概要

従来の単語音声認識においては, 主に音声の音韻的特徴が用いられてきた. しかし, 日本語では, 「箸」, 「橋」のような音韻的には同一だがアクセントの違いによって弁別できる単語が存在する. 過去の研究において, 日本語における同音異義語の音声認識の研究はあまり行われていない. そこで本研究では, 不特定話者における同音異義語の音声認識精度を調査する. 同音異義語の認識はアクセント情報と韻律的情報を含む特徴パラメータとして FBANK を用いることで行う. 実験の結果, アクセントと前後音素環境情報を用いたモデルと特徴パラメータに MFCC を用いることで, 89%の精度が得られた.

目 次

図 目 次

表 目 次

1　はじめに

従来の単語音声認識においては, 主に音声の音韻的特徴が用いられてきた. しかし, 日本語では, 「箸」, 「橋」のような音韻的には同一だがアクセントの違いによって弁別できる単語が存在する. しかし, 日本語における同音異義語の音声認識の研究はあまり行われていない [11].

過去の研究において, 韻律的特徴を用いた研究としては, 高橋ら [10] の研究がある. 高橋らの研究では音声の音韻とアクセントを別々に認識する. 具体的には, 音声からピッチパターンを抽出し単語のアクセント型を 0 型, 1 型, N 型 (0, 1 型以外) の分類で認識する. しかし, 音声から韻律情報のみを分離するのは困難である.

そのため以前の研究において, 特定話者における同音異義語の音声認識を行った. 音韻と韻律を分離せずに同時に認識するために, 単語のアクセント型の情報と各モーラ位置でのアクセントの高低情報を音素に付与しラベル分類を行った. そして, 音声認識に一般的に用いられている特徴パラメータである MFCC は音韻情報しか含んでいないため, 同音異義語の認識精度が低いと予想した. そこで, 韻律的情報を含む特徴パラメータとして FBANK を用いて MFCC と比較し評価した. 実験の結果, 特定話者においてアクセント情報と特徴パラメータとして FBANK を用いることで, 同音異義語の認識精度が高いことを確認した [12].

そこで本研究では, 不特定話者における同音異義語の音声認識精度を調査する. 具体的には, 単語のアクセント型の情報と各モーラ位置でのアクセントの高低情報を音素 HMM に付与して単語音声認識を行い, 評価データ中の同音異義語の認識結果に注目して評価する. 不特定話者認識では, 特定話者認識と比較して認識精度が低下すると考えられる. そこで, 本研究ではアクセント情報と前後音素環境情報を利用したモデルを提案し精度を評価する. また, アクセント情報を音素 HMM に付与すると音素数が増加する. そのため, 本研究において, [12] でも使用した半連続型 HMM[8] と木に基づく状態共有手法 [9] を用いた状態共有型 HMM を利用する. また, 特徴パラメータとして FBANK と MFCC を利用する.

実験の結果, アクセント情報と前後音素環境情報を用いた半連続型 HMM, MFCC, Full の同音異義語認識において 89%の精度が得られた. そして, 半連続型 HMM の認識精度は, 状態共有型 HMM の認識精度より高かった. また, 特徴パラメータとして FBANK を用いた認識精度は, MFCC を用いた認識精度より低かった.

2　音声分析

2.1　音声の生成構造

音声信号は, 人間の調音器官により生成される音響信号である. 音声の生成は, 「音源」により生成された音が「調音器官」により形成される音響的なフィルタを通過することでさまざまに変化し, 口または鼻から「放射」されるというのが基本的な構造になる. 調音フィルタは, ほとんどの場合に伝達関数が

$$H(z) = \frac{b_0}{1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_p z^{-p}} \tag{1}$$

により伝えられる全極形のシステムであると仮定され, 音声認識の分野では広く妥当な仮定として受け入れられている [2].

2.2　音声の特徴抽出

音声認識のための信号分析の目的は, 与えられた信号を生成した調音フィルタの性質を信号より推定することにあり, 信号の周波数領域における表現がその基礎を与える. 音声から連続する数十 ms 程度の時間長の信号区間を切り出し, 切り出された信号が定常確率過程に従うと仮定して, スペクトル解析を行う. すなわち, 与えられた信号 $s(n)$ に長さ N の分析窓を掛けることで以下のように信号系列 $s_w(m;l)$ を取り出す.

$$s_w(m;l) = \sum_{m=0}^{N-1} w(m)s(l+m)(l = 0, T, 2T, \cdots) \tag{2}$$

ここで, 添え字 l は, 信号の切出し位置に対応している. すなわち, l を一定間隔 Tっで増加されることで, 定常とみなされる長さ N の音声信号系列 $s_w(n)(n = 0, \cdots, N-1)$ が間隔 T で得られる. この処理はフレーム化処理と呼ばれ,N をフレーム長,T をフレーム間隔と呼ぶ. また, フレーム化処理を行う窓関数 $w(n)$ としては, ハミング窓やハニング窓がしばしば用いられる.

$$\text{ハミング窓} : w(n) = 0.54 - 0.46\cos(\frac{2n\pi}{N-1})(n = 0, \cdots, N-1) \tag{3}$$

$$\text{ハニング窓} : w(n) = 0.5 - 0.5\cos(\frac{2n\pi}{N-1})(n = 0, \cdots, N-1) \tag{4}$$

フレーム化処理によって得られた音声信号系列の短時間フーリエスペクトルは, 離散フーリエ変換 (DTFT) により以下で与えられる.

$$S(e^j\omega) = \sum_{n=0}^{N-1} s_w(n)e^{-j\omega n} \tag{5}$$

実際の信号処理過程では, 離散フーリエ変換 (DFT) をその高速算法である FFT を用いて実行し, 当該音声区間のスペクトル表現とすること t が一般的である. すなわち

$$S'(k) = S(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} s_w(n)e^{-j\frac{2\pi}{N}kn}(k = 0, \cdots, N-1) \tag{6}$$

なる複素数系列 $S'(k)$ が音声のスペクトル表現として最も一般的に用いられる. 音声信号の音素的特徴は主として調音フィルタの振幅伝達特性に含まれている. したがって, 音声認識においては, 音声信号の振幅スペクトル, あるいはその 2 乗値であるパワースペクトルが注目すべきスペクトル表現である.

2.3 ケプストラム

音声のパワースペクトラムは, 声帯の振動や, 摩擦による乱流などの音源信号に調音フィルタが畳み込まれたものであり, 音素の音響的な特徴は, 調音フィルタの振幅伝達特性によって, 主として担われている. このため, 音声信号から音素の特徴を抽出するためには, 観測された音声のパワースペクトラムから, 音源信号のスペクトルと, 調音フィルタのスペクトルを分離し, 調音フィルタの特性にのみ関連する情報を抽出すれば良い. しかし音声信号から聴音フィルタを分離する問題は, 出力信号 $y(n) = x(n) * h(n)$ から, 入力信号 $x(n)$ とシステムの伝達関数 $h(n)$ を分離する問題である.

ケプストラム (cepstrum) $c(\tau)$ は, 波形の短時間振幅スペクトル $|S(e^{j\omega})|$ の対数の逆フーリエ変換として定義される. 音源信号のスペクトラムを $G(e^{j\omega})$, 調音フィルタの伝達特性を $H(e^{j\omega})$ とすると次の関係が得られる.

$$S(e^{j\omega}) = G(e^{j\omega})H(e^{j\omega}) \tag{7}$$

この対数を取ると,

$$log|S(e^{j\omega})| = log|G(e^{j\omega})|log|H(e^{j\omega})| \tag{8}$$

となる. 次にこれをフーリエ逆変換すると,

$$c(\tau) = \mathcal{F}^{-1}log|S(e^{j\omega})| = \mathcal{F}^{-1}log|G(e^{j\omega})| + \mathcal{F}^{-1}log|H(e^{j\omega})| \tag{9}$$

となり, これがケプストラムである. 離散フーリエ変換 (DFT) で求めると,

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} log|S(k)|e^{j2\pi kn/N}(0 \leq n \leq N-1) \tag{10}$$

となる.

ケプストラムという言葉は, スペクトルを逆変換するという意味から, spectrum をもじって作った造語であり, その変数は frequency をもじってケフレンシー (quefrency) と呼ばれる [2].

従来の音声認識では, 特徴パラメータとしてケプストラムが使われてきた. ケプストラムは低次にフォルマント情報を高次にピッチ情報を含んでいる. しかしピッチ情報は正確なピッチ周波数の抽出が困難であるため, 音声認識ではフォルント情報しか用いられていない.

2.4 FBANK

FBANK は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を使用する. 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. またパワーケプストラムの全域に, 人間の聴覚の特性にあわせて低周波部分は細かく, 高周波部分は大まかに調べるためメルスケールに沿って等間隔に配置された三角関数のフィルタをかける. この三角関数の個数がフィルタバンクのチャンネルのチャンネル数 (特徴パラメータにおける次数) を表している. 周波数メル分割の式は

$$Mel(f) = 2592 \log_{10}(1 + \frac{f}{700}) \tag{11}$$

となる. 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 として使用する.

2.5 MFCC

ケプストラムパラメータには, 多様な計算方法がある. その中にはMFCC(Mel-Frequency Cepstrum Coefficient) がある. MFCC の計算では, スペクトラル分析は周波数軸上に三角窓を配置し, フィルタバンク分析により行う. すなわち, 窓の幅に対応する周波数帯域の信号のパワーを, 単一スペクトルチャンネルの振幅スペクトルの重みづけ和で求める. さらに, 窓はメル周波数軸上に等間隔に配置される.

最終的に, フィルタバンク分析により得られた帯域におけるパワーを離散コサイン変換

することで,MFCC が求められる.

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j - 0.5)) \tag{12}$$

N はフィルタバンクチャンネの数を表し,m_j は対数フィルタバンクの振幅を表す.

2.6 本研究で使用する特徴パラメータ

従来の音声認識の特徴パラメータに用いられている MFCC には韻律情報が含まれていないため, 同音異義語を認識する実験において認識精度が低いと予想される. そのため, 本研究では, 韻律情報が含まれている FBANK を特徴パラメータとして使用する. なお, 特定話者認識において FBANK を用いると MFCC より単語音声認識精度が向上することが知られている [5]. また, FBANK を用いると特定話者において同音異義語の認識精度が高いことが判っている [12].

3 HMMを用いた音声認識

3.1 HMMを用いた音声認識の理論

音声認識は, パターン認識の一分野である. 音声波形から認識に有効な特徴パラメータが抽出された後は, 通常のパターン認識の技術と本質的に変わりはない. 通常のパターン認識との違いは, 音声パターンが時系列パターンであることと言語情報の制約を受けることである. パターン認識には構造的・構文的パターン認識法と統計的・確率的パターン認識法が存在する。最近になって, 音声パターンの時系列パターンに対しての統計的・確率的パターン認識法がHMM(Hidden Markov Model; 隠れマルコフモデル) による手法である [1] .

HMMは, 出力シンボルによって一意に状態遷移先が決まらないという意味での非決定状態オートマトンとして定義される. このモデルでは, 状態と出力シンボルの2課程を考え, 状態が確率的に遷移するときに対応して確率的にシンボルを出力する. このとき観測できるのはシンボル系列だけであることからHidden(隠れ) マルコフモデルとよばれている.

HMMによる音声認識では, 各カテゴリのHMMに対して入力パターンの特徴パラメータ時系列に対する尤度を求め, それを最大にするモデルに対応するカテゴリを認識結果とするのが基本手法である.

HMMは以下の組から定義される.

- 状態の有限集合;$S = \{s_i\}$
- 出力シンボルの集合;$O = \{o_i\}$
- 状態遷移確率の集合;$A = \{a_{ij}\}$; a_{ij} は状態 s_i から 状態 s_j への遷移確率, ここで $\sum_j a_i j = 1$.
- 出力確率の集合;$B = \{b_{ij}(k)\}$; $b_i(k)$ は状態 s_i から においてシンボル k を出力する確率.
- 初期状態確率の集合;$\pi = \{\pi_i\}$;π_i は初期状態が s_i である確率,$\sum_j \pi_j = 1$.
- 最終状態の集合;F

出力シンボルを連続値として表す場合と, 有限個のシンボルの組合せで表現する場合があり, 以下のように分類される [3] .

3.1.1 連続 HMM

出現するスペクトルパターンを連続値として表す分布モデルである. 出現確率を表す方法としては単一ガウス分布や混合ガウス分布が用いられる. パラメータの自由度を減らすために無相関ガウス分布を用いることが多い.

出現確率 $b_ij(o_t)$ が混合ガウス分布に従う場合は,

- M_{ij}...状態 i から状態 j の遷移における混合数
- C_{ijm}...状態 i から状態 j の遷移における混合数のときの重み
- $\mathcal{N}(;\mu,\Sigma)$...平均ベクトルμ, 共分散行列Σをもつ混合ガウス分布

とすると, 以下のように計算される.

$$b_{ij}(o_t) = \sum_{m=1}^{M_{ij}} C_{ijm}\mathcal{N}(o_t;\mu_{ijm},\Sigma_{ijm}) \tag{13}$$

$\mathcal{N}(;\mu,\Sigma)$ は

- n...観測行列の次元数
- $(O-\mu)^t$...$(O-\mu)$ の天地行列
- $\mid \Sigma \mid$ ...Σの固有値
- Σ^{-1}...Σの逆行列

とすると, 以下の式で表現される.

$$\mathcal{N}(O;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n \mid \Sigma \mid}} \exp(-\frac{1}{2}(O-\mu)^t\Sigma^{-1}(O-\mu)) \tag{14}$$

3.1.2 離散 HMM

出現するスペクトルパターンを有限個のシンボルの組合せで表す分布モデルである. スペクトルパターンのベクトル量子化によって, 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出現確率を表す.

3.1.3 半連続型 HMM

半連続型 HMM は離散 HMM の出力確率値に分布を与えた HMM である. 半連続分布は, 離散 HMM の符号張の 1 つずつのベクトルに分布を与えたもので, 連続密度符号張 (continuous density codebook) とも呼ばれている. そこでは, 出力確率を連続密度符号張の分布の混合で表す. 符号張のなかの分布数を M とすると,

$$b_{ij}(x) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(x) \tag{15}$$

と混合正規分布で表す. ただし,

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{16}$$

である. 平均値と共分散はすべての出力確率で同一であり, 遷移 $s_i \to s_j$ での分布の重み λ_{ijm} のみが変わる [4].

3.2 HMM の例

音声認識に用いられる HMM は,left-to-right モデルと呼ばれるものである.left-to-right モデルの例を図 1 に示す.

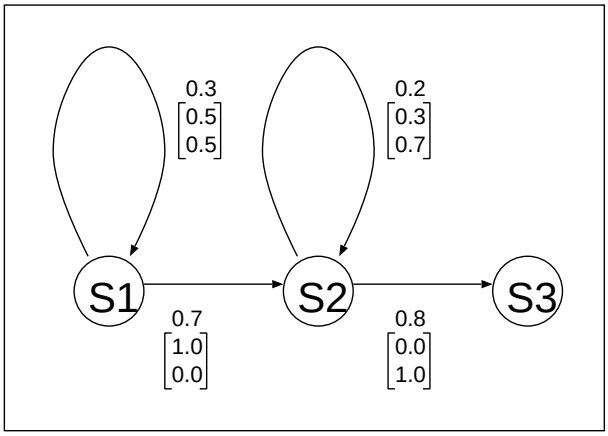

図 1: left-to-right モデルの例

例の HMM は 3 状態で構成され, 出力は有限個のシンボル a と b の 2 種類である. 最終状態を s_3 とし, 初期状態確率の集合 π を以下とする.

$$\pi = \begin{pmatrix} 1.0 & 0 & 0 \end{pmatrix} \tag{17}$$

状態遷移確率の集合 A は以下であり, 図では [] 上部の数字で示される.

$$A = \begin{pmatrix} 0.3 & 0.7 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \tag{18}$$

シンボル a の出力確率の集合 B_a は以下であり, 図では [] 内の上段の数字で示される.

$$B_a = \begin{pmatrix} 0.5 & 1.0 & 0.0 \\ 0.0 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \tag{19}$$

シンボル b の出力確率の集合 B_b は以下であり, 図では [] 内の下段の数字で示される.

$$B_b = \begin{pmatrix} 0.5 & 0.0 & 0.0 \\ 0.0 & 0.7 & 1.0 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \tag{20}$$

状態 s_1 を例にとれば, 状態 s_1 から s_2 の遷移は 0.7 の確率で行われ, 遷移の際に a を出力する確率は 1.0 であり,b を出力する確率は 0.0 である.

例の HMM の出力シンボルが”aab” である場合, 可能な状態遷移系列は $s_1s_1s_2s_3$ と $s_1s_2s_2s_3$ の 2 つで, それぞれの確率は以下のようにして求めることができる.

$$0.3 * 0.5 * 0.7 * 1.0 * 0.8 * 1.0 = 0.084 \tag{21}$$

$$0.7 * 1.0 * 0.2 * 0.3 * 0.8 * 1.0 = 0.0336 \tag{22}$$

よって, この HMM が”aab” を出力する確率は以下のようになる.

$$0.084 + 0.0336 = 0.1176 \tag{23}$$

3.3 認識アルゴリズム

一般に $P(y|M)$ の値は, 以下の trellis アルゴリズムで求められる. 符号ベクトル y_t を出力して状態 s_i にある確率を $\alpha(i,t)$ とする.

$$\alpha(i,t) = \begin{cases} \log \pi_i & (t=0) \\ \sum_j \alpha(j,t-1) a_{ji} b_{ji}(y_t) & (t=1,2,\cdots,T) \end{cases} \tag{24}$$

これを計算して, 最後に

$$P(y|M) = \sum_{i,s_i \in F} \alpha(i,T) \tag{25}$$

を求めれば良い.

$P(y|M)$ を厳密に求めないで, 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 Viterbi アルゴリズムと呼ばれる方法もある. 尤度は, 各遷移での確率値を対数変換しておくことで高速に求めることができる. このアルゴリズムを以下に示す.$i = 1, 2, \cdots, S$ において

$$f'(i,t) = \begin{cases} \log \pi_i & (t = 0) \\ \max_j \{f'(i, t-1) + \log a_{ji} b_{ji}(y_t)\} & (t = 1, 2, \cdots, T) \end{cases} \tag{26}$$

を計算し, 対数尤度

$$L = \max_{i,s_i \in F} f'(i,t) \tag{27}$$

を求める.

この Viterbi アルゴリズムによる利点は trellis 法に比べて以下のようなものである.

- 計算値のダイナミックレンジが小さく, アンダーフロー問題を解消できる.
- 計算量が少ない.
- 音声認識性能がほとんど変わらない.
- DP による効率のよい連続単語音声認識アルゴリズムに用意に適用できる.

このため Viterbi アルゴリズムは広く用いられている.

3.4 離散 HMM のパラメータ推定

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,

$$\prod_{n=1}^{N} P(y_1^T(n)|\pi_i, a_{ij}, b_{ij}(k)) \tag{28}$$

を最大化するパラメータセット $\{\hat{\pi}_i, a_{ij}, b_{ij}(k)\}$ は,Baum-Welch アルゴリズムによって, 次のように推定できる.

まず以下のような変数 $\beta(i,t)$,$\gamma(i,j,t)$ を定義する.

$\beta(i,t)$: 時刻 t に状態 s_i にあって, 以後符号ベクトル y_{t+1}^T を出力する確率

$\gamma(i,j,t)$: 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 s_i から状態 s_j へ遷移し符号ベクトル y_t を出力する確率

このとき, 以下の関係が得られる.

$$\beta(i,T) = \begin{cases} 1 & s_i \in F \\ 0 & s_i \notin F \end{cases} \tag{29}$$

$$\beta(i,t) = \sum_j a_{ij} b_{ij}(y_t) \beta(j,t+1) \quad (t = T, T-1, \cdots, 1; i = 1, 2, \cdots, S) \tag{30}$$

$$\gamma(i,j,t) = \frac{\alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{P(y_1^t|M)} \tag{31}$$

以上を用いて, パラメータ pi_i, a_ij, $b_ij(k)$ を, 以下の再推定によって求める.

$$\hat{\pi_{ij}} = \frac{\sum_j \gamma(i,j,1)}{\sum_i \sum_j \gamma(i,j,1)} \tag{32}$$

$$\hat{a_{ij}} = \frac{\sum_{t=1}^T \alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{\sum_t \alpha(i,t) \beta(j,t)} = \frac{\sum_t \gamma(i,j,1)}{\sum_t \sum_j \gamma(i,j,t)} \tag{33}$$

$$\hat{b_{ij}} = \frac{\sum_{t,y_t=k} \gamma(i,j,t)}{\sum_t \gamma(i,j,t)} \tag{34}$$

実際は, すべての学習サンプルに対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す.

3.5 連続 HMM のパラメータ推定法

連続 HMM のパラメータ推定においては、初期確率 π_i と遷移確率 a_ij の推定式は離散 HMM の場合と同じである。

3.5.1 出現確率が単一 (多次元) ガウス分布で表される場合

出現確率のガウス分布 $N(\mu_{ij}, \Sigma_{ij})$ は次式のように最尤推定できる.

$$\hat{\mu_{ij}} = \frac{\sum_{t=1}^T \gamma(i,j,t) y_t}{\sum_{t=1}^T \gamma(i,j,t)} \tag{35}$$

$$\hat{\Sigma_{ij}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t)(y_t - \mu_{ij})(y_t - \mu_{ij})^t}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{36}$$

離散 HMM の場合と同様に, この推定を値が収束するまで繰り返す.

3.5.2 出現確率が混合ガウス分布で表される場合

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

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{37}$$

ここで

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{38}$$

$$\int b_{ijm}(y)dy = 1 \tag{39}$$

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.

$$\hat{\lambda_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{40}$$

$$\hat{\mu_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m) y_t}{\sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{41}$$

$$\hat{\Sigma_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)(y_t - \mu_{ijm})(y_t - \mu_{ijm})^t}{\sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{42}$$

ただし,

$$\gamma(i,j,m,t) = \alpha(i,t-1) a_i j \lambda_{ijm} b_{ijm}(y_t) \beta(j,t) \tag{43}$$

で,m 番目の分布関数の遷移 $q_i \to q_j$ の確率 (遷移回数) を表している. これらの推定も値が収束するまで繰り返す.

3.5.3 半連続 HMM の場合

符号張の中の分布数を M として, 出現確率は次のようになる.

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{44}$$

ここで

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{45}$$

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, 遷移状態 $(s_i \to s_j)$ ごとに推定する. 平均値 μ_m および 共分散 Σ_m は, すべての出現分布で共通化してあるので, これらの推定式は,

$$\hat{\lambda_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{46}$$

$$\hat{\mu_{ijm}} = \frac{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m) y_t}{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{47}$$

$$\hat{\Sigma_{ijm}} = \frac{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)(y_t - \mu_{ijm})(y_t - \mu_{ijm})^t}{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{48}$$

となる.

3.6 連結学習

音声認識においては, 通常, 音響モデルとして音素のようなサブワードを単位とするモデルが用いられる. サブワードモデルを学習するためには, 大量の音声データが必要とされる. 音声データ中のサブワードの境界を人手でラベル付けすることはできるが, 人手で行う方法では得られるデータの量はとても限られている. このため学習において連結学習という方法が用いられる. 連結学習ではラベル付けされていない大規模なデータベースを扱うことができる. しかし, 各音声データの発話のシンボルが記述されたテキストが必要とされる. まず, 各サブワードモデルを音声データの発話のシンボルが記述されたテキストを基に連結する. このとき, 前のモデルの最終状態が次のモデルの初期状態になる. 次に, Baum-Welch アルゴリズムによって, 音 声データから連結されたモデルのパラメータの推定を行う. 連結学習では, 初期モデルが重要であり, 通常は, ラベル付けされた音声データを用いて初期モデルを作成する.

連結学習の例を図 2 に示す. 音声データの音素表記 “pau a i pau” を元にして各音素 HMM を連結し, 連結した HMM のパラメータを音声データから推定する.

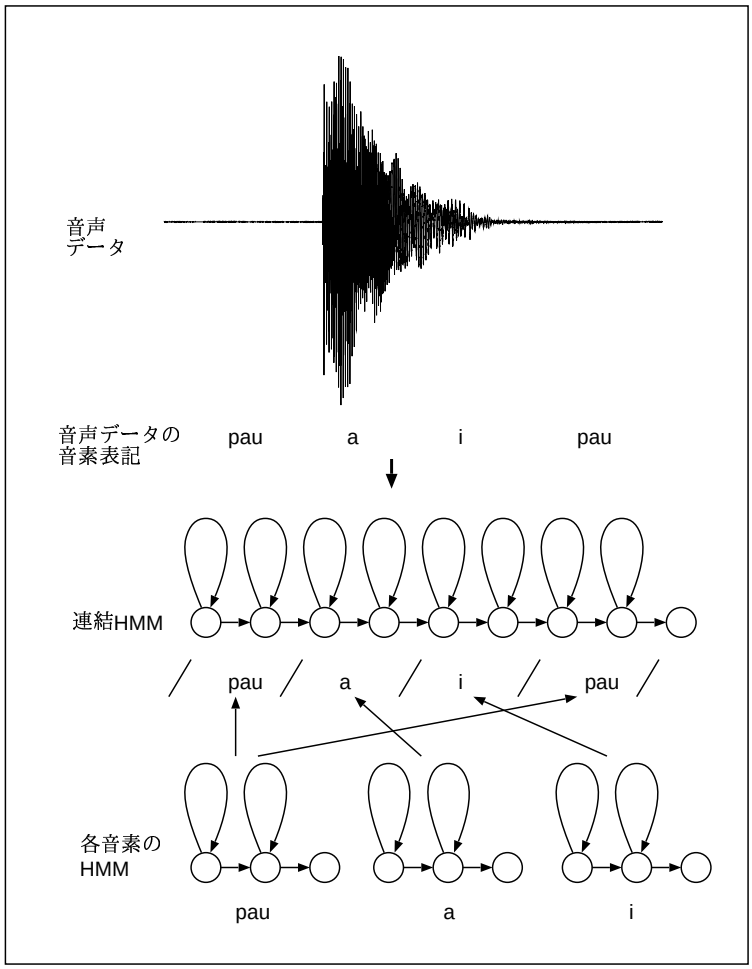

図 2: 連結学習の例

3.7 木に基づく状態共有

連続型 HMM は, ガウス分布数が固定された半連続型 HMM と違い, 状態毎にガウス分布を持つ. そして, 連続型 HMM での triphone はパラメータ数が多く, 信頼性のあるパラメータの推定が困難である. そのため, HMM の似ている状態を共有し, 1 状態あたりの学習データを増やす手法が存在する. 本研究において, 状態共有には音の決定木に基づく状態共有手法 [9] を用いる.

状態共有 HMM システムの構築手順を図 3 に示し, 手順の説明を以下に示す.

1. 単一ガウス分布のモノフォン HMM を作成する.
2. triphoneHMM をモノフォン HMM からコピーし学習することで作成する.
3. 状態のクラスタリングを行い, クラスタ化された状態集合を共有する.
4. HMM の出力確率を混合ガウス分布数にし, パラメータを学習する.

木に基づく状態共有において, 状態のクラスタリングは音の決定木によって行う. 音の決定木の例を図 4 に示す. 音の決定木はバイナリツリーであり, ノード毎に質問が付随する. 図 4 において, ルートノードの質問 “L-Nasal?” の意味は “文脈の左の音は鼻音であるか?” である. 全ての質問の形式は “左/右の音は集合 X に属するか?” となる.

音の決定木が一旦構築されると, 全ての状態はルートノードから決定木を下り, 末端のノードに集められた状態集合が共有される. 木に基づく状態共有は, 未知の triphone でも作成された音の決定木を用いることで状態を合成し作成することができる.

音の決定木の構築手順を以下に示す.

1. ルートノードに全ての状態をおく.
2. log 尤度が最大になるように親ノードの状態を分割する質問を見付け, 状態を 2 つに分ける.
3. 全ての状態を共有したときの log 尤度と, 状態を分割したときの log 尤度のを比べる. log 尤度の増加が閾値を下回れば決定木の構築を終わり, 下回らなければ状態の分割を繰り返す.

本研究では, アクセント情報を用いたモデルにおいて木に基づく状態共有手法を用いる. アクセント情報に関する質問の形式は, “音に付属するアクセント情報が集合 Y に属

表 1: 英語音素と日本語音素の対応表

英語音素	日本語音素	英語音素	日本語音素	英語音素	日本語音素	英語音素	日本語音素
aa	a	ae	a	ah	a	ao	a
aw	a	ax	a	ay	a	b	b
ch	ch	d	d	dd	d	dh	d
dx	d	eh	e	en	e	er	e
ey	e	f	f	g	g	hh	h
ih	i	iy	i	jh	j	k	k
kd	k	l	l	m	m	n	n
ng	N	ow	o	oy	o	p	p
pd	p	r	r	s	s	sh	sh
sil	pau	t	t	td	t	th	t
ts	ts	uh	u	uw	u	v	v
w	w	y	y	z	z		

するか?” とし, 集合 Y が全てのパターンを網羅するように作成する. また, モーラ情報を用いたモデルにおいて, 集合 Y が全てのモーラ情報パターンを網羅するように作成する. そして, 木に基づく状態共有の triphone モデルの質問は, HTK に付属する英語音素のための質問 (HTK-samples-3.3 の samples/RMHTK/lib/quests.hed) を対応する日本語音素に変換することで作成する. 英語音素と日本語音素の対応を表 1 に示す. また, 作成した質問の例を 2 に示す.

なお, アクセント triphone モデルの質問はアクセントモデルと triphone モデルの質問を合わせて用いる. そして, モーラ triphone モデルの質問はモーラモデルと triphone モデルを合わせて用いる. 前後音素のアクセント情報は考慮しない.

表 2: 質問の例

モーラモデル	“モーラ位置は 1 であるか?” “モーラ位置は 2,4 または 7 であるか?” “モーラ位置は 1,2,3,4,6 または 7 であるか?” “単語のモーラ数は 4 であるか?” “単語のモーラ数は 5 または 6 であるか?” “単語のモーラ数は 3,5,6 または 7 であるか?”
アクセントモデ	“モーラ位置は 3 であるか?” “モーラ位置は 1,2 または 5 であるか?” “単語のモーラ数は 4,5 または 6 であるか?” “単語のモーラ数は 1,2,3,4,5 または 6 であるか?” “単語のアクセント型は 4 型で, かつアクセントは高いか?” “単語のアクセント型は 3 または 7 型で, かつアクセントは低いか?” “単語のアクセント型は 1,3,5 または 7 型で, かつアクセントは高いか?”
triphone モデル	“左音素環境は音素 g または gy であるか?” “右音素環境は音素 a, e, m, o または r であるか?” “左音素環境は音素 by,i または gy であるか?”

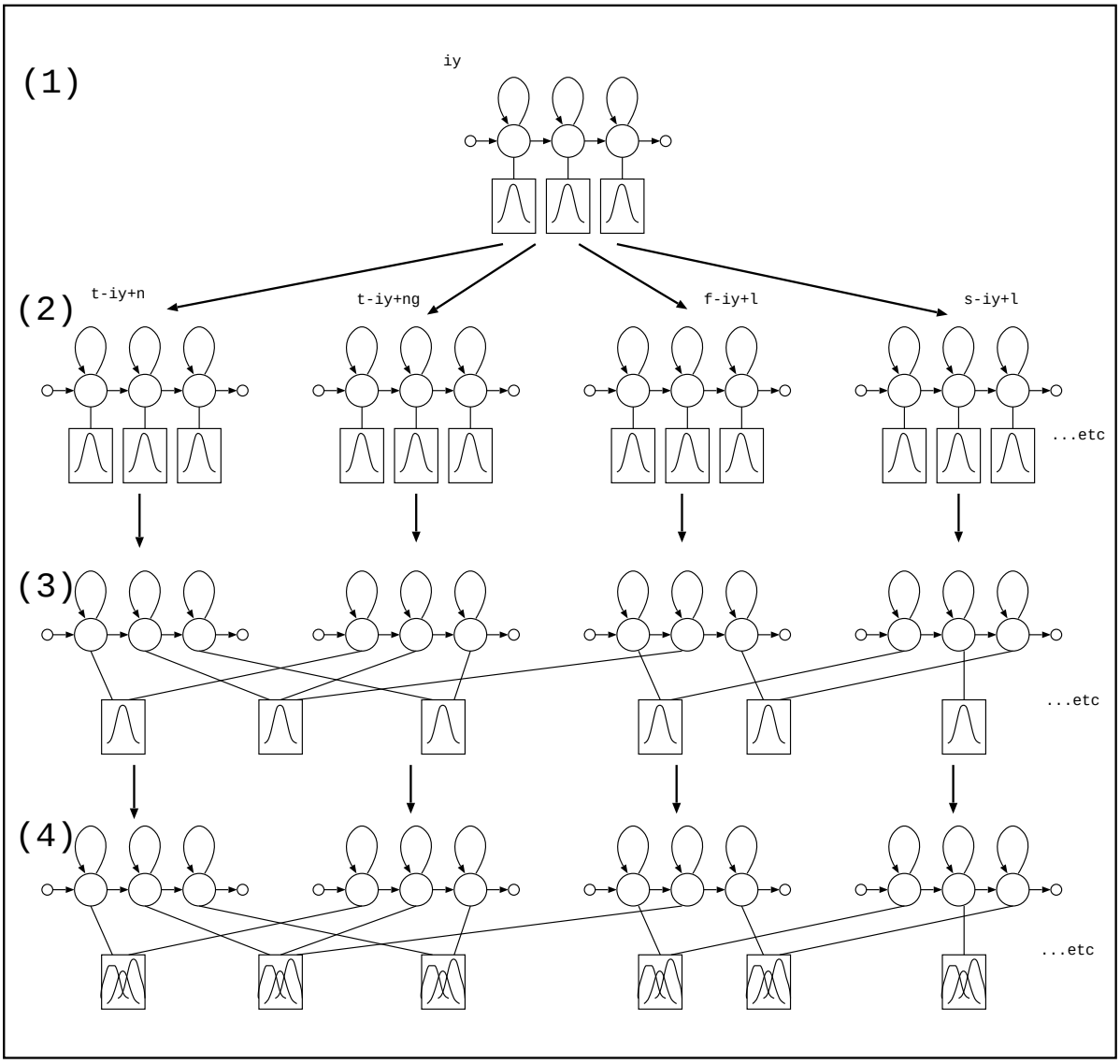

図 3: 状態共有 HMM システムの構築手順

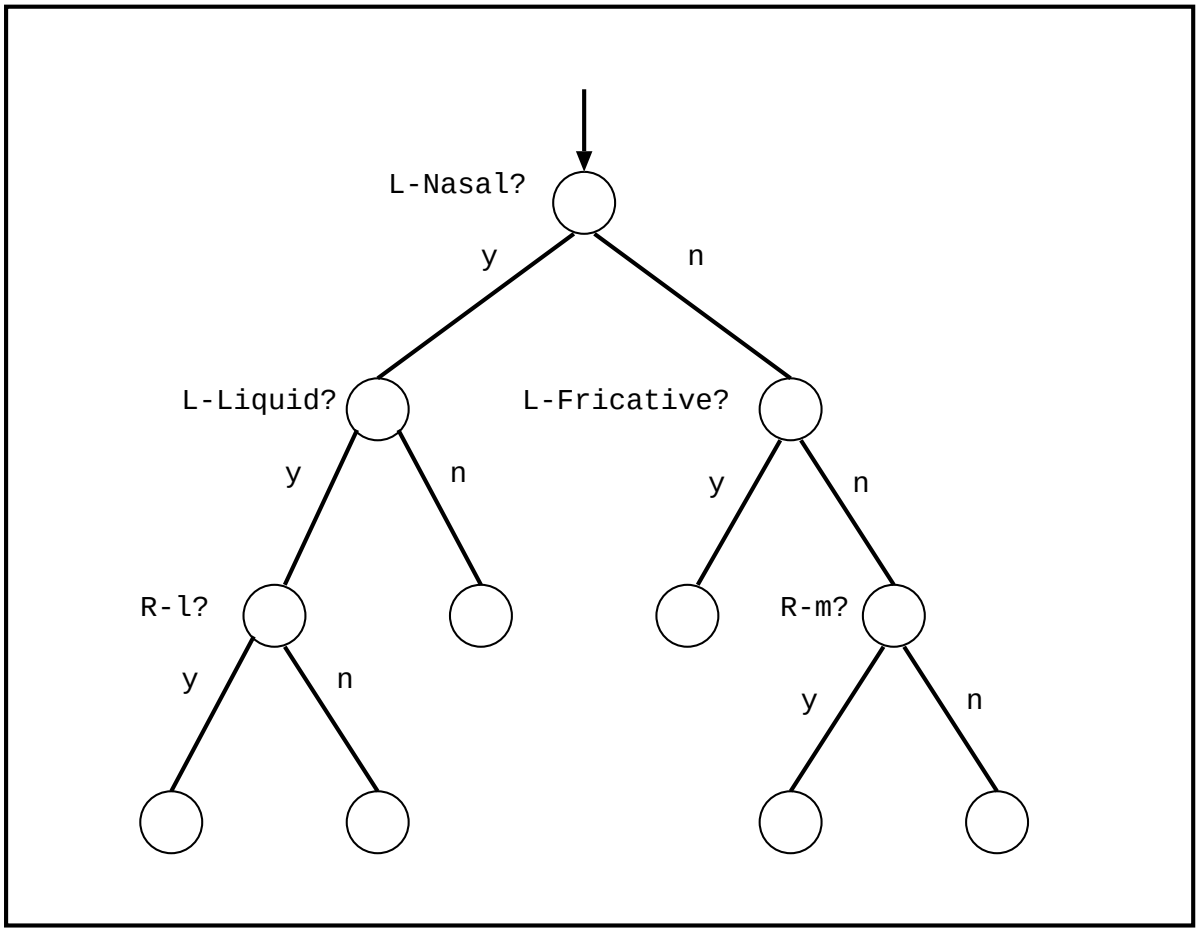

図 4: 音の決定木の例

4　アクセントとモーラ情報

4.1　アクセント

アクセントは他の単語との区別を明確にするのに用いられ, 英語においては強弱で, 日本語においては高さで表現される. 日本語の単語のアクセントは, 日本語での仮名文字単位に相当するモーラごとに高低の 2 レベルが与えられる. そして、アクセントのあるモーラの直後にレベルが高から低に移る. これをアクセント核と呼ぶ.k モーラ目に核が存在するアクセント型を k 型と呼び, 核のないものを 0 型と呼ぶ. アクセント型の例を図 5 に示す.

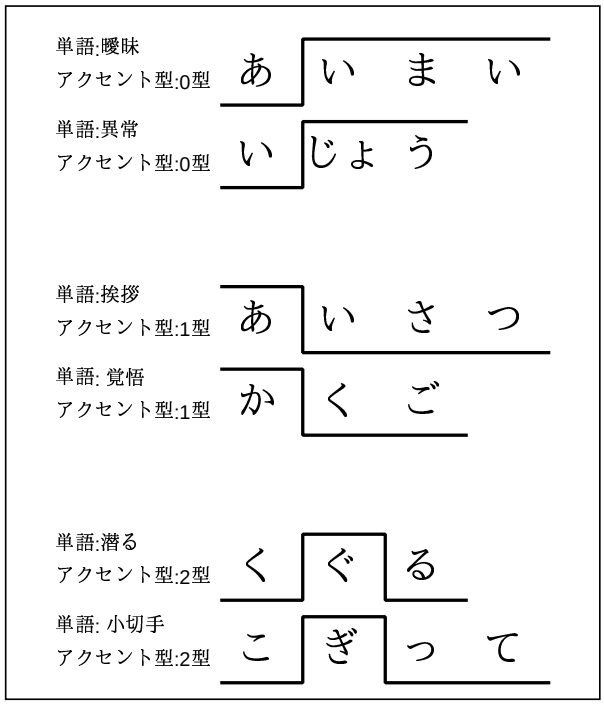

図 5: アクセント型の例

4.2 モーラ情報

モーラとは, 日本語での仮名文字単位に相当し, 和歌においての 5,7,5 の数え方をするときの単位である. 伸ばす音の長母音「ー」, 詰まる音の促音「ッ」, 跳ねる音の撥音「ン」なども 1 モーラに当たる. モーラ数は単語のモーラの総数を示し, モーラ位置は単語でのモーラの位置を示す. 本研究では, モーラ数とモーラ位置をあわせたものをモーラ情報と定義する. 単語におけるモーラ情報の例を表 3,4,5 に示す.

表 3: 単語「参加」におけるモーラ情報

単語	さ	ん	か
音素表記	sa	N	ka
モーラ数		3	
モーラ位置	1	2	3

表 4: 単語「国会」におけるモーラ情報

単語	こ	っ	か	い
音素表記	ko	q	ka	i
モーラ数		4		
モーラ位置	1	2	3	4

表 5: 単語「ジュース」におけるモーラ情報

単語	ジュ	ー	ス
音素表記	ju	u	su
モーラ数		3	
モーラ位置	1	2	3

5　評価実験

本研究では音素 HMM に単語のアクセント型と各モーラ位置のアクセントの高低の情報加えたモデル (以下, アクセントモデル)[12] を用いる. また, 本研究ではアクセントモデルに前後の音素環境情報を加えたモデル (以下, アクセント triphone モデル) を提案し, 評価する. なお, 研究において単語のアクセントは NHK 日本語発音アクセント辞典 [6] を利用する. また, 通常の音素ラベルを用いて学習した音素 HMM を基本モデル, 通常の音素ラベルにおいて前後音素環境を考慮したモデルを triphone モデルとする.

5.1　アクセントを用いた音素ラベルの分類

本研究では音素 HMM に単語のアクセント型と各モーラ位置のアクセントの高低の情報加えたモデル (以下, アクセントモデル)[12] を用いる. また, 本研究ではアクセントモデルに前後の音素環境情報を加えたモデル (以下, アクセント triphone モデル) を提案し, 評価する. なお, 研究において単語のアクセントは NHK 日本語発音アクセント辞典 [6] を利用する. また, 通常の音素ラベルを用いて学習した音素 HMM を基本モデル, 通常の音素ラベルにおいて前後音素環境を考慮したモデルを triphone モデル, モーラ情報を考慮したモーラモデル, モーラ情報と前後音素環境情報を考慮したモーラ triphone モデルとする.

アクセントモデルとアクセント triphone モデルと triphone モデルのラベルの分類例を表 6 に示す. なお, 表中のラベル表記で, + の後の音素は後音素環境を, − の前の音素は前音素環境を表現する. そして, アクセントを用いるモデルは母音, 撥音, 促音音素の後ろ 7 桁の数字でアクセントとモーラ情報を表現する. 7 桁の数字の意味を図 6 に示す.

なお, モーラ情報を用いたモデルも, 母音, 撥音, 促音音素の後ろ 4 桁の数字でモーラ情報を表現する. なお, 4 桁の数字の意味はアクセントを用いたモデルの 7 桁の数字のうちの前 4 桁と同一である.

表 6: 単語:秋 (¯a|k i) の音素ラベルの分類例

基本 モデル	a	k	i
アクセント モデル	a0201011	k	i0202010
triphone モデル	a+k	a−k+i	k−i
アクセント triphone モデル	a0201011 +k	a0201011− k+i0202010	k− i0202010
モーラ モデル	a0201	k	i0202
モーラ triphone モデル	a0201+k	a0201−k+i0202	k−i0202

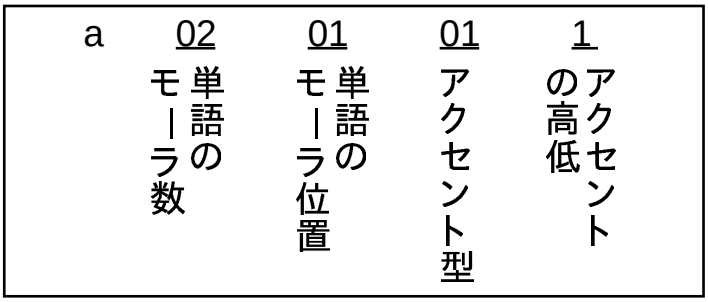

図 6: ラベル表記

5.2 音素 HMM の作成

HMM は初期モデルが重要であるため, アクセントモデルと triphone モデルの初期モデルは基本モデルから作成する. そして, アクセント triphone モデルの初期モデルは triphone モデルから作成する.

5.2.1 半連続型 HMM

半連続型 HMM の実験手順を図 7 に示す. 初めに基本モデルを作成する (図中 a). 次に, 基本モデル以外の各モデルを作成する. アクセントモデルの作成手順を以下に示す.

- 作成された基本モデルの HMM を複製してアクセントモデルの初期モデルとする (図中 b).
- 連結学習を行いアクセントモデルの HMM を作成する (図中 c).

アクセント triphone モデルの作成手順を以下に示す.

- 作成された基本モデルの HMM を複製して triphone モデル の初期モデルとする (図中 d).
- 連結学習を行い triphone モデルの HMM を作成する (図中 e).
- 作成された triphone モデルの HMM を複製してアクセント triphone モデルの初期モデルとする (図中 f).
- 連結学習を行いアクセント triphone モデルの HMM を作成する (図中 g).

5.2.2 状態共有型 HMM

木に基づく状態共有型 HMM の実験手順を図 8 に示す. 初めに, 基本モデルを作成する (図中 h). 次に, 基本モデル以外のモデルを作成する. アクセントモデルの作成手順を以下に示す.

- 作成された基本モデルの HMM を複製してアクセントモデルの初期モデルとする (図中 k).

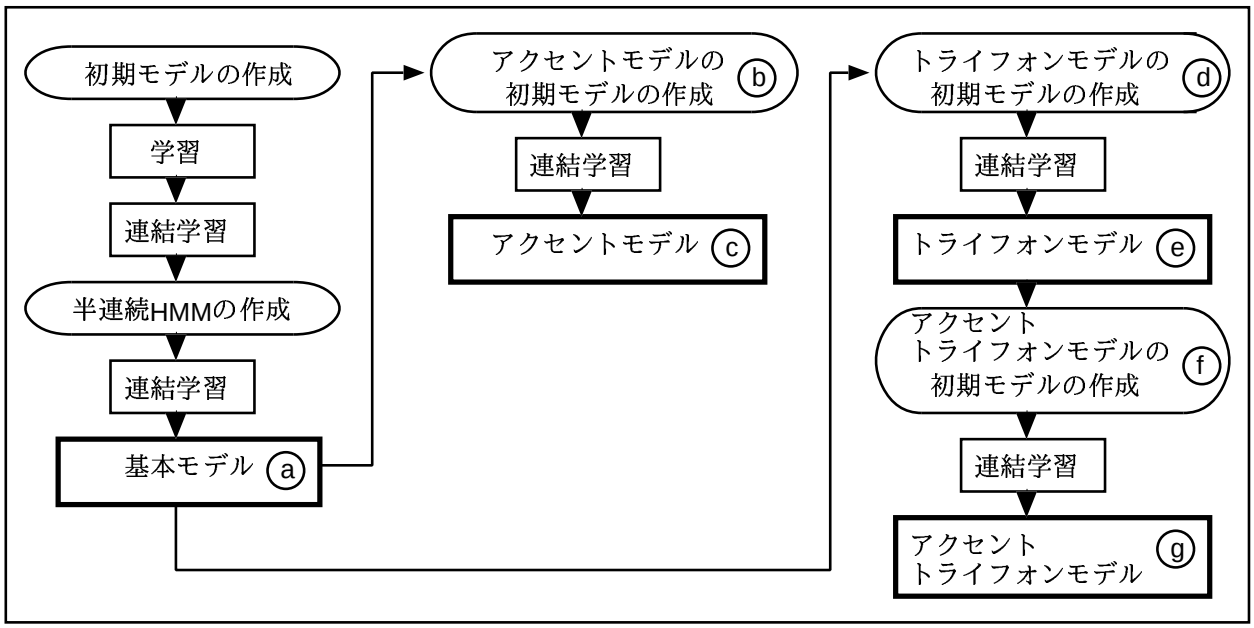

図 7: 半連続型音素 HMM の作成手順

- 状態共有を行った後に混合分布数を増加し, アクセントモデルの HMM を作成する (図中 l).

アクセント triphone モデルの作成手順を以下に示す.

- 作成された基本モデルの HMM を複製して triphone モデルの初期モデルとする (図中 m).
- 学習を行い triphone モデルの HMM を作成する (図中 n).
- 作成された triphone モデルの HMM を複製してアクセント triphone モデルの初期モデルとする (図中 o).
- 状態共有を行った後に混合分布数を増加し, アクセント triphone モデルの HMM を作成する (図中 p).

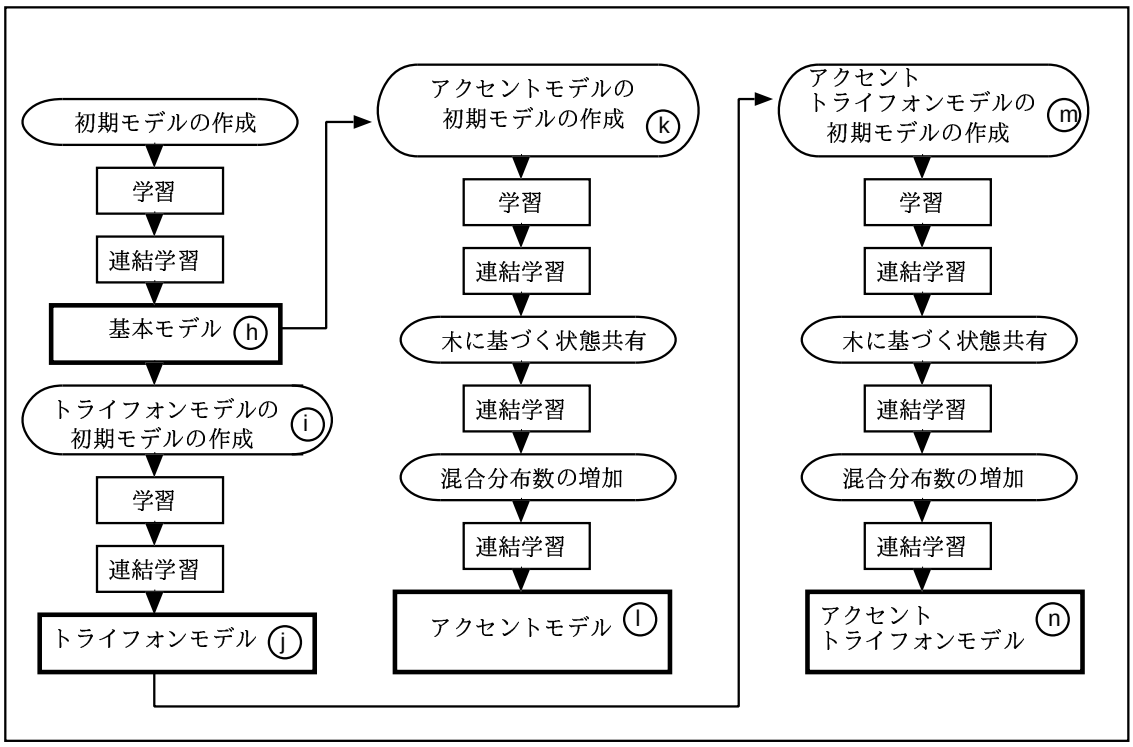

図 8: 状態共有型音素 HMM の作成手順

5.3 学習データと評価データ

本研究ではモデルの有効性検証のために, ATR 単語発話データベース Aset を用いる. なお, データベースには, 男性話者 10 名と女性話者 10 名が発話した単語の音声データが収録されている. そして, 話者毎に 5240 単語の音声データがある. また, 各音声データには, 人手によって付与された音素境界位置情報が与えられる.

各モデルにおいてのデータベースに含まれる音素の種類の数を表 7 に示す. なお, 本研究ではシンボル出力の状態数が 3 の音素 HMM を用いるので, 連続型 HMM の状態数は音素数 ×3 となる.

表 7: モデルにおける音素数

基本モデル	モーラモデル	アクセントモデル	triphone モデル	モーラ triphone モデル	アクセント triphone モデル
35	約 150	約 450	約 2300	約 6700	約 10300

本研究では, 男女別に, 認識対象話者以外の 9 話者分の奇数番を学習データに, 認識対象話者の偶数番を評価データに用いる. 同音異義語認識実験は, 2620 単語の音声認識を行い, 評価データ中のアクセントの異なる 11 組の同音異義語の認識結果で評価する. 詳細は表 8 に示す.

実験で用いられる評価データ中の同音異義語を表 10 に示す. また, 学習データ中の 31 組の同音異義語を表 9 に示す. 表中のデータ番号はデータベースにおいて付けられているデータの番号を示す. また, 括弧内の数字の 0 はアクセントの低, 1 は高を意味する.

表 8: 実験に用いたデータベース

データベース	ATR 単語発話データベース Aset　5240 単語/話者
認識対象話者	6 話者 (男性 3 話者 (mau,mmy,mnm), 女性 3 話者 (faf,ftk,fms))
学習データ	奇数番号 2620 単語/話者 ×9 話者
評価データ	偶数番号 2620 単語/話者
	(11 組のアクセントの異なる同音異義語が存在)

なお, 表 10 中の実験に用いる 6 話者の単語のアクセントは人手による聴取結果と一致することを確認した. また, 表 9 中の単語のアクセントを人手で聴取した. その結果, データ番号 10762 と 14882 の単語がアクセント辞典から決定したアクセントと異なることを確認した. 聴取結果を表 11 に示す. 聴取結果より他のデータにもアクセントの誤りがあると考えられるが, 数が多いためにアクセントの訂正は行っていない.

表 9: 認識データ中の同音異義語の対

	データ番号		データ番号	
1.	10150	ある (10)	10152	有る (10)
2.	10192	息 (10)	10194	意気 (10)
3.	10322	居る (01)	10324	射る (10)
4.	10558	置く (01)	10560	億 (10)
5.	10666	折る (10)	10668	織る (10)
6.	10734	代える (011)	10736	返る (100)
7.	10760	書く (10)	10762	角 (10)
8.	10788	欠ける (011)	10790	駆ける (010)
9.	11042	器械 (010)	11044	機械 (010)
10.	11056	利く (01)	11058	菊 (01)
11.	11062	起源 (100)	11064	機嫌 (011)
12.	11520	公演 (0111)	11522	講演 (0111)
13.	11524	公開 (0111)	11526	航海 (1000)
14.	11564	公正 (0111)	11566	構成 (0111)
15.	11830	咲く (01)	11832	柵 (01)
16.	12118	氏名 (100)	12120	指名 (011)
17.	12616	住む (10)	12618	澄む (10)
18.	12642	背 (10)	12644	性 (10)
19.	12732	千 (10)	12734	線 (10)
20.	13020	度 (01)	13022	足袋 (10)
21.	13270	付ける (010)	13272	漬ける (011)
22.	13486	解く (10)	13488	徳 (01)
23.	13858	刃 (1)	13860	歯 (1)
24.	13890	吐く (10)	13892	掃く (10)
25.	13960	放す (010)	13962	離す (010)
26.	14216	拭く (01)	14218	服 (01)
27.	14520	巻く (01)	14522	幕 (01)
28.	14880	焼く (01)	14882	約 (01)
29.	15070	因る (01)	15072	夜 (10)
30.	15142	礼 (10)	15144	零 (10)
31.	15210	沸く (01)	15212	枠 (01)

表 10: 認識データ中のアクセントの異なる同音異義語の対

1.	居る (01)	射る (10)
2.	代える (011)	返る (100)
3.	欠ける (011)	駆ける (010)
4.	機嫌 (011)	起源 (100)
5.	公開 (0111)	航海 (1000)
6.	置く (01)	億 (10)
7.	指名 (011)	氏名 (100)
8.	度 (01)	足袋 (10)
9.	徳 (01)	解く (10)
10.	付ける (010)	漬ける (011)
11.	因る (01)	夜 (10)

表 11: アクセントの聴取による評価
○:アクセント辞典と聴取結果が同一と判断
△:判断がつかないと判断
×:アクセント辞典と聴取結果が異なると判断

番号	mau	mmy	mnm	faf	fms	ftk
10762	△	×	×	×	×	△
14882	×	△	×	×	×	×
その他	○	○	○	○	○	○

表 12: 実験条件

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態 半連続分布型
stream 数	3
MFCC 特徴ベクトル	12 次 MFCC+Δ 12 次 MFCC +対数パワー+Δ対数パワー (計 26 次)
FBANK 特徴ベクトル	24 次 FBANK+Δ 24 次 FBANK +対数パワー+Δ対 5 数パワー (計 50 次)
半連続型 HMM Diagonal 混合分布数	MFCC 1024 Δ MFCC 1024 対数パワー, Δ対数パワー 64
半連続型 HMM Full 混合分布数	MFCC 128 Δ MFCC 128 対数パワー, Δ対数パワー 16
状態共有型 HMM Diagonal 混合分布数	MFCC 4 Δ MFCC 4 対数パワー, Δ対数パワー 2
FBANK の混合分布数は MFCC と同様なので省略	

5.4 実験条件

評価実験は, 男性話者 3 名と女性話者 3 名で行う. 実験には単語音声認識ツールの HTK [7] を使用する. その他の実験条件を表 12 に示す. HTK の設定ファイルは付録に示す. なお, 状態共有型 HMM において, 半連続型 HMM の Diagonal の実験条件と同一にするために, 混合分布数は 2112(1024 + 1024 + 64) とする. stream 数は 3 に設定し,MFCC を用いた実験では MFCC, Δ MFCC, 対数パワーと Δ対数パワーを, FBANK を用いた実験では FBANK, Δ FBANK, 対数パワーと Δ対数パワーをそれぞれ別の多次元ガウス分布で表現する. なお, MFCC と FBANK の特徴パラメータの次数は同じにするのが困難である. そのため, 特徴パラメータの次数は同一にしていない. また, 初期モデルの混合分布数を表 13 に示す.

表 13: 初期モデルの混合分布数

半連続型 HMM	Diagonal	MFCC 4 Δ MFCC 4 対数パワー, Δ 対数パワー 2
		FBANK 4 Δ FBANK 4 対数パワー, Δ 対数パワー 2
	Full	MFCC 1 Δ MFCC 1 対数パワー, Δ 対数パワー 1
		FBANK 1 Δ FBANK 1 対数パワー, Δ 対数パワー 1
状態共有型 HMM	Diagonal	MFCC 1 Δ MFCC 1 対数パワー, Δ 対数パワー 1
		FBANK 1 Δ FBANK 1 対数パワー, Δ 対数パワー 1

6 実験結果

6.1 同音異義語の認識精度

6.1.1 半連続型 HMM

半連続型 HMM, MFCC, Diagonal での同音異義語の認識精度を表 14 に示す. FBANK, Diagonal における同音異義語の認識精度を表 15 に示す. MFCC, Full における同音異義語の認識精度を表 16 に示す. FBANK, Full における同音異義語の認識精度を表 17 に示す.

表 14: 半連続型 HMM, MFCC, Diagonal を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	27%(6/22)	18%(4/22)
mmy	18%(4/22)	27%(6/22)
mnm	36%(8/22)	27%(6/22)
faf	23%(5/22)	18%(4/22)
fms	9%(2/22)	0%(0/22)
ftk	6%(6/22)	27%(6/22)
平均	23%(31/132)	20%(26/132)

表 15: 半連続型 HMM, FBANK, Diagonal を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	23%(5/22)	27%(6/22)
mmy	23%(5/22)	27%(6/22)
mnm	41%(9/22)	32%(7/22)
faf	23%(5/22)	23%(5/22)
fms	5%(1/22)	0%(0/22)
ftk	32%(7/22)	18%(4/22)
平均	24%(32/132)	21%(28/132)

表 16: 半連続型 HMM, MFCC, Full を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	14%(3/22)	5%(1/22)
mmy	23%(5/22)	5%(1/22)
mnm	32%(7/22)	14%(3/22)
faf	5%(1/22)	5%(1/22)
fms	9%(2/22)	9%(2/22)
ftk	27%(6/22)	27%(6/22)
平均	18%(24/132)	11%(14/132)

表 17: 半連続型 HMM, FBANK, Full を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	18%(4/22)	14%(3/22)
mmy	27%(6/22)	32%(7/22)
mnm	45%(10/22)	32%(7/22)
faf	0%(0/22)	9%(2/22)
fms	5%(1/22)	0%(0/22)
ftk	14%(3/22)	9%(2/22)
平均	18%(24/132)	16%(21/132)

6.1.2 状態共有型 HMM

状態共有型HMMにおけるMFCCでのの同音異義語の認識精度を表18に示す. FBANKでのの同音異義語の認識精度を表19に示す.

表 18: 状態共有型 HMM, MFCC, Diagonal を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	36%(8/22)	36%(8/22)
mmy	55%(12/22)	36%(8/22)
mnm	41%(9/22)	50%(11/22)
faf	27%(6/22)	14%(3/22)
fms	18%(4/22)	36%(8/22)
ftk	27%(6/22)	32%(7/22)
平均	34%(45/132)	34%(45/132)

表 19: 状態共有型 HMM, FBANK, Diagonal を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	14%(3/22)	45%(10/22)
mmy	55%(12/22)	45%(10/22)
mnm	45%(10/22)	50%(11/22)
faf	27%(6/22)	36%(8/22)
fms	36%(8/22)	27%(6/22)
ftk	33%(5/22)	41%(9/22)
平均	33%(44/132)	41%(54/132)

実験より以下の結果を得た.

1. MFCC は FBANK より同音異義語の認識精度が高い.
2. アクセント triphone モデルの方がアクセントモデルより同音異義語の認識精度が高い.
3. 状態共有型 HMM の認識率は半連続型 HMM と比べ低い.
4. 最も同音異義語を認識できた実験では平均 89%の精度が得られた.

6.2 単語音声認識精度

6.2.1 半連続型 HMM

半連続型 HMM, MFCC, Diagonal での単語認識精度を表 20 に示す. FBANK, Diagonal における単語認識精度を表 21 に示す. MFCC, Full における単語認識精度を表 22 に示す. FBANK, Full における単語認識精度を表 23 に示す.

表 20: 半連続型 HMM, MFCC, Diagonal の単語音声認識誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	12.10% (317/2620)	3.51% (92/2620)	8.63% (226/2620)	8.28% (217/2620)	4.50% (118/2620)	3.85% (101/2620)
mmy	12.98% (340/2620)	7.86% (206/2620)	11.03% (289/2620)	8.89% (232/2620)	9.89% (233/2620)	7.48% (196/2620)
mnm	11.34% (297/2620)	4.69% (123/2620)	9.66% (253/2620)	9.12% (239/2620)	5.50% (144/2620)	4.66% (122/2620)
faf	10.69% (280/2620)	6.07% (159/2620)	10.84% (284/2620)	9.66% (253/2620)	7.71% (202/2620)	6.26% (164/2620)
fms	13.21% (346/2620)	4.58% (120/2620)	9.16% (240/2620)	8.28% (217/2620)	5.00% (131/2620)	4.47% (117/2620)
ftk	11.26% (295/2620)	7.25% (190/2620)	6.95% (182/2620)	7.06% (185/2620)	8.17% (214/2620)	7.71% (202/2620)
平均	11.81% (1875/15720)	5.66% (890/15720)	9.38% (1474/15720)	8.54% (1343/15720)	6.63% (1042/15720)	5.74% (902/15720)

6.2.2 状態共有型 HMM

状態共有型 HMM における MFCC でのの単語認識精度を表 24 に示す. FBANK での単語認識精度を表 25 に示す.

実験より以下の結果を得た.

1. MFCC は FBANK より同音異義語の認識精度が高い.
2. 状態共有型 HMM の認識率は半連続型 HMM と比べ低い.
3. 半連続型 HMM の FULL の MFCC で最も高く, 94.65%の認識精度が得られた.
4. どの条件でも, 認識精度はアクセント triphone モデルまたは triphone モデル, アクセントモデル, 基本モデルの順で高い.

表 21: 半連続型 HMM, FBANK,Diagonal の単語音声認識誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	14.62% (383/2620)	10.80% (283/2620)	10.38% (272/2620)	9.43% (247/2620)	12.21% (320/2620)	12.86% (337/2620)
mmy	14.12% (370/2620)	9.50% (249/2620)	11.26% (295/2620)	13.29% (346/2620)	11.60% (304/2620)	10.73% (281/2620)
mnm	13.89% (364/2620)	9.35% (245/2620)	11.64% (305/2620)	11.98% (314/2620)	11.18% (293/2620)	9.96% (261/2620)
faf	12.21% (320/2620)	7.48% (196/2620)	9.16% (240/2620)	8.74% (229/2620)	8.89% (233/2620)	7.44% (195/2620)
fms	15.31% (401/2620)	6.26% (164/2620)	10.80% (283/2620)	10.46% (274/2620)	7.86% (206/2620)	7.48% (196/2620)
ftk	15.23% (399/2620)	12.86% (337/2620)	12.18% (319/2620)	13.32% (349/2620)	16.37% (429/2620)	15.08% (395/2620)
平均	14.23% (2237/15720)	9.38% (1474/15720)	10.90% (1714/15720)	11.19% (1759/15720)	11.35% (1785/15720)	10.59% (1665/15720)

表 22: 半連続型 HMM, MFCC,Full の単語音声認識誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	11.07% (290/2620)	3.02% (79/2620)	7.52% (197/2620)	6.22% (163/2620)	3.89% (102/2620)	3.36% (88/2620)
mmy	14.50% (380/2620)	8.28% (217/2620)	11.95% (313/2620)	9.62% (252/2620)	8.85% (232/2620)	7.52% (197/2620)
mnm	13.40% (351/2620)	5.99% (157/2620)	11.91% (312/2620)	10.61% (278/2620)	5.92% (155/2620)	5.27% (138/2620)
faf	11.60% (304/2620)	5.57% (146/2620)	9.77% (256/2620)	7.75% (203/2620)	5.95% (156/2620)	4.81% (126/2620)
fms	14.20% (372/2620)	5.34% (140/2620)	10.95% (287/2620)	8.85% (232/2620)	5.53% (145/2620)	4.50% (118/2620)
ftk	10.61% (278/2620)	5.69% (149/2620)	6.83% (179/2620)	6.64% (174/2620)	6.72% (176/2620)	6.64% (174/2620)
平均	12.56% (1975/15720)	5.65% (888/15720)	9.82% (1544/15720)	8.28% (1302/15720)	6.15% (966/15720)	5.35% (841/15720)

表 23: 半連続型 HMM, FBANK,Full の単語音声認識誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	9.16% (240/2620)	2.79% (73/2620)	5.99% (157/2620)	5.08% (133/2620)	4.08% (107/2620)	3.40% (89/2620)
mmy	16.53% (433/2620)	9.39% (246/2620)	13.40% (351/2620)	14.05% (368/2620)	11.30% (296/2620)	10.30% (270/2620)
mnm	16.37% (429/2620)	7.60% (199/2620)	13.40% (356/2620)	13.17% (345/2620)	9.66% (253/2620)	7.86% (206/2620)
faf	8.63% (226/2620)	6.03% (158/2620)	6.64% (174/2620)	5.27% (138/2620)	7.98% (209/2620)	5.42% (142/2620)
fms	12.10% (317/2620)	5.73% (150/2620)	9.05% (237/2620)	8.13% (213/2620)	8.09% (212/2620)	5.50% (144/2620)
ftk	10.31% (270/2620)	4.73% (124/2620)	7.02% (184/2620)	5.88% (154/2620)	6.60% (173/2620)	5.04% (132/2620)
平均	12.18% (1915/15720)	6.04% (950/15720)	9.28% (1459/15720)	8.59% (1351/15720)	9.05% (1423/15720)	6.25% (983/15720)

表 24: 状態共有型 HMM,MFCC,diagonal の単語音声認識誤り率

	基本モデル	アクセントモデル	アクセント triphone モデル	triphone モデル
mau	22.98% (602/2620)	15.15% (397/2620)	7.79% (204/2620)	7.86% (206/2620)
mmy	23.09% (605/2620)	17.44% (457/2620)	7.71% (202/2620)	12.71% (333/2620)
mnm	22.56% (591/2620)	16.37% (429/2620)	8.13% (213/2620)	9.58% (251/2620)
faf	21.41% (561/2620)	12.06% (316/2620)	6.60% (173/2620)	7.75% (203/2620)
fms	27.21% (713/2620)	15.04% (394/2620)	10.80% (283/2620)	11.76% (308/2620)
ftk	22.75% (596/2620)	11.26% (295/2620)	10.73% (281/2620)	11.91% (312/2620)
平均	23.33% (3668/15720)	14.55% (2288/15720)	8.63% (1356/15720)	10.26% (1613/15720)

表 25: 状態共有型 HMM,FBANK,diagonal の単語音声認識誤り率

	基本モデル	アクセントモデル	アクセント triphone モデル	triphone モデル
mau	41.64% (1091/2620)	12.56% (329/2620)	15.27% (400/2620)	14.05% (368/2620)
mmy	50.46% (1322/2620)	20.46% (536/2620)	12.82% (336/2620)	13.09% (343/2620)
mnm	42.79% (1121/2620)	16.30% (427/2620)	12.37% (324/2620)	11.72% (307/2620)
faf	34.58% (906/2620)	11.11% (291/2620)	14.01% (367/2620)	13.97% (366/2620)
fms	46.15% (1209/2620)	19.54% (512/2620)	14.39% (377/2620)	12.94% (339/2620)
ftk	51.22% (1342/2620)	15.92% (417/2620)	18.44% (483/2620)	18.97% (497/2620)
平均	44.47% (6991/15720)	15.98% (2512/15720)	14.55% (2287/15720)	14.12% (2220/15720)

7　その他の実験結果

参考として, 本研究での実験条件以外の実験結果を示す. ほとんどの実験は, 実験条件を同一にしていない.

7.1　同音異義語の認識精度

7.1.1　状態数無調整の状態共有型 HMM

状態数を調整していない状態共有型 HMM, MFCC での同音異義語認識精度を表 28 に示す. また, FBANK での同音異義語認識精度を表 29 に示す. 状態数を制御する閾値が一定であるために, 状態数は各実験条件とモデルによって異なる. 状態数以外の実験条件は, 本研究での実験条件と同一である. MFCC での状態数を表 26 に, FBANK での状態数を表 27 に示す.

表 26: モデルにおける MFCC の状態数

モーラモデル	アクセントモデル	triphone モデル	モーラ triphone モデル	アクセント triphone モデル
約 200	約 300	約 1000	約 1050	約 1100

表 27: モデルにおける FBANK の状態数

モーラモデル	アクセントモデル	triphone モデル	モーラ triphone モデル	アクセント triphone モデル
約 300	約 550	約 1500	約 1750	約 1900

実験より以下の結果を得た.

1. 得られた精度は, 同条件の半連続型 HMM の Diagonal と状態数を調整した状態共有型 HMM の精度より高い. しかし, 半連続型 HMM の Full よりは高くない.

表 28: MFCC,diagonal, 状態数無調整の状態共有型 HMM の同音異義語誤り率

	アクセントモデル	アクセント triphone モデル
mau	41%(9/22)	18%(4/22)
mmy	45%(10/22)	14%(3/22)
mnm	32%(7/22)	23%(5/22)
faf	27%(6/22)	14%(3/22)
fms	14%(3/22)	18%(4/22)
ftk	32%(7/22)	27%(6/22)
平均	32%(42/132)	19%(25/132)

表 29: FBANK,diagonal, 状態数無調整の状態共有型 HMM の同音異義語誤り率

	アクセントモデル	アクセント triphone モデル
mau	18%(4/22)	36%(8/22)
mmy	50%(11/22)	14%(3/22)
mnm	36%(8/22)	27%(6/22)
faf	27%(6/22)	14%(3/22)
fms	23%(5/22)	0%(0/22)
ftk	14%(3/22)	14%(3/22)
平均	28%(37/132)	17%(23/132)

7.1.2 特定話者実験の半連続型 HMM

特定話者における半連続型 HMM を用いた MFCC, Diagonal のアクセントモデルの同音異義語の認識精度を表 30 に示す. また, FBANK, Diagonal の同音異義語の認識精度を表 31 に示す. 表 30, 31 の実験結果は [12] での結果であり, 実験条件は [12] に示されている.

また, [12] の条件で stream 数と混合分布数が異なる実験結果を示す. MFCC, Diagonal の同音異義語認識精度を表 32 に, FBANK, Diagonal の同音異義語認識精度を表 34 に, MFCC, Full の同音異義語認識精度を表 33 に示す.

表 30: 特定話者におけるアクセントモデル, 半連続型 HMM, MFCC を用いた同音異義語誤り率

話者	Diagonal	Full
mau	18%(4/22)	9%(2/22)
mmy	14%(3/22)	9%(2/22)
mnm	14%(3/22)	5%(1/22)
faf	0%(0/22)	5%(0/22)
fms	14%(3/22)	14%(3/22)
ftk	5%(1/22)	9%(2/22)
平均	11%(14/132)	8%(10/132)

表 31: 特定話者におけるアクセントモデル, 半連続型 HMM, FBANK を用いた同音異義語誤り率

話者	Diagonal	Full
mau	5%(1/22)	5%(1/22)
mmy	18%(4/22)	0%(0/22)
mnm	9%(2/22)	9%(2/22)
faf	0%(0/22)	0%(0/22)
fms	18%(4/22)	5%(1/22)
ftk	5%(1/22)	0%(0/22)
平均	9%(12/132)	3%(4/132)

表 30, 31 の実験より以下の結果を得た.

1. MFCC より FBANK を用いた特徴パラメータの方がアクセントの認識精度が高い.

表 32: 特定話者における半連続型 HMM,MFCC,Diagonal,stream 数 1, 混合分布数 256 の同音異義語誤り率

	アクセントモデル	アクセント triphone モデル
mau	27%(6/22)	23%(5/22)
mmy	13%(3/22)	18%(4/22)
mnm	13%(3/22)	23%(5/22)
faf	22%(5/22)	27%(6/22)
fms	18%(4/22)	9%(2/22)
ftk	4%(1/22)	9%(2/22)
平均	16%(22/132)	18%(24/132)

表 33: 特定話者における半連続型 HMM,MFCC,Full,stream 数 1, 混合分布数 256 の同音異義語誤り率

	アクセントモデル
mau	31%(7/22)
mmy	4%(1/22)
mnm	27%(6/22)
faf	4%(1/22)
fms	13%(3/22)
ftk	9%(2/22)
平均	14%(20/132)

2. 最も同音異義語を認識できた FBANK, Full の実験では平均 97%の精度が得られた.

表 34: 特定話者における半連続型 HMM,FBANK,Diagonal,stream 数 3, 混合分布数 256,256,32 の同音異義語誤り率

	アクセントモデル
mau	0%(0/22)
mmy	5%(1/22)
mnm	5%(1/22)
faf	0%(0/22)
fms	9%(2/22)
ftk	14%(3/22)
平均	5%(7/132)

7.1.3 特定話者実験の状態共有型 HMM

特定話者における状態共有型 HMM を用いた MFCC,Diagonal の同音異義語の認識精度を表 37 に示す. また FBANK,Diagonal の同音異義語の認識精度を表 38 に示す. なお, 状態数を制御する閾値が一定であるために, 状態数は各実験条件とモデルによって異なる. 状態数以外の実験条件は, 本研究での実験条件と同一である. MFCC での状態数を表 35 に, FBANK での状態数を表 36 に示す.

表 35: 特定話者のモデルにおける MFCC の状態数

モーラモデル	アクセントモデル	triphone モデル	モーラ triphone モデル	アクセント triphone モデル
約 150	約 200	約 400	約 425	約 450

表 36: 特定話者のモデルにおける FBANK の状態数

モーラモデル	アクセントモデル	triphone モデル	モーラ triphone モデル	アクセント triphone モデル
約 200	約 300	約 650	約 675	約 700

表 37: 特定話者における状態共有型 HMM,MFCC,Diagonal の同音異義語誤り率

	アクセントモデル	アクセント triphone モデル
mau	14%(3/22)	9%(2/22)
mmy	14%(3/22)	5%(1/22)
mnm	9%(2/22)	14%(3/22)
faf	9%(2/22)	5%(1/22)
fms	9%(2/22)	9%(2/22)
ftk	0%(0/22)	5%(1/22)
平均	9%(12/132)	8%(10/132)

表 38: 特定話者における状態共有型 HMM,FBANK,Diagonal の同音異義語誤り率

	アクセントモデル	アクセント triphone モデル
mau	9%(2/22)	0%(0/22)
mmy	5%(3/22)	5%(1/22)
mnm	9%(2/22)	9%(2/22)
faf	5%(1/22)	5%(1/22)
fms	0%(0/22)	0%(0/22)
ftk	5%(1/22)	0%(0/22)
平均	7%(9/132)	3%(4/132)

7.2 単語音声認識精度

7.2.1 状態数無調整の状態共有型 HMM

状態数を調整していない状態共有型 HMM, MFCC での単語認識精度を表 39 に示す. また, FBANK での単語認識精度を表 40 に示す. 状態数を制御する閾値が一定であるために, 状態数は各実験条件とモデルによって異なる. 状態数以外の実験条件は, 本研究での実験条件と同一である. MFCC での状態数を表 26 に, FBANK での状態数を表 27 に示す.

表 39: 状態数無調整の状態共有型 HMM,MFCC,diagonal の単語誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	22.98% (602/2620)	4.35% (114/2620)	14.52% (383/2620)	14.27% (374/2620)	4.58% (120/2620)	4.01% (105/2620)
mmy	23.09% (605/2620)	8.97% (235/2620)	18.09% (474/2620)	16.49% (432/2620)	9.27% (243/2620)	7.10% (186/2620)
mnm	22.56% (591/2620)	4.16% (109/2620)	17.14% (449/2620)	16.49% (432/2620)	4.05% (106/2620)	3.70% (97/2620)
faf	21.41% (561/2620)	5.53% (145/2620)	14.01% (367/2620)	12.48% (327/2620)	5.73% (150/2620)	4.73% (124/2620)
fms	27.21% (713/2620)	5.92% (155/2620)	16.79% (440/2620)	14.73% (386/2620)	5.53% (145/2620)	4.96% (130/2620)
ftk	22.75% (596/2620)	9.69% (254/2620)	12.82% (336/2620)	11.79% (309/2620)	8.93% (234/2620)	8.28% (217/2620)
平均	23.33% (3668/15720)	6.44% (1012/15720)	15.58% (2449/15720)	14.38% (2260/15720)	6.35% (998/15720)	5.46% (859/15720)

表 40: 状態数無調整の状態共有型 HMM,FBANK,diagonal の単語誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	41.64% (1091/2620)	13.36% (350/2620)	12.60% (330/2620)	11.56% (303/2620)	12.67% (332/2620)	12.63% (331/2620)
mmy	50.46% (1322/2620)	7.56% (198/2620)	18.47% (484/2620)	19.43% (509/2620)	8.32% (218/2620)	7.86% (206/2620)
mnm	42.79% (1121/2620)	8.05% (211/2620)	16.26% (426/2620)	15.27% (400/2620)	7.82% (205/2620)	6.53% (171/2620)
faf	34.58% (906/2620)	8.05% (211/2620)	12.79% (335/2620)	11.11% (291/2620)	8.24% (216/2620)	6.79% (178/2620)
fms	46.15% (1209/2620)	6.64% (174/2620)	19.54% (512/2620)	17.33% (454/2620)	6.91% (181/2620)	4.69% (123/2620)
ftk	51.22% (1342/2620)	10.84% (284/2620)	16.26% (426/2620)	10.99% (288/2620)	10.46% (274/2620)	9.50% (249/2620)
平均	44.47% (6991/15720)	9.08% (1428/15720)	15.99% (2513/15720)	14.28% (2245/15720)	9.07% (1426/15720)	8.00% (1258/15720)

7.2.2 特定話者実験の半連続型 HMM

特定話者での, stream 数 3, 混合分布数 128 128 16, 半連続型 HMM の MFCC, Diagonal における単語音声認識精度を表 41 に示す. FBANK, Diagonal における単語音声認識精度を表 42 に示す. MFCC, Full における単語音声認識精度を表 43 に示す. FBANK, Full における単語音声認識精度を表 44 に示す. 表 41, 42, 43, 43 の実験結果は [12] での結果であり, 実験条件は [12] に示されている.

また, [12] の条件で stream 数と混合分布数が異なる実験結果を示す. 表 45, 表 46, 表 47 は実験条件を同一にしていない.

表 41: 特定話者における半連続型 HMM,MFCC,Diagonal,stream 数 3, 混合分布数 128,128,16 での単語誤り率

話者	基本モデル	アクセントモデル
mau	6.76%(177/2620)	3.85%(101/2620)
mmy	7.21%(189/2620)	4.58%(120/2620)
mnm	8.13%(213/2620)	4.16%(109/2620)
faf	7.33%(192/2620)	3.78%(99/2620)
fms	7.06%(185/2620)	5.23%(137/2620)
ftk	6.82%(179/2620)	4.16%(109/2620)
平均	7.22%(1135/15720)	4.29%(675/15720)

表 42: 特定話者における半連続型 HMM,FBANK,Diagonal,stream 数 3, 混合分布数 128,128,16 での単語誤り率

話者	基本モデル	アクセントモデル
mau	10.31%(270/2620)	7.02%(184/2620)
mmy	12.29%(322/2620)	7.10%(186/2620)
mnm	10.34%(271/2620)	7.33%(192/2620)
faf	8.70%(228/2620)	6.37%(167/2620)
fms	11.45%(300/2620)	8.21%(215/2620)
ftk	9.77%(256/2620)	7.18%(188/2620)
平均	10.48%(1647/15720)	7.20%(1132/15720)

表 43: 特定話者における半連続型 HMM,MFCC,Full,stream 数 3, 混合分布数 128,128,16 での単語誤り率

話者	基本モデル	アクセントモデル
mau	4.69%(123/2620)	3.21%(84/2620)
mmy	6.18%(162/2620)	3.74%(98/2620)
mnm	5.46%(143/2620)	3.40%(89/2620)
faf	4.69%(123/2620)	3.05%(80/2620)
fms	5.50%(144/2620)	3.51%(92/2620)
ftk	4.85%(127/2620)	3.40%(89/2620)
平均	5.23%(822/15720)	3.38%(532/15720)

表 44: 特定話者における半連続型 HMM,FBANK,Full,stream 数 3, 混合分布数 128,128,16 の単語誤り率

	基本モデル	モーラモデル	アクセントモデル
mau	5.21%(136/2611)	3.03%(79/2611)	2.94%(77/2611)
mmy	6.09%(159/2611)	3.18%(83/2611)	3.03%(79/2611)
mnm	5.48%(143/2611)	3.14%(82/2611)	3.14%(82/2611)
faf	4.79%(125/2611)	3.83%(100/2611)	3.33%(87/2611)
fms	5.52%(144/2611)	3.98%(104/2611)	3.87%(101/2611)
ftk	5.78%(151/2611)	3.68%(96/2611)	3.41%(89/2611)
平均	5.48%(858/15666)	3.47%(544/15666)	3.29%(515/15666)

表 45: 特定話者における半連続型 HMM,MFCC,Diagonal,stream 数 1, 混合分布数 256 の単語誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	5.31% (139/2620)	2.48% (65/2620)	4.12% (108/2620)	4.12% (108/2620)	2.25% (59/2620)	2.29% (60/2620)
mmy	6.22% (163/2620)	4.05% (106/2620)	4.54% (119/2620)	4.54% (119/2620)	4.08% (107/2620)	3.93% (103/2620)
mnm	6.07% (159/2620)	2.75% (72/2620)	4.31% (113/2620)	4.23% (111/2620)	2.79% (73/2620)	2.44% (64/2620)
faf	7.71% (202/2620)	4.62% (121/2620)	5.64% (148/2620)	5.03% (132/2620)	4.85% (127/2620)	4.35% (114/2620)
fms	5.53% (145/2620)	2.86% (75/2620)	3.96% (104/2620)	3.92% (103/2620)	2.86% (75/2620)	3.36% (88/2620)
ftk	6.56% (172/2620)	3.85% (101/2620)	5.15% (135/2620)	4.35% (114/2620)	4.20% (110/2620)	3.70% (97/2620)
平均	6.23% (980/15720)	3.43% (540/15720)	4.62% (727/15720)	4.36% (687/15720)	3.51% (551/15720)	3.35% (526/15720)

表 46: 特定話者における半連続型 HMM,FBANK,Diagonal,stream 数 3, 混合分布数 256 256 32 の単語誤り率

	基本モデル	モーラモデル	アクセントモデル
mau	5.21%(173/2620)	5.15%(135/2620)	5.92%(155/2620)
mmy	7.94%(208/2620)	6.20%(165/2620)	6.22%(163/2620)
mnm	8.17%(214/2620)	5.57%(146/2620)	5.80%(152/2620)
faf	6.79%(178/2620)	5.38%(141/2620)	5.15%(135/2620)
fms	7.06%(185/2620)	5.00%(131/2620)	5.61%(145/2620)
ftk	7.18%(188/2620)	5.42%(142/2620)	6.07%(159/2620)
平均	7.06%(1146/15720)	5.45%(860/15720)	5.80%(909/15720)

表 47: 特定話者における半連続型 HMM,MFCC,Full,stream 数 1, 混合分布数 256 の単語誤り率

	基本モデル	triphone モデル	モーラモデル	アクセントモデル
mau	3.79% (99/2611)	1.38% (36/2611)	2.75% (72/2611)	2.06% (54/2611)
mmy	4.21% (110/2611)	2.95% (77/2611)	2.83% (74/2611)	2.64% (69/2611)
mnm	3.68% (96/2611)	1.65% (43/2611)	2.25% (59/2611)	2.33% (61/2611)
faf	3.18% (83/2611)	2.14% (56/2611)	2.45% (64/2611)	1.95% (51/2611)
fms	3.87% (101/2611)	2.11% (55/2611)	2.87% (75/2611)	2.91% (76/2611)
ftk	3.19% (86/2611)	1.76% (46/2611)	2.37% (62/2611)	2.83% (74/2611)
平均	3.65% (575/15720)	1.99% (313/15720)	2.58% (406/15720)	2.45% (385/15720)

7.2.3 特定話者実験の状態共有型 HMM

状態数を調整していない特定話者における状態共有型 HMM, MFCC での単語認識精度を表 48 に示す. また, FBANK での単語認識精度を表 49 に示す. 状態数を制御する閾値が一定であるために, 状態数は各実験条件とモデルによって異なる. 状態数以外の実験条件は, 本研究での実験条件と同一である. MFCC での状態数を表 35 に, FBANK での状態数を表 36 に示す.

表 48: 特定話者における状態共有型 HMM,MFCC,diagonal の単語誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	13.70% (359/2620)	1.68% (44/2620)	7.52% (197/2620)	7.25% (190/2620)	2.25% (59/2620)	2.10% (55/2620)
mmy	14.12% (370/2620)	2.37% (62/2620)	7.98% (209/2620)	7.75% (203/2620)	2.14% (56/2620)	2.63% (69/2620)
mnm	14.43% (378/2620)	1.87% (49/2620)	8.66% (227/2620)	7.75% (203/2620)	2.14% (56/2620)	1.91% (50/2620)
faf	12.63% (331/2620)	2.14% (56/2620)	7.52% (197/2620)	6.41% (168/2620)	2.06% (54/2620)	1.91% (50/2620)
fms	16.49% (432/2620)	2.10% (55/2620)	8.93% (234/2620)	7.71% (202/2620)	1.79% (47/2620)	2.02% (53/2620)
ftk	14.16% (371/2620)	1.98% (52/2620)	7.71% (202/2620)	6.64% (174/2620)	1.54% (43/2620)	1.79% (47/2620)
平均	14.26% (2241/15720)	2.02% (318/15720)	8.05% (1266/15720)	7.25% (1140/15720)	2.00% (315/15720)	1.99% (313/15720)

表 49: 特定話者における状態共有型 HMM,FBANK,diagonal の単語誤り率

	基本モデル	triphone モデル	モーラ モデル	アクセント モデル	モーラ triphone モデル	アクセント triphone モデル
mau	28.82% (755/2620)	0.95% (25/2620)	7.59% (112/2620)	5.92% (155/2620)	1.41% (37/2620)	1.11% (29/2620)
mmy	35.23% (923/2620)	2.75% (72/2620)	8.78% (230/2620)	7.75% (203/2620)	2.37% (62/2620)	2.67% (70/2620)
mnm	32.33% (847/2620)	2.14% (56/2620)	9.16% (240/2620)	7.18% (188/2620)	1.98% (52/2620)	1.64% (43/2620)
faf	26.22% (687/2620)	2.18% (57/2620)	6.72% (176/2620)	6.49% (170/2620)	2.06% (54/2620)	1.68% (44/2620)
fms	29.81% (781/2620)	1.68% (44/2620)	8.47% (222/2620)	7.90% (207/2620)	1.79% (47/2620)	1.79% (47/2620)
ftk	26.95% (706/2620)	2.21% (58/2620)	7.52% (197/2620)	6.49% (170/2620)	2.25% (59/2620)	1.95% (51/2620)
平均	29.89% (4699/15720)	1.98% (312/15720)	7.49% (1177/15720)	6.95% (1093/15720)	1.98% (311/15720)	1.81% (284/15720)

8 考察

8.1 同音異義語の認識結果に対する考察

8.1.1 同音異義語の誤認識

多くの実験結果において, 男性話者の同音異義語認識精度は女性話者より低い. しかし, 認識精度が最も高い半連続型 HMM を用いた MFCC, Full のアクセント triphone モデルの実験においてのみ, 男性の認識精度は 92%(61/66), 女性は 86%(57/66) となる. そして, 最も高い男性の認識精度は半連続型 HMM を用いた MFCC, Full のアクセント triphone モデルでの 92%である. また, 最も高い女性の認識精度は, 半連続型 HMM を用いた FBANK, Full のアクセントモデルとアクセント triphone モデルでの 94%(62/66) である. ただし, 話者毎の認識結果には大きな偏りがある

同音異義語の実験では, アクセントが低で始まる 3 モーラ以上の単語を誤認識する結果が多い. 誤認識の例を表 50 に示す. なお, 表 50, 53 の括弧内の数字の 0 はアクセントの低, 1 は高を意味する.

表 50: 同音異義語の誤認識例

認識結果	正解
航海 (1000)	公開 (0111)
付ける (100)	漬ける (011)

8.1.2 単語の誤認識

半連続型 HMM での単語を同音異義語に誤認識した割合を表 51, 状態共有型 HMM での単語を同音異義語に誤認識した割合を表 52 に示す. 表 51 の括弧内の分母は, 表 14, 15, 16, 17 での 6 話者の同音異義語の誤り数である. また, 表 52 の括弧内の分母は, 表 18, 19 での 6 話者の同音異義語の誤り数である. 表 51, 52 の括弧内の分子は同音異義語を別のアクセントの同音異義語に誤認識した数を示す.

アクセント triphone モデルにおいて, 単語を同音異義語以外の単語に誤認識した割合は平均 39%(74/188) であり, アクセントモデルの 51%(101/200) と比べて低い. アクセントモデルの前後音素環境を考慮することによって, 単語音声認識精度が向上したためだと考えられる. 同音異義語ではない別の単語に誤認識した例を表 53 に示す.

なお, 特定話者の半連続型 HMM において, アクセント triphone モデルの認識精度は,

アクセントモデルより低い. アクセント triphone モデルにおいて HMM のパラメータ数が膨大となり, 学習データが不足するためだと考えられる.

表 51: 半連続型 HMM での単語を同音異義語に誤認識した割合

	アクセントモデル	アクセント triphone モデル
Diagonal MFCC	61%(19/31)	84%(22/26)
Diagonal FBANK	66%(21/32)	71%(20/28)
Full MFCC	50%(12/24)	79%(11/14)
Full FBANK	54%(13/24)	81%(17/21)
平均	59%(65/111)	79%(70/89)

表 52: 状態共有型 HMM での単語を同音異義語に誤認識した割合

	アクセントモデル	アクセント triphone モデル
Diagonal MFCC	31%(14/45)	49%(22/45)
Diagonal FBANK	45%(20/44)	41%(22/54)
平均	38%(34/89)	44%(44/99)

表 53: 単語を同音異義語ではない単語とした誤認識例

認識結果	正解
徳 (01)	置く (01)
堪える (010)	返る (100)
勤勉 (0111)	機嫌 (011)
酔う (10)	夜 (10)
解く (10)	億 (10)
生える (010)	代える (011)
悲劇 (100)	機嫌 (011)

8.2 単語音声認識に対する考察

8.2.1 アクセント情報と前後音素環境情報

不特定話者実験のほとんどの実験結果において, triphone モデルとアクセント triphone モデルの認識精度の差は小さい. アクセント triphone モデルは, 単語のアクセント型と各モーラ位置でのアクセントの高低情報と, triphone モデルと同じく前後音素環境情報を用いている. なお, アクセント情報のみを用いたアクセントモデルの単語音声認識精度は基本モデルより高く, アクセント情報は単語音声認識に対して有効である. また, 前後音素環境情報は単語音声認識に対して有効なことが知られている.

アクセント情報は単語のモーラ数とモーラ位置の情報を含んでいる. そして, 単語のモーラ数とモーラ位置によって, 前後音素環境はある程度定まると考えられる. つまり, アクセントの情報と前後音素環境情報が似ているために, 認識精度に差が出なかったと考えている.

8.2.2 アクセント情報とモーラ情報

半連続型 HMM において, アクセント情報を用いたモデルは同条件のモーラ情報を用いたモデルより単語音声認識精度が高い. ピッチ周波数はモーラ情報に依存し, モーラ情報を用いたモデルで単語音声認識精度が高いことは知られている [13].

アクセント情報はモーラ情報を含んでおり, アクセント情報を用いることによって, ピッチ周波数がフォルマントに与える影響をモーラ情報のみより考慮できると考えられる. そのためアクセント情報を用いたモデルの単語音声認識精度がモーラ情報を用いたモデルより高いと考えている.

8.3 FBANKとMFCC

不特定話者のほとんどの実験結果において, FBANK の同音異義語認識精度は MFCC と比べて低い. しかし, 特定話者において FBANK を用いた同音異義語認識精度は MFCC と比べ高い. FBANK は MFCC より話者とモデルの依存度が高い結果となっている.

本研究では, 評価する話者以外の多数の話者のデータを学習データとして用いている. しかし, FBANK は話者を特徴付ける韻律情報を含んでおり, 多数の韻律情報の学習によって特徴が平滑化されると認識に有効ではないと考える. 本研究で用いた手法以外の不特定話者のモデル作成方法には, 多数のモデルを作成しておき評価データから最適なモデルを選択する話者選択手法や, モデルを評価データに適合させる話者適合手法がある. 話者選択や話者適合手法を用いれば FBANK において韻律情報を有効に利用し, 不特定話者の認識精度を特定話者の認識精度に近付けることができると考えられる. そして, 特定話者において MFCC より認識精度が高い FBANK は, 不特定話者の認識精度を向上させる手法を用いると, 不特定話者の認識においても有効であると考えている.

8.4 木に基づく状態共有

8.4.1 不特定話者における実験結果の比較

表 54 に不特定話者, MFCC における状態共有型 HMM の実験結果を示す. 表 55 に不特定話者, MFCC における状態共有型 HMM の実験結果を示す. なお, 表中の状態数はおよその数である. 表 54, 55 の全てのモデルの結果において, 状態数を調整していない状態共有型 HMM の実験結果の単語と同音異義語の認識精度は, 状態数が約 200 の状態共有型 HMM の実験結果の認識精度より高い.

また, 各モデルの音素数は表 7 に示しているが, 状態共有を行わない時の状態数は表 56 となる. 状態共有によって状態数を 200 としたとき, triphone モデルの状態は 2.90%(200/6900) に, アクセントモデルは 14.81%(200/1350) に, アクセント triphone モデルは 0.65%(200/30900) になった.

実験結果より, 全てのモデルにおいて 200 程度の状態数は, 状態空間を表現するには不十分だと考えられる. また, 状態数を調整していない状態共有型 HMM の実験結果のモーラモデル, アクセントモデル以外のモデルの認識精度のほとんどは, 同条件の Diagonal の半連続型 HMM の認識精度より高く, Full の半連続型 HMM の認識精度より低い.

表 54: 不特定話者, MFCC における状態共有型 HMM の実験結果

	triphone モデル		アクセントモデル			アクセント triphone モデル		
	状態数	単語誤り率	状態数	単語誤り率	同音異義語誤り率	状態数	単語誤り率	同音異義語誤り率
表 24, 18 の実験結果	200	10.26% (1613/15720)	200	14.55% (2288/15720)	34% (45/132)	200	8.63% (1356/15720)	45% (45/132)
表 39, 28 の実験結果	1000	6.44% (1012/15720)	300	14.38% (2260/15720)	32% (42/132)	1100	5.46% (859/15720)	19% (25/132)

表 55: 不特定話者, FBANK における状態共有型 HMM の実験結果

	triphone モデル		アクセントモデル			アクセント triphone モデル		
	状態数	単語誤り率	状態数	単語誤り率	同音異義語誤り率	状態数	単語誤り率	同音異義語誤り率
表 25, 19 の実験結果	200	14.12 (2220/15720)	200	15.98% (2512/15720)	33% (44/132)	200	14.55% (2287/15720)	54% (54/132)
表 40, 29 の実験結果	1500	9.08% (1428/15720)	550	14.28% (2245/15720)	28% (37/132)	1900	8.00% (1258/15720)	17% (23/132)

表 56: モデルにおける状態数

triphone モデル	アクセントモデル	アクセント triphone モデル
約 6900	約 1350	約 30900

8.4.2 特定話者における状態共有型 HMM

特定話者の前後音素環境情報を用いた状態共有型 HMM において, 最も高い単語認識精度が得られた. また, 同音異義語認識において, FBANK, Full の半連続型 HMM と FBANK の状態共有型 HMM の認識精度が最も高い. ただし, 各実験の条件は, 混合分布数が異なっており同一ではない. また, FBANK の Diagonal では MFCC の Diagonal と比較して単語認識精度が劣ることが知られている [5],[12]. しかし, FBANK, Diagonal の状態共有型 HMM の単語認識精度が MFCC の結果より高いのは, FBANK での状態数が MFCC より適切な値となっているためだと考えている.

8.4.3 質問

triphone モデルに対する木に基づく状態共有の質問は, 英語音素用に作られた質問から英語音素を対応する日本語音素に変換して作成した. 作成した質問は, 連続型 HMM の状態数を減らし, 半連続型 HMM の混合分布数と同一にすることを可能にした. しかし, 質問が最適かどうかの評価は行っていない. より最適な日本語音素に対する質問を作成して利用することで, より高い認識精度を得ることが可能になると考える.

8.4.4 状態数

HMM の状態共有において, 状態数が多すぎれば状態あたりの学習データ数が減少し信頼性のあるパラメータ推定が行えない. また, 状態数が少なすぎれば特徴ベクトルデータを十分に表現できない. モデルや特徴ベクトル, 学習データ量等の実験条件に対して最適な HMM の状態数が存在するはずであるが, 本研究では調査していない. 認識精度向上のためには最適な状態数を調査する必要があると考える.

8.4.5 状態共有の調査

状態数を, 不特定話者で最も認識精度が高いアクセント triphone モデル, 状態数無調整の状態共有型 HMM, FBANK の話者 mau において簡単に調査した. 状態数の調査結果を表 57 に示す. なお, 表中の列の意味を以下に示す.

- “音素” の列はアクセント triphone モデルにおける音素表記の中心音素を示す. 例えば `m-a0302001+u0303001` の音素表記の中心音素は a である.

- “音素数” の列は左の列の音素を中心とする音素の数を示す.
- “状態数 2”, “状態数 3”, “状態数 4” の列はそれぞれの HMM の共有された状態数を示す. 本研究で用いる HMM は 5 状態であり, 状態数 1 と状態数 5 が存在する. しかし, 状態 1 と 5 はシンボル出力確率を持たない状態であり, 状態共有の対象にはならない.

共有された状態の一部を例として表 58 に示す. 表において, 共有状態名の ST_N_2_1 は中心音素 N の HMM の状態 2 の 17 番目の共有状態の名前を示す. 例えば, ST_N_2_17 において, a0201000-N0202001+pau, a0403001-N0404001+pau および a0403020-N0404020+pau の音素 HMM が状態 2 を共有している.

調査結果より以下の結果を得た.

1. 母音, 撥音を中心音素とする共有状態数が他の音素に比べて多い.
2. 母音・撥音において状態 3 の共有状態数が状態 2,4 と比べて多い.
3. 子音は上記の傾向と異なる結果が多い.
4. 母音において, 状態 4 の共有状態数は状態 2 より多い.
5. アクセント情報が付与されている母音, 撥音, 促音音素の $19653(302+1724+662+1235+1268+114+1246)\times 3$ の状態のうち, 1857 音素はアクセント情報に関する質問によって分類された. つまり, アクセント情報のみが異なる音素の 1857 の状態が異なる状態として共有されている. そして, 残りの $17796(19653-1857)$ の状態は前後の音素環境情報に関する質問のみによって分類された.

音素 HMM の状態 2 は音の立上り, 状態 3 は音の定常状態, 状態 4 は次の音への変化を表現すると考えられる. また, 複雑な音を表現するには, 多くの状態数が必要になると考えられる. ゆえに, 母音は音の定常部分が複雑で, 子音は音の変化部分が複雑だと考えている.

表 57: アクセント triphone モデル, 状態数無調整の状態共有型 HMM, FBANK の mau 不特定話者における状態数

音素	音素数	状態 2	状態 3	状態 4
N	302	19	22	13
a	1724	24	31	28
b	229	10	6	7
by	5	1	1	1
ch	119	5	4	4
d	150	6	5	7
e	662	14	26	25
f	30	2	2	1
g	237	6	7	11
gy	17	2	2	1
h	108	7	6	6
hy	10	1	1	1
i	1235	34	42	37
j	124	7	5	5
k	446	15	17	12
ky	38	2	4	4
m	329	11	10	8
my	9	1	1	1
n	227	6	4	8
ny	8	1	1	1
o	1268	22	41	42
p	47	2	2	2
pau	202	20	24	24
py	2	1	1	1
q	114	5	3	4
r	367	11	12	14
ry	37	2	2	2
s	252	11	11	8
sh	185	9	9	6
t	230	9	6	8
ts	98	4	3	2
u	1246	50	63	60
w	68	4	3	2
y	118	6	5	3
z	134	6	4	4
合計	10377	336	386	363

表 58: モデルにおける状態

共有状態名	共有された状態の集合
ST_N_2_1	e0301000-N0302001+g,e0301000-N0302001+k,e0301000-N0302001+ry,e0301011-N0302010+g,e0301011-N0302010+i0303010,e0301011-N0302010+k,e0301011-N0302010+ky,e0401000-N0402001+a0403001,e0401000-N0402001+g,e0401000-N0402001+i0403001,e0401000-N0402001+k,e0401000-N0402001+ky,e0401000-N0402001+ry,e0401011-N0402010+g,e0401011-N0402010+k,e0401011-N0402010+ry,e0401020-N0402021+g,e0401030-N0402031+g,e0401030-N0402031+k,e0401030-N0402031+ry,i0301000-N0302001+g,i0301000-N0302001+k,i0301011-N0302010+g,i0301011-N0302010+k,i0401000-N0402001+g,i0401000-N0402001+gy,i0401000-N0402001+i0403001,i0401000-N0402001+k,i0401000-N0402001+ky,i0401000-N0402001+ry,i0401011-N0402010+g,i0401011-N0402010+k,i0401011-N0402010+ry,i0503010-N0504010+g,u0301011-N0302010+i0303010,u0301011-N0302010+k,u0401000-N0402001+g,u0401000-N0402001+k,u0401011-N0402010+g,u0401030-N0402031+i0403031,u0401030-N0402031+ry
ST_N_2_2	e0301000-N0302001+b,e0301011-N0302010+b,e0301011-N0302010+p,e0401000-N0402001+b,e0401000-N0402001+m,e0401000-N0402001+p,e0401000-N0402001+py,e0401011-N0402010+b,e0401011-N0402010+p,i0301011-N0302010+m,i0301011-N0302010+p,i0401000-N0402001+b,i0401000-N0402001+m,i0401000-N0402001+p,i0401011-N0402010+b,i0401030-N0402031+m
...	...
ST_N_2_17	a0201000-N0202001+pau,a0403001-N0404001+pau,a0403020-N0404020+pau
ST_N_2_18	a0302001-N0303001+pau
ST_N_2_19	a0201011-N0202010+pau,a0302010-N0303010+pau,a0403010-N0404010+pau,a0403031-N0404030+pau,a0504020-N0505020+pau

9 おわりに

本研究では, 従来の単語音声認識において, あまり行われてこなかった日本語の同音異義語の音声認識を調査した. 不特定話者において同音異義語を音声認識するために, アクセント情報を用いたモデルを提案し, 単語音声認識実験を行った. そして, 評価データ中に含まれるアクセントの異なる同音異義語に注目した. なお, アクセント情報を音素ラベルに付与すると, 音素数が増加し, 信頼性のある HMM のパラメータ推定は困難である. そこで, 本研究では, 半連続型 HMM と木に基づく状態共有手法を用いた状態共有型 HMM を利用して認識を行い評価した. また, 認識精度向上のために, 前後音素環境情報も利用した. そして, 特徴パラメータに一般的に使われている MFCC は音韻的特徴しか含んでいないため, アクセントを用いた実験において認識精度が低いと予測した. そのため, 韻律的特徴を含む FBANK を用いて認識結果を MFCC と比較し評価した.

不特定話者における実験結果より以下を確認した.

1. 前後環境も考慮したアクセント triphone モデルの MFCC, Full, 半連続型 HMM において 89%の同音異義語音声認識の精度が得られた.

2. 単語音声認識精度においてもアクセント triphone モデル MFCC, Full, 半連続型 HMM の結果が最も高く 94.65%の精度が得られた.

3. 韻律情報が含まれる特徴パラメータである FBANK を用いた精度は MFCC より低いことを確認した.

4. 半連続型 HMM を用いた認識精度は混合分布数を同一にした状態共有型 HMM を用いた認識精度より高いことを確認した.

特定話者における実験結果より以下を確認した.

1. 実験条件が同一の結果では, アクセントモデルの半連続型 HMM の FBANK, Full が同音異義語認識精度と単語認識精度が高かった. そして, 単語音声認識において, 97%の精度が得られた. また, 同音異義語認識において, 96.71%の精度が得られた.

2. 実験条件が同一ではない行った全ての実験結果では, アクセントモデルの半連続型 HMM の FBANK, Full の実験結果とアクセント triphone モデルの状態共有型 HMM, FBANK, Diagonal の同音異義語認識精度が最も高く, 97%の精度が得られた.

3. 実験条件が同一ではない行った全ての実験結果では, アクセント triphone モデルの状態共有型 HMM, FBANK, Diagonal の単語認識精度が最も高く, 98.19%mp 精度が得られた.

4. 韻律情報が含まれる特徴パラメータである FBANK を用いた精度は MFCC より高いことを確認した.

今後, 認識精度を高める手法として FBANK を用いることが考えられる. FBANK は特定話者において MFCC より高い精度が得られる. ゆえに, 話者選択手法や話者適合手法によって不特定話者認識精度の改良を行い特定話者の認識精度に近付けると, FBANK の精度が MFCC より高くなると考えている. また, 状態共有において質問や状態数について評価し, 状態共有型 HMM の効果を確かめる必要がある. 状態共有型 HMM の改善を行うことで, 不特定話者における同音異義語の認識精度が改善できる可能性がある.

謝辞

最後に, ニ年間に渡って御指導, 御教授して頂きました鳥取大学工学部知能情報工学科計算機 C 研究室の池原教授と村上助教授に深くお礼申し上げます. また, 論文を執筆にあたり, 助言を頂いた徳久助手にお礼を申し上げます. 加えて, 本稿を執筆するにあたり参考にさせて頂いた論文, 本の著者の方々の皆様にもお礼申し上げます.

参考文献

[1] 中川 聖一:“確率モデルによる音声認識”, 社団法人 電子情報通信学会,(1988)

[2] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄:“音声認識システム”, オーム社,(2001)

[3] 古井 貞熙:“音声情報処理”, 森北出版株式会社,(1998)

[4] 鹿野 清宏, 中村 哲, 伊勢 史朗:“音声・音情報のディジタル信号処理”, 株式会社 昭昇堂,(1997)

[5] 谷口 勝則, 村上 仁一, 池原 悟:“FBANK を用いた孤立単語音声認識”, 日本音響学会講演論文集,3-Q-3,pp.157-158(2003-3)

[6] NHK 日本語発話アクセント辞典新版. NHK 出版, 1998. ISBN4-14-011112-7.

[7] *HTK Ver3.2 reference manual.* Cambridge University, 2002.

[8] X.D.Huang, Y.Ariki, and M.A.Jack. Hidden markov models for speech recognition. Edinburgh Univ Press, 1990.

[9] S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state tying for high accuracy acoustic modelling. *Proc. ICASSP*, pp. 307–312, 1994.

[10] 高橋, 松永, 嵯峨山. ピッチパタン情報を用いた単語音声認識. 日本音響学会講演論文集, No. 1-3-20, pp. 39–40, 1990.

[11] 村上, 荒木, 池原. 音声におけるポーズ長およびアクセント位置の情報量の考察. 日本音響学会講演論文集, No. 3-3-11, pp. 89–90, 1988.

[12] 堀田, 村上, 池原. モーラ情報およびアクセント位置をもちいた単語音声認識. 日本音響学会講演論文集, No. 3-Q-4, pp. 151–152, 2004.

[13] 堀田, 村上, 池原. モーラ情報を用いた単語音声認識. 日本音響学会講演論文集, No. 1-4-8, pp. 15–16, 2003.

付録

1. 基本モデルの実行シェルスクリプト
   (a) hanrenzoku/normal/stage2(特徴ベクトルの出力)
   (b) hanrenzoku/normal/stage2_5(初期モデルの作成)
   (c) hanrenzoku/normal/stage3(初期学習)
   (d) hanrenzoku/normal/stage5(再推定)
   (e) hanrenzoku/normal/stage7(連結学習)
   (f) hanrenzoku/normal/stage9(連続型 HMM の作成)
   (g) hanrenzoku/normal/stage12(連結学習)
   (h) hanrenzoku/normal/stage12_5(辞書, ネットワークグラマーの作成)
   (i) hanrenzoku/normal/stage13c(認識)
2. 半連続型 HMM の triphone モデルの実行シェルスクリプト
   (a) hanrenzoku/triphone/stage9t(基本モデルから音素のコピー)
   (b) hanrenzoku/triphone/stage12tt(連結学習)
   (c) hanrenzoku/triphone/stage12_5t_2(辞書, ネットワークグラマーの作成)
   (d) hanrenzoku/triphone/stage13tc(認識)
3. 半連続型 HMM のモーラモデルの実行シェルスクリプト
   (a) hanrenzoku/mora/stage9m(基本モデルからモーラモデルへの音素のコピー)
   (b) hanrenzoku/mora/stage12m(モーラモデルの連結学習)
   (c) hanrenzoku/mora/stage12_5m(モーラモデルの辞書, ネットワークグラマーの作成)
   (d) hanrenzoku/mora/stage13(モーラモデルの認識)
4. 半連続型 HMM のアクセントモデルの実行シェルスクリプト
   (a) hanrenzoku/accent/stage9mac(基本モデルから素のコピー)
   (b) hanrenzoku/accent/stage12mac(連結学習)

(c) hanrenzoku/accent/stage12_5mac(辞書, ネットワークグラマーの作成)

(d) hanrenzoku/accent/stage13mac(認識)

5. 半連続型 HMM のモーラ triphone モデルの実行シェルスクリプト

(a) hanrenzoku/moratri/stage9mt_tri(triphone モデルから音素のコピー)

(b) hanrenzoku/moratri/stage12mt(連結学習)

(c) hanrenzoku/moratri/stage12_5mt_2(辞書, ネットワークグラマーの作成)

(d) hanrenzoku/moratri/stage13mtc(認識)

6. 半連続型 HMM のアクセント triphone モデルの実行シェルスクリプト

(a) hanrenzoku/acctri/stage9mat_tri(triphone モデルから音素のコピー)

(b) hanrenzoku/acctri/stage12mat(連結学習)

(c) hanrenzoku/acctri/stage12_5mat_2(辞書, ネットワークグラマーの作成)

(d) hanrenzoku/acctri/stage13matc(認識)

7. 状態共有型 HMM の triphone モデルの実行シェルスクリプト

(a) tree/triphone/stage2_5t(初期モデルの作成)

(b) tree/triphone/stage3t(初期学習)

(c) tree/triphone/stage5t(再推定)

(d) tree/triphone/stage7t(連結学習)

(e) tree/triphone/stage8tc(認識)

(f) tree/triphone/stage9t(状態共有)

(g) tree/triphone/stage12t(連結学習)

(h) tree/triphone/stage13tc(認識)

(i) tree/triphone/stage14t(混合分布数増加)

(j) tree/triphone/stage15t(連結学習)

(k) tree/triphone/stage16tc(認識)

8. 状態共有型 HMM のモーラモデルの実行シェルスクリプト

(a) tree/mora/stage2_5m(初期モデルの作成)

(b) tree/mora/stage3m(初期学習)

(c) tree/mora/stage5m(再推定)

(d) tree/mora/stage7m(連結学習)

(e) tree/mora/stage8mc(認識)

(f) tree/mora/stage9m(状態共有)

(g) tree/mora/stage12m(連結学習)

(h) tree/mora/stage13mc(認識)

(i) tree/mora/stage14m(混合分布数増加)

(j) tree/mora/stage15m(連結学習)

(k) tree/mora/stage16mc(認識)

9. 状態共有型 HMM のアクセントモデルの実行シェルスクリプト

(a) tree/accent/stage2_5mac(初期モデルの作成)

(b) tree/accent/stage3mac(初期学習)

(c) tree/accent/stage5mac(再推定)

(d) tree/accent/stage7mac(連結学習)

(e) tree/accent/stage8macc(認識)

(f) tree/accent/stage9mac(状態共有)

(g) tree/accent/stage12mac(連結学習)

(h) tree/accent/stage13macc(認識)

(i) tree/accent/stage14mac(混合分布数増加)

(j) tree/accent/stage15mac(連結学習)

(k) tree/accent/stage16macc(認識)

10. 状態共有型 HMM のモーラ triphone モデルの実行シェルスクリプト

(a) tree/moratri/stage2_5mt(初期モデルの作成)

(b) tree/moratri/stage3mt(初期学習)

(c) tree/moratri/stage5mt(再推定)

(d) tree/moratri/stage7mt(連結学習)

(e) tree/moratri/stage8mtc(認識)

(f) tree/moratri/stage9mt(状態共有)

(g) tree/moratri/stage12mt(連結学習)

(h) tree/moratri/stage13mtc(認識)

(i) tree/moratri/stage14mt(混合分布数増加)

(j) tree/moratri/stage15mt(連結学習)

(k) tree/moratri/stage16mtc(認識)

11. 状態共有型 HMM のアクセント triphone モデルの実行シェルスクリプト

(a) tree/acctri/stage2_5mat(初期モデルの作成)

(b) tree/acctri/stage3mat(初期学習)

(c) tree/acctri/stage5mat(再推定)

(d) tree/acctri/stage7mat(連結学習)

(e) tree/acctri/stage8matc(認識)

(f) tree/acctri/stage9mat(状態共有)

(g) tree/acctri/stage12mat(連結学習)

(h) tree/acctri/stage13matc(認識)

(i) tree/acctri/stage14mat(混合分布数増加)

(j) tree/acctri/stage15mat(連結学習)

(k) tree/acctri/stage16matc(認識)

12. HTK コンフィグファイル

(a) MFCC で用いた HTK コンフィグファイル-MFCC_conf

(b) FBANK で用いた HTK コンフィグファイル-FBANK_conf

(c) triphoneモデルにおける木に基づく状態共有の質問スクリプト-triTiedState_tree.hed

(d) モーラモデルにおける木に基づく状態共有の質問スクリプト-mqTiedState_tree.hed

(e) アクセントモデルにおける木に基づく状態共有の質問スクリプト-macTiedState_tree.hed

(f) モーラ triphoneモデルにおける木に基づく状態共有の質問スクリプト-mqTiedState_tree.he

(g) アクセント triphone モデルにおける木に基づく状態共有の質問スクリプト-macTiedState_tree.hed

13. 実行シェルスクリプト中のプログラム

(a) make_T_lab2.rb(triphone モデルのラベル作成)

(b) make_M_lab.rb(モーラモデルのラベル作成)

(c) make_MAC_lab.rb(アクセントモデルのラベル作成)

(d) make_MT_lab2.rb(モーラ triphone モデルのラベル作成)

(e) make_MAT_lab2.rb(アクセント triphone モデルのラベル作成)

(f) make_Mbase.rb(基本モデルからアクセントモデルへの音素のコピー)

(g) make_Mbase2.rb(モーラ triphone モデルやアクセント triphone モデルへの音素のコピー)

(h) make_bcplist.rb(音素リストの作成)

(i) make_dic.rb(辞書の作成)

(j) make_net.rb(ネットワークグラマーの作成)

(k) make_telist.rb(評価データリストの作成)

(l) make_trainlist.rb(認識データリストの作成)

14. 実験の誤認識の出力結果 (mau 不特定話者)

06_hMF_normal 半連続型 HMM,Full,MFCC, 基本モデルの出力結果

07_hMF_mac 半連続型 HMM,Full,MFCC, アクセントモデルの出力結果

08_hMF_mat 半連続型 HMM,Full,MFCC, アクセント triphone モデルの出力結果