T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	周波数帯域ごとの重みつき尤度を用いた音声認識の検討
Title(English)	
著者(和文)	西村 義隆, 篠崎 隆宏, 岩野 公司, 古井 貞熙
Authors(English)	Takahiro Shinozaki, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2004年春季講演論文集, Vol. , No. 2-11-9, pp. 117-118
Citation(English)	, Vol. , No. 2-11-9, pp. 117-118
発行日 / Pub. date	2004, 3

Powered by T2R2 (Science Tokyo Research Repository)

周波数帯域ごとの重みつき尤度を用いた音声認識の検討 *

◎西村 義隆　篠崎 隆宏　岩野 公司　古井 貞熙　(東工大)

1　はじめに

音声認識では，特徴量として MFCC (Mel Frequency Cepstrum Coefficient) を用いることが一般的である．しかし，スペクトル領域において雑音がある帯域にのみに重畳していたとしても，MFCC ではその雑音が全ての係数に拡がってしまうという欠点がある．そこで，加法性雑音に対する頑健性を考えたとき，スペクトル領域の特徴量を用いることができれば，雑音の分離がしやすく有利である．

スペクトル領域の特徴量を用いた音声認識は MFCC と比べるとその有効性が示されていなかったが，我々はこれまでに，MFCC と同じ正規化を施すことにより，連続数字発声において MFCC と同程度の認識ができることを示した．さらに，スペクトルピークを用いた帯域ごとの尤度に対する重みづけを行うことで，その認識率が向上することも確認した [5].

本稿では，スペクトル特徴量の大語彙音声認識への適用を検討する．さらに，ニューラルネットを用いたスペクトル帯域ごとの尤度の重みづけ手法を提案し，その有効性について報告する．

2　スペクトル特徴量

図 1 に，スペクトル領域で正規化されたスペクトル特徴量を抽出するための手法を示す．E の平均エネルギー正規化, F のスペクトルピーク強調, G の対数スペクトル平均除去が通常 MFCC で行われている正規化処理をスペクトル領域で実現するための処理である．以下，E〜G の正規化について順に示す.

2.1　平均エネルギー正規化

平均エネルギー正規化では，ケプストラムの 0 次の係数の除去と同等の正規化を行う．ケプストラムの 0 次の係数には対数スペクトルの直流成分の値が現れるため，この成分を除去する処理を行う．これは，離散コサインの変換公式から，1 フレームにおける各次元の対数スペクトルの平均成分を減じることで処理できる．

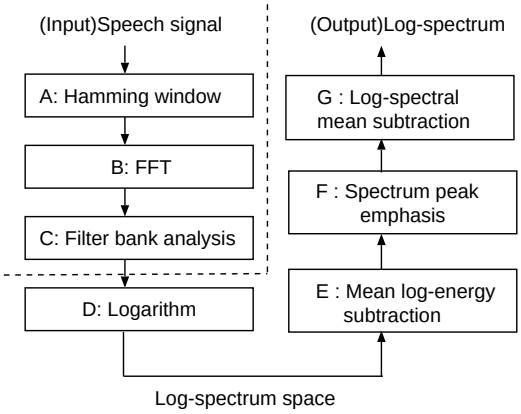

図 1. 対数スペクトル領域で正規化を行いスペクトル特徴量を抽出する手順

2.2　スペクトルピーク強調

スペクトルピーク強調では，MFCC におけるリフタリング処理に相当する正規化を行う．

MFCC におけるリフタリング処理では，まず離散コサイン変換を行い，高次の係数の打ち切りを行う．スペクトル領域での正規化では，これをメルフィルタバンクの次元数を減らす処理で近似した．

さらに，MFCC では，高域通過リフターに通すことにより，ダイナミックレンジの調整を行う．スペクトル領域においては，ケプストラムの高次の係数が強調されることにより，スペクトルの山と谷の構造が強調されるため，重要な処理と考えられる．

スペクトルピーク強調は，次のような伝達関数をもつリフタに通すことにより処理を行う．

$$H(z) = 1 - pz^{-1} \tag{1}$$

2.3　対数スペクトル平均除去

対数スペクトル平均除去は，MFCC における CMS と同等の処理を行う．CMS は，特徴量の各次元においてその時間平均を差し引くことにより処理を行うが，対数スペクトル領域においても同じように時間平均を差し引くことで処理を行うことができる．

3　大語彙音声認識による実験

3.1　実験条件

使用する音声には，CSJ より学会講演音声 797 講演の音声データを用いた．音響モデルの作成は 787 講演 (186 時間) で行い，残りの 10 講演によって認識実験を行った．付与する雑音には，電子協騒音データベースより，エレベータホール雑音を用いた．

音響特徴量には，1) MFCC12 次元+Δ 12 次元+Δ対数パワー 1 次元の計 25 次元 (以下，MFCC と略記)，2) 2 節で説明したスペクトル 13 次元+Δ 13 次元+Δ対数パワー 1 次元の計 27 次元 (以下，SPEC と略記) の 2 つを用いる．ケプストラムは 0 次係数の除去によって次元が少なくなるのに対し，スペクトル領域では次元数が変わらない．したがって，SPEC は MFCC に比べ 2 次元多いが情報量は同等であると考えられる．

音響モデルは混合数 16，状態数 3000 の triphone である．分散には対角共分散を用いた．モデルの作成には HTK を用い，デコーダには Julius を用いた．

また，スペクトルサブトラクション (SS) や MLLR を適用した場合の実験も行った．MLLR は教師なし適応を行った．SS は正規化に先立ち，文献 [4] と同様の手法で行った．

3.2　実験結果

大語彙音声認識による MFCC と SPEC の比較実験結果を表 1 に示す．+SS, +MLLR として示してあるものは，それぞれ SS または MLLR の適応化を行った結果を示す．

∗ Speech recognition using weighted likelihood of each frequency band

By Yoshitaka Nishimura, Takahiro Shinozaki, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

表 1. 学会講演音声を用いたスペクトルと MFCC の認識率の比較

SNR	∞	20dB	10dB	5dB
MFCC	69.94	64.43	33.29	16.49
SPEC	70.02	64.70	33.99	16.26
MFCC+SS	69.94	62.16	42.91	23.71
SPEC+SS	70.02	63.84	44.03	23.76
MFCC+MLLR	75.30	71.97	51.41	25.17
SPEC+MLLR	75.52	72.03	50.27	25.38

MFCC と SPEC は同等の認識性能であることが示されている．また，SS，MLLR を用いた場合にも両者に差がないことが分かる．

4 尤度に対する重みづけ実験

スペクトル領域の特徴量を用いることの利点を活かし，各周波数帯域の尤度に対する重みづけをし，音声認識を行う．重みは，音声と雑音の非定常性を考慮し，フレーム単位で変化させる．

4.1 スペクトルピークを用いた重みづけ

スペクトルピークは，1) 音声認識において認識に重要な成分である 2) 雑音に埋もれにくく，信頼性が高い などの理由から，スペクトルピークに大きな重みをおく方法を提案する．

具体的には，一定の閾値以上の対数スペクトル値に，その値に比例した重みを与える [5].

4.2 ニューラルネットを用いた重みづけ

各周波数帯域に対する重みは，各帯域の SNR に比例した値とすることが望ましい．しかし，帯域ごとの SNR を評価データから直接計算することは難しい．そこで，SNR が既知のデータを使って学習したニューラルネットによって重みを推定する．

ニューラルネットには 3 階層ネットワークを用いることとした．入力層，中間層，出力層の 3 層である．入力層には特徴ベクトルを入れることとし，出力層からは各次元の重みが出力される．したがって，入力層および出力層は 13 次元である．中間層は，4 次元として実験を行った．

4.3 実験条件

この実験では，連続数字発声を用いた．男性話者 11 名による発声で，1 名当り 1050 数字発声している．評価データにはエレベータホール雑音とステーション雑音を帯域全体での SNR が 20, 10, 5dB となるように重畳した．

音響モデルの作成には評価する話者以外の 10 名の clean な音声データを用いた．

ニューラルネットの学習には，評価データとは異なる雑音が重畳された，評価する話者以外の 10 名の音声データを用いた．ニューラルネットは帯域全体の SNR(∞, 20,10, 5dB) ごとに作成した．実験ではこの値を既知とし，該当する SNR のニューラルネットのみを用いた．

音響特徴量については大語彙音声認識と同じ条件である．モデルは，混合数 4, 状態数 376 の triphone である．分散には対角共分散を用いている．評価用のデコーダには Julian を用いた．

5 実験結果

実験結果を表 2 に示す．この結果において，Proper weight には，各帯域の SNR を既知とした理想的な重みをつけて行った認識結果を示す．Peak には，スペクトルピークに重みづけを行った結果を示す．NN には，ニューラルネットを用いた重みづけを行った結果を示す．

表 2. ニューラルネットを用いた尤度に対する重みづけを行った結果

		Elevator hall noise			Station noise		
SNR	∞	20dB	10dB	5dB	20dB	10dB	5dB
MFCC	98.96	90.72	47.78	26.54	84.74	34.75	19.53
SPEC	99.08	87.83	49.28	31.96	85.55	42.50	28.68
Proper weight	99.08	99.24	67.12	51.92	93.17	64.54	46.92
Peak	98.16	88.31	51.62	32.98	87.45	45.47	30.65
NN	99.09	89.03	50.13	35.73	89.19	49.26	33.07

理想的な重みづけができるとすると，その認識率は大きく向上することが分かる．

スペクトルピークを用いた重みづけを行うと，雑音環境下では認識率が向上するが，雑音のない環境下では認識率が低下する．この手法は簡単に重みの計算を行うことができるが，雑音のない環境下ではスペクトルピークに重みを置きすぎると認識率の低下が起こるという欠点がある．一方，ニューラルネットを用いた重みづけではどの環境下でも認識率は向上しており，スペクトルピークを用いた重みづけと比較しても有効性が示されている．

6 まとめ

本稿では，大語彙音声認識による実験において MFCC と同等の正規化を行ったスペクトル特徴量を用い，MFCC と同等の認識率が達成できることを示した．また，SS や MLLR との組合せにおいても，MFCC と同等の性能が得られることを確認した．

さらに，ニューラルネットを用いて帯域ごとの尤度に対して重みづけを行うことにより，認識率が向上することを確認した．

各帯域の SNR が完全に分かっているとすると，その認識率は大きく向上するため，今後の課題として精度の高い SNR の推定方法の検討，よりよい重みづけ方法の検討などが挙げられる．

参 考 文 献

[1] A. Hagen and A. Morris, “Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR,” *Proc. ICSLP2000*, vol.1, pp.345-348, 2000.

[2] J. Ming and F.J. Smith, “Union: A new approach for combining sub-band observations for noisy speech recognition,” *Speech Communication*, vol.34, nos.1-2, pp.41-55, 2001.

[3] S. Tibrewala and H. Hermansky, “Sub-band based recognition of noisy speech,” *Proc. ICASSP97*, vol.2, pp.1255-1258, 1997.

[4] 松本弘, “雑音環境下の音声認識手法,” 情報科学技術フォーラム FIT2003, 2003.

[5] 西村義隆, 篠崎隆宏, 岩野公司, 古井貞熙, “周波数帯域ごとの重みつき尤度を用いた雑音に頑健な音声認識,” 信学技報, vol.103, no.519, pp.19-24, 2003.