概要

話者適応は，学習データ量が少ない場合認識精度が向上するとは限らない．また，話者適応に用いる学習データ内に含まれる音素数が，認識精度に与える影響についてはあまり考察されていない．そこで本研究は，学習データに含まれる音素数に着目し，認識精度の低下を防ぐための手法として「混合 HMM」を提案する．

不特定話者の誤り率が 11.17%であるのに対して，164 単語の学習データを用いた実験では，話者適応の誤り率が 13.48%，30 個未満混合 HMM の誤り率が 8.84%となった．82 単語の学習データを用いた実験では，話者適応の誤り率が 33.08%，30 個未満混合 HMM の誤り率が 10.74%となり，混合 HMM の有効性が得られた．

また，話者適応 HMM と混合 HMM の認識精度より，音素数に偏りを持つ学習データを作成した．164 単語の音素数に偏りを持つ学習データを用いた実験では，話者適応の誤り率が 9.14%，30 個未満混合 HMM の誤り率が 8.52%となった．82 単語の音素数に偏りを持つ学習データを用いた実験では，話者適応の誤り率が，13.47%，30 個未満混合 HMM の誤り率が 9.84%となり，通常の学習データより認識精度が向上することを確認した．

目 次

図 目 次

表 目 次

1　はじめに

現在，不特定話者音声認識には複数話者の音声を 1 つの HMM に学習する手法 [1] や，複数の話者を選択的に用いる話者選択型 [2] などの手法がある．しかし，不特定話者の認識精度では不十分である．そこで認識精度を向上させる手法として，認識する話者のデータを利用する話者適応が挙げられる．しかし，認識する話者のデータを大量に収集することは困難であり，限られたデータでより効果的に話者適応を行う必要がある．

話者適応にはすでに様々な学習方法が提案されているが [3] [4]，話者適応に用いる学習データが少ない場合，認識精度が向上するとは限らない．また，話者適応に用いる学習データ内に含まれる音素数が，認識精度に与える影響についてはあまり考察されていない．

そこで本研究は，学習データ内の各音素の数に着目し，各音素の数による認識精度の変化を調査すると共に，認識精度を低下させずに話者適応を行う手法として，不特定話者 HMM と話者適応 HMM を組み合わせて作成する「混合 HMM」を用いる手法を提案し，認識実験により評価する．

不特定話者の誤り率が 11.17%であるのに対して，164 単語の学習データを用いた実験では，話者適応の誤り率が 13.48%，30 個未満混合 HMM の誤り率が 8.84%となった．82 単語の学習データを用いた実験では，話者適応の誤り率が 33.08%，30 個未満混合 HMM の誤り率が 10.74%となった．混合 HMM を用いることにより，不特定話者と話者適応より高い認識精度が得られた．

また，話者適応 HMM と混合 HMM の認識精度より，話者適応においてより効果的な学習データについて考察する．あらかじめ音素数の少ない音素を削除し，音素数の多い音素を増やした学習データを作成することで，更に認識精度が向上すると考え，音素数に偏りを持つ学習データを作成し，認識精度の向上を試みる．

164 単語の音素数に偏りを持つ学習データを用いた実験では，話者適応の誤り率が 9.14%，30 個未満混合 HMM の誤り率が 8.52%となった．82 単語の音素数に偏りを持つ学習データを用いた実験では，話者適応の誤り率が，13.47%，30 個未満混合 HMM の誤り率が 9.84%となり，通常の学習データより認識精度が向上することを確認した．

2　音声認識

2.1　音声認識の原理

2.1.1　音声認識の構成

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 1 のように通信理論の問題として，確率モデルを用いて定式化できる．話者が文を考える課程が文発声部で，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別ける．話者による発声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な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 ω 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 $\hat{\omega}$ を出力する．$\hat{\omega}$ は，事後確率 $P(\omega|y)$ が最大になるように推定する．$P(\omega|y)$ 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．

$$P(\hat{\omega}|y) = \max_{\omega} \frac{p(y|\omega)P(\omega)}{P(y)} \tag{1}$$

ここで，$P(y)$ は ω に無関係であるので無視できる．尤度 $P(y|\omega)$ は音響モデルによって得られ，文 ω が発生される事前確率 $\mathrm{P}(\omega)$ は言語モデルによって得られる．したがって音声認識では，音響モデルと言語モデルをいかに作り，$P(y|\omega)$ と $P(\omega)$ を計算するが重要となる．

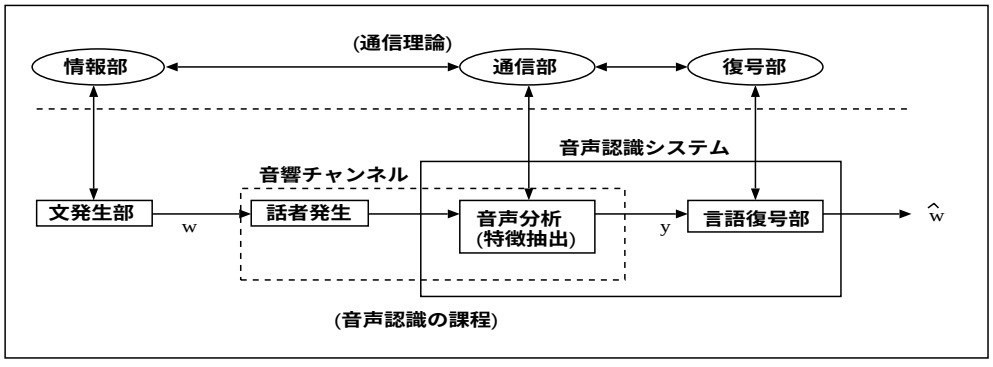

図 1: 音声認識課程の確率モデル

2.1.2 音声認識の分類

1. 対応する話者による分類

   - 特定話者：話者を特定し，学習した話者の音声を認識する．
   - 不特定話者：多数の話者で学習し，様々な話者の音声を認識できる．
   - 話者適応：最初は不特定話者であるが，話者の音声に徐々に対応させていく．

2. 発声単位による分類

   - 孤立単語音声認識：単語ごとに区切って発声した音声を認識する．
   - 連続音声認識：単語を連続して発生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によって，次の 2 つに分類できる．

   - 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 $P(\omega)$ は全て等しいと考えて，$P(y|\omega)$ とともに $P(\omega)$ が認識判定に重要な約割りを果たす．

2.2 音響分析

2.2.1 特徴抽出

音声認識を行うためには，まず，音声区間の検出を行うことが必要である．そして尤度 $P(y|w)$ を計算するには, 音声区間の時系列データ y の表現形式を決める必要がある．音声波形そのものを用いたのでは情報量が多すぎ，波形の位相情報は伝送系や録音系によって変わりやすい上，人間による音声の知覚にはほとんど寄与しないので，位相情報はむしろ取り除いたほうがよい．このため，音声波から一定周期ごとに，短時間スペクトル (密度) を抽出して用いることが多い．現在短時間スペクトル分析の手法としては，帯域フィルタ群を用いる方法，FFT を用いて直接的にスペクトルを計算する方法，相関関数を用いる方法，および LPC 分析を基礎とする方法の 4 種類にわけることができる [6].

2.2.2 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.

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

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.

2.2.3 FBANK

音声認識では，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．本研究ではフィルタバンク分析を用いる．

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．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は以下のようになる．

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{3}$$

3　HMMによる音声認識

3.1　HMMとは

HMM(隠れマルコフモデル，Hidden Marcov Model) とはマルコフモデルの確率的な自由度ををより拡大したモデルである．マフコフモデルとは確率的な生起事象の系列 (課程) を考えたとき，各事象間に関連 (相関) のある場合を考えたことをいう．HMM 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ともに，それに応じてシンボルを確率的に出力すると考える．そのとき，外部からは状態の遷移は直接的に観測できず，出力シンボルのみが観測可能であることから，隠れマルコフモデルと呼ばれる．

3.2　HMMを用いた音声認識

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．

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 $\sum_j a_ij = 1$

- 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：$\pi = \pi$：π は初期状態が s_i である確率，$\sum_j \pi_j = 1$
- 最初状態の状態：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み合わせで表現する場合があり，以下のように分類される.

3.3 HMM の種類

HMM にはスペクトルパターンの表現方法により，離散型 HMM，連続型分布型 HMM に分類される．また，離散型 HMM と連続分布型 HMM の中間的な性質を持った半連続分布型 HMM がある．以下にそれぞれの特徴を示す.

3.3.1 離散分布型 HMM(Descrete HMM)

出現され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．出現確率は，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.

3.3.2 連続分布型 HMM(Continuous 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．

$$b_{ij}(o_t) = \sum_{m=1}^{M_{ij}} C_{ijm} \mathcal{N}(o_t; \mu_{ijm}, \Sigma_{ijm}) \tag{4}$$

$\mathcal{N}(;\mu,\Sigma)$ 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．

$$\mathcal{N}(O;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n \mid \Sigma \mid}} \exp(-\frac{1}{2}(O-\mu)^t \Sigma^{-1}(O-\mu)) \tag{5}$$

3.3.3 半連続分布型 HMM(Semi-continuous 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，

$$b_{ij}(x) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(x) \tag{6}$$

と混合正規分布で表す．ただし，

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{7}$$

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．

3.4 HMM の利点と問題点

HMM が音声認識において有理な点を以下に示す.

- 個人差や調音結合，発声法 (強さ，速さ，明瞭さ) などによる音声パターンの変動を確率モデルで捉え，統計的処理で対処できる.
- 従って，統計理論や情報理論・確率課程論による論理的展開がしやすい.
- 比較的簡単なモデルのパラメータ推定法が知られている.
- 言語レベルの処理も音響処理部と同様に確率モデルで表現でき，両者を統合しやすい.
- 認識時の計算量が比較的少ない.

HMM が音声認識における問題点を以下に示す.

- モデルの設計法が確立されていなく，試行錯誤的，ノウハウ的要素が強い.
- HMM のパラメータ推定に多量の訓練用サンプルを必要とし，計算量も多い.
- 音声の過渡的パターンの表現に乏しく，時系列パターンの中の 2 時点におけるパターンの相関が考慮できない.

3.5 HMM の例 (left-to-right モデル)

HMM には，ある状態から全ての状態に遷移できる全遷移型 (Ergodic) モデルや，状態遷移が一定方向に進む left-to-right モデルがある．音声認識に用いられる HMM は，left-to-right モデルである．left-to-right モデルの例を図 2 に示す．

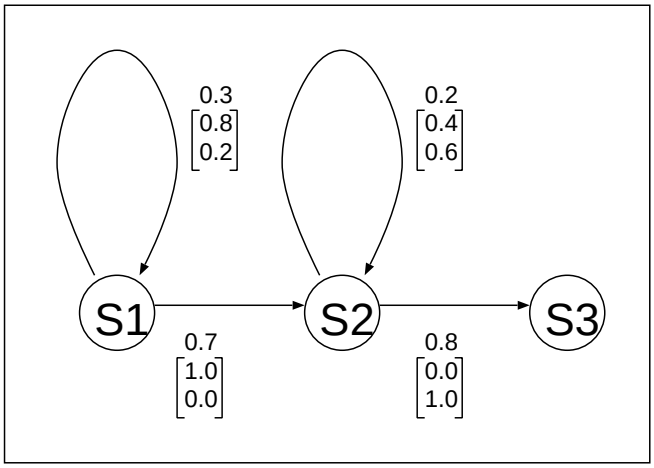

図 2: left-to-right モデルの例

この HMM は 3 つの状態で構成され，2 種類のシンボル a と b からなる．初期状態確率 $\pi_1 = 1.0, \pi_2 = 0, \pi_3 = 0$，最終状態を S_3 とし，図 2 のような遷移のみを行うものとする．a_ij は，状態 S_i から S_j への遷移確率を示し，[ ] 内の数字に上段はラベル a の出力確率，下段はラベル b の出力確率を表す．例として状態 S_1 では，状態 S_1 から状態 S_1 自身に 0.3 の確率で遷移し，遷移の際に 0.8 の確率で a を出力し，0.2 の確率で b を出力する．

出力シンボルが”aab” である場合の状態遷移系列と確率を以下に示す．

- 状態遷移系列 $S_1 - S_1 - S_2 - S_3$

$$0.3 \times 0.8 \quad \times \quad 0.7 \times 0.1 \quad \times \quad 0.8 \times 1.0 = 0.1344 \tag{8}$$

- 状態遷移系列 $S_1 - S_2 - S_2 - S_3$

$$0.7 \times 1.0 \quad \times \quad 0.2 \times 0.4 \quad \times \quad 0.8 \times 1.0 = 0.0448 \tag{9}$$

HMM が”aab” を出力する確率は合計で求まる．

$$0.1344 + 0.0448 = 0.1792 \tag{10}$$

3.6 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)$P(y/M)$ 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 3 に HMM を用いた単語音声認識の方法を示す.

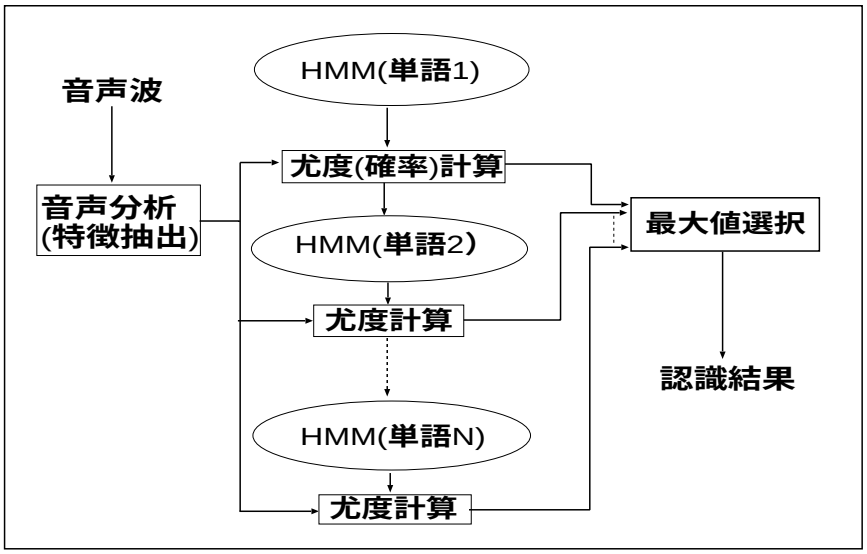

図 3: HMM を用いた単語音声認識の方法

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

$$P(y \mid M) = \sum_{i_0, i_1, \cdots, i_T} P(y \mid q, M) \cdot P(q \mid M) \tag{11}$$

と表すことができる．そして一般的に $P(y \mid M)$ の値は, トレリスアルゴリズムで求められる.

フォワード変数 α (i, t) を定義し, 符号ベクトル y_t を出力して状態 q_t にある確率とすれば, $i = 1, 2, \cdots, S$ とおいて, 以下の式を得る.

$$\alpha(i,t) = \begin{cases} \pi_i & (t = 0) \\ \sum_j \alpha(j, t-1) \cdot \alpha_{ji} \cdot b_{ji}(y_t) & (t = 1, 2, \cdots, T) \end{cases} \tag{12}$$

これを計算し, 最後に以下を求めれば良い.

$$P(y \mid M) = \sum_{i, q \in F} \alpha(i, T) \tag{13}$$

3.6.1 Viterbi アルゴリズム

Viterbi アルゴリズムはモデル λ において最適な状態系列 (最短経路) $S = s_1, s_2, \cdots, s_T$ と，この経路上での確率を求めるアルゴリズムである．

モデル λ において観測系列 $O = o_1, o_2, \ldots, o_T$ に対する最適な状態系列 $S = s_1, s_2, \cdots, s_T$ を求めるために，時刻 t で状態 i に至るまでの最適状態確率 $\delta_t(i)$ を定義する．

$$\delta_t(i) = \max_{s_1, s_2, \cdots, s_t - 1} p(s_1, s_2, \cdots, s_t = i, o_1, o_2, \cdots, o_t | \lambda) \tag{14}$$

時刻 $t+1$ における最適状態の確率は次のように導出できる．

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_{ij}(o_{t+1}) \tag{15}$$

時刻 t 状態 i において生成確率を最大にする経路 (状態遷移) を $\Psi_t(j)$, 最適経路の生成確率を p^*，最適経路上の最終状態を s_T^* とすると最適経路，およびその生成確率は以下の手順で求まる．

1. 初期化

$$\delta_0 = \pi_i \Psi_0(i) = 0 \qquad (1 < i < N) \tag{16}$$

2. 繰返し

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij} b_{ij}(o_t)] \tag{17}$$

$$\Psi_t = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij} b_{ij}] \qquad (1 < t < T), (1 < j < N) \tag{18}$$

3. 最終チェック

$$p^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{19}$$

$$s_T^* = \arg \max_{1 \le i \le N} [\delta_T(i)] \tag{20}$$

4. 経路トレース

$$s_t^* = \Psi_{t+1}(s_{t+1}) \qquad (t = T-1, \cdots, 1) \tag{21}$$

4. で求めた $s_0^*, s_1^*, \cdots, s *_T$ が最適経路となる．Viterbi アルゴリズムは，HMM の初期モデル作成と認識に使用されている．

3.6.2 Baum-Welch アルゴリズム

観測系列の生成確率を最大にするモデル λ のパラメータの局所的最適値を求める方法として，Baum-Welch アルゴリズム (パラメータ再推定法) がある．

モデル λ が観測系列 $O = o_1, o_2, \cdots, o_T$ を生成する場合において，時刻 t で状態 i から状態 j に遷移する確率 $xi_t(i,j)$ を次のように定義する．

$$\begin{aligned} \xi_t(i,j) &= P(s_{t-1} = i, s_t = j|O, \lambda) \\ &= \frac{\alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{P(O|\lambda)} \qquad (1 \leq t \leq T) \end{aligned} \tag{22}$$

ここで，シンボル生成課程で，時刻 t で状態 j にいる確率 $\gamma_t(j)$ を定義する．

$$\begin{aligned} \gamma_t(j) &= P(s_t = j|O, \lambda) \\ &= \sum_{i=1}^{N} \xi_t(i,j) \qquad (1 <\leq t \leq T) \end{aligned} \tag{23}$$

この $\gamma_t(i)$ と $\xi_t(i,j)$ からモデル λ の再推定 ($\lambda \rightarrow \overline{\lambda}$) を次のように行う．

1. 初期状態確率

$$\overline{\pi}_i = \gamma_0(i) = \frac{\alpha_0(i)\beta_0(i)}{P(O|\lambda)} \qquad (1 \leq i \leq N) \tag{24}$$

2. 状態遷移確率

$$\overline{a}_i j = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_{t-1}(i)} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} \tag{25}$$

3. シンボル出力確率

$$\overline{b}_{ij}(O_t) = \frac{\sum_{t \in (o_t = v_k)} \xi_t(i,j)}{\sum_{t=1}^{T} \xi_t(i,j)} = \frac{\sum_{t \in (o_t = v_k)} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)} \tag{26}$$

再推定された $\overline{\lambda}$ の評価は次のようになる．

1. $\overline{\lambda} = \lambda \quad \rightarrow \quad$ (局所的な) 収束状態

2. $P(O|\overline{\lambda}) > P(O|\lambda) \quad \rightarrow \quad$ シンボル系列 O を出力するより最適なモデル λ を推定

Baum-Welch アルゴリズムは，学習データの尤度を最大にするようにパラメータを学習する．本研究では，HMM 初期モデルの再推定に使用されている．

3.7 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，

$$\prod_{n=1}^{N} P(y_1^T(n)|\pi_i, a_{ij}, b_{ij}(k)) \tag{27}$$

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．

$$\beta(i,T) = \begin{cases} 1 & s_i \in F \\ 0 & s_i \notin F \end{cases} \tag{28}$$

$$\beta(i,t) = \sum_j a_{ij} b_{ij}(y_t) \beta(j, t+1) \quad (t = T, T-1, \cdots, 1; i = 1, 2, \cdots, S) \tag{29}$$

$$\gamma(i,j,t) = \frac{\alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{P(y_1^t|M)} \tag{30}$$

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．

$$\hat{\pi_{ij}} = \frac{\sum_j \gamma(i,j,1)}{\sum_i \sum_j \gamma(i,j,1)} \tag{31}$$

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T} \alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{\sum_t \alpha(i,t) \beta(j,t)} = \frac{\sum_t \gamma(i,j,1)}{\sum_t \sum_j \gamma(i,j,t)} \tag{32}$$

$$\hat{b_{ij}} = \frac{\sum_{t,\, y_t = k} \gamma(i,j,t)}{\sum_t \gamma(i,j,t)} \tag{33}$$

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．

3.8 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。

3.8.1 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.

$$\hat{\mu_{ij}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t) y_t}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{34}$$

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

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.

3.8.2 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).

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{36}$$

ここで

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{37}$$

$$\int b_{ijm}(y) dy = 1 \tag{38}$$

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.

$$\hat{\lambda_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{39}$$

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

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

ただし,

$$\gamma(i, j, m, t) = \alpha(i, t-1)a_ij\lambda_{ijm}b_{ijm}(y_t)\beta(j, t) \tag{42}$$

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.

3.8.3 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.

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{43}$$

ここで

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{44}$$

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,

$$\hat{\lambda_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i, j, t, m)}{\sum_{t=1}^{T} \gamma(i, j, t)} \tag{45}$$

$$\hat{\mu_{ijm}} = \frac{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i, j, t, m) y_t}{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i, j, t, m)} \tag{46}$$

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

となる.

3.9 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 4 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．

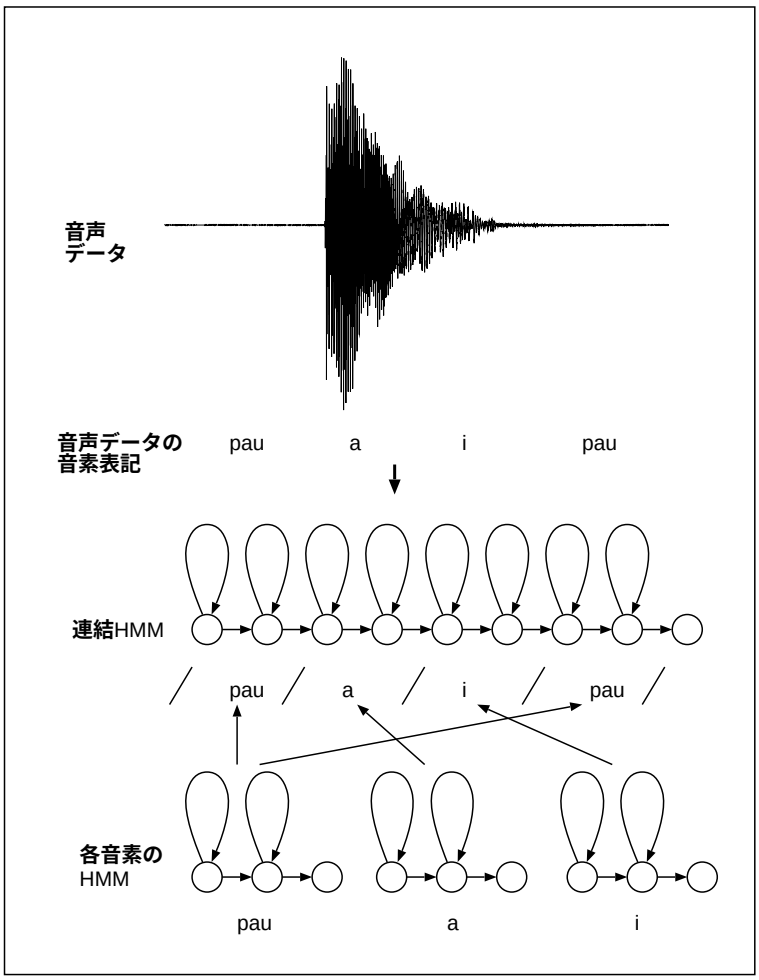

図 4: 連結学習の例

本研究は，話者適応の学習に連結学習を用いる．

4　混合 HMM

混合 HMM は，話者適応における認識精度の低下を防ぐために，話者適応 HMM と不特定話者 HMM を組み合わせて作成する．話者適応において，学習データ内の音素数が少ない音素ほど，話者適応 HMM の信頼性が低いと考えられる．よって，学習データ内の音素数が少ない音素に不特定話者 HMM を用い，音素数が多い音素に話者適応 HMM を用いる．

具体的には，音素数の基準を n 個とした場合，学習データ内の音素数が n 個未満の音素に不特定話者 HMM を用い，n 個以上の音素に話者適応 HMM を用いる．このようにして作成した混合 HMM を「n 個未満混合 HMM」と呼ぶ．

図 5 に，n 個未満混合 HMM の作成の様子を示す．図中の番号は，各 HMM の作成する順番を示している．

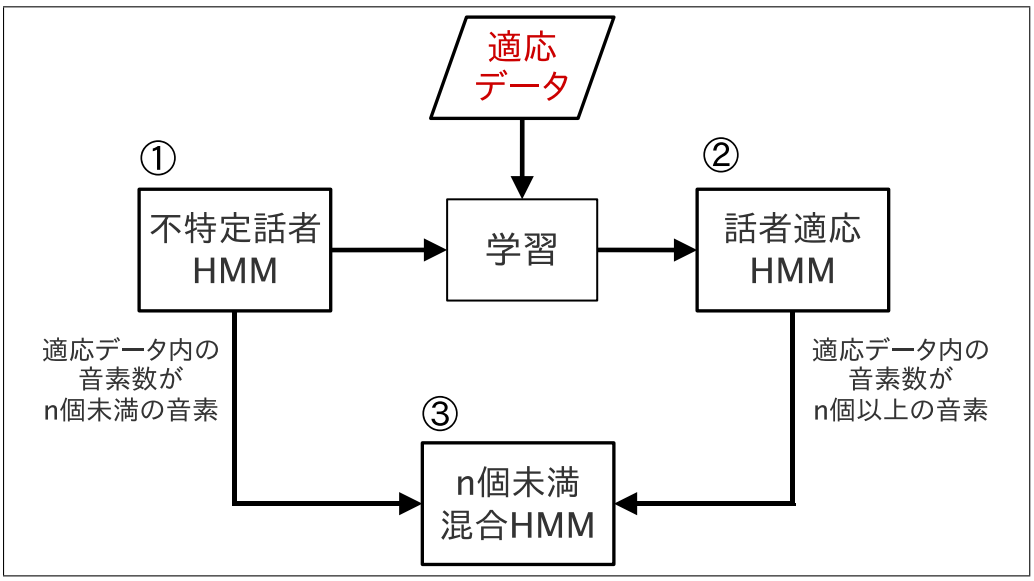

図 5: 混合 HMM の作成手順

4.1 混合 HMM の例

図 6 に 20 個未満混合 HMM の作成の様子を示す．実験に用いた音素 HMM の種類は 34 種類だが，図中には例として 5 つの音素のみ示す．

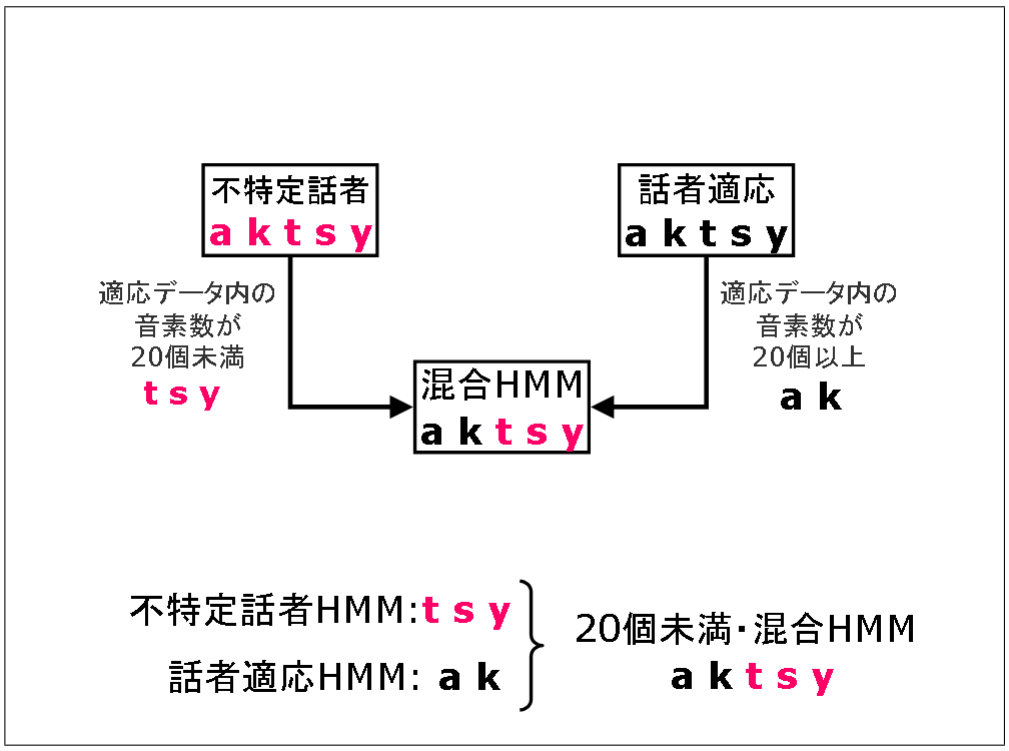

図 6: 20 個未満混合 HMM の作成手順

図中の例では，20 個未満混合 HMM なので，不特定話者 HMM と話者適応 HMM を用いる基準は 20 個となる．図 6 では 20 個未満の音素が “t”，“s”，“y” であり，20 個以上の音素は “a”，“k” とする．よって，“t”，“s”，“y” は不特定話者 HMM を用いて，“a”，“k” は話者適応 HMM を用いる．

その他の音素も同様にして混合 HMM を作成する．

5　評価実験

本研究は，不特定話者 HMM，話者適応 HMM，混合 HMM を用いて単語音声認識を行う．以下にそれぞれの HMM を作成手順を示す．

5.1　音素 HMM の作成

図 7 に話者適応 HMM の作成までの流れを示す．不特定話者 HMM と話者適応 HMM の音素 HMM の作成手順を以下に示す．なお，話者適応には，連結学習を用いて教師あり話者適応を行う．

1. 「初期モデル」を作成する．
2. 「初期モデル」に不特定話者の学習データを用いて Baum-Welch 学習を行い,「不特定話者 HMM」を作成する．
3. 「不特定話者 HMM」に話者適応の学習データを用いて連結学習を行い,「話者適応 HMM」を作成する．

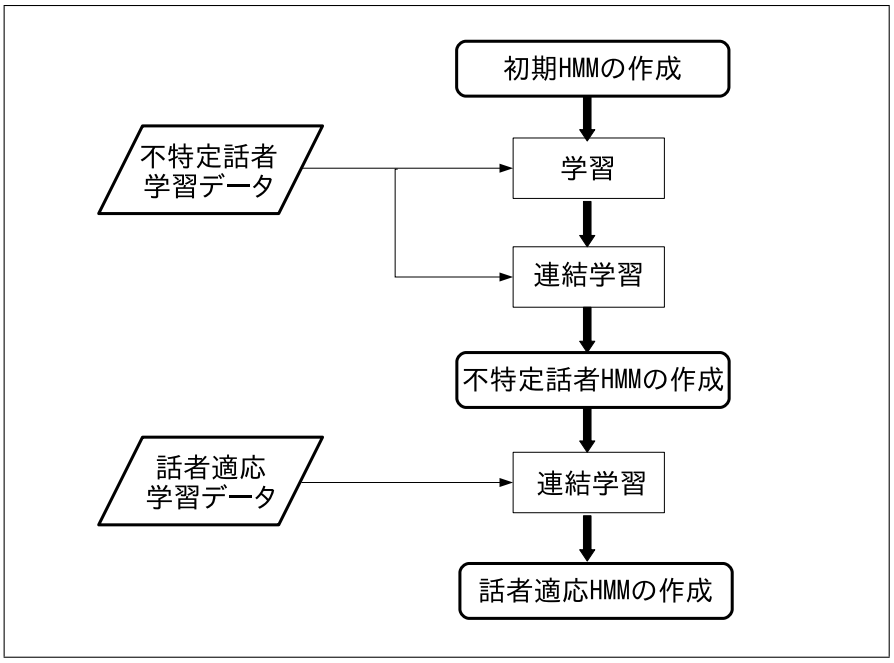

図 7: 話者適応 HMM の作成手順

5.2 混合 HMM の作成条件

本研究の実験は，学習データ内の音素数の変化による認識精度の違いを調査するため，不特定話者 HMM と話者適応 HMM を使い分ける音素数の基準を 10 個，20 個，30 個とする．

5.3 学習データと評価データ

ATR 単語発話データベース Aset(1 話者につき 5,240 単語) の男女各 10 名の音声データを使用する．評価データは，認識する話者の偶数番号データ (2,620 単語) を使用する．不特定話者に用いる学習データは，男女別に，認識する話者以外の 9 話者の奇数番号データ (1 話者につき 2,620 単語) を使用する．話者適応に用いる学習データは，認識する話者の奇数番号から，各音素の出現割合が評価データと同程度になるように選択した 164 単語と 82 単語の 2 種類を用いる．

話者適応に用いる学習データ内に含まれる各音素の範囲別の種類数を表 7 に示す．

表 1: 各範囲に含まれる音素の種類数

	0 個 (種)	10 個未満	10 個以上 20 個未満	20 個以上 30 個未満	30 個以上 40 個未満	40 個以上	合計 (種) n
164 単語	6	8	9	0	3	8	34
82 単語	6	15	6	2	1	4	34

5.4 実験条件

評価実験は，Aset の男女各 10 名のうち男性話者 3 名，女性話者 3 名で行う．本実験では認識に HTK[9] を使用する．実験環境は表 2 にまとめる．

特徴パラメータは MFCC を使用する．HMM の共分散行列には Diagonal-covarience (以下，Diagonal を使用する．HMM は連続型 HMM とする．その他の実験環境は表 2 にまとめる．stream 数は 3 に設定し，MFCC，Δ MFCC，対数パワーと Δ 対数パワーをそれぞれ多次元ガウス分布で表現する．実験でのパラメータの再推定において，データ不足により作成できない音素 HMM が存在した場合，混合分布数を下げて作成する．

表 2: 特徴パラメータの実験条件

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 (連続分布型)
stream 数	3
特徴	12 次 MFCC+ Δ 12 次 MFCC
パラメータ	+対数パワー+ Δ 対数パワー (計 26 次)
連続型 HMM の 初期モデル 混合分布数	(母音・撥音) MFCC 10, Δ MFCC 10, 対数パワー 4, Δ 対数パワー 4 (その他の子音) MFCC 4, Δ MFCC 4, 対数パワー 2, Δ 対数パワー 2

6　実験結果

6.1　不特定話者の実験結果

不特定話者 HMM を用いた単語音声認識の実験結果結果を表 3 に示す．表中の括弧内の分母は認識話者の評価単語数，分子は認識できた単語数を示す．

表 3: 不特定話者 HMM を用いた単語音声認識の実験結果

話者	不特定話者
mau	89.16% (2336/2620)
mmy	89.20% (2337/2620)
mnm	88.17% (2310/2620)
faf	89.31% (2340/2620)
fms	87.18% (2284/2620)
ftk	89.96% (2357/2620)
平均	88.83% (13964/15720)

6.2 話者適応の実験結果

話者適応 HMM を用いた単語音声認識の実験結果を表 4 に示す.

表 4: 話者適応 HMM を用いた単語音声認識の実験結果

	164 単語 話者適応	82 単語 話者適応
mau	86.87% (2276/2620)	64.69% (1695/2620)
mmy	87.63% (2296/2620)	67.29% (1763/2620)
mnm	85.50% (2240/2620)	66.64% (1746/2620)
faf	87.98% (2305/2620)	70.65% (1851/2620)
fms	84.77% (2221/2620)	68.47% (1794/2620)
ftk	86.37% (2263/2620)	63.47% (1670/2620)
平均	86.52% (13601/15720)	66.92% (10519/15720)

実験より以下の結果を得た.

(1) 164 単語の話者適応と 82 単語の話者適応の共に，不特定話者より認識精度が低い.

(2) 82 単語の話者適応は，不特定話者と比較して認識精度が大きく低下している.

6.3 混合 HMM の実験結果

164 単語の話者適応を行い作成した混合 HMM を用いた単語音声認識の実験結果を表 5 に，82 単語の話者適応を行い作成した混合 HMM を用いた単語音声認識の実験結果を表 6 示す.

表 5: 164 単語の混合 HMM を用いた単語音声認識の実験結果

	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM
mau	90.69% (2376/2620)	91.45% (1396/2620)	91.45% (2396/2620)
mmy	89.54% (2346/2620)	90.95% (2383/2620)	90.95% (2383/2620)
mnm	88.97% (2331/2620)	89.20% (2337/2620)	89.20% (2337/2620)
faf	91.30% (2392/2620)	91.95% (2409/2620)	91.95% (2409/2620)
fms	89.73% (2351/2620)	90.61% (2374/2620)	90.61% (2374/2620)
ftk	89.85% (2354/2620)	92.79% (2431/2620)	92.79% (2431/2620)
平均	90.01% (14150/15720)	91.16% (14330/15720)	91.16% (14330/15720)

表 6: 82 単語の混合 HMM を用いた単語音声認識の実験結果

	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM
mau	85.04% (2228/2620)	88.74% (2325/2620)	89.39% (2342/2620)
mmy	86.41% (2264/2620)	87.37% (2289/2620)	88.36% (2315/2620)
mnm	84.69% (2219/2620)	85.73% (2246/2620)	88.13% (2309/2620)
faf	87.86% (2302/2620)	89.05% (2333/2620)	89.16% (2336/2620)
fms	85.53% (2241/2620)	87.14% (2283/2620)	89.24% (2338/2620)
ftk	85.69% (2245/2620)	90.46% (2370/2620)	91.26% (2391/2620)
平均	85.87% (13499/15720)	88.08% (13846/15720)	89.26% (14031/15720)

不特定話者 HMM，話者適応 HMM，混合 HMM を用いた場合の，単語音声認識の 6 話者の平均誤り率を示す．164 単語の学習データの結果を図 8 に，82 単語の学習データの結果を図 9 に示す．

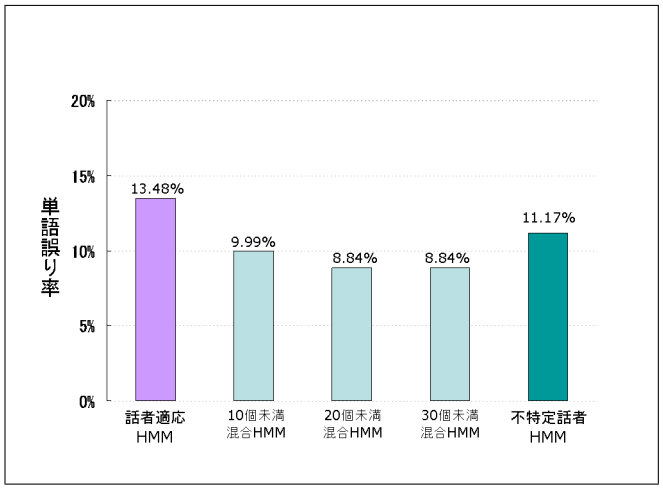

図 8: 164 単語の学習データを用いた実験結果

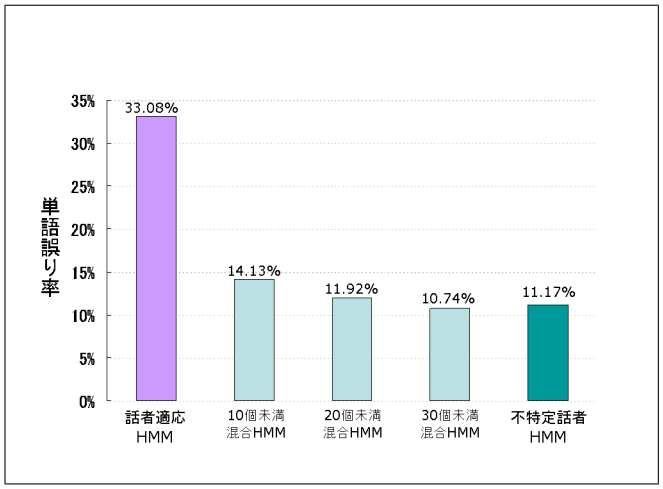

図 9: 82 単語の学習データを用いた実験結果

実験より以下の結果を得た.

(1) 164 単語の学習データでは，全ての混合 HMM において不特定話者と話者適応よりも認識精度が高い.

(2) 82 単語の学習データでは，30 個未満混合 HMM において不特定話者と話者適応よりも認識精度が高い.

(3) 164 単語の学習データと 82 単語の学習データの共に，30 個未満混合 HMM が最も認識精度が高い.

(4) 164 単語の学習データを用いた実験では，学習データ内に 20 個以上 30 未満の範囲内に音素が存在しなかったために，20 個未満と 30 個未満の混合 HMM の認識精度が同じとなった.

82 単語の話者適応の認識精度が大きく低下していることと，30 個未満混合 HMM が最も認識精度が高く，20 個未満，30 個未満となるにつれて認識精度が低下していることから，話者適応において，音素数が少ない音素を多く含む学習データほど，認識精度が低下するといえる.

7　考察

7.1　音素数に偏りを持つ学習データ

混合 HMM の実験結果より，あらかじめ音素数の少ない音素を削除し，音素数の多い音素を増やした学習データを作成することで，更に認識精度が向上すると考えられる．本節では，このように音素数に偏りを持つ学習データの作成を行い，認識精度の評価を行う．

7.1.1　音素数に偏りを持つ学習データの作成

偏りを持つ学習データは，評価データ内の音素の出現率が 2%未満の子音を含む単語を，2%以上の子音を含む単語に選択し直して作成する．通常の学習データと同様に 164 単語の学習データ，82 単語の学習データを作成する．

図 10 に評価データと通常の 164 単語の学習データと音素数に偏りを持つ 164 単語の学習データの音素の出現率を降順にソートした分布を，図 11 に評価データと通常 82 単語の学習データと音素数に偏りを持つ 82 単語の学習データの音素の出現率を降順にソートした分布を示す．

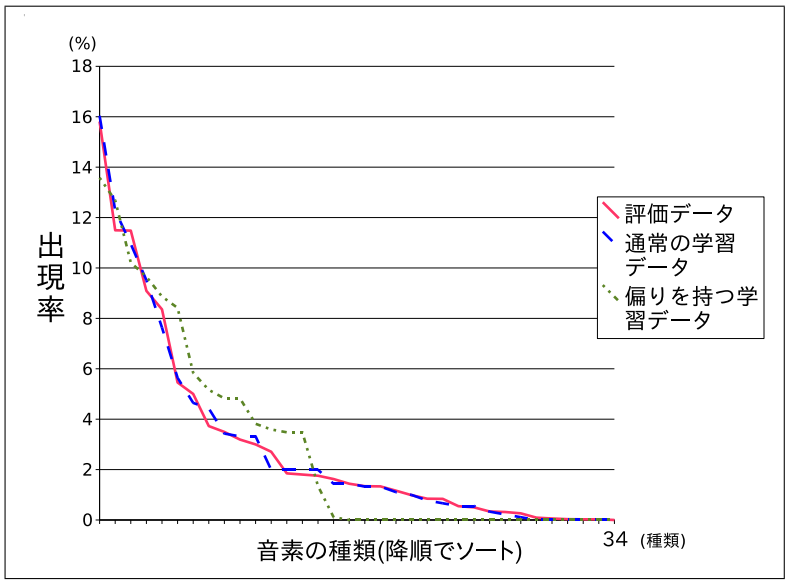

図 10: 164 単語の学習データ内の音素の出現率

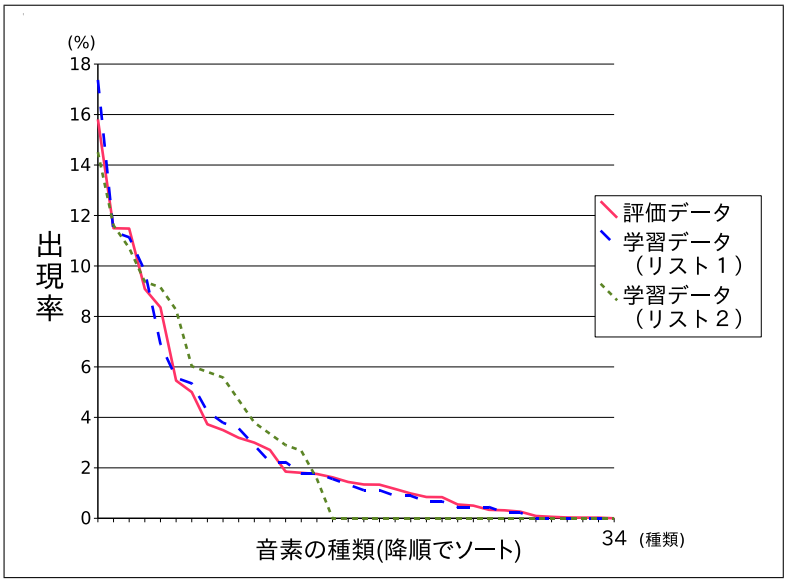

図 11: 82 単語の学習データ内の音素の出現率

- 2%以下の分布
  この範囲に分布されている音素は，音素数の少ない子音である．音素数に偏りを持つ学習データ内にはほぼ存在しない分布となる．

- 2%〜8%の分布
  この範囲に分布されている音素は，音素数の多い子音とである．この範囲では常に偏りを持つ学習データがの出現率が多い．

- 8%以上の分布
  この範囲に分布されている音素は，母音である．母音は条件を付けていないため，2 つの学習データの共に評価データとほぼ同じ分布となる．

作成した音素数に偏りを持つ学習データ内に含まれる各音素の範囲別の種類数を表 7 に示す．

表 7: 各範囲に含まれる音素の種類数

	0 個 (種)	10 個未満	10 個以上 20 個未満	20 個以上 30 個未満	30 個以上 40 個未満	40 個以上	合計 (種) n
164 単語	18	1	1	0	4	10	34
82 単語	19	1	4	4	1	5	34

7.1.2 実験結果

表 8 に，音素数に偏りを持つ学習データを用いて作成した話者適応 HMM の単語音声認識の実験結果を，表 9 に，164 単語の偏りを持つ学習データを用いて作成した混合 HMM を用いた単語音声認識の実験結果を，表 10 に，82 単語の偏りを持つ学習データを用いて作成した混合 HMM を用いた単語音声認識の実験結果を示す.

表 8: 偏りを持つ学習データを用いた話者適応 HMM の実験結果

	164 単語 話者適応	82 単語 話者適応
mau	91.64% (2401/2620)	85.57% (2242/2620)
mmy	91.30% (2392/2620)	87.21% (2285/2620)
mnm	89.85% (2354/2620)	85.76% (2247/2620)
faf	91.72% (2403/2620)	88.44% (2317/2620)
fms	89.05% (2333/2620)	86.03% (2254/2620)
ftk	91.60% (2400/2620)	86.18% (2258/2620)
平均	90.86% (14283/15720)	86.53% (13603/15720)

表 4 と表 8，表 5 と表 9，表 6 と表 10 の比較より，音素数に偏りを持つ学習データを用いることで，話者適応と全ての混合 HMM において認識精度が向上することが確認できた.

本節で作成した音素数に偏りを持つ学習データは，混合 HMM の比較を行うために 30 個未満の音素を含んでいるが，より詳細な条件を用いて作成することで，更に認識精度を向上させることができると考えている.

表 9: 164 単語の偏りを持つ学習データを用いた混合 HMM の実験結果

	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM
mau	91.64% (2401/2620)	92.06% (2412/2620)	92.06% (2412/2620)
mmy	91.30% (2392/2620)	91.11% (2387/2620)	91.11% (2387/2620)
mnm	89.85% (2354/2620)	90.15% (2362/2620)	90.15% (2362/2620)
faf	91.72% (2403/2620)	92.60% (2426/2620)	92.60% (2426/2620)
fms	89.05% (2333/2620)	89.54% (2346/2620)	89.54% (2346/2620)
ftk	91.60% (2400/2620)	93.40% (2447/2620)	93.40% (2447/2620)
平均	90.86% (14283/15720)	91.48% (14380/15720)	91.48% (14380/15720)

表 10: 偏りを持つ 82 単語の学習データを用いた混合 HMM の実験結果

	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM
mau	89.20% (2337/2620)	90.11% (2361/2620)	90.76% (2378/2620)
mmy	88.44% (2317/2620)	89.24% (2338/2620)	89.89% (2355/2620)
mnm	87.98% (2305/2620)	88.40% (2316/2620)	88.09% (2309/2620)
faf	89.92% (2356/2620)	90.95% (2383/2620)	90.76% (2378/2620)
fms	87.82% (2301/2620)	88.63% (2322/2620)	89.35% (2341/2620)
ftk	89.69% (2350/2620)	91.15% (2388/2620)	92.10% (2413/2620)
平均	88.84% (13966/15720)	89.75% (14108/15720)	90.16% (14173/15720)

通常の学習データを用いた実験と音素数に偏りを持つ学習データ実験の 6 話者の平均誤り率を示す．164 単語の学習データの結果を図 12 に，82 単語の学習データの結果を図 13 に示す．

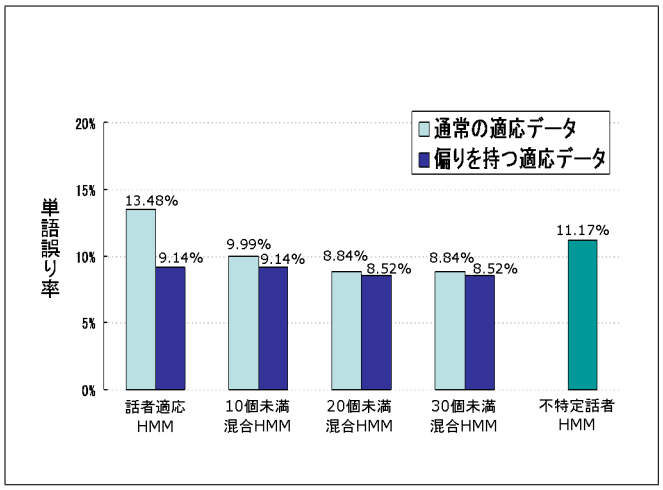

図 12: 164 単語の学習データを用いた実験結果

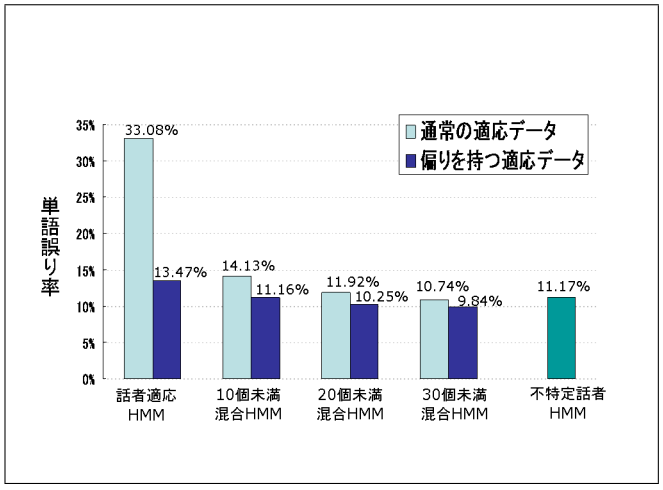

図 13: 82 単語の学習データを用いた実験結果

実験結果より，音素数に偏りを持つ学習データを用いることで，話者適応と全ての混合 HMM において認識精度が向上することが確認できた．

本節で作成した音素数に偏りを持つ学習データは，混合 HMM の比較を行うために 30 個未満の音素を含んでいるが，より詳細な条件を用いて作成することで，更に認識精度を向上させることができると考えている．

7.2 母音と子音による認識精度の違い

本研究で作成した混合 HMM は母音と子音を区別せずに作成している．本節では，母音のみ話者適応 HMM を用いた混合 HMM と，子音のみ話者適応 HMM を用いる混合 HMM を作成し，認識精度の違いを調査する．話者適応に用いる学習データは，7.1 節で作成した音素数に偏りを持つ学習データを使用する．

7.2.1 母音のみ話者適応 HMM の利用

母音の音素数が上位の音素のみ話者適応 HMM を用い，その他の音素は不特定話者 HMM を用いて混合 HMM する．母音のみ話者適応 HMM を用いて作成した混合 HMM の実験結果を表 11 に示す．比較対象として，不特定話者の実験結果を同時に示す．

表 11: 音素数の多い母音のみ話者適応 HMM を用いた混合 HMM の実験結果

		話者適応 HMM を用いた音素		
	不特定話者	上位 2(u a)	上位 4(u a i e)	上位 5(aiueo)
mau	89.16% (2336/2620)	89.58% (2347/2620)	88.89% (2329/2620)	89.24% (2338/2620)
mmy	89.20% (2337/2620)	88.28% (2313/2620)	88.44% (2317/2620)	89.16% (2336/2620)
mnm	88.17% (2310/2620)	87.63% (2296/2620)	87.29% (2287/2620)	87.98% (2305/2620)
faf	89.31% (2340/2620)	87.40% (2290/2620)	88.70% (2324/2620)	90.34% (2367/2620)
fms	87.18% (2284/2620)	88.05% (2307/2620)	87.48% (2292/2620)	88.55% (2320/2620)
ftk	89.96% (2357/2620)	90.15% (2362/2620)	91.26% (2391/2620)	91.30% (2392/2620)
平均	88.83% (/15720)	88.52% (13915/15720)	88.68% (13940/15720)	89.42% (14058/15720)

不特定話者より高い認識精度が得られたのは，母音全て (a i u e o) を入れ換えた場合のみとなった．

7.2.2 子音のみ話者適応 HMM の利用

子音の音素数が上位の音素のみ話者適応 HMM を用い，その他の音素は不特定話者 HMM を用いて混合 HMM する．子音のみ話者適応 HMM を用いて作成した混合 HMM の実験結果を表 12 に示す.

表 12: 音素数の多い子音のみ話者適応 HMM を用いた混合 HMM の実験結果

	話者適応 HMM を用いた音素			
	上位 1(k)	上位 2(k s)	上位 (k s r t)	上位 6(ksrtmg)
mau	90.65% (2375/2620)	90.73% (2377/2620)	90.19% (2363/2620)	89.08% (2334/2620)
mmy	90.15% (2362/2620)	89.47% (2344/2620)	87.67% (2297/2620)	87.33% (2288/2620)
mnm	88.21% (2311/2620)	87.94% (2304/2620)	87.71% (2298/2620)	88.63% (2322/2620)
faf	89.54% (2346/2620)	89.77% (2352/2620)	89.27% (2339/2620)	88.66% (2323/2620)
fms	88.24% (2312/2620)	87.37% (2289/2620)	86.95% (2278/2620)	86.37% (2263/2620)
ftk	90.50% (2371/2620)	90.50% (2371/2620)	89.81% (2353/2620)	89.39% (2342/2620)
平均	89.54% (14077/15720)	89.30% (14037/15720)	88.60% (13928/15720)	88.24% (13872/15720)

不特定話者より高い認識精度が得られたのは，上位 1 位 (k) と上位 2 位 (k s) となった．本研究の学習方法では，学習データ内の音素数が 30 個未満の音素が，認識精度が低下する傾向となっている．ここで用いた子音では “k” のみ 30 個以上なので，話者適応 HMM を用いる子音を増やすほど，認識精度が低下する結果となった.

7.2.3 母音と子音の認識精度の違いについて

表 13 と表 14 に話者適応 HMM を用いた母音と子音の数を示す.

表 13: 母音の音素数

	u	a	i	e	o
音素数 (個)	65	52	48	42	37

表 14: 子音の音素数

	k	s	r	t	m	g
音素数 (個)	41	37	27	26	25	21

母音は子音に比べ学習データ量が多い．しかし，母音は母音全てに話者適応 HMM を用いた場合のみ認識精度が向上している．これに対し子音は，学習データ量の多い音素があれば，1 つの子音のみ話者適応 HMM を使用した場合でも認識精度が向上している．このことから，母音は子音に比べ認識精度の改善が難しいといえる．

本研究では，学習データは 164 単語，82 単語と十分な量の学習データを用いている．しかし，学習データ量を減らした場合，母音全てに十分な音素数が含まれるとは限らない．このような場合，母音と子音で学習方法を変える，子音のみを用いて混合 HMM を作成するなどの対策が必要と考えている．

7.3 特定話者音声認識との比較

特定話者は，一般的に学習データが少なくても比較的高い認識精度が得られる．そこで，話者適応に用いた 164 単語・82 単語を特定話者の学習データとして特定話者 HMM を作成し，混合 HMM の認識精度と比較を行う．

偏りを持つ学習データは含まれる音素の種類が少ないため，実験には 164 単語・82 単語の共に通常の学習データを用いる．特定話者の実験結果を表 15 に示す．参考として 2,620 単語で学習した特定話者の結果を同時に示す．

表 15: 特定話者 HMM の単語音声認識の誤り率

学習データ量	2,620 単語	164 単語	82 単語
mau	95.80% (2510/2620)	84.81% (2222/2620)	65.69% (1721/2620)
mmy	95.27% (2496/2620)	84.96% (2226/2620)	62.79% (1645/2620)
mnm	95.08% (2491/2620)	83.36% (2184/2620)	64.89% (1700/2620)
faf	94.89% (2486/2620)	85.31% (2235/2620)	67.10% (1758/2620)
fms	95.69% (2507/2620)	81.91% (2146/2620)	61.91% (1622/2620)
ftk	95.73% (2508/2620)	83.40% (2185/2620)	65.76% (1723/2620)
平均	95.41% (14998/15720)	83.96% (13198/15720)	64.69% (10169/15720)

結果より，164 単語・82 単語の両方において，混合 HMM の方が認識精度が高い．また，特定話者の認識精度が 164 単語から 82 単語の間で大きく低下しているため，認識する話者の音声が少ないほど話者適応として用いることが有効である．

7.4 追加実験

本研究では，混合 HMM を作成する音素数の基準を $n = 10, 20, 30$ とした．しかし，164 単語の学習データにおいては，30 個以上の音素が多く存在しているため，$n = 40, 50$ として混合 HMM を作成することも可能である．本節では，164 単語の学習データを用いた際の 40 個未満混合 HMM，50 個未満混合 HMM を作成し，認識精度の調査を行った．

通常の 164 単語の学習データを用いた場合の実験結果を表 16 に，偏りを持つ 164 単語の学習データを用いた場合の実験結果を表 17 に示す．

表 16: 164 単語の学習データを用いた混合 HMM の実験結果

	話者適応	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM	40 個未満 混合 HMM	50 個未満 混合 HMM
mau	86.87% (2276/2620)	90.69% (2376/2620)	91.45% (1396/2620)	91.45% (2396/2620)	91.30% (2392/2620)	91.18% (2389/2620)
mmy	87.63% (2296/2620)	89.54% (2346/2620)	90.95% (2383/2620)	90.95% (2383/2620)	89.85% (2354/2620)	90.15% (2362/2620)
mnm	85.50% (2240/2620)	88.97% (2331/2620)	89.20% (2337/2620)	89.20% (2337/2620)	87.63% (2296/2620)	88.70% (2324/2620)
faf	87.98% (2305/2620)	91.30% (2392/2620)	91.95% (2409/2620)	91.95% (2409/2620)	90.80% (2379/2620)	90.73% (2377/2620)
fms	84.77% (2221/2620)	89.73% (2351/2620)	90.61% (2374/2620)	90.61% (2374/2620)	89.39% (2342/2620)	90.34% (2367/2620)
ftk	86.37% (2263/2620)	89.85% (2354/2620)	92.79% (2431/2620)	92.79% (2431/2620)	91.60% (2400/2620)	91.64% (2401/2620)
平均	86.52% (13601/15720)	90.01% (14150/15720)	91.16% (14330/15720)	91.16% (14330/15720)	90.10% (14163/15720)	90.46% (14220/15720)

表 17: 偏りを持つ 164 単語の学習データを用いた混合 HMM の実験結果

	話者適応	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM	40 個未満 混合 HMM	50 個未満 混合 HMM
mau	91.64% (2401/2620)	91.64% (2401/2620)	92.06% (2412/2620)	92.06% (2412/2620)	91.56% (2399/2620)	91.26% (2391/2620)
mmy	91.30% (2392/2620)	91.30% (2392/2620)	91.11% (2387/2620)	91.11% (2387/2620)	89.85% (2354/2620)	87.82% (2301/2620)
mnm	89.85% (2354/2620)	89.85% (2354/2620)	90.15% (2362/2620)	90.15% (2362/2620)	88.85% (2328/2620)	87.75% (2299/2620)
faf	91.72% (2403/2620)	91.72% (2403/2620)	92.60% (2426/2620)	92.60% (2426/2620)	91.91% (2408/2620)	90.11% (2361/2620)
fms	89.05% (2333/2620)	89.05% (2333/2620)	89.54% (2346/2620)	89.54% (2346/2620)	88.40% (2316/2620)	87.98% (2305/2620)
ftk	91.60% (2400/2620)	91.60% (2400/2620)	93.40% (2447/2620)	93.40% (2447/2620)	92.90% (2434/2620)	91.30% (2392/2620)
平均	90.86% (14283/15720)	90.86% (14283/15720)	91.48% (14380/15720)	91.48% (14380/15720)	90.58% (14239/15720)	89.37% (14049/15720)

164 単語の混合 HMM を用いた場合の 6 話者の平均誤り率を図 14 に，偏りを持つ 164 単語の混合 HMM を用いた場合の 6 話者の平均誤り率を図 15 に示す．

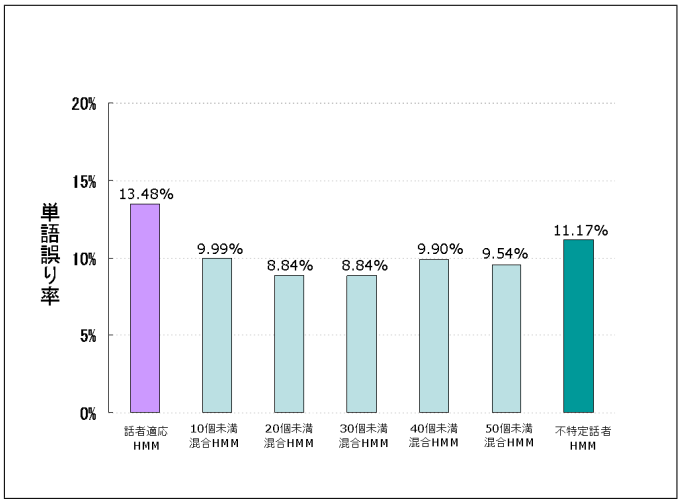

図 14: 164 単語の学習データを用いた実験結果

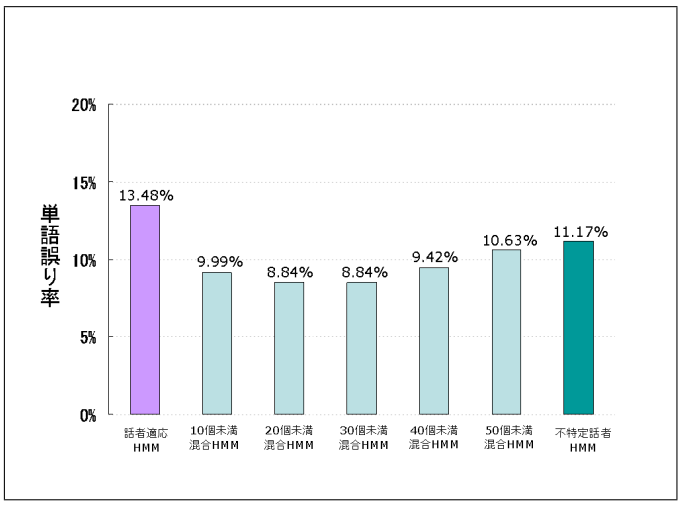

図 15: 偏りを持つ 164 単語の学習データを用いた実験結果

結果より，30 個未満混合 HMM が最も高い認識精度だとわかる．40 個未満・50 個未満混合 HMM を用いた場合，30 個未満混合 HMM より認識精度が低下し不特定話者の認識精度に近付いている．

本研究の条件では 30 個未満混合 HMM(n=30) を用いることで最も高い認識精度が得られるが，この n の最適値は，用いるデータベースや学習方法によって変化すると考えらる．また音素の種類によっても，学習データ量による認識精度の違いがあるため，音素ごとの n の値を調べる必要があると考えている．

8　おわりに

本研究では学習データに含まれる各音素の数に注目し，話者適応 HMM と不特定話者 HMM を組み合わせた混合 HMM を提案し，不特定話者 HMM，話者適応 HMM，混合 HMM を用いて単語音声認識を行い認識精度を調査した．

不特定話者の誤り率が 11.17%であるのに対して，164 単語の学習データを用いた実験では，話者適応の誤り率が 13.48%，30 個未満混合 HMM の誤り率が 8.84%となった．82 単語の学習データを用いた実験では，話者適応の誤り率が 33.08%，30 個未満混合 HMM の誤り率が 10.74%となり，混合 HMM の有効性が得られた．

また，混合 HMM の実験結果を元に，あらかじめ音素数の少ない音素を削除し，音素数の多い音素を増やした学習データを作成することで更に認識精度が向上すると考え，音素数に偏りを持つ学習データを作成した．

164 単語の音素数に偏りを持つ学習データを用いた実験では，話者適応の誤り率が 9.14%，30 個未満混合 HMM の誤り率が 8.52%となった．82 単語の音素数に偏りを持つ学習データを用いた実験では，話者適応の誤り率が，13.47%，30 個未満混合 HMM の誤り率が 9.84%となり，通常の学習データより認識精度が向上することを確認した．

今後の課題として，単語数がより少ない学習データでの認識精度向上と，MLLR や MAP 推定などの話者適応を用いた場合の認識精度の調査が挙げられる．

謝辞

最後に, 一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，村上助教授，徳久助手に深くお礼申し上げます. また，御意見と御助言をいただきました，鳥取大学知能情報工学科知識 A 研究室の清水助教授に深く感謝します.

参考文献

[1] 堀田, 村上, 池原, アクセントを用いた同音異義語の不特定話者音声認識, 電子情報通信学会技術研究報告, SP2005-195, pp. 65-70, (2006).

[2] 松浦, 村上, 池原, 話者選択型音声認識の可能性について, 日本音響学会 2007 年秋期研究発表会, 3-Q-9, pp. 134, (2007).

[3] C.Leggetter and P.Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, Vol.9, pp.171-185, (1995).

[4] G.Zavaliagkos, R.Schwartz and John McDonough, Maximum a posteriori adaptation for large-scale HMM recognizers, Proc. ICASSP-96, pp.725-728, (1995).

[5] 古井　貞熙，音声情報処理，森北出版株式会社

[6] 谷口 勝則, 村上 仁一，池原 悟:“モーラ情報を用いた単語音声認識”，鳥取大学知能情報工学科修士論文，(2002)

[7] S.Young, P.Woodland and G.Evermann, HTK Book, Cambridge University Engineering Department, (2002).

[8] 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)

[9] S.Young, P.Woodland and G.Evermann, HTK Book, Cambridge University Engineering Department, (2002).

[10] 松浦, 村上, 池原, 話者適応における学習データ内の音素数と認識精度の考察, 電子情報通信学会技術研究報告, 音声・思考と言語・福祉情報, SP2007-182, pp.87-91, (2008).

付録

1. 不特定話者の実験結果

2. 通常の学習データを用いた話者適応の実験結果 (164 単語の学習データ，82 単語の学習データ)

3. 偏りを持つ学習データの実験結果
   (164 単語の学習データ，82 単語の学習データ)

4. 通常の学習データを用いた混合 HMM の実験結果
   164 単語 (10〜50 個未満混合 HMM)
   82 単語 (10〜30 個未満混合 HMM)

5. 偏りを持つ学習データを用いた混合 HMM の実験結果
   164 単語 (10〜50 個未満混合 HMM)
   82 単語 (10〜30 個未満混合 HMM)