ランダムプロジェクションを用いた音声特徴量抽出*

◎吉井 麻里子，滝口 哲也，有木 康雄 (神戸大)，Jeff BILMES(University of Washington)

1　はじめに

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もほぼ保存される，という性質を持つ空間写像の手法である．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ランダムな n × k 行列として定義される．本稿では，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手法を提案し，従来よりも音声特徴をより表現する特徴量空間を作成する．評価は CENSREC-3 を用いた個別単語認識で行い，その有効性を示す．

2　提案手法

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．このランダムプロジェクションの有用な性質として，n 次元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 k 次元空間においても，高い確率で保存される，という性質がある．ランダムプロジェクションは，次の式で表わされる．

$$x = Ry$$

ここで，y は n 次元の元特徴量ベクトル，x は k 次元の変換後の特徴量ベクトル，R は n × k のランダム写像行列である．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，距離保存の性質を持つことが証明されている．本稿では，ランダム写像行列 R として標準正規分布 N(0,1) に従う値を要素に持つ n × k の行列を用いる．

ランダムプロジェクションはその特徴量間保存の性質から，文書圧縮や文書検索，画像処理の分野で用いられてきた [1, 2].

2.2　音声特徴量抽出

ランダムプロジェクションは，高次元空間における特徴量間の類似度が保存されるという性質を持つ．一般に，認識を行う際の特徴量は次元が多いほど多くの情報量を持つが，特徴量次元が大きすぎると認識器にかける際次元の呪いによって認識率が低下する．本研究では多くの情報量を高確率で保存したまま次元を変換することができるランダムプロジェクションを用いて，新たな音声特徴量空間を生成する．

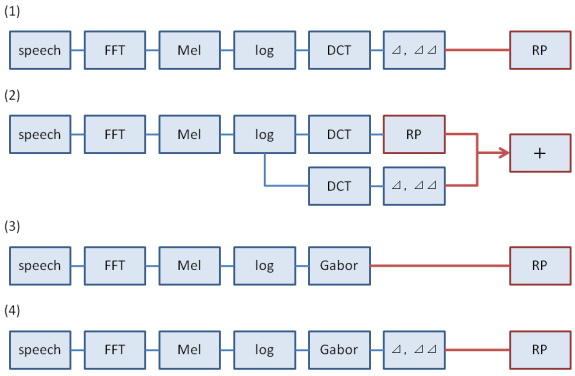

Fig. 1　特徴量抽出流れ図

最もよく用いられている音声特徴量として MFCC(Mel-Frequency Cepstrum Coefficient) があげられる．MFCC は短時間フーリエ変換で得られた周波数を，人間の聴覚特性を模したメル軸に変換しフィルタバンクを行ったものの対数をとり，離散コサイン変換したものである．MFCC は時間方向の一次微分や二次微分の値と組み合わせることで時間方向の関係性も記述することができる．

また，画像処理でよく用いられる手法の一つとして 2-D Gabor Filtering がある [3]．これは，画像で言うと特定方向のエッジを抽出する処理にあたり，音声特徴量として用いることで，時間方向の特徴量を記述できる．

これらの音声特徴量をランダムプロジェクションを用いて統合し，より多くの情報量を持った音声特徴量空間を生成する．Fig. 1 に提案手法の流れ図を示す．

(1) は，MFCC とその一次微分，二次微分に対してランダムプロジェクションを行っている．(2) は，MFCC 出力に対してランダムプロジェクションを行ったものと，MFCC の一次微分，二次微分の値を組み合わせた特徴量である．(3) は，対数メルフィルタバ

*Speech Feature Extraction Using Random Projection, by Mariko YOSHII, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University), Jeff BILMES (University of Washington)

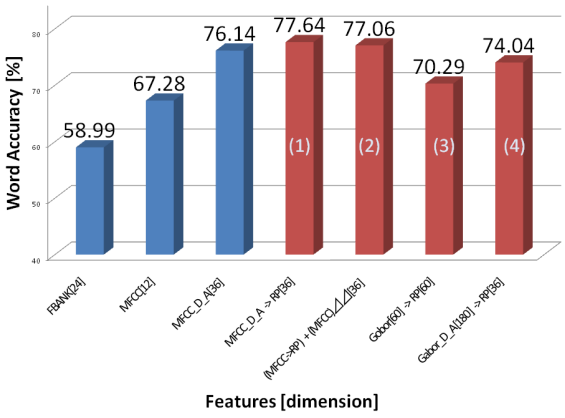

Fig. 2　単語認識率

ンク出力に対して Gabor Filtering を行ったものに対し，ランダムプロジェクションを行っている．(4) は，(3) の Gabor Filtering とその一次微分，二次微分を計算したものに対して，ランダムプロジェクションを行っている．

3　評価実験

3.1　コーパスについて

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 [4] を用いて単語認識実験を行う．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，音響モデルは音素 HMM である．また，各 HMM の状態数は 5，状態あたりの混合分布数は 32 である．

3.2　単語認識実験

単語認識実験の結果を Fig. 2 に示す．グラフ左 (青色部分) はそれぞれ対数メルフィルタバンク出力，MFCC 出力，MFCC+Δ+ΔΔ 出力となっており，ベースラインの特徴量を用いた認識率を示す．グラフ右 (赤色部分) が提案手法 (1) から (4) の特徴量を用いた単語認識結果を示している．Fig. 2 の右側 (1) から (4) が Fig. 1 の提案手法 (1) から (4) と対応している．

すべての提案手法において，変換前の認識率よりは良くなる結果となっているが，(3)，(4) の場合は，MFCC+Δ+ΔΔ 出力には及ばなかった．

また，ランダムプロジェクションを行う際のランダム写像行列は標準正規分布 N(0,1) に従う乱数を要素に持つが，作成する度に異なる写像行列が作られる．そこで，数種類のランダム写像行列を生成し，ランダムプロジェクションを行ったところ，Table 1 のように認識率が変化した．このことから，ランダム写像行列をうまく設定することにより，より音声特徴を表現する空間が生成できるのではないかと考えられる．

Table 1　ランダム写像行列による認識率の変化 [%]

	max	mean	min
(1)	77.64	75.918	75.11
(2)	77.06	76.522	75.6
(3)	70.29	67.81	64.18
(4)	74.04	70.722	67.87

4　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声特徴量抽出を行い，単語認識実験によりその有用性を示した．ランダムプロジェクションは，ランダムな写像行列から認識に有用な空間を生成できる．また，次元削減時の特徴量間類似度も高い確率で保存されることが証明されている．今後は，様々な特徴抽出手法とランダムプロジェクションを組み合わせることによる音声特徴量の統合を考えていき，より音声認識に適した空間を生成していきたい．

参考文献

[1] S. Kaski. ”Dimensionality reduction by random mapping”, In Proc. Int. Joint Conf. on Neural Networks, volume 1, p.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, p.245-250, 2001.

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

[4] 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), 2005-SLP-55, p.41-46,S 2005.