ランダムプロジェクションを用いた構音障害音声の認識および誤り単語検出*

☆吉岡利也，高島遼一，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．彼らは意図的な動作時や緊張状態にある場合に筋肉の制御が難しくなり，アテトーゼと呼ばれる不随意運動を伴う．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，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，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．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．さらに，その投票結果に基づく正誤判定手法を紹介する．

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．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 $(1 \pm \varepsilon)$ に収まることが証明されている $(0 \leq \varepsilon \leq 1)$[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 × 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

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　ROVER による結果統合手法

我々は，ランダムプロジェクションを用いた雑音に頑健な音声特徴量変換手法について研究を行ってきた [3]．ランダムプロジェクションによって複数の音声特徴量を生成し，それらを最適に統合することで安定して高い認識率を得る．本稿では，ROVER[4] を用いた結果統合手法を適用する．

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 に本稿で提案する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．

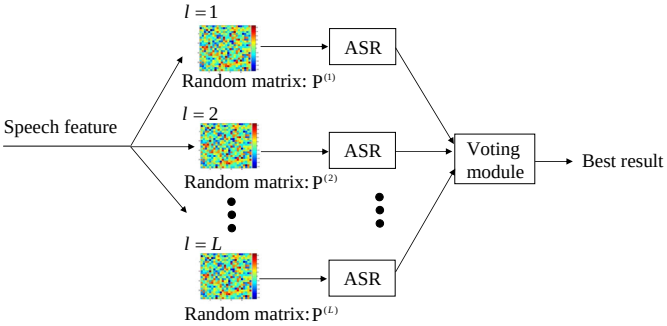

Fig. 1　Overview of vote-based random-projection combination.

4　投票結果に基づく正誤判定手法

ここでは，ROVER による投票結果に基づいた，認識結果の正誤判定手法を紹介する．投票によって選ばれた認識結果が正解の場合，少数の候補に票が集まりやすい．一方で，不正解の場合，認識結果が多様になるため，複数の候補に票がばらける傾向がある．そこで，各単語に対して正解か誤りかをラベルとし，投票結果の第 1 候補および第 2 候補の投票数を入力データとして分類モデルを作成する．

*Dysarthric speech recognition and word accuracy determination using random projection, by Toshiya Yoshioka, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Graduate School of System Informatics, Kobe University)

Table 1　Word-recognition results for a person with an articulation disorder using the proposed method. (The recognition rate for the original feature is 76.67%.)

Number of random matrices	Proposed method	RP w/o combination		
		Max.	Mean	Min.
20	80%	77.62%	75.05%	71.43%
40	79.05%	78.1%	75.14%	70.95%
60	79.52%	79.05%	75.04%	70%
80	**80.48%**	79.05%	75.15%	70%
100	**80.48%**	79.05%	75.21%	70%

5　評価実験

5.1　実験条件

構音障害者 1 名を対象とした孤立単語認識実験，および正誤判定実験を行った．実験データとして，構音障害者 1 名が発話する ATR 音素バランス単語 (210 単語) を用いた．各単語は 5 回連続発話されており，合計 1,050 単語を使用する．音声の標本化周波数は 16kHz，語長 16bit 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 2〜5 発話を学習データとして用いる．

5.2　単語認識実験

本稿で用いる音声特徴量を以下に示す．

- MFCC[12] to RP[12] + ΔMFCC[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 20, 40, 60, 80, 100 と変化させる．MFCC[12] + ΔMFCC[12] を用いた場合をベースラインとする．単語認識実験の結果を Table 1 に示す．Table 1 より，平均認識率（Mean）ではベースラインに及ばなかったが，ROVER を用いて認識結果を統合することで，安定して高い認識率が得られることが示された．また，統合特徴量数を増やすことで認識率が上がる傾向があるが，元特徴量と比べて，20〜40 個程度の統合で十分な認識率が得られている．

5.3　正誤判定実験

まず，分類モデルを作成するために第 2〜5 発話に対して提案手法を適用する．このとき，第 2 発話を評価する際は，第 3〜5 発話で音響モデルを学習する．これを各発話毎に行う．各単語の投票結果を用いて非線形 SVM を学習し，第 1 発話に対して正解か誤りかの分類を行う．

Table 2　The performance of the word accuracy determination using voting results for the 1st utterance of articulation disorders.

Number of random matrices	T/P Rate	T/N Rate	Acc.
20	0.923	0.429	0.824
40	**0.958**	0.432	0.848
60	0.928	**0.605**	**0.862**
80	0.929	0.585	**0.862**
100	0.923	0.561	0.852

実験結果を Table 2 に示す．Table 2 より，正解単語に関しては高い精度で分類できているが，不正解単語に関しては分類精度が大きく低下している．この原因として，不正解単語のサンプル数の不足，分類モデル作成のための入力データの次元数が少なすぎるといったものが考えられる．

6　おわりに

本稿では，ランダムプロジェクションを用いた構音障害者の音声認識手法を提案した．複数のランダム写像行列を用いて特徴量を変換し，各々のランダム写像行列に対する認識結果に対して投票を行うことで，最適な認識結果を求めた．実験結果より，ROVER による結果統合によって従来より安定して高い認識率が得られることが示された．また，その投票結果に基づく正誤判定手法を提案し，正解単語に対して高い分類性能を示した．今後は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.