

# Evaluation of Random-Projection-Based Feature Combination on Dysarthric Speech Recognition

Toshiya Yoshioka\*, Tetsuya Takiguchi, Yasuo Ariki

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

**Abstract** We investigated the speech recognition of persons with articulation disorders resulting from cerebral palsy. The articulation of their first speech tends to become unstable due to strain on speech-related muscles, and that causes degradation of speech recognition. In this paper, we propose a feature extraction method based on RP (Random Projection) for dysarthric speech recognition. Random projection has been suggested as a means of space mapping, where the original data are projected onto a space using a random matrix. It represents a computationally simple method that approximately preserves the Euclidean distance of any two points through the projection. Moreover, as we are able to produce various random matrices, there may be some possibility of finding a random matrix that gives better speech recognition accuracy among these random matrices. To obtain an optimal result from many random matrices, a vote-based combination is introduced in this paper. ROVER combination is applied to the recognition results obtained from the ASR (Automatic Speech Recognition) systems created from each RP-based feature. Its effectiveness is confirmed by word recognition experiments.

**Keywords** Articulation Disorders, Speech Recognition, Random Projection, ROVER

## 1. Introduction

Recently, the importance of information technology in the welfare-related fields has increased. For example, sign language recognition using image recognition technology [1][2][3], text-reading systems from natural scene images [4][5][6], and the design of wearable speech synthesizers for voice disorders [7][8] have been studied.

There are 34,000 people with speech impediments associated with articulation disorders in Japan alone, and it is hoped that speech recognition systems will one day be able to recognize their voices. 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 [9].

In this paper, we focused on persons with articulation disorders 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, 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 persons with articulation disorders who uttered each of the words several times, and investigated the influence of the unstable speaking style caused by the athetoid symptoms.

The goal of front-end speech processing in ASR is to obtain a projection of the speech signal to a compact parameter space where the information related to speech content can be extracted. In current speech recognition technology, MFCC (Mel-Frequency Cepstrum Coefficient) is being widely used. The feature is uniquely 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 [10], we proposed robust feature extraction based on PCA (Principal Component Analysis) with more stable utterance data instead of DCT in a dysarthric speech recognition task. Also, [11] used MAF (multiple acoustic frames) as an acoustic dynamic feature to improve the recognition rate of a person with articulation disorders, especially in speech recognition using dynamic features only. These methods improved the recognition accuracy, but the performance for articulation disorders was

\* Corresponding author:

yoshioka@me.cs.scitec.kobe-u.ac.jp (Toshiya Yoshioka)

Published online at <http://journal.sapub.org/ajsp>

Copyright © 2013 Scientific & Academic Publishing. All Rights Reserved

not sufficient when compared to that of persons with no disability.

Random projection has been suggested as a means of space mapping, where a projection matrix is composed of the columns defined by the random values chosen from a probability distribution. In addition, the Euclidean distance of any two points is approximately preserved through the projection. Therefore, random projection has also been suggested as a means of dimensionality reduction[12]. In contrast to conventional techniques such as PCA, which find a subspace by optimizing certain criteria, random projection does not use such criteria; therefore, it is data independent. Moreover, it represents a computationally simple and efficient method that preserves the structure of the data without introducing significant distortion[13]. Goel *et al*[13] have reported that random projection has been applied to various types of problems, including information retrieval (e.g.,[14]), image processing (e.g.,[15][16]), machine learning (e.g.,[17][18][19]), and so on. Although it is based on a simple idea, random projection has demonstrated good performance in a number of applications, yielding results comparable to conventional dimensionality reduction techniques, such as PCA.

In this paper, we investigate the feasibility of random projection for speech feature transformation in order to improve the recognition rate of persons with articulation disorders. In our proposed method, original speech features (MFCCs) are transformed using various random matrices. Then, we use the same number of dimensions for the projected space as that of the original space. There may be some possibility of finding a random matrix that gives

better speech recognition accuracy among random matrices, since we are able to produce various RP-based features (using various random matrices). Therefore, a vote-based combination method is introduced in order to obtain an optimal result from many (infinite) random matrices, where ROVER combination[20] is applied to the results from the ASR systems created from each RP-based feature.

The rest of the paper is organized as follows. Section 2 describes a feature projection method using random orthogonal matrices. In Section 3, a vote-based combination method is explained. Results and discussion for the experiments on a dysarthric speech recognition task are given in Section 4. Section 5, concludes the paper with a summary of our proposed method, contribution, and future work.

## 2. Random Orthogonal Projection

This section presents a feature projection method using random orthogonal matrices. The main idea of random projection arises from the Johnson-Lindenstrauss lemma [21]; namely, if original data are projected onto a randomly selected subspace using a random matrix, then the distances between the data are approximately preserved.

Random projection is a simple yet powerful technique, and it has another benefit. Dasgupta[17] has reported that even if distributions of original data are highly skewed (have ellipsoidal contours of high eccentricity); their transformed counterparts will be more spherical.

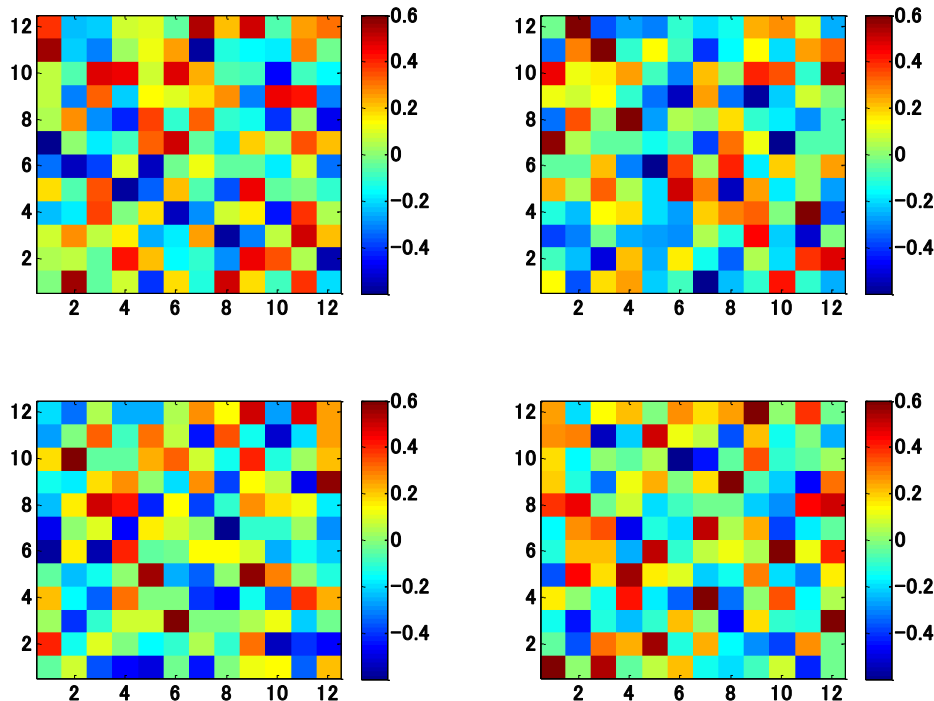


Figure 1. Examples of random matrices 12 dim. (12 × 12)

First, we choose an  $n$ -dimensional random vector,  $\mathbf{p}$ , and let  $\mathbf{P}^{(l)}$  be the  $l$ -th  $n \times d$  matrix whose columns are vectors,  $\mathbf{p}_1^{(l)}, \mathbf{p}_2^{(l)}, \dots, \mathbf{p}_d^{(l)}$ . Then, an original  $n$ -dimensional vector,  $\mathbf{X}$ , is projected onto a  $d$ -dimensional subspace using the  $l$ -th random matrix,  $\mathbf{P}^{(l)}$ , where we compute a  $d$ -dimensional vector,  $\mathbf{X}'$ , whose coordinates are the inner products,  $x'_1 = \mathbf{P}_1^{(l)} \cdot \mathbf{X}, \dots, x'_d = \mathbf{P}_d^{(l)} \cdot \mathbf{X}$ .

$$\mathbf{X}' = \mathbf{P}^{(l)\top} \mathbf{X} \quad (1)$$

In this paper, we investigate the feasibility of random projection for speech feature transformation. As described above, a random projection from  $n$  dimensions to  $d$  ( $= n$ ) dimensions is represented by an  $n \times d$  matrix,  $\mathbf{P}$ . It has been shown that if the random matrix,  $\mathbf{P}$ , is chosen from the standard normal distribution, with mean 0 and variance 1, referred to as  $N(0,1)$ . Then, the projection preserves the structure of the data[21]. In this paper we use  $N(0,1)$  for the distribution of the coordinates. The random matrix,  $\mathbf{P}$ , can be calculated using the following algorithm [13][17].

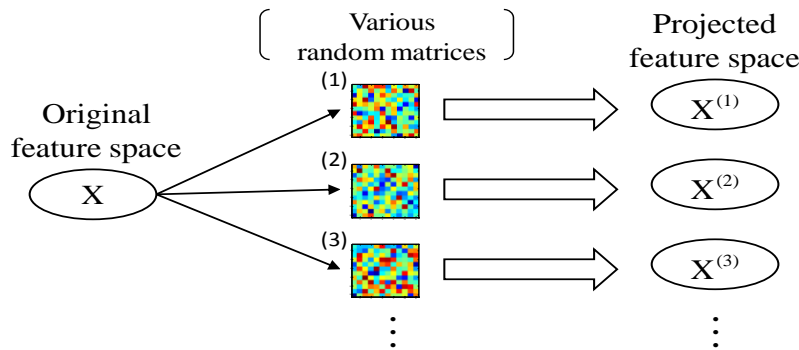
- Choose each entry of the matrix from an independent and identifiably distributed  $N(0,1)$  value.
- Make the orthogonal matrix using the Gram Schmidt algorithm, and then normalize it to unit length.

Orthogonality is effective for feature extraction because the HMMs used in speech recognition experiments consist of diagonal covariance matrices. Fig. 1 shows examples of random matrices from  $N(0,1)$ .

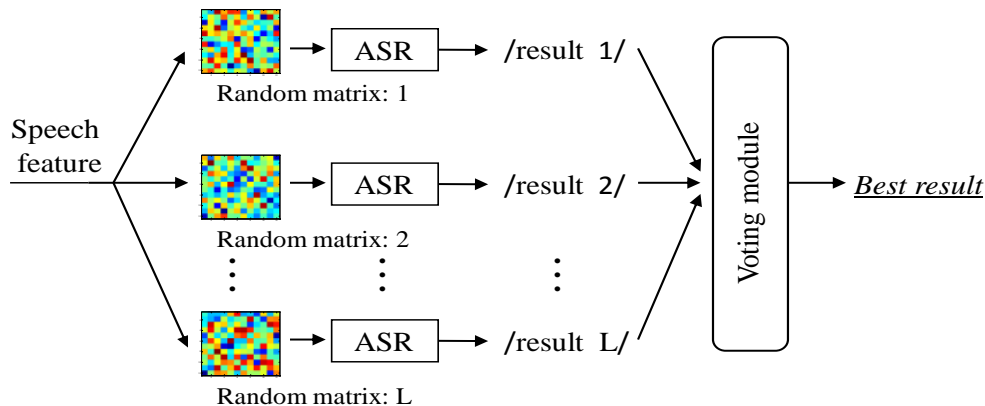
### 3. Vote-Based Combination

As described in Section 2, we can make many (infinite) random matrices from  $N(0,1)$  (Fig. 2). Since there may be some possibility of finding a random matrix that gives better performance, we will have to select the optimal matrix or the optimal recognition result from them. To obtain the optimal result, a majority vote-based combination is introduced in this paper, where ROVER combination is applied to the results from the ASR systems created from each RP-based feature.

Fig. 3 shows an overview of the vote-based combination method. First, random matrices  $\mathbf{P}^{(l)}$  ( $l=1, \dots, L$ ) are chosen from the standard normal distribution, with mean 0 and variance 1. Original speech features (MFCCs) are projected using each random matrix. An acoustic model corresponding to each random matrix is also trained. For the test utterance, using each acoustic model, an ASR system outputs the best scoring word by itself. To obtain a single hypothesis from among all the results for random projection, voting is performed by counting the number of occurrences of the best word for each RP-based feature.



**Figure 2.** Random projection on the feature domain. An original feature is transformed to various features using various random matrices. (Eq. 1)



**Figure 3.** Overview of the vote-based combination

For example, in the case of  $l = 20$ , 20 kinds of new feature vectors are calculated using 20 kinds of random matrices. Then, we train the 20 kinds of acoustic models using 20 kinds of new feature vectors. In the test process, 20 kinds of recognition results are obtained using 20 kinds of acoustic models. To obtain a single hypothesis from among 20 kinds of recognition results, voting is performed.

## 4. Evaluation

### 4.1. Experimental Conditions

The proposed method was evaluated on a word recognition task for three males with articulation disorders (Speaker A, B, C). For the conducted experiments, we recorded 210 words included in the ATR Japanese speech database for each speaker. Each of the 210 words was repeated five times (Fig. 4). The speech signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. Fig. 5 shows an example of a spectrogram spoken by a person with an articulation disorder. Fig. 6 shows a spectrogram spoken by a physically unimpaired person doing the same task.

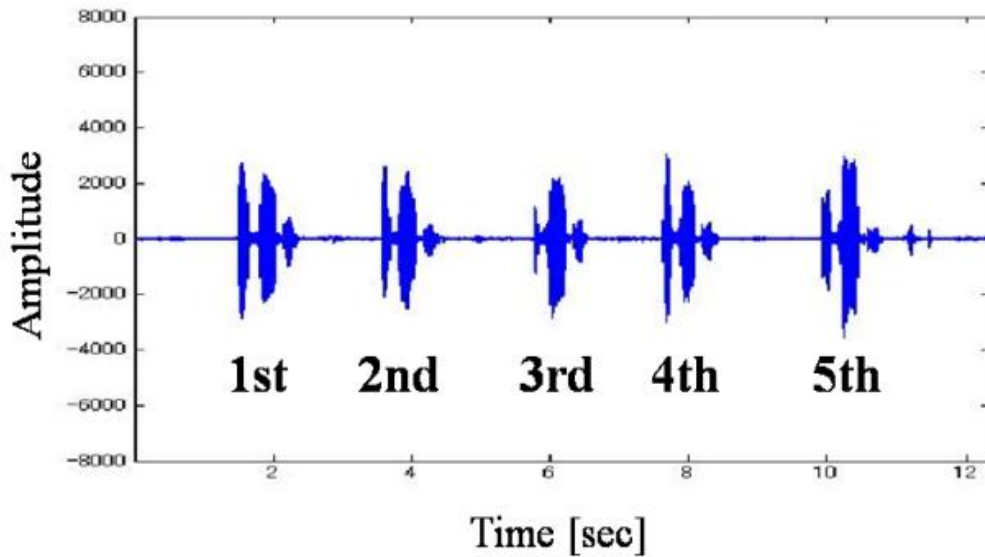


Figure 4. Example of recorded speech data

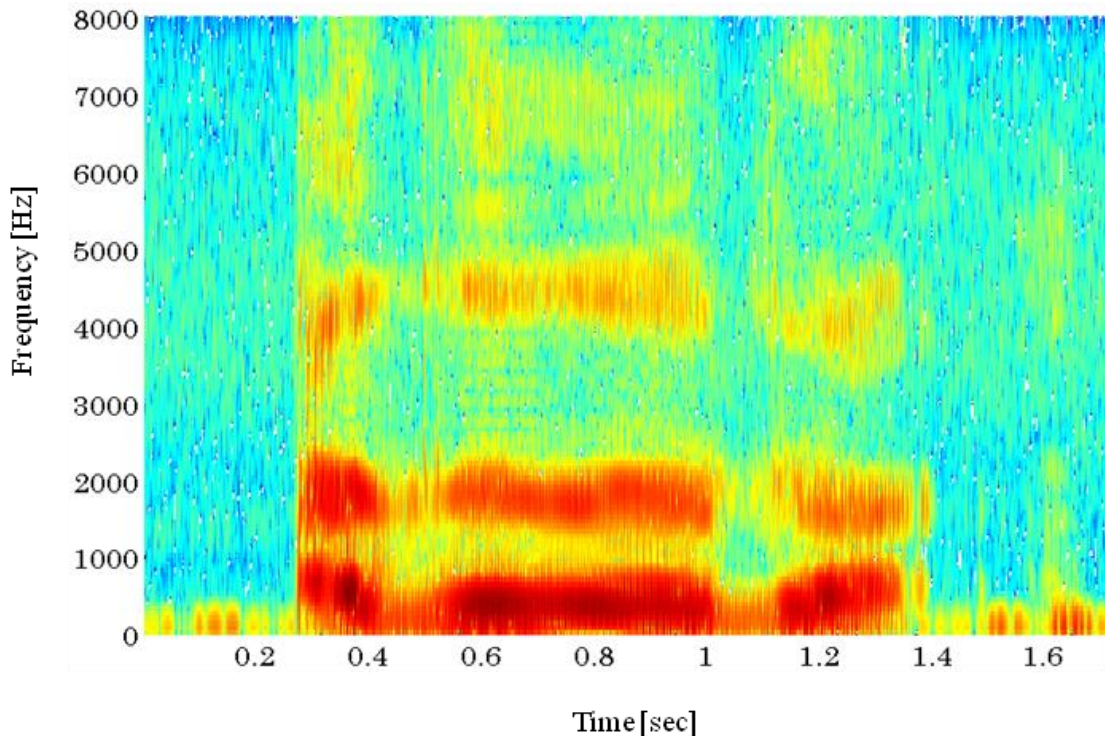
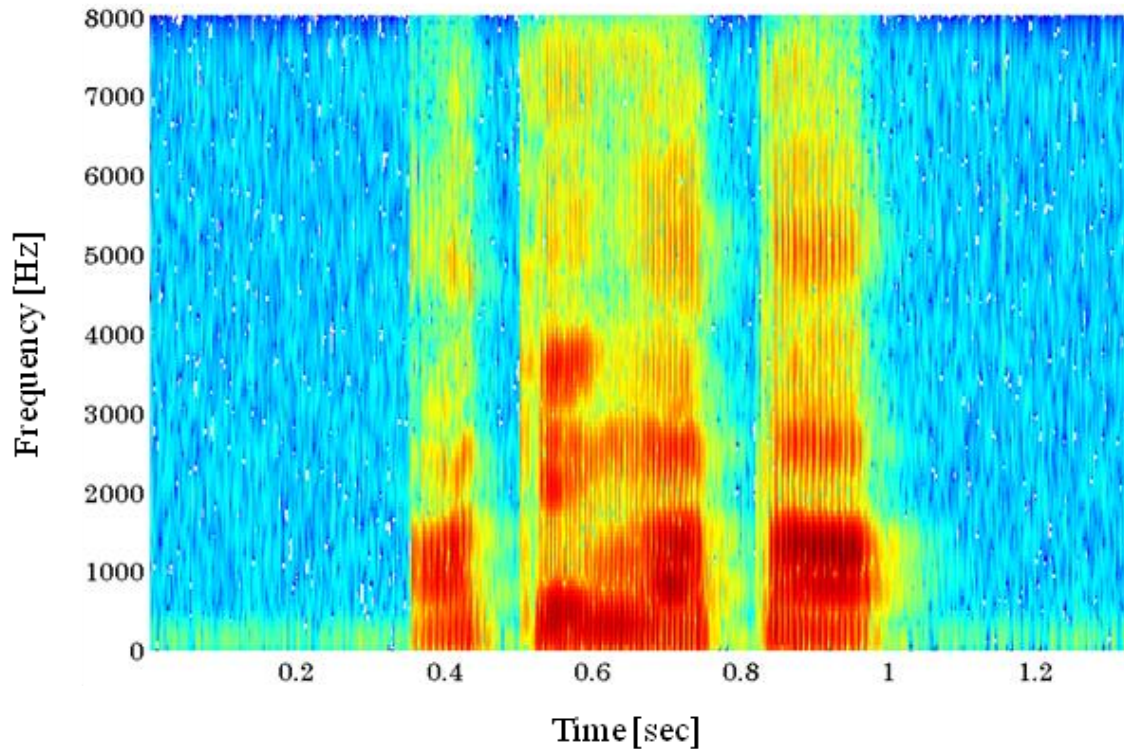
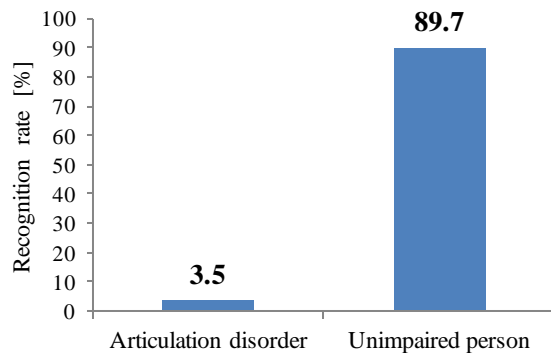


Figure 5. Example of a spectrogram from a person with an articulation disorder. (//a k e g a t a)





**Figure 6.** Example of a spectrogram from a physically unimpaired person. (/a k e t a)



**Figure 7.** Recognition results[%] for the speaker-independent model using training data uttered by unimpaired persons

The recognition results for a speaker-independent model are shown in Fig. 7, where the speech data uttered by unimpaired persons were used. As can be seen in Fig. 7, the recognition rate of a physically unimpaired person was around 90%. However, the result of a person with an articulation disorder (speaker A) was only 3.5%. It is clear that the speaking style of a person with an articulation disorder differs considerably from that of physically unimpaired persons.

Also, Fig. 7 shows that it was difficult to recognize utterances of articulation disorders using an acoustic model trained by utterances of physically unimpaired persons. Therefore, in this paper, we trained the acoustic model using the utterances of a person with an articulation disorder. The acoustic model consists of a HMM set with 54 context-independent phonemes and 8 mixture

components for each state. Each HMM has three states and three self-loops.

#### 4.2. Experiment 1

In Experiment 1, recognition results were obtained for each utterance using speaker-dependent models. Each system was trained using 24-dimensional feature vectors consisting of 12-dimensional MFCC parameters, along with their delta parameters. When we recognized the 1st utterance, the 2nd through 5th utterances were used for training. We iterated this process for each utterance.

**Table 1.** Recognition results[%] for each utterance of speaker (A)

1st	2nd	3rd	4th	5th
75.7	86.7	92.9	90.5	88.6

Table 1 shows the results obtained in Experiment 1 for speaker (A). As can be seen in Table 1, the recognition rate for the 1st utterance was 75.7%. It was significantly lower than other utterances. It is considered that the speaker experiences a more strained state during the first utterance compared to subsequent utterances because the first utterance is the first intentional movement. Therefore, athetoid symptoms occur and articulation becomes difficult. It is believed that this difficulty causes fluctuations in speaking style and degradation of the recognition rates.

Tables 2 and 3 show the recognition results for each utterance for speaker (B) and (C), respectively. As can be seen in Tables 2 and 3, a decrease in recognition rate for the first utterance due to fluctuations in speaking style was confirmed.

**Table 2.** Recognition results[%] for each utterance of speaker (B)

1st	2nd	3rd	4th	5th
85.7	91.9	91.4	93.3	95.2

**Table 3.** Recognition results[%] for each utterance of speaker (C)

1st	2nd	3rd	4th	5th
94.3	99.5	97.6	97.6	95.7

### 4.3. Experiment 2

The aim of Experiment 2 is to evaluate the improvement introduced by the use of a RP-based feature projection method for the unstable 1st utterance. In the experiments, the following RP-based features were evaluated. Random projection is applied to MFCC at the  $t$ -th frame,  $x(t) \in \mathbb{R}^{12}$ , and the new feature,  $y(t) \in \mathbb{R}^{12}$ , is obtained:

$$y(t) = \mathbf{P}^{(l)\top} x(t) \quad (2)$$

Then, the new feature also has the delta parameter of original feature,  $x(t)$ . The final system feature dimensionality is 24 (MFCC[12]  $\rightarrow$  RP[12] +  $\Delta$ MFCC[12]).

We investigated the performance of random projections for various random matrices ( $l = 20, 40, 60, 80$ , and 100) sampled from  $N(0,1)$ . Tables 4 ~ 6 show the recognition rate versus the number of random matrices for each speaker. The results of “RP w/o combination” show the maximums, means, and minimums obtained from each random projection without ROVER-based combination.

**Table 4.** Word recognition rate[%] for the 1st utterances of speaker (A) using the proposed method in various random matrices. (The recognition rate for the original features is 75.7%)

Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	80.5	80	76.8	73.8
40	81.4	80	76.8	72.9
60	<b>81.9</b>	80.5	77	72.9
80	81.4	80.5	76.9	72.9
100	81	80.5	76.8	72.9

**Table 5.** Word recognition rate[%] for the 1st utterances of speaker (B) using the proposed method in various random matrices. (The recognition rate for the original features is 85.7%)

Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	90	92.9	86.8	84.3
40	89.1	92.9	87.1	84.3
60	<b>90.5</b>	92.9	87.1	84.3
80	<b>90.5</b>	92.9	87.1	84.3
100	90	92.9	87.1	84.3

Table 4 shows the performance results versus the number of random matrices for speaker (A). As can be seen in Table 4, the results for RP-based feature indicate that the vote-based random-projection combination improved the recognition rate from 75.7% to 81.9% using the

combination of 60 random matrices, and even the means of random projections without combination for some random matrices was better than the recognition rate of the original features. Also, even if the number of random matrices is changed, we do not see that subsequent performance varies in our experiments.

Tables 5 and 6 show the performance of the proposed method for speaker (B) and (C), respectively. As can be seen in Table 5, the recognition rate of 90.5% was obtained using the combination of 60 or 80 random matrices. Also, the results in Table 6 were maintained above 95%, showing the effectiveness of our proposed method for each person with articulation disorders who participated in our experiment.

**Table 6.** Word recognition rate[%] for the 1st utterances of speaker (C) using the proposed method in various random matrices. (The recognition rate for the original features is 94.3%)

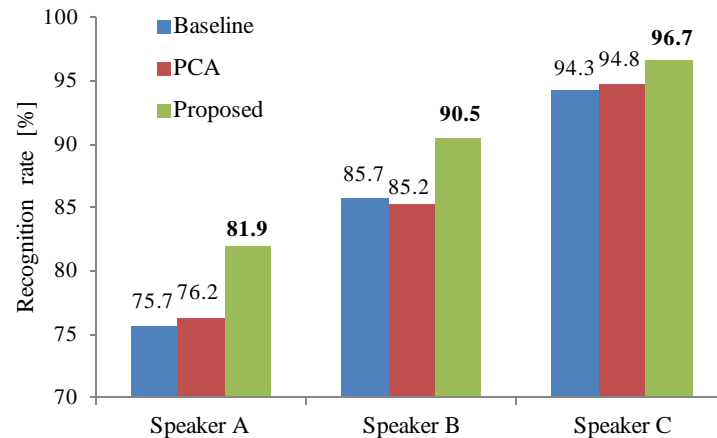
Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	96.2	97.1	95.3	93.8
40	95.7	97.1	95.3	93.8
60	<b>96.7</b>	97.6	95.5	93.8
80	<b>96.7</b>	97.6	95.5	93.3
100	95.7	97.6	95.5	93.3

### 4.4. Experiment 3

In order to show the superiority of the RP-based feature projection method, in Experiment 3, we compared the proposed method and the PCA-based feature projection method.

For Experiment 3, PCA was applied to 12-dimensional MFCC, and the new feature also had the delta coefficient of the MFCC features. Then, we computed the eigenvector matrix using the 2nd through 5th utterances (the more stable utterances) for each speaker. A comparison between the PCA-based feature projection method and the results obtained by our proposed method using the combination of 60 random matrices are shown in Fig. 8.

We can see that the combination of random projection and ROVER outperforms both the baseline method (MFCCs) and the PCA-based feature extraction method. This result gives the evidence of the improvement introduced by the speech feature extraction based on random projection and the use of ROVER to obtain an optimal result. One of the possible reasons the random projection improves the recognition rates may be that if distributions of original data are skewed (have ellipsoidal contours of high eccentricity), their transformed counterparts will become more spherical[17]. However, there were ‘bad’ projections that cause degradation of speech recognition accuracy compared with the recognition of original features (Tables 4 ~ 6). Therefore, more research will be needed to investigate the effectiveness of the random projection method for speech features.



**Figure 8.** Comparison between the PCA-based feature projection method and the results obtained by our proposed method for the 1st utterance

## 5. Conclusions

As a result of this work, a method for recognizing dysarthric speech using RP-based features has been developed. The proposed method transforms the conventional speech features such as MFCC using various random matrices. It also introduces the vote-based combination method to obtain an optimal result from the ASR systems created from each RP-based feature. Word recognition experiments were conducted to evaluate the proposed method for three males with articulation disorders. The results of the experiments showed that all the recognition rates of the proposed method outperformed the baseline rate (using MFCCs).

As future work, we will continue to investigate how to select the optimal basis vector via the use of such random matrices.

## ACKNOWLEDGEMENTS

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] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), pp. 1371-1375, 1998.
- [3] G. Fang, W. Gao and D. Zhao, "Large vocabulary sign language recognition based on hierarchical decision trees," Proceedings of the 5th international conference on Multimodal interfaces, pp. 125-131, 2003.
- [4] N. Ezaki, M. Bulacu and L. Schomaker, "Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons," ICPR 2004, pp. 683-686, 2004.
- [5] V. Wu, R. Manmatha and E. M. Riseman, "Textfinder: an automatic system to detect and recognize text images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(11), pp. 1224-1229, 1999.
- [6] 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.
- [7] T. Ohsuga, Y. Horiuchi, and A. Ichikawa, "Estimating Syntactic Structure from Prosody in Japanese Speech," IEICE Transactions on Information and Systems, 86(3), pp. 558-564, 2003.
- [8] 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.
- [9] S.T. Canale, and W.C. Campbell, "Campbell's Operative Orthopaedics," Mosby-Year Book, 2002.
- [10] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders," INTERSPEECH 2007, pp. 1565-1568, 2007.
- [11] 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, pp. 517-520, 2010.
- [12] Ella Bingham, and Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data," KDD'01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 245-250, 2001.
- [13] N. Goel, G. Bebis, and A. Nefian, "Face Recognition Experiments with Random Projection," SPIE, vol. 5779, pp. 426-437, 2005.
- [14] P. Thaper, S. Guha, and N. Koudas, "Dynamic Multidimensional Histograms," ACM SIGMOD, pp. 428-439, 2002.
- [15] L. Liu, P. Fieguth, G. Kuang, and H. Zha, "Sorted Random

- Projections for robust texture classification,” IEEE International Conference on Computer Vision, pp. 391-398, 2011.
- [16] H. T. Ho, and R. Chellappa, “Automatic head pose estimation using randomly projected dense SIFT descriptors,” IEEE International Conference on Image Processing, pp. 153-156, 2012.
  - [17] S. Dasgupta, “Experiments with random projection,” UAI, pp. 143-151, 2000.
  - [18] X. Z. Fern, and C. E. Brodley, “Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach,” the 20th Int. Conf. on Machine Learning, pp. 186-193, 2003.
  - [19] S. Lee, and A. Nedic, “Distributed Random Projection Algorithm for Convex Optimization,” IEEE Journal of Selected Topics in Signal Processing, Vol. 7, No. 2, pp. 221-229, 2013.
  - [20] J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER),” IEEE ASRU, pp. 347-352, 1997.
  - [21] R. I. Arriaga, and S. Vempala, “An algorithmic theory of learning: robust concepts and random projection,” IEEE Symposium on Foundations of Computer Science, pp. 616-623, 1999.