshioka@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: +81-78-803-6226/6226

***Abstract*—We investigated the speech recognition of a person with articulation disorders resulting from athetoid cerebral palsy. Recently, the accuracy of speaker-independent speech recognition has been remarkably improved by the use of stochastic modeling of speech. However, the use of those acoustic models causes degradation of speech recognition for a person with different speech styles (e.g., articulation disorders). In this paper, we discuss our efforts to build an acoustic model for a person with articulation disorders. The articulation of the first utterance tends to become more unstable than other utterances due to strain on speech-related muscles, and that causes degradation of speech recognition. Therefore, we propose a robust feature extraction method based on exemplar-based sparse representation using NMF (Non-negative Matrix Factorization). In our method, the unstable first utterance is expressed as a linear and non-negative combination of a small number of bases created using the more stable utterances of a person with articulation disorders. Then, we use the coefficient of combination as an acoustic feature. Its effectiveness has been confirmed by word-recognition experiments.**

I. Introduction

Recently, the importance of information technology in the welfare-related fields has increased. 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 voice disorders [3] have been studied.

As for speech recognition technology, the opportunities in various environments and situations have increased (e.g., operation of a car navigation system, lecture transcription into a document in a meeting). However, degradation can be observed in children [4], persons with speech impediments, and so on, and there has been very little research on orally-challenged people, such as those with speech impediments. 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. 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 as follows: 1) spastic type 2) athetoid type 3) ataxic type 4) atonic type 5) rigid type, and a mixture of types [5].

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 a cerebral palsy sufferers. In the case of a person with this type of articulation disorder, the first movements are sometimes more unstable than usual. That means, in the case of speaking-related movements, the first utterance is often unstable or unclear due to the athetoid symptoms, and that causes degradation of speech recognition. Therefore, we recorded speech data for a person with a speech impediment who uttered a given word several times, and we investigated the influence of the unstable speaking style caused by the athetoid symptoms.

In current speech recognition technology, MFCC (Mel Frequency Cepstral Coefficient) has been widely used. The feature is derived from the mel-scale filter bank output by DCT (Discrete Cosine Transform). The low-order MFCCs account for the slowly changing spectral envelope, while the high-order ones describe the fast variations of the spectrum. Therefore, a large number of MFCCs is not used for speech recognition because we are only interested in the spectral envelope, not in the fine structure. In [6], we proposed robust feature extraction based on PCA (Principal Component Analysis) with more stable utterance data instead of DCT. In [7], 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 propose a robust feature extraction method to utterance fluctuations of articulation disorders based on exemplar-based sparse representation using Non-negative Matrix Factorization (NMF). In the field of speech processing, NMF is a popular method for source separation and speech enhancement [8][9]. In these approaches, amplitude spectrum of the observed signal can be expressed as a linear and non-negative combination of only a small number of exemplars, called atoms. The collection of atoms is called a “dictionary”. In some approaches for source separation, a dictionary is

constructed for each source, and the mixed signal is 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. [10] also used the activity of the speech dictionary as phonetic scores (instead of the likelihoods of HMM) for speech recognition.

Our new robust feature extraction method is constructed by using NMF. In NMF, there are two approaches: an unsupervised approach and a supervised approach. Fig. 1 shows a system flow chart of robust feature extraction. In our approach, unsupervised NMF is employed to create the basis-matrix which consists of the stable utterance bases of articulation disorders at the training step. When extracting the robust feature of test signals, we use supervised NMF with the basis-matrix created at the training step, and use the weights of each basis in the sparse representation as an acoustic feature for speech recognition. In our experiment, the information about the weights of each basis is used to supplement the conventional acoustic features (e.g., MFCCs).

This paper is organized as follows. In Section 2, we describe our robust feature extraction method using NMF. Its effectiveness is evaluated in Section 3, and the final section concludes this paper and discusses the work we need to do in the future.

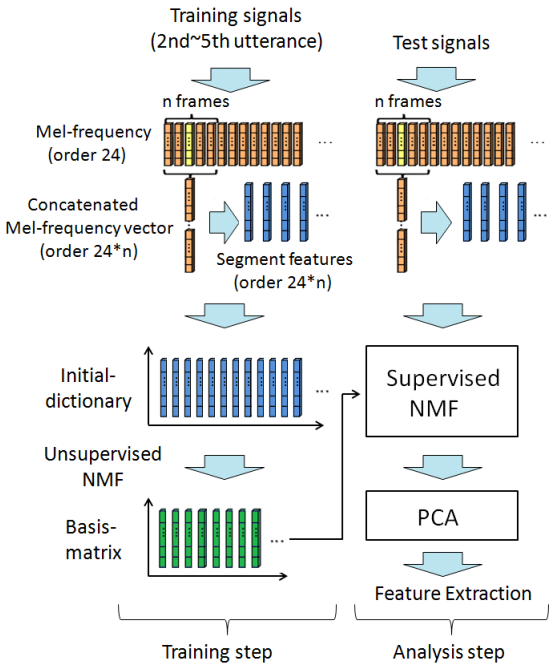

Fig. 1. Flowchart of proposed method

II. ROBUST FEATURE EXTRACTION

The first utterance of those who have articulation disorders due to athetoid cerebral palsy tends to become unstable due to the athetoid symptoms. So, we propose a robust feature extraction method for improving the recognition rate of the first utterance. As can be seen from Fig. 1, the proposed method consists of the training step, in which the basis-matrix is created using the speaker's more stable utterances, and the analysis step, in which the sparse features are extracted using supervised NMF with the basis-matrix. Each step is explained in detail below.

A. The training step

In this step, we create the basis-matrix using the more stable utterances. First, we compute the segment features of the mel-scale filter-bank output using the 2nd through 5th utterances, where we recorded 210 words, repeating each five times. Fig. 2 shows the flow of the extraction of the segment features. The frame concerned and several frames left and right of this frame are made to connect. This allows us to capture time dynamics. After that, the segment features are grouped into the initial-dictionary. Furthermore, since many similar exemplars are included in this initial-dictionary, unsupervised NMF is applied to this initial-dictionary.

Employing unsupervised NMF to the initial-dictionary $\mathbf{V}(\in \mathbf{R}^{F\times T})$, $\mathbf{V}$ is approximately decomposed into the product of basis-matrix $\mathbf{W}(\in \mathbf{R}^{F\times R})$ and activity-matrix $\mathbf{H}(\in \mathbf{R}^{R\times T})$ as follows:

$$\mathbf{V} \approx \mathbf{WH} \quad \forall i,j,k \quad \mathbf{W}_{ij} \geq 0, \mathbf{H}_{jk} \geq 0 \tag{1}$$

where F, T, and R are the numbers of bins of frequency, frames, and bases, respectively. $\mathbf{W}$ and $\mathbf{H}$ can be obtained by iteratively calculating update rules based on Euclidean divergence. The update rules for each matrix element are:

$$\mathbf{W}_{ij} = \mathbf{W}_{ij} \frac{(\mathbf{VH}^T)_{ij}}{(\mathbf{WHH}^T)_{ij}} \tag{2}$$

$$\mathbf{H}_{jk} = \mathbf{H}_{jk} \frac{(\mathbf{W}^T\mathbf{V})_{jk}}{(\mathbf{W}^T\mathbf{WH})_{jk}} \tag{3}$$

We use the updated exemplars $\mathbf{W}$ as the basis-matrix in the analysis step.

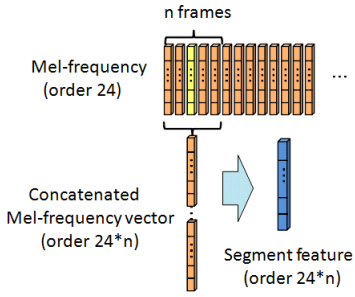

Fig. 2. Multiple acoustic frames construction

B. The analysis step

In this step, we extract robust features to the utterance fluctuations associated with articulation disorders using exemplar-based sparse representation. First, the activity-matrix of the first utterance (the unstably articulated utterance) is extracted by using supervised NMF with the basis-matrix created at the training step.

After that, Principal Component Analysis (PCA) is applied to the extracted activity-matrix for reducing the data dimensions and orthogonalization. Then we compute the filter (eigenvector matrix) using the 2nd through 5th utterances (the more stable utterances). We use the activity-matrix to which PCA is applied as an acoustic feature (called "Sparse feature" in this paper) for speech recognition.

C. Estimation of the activity

In the exemplar-based approach, we assume that the spectrum of the l-th frame of the observed signal can be expressed by a non-negative linear combination of a small number of exemplars in the dictionary and their activities:

$$\begin{aligned} \mathbf{v}_l &= \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} \\ &= \mathbf{W}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{4}$$

where $\mathbf{h}_l$ is the non-negative weights of each exemplar. J is the total number of frames. When the spectrogram is given, (4) can be written as follows:

$$\mathbf{V} = \mathbf{W}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0 \tag{5}$$

In the analysis step, $\mathbf{V}$ represents the segment features of the test signals, and $\mathbf{W}$ represents the basis-matrix created during the training step. Then, the activity-matrix $\mathbf{H}$ which consists of the non-negative weights of each basis is found by minimizing the following cost function:

$$d(\mathbf{V}, \mathbf{W}\mathbf{H}) + \|\lambda . * \mathbf{H}\|_1 \quad s.t. \quad \mathbf{H} \geq 0 \tag{6}$$

The first term measures the Kullback-Leibler (KL) divergence between $\mathbf{V}$ and $\mathbf{WH}$. The second term enforces the sparseness of $\mathbf{H}$ using l_1 norm regularization, weighted by element-wise multiplication (operator .*) of the vector $\lambda = [\lambda_1 \quad \lambda_2 \ldots \quad \lambda_J]$. The activity-matrix minimizing the cost function (6) is estimated iteratively applying the following update rule:

$$\mathbf{H} \leftarrow \mathbf{H} . * (\mathbf{W}^T(\mathbf{V}./(\mathbf{W}\mathbf{H}))) ./ (\mathbf{W}^T\mathbf{1} + \lambda) \tag{7}$$

where the vector $\mathbf{1}$ is an all-one matrix.

III. Recognition Experiment

A. Experimental setup

The proposed method was evaluated on word-recognition tasks for one person with articulation disorders. We recorded 210 words included in the ATR Japanese speech database repeating each word five times (Fig. 3). The speech signal was sampled at 16 kHz and windowed with a 32-msec Hamming window every 16 msec. Then we clipped each utterance manually. Fig. 4 shows an example of a spectrogram spoken by a person with articulation disorders. Fig. 5 shows a spectrogram spoken by a physically unimpaired person doing the same task.

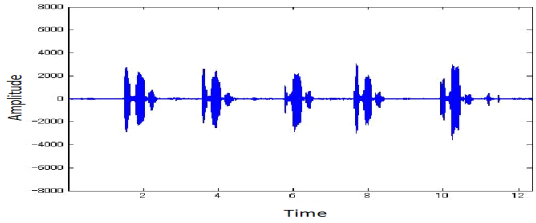

Fig. 3. Example of recorded speech data

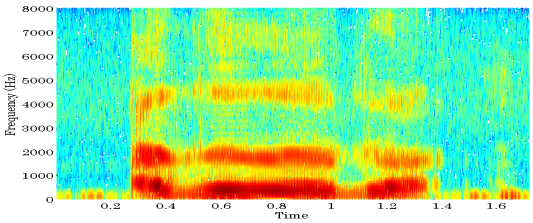

Fig. 4. Example of a spectrogram spoken by a person with articulation disorders //a k e g a t a

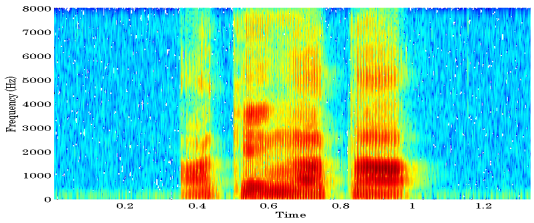

Fig. 5. Example of a spectrogram spoken by a physically unimpaired person //a k e g a t a

In the training step, we calculated the segment features (3 frames were used to make a segment) of 24 mel-scale filterbank output using the 2nd through 5th utterances, and this resulted in initial dictionary which consisted of the 91,435 exemplars. Next, we created the basis-matrix by selecting 100 bases from 91,435 exemplars using unsupervised NMF. At the analysis step, the activity-matrix of each utterance was obtained by using supervised NMF with the basis-matrix which consists of 100 bases. The sparsity parameter λ was set to 0.65, and the update rule (7) was run for 1,000 iterations, which was enough to converge.

It was difficult to recognize the utterances of a person with articulation disorders using an acoustic model trained by the utterances of a physically unimpaired person. Therefore, we trained the acoustic model using the utterances of a person with articulation disorders. When we recognize the 1st utterance, the 2nd through 5th utterances were used for training. We iterated this process for each utterance. The

acoustic model consists of an HMM set with 54 context-independent phonemes and 10 mixture components for each state. Each HMM has three states and three self-loops.

B. Recognition results

In the proposed feature extraction, PCA was applied to the activity-matrix of each utterance to reduce the data dimensions. In order to determine the optimal number of dimensions, speech recognition was performed to change the number of principal components from 14 to 24 dimensions. Fig. 6 shows the recognition rates of the 1st utterance using each dimensional Sparse feature.

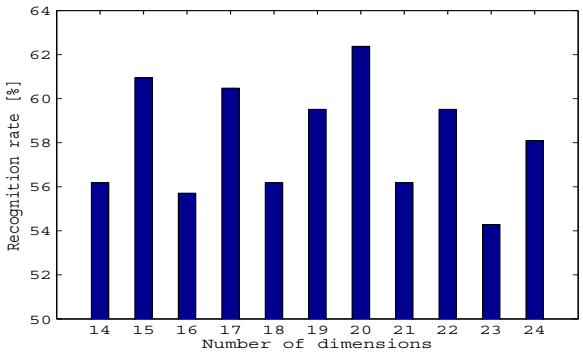

Fig. 6. Recognition rate for the 1st utterance using the proposed method

In Fig. 6, the recognition rate for the 20 principal components was the highest. So, the extracted 100 dimensional activity-matrix was reduced to 20 dimensions using PCA in our experiment.

Next, we compared the proposed method with the following three kinds of features:

1) 24-dimensional MFCC-delta feature (12-order MFCCs and their delta)
2) 20-dimensional Sparse feature (extracted using proposed method)
3) 44-dimensional Sparse feature + MFCC-delta (connected Sparse feature and MFCC-delta)

TABLE I shows the recognition rates for each utterance. As can be seen from TABLE I, the use of the proposed method improves the recognition rates for the 1st utterance from 78.1% to 81.9% (feature (1) and feature (3)). These results clearly show that using the information of the activities achieves good performance. It can be expected that the unstable 1st utterance gets closer to the subsequent stable utterances using exemplar-based sparse representation, so the recognition rate of the 1st utterance has been improved by combining the sparse feature and MFCC-delta. In addition, the recognition rates of the other utterances were almost equal to MFCC-delta. However, the performance of using sparse features only was poor compared to the results of MFCC-delta. It can be expected that the quality and number of bases in the basis-matrix were insufficient, and this caused degradation of the quality of the presumed activity-matrix.

TABLE I
RECOGNITION RATES FOR EACH UTTERANCE OF ARTICULATION DISORDERS

Feature	1st	2nd	3rd	4th	5th
MFCC-delta	78.1	89.52	92.86	88.57	90
Sparse	62.38	76.67	73.81	77.14	66.19
Sparse+MFCC-delta	81.9	90	92.86	90.48	90.95

IV. CONCLUSIONS

The first utterance of a person with articulation disorders tends to become unstable due to strain on their speech-related muscles. In this paper, we proposed a robust feature extraction method for improving the recognition rate of the first utterance, and the effectiveness was confirmed by word-recognition experiments. The proposed method resulted in an improvement of 3.8% (from 78.1% to 81.9%) in the recognition rate of the first utterance compared to the conventional method, MFCC. This result clearly shows that sparse representations suppressed the instability of the first utterance, resulting in an improved recognition rate. However, the results of using the sparse features only were inferior to that of MFCC. So, in future work, we will seek to improve our method of creating the dictionary bases. In this study, there was only one subject speaker, so, in future experiments, we will increase the number of subjects and further examine the effectiveness of the proposed method.

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] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," ICASSP2003, pp. 137-140, 2003.
[5] S.T. Canale and W.C. Campbell, "Campbell's Operative Orthopaedics," Mosby-Year Book, 2002.
[6] 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.
[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," 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP'10), pp. 517-520, 2010.
[8] 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, Issue 3, pp. 1066-1074, 2007.
[9] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," INTERSPEECH, pp. 2614-2617, 2006.
[10] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," IEEE Trans. Audio, Speech, Lang. Process., Vol. 19, Issue 7, pp. 2067-2080, 2011.