情報処理学会研究報告
IPSJ SIG Technical Report

ランダムプロジェクションを用いた音響モデルの線形変換

吉井　麻里子[†1]　滝　口　哲　也[†2]　有　木　康　雄[†2]

本稿では，ランダムプロジェクションを用いて音響モデルの線形変換を行い，複数の特徴量を用いた音声認識を効率良く行う手法を提案する．ランダムプロジェクションとは，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高い確率で保存される，という性質を持つ空間写像の一手法である．また，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 $n \times k$ 行列として定義される．本稿では音声特徴量をランダムプロジェクションを用いて変換し，ランダムプロジェクション特徴量を作成するが，得られた特徴量で音響モデルを学習するのではなく，変換前の特徴量で学習した音響モデルに対してランダムプロジェクションを行うことで，特徴量ごとの音響モデルを低コストで作成する．評価実験は CENSREC-3 を用いた単語音声認識を行い，提案手法の有効性を示す．

Acoustic Model Adaptation using Random Projection

Mariko Yoshii,[†1] Tetsuya Takiguchi[†2]
and Yasuo Ariki[†2]

This paper proposes a novel model adaptation 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. Moreover, as we are able to produce various random matrices, there may be some possibility of finding a transform matrix that is superior to conventional transformation matrices among these random matrices. In this paper, we adapt linear transformation to an acoustic model using random projection. Its efectiveness is confirmed by word recognition experiments on noisy speech.

†1 神戸大学大学院 自然科学研究科
Graduate School of Science and Technology, Kobe University.
†2 神戸大学 自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University.

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離散コサイン変換 (Discrete Cosine Transform, DCT) を行うことで得られる事前学習無しの特徴量であり，正規化手法や特徴量の線形回帰係数 Δ MFCC や ΔΔ MFCC と組み合わせることで高い音声認識性能を示している．しかし，観測される音声信号には音素時系列情報だけではなく，発話者の身体や感情の情報，環境情報など様々な情報が混在する．このような情報の中から，音声認識に必要な音素時系列情報を取り出すという問題に対しては 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，このような問題に対処するために様々な音声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)]，線形判別分析 (Linear Discriminant Analysis, 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機械学習の分野で提案され，画像処理や文書圧縮の手法として用いられてきたランダムプロジェクションという手法がある[4)–7)]．この手法は n 次元ユークリッド空間から $n \geq k$ の k 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点に特徴を持つ．また，ランダムプロジェクションには変換前と変換後である 2 点間の距離が高い確率で保存されるという性質や，事前計算が不要で計算量が少ないという特徴がある．

我々は以前，このランダムプロジェクションを用いた音声特徴量抽出を提案した．雑音環境下において，ランダムプロジェクション変換前の特徴量を用いた音声認識性能よりもランダムプロジェクションを行って得た音声特徴量を用いたときの音声認識性能の方が高くなり，雑音に対して従来より頑健な音声特徴量を得ることができた．さらに，無限に得られるランダム写像行列を用いて音声特徴量を生成し，複数の音声特徴量の統合を行うことで安定して高い認識性能を示すことができた．しかしながら，複数の音声特徴量からそれぞれの音響モデルを学習することになり，学習コストが大きくなるという問題があった．

本稿では，複数の音響モデルをそれぞれの特徴量で学習するのではなく，変換前の音声特

徴量を用いて学習した音響モデルに対してランダムプロジェクションを行い，変換後の音声特徴量を用いて認識を行い，複数の特徴量の統合を行う手法を試みる．音響モデルに対してランダムプロジェクションを行い，学習コストの低減と共に音声認識性能の向上を目指す．評価実験としては，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．

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抽出と，ランダムプロジェクションを用いた音響モデルの線形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 ($k \leq n$) 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 $f : R^n \rightarrow R^k$ 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．

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

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．

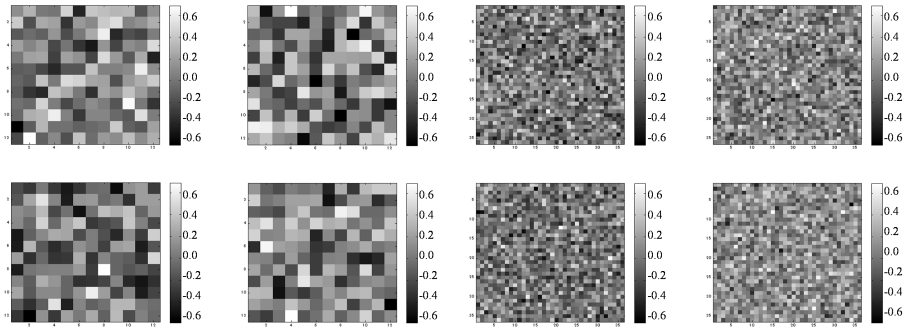

図 1 12 次元ランダム写像行列の例
Fig. 1 Examples of RM, 12 dim.

図 2 36 次元ランダム写像行列の例
Fig. 2 Examples of RM, 36 dim.

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．これは，写像後の k 次元ベクトルの各成分が，同じく正規分布 $N(0,1)$ に従って分布する「正規分布の再生性」に基づくものである．また，{-1,+1} 上の一様分布 U(-1,+1) に従う成分を持つランダムな行列も距離保存の性質を持つことが示されている．一般にはどのような分布を用いれば得られる行列が距離保存の性質を持つのか明らかではない．これらのランダム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，4)，5) で用いられている，次のような方法でランダム写像行列 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．

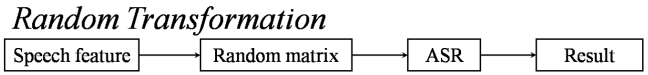

図 3 ランダムプロジェクションを用いた音声特徴量変換
Fig. 3 Overview of Random Transformation on speech feature

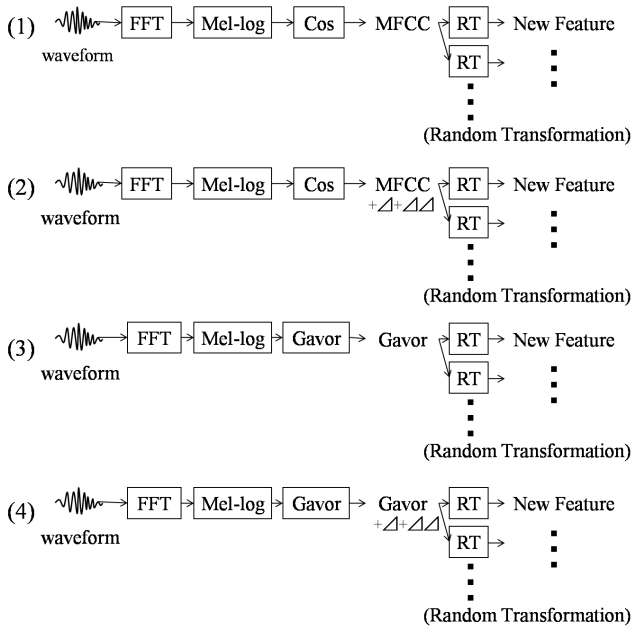

図 4 本稿で用いる音声特徴量変換のブロック線図
Fig. 4 Block diagram of random-transformation-based features evaluated in this paper

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

3. 提 案 手 法

3.1 ランダムプロジェクションを用いた音声特徴量抽出

ランダムプロジェクションを用いた音声特徴量変換では，**図 3** のように音声特徴量に対して式 (1) を用いてランダムプロジェクションを行い，得られた音声特徴量を用いて音声認識を行う．

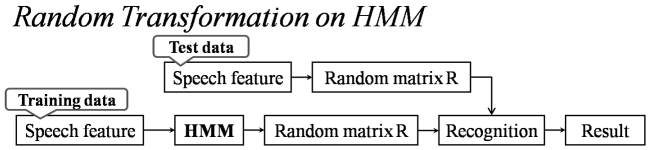

図 5 ランダムプロジェクションを用いた音声特徴量と HMM の変換
Fig. 5 Overview of Random Transformation on features and HMM

図 4 には本稿で用いる音声特徴量を示している．(1) は MFCC に対してランダムプロジェクションを行っている．(2) では MFCC とその線形回帰係数である ΔMFCC，ΔΔMFCC を組み合わせた特徴量に対してランダムプロジェクションを行っている．(1)，(2) では元の特徴量次元数を変換後も変えず，同次元の空間写像を行う．(3) 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，60 次元の特徴量で表わされている．(4) は 2-D Gavor 特徴量とその Δ，ΔΔ を組み合わせた特徴量で，180 次元のものとなる．(3) と (4) では次元が大きくそのままでは音声認識に適さないため，ランダムプロジェクションを用いて次元削減を行う．

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する．

3.2 音響モデルの線形変換を用いた音声特徴量統合

ランダム写像行列 R は無限に生成することができ，ランダムプロジェクション特徴量を用いた音声認識ではランダム写像行列 R によって認識率にばらつきが生じる．我々は以前，複数のランダム写像行列を用いて複数のランダムプロジェクション特徴量を生成し，それらを ROVER[10] を用いて統合することで安定して高い認識率を得る手法を提案した．しかしながら，複数得られた音声特徴量を用いて複数回音響モデルを学習する必要があったため，学習コストが高くなるという問題があった．

本稿では得られた特徴量ごとに音響モデルを学習するのではなく，あらかじめランダムプロジェクションを行う前の音声特徴量で音響モデルを学習しておき，その音響モデルに対してランダムプロジェクションを行うことで複数回音響モデルを学習することなくランダムプロジェクション特徴量での認識を可能にする．

図 5 に特徴量と音響モデルの変換の流れ図を示す．あらかじめ変換前の学習データで音

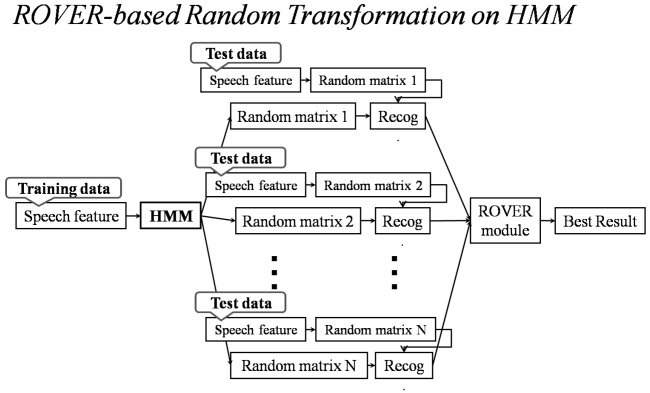

図 6 ROVER を用いた提案手法の流れ図
Fig. 6 Overview of Random Transformation on features and HMMs, and conbine them using ROVER module

響モデルを学習する．そして，認識を行うデータに対してランダムプロジェクションを行いランダムプロジェクション特徴量を得ると同時に，学習された音響モデルに対しても同様のランダム写像行列でランダムプロジェクションを行い，得られた音響モデルで認識を行う．

図 6 では，複数のランダム写像行列を用いて変換したランダムプロジェクション特徴量を ROVER を用いて統合する方法を示す．同じ元の音響モデルに対してそれぞれのランダム写像行列でランダムプロジェクションを行い，同種のランダム写像行列で得られた音声特徴量で認識を行う．得られた認識結果を ROVER を用いて統合することで最適な認識結果を得る．

このように音響モデルに対してランダムプロジェクションを行うことで，特徴量ごとに音響モデルを学習するコストを省き，効率良く有効な音声認識を行うことが可能となる．

3.3 ランダムプロジェクションを用いた音響モデルの線形変換

本稿では音響モデルに HMM(Hidden Markov Model) を用いる．HMM は混合正規分布で表現されているため，各々の正規分布の平均値ベクトルと分散共分散行列に対してランダムプロジェクションを行う．正規分布の平均と分散は，特徴量ベクトルを用いて式 (4)，式 (5) のように表現される．

$$\mu_x = \frac{1}{N}\sum_{i=1}^{n} x_i. \tag{4}$$

$$\Sigma_x = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu_x)(x_i-\mu_x)^T. \tag{5}$$

従って，それぞれの特徴量ベクトルに対して式 (1) を用いて以下の式 (6)，式 (7) のように変換できる．

$$\begin{aligned}\mu_y &= \frac{1}{N}\sum_{i=1}^{n} Rx_i \\ &= R\mu_x.\end{aligned} \tag{6}$$

$$\begin{aligned}\Sigma_y &= \frac{1}{N}\sum_{i=1}^{n}(Rx_i-R\mu_x)(Rx_i-R\mu_x)^T \\ &= R\Sigma_x R^T.\end{aligned} \tag{7}$$

このようにすべての分布に対してランダムプロジェクションを行い，音響モデルの変換を行う．

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

表 1 ランダムプロジェクションを用いた音声特徴量変換
Table 1 Feature Transformation using Random Projection

(1)	$MFCC(12dim.) -> RP(12dim.)$
(2)	$MFCC + \Delta + \Delta\Delta(36dim.) -> RP(36dim.)$
(3)	$Gavor(60dim.) -> RP(36dim.)$
(4)	$Gavor + \Delta + \Delta\Delta(180dim.) -> RP(36dim.)$

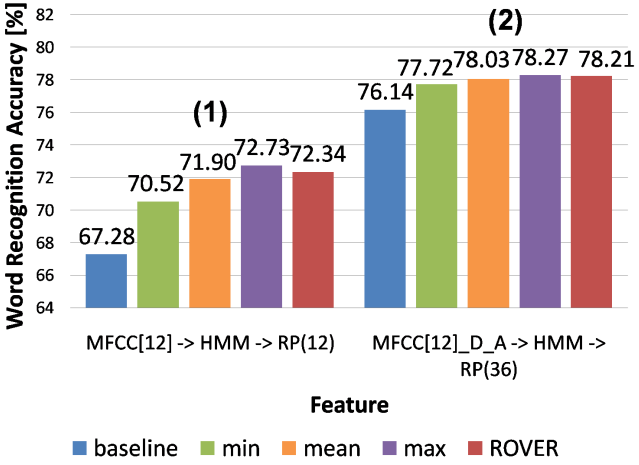

図 7 MFCC，MFCC+Δ + ΔΔ に対するランダムプロジェクションの単語音声認識結果
Fig. 7 Word Accuracy of Random Transformation for MFCC and MFCC+Δ + ΔΔ

おく．

音響モデルは音素の triphone-HMM で，各 HMM の状態数は 3，状態あたりの混合分布数は 32 である．

4 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) 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 種類で，事前に作成した．

4.2 単語音声認識実験結果

図 7 に (1)，(2) の音声認識結果を示す．それぞれ，MFCC 特徴量から得られた HMM

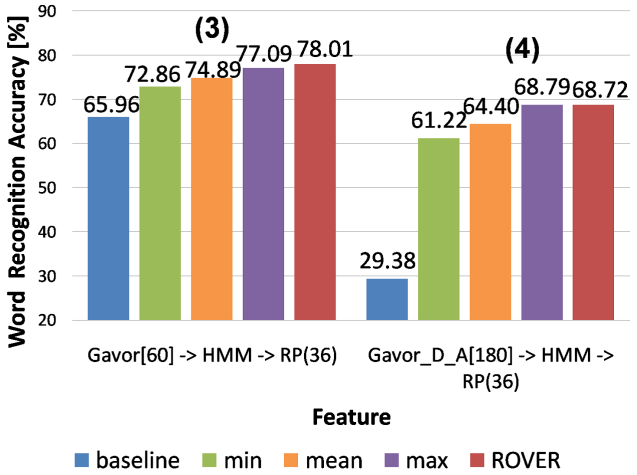

図 8 Gavor，Gavor+Δ + ΔΔ に対するランダムプロジェクションの単語音声認識結果
Fig. 8 Word Accuracy of Random Transformation for Gavor and Gavor+Δ + ΔΔ

に対してランダムプロジェクションを行ったものと，MFCC+Δ + ΔΔ からなる HMM に対してランダムプロジェクションを行ったものである．100 種類のランダム写像行列を用いたときの最小，平均，最大認識率や，ROVER を用いた際の認識率をみると従来手法よりも高い認識率が得られることがわかった．

図 8 は (3)，(4) の結果を示している．(3) は Gavor 特徴量に対してランダムプロジェクションを行ったもので，Gavor 特徴量 60 次元で学習した HMM に対して，ランダムプロジェクションで 36 次元に次元削減を行ったものである．同じく (4) では Gavor+Δ + ΔΔ の 180 次元で学習した HMM に対してランダムプロジェクションで 36 次元に次元を落としている．これらの結果も変換前の認識率と比べるとランダムプロジェクションを用いることで音声認識性能が改善していることがわかる．

また，ランダムプロジェクションでは無限の写像行列を生成することができるが，ランダム写像行列の種類数による認識率の比較を**表 2**，**表 3** に示している．統合するランダム写像行列数を 20，40，60，80，100 と変えてみたところ，特徴量数の増加によって認識率の向上が見られた．しかしながら表 2 は MFCC に対するランダムプロジェクションの結果だが，20 種類の ROVER 統合でも変換前の従来 MFCC の認識率 67.28 % を上回る認識率と

表 2 MFCC-HMM に対するランダムプロジェクションの，統合特徴量数による認識率の比較
Table 2 Comparison of recognition accuracy for the number of random matrices.

Numver of Random Matrices	RP based on ROVER	RP		
		Min.	Mean	Max.
20	72.14	70.86	71.80	72.49
40	72.37	70.86	71.92	72.49
60	72.30	70.86	71.89	72.49
80	72.32	70.86	71.88	72.62
100	72.34	70.52	71.89	72.73

表 3 MFCC+Δ + ΔΔ-HMM に対するランダムプロジェクションの，統合特徴量数による認識率の比較
Table 3 Comparison of recognition accuracy for the number of random matrices.

Numver of Random Matrices	RP based on ROVER	RP		
		Min.	Mean	Max.
20	78.11	77.84	78.03	78.27
40	78.16	77.78	78.02	78.27
60	78.18	77.78	78.03	78.27
80	78.20	77.72	78.02	78.27
100	78.21	77.72	78.03	78.27

なった．同様に，表 3 の従来 MFCC+Δ + ΔΔ の認識率は 76.14 % で，こちらも 20 種類の ROVER による認識率 78.11 % の方が良い結果となった．ROVER の統合では 20 種類程度で十分な認識が得られることがわかった．

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音響モデルの線形変換について提案した．音響モデルに対してランダムプロジェクションを行うことで，複数特徴量を統合する際の学習計算量を大幅に減らし，なおかつ自動車内雑音環境下で従来の認識率よりも高い性能を得ることができた．

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．

参 考 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.