社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ランダムプロジェクションを用いた音声特徴量変換

吉井　麻里子† 　滝口　哲也†† 　有木　康雄†† 　Jeff Bilmes†††

† 神戸大学大学院工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 神戸大学自然科学系先端融合研究環　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
††† Department of Electrical Engineering, University of Washington,
Box 352500, Seattle, WA 98195-2500, USA
E-mail: †mariko0901@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp,
†††bilmes@ee.washington.edu

あらまし　本稿では、ランダムプロジェクションを用いた音声特徴量変換を提案する。ランダムプロジェクションとは，次元削減の手法として従来用いられており、高次元空間における任意の 2 点間のユークリッド距離が射影先の低次元空間においてもほぼ保存される，という性質を持つ空間写像の一手法である．ランダムプロジェクションで用いる写像行列は，各成分が独立にある確率分布に従うランダムな n × k 行列として定義される．本稿では，複数のランダムマトリックスを用いて機械的に音声特徴量を変換し，各々のランダム写像に対する音声認識結果に投票を行い，最適な認識結果を求める．評価は CENSREC-3 で行い，その有効性を示す.

キーワード　音声特徴量変換，ランダムプロジェクション，ランダム写像行列，音声認識

Random-Projection-Based Feature Transformation

Mariko YOSHII†, Tetsuya TAKIGUCHI††, Yasuo ARIKI††, and Jeff BILMES†††

† Graduate School of Engineering, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
†† Organization of Advanced Science and Technology, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
††† Department of Electrical Engineering, University of Washington,
Box 352500, Seattle, WA 98195-2500, USA
E-mail: †mariko0901@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp,
†††bilmes@ee.washington.edu

Abstract　This paper proposes a novel feature transformation method for speech recognition based on random projection. Random projection has been suggested as a means of dimensionality reduction, where the original data are projected onto a subspace using a random matrix. In this paper, we investigate the feasibility of random projection for speech feature extraction. Its effectiveness is confirmed by word recognition experiments on noisy speech.

Key words　feature transformation, random projection, random matrix, speech recognition

1. はじめに

近年，音声認識システムにおいて，音声特徴量として MFCC (Mel-Frequency Cepstrum Coefficient) が広く使われている．これは，対数メルフィルタバンク出力に対して DCT (Discrete Cosine Transform，離散コサイン変換) を行うことにより得られる特徴量であり，正規化手法や特徴量の線形回帰係数である $\Delta MFCC$ や $\Delta\Delta MFCC$ と組み合わせることで，音声認識において高い認識率を示している．しかし，観測される音声信号には発話者の話者性や録音環境による環境雑音など，音声認識を行う上で必要とする情報以外の様々な情報が混

在する．このような問題に対しては MFCC では対処し切れておらず，音響モデルや言語モデルに頼らざるを得ない状況である．

他にも，このような問題を解決するために様々な特徴量が提案されてきている．特に，主成分分析 (Principal Component Analysis，PCA) [1]，判別分析 (Linear discriminantanalysis，LDA) [2]，独立成分分析 (Independent component analysis，ICA) [3] などの統計的手法をベースとした特徴量抽出手法が提案され，その効果が確認されている．これらは統計的に最適な空間を探し出し，その空間への写像を行うことで新たな特徴量を得ている．

空間写像の手法として，画像処理や文書圧縮の手法として用いられていたランダムプロジェクションというものがある [4], [5], [6], [7]．この手法は n 次元ユークリッド空間から d 次元ユークリッド空間へ写像を行う空間写像の手法であるが，その変換行列を各成分が確率的にある値をとるランダムな行列として定義している．画像処理や文書圧縮の分野では，その変換における距離保存の性質と，事前計算が不要で計算量が少ないということをメリットとしてこの手法が用いられてきた．

我々は，このランダムプロジェクションで作り出される空間がどのようなものかということに興味を持ち，音声特徴量抽出への応用を考えた．ランダムプロジェクションはその変換の容易さにも関わらず，特徴量間距離を保存し，主成分分析や判別分析，独立成分分析などと比較しても遜色ない変換を行うことが可能である．音声特徴量抽出の課題である，話者性や雑音を除く音声成分のみを得る特徴量抽出に役立てることができるのではないかと考えた．

本稿では，MFCC 等の音声特徴量とランダムプロジェクションを組み合わせた，新たな音声特徴量を得る手法を提案する．評価実験として，自動車内音声データベース CENSREC-3 [11] を用いた単語音声認識を行い，従来手法との比較を行う．ランダムプロジェクションを用いることで，雑音を含む音声信号から従来よりも音声特徴をより表現する特徴量空間が得られることを示し，ランダムプロジェクションの有効性を示す．

以降の 2 章ではランダムプロジェクション手法について解説し，3 章ではランダムプロジェクションを用いた音声特徴量変換の手法について述べる．4 章で，評価実験の条件とその結果を報告し，最後に 5 章で，結論と今後の課題について述べる．

2. ランダムプロジェクション

2.1 ランダムプロジェクション

ランダムプロジェクションは n 次元ユークリッド空間から k 次元ユークリッド空間へランダムに写像する空間写像の手法である．その式は単純で，ある n 次元の元特徴量ベクトル y が与えられたとき，k 次元 ($k \leqq n$) の変換後の特徴量ベクトル x は次のように表わされる．

$$x = Ry \tag{1}$$

ここで R は n × k の写像行列である．

このランダムプロジェクションは 以下の Johnson-Lindenstrauss lemma [8] から発想を得ている．

定理 1 ($Johnson - Lindenstrauss\ lemma$)

今，ϵ を $0 < \epsilon < 1$，n を整数として，k を次のようにおく．

$$k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-k} \ln(n) \tag{2}$$

このとき，n 次元空間 R^n から k 次元空間 R^k への空間写像を考え，空間写像を写像関数 $f : R^n \rightarrow R^d$ で表わす．R^n の任意の 2 点 u, v を考えるとき，この 2 点間はの距離は次のように保存される．

$$(1-\epsilon)||u-v||^2 \leqq ||f(u)-f(v)||^2 \leqq (1+\epsilon)||u-v||^2 \tag{3}$$

この定理は，n 次元空間から $O(logn/\epsilon^2)$ 次元の空間へ写像するとき，ある 2 点間のユークリッド距離が極めて高い確率 (係数 $(1 \pm \epsilon)$) で保存されることを示している．さらに, [9] により，この写像関数 f は任意のランダムな値によって得られることが分かっている．

2.2 ランダム写像行列 R

ランダム写像行列 R は，各成分が確率的にある値をとる行列として定義されるが，各成分が単に標準正規分布 $N(0,1)$ に従って独立に選ばれることによって定まるランダムな行列が，上記距離保存の性質を持つことが証明されている．これらの写像行列はデータに依存せず，高速に得ることができる．

本稿では，次のようにランダム写像行列 R を設定する．

- 標準正規分布 $N(0,1)$ に従う要素を持つ n × k の行列 R を作成する．
- グラムシュミットの直交化手法を用いて R を直交化し，列ベクトルを大きさ 1 で正規化する．

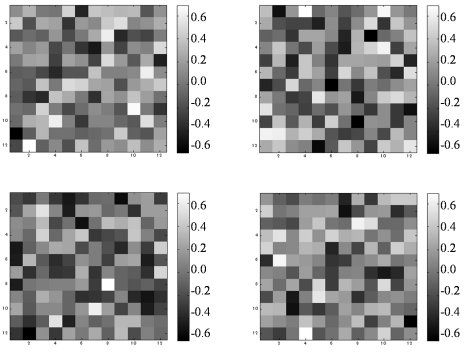


図 1 Example of random matrices 12 dim. (12 × 12)

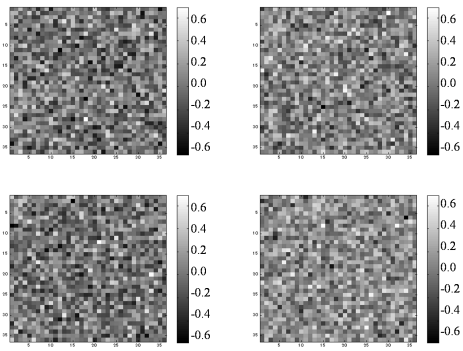


図 2　Example of random matrices 36 dim. (36 × 36)

ランダム写像行列 R は標準正規分布 $N(0,1)$ から無限に得ることができる．図 1 ，2 に，本稿で用いたランダム写像行列 R の例を示す．

3.　音声特徴量抽出

2 章で示したように，ランダム写像行列 R は標準正規分布 $N(0,1)$ から無限に生成できる．従って，ランダムプロジェクションを用いて音声特徴量抽出を行う際，無限のランダム写像行列からもっとも音声認識に適したランダム写像行列を見つけるか，もしくは複数のランダム写像行列から最適な音声認識結果を求める必要がある．本稿では最適な音声認識結果を得るために，ROVER [10] を用いた特徴量統合の手法を検討する．

3.1　音声特徴量変換

ROVER とは，複数の音声認識システムから得た認識結果に対して投票を行い，最適な認識結果を出力する手法である．ランダム写像行列を複数用いて音声認識を行い，その結果を投票により統合することで，個々のシステムから得たものよりも精度のよい音声認識結果が得られると考えられる．

図 3 に，提案する特徴量統合の手法の流れ図を示す．上図では，基本的なランダムプロジェクションを用いた音声特徴量抽出手法を示している．まず，あらかじめランダム写像行列を標準正規分布 $N(0,1)$ から生成する．次に音声特徴量を入力として，得られた写像行列を用いたランダムプロジェクションを行い，新たな音声特徴量を得る．その新たな特徴量で音声認識を行い，認識結果を得る．

ランダム写像行列は複数 (無限に) 生成できるため，本稿では図 3 下に示す統合手法を用いる．複数得られたランダム写像行列を用いて特徴量抽出を行い，音声認識を行った後に ROVER を用いて統合を行う．このように統合することで，ランダム写像行列に優劣をつけることなく，最適な音声認識結果を得ることが可能となる．

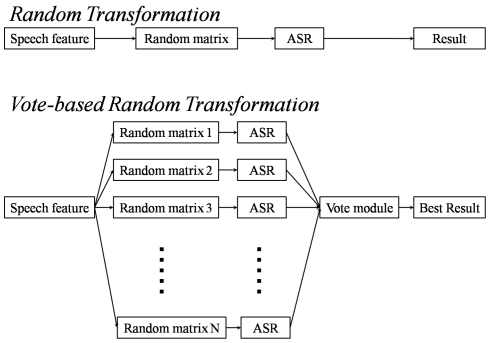


図 3　Overview of Random Transformation

3.2　音声特徴量

本稿では，様々な音声特徴量とランダムプロジェクションを組み合わせることによって，新たな音声特徴量を生成する．

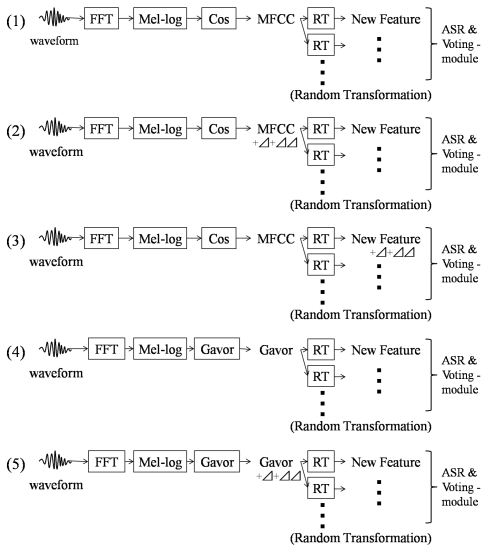


図 4　Block diagram of random-transformation-based features

図 4 では本稿で用いる音声特徴量とランダムプロジェクションの組み合わせ手法のブロック線図を示す．

(1) では，音声波形に対して高速フーリエ変換を行い，その対数メルフィルタバンク出力に対して DCT を行った MFCC 特徴量を入力としている．MFCC 特徴量に対して複数のランダムプロジェクションを行い，複数の新しい音声特徴量を得ている．(2) では同じく MFCC 特徴量と，その線形回帰係数である Δ と ΔΔ を入力特徴量として用いている．(3) では，(1) でランダムプロジェクションを行うことによって得られた音声特徴量に対して，その線形回帰係数 Δ ， ΔΔ を計算している．(2) と比べることにより，MFCC 特徴量の $\Delta, \Delta\Delta$ と，ランダムプロジェクションから得られた特徴量に対する $\Delta, \Delta\Delta$ の比較を行うことができる．(4) では 2-D Gavor 特徴量 [12] を音声特徴量として用いている．これは，時間-周

波数軸上での音響特性の変化を表現した特徴量で，音声の時間方向の音響特性が 60 次元の特徴量で表わされる．(5) は 2-D Gavor 特徴量と，その Δ， ΔΔ を同時に用いている．

これらの特徴量とランダムプロジェクションを組み合わせることで，新たな音声特徴量を抽出し，音声認識を行う．

4. 評価実験

4.1 実験条件

提案手法の評価を行うために，自動車内音声認識の評価用データベース CENSREC-3 [11] を用いて単語音声認識実験を行う．音声認識評価環境には Condition 4 を用い，その学習データはアイドリング走行時の遠隔マイクロホン音声 3608 発話 (男性 202 名，女性 91 名)，評価データは低速，高速走行時の遠隔マイクロホン音声 8836 発話 (男性 8 名，女性 10 名) である．評価用の音声データは 50 単語からなり，学習データは音素バランス文となっている．

音声の標本化周波数は 16kHz，語長 16bit であり，音響分析には Hamming 窓を使用した．フレーム幅，シフト幅はそれぞれ 20ms，10ms である．また，自動車雑音特有の低周波成分に対処するため，メルフィルタバンク分析時に 250kHz 以下の低周波成分を取り除いている．対数メルフィルタバンク特徴量の次元数は 24，MFCC 特徴量の次元数は 12，Gavor 特徴量の次元数は 60 である．それぞれの特徴量はあらかじめ平均 0，分散 1 に正規化しておく．

音響モデルは音素 HMM で，各 HMM の状態数は 5，状態あたりの混合分布数は 32 である．

5 種類のランダムプロジェクションを用いた特徴量変換の手法を表 1 に示す．それぞれ図 4 の (1), (2), (3), (4), (5) と対応している．

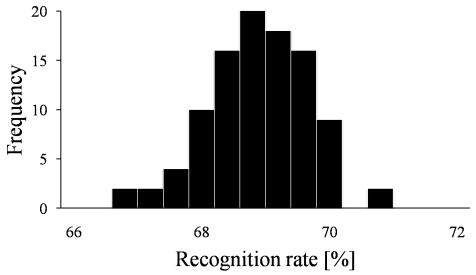
変換に用いるランダム写像行列はそれぞれ 100 個用い，事前に作成した．

表 1　Feature Transformation using Random Projection

(1)	$MFCC$(12 次元)$->RP$(12 次元)
(2)	$MFCC+\Delta+\Delta\Delta$(36 次元)$->RP$(36 次元)
(3)	$MFCC$(12 次元)$->RP$(12 次元)$+\Delta+\Delta\Delta$(36 次元)
(4)	$Gavor$(60 次元)$->RP$(30 次元)
(5)	$Gavor+\Delta+\Delta\Delta$(180 次元)$->RP$(30 次元)

4.2 単語音声認識実験結果

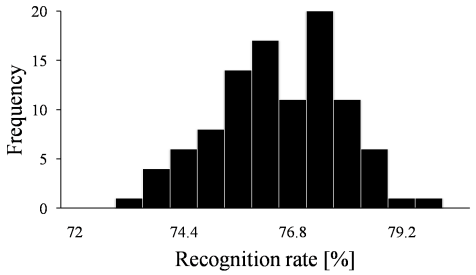
(1) の音声認識結果を図 5 に示す．ROVER を用いた特徴量統合の手法では，従来の MFCC 特徴量を用いた場合の認識率 67.28 % から 71.57 % まで認識率が改善した．しかしながら，100 個のランダム写像行列の中にはいくつか従来の MFCC 特徴量の認識率よりも劣るものも存在した．



Random transformation				Baseline
Vote	Max.	Mean	Min.	
71.57%	70.64%	68.68%	66.57%	67.28%

図 5　100 trials of random transformation for MFCC

図 6 は (2) の結果を示している．(1) の場合と同じく，従来の MFCC + Δ + ΔΔ 特徴量の場合 (76.14 %) と比べて 2.7 % 認識率が改善された．この場合は，特徴量統合の場合よりもある一つのランダム写像行列を用いた場合の認識率 (79.20 %) の方がよりよい結果となっている．



Random transformation				Baseline
Vote	Max.	Mean	Min.	
78.81%	79.20%	76.17%	72.77%	76.14%

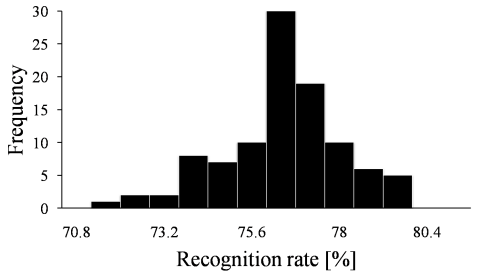
図 6　100 trials of random transformation for MFCC +Δ+ ΔΔ

図 7 は (3) の結果を示している．MFCC に対しての一次微分，二次微分だけでなく，ランダムプロジェクションを行った特徴量に対しても，一次微分，二次微分が有効であるということを示している結果となった．この場合も特徴量統合の場合よりもある一つのランダム写像行列を用いた場合の認識率の方がわずかに良い結果となっている．

図 8, 9 は (4), (5) の結果を示している．これらは 2-D Gavor feature を元の特徴量に使ったものであるが，変換前の次元数が大きいものから低次元に変換 (圧縮) を行った場合も，ランダムプロジェクションが有効であることが示されている．元特徴量での認識率と比べて，ランダムプロジェクション特徴量統合によって大幅に認識率が改善していることがわかった．

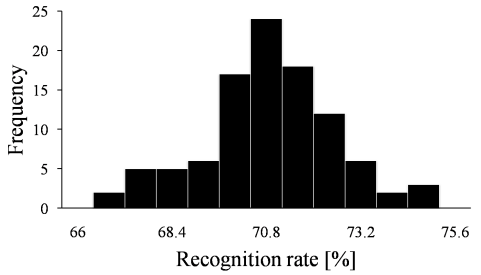
表 2　Recognition rates [%] of the vote-based random-transformation combination compared with the baseline

		(1)	(2)	(3)	(4)	(5)
Car speed	In-car condition	(MFCC)	(MFCC $+\Delta+\Delta\Delta$)	(MFCC $+\Delta+\Delta\Delta$)	(Gavor)	(Gabor $+\Delta+\Delta\Delta$)
Low speed	Normal	88.21 (82.31)	94.22 (91.16)	93.87 (91.16)	92.92 (85.50)	92.81 (45.05)
	Fan(low)	86.24 (82.82)	90.82 (89.88)	90.82 (89.88)	89.88 (82.35)	90.24 (39.06)
	Fan(high)	72.63 (71.84)	74.41 (72.40)	74.97 (72.40)	77.21 (67.71)	78.32 (23.46)
	Audio(on)	62.54 (59.01)	77.03 (73.62)	78.09 (73.62)	67.14 (53.24)	68.43 (26.86)
	Window(open)	68.78 (64.55)	77.15 (74.25)	78.48 (74.25)	72.13 (63.10)	74.58 (25.75)
High speed	Normal	79.78 (70.33)	88.67 (83.56)	88.33 (83.56)	89.22 (80.33)	89.67 (37.67)
	Fan(low)	80.11 (73.89)	86.89 (83.78)	85.89 (83.78)	86.00 (77.00)	87.44 (30.67)
	Fan(high)	70.33 (68.22)	71.33 (70.00)	73.56 (70.00)	73.11 (64.00)	75.22 (22.11)
	Audio(on)	57.95 (51.84)	76.20 (73.30)	76.31 (73.30)	69.86 (56.84)	71.64 (24.58)
	Window(open)	50.33 (49.22)	52.78 (50.89)	53.90 (50.89)	51.11 (41.98)	51.34 (13.47)
Overall		71.57 (67.28)	78.81 (76.14)	79.29 (76.14)	76.75 (67.10)	77.87 (28.73)



Random transformation				Baseline
Vote	Max.	Mean	Min.	
79.29%	79.33%	76.03%	70.93%	76.14%

図 7　100 trials of random transformation for MFCC. The new feature also has its Δ and ΔΔ .



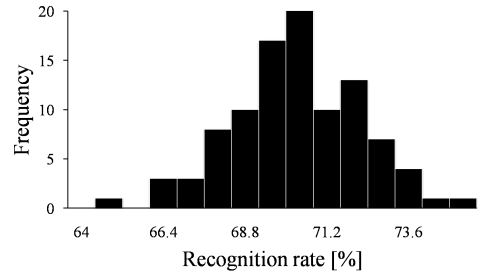
Random transformation				Baseline
Vote	Max.	Mean	Min.	
76.75%	74.68%	70.43%	66.24%	67.10%

図 8　100 trials of random transformation for Gabor

図 2 は，CENSREC-3 における様々な車内雑音環境ごとの認識率を示している．括弧内の数値が従来手法である．どの雑音環境においてもランダムプロジェクションによって認識率の改善が見られる．ランダムプロジェクションによって，音声特徴量空間が変化し，認識に適した空間が作られていると考えられる．

4.3　ランダムプロジェクションによる音声特徴量空間

4.2 の結果から，ランダムプロジェクションによって得られた空間は音声認識に適していることが考えられる．一般的に，特徴量空間は，多次元の球面状に特徴量が分布していると認識に適していると言われているが，ラ



Random transformation				Baseline
Vote	Max.	Mean	Min.	
77.87%	74.41%	69.90%	64.76%	28.73%

図 9　100 trials of random transformation for Gabor + Δ + ΔΔ

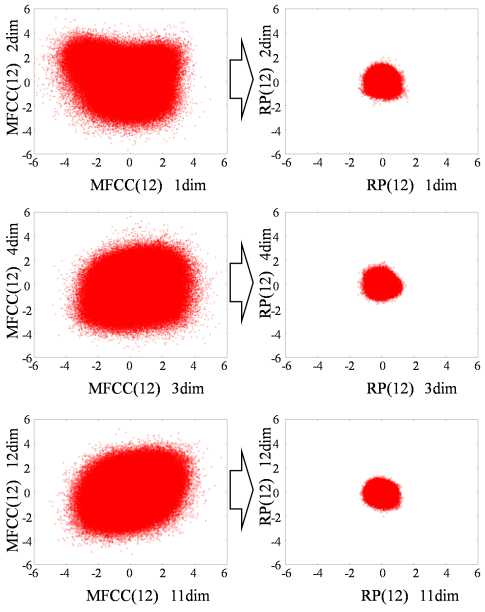


図 10　Distribution of MFCC-space and RP-space

ンダムプロジェクションによってどのような空間が形成されているのかを調べてみた．

図 10 は，MFCC 特徴量と 提案手法 (1)$MFCC->RP$ 特徴量の特徴量分布を二次元ごとに記録した図である．これらの図から，ランダムプロジェクションを行うことにより，分散が小さく，より球に近い形状に特徴量が集中していることがわかる．このことが，認識率が改善している理由の一つの説明になるのではと考えられる．

5. おわりに

本稿では，ランダムプロジェクションを用いた新しい特徴量抽出手法について提案した．我々は複数のランダムマトリックスを用いて機械的に音声特徴量を変換し，各々のランダム写像に対する音声認識結果に投票を行うことで，最適な認識結果を求めた．ランダムプロジェクションは単純で容易に得られる写像行列から空間写像を行うにも関わらず，認識に有益な特徴量空間を生成することが示された．今後は，多くのランダムプロジェクション特徴量を統合する方法をさらに考えると共に，有用なランダム写像行列とそうでないランダム写像行列を見分ける方法や，一つのランダム写像行列の中から，認識に適した次元を選び出す手法を考えていく必要がある．より有効なランダムプロジェクションの活用方法を提案していきたい．

文　献

[1] T. Takiguchi and Y. Ariki,“ PCA-Based Speech Enhancement for Distorted Speech Recognition, ” Journal of Multimedia, Vol. 2, Issue 5, pp. 13-18, 2007.

[2] S. S. Kajarekar, B. Yegnanarayana, and H. Hermansky,“ A study of two dimensional linear discriminants for ASR, ” Proc. ICASSP, Vol. 1, pp. 137-140, 2001.

[3] O. W. Kwon, T. W. Lee,“ Phoneme recognition using ICA-based feature extraction and transformation, ” Signal Processing, Vol. 84 (6), pp. 1005-1019, 2004.

[4] N. Goel, G. Bebis, and A. Nefian,“ Face Recognition Experiments with Random Projection, ” Proc. of the SPIE, Vol. 5779, pp. 426- 437, 2005.

[5] S. Dasgupta,“ Experiments with random projection, ” in Uncertainty in Artificial Intelligence, pp. 143-151, 2000.

[6] X. Z. Fern and C. E. Brodley,“ Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, ” Proc. of the 20th Int. Conf. on Machine Learning, pp. 186-193, 2003.

[7] R. I. Arriaga and S. Vempala,“ An algorithmic theory of learning: robust concepts and random projection, ” Proc. IEEE Symposium on Fundations of Computer Science, pp. 616-623, 1999.

[8] W. B. Johnson and J. Lindenstrauss,“ Extensions of Lipshitz Mapping into Hilbert Space, ”in Conference modern analysis and probability, volumn 26 of Contemporary Mathematics, pp. 189-206, 1984.

[9] S. Kaski, “ Dimensionality Reduction by Random Mapping, ”Proc. Int. Joint Conf.On Neural Networks, pp. 413-418, 1998.

[10] J. G. Fiscus,“ A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER), ”Proc. IEEE ASRU Workshop, pp. 347-352, 1997.

[11] 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, ”Proc. RWCinME, pp. 53-60, 2005.

[12] M. Kleinschmidt and D. Gelbart,“ Improving Word Accuracy with Gabor Feature Extraction, ” Proc. ICSLP, pp. 25-28, 2002.