社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

モーラ情報を用いたフィルタバンクによる孤立単語認識

谷口　勝則† 　村上　仁一† 　池原　悟†

† 鳥取大学工学部　〒 680-0945 鳥取県鳥取市湖山町南 4-101
E-mail: †{ktanigut,murakami,ikehara}@ike.tottori-u.ac.jp

あらまし　ケプストラムを用いた孤立単語認識では, モーラ情報を用いることによりピッチの影響が分離されるので, 認識精度を向上させることが可能であると知られている．本研究では，ピッチの中にも認識に有効な情報が含まれていることに着目し，ピッチの情報が直接利用できる方法について検討した．ケプストラムを用いるとホルマントとピッチが分離され，ピッチの情報が利用できない．そのため，本研究ではパワースペクトラムを使用した．さらに，人間の聴覚特性を考慮してパワースペクトラムを少ない次数で効率的に表現するために，メル分割されたフィルタバンクの対数パワーを用い，モーラ情報を考慮した音素ラベルを作成して孤立単語認識を行った結果，全ての話者に対して認識精度が向上することが分かった．

キーワード　孤立単語認識, メルフィルタバンク, モーラ数, モーラ位置

Isolated-word recognition with Filter Bank using Mora Position and Mora Length

Katsunori TANIGUCHI†, Jin'ichi MURAKAMI†, and Satoru IKEHARA†

† Faculty of Engineering,Tottori University,Koyama-cho 4-101,Tottori,680-0945 Japan
E-mail: †{ktanigut,murakami,ikehara}@ike.tottori-u.ac.jp

Abstract　In isolated-word recognition using cepstrum，the influence of pitch is divided by using mora information，and it is known that it is possible to improve recognition accuracy．In this paper，we notice that effective information for recognition is included in pitch，and we examined how to use the information on pitch directly．If cepstrum is used，formant and pitch is divided，and the information on a pitch cannot be used．In this paper，I considered use of power spectrum．Then，we performed isolated-word recognition that was used Mel Filter Bank instead of cepstrum．Consequentry，we found that recognition accuracy improved to many speakers.

Key words　Isolated-word recognition，Mel Filter Bank，Mora length，Mora position

1. はじめに

現在の孤立単語認識では，特徴パラメータにケプストラムなどが使用されている．ケプストラムは音源パルスの周期が長く，ホルマントが相互によく分離しているときはよい結果が得られるが，音源ピッチが高く，ホルマントが互いに接近しているときは両成分の分離が不完全となり，誤差が生じてしまう [1].

この問題を解決する方法として，音素ラベリングでは，音素ラベル中の母音・撥音を単語のモーラ数および単語のモーラ位置で分類し，ピッチの影響を分離する方法が提案されており，ラベリング精度が向上することが知られている [2]．また，この結果を受けて音声認識にも単語のモーラ数・モーラ位置で音素ラベルの分類を行ったところ，認識精度が向上したと報告されている [3].

本研究では，ピッチを分離するのではなく，ピッチの情報を直接利用できる方法について検討する．ケプストラムを使用するとホルマントとピッチが分離されてピッチの情報が利用できないため，パワースペクトラムの使用を考える．さらに，人間の聴覚特性を考慮し，パワースペクトラムを少ない次数で効率的に表現するために，メル分割されたフィルタバンクの対数パワーを使用する．また，音素ラベル中の母音・撥音を単語のモーラ数および単語のモーラ位置で分類し，音素 HMM を学習する．

フィルタバンクは，音声の自動ラベリングにおいてラベリング精度が向上することが報告されている [4]．本研究においても，フィルタバンクを使用して孤立単語認識を行い，精度向上の効果と有効性を検証する．

2. モーラ情報の利用

2.1 モーラ情報とピッチ情報の関係

特定話者の単語の発音において，単語のモーラ数および単語のモーラ位置 (本論文では以後，単語のモーラ数と単語のモーラ位置をモーラ情報と呼ぶ) が決まれば，ピッチ周波数はほぼ決まることが知られている [5].

図 1 は [5] から引用したもので，単一話者のナレータが発声した 5 モーラ語の地名 (固有名詞)2800 件のピッチ周波数の平均値と分散を示している．ピッチ周波数の解析には xwave ＋ [6] を使用している．図 1 より，ピッチ周波数は単語に関係なく単語のモーラ数および単語のモーラ位置でほぼ決定することが分かる．また，4,6 モーラ語も同様の傾向を示し，分散も 5 モーラ語と同程度であったと報告されている．

また，固有名詞だけでなく，普通名詞においても図 1 と同様の傾向があり，ピッチ周波数を単語のモーラ数および単語のモーラ位置である程度決定できることが報告されている [7].

本研究では，同じ種類の音素の中でも，モーラ数，モーラ位置によってピッチ周波数が異なることに注目した．音素ラベルをモーラ数，単語のモーラ位置で分類することによって，同程度のピッチ周波数を持つ音素が集まる．これにより，ピッチの情報が音素 HMM に反映され，認識精度が向上すると期待できる．

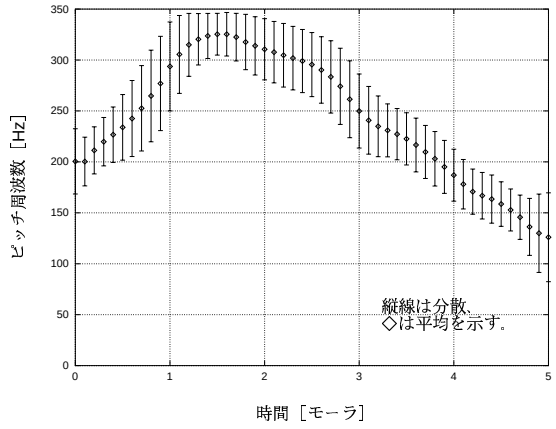

図 1　モーラ情報とピッチ周波数の関係
Fig. 1　Relation between mora information and pitch frequency

2.2 フィルタバンクを使用した特徴パラメータ

2.2.1 使用するフィルタバンク (FBANK)

本研究では，音声波形の中に含まれるピッチの情報を認識に利用することを考える．そこで，特徴パラメータにパワースペクトラムを使用する．さらに，人間の聴覚特性を考慮し，パワースペクトラムを少ない次数で効率的に表現するために，メル分割されたフィルタバンクの対数パワーを使用する．使用するフィルタバンクは図 2 のような三角形の形をしていて，メルスケールに沿って等間隔に置かれている [8]．周波数メル分割の式を (1) に示す．

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (1)$$

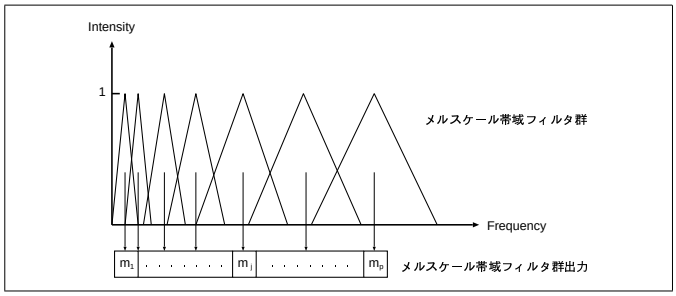

図 2　メルスケールフィルタバンク
Fig. 2　Mel scale filter bank

本研究では，音声データをフーリエ変換し，周波数成分の全域にこのフィルタをかけ，その対数パワー成分を特徴パラメータとして使用する．

2.2.2 従来の特徴パラメータ (MFCC)

一般の音声認識では，特徴パラメータとして式 (2) に示すようなケプストラムが主に使用されている．m_j は対数フィルタバンクの振幅を表し，N はフィルタバンクのチャンネル数を表している [8].

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos \frac{\pi i}{N}(j - 0.5) \qquad (2)$$

しかし，ケプストラムを使用するとホルマントとピッチが分離されてしまうためにピッチの情報を認識に利用することができない．

本研究では，パワースペクトラムにメル分割されたフィルタバンクの対数パワーを使用して孤立単語認識を行うとともに，従来から使用されているケプストラムを使用した場合についても同様に孤立単語認識を行い，フィルタバンクの有効性を調査する．

3. 評 価 実 験

3.1 音声データベース

音声データベースとして，ATR の単語発話データベース Aset(5240 単語) を使用する．話者には男性話者 3 名 (mau,mmy,mtk)，女性話者 3 名 (faf,ftk,fyn) の計 6 話者を使用する．そして，Aset のデータ 5240 単語を奇数番と偶数番に分け，奇数番の単語で音素 HMM を学習し，偶数番の単語を認識する．

3.2 モーラ情報を使用したラベルファイルの作成

本研究では，音声波形データファイルと音声ラベルファイルを使用して，学習および認識を行う．ピッチの情報を認識に利用するために，データベース中の全音声ラベルファイルの母音と撥音を単語のモーラ数および単語のモーラ位置で分類し，モーラ情報を含む音声ラベルファイルを作成する．分類方法は，母音・撥音の前方に単語のモーラ数，後方にモーラ位置情報を付け加えて分類する．子音については，ピッチの情報が少なく，モーラ情報の効果が小さいと考えたため，分類せずに使用する．分類例を表 1 に示す．

音声ラベルが atama である場合，単語のモーラ数は 3 なの

表 1　母音と撥音の分類例
Table 1　Examples of classifications of vowels and syllabic nasals

分類前	a	t	a	m	a		
分類後	3a1	t	3a2	m	3a3		
分類前	a	r	u	b	a	m	u
分類後	4a1	r	4u2	b	4a3	m	4u4
分類前	a	ng	k	e	e	t	o
分類後	5a1	5ng2	k	5e3	5e4	t	5o5

で母音の前方に 3 を付け，後方にモーラ位置を付ける．1,2,3 番目の音素 a は，分類後はそれぞれ 3a1,3a2,3a3 という音素に置き換える．置き換えられた音素 a はモーラ位置がそれぞれ異なるため，異なった音素として扱う．

3.3　音素 HMM の作成方法

3.3.1　使用する音素 HMM

母音と撥音をモーラ情報を使用して分類することによって，作成される音素 HMM の数は増加する．しかし，学習データの数は一定であるために，音素 HMM1 つあたりの学習データ数が減少し，音素 HMM の信頼度が低下してしまう．これに対処するために，本研究では半連続型 HMM を使用する．これにより，ガウス分布の数を固定し，音素 HMM の信頼度の低下を防ぐことが可能となる [9].

3.3.2　音素 HMM の作成手順

本研究では，音素 HMM を図 3 で示す手順によって学習する．モーラ情報を使用した音素 HMM は，(a) 連続型 HMM の学習，(b) 半連続型 HMM の学習，(c) モーラ情報を使用した音素 HMM の学習の 3 つのステップから作成される．モーラ情報を使用しない場合は，(a),(b) の手順で音素 HMM を作成する．

(a) 連続型 HMM の作成

学習データにモーラ情報を使用していないラベルファイルと波形データを使用する．この学習データをもとに Viterbi alignment を用いて初期モデルを作成する．この初期モデルを Baum-Welch アルゴリズムを用いて再推定し，連結学習を行って連続型 HMM を作成する．

(b) 半連続型 HMM の作成

連続型 HMM から，全ての音素 HMM の混合ガウス分布を共通にした半連続型 HMM を作成し，連結学習を行う．

(c) モーラ情報を使用した音素 HMM の作成

半連続型 HMM から，母音，撥音の音素 HMM を複製して，それらにモーラ情報を付与し，これをモーラ情報を使用した音素 HMM の初期モデルとする．これにより，音素の種類は 26 種類から 114 種類となる．さらに連結学習を行い，モーラ情報を使用した半連続型 HMM を作成する．

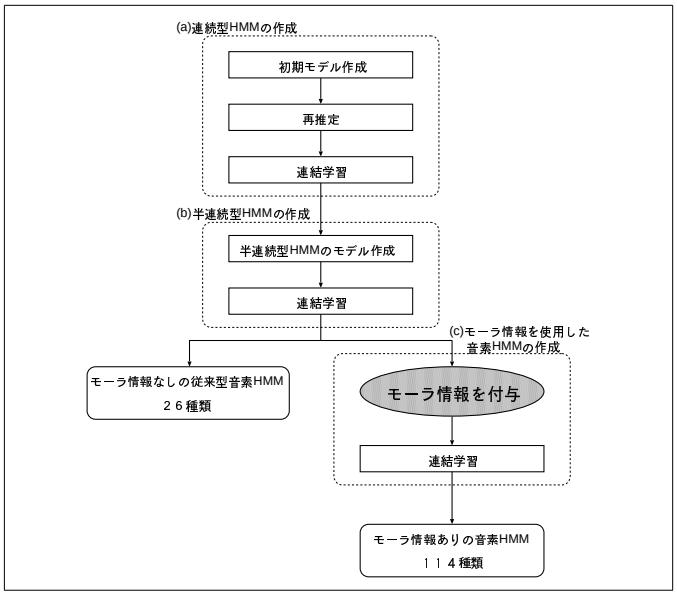

図 3　音素 HMM の作成手順
Fig. 3　The procedure of creation of Phoneme HMM

3.4　実験条件

孤立単語認識を行うツールには HTK [8] を使用し，実験は表 2 の条件で行う．特徴パラメータには 3 章で述べたメル分割されたフィルタバンクの対数パワー (FBANK) に，対数パワー，Δ対数パワー，Δ FBANK を合わせたものを使用する．また，stream 数を 3 に設定し，FBANK，対数パワーとΔ対数パワー，Δ FBANK をそれぞれ別の多次元ガウス分布で表現する．

連続型 HMM の初期モデルの混合分布数は，モーラ情報を使用する母音と撥音の音素については 4 とし，それ以外の音素に

表 2　実験条件
Table 2　Experimental conditions

基本周波数	16kHz	
分析窓	Hamming 窓	
分析窓長	20ms	
フレーム周期	5ms	
音響モデル	3 ループ 4 状態	
	半連続分布型	
stream 数	3	
特徴パラメータ	FBANK(16 次) +	
	Δ FBANK(16 次) +	
	対数パワー (1 次), Δ対数パワー (1 次)　(計 34 次)	
連続型 HMM の	(母音・撥音)	
初期モデルの	FBANK	4
混合分布数	Δ FBANK	4
	対数パワー, Δ対数パワー	2
	(それ以外の音素)	
	FBANK	2
	Δ FBANK	2
	対数パワー, Δ対数パワー	2
半連続型 HMM の	FBANK	256
混合分布数	Δ FBANK	256
	対数パワー, Δ対数パワー	32
学習 DB	2620 単語	
音素数	約 15500	
母音数	約 8000	
評価 DB	2620 単語	
音素数	約 15500	
母音数	約 8000	

ついては 2 とする．Δ FBANK の混合分布数については，母音・撥音の音素を 4，それ以外の音素を 2 とする．また，対数パワー，Δ対数パワーの混合分布数は全ての音素について 2 とする．

半連続型 HMM モデルの混合分布数については，FBANK と Δ FBANK を 256 とし，対数パワー，Δ対数パワーを 32 とする．

音素 HMM の混合ガウス分布には，Diagonal-covariance(以下，Diagonal) と Full-covariance(以下，Full) の 2 種類を使用する．本研究では，Diagonal と Full のそれぞれについて，孤立単語認識における FBANK パラメータの効果を調査する．

3.5 評価方法

モーラ情報を使用した場合とモーラ情報を使用しない場合で孤立単語認識を行い，正しく認識できた単語数と誤った認識をした単語数から認識率を求める．また，モーラ情報を使用することによって改善された誤りの度合を示す改善率も併せて求める．認識率と改善率からモーラ情報の効果を調査する．認識率と改善率の式を式 (3)，式 (4) に示す．

$$\text{認識率 (\%)} = \frac{\text{正しく認識できた単語数}}{\text{評価単語総数}} \times 100 \quad (3)$$

$$\text{改善率 (\%)} = \frac{\text{誤り数}_{\text{モーラ無し}} - \text{誤り数}_{\text{モーラ有り}}}{\text{誤り数}_{\text{モーラ無し}}} \times 100 \quad (4)$$

ここで 誤り数$_{\text{モーラ有り}}$ は，モーラ情報を使用したときの誤り数を，誤り数$_{\text{モーラ無し}}$ はモーラ情報を使用していないときの誤り数をそれぞれ示している．また，従来から使用されているメルケプストラム (以下 MFCC) においても，表 2 と同様の条件で孤立単語認識を行い，FBANK との比較を行う．

3.6 実験結果

表 2 の条件で音素 HMM の混合ガウス分布に Diagonal を使用した孤立単語認識の結果を表 3，Full を使用した結果を表 4 に示す．モーラ無しはモーラ情報を使用せずに音素 HMM の学習を行った結果で，モーラ有りはモーラ情報を使用して行った結果である．表中には，認識率と改善率を示した．

Diagonal の場合では，モーラ情報を用いることによって 6 話者の平均で 94.85%の認識率が得られ，29.18%の誤りの改善が見られた．Full の場合では，モーラ情報を用いることによって 6 話者の平均で 97.59%の認識率が得られ，34.55%の誤りの改善が見られた．

Diagonal と Full を比較すると，Full の方が認識率で約 3%高く，改善率においても約 5%増加している．このことから，モーラ情報は Full において効果が大きいことが分かる．

4. 考　　察

4.1 モーラ情報による認識誤りの改善

4.1.1 連続母音に対する効果

モーラ情報を使用することによって正しく認識できた単語の多くは，連続母音を含むものであった．特に，長母音を含む単語を短母音に誤認識してしまう単語に効果が見られた．表 5 に話者 mau において改善された単語例を示す．改善された理由として，長母音をモーラ位置で異なる音素として区別することで認識できるようになったためと考えられる．

表 3　FBANK の実験結果 (Diagonal-covariance)

Table 3　The results of the experiment of FBANK(Diagonal-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	93.13%(2413/2591)	95.29%(2469/2591)	31.46%
mmy	90.78%(2352/2591)	93.17%(2414/2591)	25.94%
mtk	92.78%(2404/2591)	95.06%(2463/2591)	31.55%
faf	92.78%(2404/2591)	95.02%(2462/2591)	31.02%
ftk	93.28%(2417/2591)	95.37%(2471/2591)	31.03%
fyn	93.59%(2425/2591)	95.18%(2466/2591)	24.70%
平均	92.72%(14415/15546)	94.85%(14745/15546)	29.18%

表 4　FBANK の実験結果 (Full-covariance)

Table 4　The results of the experiment of FBANK(Full-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	96.68%(2505/2591)	98.30%(2547/2591)	48.84%
mmy	95.91%(2485/2591)	97.11%(2516/2591)	29.25%
mtk	96.18%(2492/2591)	97.30%(2521/2591)	29.29%
faf	97.07%(2515/2591)	98.46%(2551/2591)	47.37%
ftk	95.95%(2486/2591)	97.41%(2524/2591)	36.19%
fyn	96.10%(2490/2591)	96.95%(2512/2591)	21.78%
平均	96.31%(14973/15546)	97.59%(15171/15546)	34.55%

表 5　改善された単語例 (話者 mau)

Table 5　The improved examples of words(Speaker mau)

改善された単語			誤認識した単語	
単語	音素列		単語	音素列
構成	k o o s e i	→	個性	k o s e i
習慣	sh j u u k a ng	→	主観	sh j u k a ng
葬式	s o o sh i k i	→	組織	s o sh i k i
統計	t o o k e i	→	時計	t o k e i
風刺	h u u sh i	→	節	h u s i
報道	h o o d o o	→	歩道	h o d o o
誘拐	y u u k a i	→	愉快	y u k a i

表 6　改善されなかった単語例 (話者 mau)

Table 6　The examples of words which have not been improved(Speaker mau)

正しい単語			誤認識した単語	
単語	音素列		単語	音素列
回覧	k a i r a ng	→	階段	k a i d a ng
該当	g a i t o o	→	解答	k a i t o o
機会	k i k a i	→	期待	k i t a i
パイプ	p a i p u	→	タイプ	t a i p u
悲観	h i k a ng	→	時間	zh i k a ng
町	m a ch i	→	マッチ	m a q ch i
利息	r i s o k u	→	規則	k i s o k u

4.1.2 改善されなかった単語

モーラ情報を使用しても改善されなかった単語例を，話者 mau の場合について，表 6 に示す．改善されなかった単語の多くは，子音の誤認識であった．本研究では，母音と撥音のみに

モーラ情報を使用し，その他の音素については従来の音素モデルと同じであったために改善されなかったと考えられる．

4.2 MFCC との比較

表 2 と同様の条件で，特徴パラメータに MFCC を使用し，音素 HMM の混合ガウス分布に Diagonal を使用した孤立単語認識の結果を表 7，Full を使用した結果を表 8 に示す．また，表 3，表 4 の FBANK の孤立単語認識の結果と表 7，表 8 の MFCC の孤立単語認識の結果を 6 話者の平均でまとめたものを図 4 に示す．

図 4 より，FBANK と MFCC の認識率を比較すると，モーラ情報の有無にかかわらず，Diagonal では MFCC の方が高く，Full では FBANK の方が高くなった．FBANK は MFCC に比べパラメータが独立していないので，Diagonal では高い認識率が得られなかったが，Full を使用することで音素 HMM の特徴を表現でき，認識率向上に大きな効果をもたらしたと考えられる．

表 7　MFCC の実験結果 (Diagonal-covariance)
Table 7　The results of the experiment of MFCC(Diagonal-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	95.95%(2486/2591)	96.84%(2509/2591)	21.90%
mmy	93.05%(2411/2591)	95.06%(2463/2591)	28.89%
mtk	94.48%(2448/2591)	96.14%(2491/2591)	30.07%
faf	93.40%(2420/2591)	95.68%(2479/2591)	34.50%
ftk	93.09%(2412/2591)	95.72%(2480/2591)	37.99%
fyn	94.87%(2458/2591)	95.75%(2481/2591)	17.29%
平均	94.14%(14635/15546)	95.86%(14903/15546)	29.42%

表 8　MFCC の実験結果 (Full-covariance)
Table 8　The results of the experiment of MFCC(Full-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	96.57%(2502/2591)	97.57%(2528/2591)	21.90%
mmy	94.79%(2456/2591)	96.41%(2498/2591)	31.11%
mtk	95.91%(2485/2591)	97.41%(2524/2591)	36.79%
faf	96.02%(2488/2591)	97.11%(2516/2591)	27.18%
ftk	95.18%(2466/2591)	97.11%(2516/2591)	40.00%
fyn	95.99%(2487/2591)	96.91%(2511/2591)	23.08%
平均	95.74%(14884/15546)	97.09%(15093/15546)	31.57%

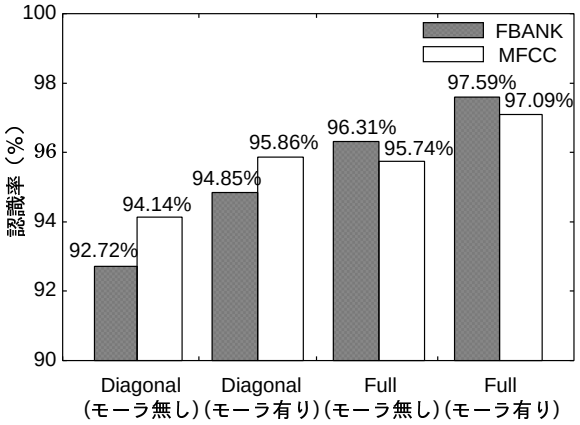

図 4　FBANK と MFCC の比較
Fig. 4　Comparison of FBANK and MFCC

4.3 FBANK における男性話者と女性話者の比較

表 4 の FBANK の結果から男性話者と女性話者でそれぞれ平均したものを表 9 に示す．また，表 8 の MFCC の結果から男性話者と女性話者でそれぞれ平均したものを表 10 に示す．

表 10 の MFCC を使用した場合では，女性話者は男性話者に比べて認識率が低くなっているのに対し，表 9 の FBANK を使用を使用した場合では，女性話者の認識率が高くなっていることが分かる．表 9，表 10 は Full の結果であるが，Diagonal においても同様の傾向が見られた．このことから，FBANK は女性話者において特に有効な特徴パラメータであると考えられる．

女性話者の音声は男性話者の音声に比べピッチ周波数が高いことが知られている．そのため，MFCC ではピッチの影響が完全に分離できていなかったこともあり，良い結果が得られなかった．しかし，FBANK を用いてピッチを直接使用することにより，ピッチの情報が認識に利用でき，そのために女性話者において良い効果をもたらしたと考えられる．

表 9　FBANK における男性，女性話者の比較 (Full-covariance)
Table 9　Comparison of male speakers and female speakers for FBANK(Full-covariance)

話者	モーラ無し	モーラ有り	改善率
男性話者	96.26%(7482/7773)	97.57%(7584/7773)	35.05%
女性話者	96.37%(7491/7773)	97.61%(7587/7773)	34.04%

表 10　MFCC における男性，女性話者の比較 (Full-covariance)
Table 10　Comparison of male speakers and female speakers for MFCC(Full-covariance)

話者	モーラ無し	モーラ有り	改善率
男性話者	95.75%(7443/7773)	97.13%(7550/7773)	32.42%
女性話者	95.73%(7441/7773)	97.04%(7543/7773)	30.72%

4.4 連続型 HMM との比較

本節では，半連続型 HMM との比較を行うために連続型 HMM での孤立単語認識を行った．実験は表 2 と同様の条件で行い，連続型 HMM の作成は図 3 の (a) の手順により行った．また，FBANK と比較するために MFCC においても同様に実験を行った．Diagonal を使用した場合の結果を表 11，Full を使用した場合の結果を表 12 に示す．なお，連続型 HMM にモーラ情報を使用した場合では，学習データの不足により音素 HMM の学習が行えない場合が多いため，モーラ情報を使用しない場合でのみ実験を行った．

表 11，表 12 より，FBANK と MFCC で認識率を比較すると，半連続型 HMM の場合と同様に Diagonal では MFCC の方が高く，Full では FBANK の方が高くなった．このことから，FBANK は連続型 HMM においても Full において効果が高いことが分かる．

また，表 11，表 12 の FBANK の結果を表 3，表 4 の半連続型 HMM の結果と比較すると，モーラ情報を使用しない場合，

表 11　連続型 HMM の実験結果 (Diagonal-covariance)
Table 11　The results of experiment of continuous HMM(Diagonal-covariance)

話者	FBANK	MFCC
mau	91.32%(2366/2591)	95.10%(2464/2591)
mmy	87.34%(2263/2591)	93.17%(2414/2591)
mtk	90.20%(2337/2591)	93.90%(2433/2591)
faf	90.74%(2351/2591)	92.32%(2392/2591)
ftk	89.77%(2326/2591)	93.05%(2411/2591)
fyn	92.67%(2401/2591)	94.02%(2436/2591)
平均	90.34%(14044/15546)	93.59%(14550/15546)

表 12　連続型 HMM の実験結果 (Full-covariance)
Table 12　The results of experiment of continuous HMM(Full-covariance)

話者	FBANK	MFCC
mau	96.99%(2513/2591)	96.53%(2501/2591)
mmy	95.79%(2482/2591)	95.45%(2473/2591)
mtk	95.95%(2486/2591)	95.99%(2487/2591)
faf	96.95%(2512/2591)	96.68%(2505/2591)
ftk	96.10%(2490/2591)	95.52%(2475/2591)
fyn	96.33%(2496/2591)	96.10%(2490/2591)
平均	96.35%(14979/15546)	96.04%(14931/15546)

Diagonal では全ての話者に対して半連続型 HMM の方が認識率が高いのに対し，Full では話者 mmy を除く 5 話者について連続型 HMM の方が認識率が高くなった．しかし，モーラ情報を使用した場合と比較すると，全ての話者に対して半連続型 HMM の方が高い認識率となった．したがって，FBANK にはモーラ情報を使用した半連続型 HMM が最も有効であると考えられる．

4.5　音素 HMM における stream 数

過去の研究では，一般的に音素 HMM を作成する際に，stream 数を 1 に設定して学習を行っている [3]．しかし，FBANK を使用した場合に stream 数を 1 にすると，パラメータ推定の際に逆行列の計算ができない場合があった．

この問題を解決するために，stream 数を 3 に設定して FBANK，Δ FBANK，対数パワーをそれぞれ独立した成分としてパラメータ推定を行ったところ，音素 HMM の作成が可能となった．そのため，本研究では stream 数を 3 に設定して実験を行った．

5. ま と め

本研究では，モーラ情報を用いたフィルタバンクによる孤立単語認識について検討した．ピッチの情報を音声認識に利用するために，単語の母音・撥音を単語のモーラ数および単語のモーラ位置で分類し，特徴パラメータに FBANK を使用した．その結果，Diagonal では 6 話者平均で 94.85%の認識率が得られ，29.18%の誤りが改善された．Full では 6 話者平均で 97.59%の認識率が得られ，34.55%の誤りが改善された．

FBANK と MFCC で認識率を比較した場合，Diagonal では MFCC の方が良く，Full では FBANK の方が良い結果が得られたことから，FBANK は Full において効果が高いことが分かった．

また，男性話者と女性話者で認識率を比較した場合，Diagonal，Full 両方共に女性話者の方が良い結果が得られた．このことから，FBANK は女性話者に有効な特徴パラメータであることが分かった．

さらに，連続型 HMM による孤立単語音声認識を行い，半連続型 HMM との比較を行った．その結果，モーラ情報を使用した半連続型 HMM において認識率が最も高くなった．したがって，FBANK にはモーラ情報を付与した半連続型 HMM が最も有効であることが分かった．

今後の課題として，FBANK にモーラ情報を加えた音素 HMM を使用し，前後の音素を考慮した tri-phone モデルでの孤立単語認識や，雑音下における孤立単語認識を検討することなどが挙げられる．

文　献

[1] 中田和男:改訂 音声，日本音響学会編，コロナ社 (1977)
[2] 前田智広，村上仁一，池原悟:モーラ情報を用いた音素ラベリング方式の検討，信学技報，SP2001-53，(2001-8)
[3] 妹尾貴宏，村上仁一，前田智広，池原悟:モーラ情報を用いた単語音声認識の研究，信学技報，SP2001-45，(2001-8)
[4] 米澤朋子，水野秀之，阿部匡伸:HMM 音素モデルによる自動ラベリングのロバスト性の検討，信学技報，SP2002-74，(2002-8)
[5] 水澤紀子，村上仁一，東田正信:音節波形接続による単語音声合成，信学技報，SP99-2，(1999-05)
[6] Introducing ESPS/waves+ with EnSigTM Entropic Research Laboratory,Inc.
[7] 石田隆浩，村上仁一，池原悟:音節波形接続型音声合成の普通名詞への応用，信学技報，SP2002-25，(2002-5)
[8] HTK Ver2.2 reference manual，1997 Cambridge University
[9] X.D.Huang, Y.Ariki, M.A.Jack, HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION，Edinburgh University Press，1990