ACOUSTIC MODEL TRANSFORMATIONS BASED ON RANDOM PROJECTIONS

Tetsuya Takiguchi, Mariko Yoshii, Yasuo Ariki

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

Jeff Bilmes

Department of Electrical Engineering
University of Washington
Seattle WA, 98195, USA

ABSTRACT

This paper proposes a novel acoustic model transformation method for speech recognition based on random projections. Random projections have been suggested as a means of dimensionality reduction, where the original data are projected onto a subspace using a random matrix. Moreover, as we are able to produce various random matrices, it may be possible to find a transform matrix that is superior to conventional transformation matrices among random matrices. In our previous work, a random-projection-based feature combination technique has been proposed but had a high computational cost. In order to deal with this cost, in this paper, we introduce random projections on the acoustic model domain, where linear transformations are applied to an acoustic model using random matrices. Its effectiveness is confirmed by word recognition experiments on noisy speech.

Index Terms— acoustic model transformation, random projection, random matrix, model domain

1. INTRODUCTION

Random projections have 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 Principal Component Analysis (PCA), which find a subspace by optimizing certain criteria, random projections do not use such criteria; therefore, they are data independent. Moreover, they represent 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 projections have been applied to various types of problems, including information retrieval (e.g. [2]), machine learning (e.g. [3, 4]), and so on. Although based on a simple idea, random projections have demonstrated good performance in a number of applications, yielding results comparable to conventional dimensionality reduction techniques, such as PCA.

In our previous work [5], we investigated the feasibility of random projections for speech feature extraction, where a speech feature is projected using various random matrices, and the parameters of the acoustic model corresponding to each random matrix are estimated from the projected features. Experimental results showed that the random-projection-based feature combination (on the feature domain) provides better performance but with a high computation cost because of the acoustic model training process for each projected feature set.

In this paper, we introduce an approach in which random projections are carried out on the acoustic model domain instead of the feature domain, where linear transformations are applied to an original acoustic model using random projection matrices. The random projections on the model domain does not require training of the parameters of the acoustic model. It only requires a model transformation process using random matrices. The computational complexity of the model transformation is very low, unlike the random projections on the feature domain which are computationally expensive as mentioned above.

This rest of this paper is organized as follows. In section 2, random projections on the feature domain are described, and random projections on the acoustic model domain is introduced in section 3. Section 4 describes the results of experiments on a noisy speech recognition task.

2. RANDOM PROJECTIONS ON THE FEATURE DOMAIN

This section describes a feature projection (extraction) method using random orthogonal matrices [5]. The main idea of random projections 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 [6].

Random projections are a simple yet powerful technique, and it has another benefit. Dasgupta [3] has reported that even if the distributions of the 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-

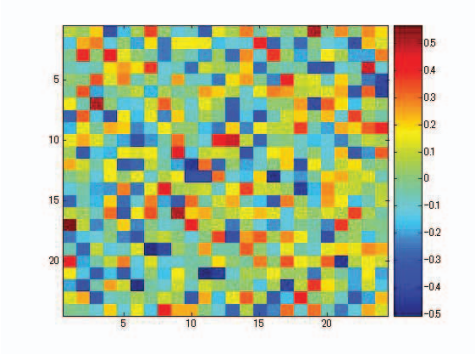

Fig. 1. An example of a random matrix

th random matrix, $\mathbf{P}^{(l)}$, where we compute a d-dimensional vector, $\mathbf{x}^{(l)}$, whose coordinates are the inner products $x_1^{(l)} = \mathbf{p}_1^{(l)} \cdot \mathbf{x}, \ldots, x_d^{(l)} = \mathbf{p}_d^{(l)} \cdot \mathbf{x}$.

$$\mathbf{x}^{(l)} = \mathbf{P}^{(l)^T} \mathbf{x} \tag{1}$$

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 [6]. In this paper, we use $N(0,1)$ for the distribution of the coordinates. The random matrix, $\mathbf{P}$, can be calculated extremely simply 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.
- Create the orthogonal matrix using the Gram-Schmidt algorithm, and then the columns are normalized to unit length.

The orthogonality may be effective for feature extraction because the hidden Markov models (HMMs) used in our experiments utilize diagonal covariance matrices.

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 was introduced in [5], where ROVER combination [7] was applied to random-projection-based features.

In the vote-based random-projection combination, first, random matrices, $\mathbf{P}^{(l)}$ ($l = 1, ..., L$), are produced as described above. Speech features are then projected using each random matrix. An acoustic model corresponding to each random matrix is also trained, as shown in Fig. 2. For the test utterance, using each acoustic model, a speech recognition system outputs the best scoring word by itself. To obtain a single hypothesis from among the different systems, voting is performed by counting the number of occurrences of the best word for each random-projection-based system.

Experimental results showed that the random-projection-based feature combination (on the feature domain) provided better performance [5] but with a high computational training cost because of the acoustic model training process needed for each projected feature, as shown in Fig. 2. In order to deal with computational cost, in the following section, we introduce random projections on the acoustic model domain (instead of the feature domain), where a linear transformation is applied to an acoustic model itself using random matrices.

3. RANDOM PROJECTIONS ON THE MODEL

Fig. 3 outlines the process of random projections on the acoustic model domain. An acoustic model for the original feature space is transformed using various random matrices. Therefore, it does not require the re-estimation process of the acoustic model in order to obtain a new acoustic model for a feature space projected using a random matrix.

In our work, an HMM is used as an acoustic model. HMMs are widely-used for speech recognition for many reasons including their tractability and their having simple maximum likelihood parameter estimation techniques. In common with most other HMM-based systems, the output probability density function is made up of a mixture of Gaus-

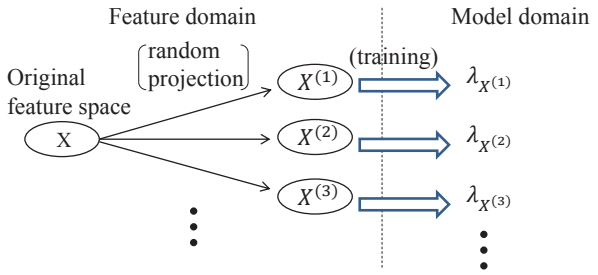

Fig. 2. Random projection on the feature domain. Acoustic models are trained for every different feature space.

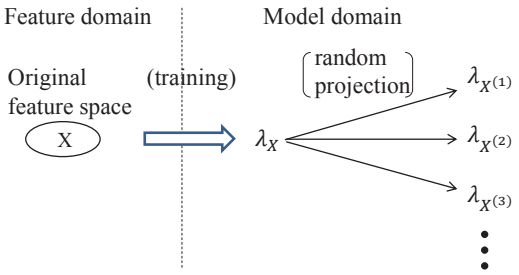

Fig. 3. Random projection on the acoustic model domain. An acoustic model is transformed using various random matrices.

sian densities as follows:

$$b(\mathbf{x}) = \sum_m c_m N(\mathbf{x}; \mu_m, \Sigma_m) \quad (2)$$

where c_m is the weight of the m-th component of a state, and $N(*; \mu, \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ.

The aim of our work is to transform the mean vectors and covariance matrices from the original space to a new space that is projected using a random matrix. For the m-th mixture component, a new estimate of the mean and covariance is found by

$$\mu_m^{(l)} = \mathbf{P}^{(l)^T} \mu_m \quad (3)$$

$$\Sigma_m^{(l)} = \mathbf{P}^{(l)} \Sigma_m \mathbf{P}^{(l)^T} \quad (4)$$

where $\mathbf{P}^{(l)}$ is the l-th random matrix, which is computed from $N(0, 1)$ using the same method as described in section 2. We used the same number of dimensions for the projected space as that of the original space in this paper. As can be seen, the computational cost of the linear transformations in Eq. (3) and (4) is very low, unlike the random-projection-based feature combination.

For the test utterances, the same low computation cost is required since it is only the features that need to be randomly transformed. To obtain a single hypothesis from among all the transformed models (Fig. 4) voting is again performed by counting the number of occurrences of the best word for each random-projection-based feature.

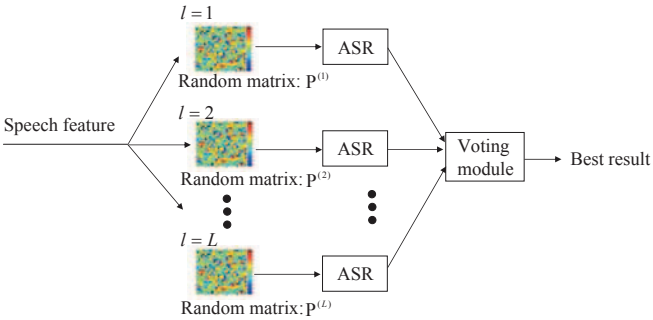

Fig. 4. Vote-based combination for decoding

4. EXPERIMENTS

4.1. Experimental Conditions

The random projections for model transformation method was evaluated on a noisy speech recognition task. Noisy speech data were taken from the CENSREC-3 (Corpus and Environments for Noisy Speech RECognition) database [8]. 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 testing in this paper. The training data were composed of 3,608 phonetically-balanced sentences, and the total number of speakers for the 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 the testing data was 18 speakers (8 males and 10 females). The tests were carried out on a 50-word recognition task.

Speech 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 the 36-dimensional feature vectors consisting of 12-dimensional MFCC parameters, along with their delta and delta-delta parameters (window lengths were ± 3 and ± 2, respectively). The baseline recognition accuracy was 76.14%. In the experiments, we used the same number of dimensions for the projected space as that of the original space.

Table 1. Random projection using model transformation. (The recognition rate for the original feature is 76.14%.)

Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	78.11%	78.27%	78.03%	77.84%
40	78.16%	78.27%	78.02%	77.78%
60	78.18%	78.27%	78.03%	77.78%
80	78.20%	78.27%	78.02%	77.72%
100	78.21%	78.27%	78.03%	77.72%

Table 2. Comparison of recognition accuracy [%] between the feature-based RP and the model-based RP

w/o RP	Feature-RP	Model-RP
76.14	78.81	78.21

4.2. Experimental Results

We investigated the performance of random projections for various random matrices (l = 20, 40, 60, 80, and 100) sampled from $N(0, 1)$. Table 1 shows the recognition rate versus the number of random matrices. The results of "RP w/o combination" shows the maximums, means, and minimums obtained from each random projection without ROVER-based combination. As shown in Table 1, experimental results indicate

that the vote-based random-projection combination improves the recognition rate from 76.14% to 78.11% using the combination of 20 random matrices, and even the minimum result of RP (Random Projection) without combination for random matrices was better than the recognition rate of the original feature. Also, even if the number of random matrices is increased, we do not show further performance increases in our experiments. This characteristic is the same as that shown in our previous work (random projection on the feature domain) [5]. Table 2 shows a comparison of recognition accuracy between the feature-based RP and the model-based RP, where 100 random matrices are used. The experiment result indicates that model-based random projection can improve the recognition rate from the baseline to basically the same degree as feature-based RP.

Table 3 shows a comparison of recognition accuracy for PCA-based features. A 2-D Gabor feature (60 dimensions) [9] is computed in the filter-bank output domain, the 36-dimensional Gabor feature is obtained from Gabor + ΔGabor + $\Delta\Delta$Gabor using PCA. Also, the 32-dimensional multiple-frame feature (e.g. [10]) is computed from 11 successive frames of 12-dimensional MFCC using PCA. As shown in this table, both random projection approaches can improve the recognition rate from the baselines. As mentioned above, one possible reason 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]. It is also interesting to note that ROVER-based combination shows benefits even if combining results from essentially the same system that has been repeatedly subjected to different random projections.

Table 3. Comparison of recognition accuracy [%] for PCA-based features

	w/o RP	Feature-RP	Model-RP
Gabor	72.39	78.45	78.21
Multiple-frame	81.09	81.44	84.73

5. CONCLUSION

This paper has described a random projection method based on acoustic model transformations. We might expect to find a projection matrix that gives a better speech recognition accuracy among random matrices since the space of likely instances is extremely diverse, more so perhaps than that of data-driven transformations. The method of random projections based on model transformation provides better performance in comparison with the baseline, just as random projections on the feature domain does. Also, the computational cost of the model based transformation is very low, while the feature domain method is computationally costly due to the need to retrain each system individually. For decoding, the same low computation cost is required since only linear transformations on the features are used. Further speedups could be obtained in practice using fast matrix-matrix multiply routines making the proposed technique feasible for large-scale speech recognition systems. In future research, we will continue to investigate how to select the optimal basis vectors via the use of such random matrices.

Acknowledgment: This research was supported in part by MIC SCOPE.

6. REFERENCES

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

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

[3] S. Dasgupta, "Experiments with random projection," in *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 *the 20th Int. Conf. on Machine Learning*, 2003, pp. 186–193.

[5] T. Takiguchi, J. Bilmes, M. Yoshii, and Y. Ariki, "Evaluation of random-projection-based feature combination on speech recognition," in *ICASSP*, 2010, pp. 2150–2153.

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

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

[8] 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 *RWCinME*, 2005, pp. 53–60.

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

[10] S. Nakagawa and K. Yamamoto, "Evaluation of segmental unit input HMM," in *ICASSP*, 1996, pp. 439–442.