論文 / 著書情報
Article / Book Information

論題(和文)	重みつきスペクトル特徴量を用いた雑音に頑健な音声認識
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 2003年秋季講演論文集, Vol. , No. 1-6-3, pp. 5-6
Citation(English)	, Vol. , No. 1-6-3, pp. 5-6
発行日 / Pub. date	2003, 9

重みつきスペクトル特徴量を用いた雑音に頑健な音声認識 *

◎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，その認識をするための特徴量としてケプストラム領域の特徴量 MFCC(Mel Frequency Cepstrum Coefficient) を用いることが一般的である．ケプストラム領域は，対数スペクトルをフーリエ変換した領域であるため，スペクトルの領域においてある箇所のみに重畳していた雑音であっても，ケプストラム領域ではその雑音が拡がってしまい，ケプストラムの全ての項に対して雑音の影響を与えてしまう欠点がある．このため，加法性の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．

スペクトル特徴量を用いた音声認識はこれまでも試みられているが，狭帯域雑音などの特定の条件下でしか有効性が示されていない [1-5].

そこで本稿では，従来用いられていたスペクトル特徴量と MFCC 特徴量の比較を行い，MFCC と同程度の認識ができる対数スペクトル特徴量を提案する．さらに，スペクトルピークへの重みづけを加えることにより [6, 7]，広帯域雑音環境下において MFCC よりも高い認識率を確認した．

2　スペクトル特徴量

2.1　MFCC とスペクトル

音声認識で一般的に用いられているケプストラム領域の特徴量である MFCC は，スペクトル特徴量を離散コサイン変換し，ケプストラム領域において 3 つの処理 (直流成分の除去，リフタリング処理，ケプストラム平均除去) することで得られる．図 1 にその流れを示す．以下，これらの処理について説明する．

(1) スペクトルの直流成分の除去．スペクトルを離散コサイン変換すると，一番低次の項 (C_0 項) には，そのフレームにおけるスペクトルの直流成分が入る．音声認識では，スペクトル構造により音素の違いを識別するため，直流成分の情報は不要である．したがって，この項を抜き取ることによって，フレームごとの直流成分の大きさによる差を吸収し，正規化を行うことができる．

(2) リフタリング処理．リフタリング処理では，スペクトルの各次元を時間系列と見立てたときに，高域通過フィルタにかけることによってフィルタリング処理を行っている．スペクトルピークの形を平滑化し，よりはっきりさせる．高域通過フィルタにかけることにより，ケプストラムの高次の項が強調され，山と谷をはっきりとさせる効果がある．音声認識ではスペクトルピークが認識における重要な特徴となっているため，リフタリング処理により，認識精度を高めることができる．

(3) ケプストラム平均除去 (CMS)．CMS では，各ケプストラム項ごとにその時間平均の値を差し引く．雑音には加法性の雑音と乗法性の歪みの 2 種類があるが，乗法性の歪みには回線特性やマイクの特性などが含まれる．これらはどの時間においてもほぼ一定であるため，時間平均音声スペクトルの変化によって観測できる．音声信号やスペクトルなどの領域で乗法性であるこれらの歪みは，スペクトルの対数をとった時点で，加法性になる．ケプストラム領域でも加法性の雑音となるので，CMS によって，これらを抑える効果が期待できる．

2.2　認識に用いるスペクトル特徴量

スペクトル特徴量を用いた音声認識をするためには，MFCC において離散コサイン変換 (DCT) によってケプストラム領域に変換する前の特徴量を用いることもできるが，DCT した後の 3 つの処理は音声認識の認識率向上に大きく貢献しており，この処理を省略すると，大きな認識率低下につながる．

そこで，本稿ではスペクトル特徴量を用いた音声認識においても，これらケプストラム領域での処理を行い，MFCC を単純に逆離散コサイン変換 (IDCT) をした対数スペクトルを特徴量として用いることとした．図 1 にスペクトル特徴量を求める手順を示す．

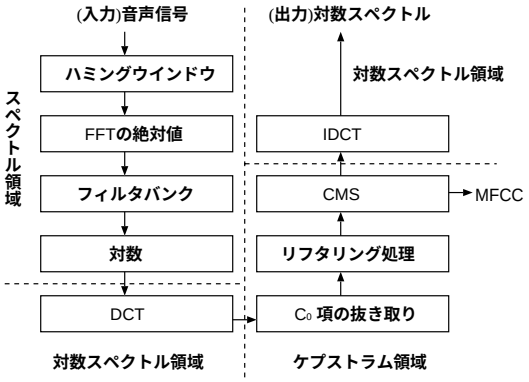

図 1. 対数スペクトルを求める手順

3　スペクトル特徴量の重みづけ

3.1　マルチバンド音声認識

本稿ではマルチバンド音声認識の枠組みに基づき，入力された特徴量に重みづけを行う．スペクトル特徴量を用いると，ある帯域に雑音が重畳していたとき，雑音の重畳していない帯域に重みをおいて認識することにより耐雑音性の向上が可能となる．これにより頑健な音声認識を目指す．

マルチバンド音声認識を用いたスペクトルの重みづけにおいて，次元 d に対する重みを w_d とすると，モデル Λ に対するある特徴量

$$\mathbf{y} = [\ y_1 \quad y_2 \quad \cdots \quad y_d \quad \cdots \quad y_D\ ]^{\mathbf{t}}$$

の重みつき尤度は，対角共分散行列を用いた正規分布を仮定し，

$$P(\mathbf{y}|\Lambda) = \sum_{d=1}^{D} w_d P_d(y_d|\Lambda) \qquad (1)$$

∗ Noise-robust speech recognition using weighted spectral features

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

として求める．$P_d(y_d|\Lambda)$ は次元 d の混合正規分布に対する対数尤度である．

重みづけには，スペクトル構造に着目した方法を用いた．人間の音声のスペクトル構造には声道の共振周波数に対応するフォルマントがあり，音素を識別するのに大きな役割を果たしている．このフォルマント位置に大きな重みを置くような重みづけを行った．

3.2 重み決定方法

重みの決定には，マルチバンドの尤度に対し，スペクトルの大きさと同じ割合での重みづけを行った．スペクトル系列を

$$\mathbf{s} = [\ s_1 \quad s_2 \quad ... \quad s_N\ ]^t \tag{2}$$

として表すとき，重み w を次のようにして決定する．

$$s'_i = \begin{cases} 1 + \alpha s_i & , s_i \geq 0 \\ 1 & , s_i < 0 \end{cases} \tag{3}$$

$$w_i = \frac{s'_i}{\sum_{j=1}^{N} s'_j} N \tag{4}$$

本稿では対数スペクトルにおいて信頼性の低い負の値は一定値 1 と固定し, 正の対数スペクトルに対して一定値 α を乗じた値を加算することとした．α は $0 < \alpha \leq 1$ とした．

さらに，重みづけを行う前と行った後で，音響モデルに対する尤度がほぼ同じになるよう N を乗じ, 正規化を施した．

4 実験条件

学習用および評価用に用いた音声データは clean な環境で録音した男性話者 11 名による連続数字音声であり，全ての話者は 2 桁から 8 桁の連続数字をそれぞれ 30 回 (210 連続数字，合計 1,050 数字) 発声している．

実験方法には leave-one-out 法を用い，男性話者 10 名の clean な音声で学習を行い，残りの話者 1 名によって認識を行う．

雑音には電子協騒音データベースのエレベータホール雑音を用いた．重みつきスペクトルの評価実験にはさらにステーション雑音も用いた．ともにバブル雑音である．雑音は SNR5, 10, 20dB の 3 種類の大きさで音声に重畳した．

音響特徴量に MFCC12 次元，ΔMFCC12 次元，Δ 対数パワーの計 25 次元を用いた．スペクトルには，スペクトル 12 次元に Δ スペクトル 12 次元，Δ 対数パワーの計 25 次元を用い，スペクトルは MFCC を IDCT したものである．CMS は複数の発声をまとめて (平均 50 数字) かけている．

学習用の音響モデルは混合数 4，状態数 376 のトライフォンを使用し，作成には HTK を用いた．言語モデルにはネットワーク文法を用いた．MFCC とスペクトルの比較実験の評価用デコーダには HTK を用い，重みつきスペクトルの評価実験には Julian を用いた．これには，スペクトル特徴量の重みづけに対応させるため，改良を行ってある．

5 実験結果

表 1 に MFCC 特徴量とスペクトル特徴量 (SPEC) の比較を示した．表 1 における SPEC の括弧内は 2.1 節で示した 3 つの処理を示している．“C_0 cut” は C_0 項を抜くことによる直流成分の除去を，“Lifter” はリフタリング処理を，“CMS” はケプストラム平均除去の処理を行ったことを示す．

表 1. スペクトルと MFCC の認識率の比較

SNR	∞	20dB	10dB	5dB
MFCC	99.31	91.55	48.91	27.25
SPEC(-)	40.49	31.47	22.38	19.01
SPEC(Lifter+CMS)	93.12	66.87	35.63	26.48
SPEC(C_0 cut+CMS)	98.13	86.88	39.83	24.96
SPEC(C_0 cut+Lifter)	99.23	92.57	52.39	31.74
SPEC(C_0 cut+Lifter+CMS)	99.51	91.78	53.69	34.37

表 2. 重みつきスペクトルを用いた効果

		エレベータホール雑音			ステーション雑音		
SNR	∞	20dB	10dB	5dB	20dB	10dB	5dB
MFCC	98.92	90.70	46.85	26.08	76.67	30.83	18.37
SPEC	98.51	90.62	49.35	32.28	82.43	41.97	26.67
Weighted SPEC	97.99	91.82	55.29	35.59	85.78	41.25	28.79

表 2 は重みづけにより平均的に最もよかった結果の $\alpha = 0.2$ とした場合の重みつきスペクトルと重みづけを行わないスペクトル，MFCC の値を比較したものである．

これらの結果より，雑音環境下においては対数スペクトルおよびその重みづけを用いることによって大きく認識率の向上を得ることができた．

6 まとめ

提案した手法により，スペクトル特徴量を用いることにより，MFCC 特徴量と clean な環境において同程度，雑音環境下においては大きく認識率が向上することを確認した．

今後は，大語彙音声認識に対するタスクに本手法を適用する予定である．さらに，スペクトル領域における適応化手法などを組み合わせることにより，より頑健なシステムの構築を目指す．

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] A. Hagen, A. Morris, and H. Bourlard, “From multi-band full combination to multi-stream full combination processing in robust ASR,” *Proc. ISCA ITRW ASR2000*, pp.175-180, 2000.

[3] A. Hagen, H. Bourlard, and A. Morris, “Adaptive ML-weighting in multi-band recombination of gaussian mixture ASR,” *Proc. ICASSP2001*, vol.1, pp.257-260, 2001.

[4] A. Hagen, A. Morris, H. Bourlard, and H. Glotin, “Multi-stream adaptive evidence combination for noise robust ASR,” *Speech Communication*, vol.34, nos.1-2, pp.25-40, 2001.

[5] 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.

[6] 古井 貞熙，ディジタル音声処理，東海大学出版会.

[7] 杉山 雅英，鹿野 清宏，“ピークに重みをおいた LPC スペクトルマッチング尺度，” 信学論，vol.J64-A, no.5, pp.409-416, 1981.