EVALUATION OF RANDOM-PROJECTION-BASED FEATURE COMBINATION ON SPEECH RECOGNITION

Tetsuya Takiguchi[#], Jeff Bilmes[*], Mariko Yoshii[#], Yasuo Ariki[#]

[#]Department of Computer Science and Systems Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe, 6578501, Japan
[*]Department of Electrical Engineering, University of Washington
Seattle WA, 98195, USA

ABSTRACT

Random projection has been suggested as a means of dimensionality reduction, where the original data are projected onto a subspace 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 a better speech recognition accuracy among these random matrices. In this paper, we investigate the feasibility of random projection for speech feature extraction. To obtain an optimal result from among many (infinite) random matrices, a vote-based random-projection combination is introduced in this paper, where ROVER combination is applied to random-projection-based features. Its effectiveness is confirmed by word recognition experiments.

Index Terms— feature extraction, feature combination, random projection

1. INTRODUCTION

The goal of front-end speech processing in automatic speech recognition 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). There are also other methods for feature extraction such as PCA, and those conventional features are uniquely obtained based on certain criteria. (For example, PCA finds a subspace that maximizes the variance in the data.) The effectiveness of those conventional techniques has been confirmed in speech recognition or speech enhancement experiments, but it still remains a problem of mismatch conditions such as speaker variations, noise variations and so on.

Random projection has been suggested as a means of dimensionality reduction, where a random projection matrix is used to project data into low-dimensional spaces. 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 [1]. Goel et al. [1] have reported that random projection has been applied to various types of problems, including information retrieval (e.g. [2]), machine learning (e.g. [3, 4]), 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 extraction. There may be some possibility of finding a random matrix that gives a better speech recognition accuracy among random matrices, since we are able to produce various random-projection-based features (using various random matrices). In this paper, a vote-based random-projection combination is introduced in order to obtain an optimal result from among many (infinite) random matrices, where ROVER combination is applied to random-projection-based features. Its effectiveness is confirmed by word recognition experiments.

2. RANDOM ORTHOGONAL PROJECTION

This section presents a feature projection (extraction) method using random orthogonal matrices. The main idea of random projection arises from the Johnson-Lindenstrauss lemma; namely, if original data are projected onto a randomly selected subspace using a random matrix, then the distances between the data are approximately preserved [5].

Random projection is a simple yet powerful technique, and it has another benefit. Dasgupta [3] 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.

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, $\mathbf{p}_1^{(l)}$, $\mathbf{p}_2^{(l)}$, $\ldots$, $\mathbf{p}_d^{(l)}$. 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 $x'_1 = \mathbf{p}_1^{(l)} \cdot \mathbf{x}, \ldots, x'_d = \mathbf{p}_d^{(l)} \cdot \mathbf{x}$.

$$\mathbf{x}' = \mathbf{P}^{(l)^T} \mathbf{x} \tag{1}$$

In this paper, we investigate the feasibility of random projection for speech feature extraction. 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 [5]. 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 [1, 3].

- Choose each entry of the matrix from an independent and identically distributed (i.i.d.) $N(0, 1)$ value.

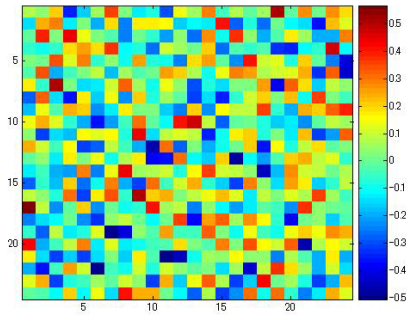


Fig. 1. An example of random matrix

- Make the orthogonal matrix by using the Gram-Schmidt algorithm, and then normalize it to unit length.

The orthogonality is effective for the feature extraction because HMMs used in speech recognition experiments consist of diagonal covariance matrices.

3. VOTE-BASED RANDOM-PROJECTION COMBINATION

Fig. 1 shows an example of the random matrix from $N(0, 1)$. As shown in Fig. 1, a random matrix is composed of various random vectors. As we can make many (infinite) random matrices from $N(0, 1)$, we will have to select the optimal matrix or the optimal recognition result from among them. To obtain the optimal result, a vote-based random-projection combination is introduced in this paper, where ROVER combination [6] is applied to random-projection-based features.

Fig. 2 shows an overview of the vote-based random-projection combination. First, random matrices, $\mathbf{P}^{(l)}$ (l = 1, ..., L), are chosen from the standard normal distribution, with mean 0 and variance 1. Speech features 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, a speech recognition system outputs the best scoring word by itself. To obtain an optimal result from among all the results for random projection, voting is performed by counting the number of occurrences of the best word for each random-projection-based feature.

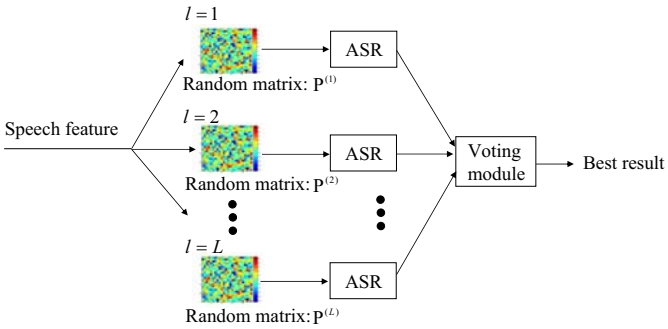
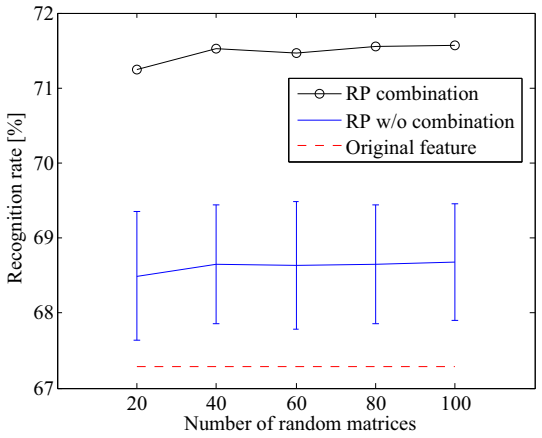


Fig. 2. Overview of vote-based random-projection combination



Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	**71.24%**	69.68%	68.49%	66.57%
40	**71.53%**	69.79%	68.65%	66.57%
60	**71.47%**	70.64%	68.63%	66.57%
80	**71.56%**	70.64%	68.64%	66.57%
100	**71.57%**	70.64%	68.68%	66.57%

Fig. 3. Random projection for MFCC (The recognition rate for the original feature is 67.28%.) The final system feature dimensionality is 12. "Max." "Mean" and "Min." in the lower table denote the max, mean, and min of the accuracies, respectively.

4. EXPERIMENTS

4.1. Experimental Conditions

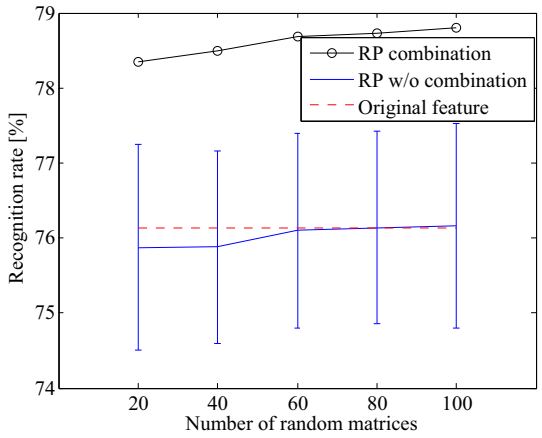
The random-projection combination method was evaluated on noisy speech recognition tasks. Noisy speech data were taken from the CENSREC-3 (Corpus and Environments for Noisy Speech RECognition) database [7]. All speech data were collected in car environments (idling, low speed, and high speed). The "condition 4" of the CENSREC-3 was used for training and test in this paper. Therefore, the training data were composed of 3,608 phonetically-balanced sentences, and the total number of speakers for training data was 293 (202 males and 91 females). The test data were composed of 8,836 utterances, and the total number of speakers for testing data was 18 speakers (8 males and 10 females). The tests were carried out on a 50-word recognition task.

The speech signal was sampled at 16 kHz and windowed with a 20-msec Hamming window every 10 msec. In the mel-filter bank analysis, a cut-off was applied to frequency components lower than 250 Hz, and the total number of dimensions of the filter-bank output was 24. In this paper, cepstral mean subtraction was applied to the MFCC-based feature vectors.

The acoustic models consist of triphone HMMs that have five states with three distributions. Each distribution was represented with 32-mixture Gaussians. The baseline system was trained using 36-dimensional feature vectors consisting of 12-dimensional MFCC parameters, along with their delta and delta-delta parameters. The baseline recognition accuracy was 76.14%.

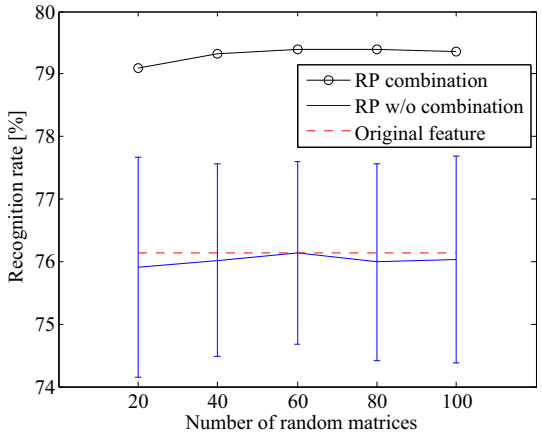
Four random-projection-based features were evaluated. Each feature description is found below.

(I) Random projection is applied to MFCC at t-th frame, $\mathbf{x}(t) \in$



Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	78.35%	**78.60%**	75.87%	73.65%
40	78.49%	**78.60%**	75.88%	73.65%
60	**78.69%**	78.60%	76.10%	73.64%
80	**78.73%**	78.60%	76.14%	73.64%
100	78.80%	**79.20%**	76.17%	72.77%

Fig. 4. Random projection for MFCC+Δ+ΔΔ (The recognition rate for the original feature is 76.14%.) The final system feature dimensionality is 36.



Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	**79.10%**	79.04%	75.91%	70.93%
40	**79.32%**	79.04%	76.02%	70.93%
60	**79.40%**	79.04%	76.14%	70.93%
80	**79.40%**	79.04%	75.99%	70.93%
100	**79.36%**	79.33%	76.03%	70.93%

Fig. 5. Random projection for MFCC (The recognition rate for the original feature is 76.14%.) The new feature also has its Δ and ΔΔ. The final system feature dimensionality is 36.

R^{12}, and the new feature, $\mathbf{y}(t) \in \mathrm{R}^{12}$, is obtained.

$$\mathbf{y}(t) = \mathbf{P}^{(l)^T}\mathbf{x}(t) \tag{2}$$

(II) Random projection is applied to MFCC + ΔMFCC + ΔΔMFCC, $\mathbf{x}(t) \in \mathrm{R}^{36}$, and the new feature, $\mathbf{y}(t) \in \mathrm{R}^{36}$, is obtained.

(III) Random projection is applied to MFCC, $\mathbf{x}(t) \in \mathrm{R}^{12}$, and the new feature, $\mathbf{y}(t) \in \mathrm{R}^{12}$ is obtained. Then the delta and acceleration coefficients of $\mathbf{y}(t)$ are calculated.

(IV) Random projection is applied to a 2-D Gabor feature (60-dimension) [8] + ΔGabor + ΔΔGabor, $\mathbf{x}(t) \in \mathrm{R}^{180}$, in the filter-bank output domain, and the new feature, $\mathbf{y}(t) \in \mathrm{R}^{30}$, is obtained.

The number of random matrices is l = 20, 40, 60, 80, and 100. For example, in the case of l = 20, 20 kinds of the new feature vectors are calculated using 20 kinds of random matrices. Then, we train the 20 kinds of acoustic models using 20 kinds of the new feature vectors. In the test process, 20 kinds of recognition results are obtained using 20 kinds of acoustic models.

4.2. Experimental Results

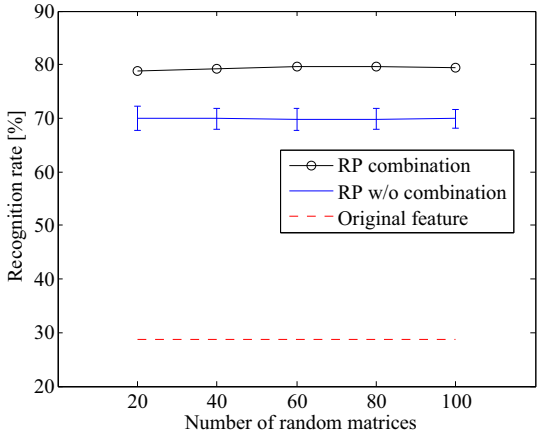
We investigated the performance of random projection for various random matrices (l = 20, 40, 60, 80, and 100) from $N(0, 1)$. Fig. 3 shows the recognition rate versus the number of random matrices for (I). The plot of "RP w/o combination" shows the means and the standard deviations of the random-projection-based features without ROVER-based combination. As shown in Fig. 3, the results of the experiment indicate that the vote-based random-projection combination improves the recognition rate from 67.28% to 71.24% using the combination of 20 random matrices, although the performance of RP (Random Projection) without combination for some random matrices was lower than the recognition rate of the original feature (MFCC). Also, even if the number of random matrices increases, we could not obtain a large performance increase in our experiments.

Fig. 4 and Fig. 5 show the performance for (II) and (III), respectively. The vote-based random-projection combination improved the recognition rate from the baseline rate, but the performance was slightly lower than the maximum recognition rate in Fig. 4. Fig. 6 shows the performance for (IV). The experiment results indicate that the vote-based random-projection combination can improve the recognition rate from the baseline and any random projection.

Table 1 shows the recognition results of the vote-based random-projection combination for each in-car condition of the CENSREC-3 database (condition 4), where speech data were recorded under 5 kinds of in-car environments (normal, with air-conditioner on (fan low/high), with audio CD player on, and with window open, and 20 random matrices are used. The recognition rate inside the ( ) indicates the baseline. The original feature in (III) is MFCC, but we compared the performance of the proposed method with that of MFCC + Δ + ΔΔ in (III) because the proposed feature in (III) was composed of the random-projection feature, its Δ, and ΔΔ. The experiment results indicate that the vote-based random-projection combination improves the average recognition rate for all noise conditions of CENSREC-3. One of the 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 [3]. More research will be needed to investigate the effectiveness of the random projection for speech features.

Table 1. Recognition rates [%] of the vote-based random-projection combination compared with the original feature. The number of random matrices was 20.

Car speed	In-car condition	(I) (MFCC)	(II) (MFCC+Δ+ΔΔ)	(III) (MFCC+Δ+ΔΔ)	(IV) (Gabor+Δ+ΔΔ)
Low speed	Normal	**87.97** (82.31)	**93.16** (91.16)	**93.40** (91.16)	**93.87** (45.05)
	Fan (low)	**85.76** (82.82)	**91.18** (89.88)	**90.59** (89.88)	**91.65** (39.06)
	Fan (high)	**72.07** (71.84)	**74.41** (72.40)	**74.97** (72.40)	**79.33** (23.46)
	Audio (on)	**61.72** (59.01)	**76.91** (73.62)	**77.15** (73.62)	**69.73** (26.86)
	Window (open)	**68.56** (64.55)	**77.26** (74.25)	**77.70** (74.25)	**76.48** (25.75)
High speed	Normal	**79.44** (70.33)	**88.44** (83.56)	**89.33** (83.56)	**90.44** (37.67)
	Fan (low)	**80.11** (73.89)	**85.56** (83.78)	**85.78** (83.78)	**88.00** (30.67)
	Fan (high)	**70.33** (68.22)	69.78 (**70.00**)	**72.89** (70.00)	**76.67** (22.11)
	Audio (on)	**57.73** (51.84)	**75.97** (73.30)	**75.75** (73.30)	**71.19** (24.58)
	Window (open)	**49.89** (49.22)	**52.23** (50.89)	**54.68** (50.89)	**52.00** (13.47)
Overall		**71.24** (67.28)	**78.35** (76.14)	**79.10** (76.14)	**78.84** (28.73)



Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	**78.84%**	73.95%	69.99%	65.86%
40	**79.30%**	73.95%	69.92%	65.86%
60	**79.59%**	74.41%	69.88%	64.76%
80	**79.57%**	74.41%	69.86%	64.76%
100	**79.48%**	74.41%	69.90%	64.76%

Fig. 6. Random projection for Gabor+Δ+ΔΔ (The recognition rate for the original feature is 28.73%.) The final system feature dimensionality is 30.

5. CONCLUSION

This paper has described a random-projection-based feature combination technique using random matrices. We can expect to find a projection matrix that gives a better speech recognition accuracy, among random matrices, since we are able to produce various random matrices. From our recognition results, it has been shown that the use of the vote-based random-projection combination provides better performance but with a high computation need. In future research, we will continue to investigate how to select the optimal basis vector from a random matrix.

6. REFERENCES

[1] N. Goel, G. Bebis, and A. Nefian, "Face recognition experiments with random projection," in *Proc. SPIE*, vol. 5779, 2005, pp. 426–437.

[2] P. Thaper, S. Guha, and N. Koudas, "Dynamic multidimensional histograms," in *Proc. ACM SIGMOD*, 2002, pp. 428–439.

[3] S. Dasgupta, "Experiments with random projection," in *Proc. UAI*, 2000, pp. 143–151.

[4] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. the 20th Int. Conf. on Machine Learning*, 2003, pp. 186–193.

[5] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: robust concepts and random projection," in *Proc. IEEE Symposium on Fundations of Computer Science*, 1999, pp. 616–623.

[6] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover)," in *Proc. IEEE ASRU*, 1997, pp. 347–352.

[7] M. Fujimoto, S. Nakamura, K. Takeda, S. Kuroiwa, T. Yamada, N. Kitaoka, K. Yamamoto, M. Mizumachi, T. Nishiura, A. Sasou, C. Miyajima, and T. Endo, "Censrec-3: An evaluation framework for japanese speech recognition in real driving car environments," in *Proc. IEEE RWCinME*, 2005, pp. 53–60.

[8] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. ICSLP*, 2002, pp. 25–28.