平成14年度　修士論文

FBANKを用いた孤立単語音声認識

Isolated-word speech recognition using FBANK

指導教官	池原 悟	教授
	村上 仁一	助教授
	徳久 雅人	助手

鳥取大学大学院 工学研究科 知能情報工学専攻

M01T2019H　谷口勝則

内容梗概

ケプストラムを用いた孤立単語音声認識では，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 (FBANK) 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音声認識を行った結果，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．

目 次

図版目次

表目次

第1章　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良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 (FBANK) 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音声認識を行い，精度向上の効果と有効性を検証する．

第2章　音声認識

2.1　音声認識の原理

2.1.1　音声認識の基本的構成

音声認識とは，音声波に含まれる意味内容に関する情報 (言語情報) を，コンピュータや電気回路によって抽出し，判定することである．音声認識装置には次のような利点がある．

- 音声で入力できれば，タイプライタや押しボタンなどに比べて，操作に熟練がいらないので使いやすい．
- 情報の入力速度がタイプライタの約 3〜4 倍，手書き文字入力の約 8〜10 倍と速い．
- 手足，眼，耳などの器官を同時に別の作業に使いながら並列的に，あるいは動きまわりながら情報の入力ができる．
- マイクロホン，電話機などを入力端末に使えるので，経済的であり，既存の電話機をそのまま用いて，遠隔地から入力することができる．

一般に人が発声した音声をコンピュータなどで認識する課程は，図 2.1 のように，通信理論 (情報処理論) の問題として，確率モデルを用いて定式化することができる．話者が文を考える課程が文発生部で，これを通信理論の情報源に対応させる．音声認識システムを音響処理部と言語復号部に分ける．話者による発生部と音響処理部を合わせて，一つの音響チャンネルとしてモデル化し，これを歪み (雑音) のある通信路に対応させる．音声認識システムの主たる部分である言語復号部を復号部に対応させる．話者はまず，情報源に対応する文 w を頭の中で組み立て，それに基づいて，その話者の発話習慣に従って音声波形 s を生成する．s には通常，話者の個人差，負荷雑音，伝送歪みなどが重畳している．音響処理部は音声波形データの分析・変換を行って，例えば短時間スペクトルなどの時系列データ (ベクトル系列)y を出力する．言語復号部は y から送信文の推定値として $\hat{w}$ を出力する．$\hat{w}$ は，事後確率 $P(w|y)$ が最大になるように推定する．$P(w|y)$ を直接求めるのは，通常困難であるので，ベイズ則によって，次式を満たすように推定する．

$$P(\hat{w}|y) = \max_{w} \frac{P(y|w)P(w)}{P(y)} \tag{2.1}$$

ここで，$P(y)$ は w に無関係であるので無視することができる．尤度 $P(y|w)$ は音響モデルによって得られ，文 w が発生される事前確率 $P(w)$ は言語モデルによって得ら

れる．したがって，音声認識のポイントは，$P(y|w)$ と $P(w)$ をいかに計算するか，言い換えると音響モデルと言語モデルをいかに作るかにある．

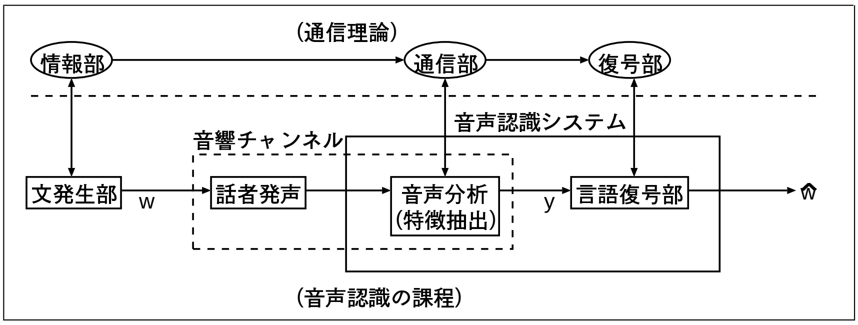

図 2.1: 音声認識課程の確率モデル

2.1.2　音声認識の分類と課題

音声認識の形態は，次のように分類できる．

(a) 認識対象音声による分類

- 孤立単語音声認識：区切って発声した単語を認識する．
- 連続音声認識：単語を連続して発声した音声を認識する．

連続音声認識は，語彙以外の言語的知識を用いるか否かによって，次のように分類される．

- 連続単語音声認識：連続数字音声認識のように，比較的少数の語彙を対象とし，言語的知識は用いず音響的特性によって認識する．
- 文音声認識，会話音声認識：比較的多数の語彙を対象とし，言語的知識を用いて，その意味内容を理解しようとする．

この内，孤立単語音声認識と連続単語音声認識では，通常，$P(w)$ は全て等しいと考えて，$P(y|w)$ とともに $P(w)$ が認識判定に重要な役割を果たす．

(b) 対象話者による分類

- 不特定話者型：誰の音声でも認識できる．
- 特定話者型：学習を行った特定の話者の音声のみを認識する．
- 話者適応型：新しい話者の短い音声を用いて認識装置を自動的にその話者の声に適応させた後に (あるいは適応させると同時に) 認識する．

音声認識の難しさは，次の 4 点に集約することができる．

(1) 音響モデルに関連して
- 調音結合の効果…単語や文章を発声したときの各音素のスペクトルが，調音結合のために前後の音素の影響を受けて変形する．
- 区分化の難しさ…音声は草書で書かれた文字列のようなもので，音素や単語の境界を決定する区分化が難しい．音声の始端と終端を検出するのも難しい．
- 個人差による変動…同じ言葉でも，話し方や発声器官の違いのため，人によって音声が異なる．同じ人の音声でも雑音や伝送歪みなどによって変動する．

(2) 言語モデルに関連して
- 言語モデルのあいまい性…話し言葉には書き言葉と違う種々の文法的ゆらぎがあり，モデル化が極めて難しい．

2.2 HMM

2.2.1 HMM の理論

隠れマルコフモデル (Hidden Markov Model：HMM) は，2.1 節で述べた，音声認識の通信理論 (情報処理理論) に基づく定式化において，$P(y|w)$ を求めることを目的とに考え出された方法である．DP マッチングにおいては，単語や音素に関して，標準的な時系列を標準パターンとして用いるが，HMM では，各単語や音素を標準的な確率状態遷移モデル (マルコフモデル) で表現する．非定常信号源である音声を，定常信号源の連結で表す統計的信号源モデルである．HMM は DP マッチングによる方法に比べて，スペクトル時系列の統計的変動をモデルのパラメータに反映させることができる特徴があるが，逆にモデルパラメータを決定するための学習処理がやや複雑になる．

音声認識に用いられる HMM は，left-to-right 型で一つの初期状態と一つの最終状態がある構造が多く，図 2.2 は最もよく用いられるベイキス (Bakis) モデルと呼ばれる型の例である．q_i は状態，a_{ij} は状態遷移行列，b_ij は出現確率を示している．また，図の状態遷移のアークに付けられた数値 a_{ij} は，状態 q_i から状態 q_j への状態遷移確率を表し，状態数を S とすると $S \times S$ の行列で表現できる．通常，音声パターンには，時間的に非可逆性の性質があるので，$i > j$ なら $a_{ij} = 0$ である．各状態 q_i の初期確率を π_i で表し，最終状態の集合を F で表す．

$b_{ij}(y)$ は，状態 q_i から状態 q_j への遷移で，スペクトルパターン y が観測 (出力) される観測 (出現) 確率を表し，$\{b_{ij}(y)\}$ を出現確率行列と呼ぶ．出現確率が状態遷移に独立に遷移元の状態によってのみ決定されるモデルも考えられ，この場合は $b_i(y)$ と書くことができる．この両者は数学的に等価変換が可能である．出現するスペクトルパターンに関しては，連続値として表す場合 (連続分布モデル，連続型 HMM：continuous HMM) と，有限個 (K 個) のシンボルの組合せで表現する場合 (離散分布モデル，離散型 HMM：discrete HMM) ，また連続分布モデルと離散分布モデルの中間の性質を持つ半連続分布モデル (半連続型 HMM：semi-continuous HMM) がある．以下にそれぞれの特徴を示す．

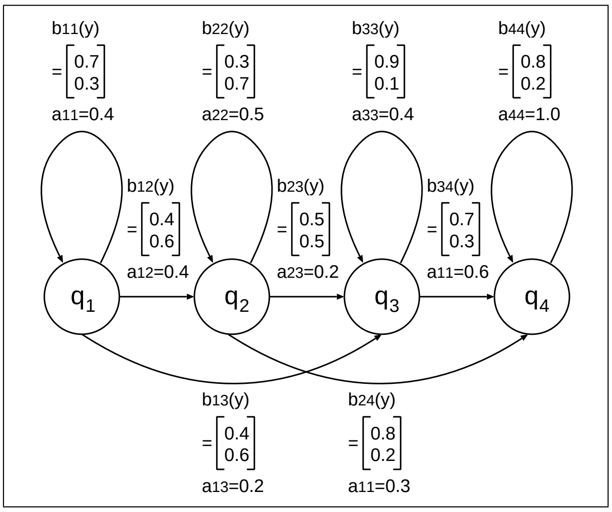

図 2.2: ベイキス型 HMM の例

- 離散分布モデル (離散型 HMM：descrete HMM)
出現するスペクトルパターンは，有限個のシンボルの組合せで表現される．出力確率は，スペクトルパターンのクラスタ化 (ベクトル量子化) によって，代表スペクトルパターン (符号ベクトル) を生成し，各符号ベクトルの出現確率の組合せによって表現する．

- 連続分布モデル (連続型 HMM：continuous HMM)
出現するスペクトルパターンは，連続値で表現される．出力確率は単一ガウス分布 (正規分布) または混合ガウス分布で表現される．パラメータの自由度を減らすために無相関ガウス分布が用いられることが多い．

- 半連続分布モデル (半連続型 HMM：semi-continuous HMM)
連続分布モデルと離散分布モデルの中間の性質を持つ．これは，連続分布モデルにおける混合ガウス分布を，全てのモデルの全ての状態で共通にし，各分布の重みだけを変えるようにしたものである．結び混合分布モデル (tied-mixture model) とも呼ばれる．離散分布モデルにおける各符号ベクトルに確率分布を持たせたものと見ることもできる [5)]．

実際の音声認識に用いる HMM においては，対象に応じて適切に状態数やモデル構造 (遷移構造) を決定し，スペクトルパターンの表現法 (離散分布モデルの場合はその種類 K，連続分布モデルの場合はそのモデル化の方法) を決定する必要がある．これらの数や複雑さ，すなわちモデルの自由度を大きくすれば，きめ細かい変動が表現できるが，推定すべきモデルパラメータが多くなり，推定精度が悪くなる．図 2.2 における数値例は，以後の説明のために特に簡略化したものであり，離散分布モデルで出力符号ベクトルを $\{a, b\}$ の二つに限り，図の [ ] 内にそれぞれの出現確率を示している．この例では，遷移確率行列は，

$$A = (a_{ij}) = \begin{bmatrix} 0.4 & 0.4 & 0.2 & 0.0 \\ 0.0 & 0.5 & 0.2 & 0.3 \\ 0.0 & 0.0 & 0.4 & 0.6 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \tag{2.2}$$

となり，$\pi_1 = 1$，$\pi_i = 0(i > 1)$，$F = \{q_4\}$ である．

2.2.2　認識アルゴリズム

$y = \{y_1, y_2, \cdots, y_T\}$ を観測 (出力) 系列とする．具体的には，スペクトルやケプストラムベクトルの時系列である．このとき，各 HMM モデルによって y が生起する確率 (尤度) $P(y/M)$(M は HMM によって表現される単語や音素に対応) を求め，最大確率 (最大尤度) を与えるモデルを選んで，これを認識結果とする．図 2.3 に HMM を用いた孤立単語音声認識の方法を示す．

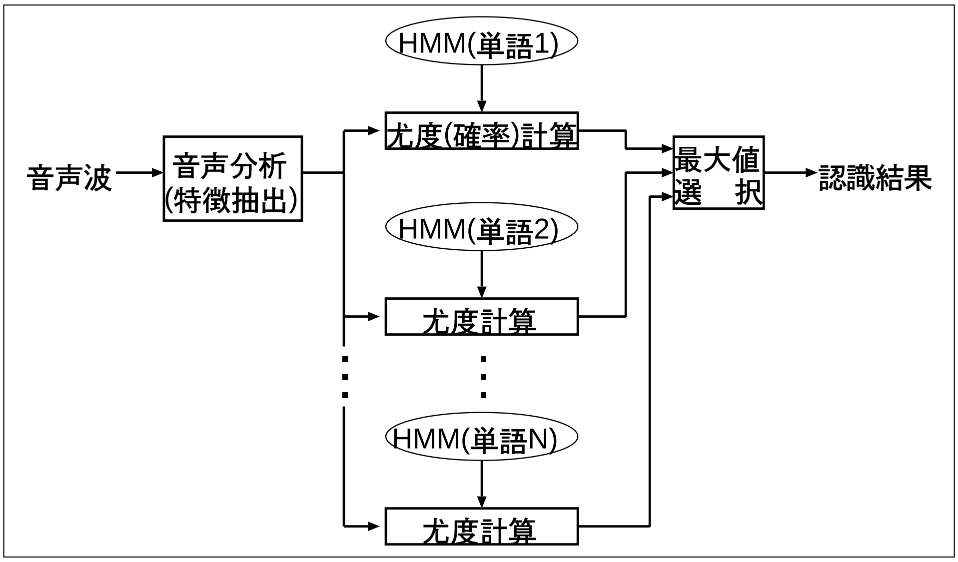

図 2.3: HMM を用いた孤立単語音声認識の方法

$q = q_{i0}, q_{i1}, \cdots, q_{iT}$ を状態遷移系列 (ただし $q_iT \in F$) とすれば，

$$P(y|M) = \sum_{i_0, i_1, \cdots, i_T} P(y|q, M) \cdot P(q|M) \tag{2.3}$$

と書ける．ここで図 2.2 の例について，符号ベクトル系列「aba」が出力される確率を求めると，この場合の状態遷移系列は，時間を横に状態を縦に並べた図 2.4 の平面で，左上隅から右下隅に至る経路に対応し，次の 7 通りである．なお，図 2.4 中の太線の部

分は，最も可能性の高い単独の状態遷移系列，すなわちビタービ法による系列を示している．

$$
\begin{array}{ll}
q_1 \to q_1 \to q_2 \to q_4 & P_1 = 0.4 \times 0.7 \times 0.4 \times 0.6 \times 0.3 \times 0.8 = 0.016128 \\
q_1 \to q_1 \to q_3 \to q_4 & P_2 = 0.4 \times 0.7 \times 0.2 \times 0.6 \times 0.6 \times 0.7 = 0.014112 \\
q_1 \to q_2 \to q_2 \to q_4 & P_3 = 0.4 \times 0.4 \times 0.5 \times 0.7 \times 0.3 \times 0.8 = 0.013440 \\
q_1 \to q_1 \to q_3 \to q_4 & P_4 = 0.4 \times 0.4 \times 0.2 \times 0.5 \times 0.6 \times 0.7 = 0.006720 \\
q_1 \to q_2 \to q_4 \to q_4 & P_5 = 0.4 \times 0.4 \times 0.3 \times 0.2 \times 1.0 \times 0.8 = 0.007680 \\
q_1 \to q_3 \to q_3 \to q_4 & P_6 = 0.2 \times 0.4 \times 0.4 \times 0.1 \times 0.6 \times 0.7 = 0.001344 \\
q_1 \to q_3 \to q_4 \to q_4 & P_7 = 0.2 \times 0.4 \times 0.6 \times 0.3 \times 1.0 \times 0.8 = 0.011520
\end{array}
$$

それぞれの確率は上に示す通りとなるので，

$$P(aba|M) = P_1 + P_2 + P_3 + P_4 + P_5 + P_6 + P_7 = 0.070944 \tag{2.4}$$

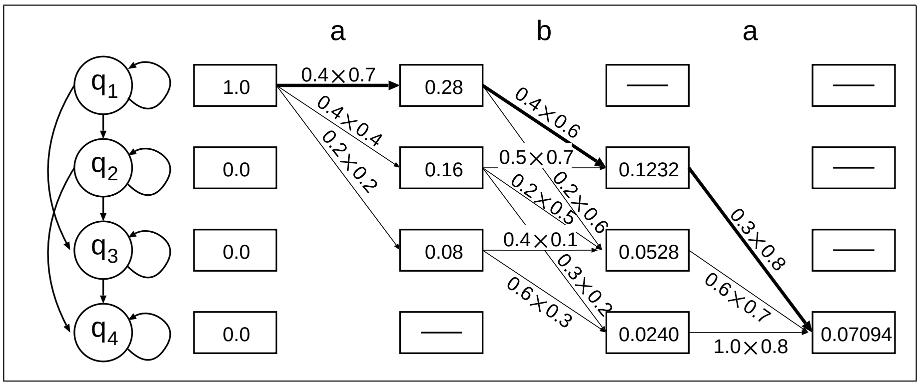

図 2.4: 図 2.2 の HMM から符号ベクトル系列「aba」が出力される確率をトレリス法によって計算する手順

一般に $P(y|M)$ の値は，以下に述べるトレリス (trellis) アルゴリズムで求められる．フォワード (forward) 変数 $\alpha(i,t)$ を定義し，符号ベクトル y_t を出力して状態 q_i にある確率とすれば，$i = 1, 2, \cdots, S$ において，

$$\alpha(i,t) = \begin{cases} \pi_i & (t = 0) \\ \sum_j \alpha(j, t-1) \cdot a_{ji} \cdot b_{ji}(y_t) & (t = 1, 2, \cdots, T) \end{cases} \tag{2.5}$$

であるから，これを計算して，最後に

$$P(y|M) = \sum_{i, q_i \in F} \alpha(i, T) \tag{2.6}$$

を求めればよい．図 2.4 の数値は，実際にトレリス法を実行した結果を示したものである．

$P(y|M)$ を厳密に求めないで，近似的に，モデル M が符号ベクトル系列 y を出力するときの，最も可能性の高い状態系列上での出現確率を用いることも考えられる．図

2.4 の例では，太線で示した $q_1 \rightarrow q_1 \rightarrow q_2 \rightarrow q_4(P_1)$ がこれに相当する．この場合の出現確率 (尤度) は，各遷移での確率値を対数変換しておくことによって，次のように，加算と大小判定のみからなる DP 演算によって高速に求めることができる．すなわち，$i = 1, 2, \cdots, S$ において

$$f'(i,t) = \begin{cases} \log \pi_i & (t = 0) \\ \max_j \{ f'(i, t-1) + \log a_{ji} b_{ji}(y_t) \} & (t = 1, 2, \cdots, T) \end{cases} \tag{2.7}$$

を計算し，対数尤度

$$L = \max_{i, q_i \in F} f'(i, T) \tag{2.8}$$

を求めればよい．対数を用いた計算なので，トレリス法を用いる場合に比べて計算値のダイナミックレンジが小さくてすみ，アンダーフローの問題が解消できるという長所がある．この方法を，ビタービ (Viterbi) アルゴリズムと呼び，計算量が少ないにもかかわらず，トレリス法と音声認識性能はほとんどかわらないことが，実験的に証明されている．さらに，このビタービアルゴリズムを用いると，DP による効率のよい単語音声認識アルゴリズムが容易に適用できるため，広く用いられている．

2.2.3　離散型 HMM のパラメータ推定法

まず，離散型 HMM のパラメータ推定法について説明する．学習用音声として，N 個の観測符号ベクトル系列 $\{ y_1^{T(n)} = y_1, Y_2, \cdots, y_{T(n)} \}_{n=1}^{N}$ が与えられたとき，

$$\prod_{n=1}^{N} P(y_1^{T(N)} | \pi_i, a_{ij}, b_{ij}(k)) \tag{2.9}$$

を最大化するパラメータセット $\{\hat{\pi}_i\}, \hat{a}_{ij}, \hat{b}_{ij}(k)$ は，次のようなバウムウェルチ (Baum-Welch) アルゴリズム (フォワードバックワード (forward- backward) アルゴリズム) によって，推定することができる．

バックワード (backward) 変数 $\beta(i,t)$ を，時刻 t に状態 q_i にあって，以後符号ベクトル y_{t+1}^{T} を出力する確率と定義し，$\gamma(i,j,t)$ を，モデル M が y_1^T を出力する場合において，時刻 t に状態 q_i から q_j へ遷移し符号ベクトル y_t を出力する確率と定義する．このとき，

$$\beta(i,T) = \begin{cases} 1 & q_i \in F \\ 0 & q_i \notin F \end{cases} \tag{2.10}$$

$$\beta(i,t) = \sum_j a_{ij} \cdot b_{ij} \cdot \beta(j, t+1) \quad (t = T, T-1, \cdots, 1; i = 1, 2, \cdots, S) \tag{2.11}$$

$$\gamma(i,j,t) = \frac{\alpha(i, t-1) \cdot a_{ij} \cdot b_{ij}(y_t) \cdot \beta(j,t)}{P(y_1^t | M)} \tag{2.12}$$

の関係が得られる．これらを用いて，パラメータ $\pi_i, a_{ij}, b_{ij}(k)$ を，次の再推定の繰り返しによって求める．

$$\hat{\pi}_i = \frac{\sum_j \gamma(i,j,1)}{\sum_i \sum_j \gamma(i,j,1)} \tag{2.13}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha(i,t-1) \cdot a_{ij} \cdot b_{ij}(y_t) \cdot \beta(j,t)}{\sum_t \alpha(i,t) \cdot \beta(j,t)} = \frac{\sum_t \gamma(i,j,1)}{\sum_t \sum_j \gamma(i,j,t)} \tag{2.14}$$

$$\hat{b}_{ij}(k) = \frac{\sum_{t,y_t=k} \gamma(i,j,t)}{\sum_t \gamma(i,j,t)} \tag{2.15}$$

これは，仮定されたパラメータ (初期確率，遷移確率，出現確率) による HMM を用いて，与えられた学習用サンプルに対する状態間の遷移確率を計算し，この遷移確率を確率的回数と仮定して，学習用サンプルに対する出現確率の最尤推定を行い，初期確率，遷移確率，出現確率の値を更新する手続きになっている．実際には，全ての学習用サンプルに関してこの計算を行ってから 1 回パラメータを更新するというサイクルを，値が収束するまで繰り返して，パラメータの値を決定する．期待値の計算と尤度最大化の計算が交互に繰り返されるので，このような推定アルゴリズムは一般に EM(expectation-maximization) アルゴリズムと呼ばれ，少なくとも局所的最大値に収束することが分かっている．

2.2.4　連続型 HMM のパラメータ推定

連続型 HMM の場合の学習用音声の観測 (出力) 系列 $y = \{y_1, y_2, \cdots, y_T\}$ の y_t は，多次元ベクトルである．出現確率が単一 (多次元) ガウス分布で表される場合の，初期確率 π_i と遷移確率 a_{ij} の推定式は，2.2.3 節で述べた離散型 HMM の場合と同じである．出現確率のガウス分布 $N(\mu_{ij}, \sum_{ij})$ は，次式のように最尤推定できる．

$$\mu_{ij} = \frac{\sum_{t=1}^{T} \gamma(i,j,t) y_i}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{2.16}$$

$$\hat{\Sigma}_{ij} = \frac{\sum_{t=1}^{T} \gamma(i,j,t)(y_t - \mu_{ij})(y_t - \mu_{ij})^t}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{2.17}$$

離散型 HMM の場合と同様に，この推定を EM アルゴリズムが収束するまで繰り返す．

混合ガウス分布の場合の出現確率は，次のように表される (ガウス分布の数を M とする)．

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{2.18}$$

ここで，

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{2.19}$$

$$\int b_{ijm}(y) dy = 1 \tag{2.20}$$

である．このときの初期確率 π_i と遷移確率 a_{ij} の推定式は，離散型 HMM の場合と同じである．混合ガウス分布の出現確率は，単一ガウス分布の場合と同様に次式で表せる．

$$\hat{\lambda}_{ijm} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{2.21}$$

$$\hat{\mu}_{ijm} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m) y_t}{\sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{2.22}$$

$$\hat{\Sigma}_{ijm} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)(y_t - \mu_{ijm})(y_t - \mu_{ijm})^t}{\sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{2.23}$$

ただし,

$$\gamma(i,j,m,t) = \alpha(i,t-1) \cdot a_{ij} \cdot \lambda_{ijm} \cdot b_{ijm}(y_t) \cdot \beta(j,t) \tag{2.24}$$

で, m 番目の分布関数の遷移 $q_i \rightarrow q_j$ の確率 (遷移確率) を表している. これらの推定も, EM アルゴリズムが収束するまで繰り返す.

半連続型 HMM の場合は, 符号帳の中の分布数を M として, 出現確率は次のように表される.

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{2.25}$$

ここで,

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{2.26}$$

である. この混合分布のパラメータの内, 分布の重み λ_{ijm} は, 遷移確率 $(q_i \rightarrow q_j)$ ごとに推定する. 平均値 μ_m および共分散 Σ_m は, 全ての出現分布で共通であるので, これらの推定式は,

$$\hat{\lambda}_{ijm} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{2.27}$$

$$\hat{\mu}_m = \frac{\sum_{all(q_i \rightarrow q_j)} \sum_{t=1}^{T} \gamma(i,j,t,m) y_t}{\sum_{all(q_i \rightarrow q_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{2.28}$$

$$\hat{\Sigma}_m = \frac{\sum_{all(q_i \rightarrow q_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)(y_t - \mu_m)(y_t - \mu_m)^t}{\sum_{all(q_i \rightarrow q_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{2.29}$$

となる. 半連続型 HMM では, 分布の数 (M) が必然的に極めて大きくなるため, 式 (2.25) において, 通常, 確率値の大きい分布を上位数個だけ用いる.

2.2.5　連結学習による音素 HMM の作成

連結学習は, 発声内容のテキストだけが与えられた大量の音声データによって, 音素モデルの学習を行う方法で, 学習にラベルデータを必要としないことから, ラベル無し連結学習とも呼ばれている. テキストに基づいて, 前の音素モデルの最終状態が次の音素の初期状態になるように音素モデルを連結し, 文モデルを作成する. これを用いて, バウムウェルチアルゴリズムにより, HMM パラメータの再学習を行う. 連結学習は, 本研究では再推定したモデルの学習に使用する.

第3章　モーラ情報の利用

3.1　モーラ情報について

モーラとは，仮名文字単位に相当し，音節とはやや異なっている．単語の仮名文字の個数 (モーラの数) をモーラ数，単語の仮名文字の位置 (モーラの位置) をモーラ位置という．本論文では，このモーラ数とモーラ位置を合わせてモーラ情報と表現している．

例を表 3.1，3.2 に示す．表 3.1 は，単語が「赤い」(akai) の場合で，音素記号列に含まれる母音と撥音の数は 3 なので，モーラ数は 3，モーラ位置は,「a」は 1,「ka」は 2,「i」は 3 となる．また，表 3.2 のように音素記号列が「kimari」である場合は，モーラ数は 3，モーラ位置は,「ki」は 1,「ma」は 2,「ri」は 3 となる．

表 3.1: 音声ラベル「akai」のモーラ情報

仮名文字単位	a	ka	i
モーラ位置	1	2	3
モーラ数		3	

表 3.2: 音声ラベル「kimari」のモーラ情報

仮名文字単位	ki	ma	ri
モーラ位置	1	2	3
モーラ数		3	

3.2　モーラ情報とピッチ周波数の関係

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 [6)].

図 3.1 は [6)] 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．なお，時間軸はモーラ位置により正規化されている．図中の◇はピッチ周波数の平均値を示し，縦線の部分はピッチ周波数の分散を示している．ピッチ周波数の解析には xwave+[7)] を使用して

いる．図 3.1 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できることが分かる．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．4,6 モーラ語の場合のピッチ周波数との関係を図 3.2，図 3.3 に示す．

また，固有名詞だけでなく，普通名詞においても図 3.1，図 3.2，図 3.3 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 [8)]．

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．音素ラベルを単語のモーラ数，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．

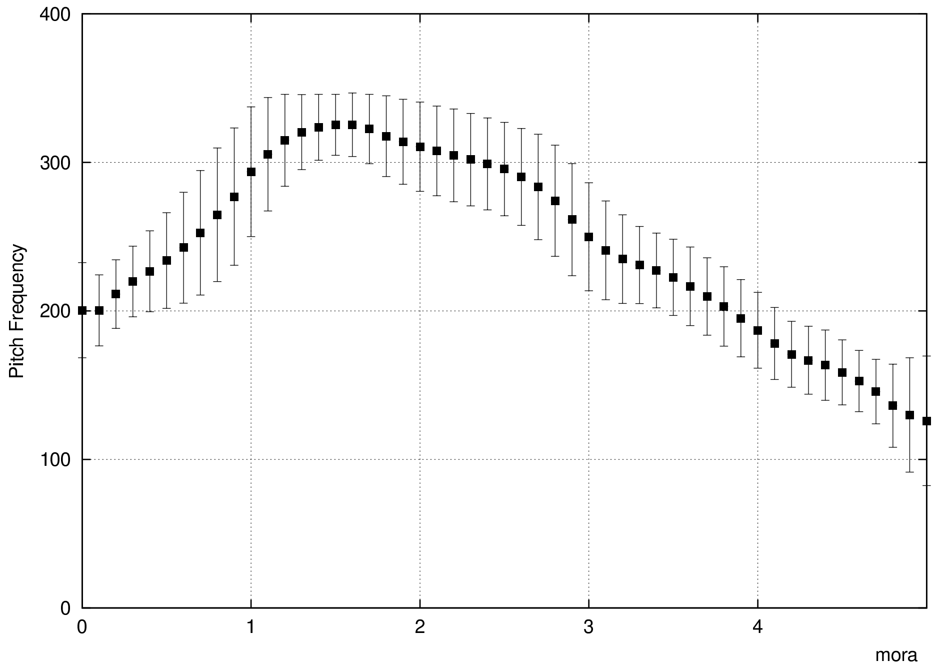

図 3.1: モーラ情報とピッチ周波数の関係 (5 モーラ語)

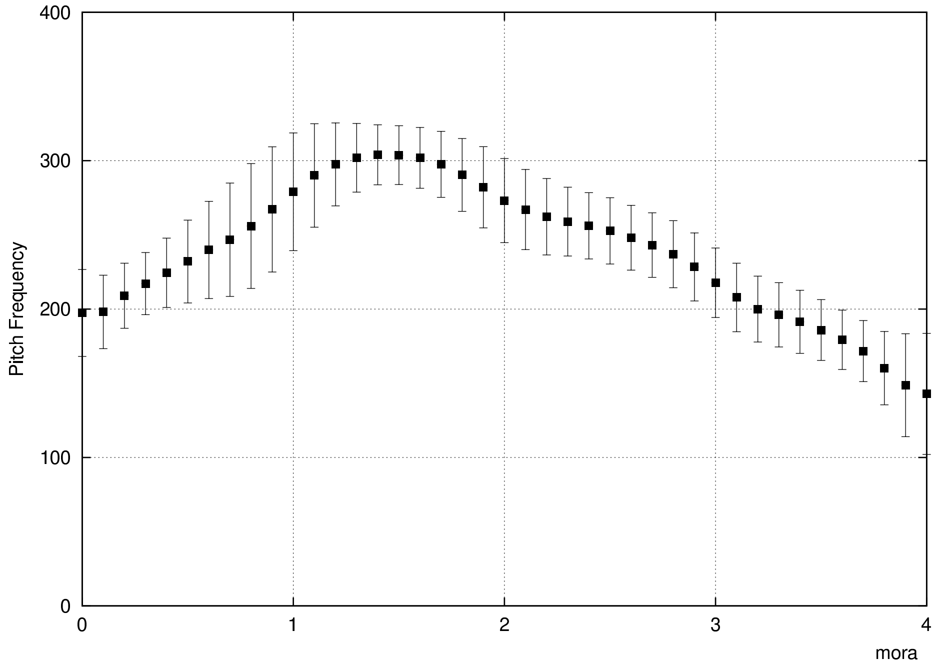

図 3.2: モーラ情報とピッチ周波数の関係 (4 モーラ語)

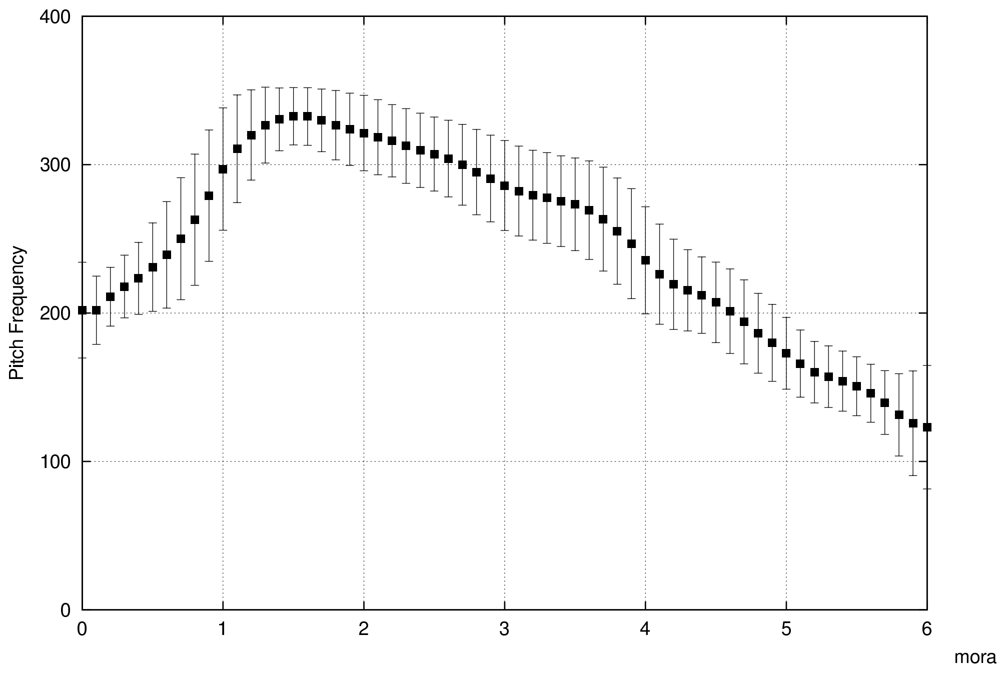

図 3.3: モーラ情報とピッチ周波数の関係 (6 モーラ語)

第4章　FBANKを使用した特徴パラメータ

4.1　従来の特徴パラメータ

音声認識では，通常，音声データに対して特徴パラメータ抽出を行い，スペクトルパラメータに変換したものを扱う．特徴パラメータ抽出を行う方法として，フィルタバンク分析 (filter bank analysis) と線形予測符号化 (linear predictive cording) がある．本研究では，ハードウェアによる実時間分析の実現が容易であることから，フィルタバンク分析を用いて特徴パラメータ抽出を行う．

現在の音声認識では，特徴パラメータとしてケプストラムが主に用いられている．人の聴覚は，音の高さに関して，メル (mel) 尺度と呼ばれる対数に近い非線形の特性を示し，低い周波数では細かく，高い周波数では荒い周波数分解能を持つ．この性質をケプストラムに利用したものを，メル周波数ケプストラム係数 (Mel Frequency Cepstrum Coefficient：MFCC) と呼び，特徴パラメータとして一般的に使用されている．

4.2　FBANKについて

本研究では，ピッチの中にも認識に有効な情報が存在すると仮定し，ピッチの情報を音声認識に直接利用することを試みる．ケプストラムを使用すると，ピッチを分離してしまい，ピッチの情報が利用できないので，パワースペクトラムを使用する．さらに，3.2 節で述べた，単語のモーラ数・モーラ位置が決まればピッチ周波数がほぼ決定できることに着目し，音素ラベルをモーラ数・モーラ位置で分類する．これにより，ピッチを含めた音声の特徴をより正確に表現できると考える．

本研究で特徴パラメータに使用する FBANK について説明する．FBANK は図 4.1 の方法により作成する．音声波形をフーリエ変換して得られたパワースペクトラムの周波数の全域に，メルスケールに沿って等間隔に配置された三角形のフィルタをかける．この三角形の個数がフィルタバンクのチャンネル数 (特徴パラメータにおける次数) を表している．そして，フィルタバンクの出力に log 対数をとったものを FBANK とし，特徴パラメータとして使用する．周波数メル分割の式を式 4.1 に示す．

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{4.1}$$

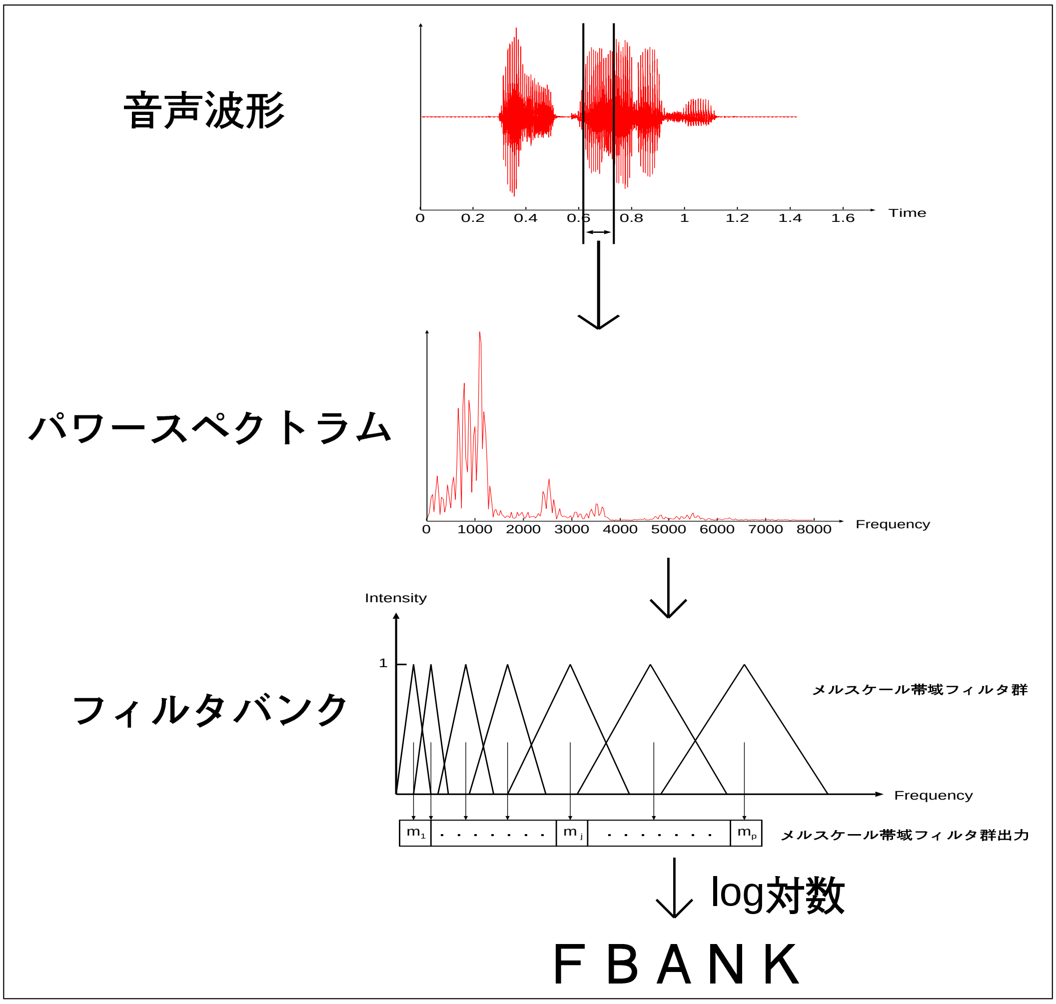

図 4.1: FBANK を使用した特徴パラメータの作成

第5章　評価実験

5.1　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．

5.2　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 5.1 に示す．

表 5.1: 母音と撥音の分類例

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

音声ラベルが atama である場合，単語のモーラ数は 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．

5.3 音素 HMM の作成方法

5.3.1 半連続型 (semi-continuous)HMM の使用

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)]．

5.3.2 音素 HMM の作成手順

本研究では，音素 HMM を図 5.1 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．

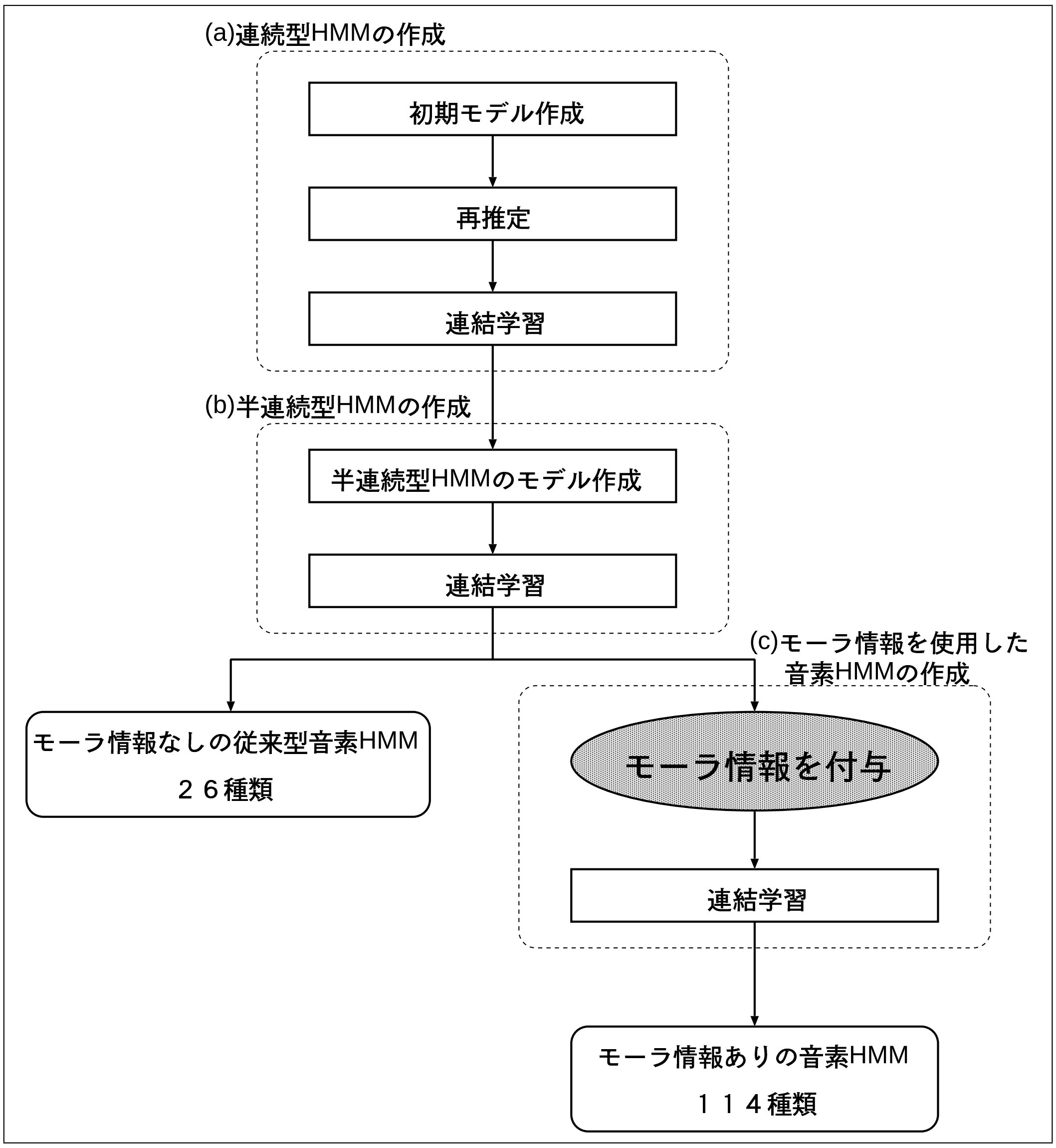

図 5.1: 音素 HMM の作成手順

5.4　実験条件

孤立単語音声認識を行うツールには HTK[10)] を使用し，実験は表 5.2 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，孤立単語音声認識における FBANK パラメータの効果を調査する．

5.5　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音声認識を行い，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 (5.1)，式 (5.2) に示す．

$$\text{認識率}\,(\%) = \frac{\text{正しく認識できた単語数}}{\text{評価単語総数}} \times 100 \tag{5.1}$$

$$\text{改善率}\,(\%) = \frac{\text{誤り数}_{\text{モーラ無し}} - \text{誤り数}_{\text{モーラ有り}}}{\text{誤り数}_{\text{モーラ無し}}} \times 100 \tag{5.2}$$

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 (MFCC) においても，表 5.2 と同様の条件で孤立単語音声認識を行い，FBANK との比較を行う．

表 5.2: 実験条件

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(24 次) +
	Δ FBANK(24 次) +
	対数パワー (1 次), Δ対数パワー (1 次)　(計 50 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

5.6　実験結果

表 5.2 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音声認識の結果を表 5.3，Full を使用した結果を表 5.4 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 94.96%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．

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．

表 5.3: FBANK の実験結果 (Diagonal-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	93.48%(2422/2591)	95.52%(2475/2591)	31.36%
mmy	91.20%(2363/2591)	93.48%(2422/2591)	25.88%
mtk	93.79%(2430/2591)	95.56%(2476/2591)	33.54%
faf	93.59%(2425/2591)	94.87%(2458/2591)	19.88%
ftk	93.67%(2427/2591)	95.52%(2475/2591)	29.27%
fyn	93.86%(2432/2591)	94.83%(2457/2591)	15.72%
平均	93.27%(14499/15546)	94.96%(14763/15546)	25.21%

表 5.4: FBANK の実験結果 (Full-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	96.91%(2511/2591)	97.99%(2539/2591)	35.00%
mmy	95.99%(2487/2591)	97.61%(2529/2591)	40.38%
mtk	95.91%(2485/2591)	97.22%(2519/2591)	32.08%
faf	97.07%(2515/2591)	97.88%(2536/2591)	27.63%
ftk	96.14%(2491/2591)	97.61%(2529/2591)	38.00%
fyn	96.14%(2491/2591)	96.60%(2503/2591)	12.00%
平均	96.36%(14980/15546)	97.48%(15155/15546)	30.92%

第6章　考察

6.1　モーラ情報による認識誤りの改善

6.1.1　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 6.3 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 6.1: 改善された単語例 (話者 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
多い	o o i	→	老い	o i
会場	k a i zh j o o	→	解除	k a i zh j o
苦労	k u r o o	→	黒	k u r o
行為	k o o i	→	故意	k o i
講師	k o o sh i	→	腰	k o sh i
構成	k o o s e i	→	個性	k o s e i
仕入れる	sh i i r e r u	→	知れる	sh i r e r u
招待	sh j o o t a i	→	書体	sh j o t a i
誠意	s e i i	→	性	s e i
地位	ch i i	→	血	ch i
陶器	t o o k i	→	時	t o 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 sh i
誘拐	j u u k a i	→	愉快	j u k a i
幼稚	j o o ch i	→	予知	j o ch i

6.1.2　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 6.4 に示す．改善されなかった単語の多くは，子音の誤認識によるものであった．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．

表 6.2: 改善されなかった単語例 (話者 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
頑丈	g a ng zh j o o	→	感情	k a ng zh j o o
機会	k i k a i	→	期待	k i t a i
行政	g j o o s e i	→	強制	k j o o s e i
憲法	k e ng p o o	→	検討	k e ng t o o
信頼	sh i ng r a i	→	寝台	sh i ng d a i
主体	sh u t a i	→	死体	sh i t a i
パイプ	p a i p u	→	タイプ	t a i p u
プラス	p u r a s u	→	クラス	k u r a s u
ペン	p e ng	→	点	t e ng
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
来年	r a i n e ng	→	体面	t a i m e ng
利息	r i s o k u	→	規則	k i s o k u

6.2 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)，Δ MFCC(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 5.2 と同様にして実験を行った．Diagonal の結果を表 6.3，Full の結果を表 6.4 に示す．また，表 5.3，表 5.4 の FBANK の孤立単語音声認識の結果と表 6.3，表 6.4 の MFCC の孤立単語音声認識の結果を 6 話者の平均でまとめたものを図 6.1 に示す.

図 6.1 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．

表 6.3: MFCC の実験結果 (Diagonal-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	95.48%(2474/2591)	96.57%(2502/2591)	23.93%
mmy	92.90%(2407/2591)	95.18%(2466/2591)	32.07%
mtk	94.02%(2436/2591)	96.49%(2500/2591)	41.29%
faf	93.36%(2419/2591)	95.75%(2481/2591)	36.05%
ftk	94.94%(2460/2591)	96.06%(2489/2591)	22.14%
fyn	93.75%(2429/2591)	95.91%(2485/2591)	34.57%
平均	94.08%(14625/15546)	95.99%(14923/15546)	32.36%

表 6.4: MFCC の実験結果 (Full-covariance)

話者	モーラ無し	モーラ有り	改善率
mau	96.68%(2505/2591)	98.07%(2541/2591)	41.86%
mmy	94.91%(2459/2591)	96.53%(2501/2591)	31.82%
mtk	94.09%(2438/2591)	96.37%(2497/2591)	38.56%
faf	95.95%(2486/2591)	97.68%(2531/2591)	52.38%
ftk	95.83%(2483/2591)	97.18%(2518/2591)	23.15%
fyn	95.02%(2462/2591)	96.49%(2500/2591)	29.46%
平均	95.41%(14833/15546)	97.05%(15088/15546)	35.76%

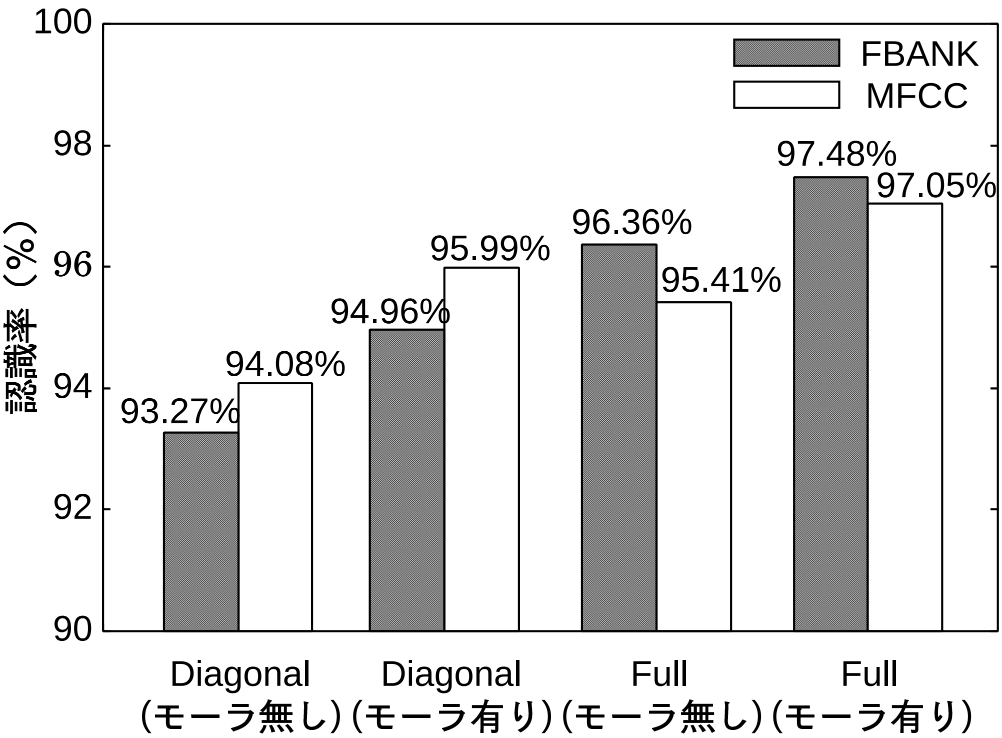

図 6.1: FBANK と MFCC の比較

6.3 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．Triphone モデルを使用した音素ラベルの分類例を表 6.5 に示す．

Triphone モデルのラベルファイルは，焦点となる音素の前に前音素を，後ろに後音素を − と + でつなぐことによって分類する．音素ラベルが atama である場合について分類例を示す．1 番目の音素 a は前音素が無く，後音素が t なので a+t と置き換えられる．また，3 番目の音素 a は前音素が t で後音素が m なので，t−a+m と置き換えられる．

実験は表 5.2 と同じ条件で行った．Triphone モデルの結果を表 6.6 に示し，表 5.3，表 5.4 の結果と比較したものを図 6.2，図 6.3 に示す．また，モーラ情報を使用したモデルと Triphone モデルとの間で，認識できる単語の違いについて調査を行った．モーラ情報を使用したモデルで認識できた単語例を表 6.7 に示し，Triphone モデルで認識できた単語例を表 6.8 に示す．

図 6.2，図 6.3 より，Diagonal では Triphone モデルの認識率が最も高くなったのに対して，Full ではモーラ情報を使用したモデルの認識率が最も高くなった．Diagonal と Full では，Full の方が高い認識率が得られることから，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．しかし，表 6.7，表 6.8 からは，モーラ情報を使用したモデル，Triphone モデル共に脱落誤りや置換誤り，挿入誤りについて大きな違いは見られなかった．

表 6.5: Triphone モデルの分類例

分類前	a	t	a	m	a		
分類後	a+t	a−t+a	t−a+m	a−m+a	m−a		
分類前	a	r	u	b	a	m	u
分類後	a+r	a−r+u	r−u+b	u−b+a	b−a+m	a−m+u	m−u
分類前	a	ng	k	e	e	t	o
分類後	a+ng	a−ng+k	ng−k+e	k−e+e	e−e+t	e−t+o	t−o

表 6.6: Triphone モデルの実験結果

話者	Diagonal	Full
mau	97.45%(2525)	98.53%(2553)
mmy	95.25%(2468)	96.59%(2513)
mtk	96.68%(2505)	97.65%(2530)
faf	96.87%(2510)	96.02%(2488)
ftk	96.64%(2504)	96.41%(2498)
fyn	96.72%(2506)	95.64%(2478)
平均	96.60%(15018)	96.87%(15060)

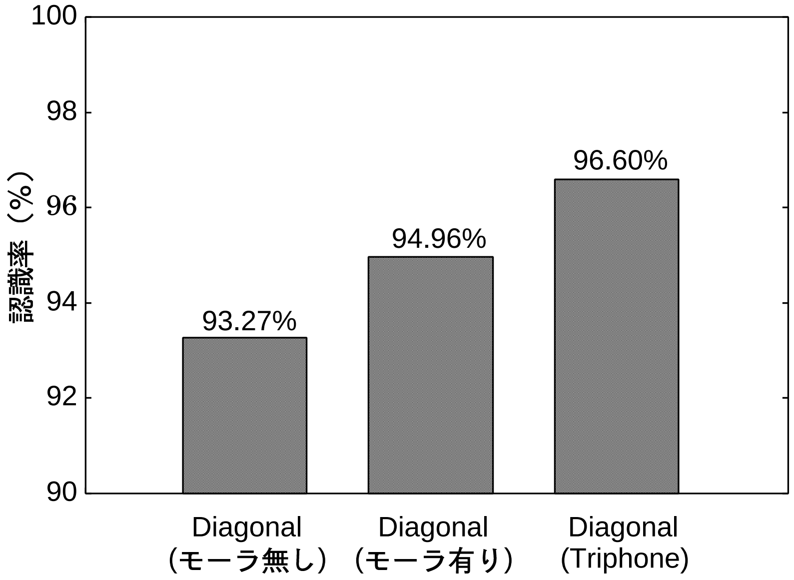

図 6.2: Triphone モデルとの比較 (Diagonal-covariance)

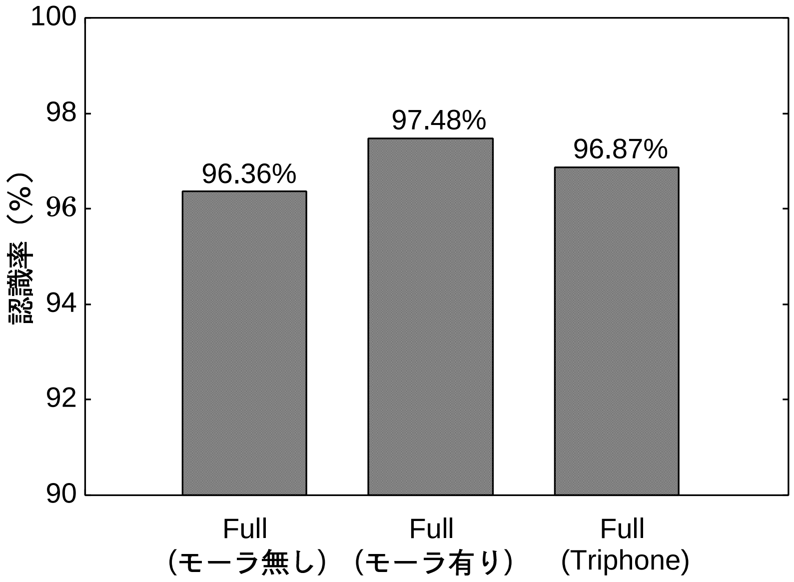

図 6.3: Triphone モデルとの比較 (Full-covariance)

表 6.7: モーラ情報を使用したモデルで認識できた単語例 (話者 mau)

モーラ情報で認識できた単語			Triphone で誤認識した単語	
単語	音素列		単語	音素列
家屋	k a o k u	→	科目	k a m o k u
奇麗	k i r e i	→	霧	k i r i
芸術	g e i zh u ts u	→	現実	g e ng zh i ts u
出演	sh u ts u e ng	→	失恋	sh i ts u r e ng
旋風	s e ng p o o	→	先頭	s e ng t o o
通訳	ts u u j a k u	→	墜落	ts u i r a k u
綱	ts u n a	→	妻	ts u m a
粒	ts u b u	→	蔓	ts u r u
似る	n i r u	→	握る	n i g i r u
部屋	h e j a	→	平野	h e i j a
無用	m u j o o	→	無料	m u r j o o
湾	w a ng	→	番	b a ng

表 6.8: Triphone モデルで認識できた単語例 (話者 mau)

Triphone で認識できた単語			モーラ情報で誤認識した単語	
単語	音素列		単語	音素列
一生	i q sh j o o	→	衣装	i sh j o o
埋まる	u m a r u	→	丸	m a r u
応援	o u e ng	→	公園	k o o e ng
記憶	k i o k u	→	気力	k i r j o k u
高校	k o o k o o	→	国交	k o q k o o
審議	sh i ng g i	→	心理	sh i ng r i
辞表	zh i h j o o	→	辞書	zh i sh j o
適応	t e k i o o	→	提供	t e i k j o o
比喩	h i j u	→	昼	h i r u
不公平	h u k o o h e i	→	コーヒー	k o o h i i
紛失	h u ng sh i ts u	→	本質	h o ng sh i ts u
申し込む	m o o sh i k o m u	→	押し込む	o sh i k o m u

6.4　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 での孤立単語音声認識を行った．実験は表 5.2 と同様の条件で行い，連続型 HMM の作成は図 5.1 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 6.9，Full を使用した場合の結果を表 6.10 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 6.9: 連続型 HMM の実験結果 (Diagonal-covariance)

話者	FBANK	MFCC
mau	92.40%(2394)	95.10%(2464)
mmy	88.50%(2293)	92.47%(2396)
mtk	90.39%(2342)	93.48%(2422)
faf	91.66%(2375)	92.90%(2407)
ftk	90.47%(2344)	93.32%(2418)
fyn	92.09%(2386)	93.28%(2417)
平均	90.92%(14134)	93.43%(14524)

表 6.10: 連続型 HMM の実験結果 (Full-covariance)

話者	FBANK	MFCC
mau	97.07%(2515)	96.22%(2493)
mmy	95.95%(2486)	94.98%(2461)
mtk	95.45%(2473)	95.41%(2472)
faf	96.95%(2512)	96.14%(2491)
ftk	96.22%(2493)	95.52%(2475)
fyn	95.87%(2484)	95.95%(2486)
平均	96.35%(14963)	95.70%(14878)

表 6.9，表 6.10 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 5.3，表 5.4 の半連続型 HMM の結果と比較したものを図 6.4，図 6.5 に示す．図 6.4，図 6.5 より，Diagonal，Full の両方ともにモーラ情報を使用した半連続型 HMM の認識率が最も高くなった．したがって，FBANK にはモーラ情報を使用した半連続型 HMM が最も有効であると考えられる．

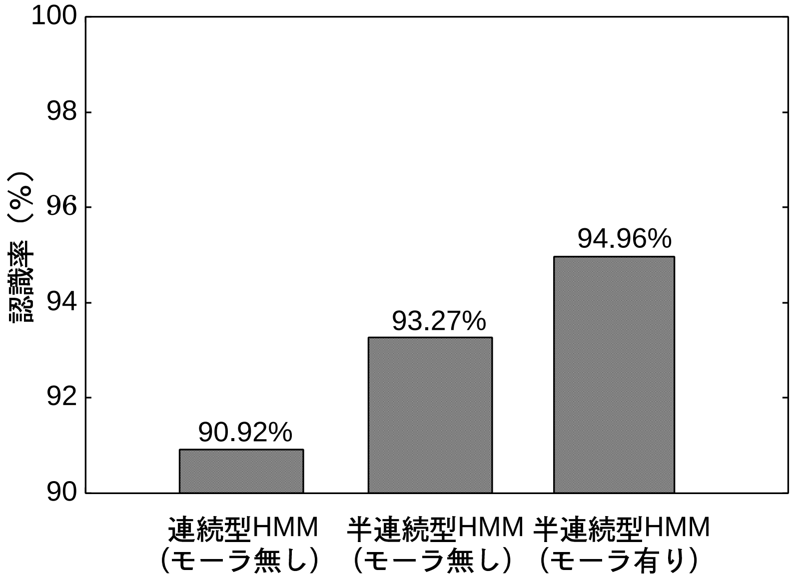

図 6.4: 連続型 HMM と半連続型 HMM の比較 (Diagonal-covariance)

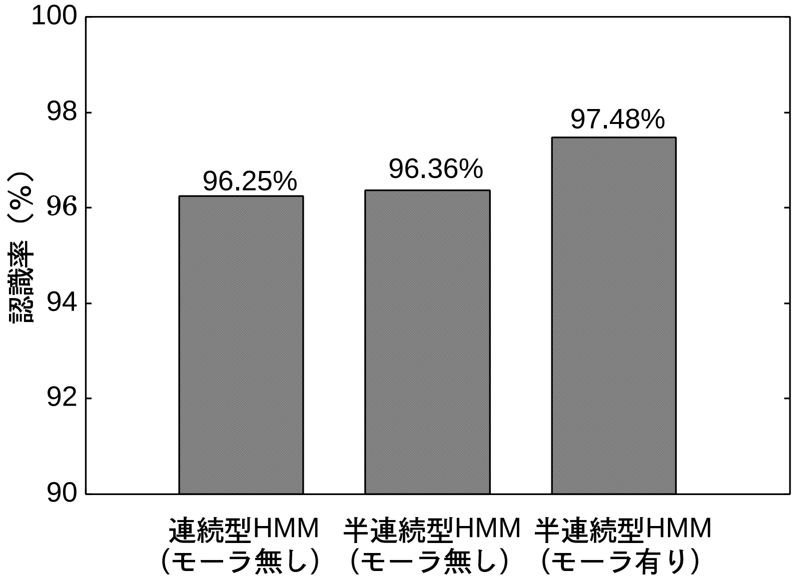

図 6.5: 連続型 HMM と半連続型 HMM の比較 (Full-covariance)

6.5　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．

第7章　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．

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，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．

謝 辞

本研究を進めるにあたり，適切な御指導と御助言を賜わりました鳥取大学工学部知能情報工学科 池原 悟 教授ならびに 村上 仁一 助教授に深く感謝致します．

また，有益な御助言を頂いた 徳久 雅人 助手に感謝致します．

最後に，実験に協力してくださった大学院生ならびに学部生の皆さんに感謝し，皆様の御健康と御発展を心よりお祈り申し上げます．

平成15年2月 谷口 勝則

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] 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] 吉井貞熙：音声情報処理，森北出版 (1998)

[6] 水澤紀子，村上仁一，東田正信：HMM 音素モデルによる単語音声合成，信学技報，SP99-2，(1999-05)

[7] Introducing ESPS/waves+ with EnSigTM Entrotic Research Laboratory,Inc.

[8] 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)

[9] X.D.Huang，Y.Ariki，M.A.Jack，HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION，Edinburgh University Press，1990

[10] HTK Ver2.2 reference manual，1997 Cambridge University

[11] 谷口勝則，村上仁一，池原悟：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，信学技報，SP2002-131，pp.63-68(2002-12)

[12] 谷口勝則，村上仁一，池原悟：FBANK を用いた孤立単語音声認識，日本音響学会 2003 年春季研究発表会 (発表予定)，3-Q-3，pp.157-158(2003-3)

[13] 谷口勝則，村上仁一，池原悟：FBANK を用いた孤立単語音声認識，電子情報通信学会，論文誌投稿中 (2003-2)