論文 / 著書情報
Article / Book Information

論題(和文)	マルチバンド音声認識における尤度重み推定法の検討
Title(English)	
著者(和文)	小島要, 岩野公司, 古井貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2005年秋季講演論文集, Vol. , No. 2-7-4, pp. 65-66
Citation(English)	, Vol. , No. 2-7-4, pp. 65-66
発行日 / Pub. date	2005, 9

マルチバンド音声認識における尤度重み推定法の検討*

◎小島　要，岩野公司，古井貞熙 (東工大)

1　はじめに

音声認識では，ケプストラム特徴量である MFCC を用いることが一般的である．しかし，音声に重畳する雑音のスペクトルが特定の帯域に集中している場合には，ケプストラム領域ではその雑音の影響が全係数に拡散してしまうことから，雑音対策が講じにくくなるという欠点がある．そこで，スペクトル領域の特徴量を利用し，周波数帯域ごとに信頼度に応じた重みづけを行うことで雑音による認識性能の劣化を抑制する，マルチバンド音声認識の研究がなされている．我々はこれまでに，MFCC と同程度の認識性能を有するスペクトル特徴量の提案と [1]，ニューラルネットを用いた帯域ごとの尤度への重みづけ手法の提案 [2] を行っており，それらを利用したマルチバンド音声認識の有効性を確認している．

本稿では新たに，線形判別分析と尤度平均化を用いた，スペクトル特徴量における尤度重み推定法を提案し，マルチバンド音声認識による性能の評価を行う．また，MFCC の各次元に対して提案手法により重みづけを行った場合の性能と，スペクトル特徴量を用いた場合の性能とを比較することで，スペクトル領域の特徴量を用いることの有効性を検証する．

2　マルチバンド音声認識

我々は，MFCC と同様の各種正規化処理を対数スペクトル領域で行うことで抽出されるスペクトル特徴量が，MFCC と同等の認識性能を有することを確認している [1, 2]．本研究では，文献 [2] で提案したスペクトル特徴量を利用する．以降では，このスペクトル特徴量を SPEC と呼ぶ．

音声認識に用いる特徴ベクトルとしては，SPEC 13 次元，ΔSPEC 13 次元，Δ 対数パワー 1 次元の 27 次元ベクトルを用いる．このうち，13 次元の SPEC については，各次元を独立したストリームとしたマルチストリーム HMM を利用して，マルチバンド音声認識を実現する．その際，ΔSPEC と Δ 対数パワーの 14 次元は，まとめて 1 つのストリームと見なし，合計 14 ストリームによるマルチストリーム音声認識を行うこととした．

マルチストリーム HMM における，ある時刻 t における特徴量 $\mathbf{o}_t$ が与えられた場合の出力確率の対数値 $b(\mathbf{o}_t)$ は，

$$b(\mathbf{o}_t) = \sum_{s=1}^{S} \lambda_s \cdot b(\mathbf{o}_{st}) \tag{1}$$

と表すことができる．$b(\mathbf{o}_{st})$ はストリーム s の特徴ベクトル $\mathbf{o}_{st}$ の時刻 t での出力確率の対数値，S は総ストリーム数，λ_s は各ストリームに対するストリーム重みを表す．このストリーム重みを調整することによって，帯域ごとに重みづけを行う．

3　尤度重み推定手法

3.1　線形判別分析による重み推定

式 (1) で示した通り，認識に用いる対数尤度は各ストリームから出力される対数尤度の重みつき線形和の形で表される．線形判別分析（LDA）を用いて得られる判別関数も同じ線形和の形をしていることから，その係数をストリーム重みに見立てることができる．これは，モデルに対する入力（特徴量）が正しい事例と間違っている事例の判別性能が最大になるように，帯域ごとの信頼度が推定されることを意味している．

まず，重み推定用のデータについて，単語（音素）系列 $w_1, w_2, \cdots, w_N$（N は単語または音素数）とその時間ラベルを用意する．時間ラベルは，重み推定前の初期モデルによる強制切り出しによって作成する．次に，w_n に相当する特徴ベクトル系列 $\mathbf{o}^{w_n}$ を重み推定前のモデル m_v（$v = 1, 2, \cdots, V$，V は総モデル数）に入力し，各ストリーム s から出力される対数尤度 $b_{m_v,s}(\mathbf{o}^{w_n})$ のフレーム平均を求め，x_s 座標（S 次元空間）に点をプロットする．全ての w_n, m_v の組み合わせについてプロットを行った後，特徴量が正しいモデルに入力されたとき（$w_n = m_v$）と，異なるモデルに入力されたとき（$w_n \neq m_v$）の 2 つの分布を識別するための線形判別関数を LDA によって求める．得られた判別関数は，

$$a_0 + \sum_{s=1}^{S} a_s x_s = 0 \tag{2}$$

となる．ストリーム重み λ_s は判別関数の係数 a_s を用いて，以下のように算出する．

$$\lambda_s = S \cdot \frac{a'_s}{\sum_{i=1}^{S} a'_i}, \quad a'_i = \begin{cases} a_i & a_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

このストリーム重みは，全てのモデルに対して共通に使用される．

本実験では，ΔSPEC と Δ 対数パワーで構成されるストリーム重みは常に 1.0 に固定することから，SPEC の各次元に相当する 13 個のストリームについてのみ重み推定を行っている．したがって，重み推定に用いる総ストリーム数（LDA で使用する次元数）S は 13 である．

3.2　尤度平均化によるモデル間での重み調整

雑音下のように入力特徴量の信頼性が低い環境で音声認識を行った場合，モデル間の出力尤度に極端なばらつきや偏りが生じ，それにより認識性能が劣化する可能性がある [3]．そこで，各モデルの出力尤度が平均的に等しくなるようにモデルごとに重みの調整を行い，出力尤度の偏りを軽減することで認識性能の向上を図る．

3.1 節と同様に，単語（音素）ごとに切り分けされた重み推定用データを利用し，モデル m_v について，単語（音素）w_n に相当する特徴ベクトル系列 $\mathbf{o}^{w_n}$

---

*Likelihood weight estimation methods for multiband speech recognition.
by Kaname Kojima, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

Table 1　各重み推定手法によるマルチストリーム音声認識の結果（数字は数字正解精度 (%)）

特徴量	手法	clean	エレベータホール雑音			ステーション雑音		
			20dB	10dB	5dB	20dB	10dB	5dB
SPEC	NONE	99.4	88.5	47.9	31.5	85.8	40.7	26.4
	OLN	99.1	87.7	48.6	32.3	86.0	42.0	27.9
	LDA	99.3	92.4	53.5	35.5	90.5	46.9	31.5
	LDA+OLN	98.9	91.7	55.9	35.3	91.7	48.7	30.2
MFCC	NONE	99.3	92.1	47.6	26.2	85.1	31.8	18.3
	OLN	97.5	90.8	53.5	33.9	84.0	40.7	27.0
	LDA	99.0	93.2	53.5	32.3	85.4	34.5	19.2
	LDA+OLN	97.6	90.2	55.2	37.0	84.7	43.6	28.0

に対する対数尤度 $b_{m_v}(\mathbf{o}^{w_n})$ を算出する．データの総フレーム数を T_A とすると，モデル m_v から得られる全データに対するフレーム平均の対数尤度 $\bar{b}_{m_v}$ は，

$$\bar{b}_{m_v} = \sum_{n=1}^{N} b_{m_v}(\mathbf{o}^{w_n})/T_A \tag{4}$$

と計算される．初期モデルにおける，モデル m_v, ストリーム s の重みを $\lambda'_{v,s}$ とすると，最終的な重み $\lambda_{v,s}$ は

$$\lambda_{v,s} = \lambda'_{v,s} \cdot V \cdot \frac{1/\bar{b}_{m_v}}{\sum_{v=1}^{V} 1/\bar{b}_{m_v}} \tag{5}$$

となる．なお，Δ 項のストリームに対する重みは，本手法においても更新せず 1.0 とする．したがって，実際には $b_{m_v}(\mathbf{o}^{w_n})$ として Δ 項以外の 13 ストリームから算出された尤度を利用している．

LDA による重み推定を行ったあとで，本手法を適用することにより，モデル，ストリームごとの細かい重み推定を行うことが可能である．

4　評価実験

4.1　実験条件

実験には男性話者 11 人による clean 環境での数字発話音声を用いた．各話者はそれぞれ 2 ～ 8 桁の連続数字を 30 回発声しており，1 人当り合計 1,050 の数字発声がなされている．実験は，話者に対する leave-one-out 法により行った．つまり，10 話者による clean 音声により HMM の学習を行い，残りの 1 話者の音声を評価する．11 人全ての話者の評価を行い，数字正解精度の平均を最終的な評価値とした．モデルとしては，triphone HMM を用いた．評価データには，電子協騒音データベースのエレベータホール雑音，ステーション雑音を SNR = 5, 10, 20dB の条件で重畳したものを用意した．

実験では，SPEC と MFCC との比較を行った．MFCC を使用する場合には，特徴ベクトルとして 25 次元（MFCC 12 次元，ΔMFCC 12 次元，Δ 対数パワー 1 次元）ベクトルを利用した．SPEC は 27 次元であるが，これは MFCC ではケプストラムの 0 次の係数が除かれているためであり，情報量としては同等である．MFCC についてもマルチストリーム化を行い，Δ 項を除く 12 ストリームについて同様の重み推定を行った．なお，音響分析は 16bit 量子化，16kHz サンプリングの音声波形に対して，窓長 25ms, フレーム周期 10ms で行っており，MFCC, SPEC で同じ条件である．また，MFCC では CMS を行っている．

重み推定用のデータには，評価データと同じ雑音条件にした学習データを利用した．したがって，今回の実験では，雑音条件が既知となっている．音素の切り出し位置の情報は，clean 音声と重み推定前の初期モデルによる強制切り出しにより得た．

4.2　実験結果

初期モデルとして全てのストリーム重みを 1.0 としたものを利用し，線形判別分析による重み推定のみを行った場合（LDA），同様の初期モデルで尤度平均化による重み調整のみを行った場合（OLN），LDA による重み推定を行ったモデルを初期モデルとして，尤度平均化による重み調整を行った場合（LDA+OLN），重みづけを行わず全てのストリーム重みを 1.0 とした場合（NONE）の 4 通りについて実験を行った．特徴量別（MFCC, SPEC）の認識結果を表 1 に示す．なお，挿入ペナルティについては，実験ごとに最適なものを選択している．

SPEC による実験結果を見ると，全ての雑音条件において，線形判別分析による重み推定（LDA），尤度平均化による重み調整（OLN），共に有効であり，特に前者の効果が大きいことがわかる．また，表中の多くの条件において，両者を組み合わせることによる性能改善も確認できる．MFCC と SPEC について，線形判別分析による重み推定を行った場合の結果を比較すると，SPEC の方が重みづけによる性能改善の効果が大きいことがわかる．これは，スペクトル帯域ごとの重みづけによる，帯域性の雑音の抑制効果が現れたものと考えられる．

5　まとめ

本稿では，スペクトル特徴量を用いたマルチバンド音声認識における重み推定方法として，線形判別分析による重み推定法と，尤度平均化による重み調整法の提案を行い，雑音環境における連続数字認識によって，その有効性を示した．また，本マルチバンド音声認識の枠組みでは，ケプストラム特徴量より，スペクトル特徴量の方が重みづけによる耐雑音性の向上が大きいことも確認された．今回の実験では，重み推定を，雑音条件既知，推定用データの音素ラベル既知という条件で行っているが，今後は，これらの情報が未知の場合における手法の有効性の確認を行う必要がある．また，提案手法を MLLR や SS 法などと併用した場合の効果についても検証する必要がある．

参考文献

[1] 西村他，信学技報，vol.103, no.519, pp.19-24 (2003-12).
[2] 西村他，音講論，vol.1, pp.117-118 (2004-3).
[3] 田村他，音講論，vol.1, pp.145-146 (2004-9).