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，滝口哲也，有木康雄 (神戸大)

1　はじめに

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，複数の音声特徴量を用いた音声認識を効率良く行う手法について提案する．ランダムプロジェクションとは，元の特徴量空間における任意の 2 点間のユークリッド距離が，写像先の特徴量空間においても高い確率で保存される，という性質を持つ空間写像の手法で，その写像行列の各要素がある確率分布に従うランダムな値として定義される点に特徴を持つ．我々は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の研究を行ってきた [1]．本稿では複数のランダム写像行列を用いて音声特徴量を変換すると同時に，変換前の音声特徴量で学習した音響モデルに対しても同様のランダム写像行列で線形変換を行うことで，各々の音声特徴量に対する音響モデルを低コストで実現する．さらに，複数の音響モデルから得た認識結果を投票により統合することで最適な認識結果を得る手法を報告する．

2　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．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．

$$x = Ry$$

ここで R は n × k の写像行列である．ランダムプロジェクションでは，任意の 2 点間の距離が高い確率で $(1 \pm \varepsilon)$ に収まることが証明されている $(0 \leq \varepsilon \leq 1)$．また，写像行列 R は確率的にある値をとる行列として定義されるが，R の各要素が N(0,1) に従うランダムな値からなるとき，その距離保存の性質が成り立つことが証明されている [2, 3]．本稿では [4] に示さ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．

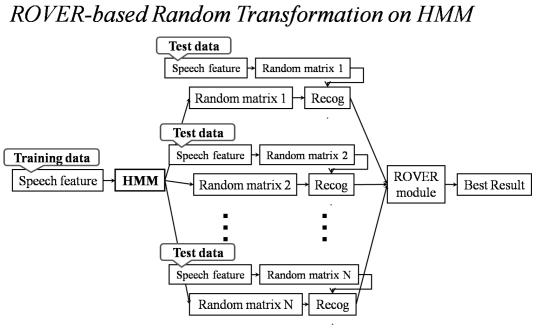

Fig. 1　Overview of Random Transformation on features and HMMs, and combine them using ROVER module

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．

3　ランダム写像行列を用いた音響モデルの線形変換

2 章で示した手法により，ランダム写像行列 R は無限に生成することができる．我々は，複数のランダムプロジェクション特徴量を生成し，ROVER[6] 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，複数の音声特徴量からそれぞれ音響モデルを学習する必要があるため，学習コストが高くなるという問題があった．本稿では，特徴量ごとに音響モデルを学習するのではなく，ランダムプロジェクションを行う前の音声特徴量で学習した音響モデルの平均値ベクトルと共分散行列に対して，ランダム写像行列を用いた線形変換行うことで，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手法を提案する．

Fig. 1 に本稿で提案する統合システムの流れ図を示す．あらかじめ変換前の学習データで音響モデルを学習する．次に，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線形変換を行い，得られた音響モデルで認識を行う．これを複数のランダム写像行列に対して同様に行い，複数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．

---

*Acoustic Model Adaptation using Random Projection, by Mariko Yoshii, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

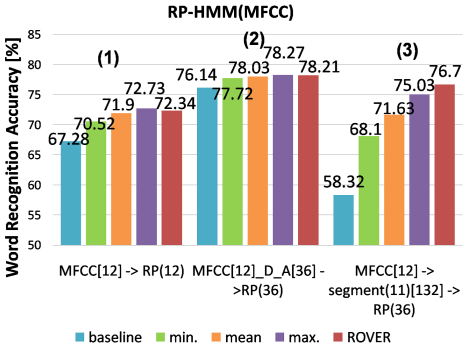

Fig. 2　Word recognition rate of MFCC-RP

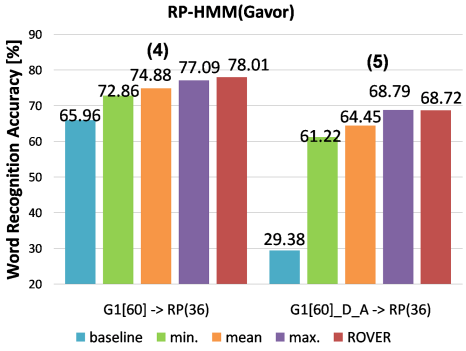

Fig. 3　Word recognition rate of Gabor-RP

4　評価実験

4.1　実験条件

提案手法の有効性を示すために，自動車内音声認識の評価用データベース CENSREC-3[7] 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 Condition4 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 3,608 発話，評価データは低速，高速走行時の遠隔マイクロホン音声 8,836 発話である．評価対象語彙数は 50 単語からなり，学習音声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，音響モデルは音素 triphone-HMM である．また，各 HMM の状態数は 3，状態あたりの混合分布数は 32 である．

本稿で用いる音声特徴量を以下に示す．(1) は，MFCC に対して，(2) は MFCC+Δ+ΔΔ に対してランダムプロジェクションを行っている．(3) は MFCC を複数フレーム並べ，セグメント化を行い，その特徴量に対してランダムプロジェクションを行い次元を削減している．(4) は 2-D Gabor 特徴量 [5] に対して，(5) はそれに Δ,ΔΔ を組み合わせたものに対してランダムプロジェクションを行い次元削減を行っている．

4.2　単語認識実験

単語認識実験の結果を Fig. 2, 3 に示す．それぞれ，左端にベースラインの変換前の認識率，中央にランダム写像行列を 100 種類用いた最小認識率，平均認識率，最大認識率を示し，右端の値が ROVER を用いて特徴量統合を行った際の認識率である．どの特徴量に対しても，ランダムプロジェクションを行うことで元の特徴量による認識率よりも高くなっている．また，ROVER を用いて統合を行うことで安定して高い認識率が得られている．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も特徴量に対するものと同様の結果を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に要する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特徴量を用いた認識率よりも高い性能を得ることができた．今後は有用なランダム写像行列を見分ける方法や，音声認識に適したランダム写像行列を生成する方法を考えていきたい．

参考文献

[1] 吉井，他，Random Projection を用いた音声特徴量抽出における Random Matrix の統合，音講論 (秋)，pp. 159-160，2009.

[2] S. Kaski. “Dimensionality reduction by random mapping,” In Proc. Int. Joint Conf. on Neural Networks, volume 1, pp. 413-418, 1998.

[3] E. Bingham, H. Mannila, “Random projection in dimensionality reduction: applications to image and text data,” In Proc. of the seventh ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, pp. 245-250, 2001.

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

[5] M. Kleinschmidt and D. Gelbart, “Improving word accuracy with Gabor feature extraction,” ICSLP, 2002.

[6] 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.

[7] 藤本，他，実走行車内単語音声データベース CENSREC-3 と共通評価環境の構築，第 55 回音声言語情報処理研究会 (SLP), pp. 41-46, 2005.