# Random Projection を用いた構音障害者の音声認識*

☆高塚智敬，滝口哲也，有木康雄 (神戸大)，李義昭 (追手門大)

## 1 はじめに

音声認識技術は近年の発展に伴い，様々な環境下や場面での利用が期待されている．例えばカーナビゲーションの操作や会議音声の議事録化などへの応用も進んでいる．しかしこれらの多くは健常者を対象としており，より必要としている言語障害者の利用は想定していない．文献 [1, 2] では，言語障害音声を対象とした特徴量抽出や音響モデル適応と構築を行っているが，言語障害に関する研究はまだまだ少ない．そこで我々は言語障害者を対象とした音声認識の実用化を目指した研究を行なっている．

言語障害の原因は種々あるが，その一つとして脳性麻痺が考えられている．意図的な動作を行う場合や緊張状態にある場合に筋肉の制御が難しくなり，不随意運動を伴う．本研究ではこの運動障害により正しく構音できない言語障害を特に構音障害と呼ぶ．構音障害は発話を不安定にするため，人にとっても機械にとっても音声認識が困難である．そこで本研究では，脳性麻痺による構音機能障害を持つ被験者 (構音障害者) を対象に音声認識精度の改善を目指している．

我々は，不随意運動を伴う構音時の音声を雑音環境下の音声と仮定し，雑音除去手法を用いた構音障害者の音声認識を考えた．そこで文献 [3] で効果を示されている，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に注目した．本稿では，この手法を構音障害音声に適用した結果を報告する．

## 2 Random Projection

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．

$$x = R^{\mathrm{T}} y$$

ここで R は $n \times k$ の写像行列である．ランダムプロジェクションでは，射影元の空間における任意の 2 点間の距離が高確率で射影先の空間においてもほとんど保存される．射影先での距離は射影元での距離の $1 \pm \varepsilon$ 倍 ($0 \leq \varepsilon \leq 1$) 以内に収束する [4]．写像行列 R は各要素が標準正規分布 N(0,1) に従うランダム値から作成される [5]．本稿では次の手順でランダム写像行列 R を生成した．

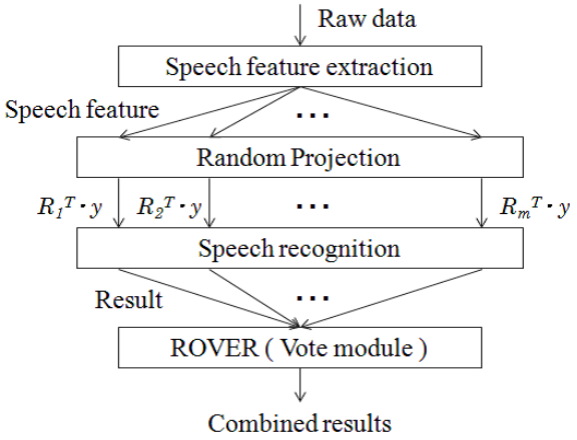


Fig. 1 System overview

(1) 標準正規分布 N(0,1) に従うランダム値のみを要素に持つ $n \times k$ 行列 R を作成

(2) Gram-Schmidt の直交化手法を用いて R を直交化

(3) 列ベクトルを大きさ 1 で正規化

このようにしてランダム写像行列 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

ランダムプロジェクションによって複数生成された音声特徴量を最適に統合することで認識精度の向上が見込まれる．その方法として次の 2 つが考えられる．1 つは無限のランダム写像行列から音声認識に最適なものを発見する方法で，もう 1 つは複数のランダム写像行列による認識結果を統合して最適な結果を求める方法である．本稿では後述の手法の一つである ROVER[6] 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．Fig. 1 に本稿で提案する統合システムの流れ図を示す．まず任意の音声特徴量を抽出する．そして，それらに複数のランダム写像行列を用いて，特徴量変換を行う．各々の特徴量から得られる認識結果を，ROVER によって統合する．このように ROVER の結果統合を用いることによって，ランダム写像行列の音声認識に対する有効性の評価を必要とせずに最適な認識結果を得ることが可能である．

*Dysarthric speech recognition using random projection, by Norihiro Takatsuka, Tetsuya Takiguchi, Yasuo Ariki (Kobe Univ.), Li Ichao (Otemon Univ.)

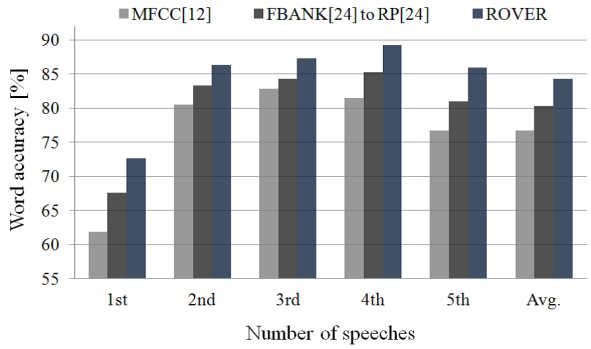


Fig. 2 Word accuracy for MFCC (baseline), FBANK[24] to RP[24] (max), and ROVER

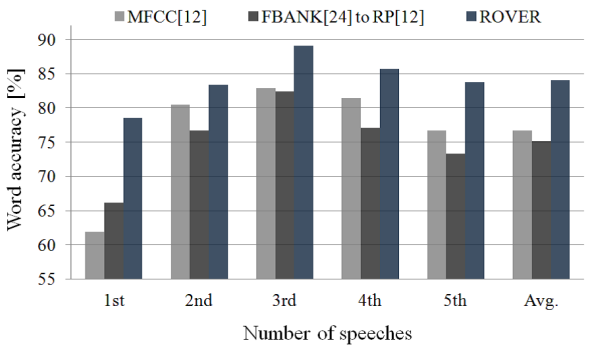


Fig. 3 Word accuracy for MFCC (baseline), FBANK[24] to RP[12] (max), and ROVER

## 4 評価実験

提案手法を評価するために孤立単語認識実験を行った．実験データには構音障害者が発声する ATR 音素バランス単語 (210 単語) を用いた．ただし各単語は連続で 5 回発話されており，実験では合計 1,050 単語を使用する．その他音声データの詳細は Table 1 に示す．評価は 5 回発話の各発話に対するクロスバリデーションで行う (例えば 1 回目の発話を認識するために 2 ～ 5 回目の発話で学習を行う)．また，ランダム写像行列は 100 個用意した．ランダムプロジェクションを行った各音声特徴量での推定単語を用いて ROVER による結果統合を行った．音声特徴量には以下の 2 つを用いた．また，特徴量 (1)，(2) を用いた実験結果をそれぞれ Fig. 2 と Fig. 3 に示す．

(1) FBANK[24] to RP[24] : 24 次元対数メルフィルタバンクを 24 × 24 のランダム写像行列で変換した 24 次元特徴量

(2) FBANK[24] to RP[12] : 24 次元対数メルフィルタバンクを 24 × 12 のランダム写像行列で圧縮した 12 次元特徴量

実験結果より一回目の発話の認識率が他と比べ低いことがわかる．これは，初めに意図的な動作を行う場合の不安定な筋肉制御に由来するものだと考えられる．ランダムプロジェクションによって生成された特徴量では一回目の発話において認識率がベースラインを上回っている．さらに他の発話と比べて大きく認識率が向上していることが分かる．これは，不随意運動を伴う構音時の音声が雑音を含んだ音声であると捉えた場合，ランダムプロジェクションの雑音頑健性を示す結果となっている．

また ROVER による結果統合は認識率の向上に大きく貢献した．Fig. 3 の一回目の発話ではベースラインに比べ 16.7 ポイントもの改善が見られた．

Table 1 Experiment condition

| | |
|---|---|
| 評価データベース | 構音障害者 1 名 (男性) の 210 単語 × 5 回発話 |
| サンプリング周波数 | 16 kHz |
| フレーム窓長 | 25 msec |
| フレーム周期 | 10 msec |

## 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構音障害者の音声認識手法を提案した．ランダムプロジェクションにより生成された音声特徴量を用いて音声認識を行い，その各々の認識結果を投票 (ROVER) によって最適な認識結果として求めた．実験結果より，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．

## 参考文献

[1] 中村 他, “発話障害者音声を対象にした健常者音響モデルの適応と検証,” 音講論 (秋), 109-110, 2005.

[2] H. Matsumasa et al., “Integration of Metamodel and Acoustic Model for Speech Recognition,” Interspeech2008, 234-2237, 2008.

[3] 吉井, 他, “ランダムプロジェクションを用いた音声特徴量抽出,” 音講論 (春), pp. 159-160, 2009.

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

[5] 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.

[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.