論　文

変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識

金寺　登[†]　荒井　隆行[††]　船田　哲男[†††]

Robust Automatic Speech Recognition Emphasizing Important Modulation Spectrum

Noboru KANEDERA[†], Takayuki ARAI[††], and Tetsuo FUNADA[†††]

あらまし　CMS 法や動的特徴量を用いることにより，音声認識性能が向上することが知られている．これらの手法では特徴パラメータの時間軌跡を操作している．この時間軌跡を周波数次元で表したものは変調スペクトルと呼ばれる．よって CMS 法や動的特徴量は，変調スペクトルを操作しているものとみなせる．また音声認識情報のほとんどが 1〜16 Hz の変調周波数バンドに存在することが明らかになってきた．そこで本研究では，音声認識情報を担う変調スペクトル成分のみを特徴量として用い，数字音声認識実験を行った．広く用いられている RASTA では IIR フィルタを用いて約 1〜12 Hz の変調周波数バンドを抽出しているのに対し，本論文では位相ひずみの少ない FIR フィルタを用いることにより認識性能が向上することを確認した．また，この特徴量と一般によく用いられている動的特徴量を含めた MFCC を種々の雑音環境（SNR 10 dB）において比較した結果，認識誤り率が平均 3%改善されることを確認した．更に重要な変調周波数バンドを複数のバンドに分割すると，認識誤り率が平均 8%改善された．

キーワード　変調スペクトル，変調周波数，頑強な音声認識，特徴抽出

1. まえがき

現在広く用いられている CMS（cepstral mean subtraction）法 [1] や動的特徴量 [2] は，いずれも特徴パラメータの時間変化に注目している．この時間変化を周波数次元で表したものが変調スペクトルであり，その周波数次元は変調周波数と呼ばれる．図 1 は，CMS，動的特徴量 [2]，RASTA（RelAtive SpecTrAl processing）[3] の変調周波数特性を示している．CMS ではケプストラムの時間軌跡の直流成分を取り除く．1 秒間のケプストラム平均を引いた場合，CMS の変調周波数特性は図 1 のようになる．これによりマイクの周波数特性や通信伝送路におけるチャネル特性などによる乗法性雑音の影響を軽減することができる．動的特徴量の計算においては，ケプストラム係数の時間軌跡に対して回帰係数を求めている．これは時間軌跡に対するフィルタリングと等価であり，相対的に 10 Hz 付近の変調周波数成分が強調されるのに対し，その他の成分は軽減される．RASTA においては，約 1〜12 Hz の変調周波数成分が強調される．このように現在広く用いられているこれらの処理は，音声の変調スペクトル成分を効果的に加工している [3].

また知覚実験 [5], [6] により，一部の変調スペクトル成分が他の成分に比べて重要であることが知られている．この事実は日本語 [7] や英語 [10] においても確認されている．Drullman ら [5], [6] は，16 Hz 以下の低域通過フィルタリングや 4 Hz 以上の高域通過フィルタリングによって，音声のめいりょう度が低下しないことを示している．荒井ら [7], [8] は，Drullman らの研究をケプストラムに対応する対数領域に拡張し，低域/高域通過フィルタばかりでなく帯域フィルタを適用した．この結果，めいりょう度を保持するために必要なほとんどの情報が 1〜16 Hz の変調周波数バンドに存在することが明らかとなった．

[†] 石川工業高等専門学校，石川県
Ishikawa National College of Technology, Ishikawa-ken, 929–0392 Japan
[††] 上智大学，東京都
Sophia University, Chiyoda-ku, Tokyo, 102–8554 Japan
[†††] 金沢大学，金沢市
Kanazawa University, Kanazawa-shi, 920–8667 Japan

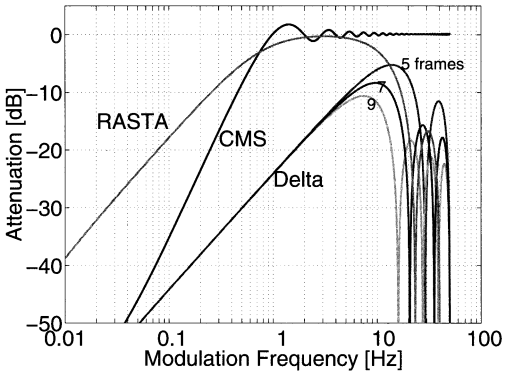


図 1　CMS, Delta, RASTA の変調周波数特性
Fig. 1　Modulation-frequency characteristics of CMS, Delta, and RASTA.

ASR（automatic speech recognition）に対して，金寺ら [12]〜[14] は重要な変調スペクトル成分を調査した．この結果，ASR にとって重要な情報のほとんどが 1〜16 Hz の変調周波数バンドに存在し，その中でも音声の音節速度（syllabic rate）に対応する 4 Hz 付近が最も重要であるという事実 [3], [8], [10] が再確認された．また雑音環境においては，2 Hz 以下や 16 Hz 以上の変調スペクトル成分が認識性能を劣化させることがあることがわかった．特に 1 Hz 以下の変調スペクトル成分は認識性能を著しく低下させる．

重要な変調周波数バンドのみ通過させ認識性能を向上させる方法として RASTA が知られている．RASTA では IIR フィルタを用いて約 1〜12 Hz の変調周波数バンドを抽出する．しかし RASTA を用いた場合，位相歪が原因で認識性能が劣化することが指摘されている [15].

そこで本論文では，知覚実験，ASR 実験により明らかになった音声認識情報を担う変調スペクトル成分のみを位相ひずみの少ないフィルタを用いて選択的に取り出し，その他の成分を取り除くことによって，雑音に強い音声認識を実現できると考え，様々な雑音環境下での音声認識実験を行った結果を報告する．良好な認識性能を得るためには，位相ひずみが少なく，0 Hz 付近での減衰量が大きい変調フィルタリングが必要である．そこで，位相ひずみの少ない直線位相 FIR フィルタを用いて以下の 2 種類の実験を行った．第 1 に FIR フィルタのタップ数を大きく設定し，0 Hz 付近での減衰量が大きい理想的な条件で，RASTA などの従来法との比較を行う．しかし 0 Hz 付近での急しゅんな減衰特性を保つためタップ数を大きく設定すると時間遅延が生じ実時間性が損なわれる．よって第 2 に実時間性を高めるため，タップ数を少なくし，比較的緩やかな特性をもつ変調フィルタを用いた場合についても認識実験を行う．

以下，**2.** では，各変調スペクトル成分の重要性を表す尺度として貢献度を定義する．また実時間での ASR を実現する上で必要となる緩やかな変調周波数特性をもった変調スペクトル成分抽出フィルタを用いた場合の各変調スペクトル成分の貢献度を調査する．この結果を急しゅんな変調周波数特性をもった変調スペクトル成分抽出フィルタにより調査した知覚実験や ASR 実験の結果と比較し，変調スペクトル成分抽出フィルタ特性による影響を明らかにする．**3.** では，音声認識にとって重要な変調スペクトル成分のみを特徴量として用いた場合の音声認識実験結果を示す．また緩やかな変調周波数特性をもつフィルタを使用し，重要な変調周波数バンドを複数のバンドに分割する方法を一般によく用いられている動的特徴量を用いた MFCC などと比較した結果についても報告する．

2.　重要な変調スペクトル成分

2.1　変調スペクトル成分の貢献度

本節では各変調スペクトル成分の重要性を表す尺度として貢献度を定義する．いくつかのバンドから得られた複数の認識率が与えられているとき，個々のバンドが認識性能にどの程度貢献するかを推定することが目的である．

まず，あらかじめケプストラム等の時間軌跡に種々の帯域フィルタを適用して得られたパラメータによる認識誤り率 $q(f_L, f_U)$ が得られているものとする．このとき認識誤り率は時間軌跡に対する帯域フィルタの低域遮断周波数 f_L と高域遮断周波数 f_U の関数である．オーバラップしない二つのバンド 1, 2 による認識誤り率をそれぞれ q_1, q_2 とする．ここで，オーバラップしないバンドは独立に認識結果に貢献すると仮定する．このとき，バンド 1, 2 を両方用いたときの認識誤り率 q_A は $q_A = q_1 q_2$ のようにそれぞれのバンドの誤り率の積になる．ここで A は，$A = \{1, 2\}$ のようにバンド番号の自然数を要素とする集合を表す．

一般に任意のバンドの集合 A を用いたときの誤り率は，

$$q_A = \prod_{i \in A} q_i \tag{1}$$

となる．積が和になるように両辺を対数に変換すると，

$$Q_A = \sum_{i \in A} Q_i \tag{2}$$

となる．ここで $Q_i = \log q_i$ である．式 (2) は次式のように変形できる．

$$Q_A = \sum_{\text{all } i} Q_i X_A(i) \tag{3}$$

ここで，$X_A(i)$ は，バンド i が A に属するかどうかを示す関数で次式で定義される．

$$X_A(i) = \begin{cases} 1 & i \in A \\ 0 & i \notin A \end{cases}$$

バンドの集合 $A, B, \cdots$ のそれぞれから得られた認識誤り率の対数 $Q_A, Q_B, \cdots$ がいくつか与えられているときに，$A, B, \cdots$ のすべての場合に式 (3) をなるべく満たす（2 乗誤差を最小にする）ような Q_i の推定量 $\widehat{Q}_i$ を求めたい．ところで，式 (3) は直線回帰の形式であるため，一般的な回帰計算法により回帰重み係数としての $\widehat{Q}_i$ とその信頼区間を求めることができる．結局，この $\widehat{Q}_i$ はバンド i が認識性能にどの程度貢献するかを対数尺度で表している．よって各変調スペクトル成分の認識性能への貢献度 C_i を

$$C_i = \exp(-\widehat{Q}_i) \tag{4}$$

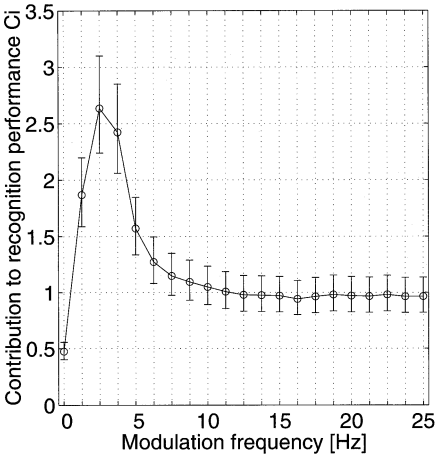
と定義する．

以上をまとめると，まずいくつかのバンドの集合 $A, B, \cdots$ から得られた認識誤り率の対数 $Q_A, Q_B, \cdots$ より式 (3) の回帰重み係数 $\widehat{Q}_i$ を求める．次に式 (4) により各変調スペクトル成分の認識性能への貢献度 C_i が求められる．

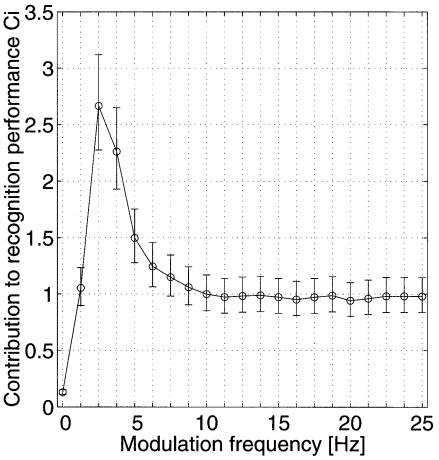
2.2 緩やかな変調周波数フィルタ特性を使用した場合の変調スペクトル貢献度

文献 [14] の実験においては，理想的な鋭い変調周波数特性が得られる条件で変調スペクトル貢献度を調査するため，長いタップの直線位相 FIR フィルタを使用した．しかしながら長いタップの FIR フィルタは長い時間遅れを生ずる．したがって実際の ASR 環境においては，短いタップをもつフィルタが望ましい．一般に短いタップのフィルタは鋭い周波数特性を得ることが難しい．よって，緩やかな周波数特性をもつフィルタを使用したとき，鋭い変調周波数特性をもつフィルタに比べ，各変調スペクトルの貢献度が変化するかどうかを確認する必要がある．そのため，短いタップのフィルタを用いて変調スペクトルの貢献度を調べた．ここでは，短いタップのフィルタの一種として少ない点数の DFT（Discrete Fourier Transform）を用いて以下の音声認識実験を行った．

まず 8 次の PLP（Perceptual Linear Predictive coding）[11] と対数パワーを求め，これらの各時間軌跡について，64 フレームを切り出し，ハミング窓を適用後，64 点の DFT を計算した．次に対象とする変調周波数バンドに対応する成分のみを得られた DFT の結果から抽出し，その時刻における特徴量とした．更



(a) Clean（雑音なし）



(b) Noisy（雑音あり）

図 2　DFT フィルタリングによる変調スペクトル貢献度
Fig. 2　Contribution to recognition performance for DFT filtering.

表 1　音声認識実験条件
Table 1　Conditions of ASR experiments.

Task	13 words Bellcore digit database (0–9, zero, oh, yes, no)
Recognizer	HMM (HTK Toolkit)
Training	150 speakers (75 males and 75 females)
Test	50 speakers (25 males and 25 females)
Sampling frequency	8 kHz
Window length	25 ms
Frame period	12.5 ms

表 2　使用した付加雑音
Table 2　Added noise.

babble	Voice Babble
buccaneer1	Buccaneer jet traveling at 190 knots; cockpit noise
buccaneer2	Buccaneer jet traveling at 450 knots; cockpit noise
destroyerengine	Destroyer: Engine room noise
destroyerops	Destroyer: Operations Room
f16	F-16 cockpit noise
factory1	Noise on floor of car factory
factory2	Noise in car production hall
hfchannel	HF Radio Channel Noise
leopard	Leopard 2 military vehicle noise
m109	M109 tank noise
machinegun	Machine Gun
pink	Pink Noise
volvo	Passenger compartment noise
white	White Noise

に時間軌跡切出し位置を 1 フレームずつシフトすることにより，すべてのフレームにおいて対象とする変調周波数バンドに対応する特徴量を抽出した．対象とする変調周波数バンドを様々に変化させ，対応するシステムの認識率を求めれば，**2.1** の方法により，各変調周波数成分が認識性能に寄与する貢献度 C_i を求めることができる．

図 2 は，単語音声に対する各変調スペクトル成分の貢献度を 95%信頼区間付きで示している．横軸は各 DFT フィルタの中心変調周波数を表している．この実験には，Bellcore digit database を使用した．図 2 (a) は雑音が少ない環境での結果を示しているのに対し，図 2 (b) においては，評価データが加法性雑音（コンピュータ雑音，SNR 10 dB）と乗法性雑音（HPF, 6 dB/oct）によって劣化された場合の結果を示している．その他の詳細な条件を表 1 に示す．

図中の貢献度 C_i は，対応する変調周波数バンドを含めることで，誤り率が 1/(貢献度）になることを表している．したがって，貢献度が 1 より大きければシステム性能が向上し，1 未満であればシステム性能が低下することを意味する．図 2 より，2〜10 Hz はクリーンな環境と雑音環境の両方で重要であった．また雑音環境では 2 Hz 未満の変調周波数成分の重要性は低くなった．特に 1 Hz 未満の変調周波数成分は著しく認識率を劣化させることがわかった．

一方，鋭い変調周波数特性をもつフィルタを使用した文献 [14] の結果では 1〜16 Hz，特に 2〜8 Hz が重要であった．この結果は今回の実験結果と一致することから，時間遅れが少ない緩やかな周波数特性をもつフィルタを使用しても，各変調スペクトルの貢献度の傾向は変化しないことがわかった．

3. 変調スペクトルの重要な成分のみを選択的に用いた音声認識

本章では，**2.2** の実験により明らかになった音声認識情報を担う変調スペクトル成分のみを選択的に取り出し，その他の成分を取り除くことによって，音声認識の耐雑音性がどのように変化するかを調査した結果について述べる．

3.1 実験条件

表 1 に示すように Bellcore digit database を用い，雑音データには NOISEX-92 database [17] を用いた．学習データには雑音を付加しないクリーンなデータを使用した．一方評価データには，付加雑音を SNR 10 dB になるように音声データに波形レベルで加算したものを用いた．付加雑音には表 2 に示す各雑音データの中よりランダムに切り出したものを用いた．これらの学習・評価データを Jack-knife 方式で 4 組用意した．HMM には単語単位のモデル（8 状態 6 出力分布，混合数 2）を用い，離散単語認識を行った．また学習・評価データにはあらかじめ切り出された音声を使用し，音声区間の検出は行わないこととした．

3.2 重要な変調周波数バンドのみを用いた音声認識

種々の雑音環境下において，すべての変調周波数バンドを用いた場合と重要な変調周波数バンドのみを用いた音声認識実験結果を表 3 に示す．表中の誤り率は **3.1** の 4 組の学習・評価データによる平均単語誤り率を示している．MFCC（Mel-Frequency Cepstral Coefficients）の次数は 12，PLP の次数は 8 とした．

表 3　重要な変調周波数バンドに対するフィルタリング前後での単語誤り率
Table 3　Word error rate with and without filtering important modulation frequency band.

	各特徴量の単語誤り率 [%]				
	フィルタリング前		フィルタリング後		
	MFCC +CMS	PLP +CMS	MFCC +FIR	PLP+FIR	
filter length	–	–	511	511	63
clean	1.7	1.5	2.2	2.5	2.1
[付加雑音]					
babble	21.5	27.5	22.1	22.0	21.3
buccaneer1	21.7	27.5	15.5	13.6	14.9
buccaneer2	21.8	26.6	17.7	15.8	17.0
destroyerengine	19.0	26.3	17.5	20.0	22.3
destroyerops	16.9	21.9	12.9	13.0	13.8
f16	21.5	27.5	18.8	15.2	17.7
factory1	20.9	27.0	17.2	15.2	17.0
factory2	16.0	18.0	13.7	10.4	12.3
hfchannel	23.1	22.8	20.8	16.6	17.3
leopard	15.5	18.7	14.9	11.4	12.5
m109	15.8	19.3	12.5	10.7	11.5
machinegun	50.2	44.6	37.0	35.6	35.0
pink	19.0	22.6	16.5	12.8	15.7
volvo	7.0	6.1	7.2	4.0	4.5
white	19.6	19.7	18.0	14.2	16.6
mean	20.6	23.7	17.5	15.4	16.6
feature size	39	27	39	27	27

MFCC と PLP には CMS を施した．また，いずれの特徴量も動的特徴量 (Δ, Δ^2) を併用した．MFCC+FIR や PLP+FIR は，MFCC や PLP の時間軌跡を FIR フィルタにかけた場合を示している．この FIR フィルタは図 3 の変調周波数特性をもつ 511 タップの直線位相 FIR フィルタで，2〜10 Hz の変調周波数バンドを通過させる帯域フィルタである．長いタップの FIR フィルタにより長い遅延を生じ，単語によっては単語長をオーバしてしまいフレーム方向のデータが不足することがある．そこで，本実験では音声の始端部分の数フレーム（第 2 フレームから第 5 フレーム）を音声の前後にフィルタリングに必要なフレーム数分繰り返しコピーした．表 3 中の clean は評価データに雑音を付加しない場合である．その他は，表 2 に示す雑音を付加した場合に対応している．mean は，雑音環境下での平均誤り率を示している．

重要な変調周波数バンドのフィルタリング前後での単語誤り率を比較すると，MFCC，PLP ともにフィルタリングにより雑音環境下での認識性能が向上していることがわかる．有意水準 1%で χ^2 検定を行った結果，clean（付加雑音なし）についてフィルタリング前後での認識性能に有意な差がなかった．一方，付加雑音がある場合については重要な変調周波数バンドのみ通過させる変調フィルタリングを用いた方が優れていることが確認できた．

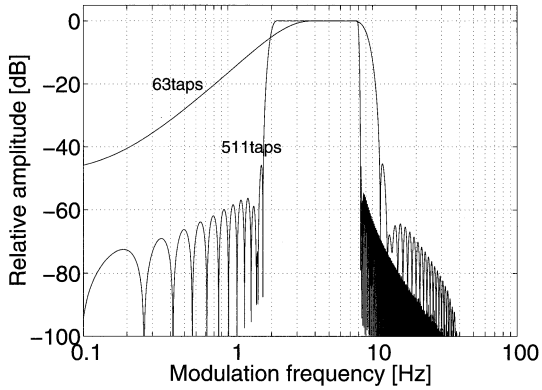


図 3　使用した FIR フィルタの変調周波数特性
Fig. 3　Modulation frequency characteristics of FIR filter used.

しかし，タップの長いフィルタは時間遅延が生じるため実用的ではない．そこで，短いタップのフィルタとして，63 タップの直線位相 FIR フィルタを用いた場合についても認識実験を行った．ここで，位相情報を保持しつつ，0 Hz において 40 dB 以上の減衰が得られる直線位相 FIR フィルタの中で 63 が最も小さいタップ数であったため，タップ数を 63 にした．12.5 ms のフレームシフトを用いた場合，63 タップのフィルタは 388 ms の遅延となる．図 3 より，63 タップのフィルタは 511 タップ（3188 ms の遅延）のフィルタに比べてかなり変調周波数特性が緩やかになっている．表 3 の実験結果より，雑音を付加した場合の平均単語誤り率について有意水準 1%で χ^2 検定を行った結果，両者に有意差はなかった．すなわち 63 タップ程度の緩やかな変調周波数特性をもつフィルタを用いても，鋭い変調周波数特性をもつフィルタを用いた場合と同程度の認識性能が得られることがわかった．

重要な変調周波数バンドのみ通過させ認識性能を向上させる方法として RASTA が知られている．RASTA では IIR フィルタを用いて約 1〜12 Hz の変調周波数バンドを抽出する．RASTA を用いて PLP+FIR と同一条件で音声認識実験を行った結果，雑音環境下での平均単語誤り率は 20.9%であった．この結果はフィルタリング前の PLP の結果よりも改善されているが，PLP+FIR の結果の方が優れていることがわかった．

PLP+FIR で使用したフィルタの帯域もこの RASTA フィルタとほぼ同じであるにもかかわらず，このような違いが得られた理由としては，PLP+FIR において位相ひずみの少ないフィルタ（直線位相 FIR フィルタ）を用いたためと考えられる．すなわち，変調スペクトルにおいては位相情報を保持することが重要であることを示唆している．

3.3 複数の変調スペクトル解像度を用いた音声認識

3.2 では，重要な変調周波数バンドのみを用いることによって耐雑音性が向上することを確認した．また，重要な変調周波数バンドを抽出する際に，遅延を少なくし実時間性を高めるため，63 タップ程度の緩やかな変調周波数特性をもつフィルタを用いても，鋭い変調周波数特性をもつフィルタを用いた場合と同程度の認識性能が得られることを確認した．本節では，重要な変調周波数バンドを複数のバンドに分割した場合の効果について述べる．

複数のバンドに分割する際に，重要な変調周波数バンド（2〜10 Hz）を対数的に等間隔になるように 2〜6 のバンドに分割した．図 2 を見ると，高い変調周波数に比べて低い変調周波数が重要であるといった偏りが見られる．よって重要な変調周波数バンドを効率的に分割するため，対数的に等間隔になるように変調周波数バンドを分割することとした．変調周波数バンドを抽出する変調フィルタには 63 タップの直線位相 FIR フィルタを用いた．8 次の PLP 及び対数パワーを複数の変調周波数バンドに分割した場合の ASR 実験結果を表 4 に示す．

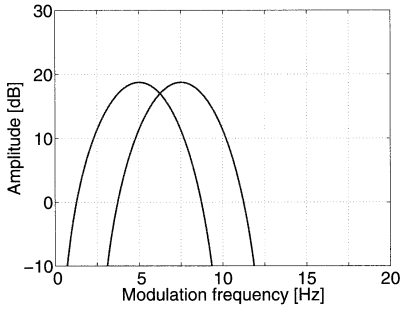
実験結果より，バンド数を増やすに従って認識性能が高くなることがわかった．表 3 の PLP+FIR（511 タップ）とこれらの結果を比較したところ，clean（付加雑音なし）については各特徴量間に有意な差はなかった．一方，雑音環境下においては，バンド数を 4 以上に分割したものが PLP+FIR に比べて優れていることが確認できた．バンド数を 3 以下に分割したものは PLP+FIR と有意な差がなかった．これより，重要な変調周波数バンドを複数のバンドに分割すると性能が向上することが確認された．また，この実験ではバンド分割数が 4〜6 の中で有意差はなかった．

なお，PLP+FIR の実験では動的特徴量 (Δ, Δ^2) を併用した．動的特徴量は図 1 のような変調周波数特性をもっているため，PLP+FIR の実験は，バンド数を 3 に分割した場合に対応する．

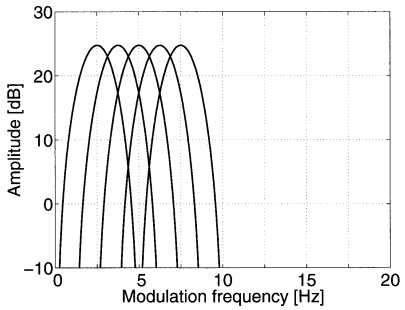
表 4　63 タップ FIR フィルタを用いて複数の変調周波数バンドに分割した場合の単語誤り率
Table 4　Word error rate using multiple modulation-frequency bands extracted by 63-tap FIR filters.

	単語誤り率 [%] 分割バンド数				
	2	3	4	5	6
clean	2.0	2.0	2.0	2.3	2.2
[付加雑音]					
babble	21.8	19.7	18.8	18.3	19.0
buccaneer1	14.5	13.7	11.5	11.5	11.4
buccaneer2	17.7	15.7	13.9	13.7	13.7
destroyerengine	23.7	19.8	18.2	18.5	18.2
destroyerops	12.5	12.3	11.3	11.6	11.7
f16	17.0	15.2	13.5	13.5	13.3
factory1	16.0	14.4	12.3	12.1	12.3
factory2	10.7	9.9	8.6	8.6	8.8
hfchannel	18.7	16.7	14.9	14.6	14.6
leopard	13.5	14.2	12.3	11.4	11.5
m109	11.0	10.4	9.4	9.6	9.2
machinegun	32.1	31.8	30.5	29.5	30.5
pink	13.8	12.5	11.3	10.8	10.7
volvo	5.2	5.4	5.0	4.8	4.7
white	15.0	13.3	11.8	11.8	12.0
mean	16.2	15.0	13.5	13.4	13.4
feature size	18	27	36	45	54

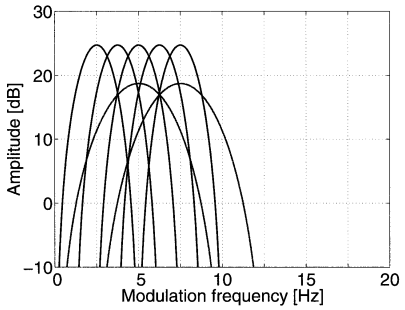
上記の実験では複数の変調周波数バンドに分割するために 63 タップの直線位相 FIR フィルタを使用したが，DFT を用いても変調周波数バンドに分割できる．そこで，短いタップのフィルタで，ある程度の周波数分離が可能なフィルタの一種として，図 4 に示す変調周波数特性をもつ 32 点及び 64 点 DFT フィルタリングを用いた音声認識実験を行った．図中の (a) は，12.5 ms のフレームシフトを用いた場合の 32 点 DFT の第 2，第 3 成分に対応するフィルタの変調周波数特性を示している．これらの成分の中心変調周波数は，5 Hz，7.5 Hz である．(b) は，64 点 DFT の第 2〜第 6 成分の変調周波数特性を示している．(c) は，(a) と (b) の両方の成分の変調周波数特性を示している．(c) は 32 点と 64 点の変調スペクトル成分を用いることにより，複数の解像度（バンド幅）による特徴を表現できるが，特徴量の数が増大してしまう．(d) では，(a) の 32 点 DFT の第 2，第 3 成分に加えて，低周波数成分を表現するため 64 点 DFT の第 2 成分を用いている．(d) のように，複数の解像度を併用する際に 16 点以下の DFT を用いることも考えられる．しかし 16 点以下の DFT を用いた場合，低域の変調周波数において十分な遮断特性が得られないため，64 点 DFT と 32 点 DFT を併用することとした．



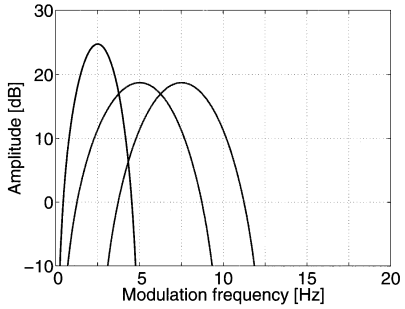
(a)32 点 DFT の第 2, 第 3 成分



(b)64 点 DFT の第 2〜6 成分



(c)32 点 DFT の第 2, 第 3 成分と
64 点 DFT の第 2〜6 成分



(d)32 点 DFT の第 2, 第 3 成分と
64 点 DFT の第 2 成分

図 4　DFT を用いた変調周波数フィルタ特性例
Fig. 4　Modulation-frequency characteristics of DFT filters.

表 5　DFT フィルタを用いて複数の変調周波数バンドに分割した場合の単語誤り率
Table 5　Word error rate using multiple modulation-frequency bands extracted by DFT.

	単語誤り率 [%]			
	(a)	(b)	(c)	(d)
clean	2.5	2.5	2.2	1.7
[付加雑音]				
babble	26.3	18.1	18.7	17.9
buccaneer1	15.0	10.1	10.3	10.0
buccaneer2	15.3	11.7	11.9	12.9
destroyerengine	22.3	16.6	17.3	16.2
destroyerops	14.0	10.8	11.0	10.7
f16	18.8	12.1	12.5	13.0
factory1	16.9	11.1	11.5	12.1
factory2	12.1	7.7	8.3	8.0
hfchannel	17.3	12.9	12.8	12.4
leopard	11.6	10.8	10.3	12.8
m109	11.9	9.0	9.2	9.0
machinegun	38.5	26.2	26.8	28.5
pink	14.3	9.5	9.9	9.9
volvo	5.4	4.9	4.1	4.3
white	15.1	10.7	10.5	10.3
mean	17.0	12.1	12.3	12.5
feature size	36	90	126	54

8 次の PLP と対数パワーを (a)〜(d) の複数の変調周波数バンドに分割した場合の ASR 実験結果を表 5 に示す．表 3 の PLP+FIR と (a)〜(d) を比較したところ，clean（付加雑音なし）については各特徴量間に有意な差はなかった．一方，雑音環境下においては，PLP+FIR に比べて (b)〜(d) が優れていることが確認できた．また，(b)〜(d) 間には有意な差はなかった．これより，重要な変調周波数バンドを複数のバンドに分割すると更に性能が向上することが DFT フィルタを用いた場合についても確認された．

性能的に有意差のない (b)〜(d) の中で，各特徴量の次元数（feature size）は表 5 の最下段に示すように (d) が最も小さいため，(d) が最も実用的である．(d) の結果は表 4 のバンド分割数 6 の結果よりも多少改善されている．

図 5 は，各特徴量の雑音環境下での平均誤り率を示している．図中の「modulation FT」は表 5 (d) に対応する．雑音環境下において，MFCC から重要な変調周波数バンドのみを抽出する（MFCC+FIR）ことによって，MFCC のみを用いる場合に比べ約 3%認識性能が向上した．また PLP から重要な変調周波数バンドのみを抽出する（PLP+FIR）ことによって，PLP のみを用いる場合に比べ約 8%認識性能が向上した．複数の変調周波数バンドに分割することによって，分

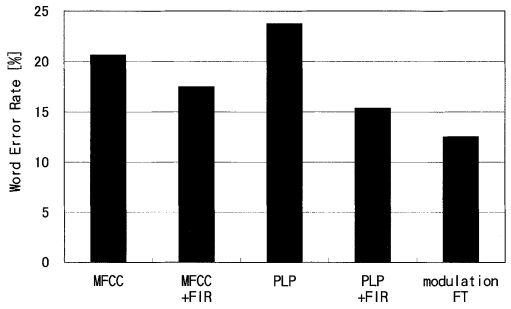


図 5　雑音環境下における平均単語誤り率
Fig. 5　Average word error rate for various noise conditions.

割しない場合（PLP+FIR）に比べ更に約 3%認識性能が向上した．この「modulation FT」は結局，一般によく用いられている MFCC に CMS 及び動的特徴量を併用した特徴量と比べて，約 8%認識性能が改善された．

4. むすび

ケプストラムや対数スペクトルの時間軌跡のフーリエ変換である変調スペクトル成分の中で，特に 2〜10 Hz の変調周波数バンドにほとんどの音声認識情報が存在するという実験結果に基づき，このバンドのみ通過させる位相ひずみの少ない変調フィルタリングを用いることで雑音環境下での音声認識性能が向上することが確認できた．また，重要な変調周波数バンドを複数のバンドに分割すると更に性能が向上することがわかった．

今後は，音声認識情報が存在する変調周波数成分を更に効率的に表現可能な特徴量を検討したい．

謝辞　多くの有益な示唆と音声認識実験環境を提供して下さった Oregon Graduate Institute of Science and Technology (OGI) & ICSI の Hynek Hermansky 教授，ICSI & University of California, Berkeley の Nelson Morgan 教授，Steven Greenberg 教授に深く感謝致します．また，OGI の Misha Pavel 教授，Sangita Sharma（現在 Intel），Narendranath Malayath, Sarel van Vuuren，そして University of California, Davis の Carlos Avendano の協力に深く感謝致します．有益な御助言を頂きました東京理科大学の藤崎博也教授，Indian Institute of Technology の B. Yegnanarayana 教授にも深く感謝致します．本研究の一部は，平成 12 年度科学技術振興事業団地域研究開発促進拠点事業の一環により行われた．

文　献

[1] B.S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” J. Acoust. Soc. Am., vol.55, no.6, pp.1304–1312, June 1974.

[2] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. Acoust. Speech & Signal Process., vol.ASSP-34, no.1, pp.52–59, Feb. 1986.

[3] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech & Audio Process., vol.2, no.4, pp.578–589, Oct. 1994.

[4] T. Houtgast and H.J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” J. Acoust. Soc. Am., vol.77, no.3, pp.1069–1077, March 1985.

[5] R. Drullman, J.M. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” J. Acoust. Soc. Am., vol.95, no.2, pp.1053–1064, Feb. 1994.

[6] R. Drullman, J.M. Festen, and R. Plomp, “Effect of reducing slow temporal modulations on speech reception,” J. Acoust. Soc. Am., vol.95, no.5, pp.2670–2680, May 1994.

[7] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, “Intelligibility of speech with filtered time trajectories of spectral envelopes,” Proc. ICSLP, pp.2490–2493, Philadelphia, 1996.

[8] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” J. Acoust. Soc. Am., vol.105, no.5, pp.2783–2791, May 1999.

[9] H. Hermansky, N. Morgan, and H. Hirsch, “Recognition of speech in additive and convolutional noise based on RASTA spectral processing,” Proc. IEEE ICASSP, pp.II-83–II-86, Minneapolis, MN, 1993.

[10] S. Greenberg, “Understanding speech understanding — Towards a unified theory of speech perception,” Proc. ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, pp.1–8, Keele, England, 1996.

[11] H. Hermansky, “Perceptual linear predictive (PLP) analysis for speech,” J. Acoust. Soc. Am., vol.87, no.4, pp.1738–1752, April 1990.

[12] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the importance of various modulation frequencies for speech recognition,” Proc. Eurospeech, pp.1079–1082, Rhodes, Greece, Sept. 1997.

[13] N. Kanedera, H. Hermansky, and T. Arai, “On properties of modulation spectrum for robust automatic speech recognition,” Proc. IEEE ICASSP, pp.II-613–II-616, Seattle, WA, May 1998.

[14] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the relative importance of various components of

the modulation spectrum for automatic speech recognition," Speech Commun., vol.28, pp.43–55, May 1999.
[15] V. Johan and B. Louis, "Channel normalization techniques for automatic speech recognition over thetelephone," Speech Commun., vol.25, pp.149–164, 1998.
[16] 金寺　登，荒井隆行，H. Hermansky, 船田哲男，"ロバストな音声認識実現を目的とした変調スペクトル特性の検討，" 信学技報，SP97-70, Dec. 1997.
[17] A. Varga and H. J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., vol.12, no.3, pp.247–251, 1993.
[18] 金寺　登，荒井隆行，船田哲男，"複数の変調スペクトル解像度を用いた音声認識の耐雑音性，" 信学技報，SP98-51, July 1998.

（平成 12 年 7 月 11 日受付，13 年 1 月 5 日再受付）

金寺　　登　（正員）

昭 60 電通大・通信卒．昭 62 東大大学院工学系研究科電子工学専攻（修士課程）了．同年石川高専助手．現在同高専助教授．音声認識の研究に従事．IEEE，日本音響学会，情報処理学会各会員．博士（工学）．

荒井　隆行　（正員）

1989 上智大・理工卒．1994 同大大学院理工学研究科電気・電子工学専攻（博士後期課程）了．同年上智大助手．1992〜1993 並びに 1995〜1996 Oregon Graduate Institute of Science and Technology（USA）客員研究員．1997〜1998 California 大学 Berkeley 校付属研究機関 International Computer Science Institute（USA）客員研究員．1998 上智大専任講師．現在同大助教授．音声・聴覚・信号処理などの研究に従事．共著「ディジタル信号と超関数」，監訳「音声の音響分析」，「音声・聴覚のための信号とシステム」．IEEE，アメリカ音響学会，日本音響学会等各会員．博士（工学）．

船田　哲男　（正員）

昭 41 金沢大・工・電子卒．昭 46 名大大学院博士課程了．昭 46 金沢大・講師．現在同大教授．生体情報処理，音声情報処理の研究に従事．共著「情報科学の基礎」，「数値解析の基礎」など．IEEE，日本音響学会，日本 ME 学会，情報処理学会各会員．