

様々な Random 行列を用いた構音障害者の音声特徴量抽出*

☆片岡悠一郎, 吉岡利也, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

本研究では, アテトーゼ型の脳性麻痺による構音障害者を対象とした音声認識の実現を目指している. アテトーゼ型の構音障害者の発話スタイルは健常者と大きく異なり, 認識精度が著しく低下する.

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一手法で, その変換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.

提案手法では, 様々な分布から作成したランダム写像行列を用いて音声特徴量を変換することで, 認識結果がどのように変化するかを見る. 各々の特徴量を用いて音声認識を行い, 各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.

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 は次のように表される.

$$x = Ry$$

ここで R は $n \times k$ の写像行列である. 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, 任意の2点間の距離が高い確率で $(1 \pm \epsilon)$ に収まることが証明されている ($0 \leq \epsilon \leq 1$) [1, 2]. 本稿では次のようなルールでランダム写像行列 R を作成する.

- 標準正規分布 $N(0,1)$ に従う要素を持つ $n \times k$ の行列 R を作成
- Gram-Schmidt の直交化手法を用いて R を直交化
- 列ベクトルを大きさ1で正規化

本来, ランダムプロジェクションでは写像行列を標準正規分布 $N(0,1)$ から生成するが, 本稿では, 他の分布であるガンマ分布, ベータ分布から写像行列を生成することで, 認識精度にどのような変化が見られるかを調べる.

3 ROVER による結果統合手法

ROVER とは, 複数の音声認識システムの出力結果に対して投票を行い, 最適な認識結果を選択する手法である. Fig. 1 に統合システムの流れ図を示す. まず, 入力音声から任意の音響特徴量を抽出する. その後, 複数のランダム写像行列を用いて特徴量変換を行い, 各々の特徴量で音響モデルの学習, 認識を行う. そして, 各認識結果を ROVER によって統合する. このように投票形式で結果統合を行うことで, ランダム写像行列に優劣をつけることなく最適な認識結果が得られる.

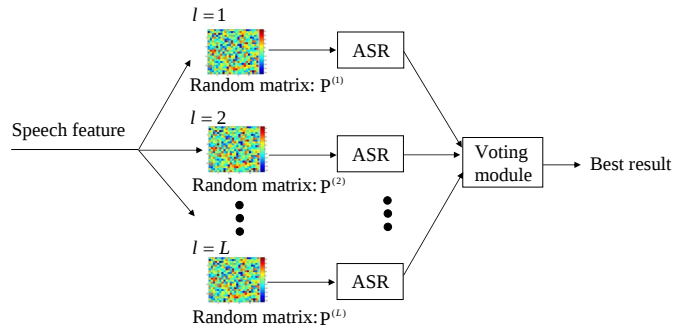


Fig. 1 Overview of vote-based random-projection combination.

4 評価実験

ガンマ分布について, (形状パラメータ, スケールパラメータ) = (1,1), (2,1), ベータ分布について, $(\alpha + \beta) = (0.5, 0.5)$, (2,2) の4種類の分布から写像行列を生成し, 標準正規分布から写像行列を作成した場合の認識率と比較する. このとき, ガンマ分布2種類は左右非対称の確率密度関数であり, ベータ分布は左右対称の確率密度関数である.

4.1 実験条件

構音障害者4名を対象とした孤立単語認識実験を行った. 実験データとして, 構音障害者4名が発話する ATR 音素バランス単語を用いた. 各単語は3回連続発話されている. 音声の標準化周波数は 16kHz であり, 音響モデルは monophone-HMM で, 各 HMM の状態数は5, 状態あたりの混合分布数は8である. ここでは, 第1発話を評価データ, 第2, 3発話を学習データとして用いる.

*Speech feature extraction of dysarthria using various random matrices, by Yuichiro Kataoka, Toshiya Yoshioka, Tetsuya Takiguchi, Yasuo Ariki (Graduate School of System Informatics, Kobe University)

本稿で用いる音声特徴量を以下に示す.

- MFCC[12] to RP[12] + Δ MFCC[12] : 12 次元 MFCC 特徴量を 12×12 のランダム写像行列で変換し, Δ 特徴量を組み合わせた 24 次元特徴量

ランダム写像行列の数は, 10 から 100 まで 10 刻みに変化させる. MFCC[12] + Δ MFCC[12] を用いた場合をベースラインとする.

4.2 実験結果

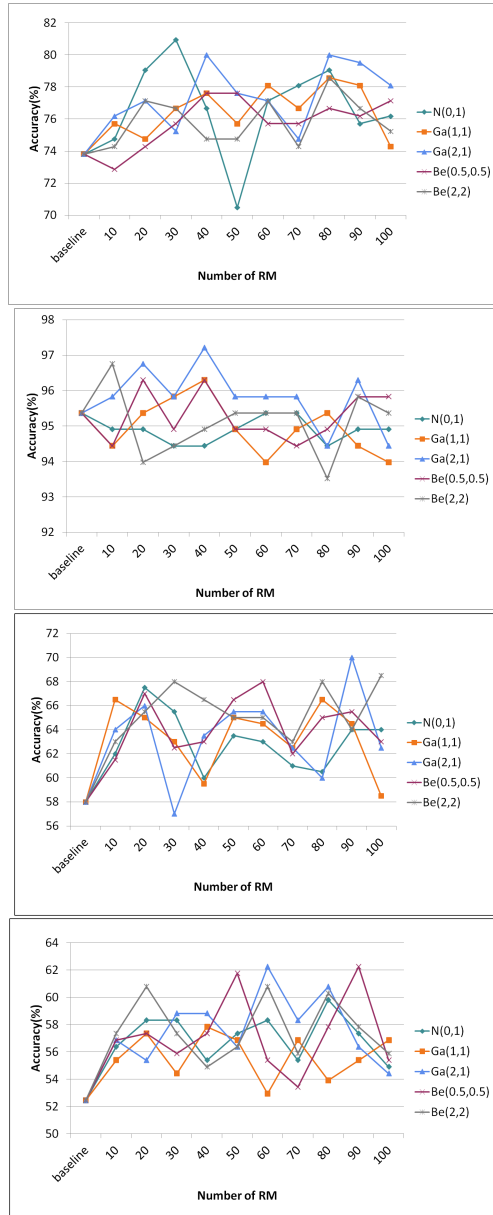


Fig. 2 Recognition results

Fig. 2 に構音障害者 4 名の認識精度のグラフを示す. 上のグラフから順に発話者 A, B, C, D とすると, B についてはベースラインの認識精度が 100% に近いので, ランダムプロジェクション後の精度が落ちているが, その他の話者については, ランダムプロ

ジェクションにより, 精度が向上している. 注目すべき点は, 分布を変化させても精度がほとんど変化しない点である. 特にガンマ分布の確率密度関数は左右非対称のグラフであるが, 他の結果と比べて変化は見られなかった.

5 おわりに

本稿では, ランダムプロジェクションにおいて, 分布を変化させた場合の構音障害者の音声認識実験を行った. 実験結果より, 分布を変化させた時も, 標準正規分布を用いた場合と認識精度は変化しないという結果が得られた. これは, 興味深い結果であるといえる. 今後は引き続き, より認識に有効なランダム写像行列の生成手法を検討する.

参考文献

- [1] S. Kaski. “Dimensionality reduction by random mapping,” In Proc. Int. Joint Conf. on Neural Networks, volume 1, pp. 413-418, 1998.
- [2] 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.
- [3] T. Takiguchi, J. Bilmes, M. Yoshii, and Y. Ariki, “Acoustic model transformations based on random projections,” In ICASSP, pp. 1933-1936, 2012.
- [4] 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.