社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49　（４）
2003／12／18

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

西村　義隆†　　篠崎　隆宏†　　岩野　公司†　　古井　貞熙†

† 東京工業大学大学院情報理工学研究科 〒152-8552 東京都目黒区大岡山 2-12-1
E-mail: †{nisshi,staka,iwano,furui}@furui.cs.titech.ac.jp

あらまし　音声認識では，認識のための音声特徴量としてケプストラム領域の特徴量である 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．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．実験の結果，スペクトルピークへの重みづけを加えることにより，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．

キーワード　スペクトル，MFCC，マルチバンド音声認識，スペクトルピーク

Noise-robust speech recognition using band-dependent weighted likelihood

Yoshitaka NISHIMURA†, Takahiro SHINOZAKI†, Koji IWANO†, and Sadaoki FURUI†

† Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: †{nisshi,staka,iwano,furui}@furui.cs.titech.ac.jp

Abstract　In most of the state-of-the-art automatic speech recognition (ASR) systems, speech is converted into a time function of the MFCC (Mel Frequency Cepstrum Coefficient) vector. However, the MFCC has a problem in that noise effects spread over all the coefficients even when the noise is limited within a narrow frequency range. If a spectrum feature is directly used, this problem can be avoided and thus robustness against noise could be expected to increase. Although various researches on using spectral domain features have been conducted, improvement of recognition performances has been reported only in limited noise conditions. This paper proposes a novel multi-band ASR method using a new log-spectral domain feature. Experimental results using bubble noise-added speech show that recognition performance is improved by the proposed method in comparison with the MFCC-based method. The performance is further improved by a spectral-peak weighting technique.

Key words　spectrum, MFCC, multi-band ASR, spectral-peak

1. はじめに

今日，雑音のない理想的な環境下での音声認識では非常に高い認識率を得ることができるが，雑音の存在する実環境下ではその認識率がたちまち下がってしまう．音声認識をカーナビなどのアプリケーションに組み込んでいくためには，雑音の存在する実環境下においてもきちんと認識できることが必要である．

音声認識では，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である 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，スペクトル領域に於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という欠点がある．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，ある特定の条件のもとでのみ有効性を示していた．つまり，スペクトル特徴量を用いた場合，ある帯域に限られた狭帯域雑音環境に対しては MFCC を上回る認識率を達成することができたが，雑音のないクリーンな環境または広帯域雑音環境に対しては MFCC を上回ることができなかった [1-7]．

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，従来用いられてきたスペクトル特徴量が MFCC と比べて有効性を示さない原因を調査し，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 [8, 9] を加えることにより，広帯域雑音環境下において MFCC よりも高い認識率を目指す．

2. スペクトル特徴量の抽出

2.1 MFCC 特徴量とスペクトル特徴量の比較

ケプストラム領域での正規化処理を含むスペクトル特徴量の抽出方法を図 1 に示す．入力信号は，ハミングウインドウ (A) にかけられ，高速フーリエ変換 (B) されて，スペクトル領域に変換される．さらに，フィルタバンク (C) により，特徴量として用いられる適当な次元のスペクトル系列に変換され，対数変換 (D) により，線形領域のスペクトルから，対数スペクトルへと変換される．離散コサイン変換 (E) により，対数スペクトル領域の系列はケプストラム領域に変換され，C_0 除去 (F)，リフタリング処理 (G)，ケプストラム平均除去 (H) の処理が加えられる．

従来のスペクトル特徴量と，ケプストラム特徴量を比較した際，スペクトル特徴量では，ケプストラム領域における処理，すなわち図 1 では，C_0 除去 (F)，リフタリング処理 (G)，ケプストラム平均除去 (H) の処理がなされていない．

C_0 除去 (F) では，スペクトルの直流成分すなわち平均エネルギーの除去を行う．この成分は，周囲の雑音や録音環境などにより変化するため，これを除去して正規化することにより，音

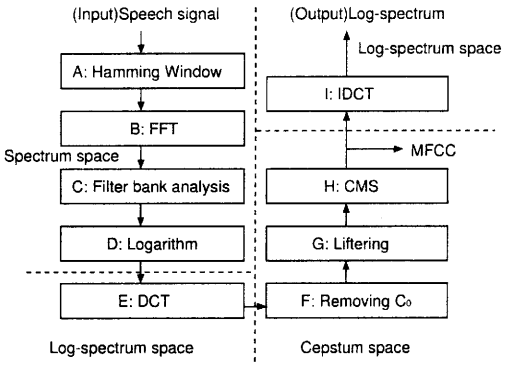

図 1　ケプストラム領域で正規化を行いスペクトル特徴量 (SPEC) を抽出する手順

声認識によい効果があると考えられる．リフタリング処理 (G) では，スペクトル構造の山と谷の強調を行う．音声認識では，スペクトル構造の山と谷により認識を行うため，非常に重要な部分である．ケプストラム平均除去 (H) では，対数スペクトル領域の平均値を引くことにより，音源からマイクまでの音響特性やマイクの回線特性など，乗算性の成分として信号に加わる雑音に対して効果がある．

そこで，スペクトル領域の特徴量を抽出する際，図 1 に示されている逆離散コサイン変換 (I) により，MFCC をケプストラム領域からスペクトル領域へ戻し，対数スペクトル特徴量を抽出した．後に示す実験では，この特徴量を “SPEC” としている．

さらに，ケプストラム領域における 3 つの処理，C_0 除去 (F)，リフタリング処理 (G)，ケプストラム平均除去 (H) のうち，どの処理が認識率に大きな影響を与えているか調べるため，これらの処理のうち，それぞれ 1 つのみを行わなかった場合のスペクトル特徴量を抽出し，実験を行った．

2.2 スペクトル領域のみの処理により抽出するスペクトル特徴量

2.1 節では，MFCC を単純に逆離散コサイン変換することにより，スペクトル特徴量を抽出した．しかし，ケプストラム領域における 3 つの処理，C_0 除去 (F)，リフタリング処理 (G)，ケプストラム平均除去 (H) と同等の処理は，スペクトル領域においても行うことが可能である．そこで本節では，ケプストラム領域への変換は行わず，スペクトル領域のみでケプストラム領域での正規化と同等の正規化を行う方法について示す．

まず，C_0 除去 (F) であるが，これは離散コサイン変換した際に出て来る 0 番目の項を除去する処理である．C_0 を除去する操作は，対数スペクトルの平均を引くことによって同等の処理とすることができる．つまり，フレーム t における i 次元目のスペクトルを s_{ti} とすると，フレーム t の平均スペクトル $\overline{s_t}$ は，

$$\overline{s_t} = \frac{1}{N}\sum_{i=1}^{N} s_{ti} \tag{1}$$

となる．N は 1 フレームにおける帯域数を表す．よって，対数

スペクトルの平均除去によって正規化されたスペクトル $\tilde{s}_{ti}$ は

$$\tilde{s}_{ti} = s_{ti} - \overline{s_t} \tag{2}$$

となる．これを直流成分除去 (F') として表す．

次に，リフタリング処理 (G) であるが，この処理では初めにケプストラム項の高次の項を切り落とし，その後高域通過リフターにかける．これはスペクトル系列を波形と見立てたときに，帯域通過フィルタにかけることと同じである．従って，フィルタバンクの数を少なくすることにより，ケプストラムの高次の項を切り落とすことと同等のことを行い，さらにスペクトル系列を入力信号系列として見たとき，何らかの高域通過フィルタにかけることにより，同様なスペクトル構造の山と谷の強調を行うことができると考えられる．本稿では式 (3) の伝達関数により一次の IIR フィルタを設計し，そのパラメータ p は，実験的に最適値を選ぶこととした．

$$H(z) = 1 - pz^{-1} \tag{3}$$

フレーム t のスペクトル系列を

$$\mathbf{s_t} = [\ s_{t1} \quad s_{t2} \quad \ldots \quad s_{tN}\ ] \tag{4}$$

スペクトルピーク強調されたスペクトル系列を $\tilde{s}_{tx}$，$H(z)$ のインパルス応答を h_x とし，

$$\tilde{s}_{tx} = h_x * s_{tx} \tag{5}$$

として処理を行うこととした．$*$ は畳み込み演算を表す．これをスペクトルピーク強調 (G') として表す．

最後にケプストラム平均除去 (H) であるが，これは，ケプストラム上においてその平均値をとっても，対数スペクトル上においてその平均値をとっても，同じ結果となる．よって，対数スペクトル上において時間平均をとり，その平均を引くことによって同じ処理を行った．フレーム t における i 次元目のスペクトルを s_{ti} とすると，i 次元目のスペクトルの時間平均 $\overline{s_i}$ は，

$$\overline{s_i} = \frac{1}{T}\sum_{t=1}^{T} s_{ti} \tag{6}$$

となる．よって，対数スペクトル平均除去において正規化されたスペクトル $\tilde{s}_{ti}$ は

$$\tilde{s}_{ti} = s_{ti} - \overline{s_i} \tag{7}$$

となる．この処理を対数スペクトル平均除去 (H') として表す．

図 2 に，スペクトル領域のみで抽出するスペクトル特徴量の抽出方法を示す．図における F', G', H' が，図 1 の F, G, H の処理にそれぞれ対応する．後の実験では，このスペクトル特徴量を "SPEC2" としている．

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．

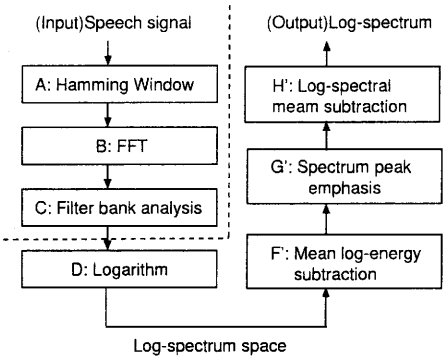

図 2　対数スペクトル領域で正規化を行いスペクトル特徴量 (SPEC2) を抽出する手順

3.2 重みづけ方法

音素 q_k に対するあるフレーム t の特徴量

$$\mathbf{s_t} = [\ s_{t1} \quad s_{t2} \quad \ldots \quad s_{tN}\ ]$$

の対数尤度は，

$$L(\mathbf{s_t}|q_k) = \sum_n L(s_{tn}|q_k) \tag{8}$$

として求めることができる．

重みを w とすると，重みつき対数尤度は

$$L(\mathbf{s_t}|q_k) = \sum_n w_d L(s_{tn}|q_k) \tag{9}$$

として求めることとする．

3.3 フォルマント構造に着目した音声領域の検出

重みづけを行う方法として，雑音検出を行い，その信号との SNR の大きさによって重みを決定する方法がある．雑音検出方法にはさまざまなものがあり，それらを用いることによって，各帯域の SN 比 (SNR) を推定することができる．雑音検出方法の多くはその音圧の大きさを用いて音声信号であるか雑音信号であるかを判別する．また，その判別には閾値が必要であり，この閾値を求めるのにいくつかのフレームのデータを集める必要などが生じ，時間遅れなどの問題がある．

本稿では，最も簡単に，しかも高性能で重みづけができる方法として，スペクトル構造に着目した重みづけを行った．人間の音声のスペクトル構造にはフォルマントとよばれる声道の共振周波数が存在しており，これは発せられている音素を識別するのに大きな役割を果たしている．音声信号には 3500Hz 以下に約 3 つのフォルマントが存在している．本稿では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．重みづけの具体的な方法は次節で述べる．

3.4 重み決定方法

重みの決定には，スペクトルの特徴量に対し，そ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

$$\mathbf{s_t} = [\ s_{t1} \quad s_{t2} \quad \ldots \quad s_{tN}\ ]^t \tag{10}$$

として表すとき，重み w を次のようにして決定する．

$$w_{ti} = \frac{s_{ti}}{\sum_{j=1}^{N} s_{tj}} \tag{11}$$

音圧の大きな部分は，音声信号がある確率が高く，しかもフォルマントが含まれている確率も高い．したがって，スペクトルの大きなところには大きな重みをおくことにより，音声認識における重要な特徴量をより強調することができる．

本稿では特徴量として対数スペクトルを用いている．そのため，負の値が生じた場合，負の重みを生ずる問題がある．負の値が発生したときにどのように重みづけするかが問題となるが，閾値 γ を設け，γ より小さなものは一定値 β，γ より大きなものは $s_{ti} - \gamma$ とした．つまり，重み w は

$$s'_{ti} = \begin{cases} s_{ti} - \gamma & , s_{ti} \geq \gamma \\ \beta & , s_{ti} < \gamma \end{cases} \tag{12}$$

$$w_{ti} = \frac{s'_{ti}}{\sum_{j=1}^{N} s'_{tj}} \tag{13}$$

として求める．

この β の値を固定すると，重み w は，正の対数スペクトルの値が大きいときには相対的に小さな値となり，正の対数スペクトルの値が小さいときには相対的に大きな値となる．どのような値を与えると最適となるかは，実験により決定しなくてはならない．そこで，β の値を 1 に固定する一方，パラメータ α を導入し，γ より大きな対数スペクトルに対し一定値 α を乗ずることとした．

$$s''_{ti} = \begin{cases} 1 + \alpha(s_{ti} - \gamma) & , s_{ti} \geq \gamma \\ 1 & , s_{ti} < \gamma \end{cases} \tag{14}$$

$$w_{ti} = N \frac{s''_{ti}}{\sum_{j=1}^{N} s''_{tj}} \tag{15}$$

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．後の実験では，$\gamma = 0$ として実験を行った．

4. 連続数字音声を用いた実験

4.1 音声データ

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．これを評価する話者を換えながら 11 回行い，その平均を認識率とした．

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．スペクトルの重みづけの実験にはさらにステーション雑音を用いた．両方ともバブル雑音である．雑音は SNR5, 10, 20dB の 3 種類の大きさを用いた．

4.2 音響特徴量

音響特徴量は MFCC には，MFCC12 次元に Δ MFCC12 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 13 次元に Δ スペクトル 13 次元，Δ 対数パワーの計 27 次元を用いた．MFCC では 2.1 節に示した C_0 項の除去においてどのフレームにおいても $C_0 = 0$ と一定であるため，この項を用いない．しかし，これをスペクトル領域に戻す際には，$C_0 = 0$ として逆離散コサイン変換を行わなければならないため，同じ特徴量を表すためにはスペクトル領域では 1 次元増えることとなる．それに伴って Δ 項も 1 次元増えるため，MFCC25 次元と比べるとスペクトル特徴量では 2 次元増えることとなる．

2.1 節で示した SPEC では，フィルタバンクの数は 24 であり，離散コサイン変換の後 13 次元のケプストラムにし，ケプストラム領域での正規化を行った後，13 次元の対数スペクトルに変換している．一方，2.2 節で示した SPEC2 では，フィルタバンクの数は 13 であり，13 次元の対数スペクトルを抽出している．

スペクトルの重みづけはスペクトル 13 次元の部分に対して行い，Δ スペクトルおよび Δ 対数パワーの 14 次元の部分に対しては重みは変えず 1 のままとした．

4.3 モデルおよびデコーダ

HMM モデルには 8 混合正規分布を用い，分散には対角共分散を用いた．

学習用の音響モデルの作成には HTK を用いた．音声データは 0 から 9 までの 10 種類の数字発声であるため，言語モデルにはネットワーク文法を用いている．

MFCC とスペクトルの比較実験には評価用のデコーダに HTK を用い，スペクトルの重みづけの評価実験および MLLR による適応化の評価実験にはネットワーク文法対応の Julian を用いた．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. 連続数字音声を用いた実験結果

5.1 MFCC とスペクトルの比較

表 1 に MFCC 特徴量を用いた音声認識結果と，2.1 節で示した処理を行わないスペクトル特徴量，処理を行ったスペクトル特徴量の比較を示した．

表 1 における SPEC(Nothing) は 2.1 節で示した 3 つの処理を全て行わなかったときの結果を示し，SPEC(C_0 cut+Lifter+CMS) は C_0 項の除去，リフタリング，CMS3 つ全ての処理を行った結果を示す．SPEC(Lifter+CMS) は C_0 項の抜き取り処理を省略し，リフタリングと CMS のみ行った結果，SPEC(C_0 cut+CMS) はリフタリングを行わず，C_0 項の抜き取りと CMS のみ行った結果，SPEC(C_0 cut+Lifter) は CMS のみ行わず，C_0 項の抜き取りとリフタリングのみ行った結果を示している．

この結果より，ケプストラム領域での 3 つの処理はどれも音声認識に大きな影響を与えており，重要な処理であることが分かる．また，3 つの正規化処理を施したスペクトル特徴量は，

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(Nothing)	50.63	33.94	19.21	13.19
SPEC(Lifter+CMS)	97.40	74.62	32.71	24.41
SPEC(C_0 cut+CMS)	98.46	86.82	37.86	23.03
SPEC(C_0 cut+Lifter)	96.08	64.50	39.65	26.43
SPEC(C_0 cut+Lifter+CMS)	99.58	91.23	50.90	31.75

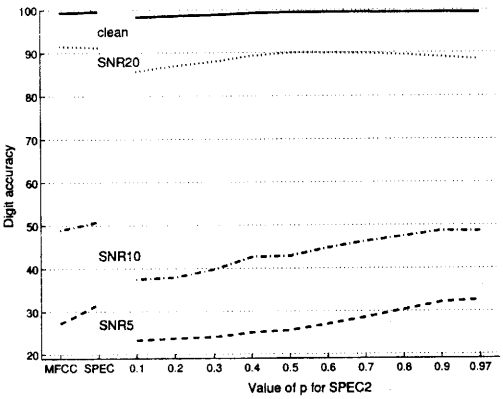

図 3　p の変化による認識率の変化

表 2　提案手法のスペクトルと MFCC の認識率の比較

		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	98.92	89.74	48.92	30.16	80.72	39.12	23.63
SPEC2	99.08	87.83	49.28	31.96	85.55	42.50	28.68
Weighted SPEC	98.58	89.53	50.16	31.81	79.92	39.28	24.57
Weighted SPEC2	98.16	88.31	51.62	32.98	87.45	45.47	30.65

MFCC よりも高い認識率を示している．

5.2　スペクトルピーク強調におけるパラメータを変化させた際の認識率の比較

2.2 節のスペクトルピーク強調において，パラメータ p を変化させることによるスペクトル特徴量の認識率の比較を図 3 に示す．

図 3 の MFCC は MFCC の認識結果，SPEC は 2.1 節で示したスペクトル特徴量で，SPEC2 は 2.2 節で示したスペクトル特徴量である．2.2 節で示した p の値を変えて実験を行った結果が示してある．図 3 の結果より，総合的に見て最適な $p = 0.9$ をスペクトルピーク強調のためのパラメータとして用いることとした．

5.3　MFCC と SPEC，SPEC2 およびマルチバンド音声認識を用いた認識の比較

表 2 は MFCC と SPEC(ケプストラム領域の正規化をしたスペクトル)，SPEC2(対数スペクトル領域で正規化をしたスペクトル) および，その重みづけを行った認識の結果を示している．α の値は，実験結果より最適な値で認識を行っている．SPEC の方では $\alpha = 0.1$, SPEC2 の方では $\alpha = 4.0$ とした結果である．

雑音のない環境下では，MFCC と同程度の認識率しか達成できなかったが，雑音環境下では，SPEC，SPEC2 ともに，その重みづけを用いることにより認識率が向上していることが分かる．

表 3　理想的な重みの検討

		Elevator hall noise			Station noise		
SNR	∞	20dB	10dB	5dB	20dB	10dB	5dB
SPEC2	99.08	87.83	49.28	31.96	85.55	42.50	28.68
Weighted SPEC2	99.08	92.94	67.12	51.92	93.17	64.54	46.92

表 4　スペクトルと MFCC の MLLR を行った際の比較

		Elevator hall noise			Station noise		
SNR	∞	20dB	10dB	5dB	20dB	10dB	5dB
MFCC+MLLR	99.43	97.91	80.80	51.96	97.41	69.66	37.13
SPEC2+MLLR	99.56	97.08	77.96	47.79	96.93	70.42	37.76

5.4　理想的な重みの検討

本稿では，3.4 節の方法によりマルチバンド音声認識における重みを求めたが，2.2 節に示した処理後のスペクトル特徴量の SN 比より理想的な重みを求めることを行った．具体的には，3.4 節における式 (14) から (16) の s_i にスペクトルではなく，SNR の値を代入し，α および γ の値を調整することにより，最適となる重みを選んだ．その結果を表 3 に示す．

表 2 の結果と比べると，理想的な重みを求めること，つまり精度よく正規化後のスペクトルの SN 比を求めることができれば，雑音環境下においても高い認識率で認識ができることが分かる．

5.5　MLLR を用いた適応化

MFCC と，2.2 節で示した重みつけはしていない SPEC2 とを用いた音声認識実験において，MLLR による適応化を行った．表 4 に実験結果を示す．SPEC2 は，適応化前は MFCC と比べると高い認識率を示していたが，MLLR による適応化を行った場合はやや異なる結果となった．即ち，雑音のない clean な状況，ステーション雑音の SNR10dB およびステーション雑音の SNR5dB では若干高い結果を示しているが，エレベータホール雑音では劣る結果となった．ただ，MLLR による適応を行った場合において，SPEC2 を用いても，MFCC とほぼ同じ認識率で認識を行うことができることが確認された．

6.　学会講演音声を用いた実験

6.1　実験条件

4. 節および 5. 節では数字発声を対象とした実験結果を示したが，本節では学会講演音声を用いた実験結果を示す．

音響モデルの学習および認識実験を行う際の音声データは 797 講演の音声データを用いた．このうち，787 講演 (186 時間) で音響モデルの学習を行い，10 講演で認識を行っている．雑音は，エレベータホール雑音を付加してある．HMM の総状態数は 3000, 16 混合の triphone である．

言語モデルは，順向き単語 bigram と逆向き単語 trigram を使用した．認識には 2 パスデコーダである Julius を利用し，ファーストパスで順向き bigram, セカンドパスで逆向き trigram を使用する．また，マルチバンド音声認識では，重みづけに対応するよう Julius を改良してある．

表 5　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	69.02	63.28	32.68	14.83
SPEC2	70.02	64.70	33.99	16.26
Weighted SPEC2	65.61	63.83	35.37	17.72

表 6　SS を適用した際のスペクトルと MFCC の認識率の比較

SNR	∞	20dB	10dB	5dB
MFCC+SS	69.94	62.16	42.91	23.71
SPEC2+SS	70.02	63.84	44.03	23.76

6.2　重みづけの効果

表 5 に MFCC, SPEC, SPEC2 および重みつきの SPEC2 の結果を示す．重みつき SPEC2 では数字発声において最もよい結果を示した $\alpha = 4.0$ の値を用いている．重みつき SPEC2 では，SNR10dB や SNR5dB といった雑音の大きな環境では認識率が上がっているが，雑音の小さな環境では逆に認識率が下がってしまっている．α の値は，現段階ではさまざまな値で試していないため，総合的によい結果を示す α の値はまだ検討できていない．α の値は，小さくすればするほど，全体的に重みが 1 に近づき，大きくすればするほど，スペクトルの大きさに応じた値に近づく．また，雑音が大きくなればなるほど，周波数帯域間に重みの差がついている方が認識率が向上する傾向がある．従って，α の値を調整することでどのような雑音環境下でも対応が可能である．例えば，$\alpha = 0.1$ とすると，SNR+∞ で 70.00%, $\alpha = 0.5$ とすると SNR+20dB で 65.10% となり，改善することができる．ただし，α の値を固定してしまうと，現段階では雑音の大きな環境下でしかよい結果を示さず，これをどうするかは今後の課題である．雑音の大きさに応じて α の値を自動的に変動させるような仕組みの検討または，重みづけ方法そのものの検討などが今後必要である．

6.3　スペクトルサブトラクションを用いた実験結果

学会講演を用いた音声認識に，さらにスペクトルサブトラクションの処理を施して実験を行った．あらかじめ雑音の平均スペクトルは分かっているものとし，スペクトルサブトラクションは，次の式により行っている．

$$|X(f)| = \max\{|X(f)| - \sqrt{a}|\overline{N}|, \sqrt{0.1}|\overline{N}|\} \qquad (16)$$

ここで，$X(f)$ は，音声認識において入力された雑音と音声の混合した入力信号のスペクトルを表す．また，$\overline{N}$ は，雑音信号の平均スペクトルを表す．$a = 1$ として実験を行った．

表 6 に実験結果を示す．clean な環境下では，あらかじめ雑音はないものと分かっているため，スペクトルサブトラクションを適用する前と結果は変わらない．SNR20dB の環境では，スペクトルサブトラクションを適用することにより，認識率が若干下がっている．これは，スペクトルサブトラクションの適用において，a の値を大きくしすぎたためと考えられる．他の雑音環境下では，認識率が向上している．スペクトルサブトラクションを加えた場合においても，MFCC と比べて SPEC2 では若干よい結果を示していることが分かる．

7. ま と め

本稿では，雑音に頑健な音声認識を行うために新たなスペクトル特徴量の抽出方法を提案した．スペクトル特徴量を抽出するため，対数スペクトル領域において，従来の MFCC を抽出するために用いる 3 つの正規化を行っている．すなわち，対数スペクトルの直流成分の除去，スペクトルの山と谷の強調，対数スペクトルの平均の除去である．これらの処理を行うことにより，MFCC と同等の認識性能が達成できることを確認した．

さらに，重みづけにより，雑音環境下において MFCC よりも高い認識率を確認した．

今後は 3.4 節に示した α の値や閾値 γ の検討，MLLR による学会講演音声への適応などがある．また，SNR により理想的な重みを求めておき，その重みと 3. 節の方法により求めた重みの比較を行い，どの程度正しく重みづけが行われているか検討を行う必要があると考えられる．さらに，理想的な重みに近づけるにはどうしたらよいか重みづけ手法の検討を行うことを予定している．そして，最終的には，スペクトルサブトラクションを行った後に本稿で示したマルチバンド音声認識を用い，雑音に頑健な音声認識を実現することを目標としている．

文　献

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

[2] A. Hagen, A. Morris, 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, 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, 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] H. Hermansky, S. Sharma and P. Jain, "Data-derived nonlinear mapping for feature extraction in HMM", Proc. ASRU99, 63-66, 1999.

[7] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech", Proc. ICASSP-97, 11:1255- 1258, 1997.

[8] 古井 貞熙, "ディジタル音声処理", 東海大学出版会, 1985.

[9] 杉山雅英, 鹿野清宏,"WLR 尺度による単語音声認識", 電子情報通信学会論文誌,J66-D,4,pp.385-392,1983.