

MDL 基準と ICA を用いた統合音素部分空間による音声特徴量抽出の検討*

◎朴玄信, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

近年の音声認識システムにおいて、音響モデルの学習データの収録環境と認識時の環境条件間のミスマッチによるパフォーマンスの低下を軽減するため、信号処理、特徴量抽出、音響モデル、言語モデル、デコーダーなど様々な側面からのアプローチがなされている。特徴量抽出において MFCC はよく使われているが、この特徴量はミスマッチの環境では性能の低下が激しい。そのために統計的手法である主成分分析や独立成分分析を用いた、データ依存型特徴量抽出手法が提案されていて、ミスマッチの環境でパフォーマンス向上が得られている。

我々は音素ごとの部分空間を推定し統合する手法を提案してきた [1]。本論文では、MDL 基準を用いた部分空間推定 [2] を用いて各音素部分空間を推定し、ICA を用いた音素部分空間統合による音声特徴量抽出法を説明する。残響の単語音声認識実験の結果を報告し提案手法の有効性を示す。

2 提案手法

Fig. 1 (a) は MFCC と提案手法の特徴量抽出の流れを示している。MFCC は対数メルフィルタバンク出力に離散コサイン変換 (DCT) を行うことで得られるが、提案する特徴量は対数メルフィルタバンク出力に IPS(Integrated phoneme subspace) 変換を行うことで得られる。IPS は Fig. 1 (b) のように各音素部分空間への射影と統合変換で構成される。この二つの変換は線形変換行列を対数メルフィルタバンク出力ベクトルにかけることで表現できる。

2.1 MDL 基準による音素部分空間

主成分分析による部分空間推定において、MDL(Minimum Description Length) 基準による最適部分空間推定手法が提案されている [2]。主成分分析における MDL 基準は次式のように定義される。

$$MDL(p) = -\ln \left\{ \frac{\prod_{j=p+1}^m \hat{\lambda}_j^{\frac{1}{m-p}}}{\frac{1}{m-p} \sum_{j=p+1}^m \hat{\lambda}_j} \right\}^{(m-p)N} + M \cdot \left(\frac{1}{2} + \ln[\gamma] \right) - \frac{M}{p} \sum_{j=1}^p \ln \left[\hat{\lambda}_j \sqrt{2/N} \right] \quad (1)$$

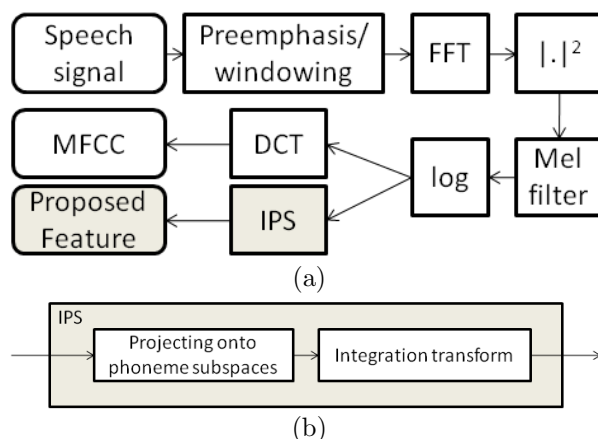


Fig. 1 (a)MFCC と提案手法の特徴量抽出 (b) 統合音素部分空間 (Integrated phoneme subspace:IPS)

ここで、 p は部分空間次元数、 m は元の次元数、 λ は固有値、 N は解析フレーム数、 $M = pm - p^2/2 + p/2 + 1$ である。最適次元数は $\hat{p} = \arg \min_p MDL(p)$ のように決まる。各音素データ集合に対して主成分分析を行い MDL 基準により部分空間を選択した後、 K 個の音素部分空間 Φ^i を $V = [\Phi^1, \Phi^2, \dots, \Phi^K]$ のようにまとめる。IPS の 1 つ目の変換により求まる特徴量 $\mathbf{y}$ は対数メルフィルタバンク出力 $\mathbf{x}$ から次式のように得られる。

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \\ \vdots \\ \mathbf{y}^K \end{bmatrix} = \begin{bmatrix} \Phi^{1T} \mathbf{x} \\ \Phi^{2T} \mathbf{x} \\ \vdots \\ \Phi^{KT} \mathbf{x} \end{bmatrix} = V^T \mathbf{x} \quad (2)$$

2.2 音素部分空間の統合

式 (2) の $\mathbf{y}$ は高次元ベクトルであり、各要素間には相関が存在する。リアルタイム音声認識では特徴ベクトルが低次元であることが望ましい。ここでは音素部分空間を統合する変換行列を用いて特徴ベクトル次元圧縮を行う。この変換行列は主成分分析や独立成分分析により求まるが、主成分分析は特徴ベクトルの要素間に無相関制約、独立成分分析は要素間に相互独立制約を与える。独立成分分析は FastICA [3] を用いた。統合する変換行列を W とし、IPS による特徴量 $\mathbf{s}$ は次のように得られる。

$$\mathbf{s} = W\mathbf{y} = WV^T \mathbf{x} \quad (3)$$

* Integration of Phoneme-subspaces obtained by MDL principal for speech feature extraction by Hyunsin PARK, Tetsuya TAKIGUCHI, and Yasuo ARIKI (Kobe Univ.)

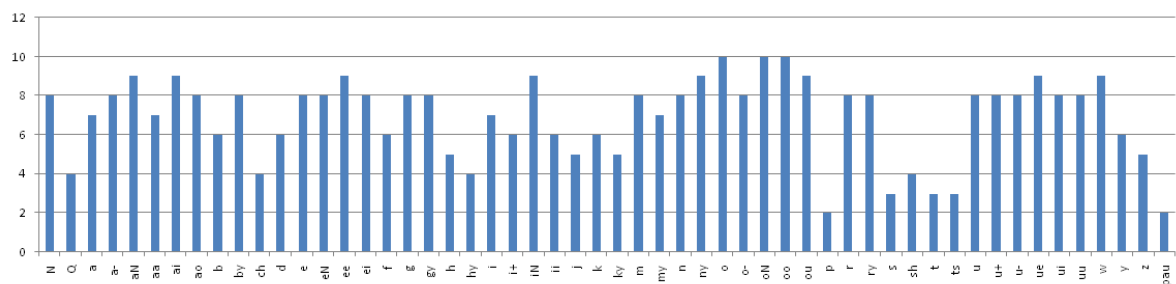


Fig. 2 Phonemes and Optimal subspace dimension

3 評価実験

評価実験として孤立単語音声認識を行った。データはATRのSpeechDatabase SET-Aから男女5名ずつ10人のデータを用いた。学習には各話者2620単語、テストには各話者1000単語を用いた。音声信号は16 bit/12 kHzでサンプリングされ、32 msのハミング窓を8 msシフトさせながらフレーム化した。24次元の対数メルフィルタバンク出力に対して離散コサイン変換したMFCC、全ての音声データを用いて主成分分析を適用した場合(PCA)、独立成分分析を適用した場合(ICA)、主成分分析を用いた音素部分空間統合による特徴量IPS1、独立成分分析を用いた音素部分空間統合による特徴量IPS2、(すべての特徴量は平均0の12+ Δ 12次元)を用いて、3状態20混合の54(KZ)個音素HMMの学習と認識を行った。学習した統合音素部分空間とHMMはクリーン音声データを用いた10人共通モデルである。テストはクリーン音声とインパルス応答を畳み込みした残響音声(380 ms, 600 ms)[4]を用いて行った。

Table 1 特徴量IPS1を用いたクリーン単語音声認識結果：各音素部分空間の次元数による認識率の変化。

次元数	14	12	10	8	6	MDL
認識率	96.8	96.8	96.9	96.9	96.8	96.9

Table 1はIPS1特徴量を用いた、クリーン単語音声の認識結果である。MDL基準により推定された音素部分空間次元数はFig. 2に示す。各音素部分空間の次元数を全て一定とした時と、MDL基準により決めた時の認識率の比較すると、MDL基準を用いた結果が96.9%で、各音素において次元数固定の場合と同じ結果が得られた。

Fig. 3は5つの特徴量によるクリーンまたは残響音声の認識結果である。#_AVEはICAの初期値を3回変えて行った認識結果の平均を表す。実験結果より、IPS1は3つのテスト条件で最も高い結果を示した。また、IPS2は残響条件においてIPS1に続く2番目に高い認識結果を示した。

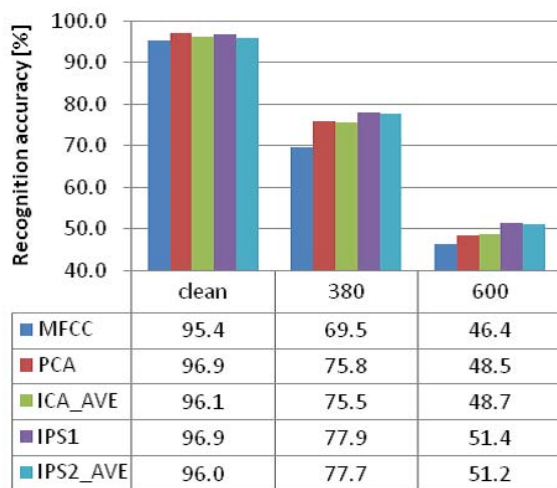


Fig. 3 孤立単語音声認識結果：10人の平均認識率
横軸：残響時間、縦軸：認識率 [%]

今回用いた音響モデルHMMは生成モデルであり、提案特徴量は音素識別的な情報になっている。今後識別的なモデルである条件付き確率場(Conditional Random Field: CRF)を用いて音声認識を行うことで、認識率の改善が期待できる。

4 おわりに

主成分分析とMDL基準を用いて各音素の部分空間を推定した後、主成分分析または独立成分分析による部分空間統合を行い、音声特徴量を抽出する手法を提案した。実験結果より、MDL基準による音素部分空間の推定が有効であることが示された。今後は、HMMの代わりに識別的なモデルであるCRFと提案手法を組み合わせる単語音声認識実験を行う予定である。

参考文献

- [1] 朴 他, 音講論(春), 199-200, 2008
- [2] R. Vetter et al, Proc. Eurospeech, 2411-2414, 1999.
- [3] A. Hyvarinen, E. Oja, Neural networks, Vol. 13, 411-430, 2000
- [4] S. Nakamura et al, Proc. LREC, Vol. 2, 965-968, 2000