# FBANK を用いた孤立単語音声認識

○谷口勝則，村上仁一，池原悟 (鳥取大)

## 1. はじめに

現在の孤立単語音声認識では，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，音源ピッチが高く，ホルマントが互いに接近しているときに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ため，ピッチ周波数とモーラ情報に強い相関があることを利用して，単語のモーラ数および単語のモーラ位置で音素ラベルを分類する方法が提案されている．この方法によりピッチの影響を HMM において分離することができ，それらの HMM を認識に使用した場合，認識精度の向上が報告されている [2].

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．そこで，ケプストラムの代わりにメル分割されたフィルタバンクの対数パワー (FBANK) を特徴パラメータとして使用する．さらに，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 を学習して孤立単語音声認識を行い，精度向上の効果と有効性を検証する．

なお，FBANK 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 [3].

## 2. 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 [4]．これは固有名詞に限らず，普通名詞においても同様に決定できることが報告されている [5]．また，これを利用した音素ラベリングや孤立単語音声認識の研究も行われており，モーラ情報の有効性が確認されている [2][6].

## 3. FBANK を使用した特徴パラメータ

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，ケプストラムの代わり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．フィルタバンクは三角形の形をしており，メルスケールに沿って等間隔に配置されている [7]．周波数メル分割の式を式 (1) に示す.

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

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フィルタをかけ，その対数パワー成分を FBANK として特徴パラメータに使用する．

## 4. 評価実験

### 4.1 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．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．音声ラベルを atama とした場合の分類例を表 1 に示す.

表 1: 母音と撥音の分類例

| | | | | | |
|---|---|---|---|---|---|
| モーラ情報無し | a | t | a | m | a |
| モーラ情報有り | 3a1 | t | 3a2 | m | 3a3 |

atama のモーラ数は 3 なので母音の前方に 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.

### 4.2 半連続型 (semi-coninuous)HMM の使用

モーラ情報を使用して音素ラベルを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.

### 4.3 音素 HMM の作成方法

本研究では，以下の 3 ステップで作成する.

**(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**

(a) 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**

(b) から，母音，撥音の音素 HMM を複製し，モーラ情報を使用した音素 HMM の初期モデルとする．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.

### 4.3 実験条件

孤立単語音声認識を行うツールには HTK[7] を使用する．実験条件を表 2 に示す．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 のデータベース 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に使用する.

また，Diagonal-covariance(以下，Diagonal) と Full-covariance(以下，Full) の 2 種類を音素 HMM の混合ガウス分布に使用して実験を行う.

表 2: 実験条件

| | | | |
|---|---|---|---|
| 基本周波数 | 16kHz | 音響モデル | 3 ループ 4 状態 |
| 分析窓 | Hamming | | 半連続分布型 |
| 分析窓長 | 20ms | stream 数 | 3 |
| フレーム周期 | 5ms | | |
| 特徴パラメータ | FBANK(24 次), Δ FBANK(24 次), 対数パワー (1 次), Δ対数パワー (1 次) (計 50 次) | | |
| 半連続型 HMM の混合分布数 | FBANK | | 256 |
| | Δ FBANK | | 256 |
| | 対数パワー, Δ対数パワー | | 32 |

### 4.4 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音声認識を行い，正しく認識でき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 (2)，式 (3) に示す.

$$\text{認識率 (\%)} = \frac{\text{正しく認識できた単語数}}{\text{評価単語総数 (2591 単語)}} \times 100 \quad (2)$$

$$\text{改善率 (\%)} = \frac{\text{誤り数}_{\text{モーラ無し}} - \text{誤り数}_{\text{モーラ有り}}}{\text{誤り数}_{\text{モーラ無し}}} \times 100 \quad (3)$$

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．

### 4.5 実験結果

表 2 の条件で Diagonal における孤立単語音声認識の結果を表 3 に示す．表中の括弧内には正しく認識できた単語数を示す．また，Full を使用した結果を同様に表 4 に示す．

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 94.96%の認識率が得られ，25.21%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 97.48%の認識率が得られ，30.92%の誤りの改善が見られた．

表 3，表 4 の結果より，FBANK にモーラ情報を使用することによって認識率が向上し，Full においてその効果が大きいことが分かった．

表 3: FBANK の実験結果 (Diagonal-covariance)

| 話者 | モーラ情報無し | モーラ情報有り | 改善率 |
|---|---|---|---|
| mau | 93.48%(2422) | 95.52%(2475) | 31.36% |
| mmy | 91.20%(2363) | 93.48%(2422) | 25.88% |
| mtk | 93.79%(2430) | 95.56%(2476) | 33.54% |
| faf | 93.59%(2425) | 94.87%(2458) | 19.88% |
| ftk | 93.67%(2427) | 95.52%(2475) | 29.27% |
| fyn | 93.86%(2432) | 94.83%(2457) | 15.72% |
| 平均 | 93.27%(14499) | 94.96%(14763) | 25.21% |

表 4: FBANK の実験結果 (Full-covariance)

| 話者 | モーラ情報無し | モーラ情報有り | 改善率 |
|---|---|---|---|
| mau | 96.91%(2511) | 97.99%(2539) | 35.00% |
| mmy | 95.99%(2487) | 97.61%(2529) | 40.38% |
| mtk | 95.91%(2485) | 97.22%(2519) | 32.08% |
| faf | 97.07%(2515) | 97.88%(2536) | 27.63% |
| ftk | 96.14%(2491) | 97.61%(2529) | 38.00% |
| fyn | 96.14%(2491) | 96.60%(2503) | 12.00% |
| 平均 | 96.36%(14980) | 97.48%(15155) | 30.92% |

## 5. 考察

### 5.1 MFCC との比較

比較のために，MFCC を使用した場合について孤立単語音声認識を行った．MFCC を使用した孤立単語音声認識では，フォルマント成分のみを使用し，ピッチの情報は通常使用しない．そのため，本研究においてもフォルマント成分のみを使用した．特徴パラメータは MFCC(12 次)，ΔMFCC(12 次)，対数パワー (1 次)，Δ対数パワー (1 次) とした．その他の条件については，表 2 と同様にして実験を行った．Diagonal の結果を表 5，Full の結果を表 6 に示す．

表 5，表 6 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 を使用することでパラメータの特徴をより良く表現でき，MFCC よりも高い認識率が得られたと考えている．

### 5.2 Triphone モデルとの比較

Triphone は前後の音素環境を考慮したモデルであり，現在の音声認識の中で最も標準的な方法として知られている．そこで，FBANK に Triphone モデルを使用して孤立単語音声認識を行い，モーラ情報を使用した場合との比較を行った．表 2 と同様の実験条件を用い，Full のみで実験を行った．その結果を表 7 に示す．

表 5: MFCC の実験結果 (Diagonal-covariance)

| 話者 | モーラ情報無し | モーラ情報有り | 改善率 |
|---|---|---|---|
| mau | 95.48%(2474) | 96.57%(2502) | 23.93% |
| mmy | 92.90%(2407) | 95.18%(2466) | 32.07% |
| mtk | 94.02%(2436) | 96.49%(2500) | 41.29% |
| faf | 93.36%(2419) | 95.75%(2481) | 36.05% |
| ftk | 94.94%(2460) | 96.06%(2489) | 22.14% |
| fyn | 93.75%(2429) | 95.91%(2485) | 34.57% |
| 平均 | 94.08%(14625) | 95.99%(14923) | 32.36% |

表 6: MFCC の実験結果 (Full-covariance)

| 話者 | モーラ情報無し | モーラ情報有り | 改善率 |
|---|---|---|---|
| mau | 96.68%(2505) | 98.07%(2541) | 41.86% |
| mmy | 94.91%(2459) | 96.53%(2501) | 31.82% |
| mtk | 94.09%(2438) | 96.37%(2497) | 38.56% |
| faf | 95.95%(2486) | 97.68%(2531) | 52.38% |
| ftk | 95.83%(2483) | 97.18%(2518) | 23.15% |
| fyn | 95.02%(2462) | 96.49%(2500) | 29.46% |
| 平均 | 95.41%(14833) | 97.05%(15088) | 35.76% |

表 7: Triphone モデルの実験結果 (Full-covariance)

| 話者 | 認識率 |
|---|---|
| mau | 98.53%(2553) |
| mmy | 96.59%(2513) |
| mtk | 97.65%(2530) |
| faf | 96.02%(2488) |
| ftk | 96.41%(2498) |
| fyn | 95.64%(2478) |
| 平均 | 96.87%(15060) |

表 7 より，6 話者の平均で比較すると，モーラ情報を使用した方が高い認識率が得られることが分かった．このことから，FBANK にはモーラ情報を使用した Full のモデルが最も有効であると言える．

## 6. まとめ

本研究では，FBANK による孤立単語音声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，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，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 94.96%の認識率が得られ，25.21%の誤りが改善された．Full では 6 話者平均で 97.48%の認識率が得られ，30.92%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，Triphone モデルとモーラ情報を使用したモデルとで比較した結果，モーラ情報を使用したモデルの方が高い認識率が得られることが分かった．

今後の課題として，不特定話者における孤立単語音声認識を検討することなどが挙げられる．

## 参考文献


[1] 中田和男：改訂 音声, 日本音響学会, コロナ社 (1977)
[2] 妹尾, 村上, 池原：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)
[3] 米澤, 水野, 阿部：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)
[4] 水澤, 村上, 東田：音節波形接続による単語音声合成, 信学技報 SP99-2,(1999)
[5] 石田, 村上, 池原：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)
[6] 前田, 池原, 村上：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, pp.25-30 (2001).
[7]HTK Ver2.2 reference manual，1997 Cambridge University