概要

従来の不特定話者音声認識は，複数の話者の音声を 1 つの HMM に学習している．従来手法の FBANK，Diagonal-covariance を用いた実験の認識率は 85.77%であった．そして，認識率の向上が課題とされている [1].

本研究は，音質が類似する話者の HMM を用いることにより認識率が向上するのではないかと考え,「複数の特定話者の HMM を選択的に用いる」という話者選択型の不特定話者音声認識を行った．特定の単語を発話し，認識率から話者選択を行う教師ありの話者選択と，任意の単語を発話し，尤度から話者選択を行う教師なしの話者選択の，2 つの方法で認識した.

実験の結果，教師ありの話者選択，Diagonal-covariance の場合に，男女 20 話者平均で 79.21%という認識率が得られた．しかし，本提案の手法は従来の手法と比較すると認識率は低い結果となった．このことから音質が類似する話者を選択するよりも，学習データ量が多いほうが有効であると考えられる．また，教師ありと教師なしの話者選択を比較したところ，全ての条件において教師なしが低くなったが，差は 2%程度となった．教師ありと教師なしの話者選択では大きな差はないとわかった.

目 次

図 目 次

表 目 次

1　はじめに

従来の不特定話者音声認識は，複数の話者の音声を 1 つの HMM に学習し，様々な話者の音声を認識できるようになっている．先行研究において，特徴パラメータに FBANK，共分散行列に Diagonal-covariance を用いた不特定話者音声認識の認識率は 85.77%であった．そして，認識率の向上が課題とされている [1]．本研究は，音質が類似する話者の HMM を用いることにより認識率が向上するのではないかと考え，「複数の特定話者の HMM を選択的に用いる」という話者選択型の不特定話者音声認識を試みる．基礎研究として，特徴パラメータに FBANK を使い従来の手法と比較する．

また，話者選択型不特定話者認識では，話者を選択するパラメータとして認識率と尤度があるので，2 つの手法が挙げられる．規定の単語を発話し，認識率から話者選択を行う教師ありの話者選択と，任意の単語を発話し，尤度から話者選択を行う教師なしの話者選択である．教師ありの話者選択と，教師なしの話者選択の精度の比較も行う．

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．

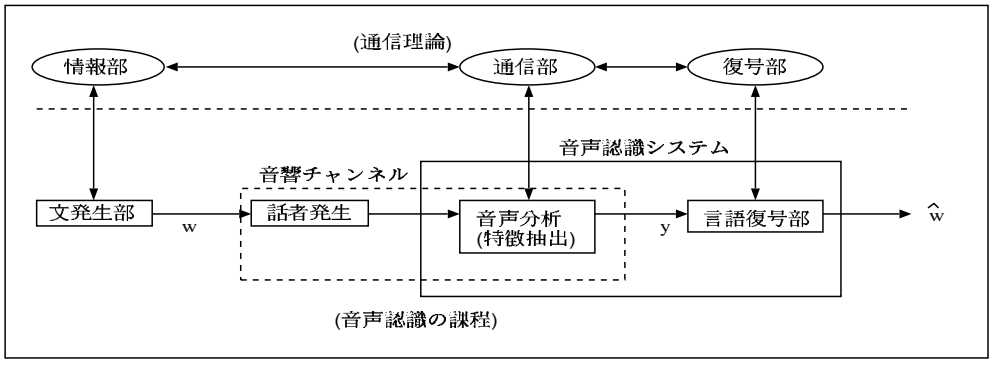

図 1: 音声認識課程の確率モデル

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 [4].

2.2.2 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．

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2}$$

2.3 HMM

2.3.1 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．

2.3.2 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.

- 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.

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

- 半連続分布型 HMM(Semi-continuous HMM)
  連続分布モデルと離散分布モデルの中間の性質を持つ．これは，連続分布モデルにおける混合ガウス分布を，すべてのモデルのすべての状態で共通にし，各分布の重みだけを変えるようにしたものである．結び混合分布モデル (tied-mixture model) とも呼ばれる．また，離散分布モデルにおける各符号に確率分布を持たせたものということもできる.

2.3.3 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.

2.3.4 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．この HMM は 3 つの状

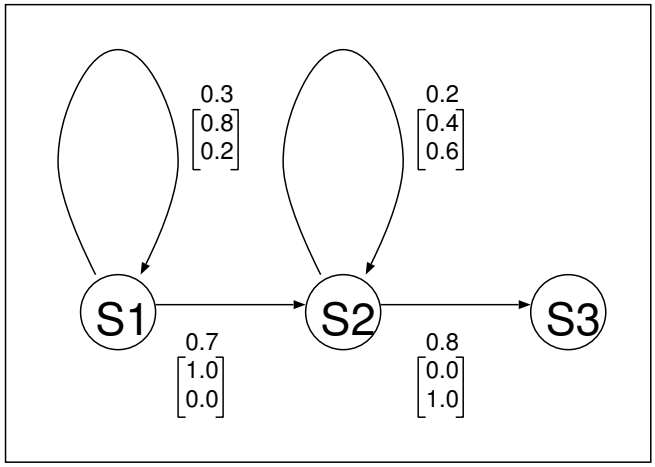

図 2: left-to-right モデルの例

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??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 $S_1 - S_1 - S_2 - S_3$

$$0.3 \times 0.8 \quad \times \quad 0.7 \times 0.1 \quad \times \quad 0.8 \times 1.0 = 0.1344 \tag{3}$$

- 状態遷移系列 $S_1 - S_2 - S_2 - S_3$

$$0.7 \times 1.0 \quad \times \quad 0.2 \times 0.4 \quad \times \quad 0.8 \times 1.0 = 0.0448 \tag{4}$$

HMM が”aab” を出力する確率は合計で求まる．

$$0.1344 + 0.0448 = 0.1792 \tag{5}$$

2.4 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.

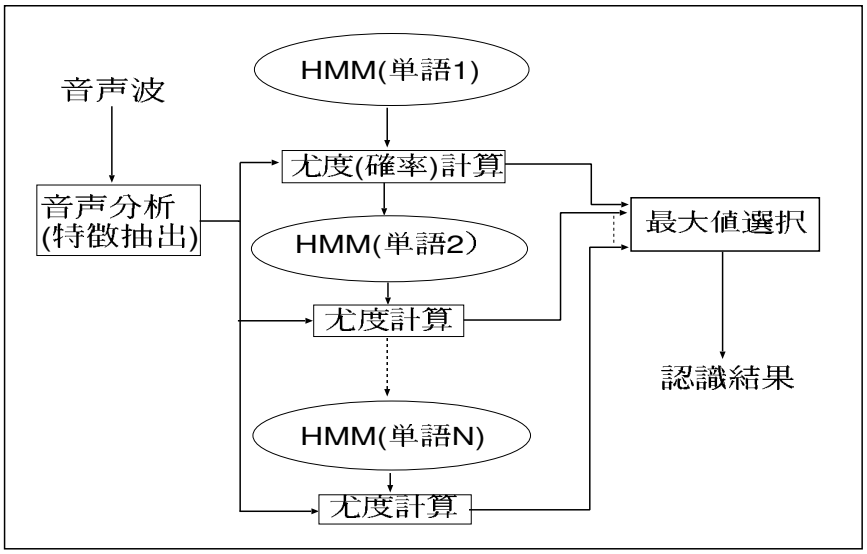

図 3: HMM を用いた単語音声認識の方法

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

$$P(y \mid M) = \sum_{i_0, i_1, \cdots, i_T} P(y \mid q, M) \cdot P(q \mid M) \tag{6}$$

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.

$$\alpha(i,t) = \begin{cases} \pi_i & (t = 0) \\ \sum_j \alpha(j, t-1) \cdot \alpha_{ji} \cdot b_{ji}(y_t) & (t = 1, 2, \cdots, T) \end{cases} \tag{7}$$

これを計算し, 最後に以下を求めれば良い.

$$P(y \mid M) = \sum_{i, q \in F} \alpha(i, T) \tag{8}$$

2.4.1 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．

$$\delta_t(i) = \max_{s_1, s_2, \cdots, s_t - 1} p(s_1, s_2, \cdots, s_t = i, o_1, o_2, \cdots, o_t | \lambda) \tag{9}$$

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．

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_{ij}(o_{t+1}) \tag{10}$$

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

$$\delta_0 = \pi_i \Psi_0(i) = 0 \qquad (1 < i < N) \tag{11}$$

2. 繰返し

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij} b_{ij}(o_t)] \tag{12}$$

$$\Psi_t = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij} b_{ij}] \qquad (1 < t < T), (1 < j < N) \tag{13}$$

3. 最終チェック

$$p^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{14}$$

$$s_T^* = \arg \max_{1 \le i \le N} [\delta_T(i)] \tag{15}$$

4. 経路トレース

$$s_t^* = \Psi_{t+1}(s_{t+1}) \qquad (t = T-1, \cdots, 1) \tag{16}$$

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．

2.4.2 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．

$$\begin{aligned} \xi_t(i,j) &= P(s_{t-1} = i, s_t = j|O, \lambda) \\ &= \frac{\alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{P(O|\lambda)} \qquad (1 \leq t \leq T) \end{aligned} \tag{17}$$

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．

$$\begin{aligned} \gamma_t(j) &= P(s_t = j|O, \lambda) \\ &= \sum_{i=1}^{N} \xi_t(i,j) \qquad (1 <\leq t \leq T) \end{aligned} \tag{18}$$

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

$$\overline{\pi}_i = \gamma_0(i) = \frac{\alpha_0(i)\beta_0(i)}{P(O|\lambda)} \qquad (1 \leq i \leq N) \tag{19}$$

2. 状態遷移確率

$$\overline{a}_i j = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_{t-1}(i)} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} \tag{20}$$

3. シンボル出力確率

$$\overline{b}_{ij}(O_t) = \frac{\sum_{t \in (o_t = v_k)} \xi_t(i,j)}{\sum_{t=1}^{T} \xi_t(i,j)} = \frac{\sum_{t \in (o_t = v_k)} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)} \tag{21}$$

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　不特定話者音声認識について

3.1　話者選択型不特定話者音声認識

話者選択型不特定話者音声認識では，あらかじめ特定話者と同様に各話者に対するHMM を作成する．認識する際は，話者に適切な HMM を選択する話者選択を行い，選択した HMM を用いて認識する．話者を選択するパラメータとして認識率と尤度が挙げられるので，話者選択は 2 つの方法が考えられる．

- 教師ありの話者選択：特定の単語を発話し，認識率で HMM を選択する方法
- 教師なしの話者選択：任意の単語を発話し，尤度で HMM を選択する方法

3.1.1　教師ありの話者選択

教師ありの話者選択の実験の流れを図 4 に示す．具体的な実験方法を以下に示す．

1. 認識する話者以外の 9 名の話者の音声から各 HMM を作成する．
2. 認識する話者の奇数番号の音声を各話者の HMM を利用して認識する．
3. 2 において，認識率が最も高い話者を選択する．
4. 選択した話者の HMM を利用して偶数番号の音声を認識し，認識率を求める．

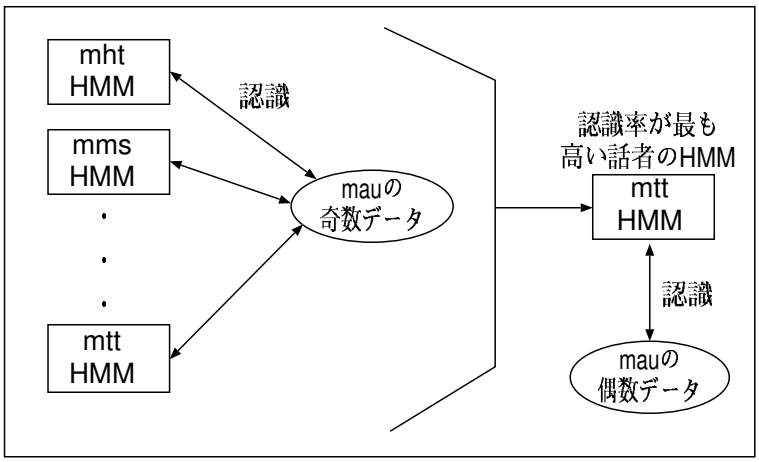

図 4: 教師ありの話者選択の実験の流れ

3.1.2 教師なしの話者選択

教師ありの話者選択の実験の流れを図 5 に示す．具体的な実験方法を以下に示す．

1. 認識する話者以外の 9 名の話者の音声から各 HMM を作成する．
2. 奇数番号の全ての単語認識における尤度の総和を求める．
3. 尤度の総和が最も高い話者を選択する．
4. 選択した話者の HMM を使用して偶数番号の音声を認識し，認識率を求める．

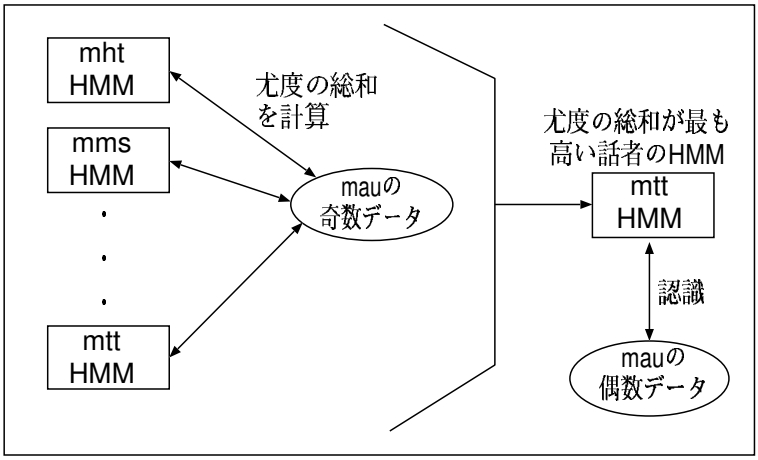

図 5: 教師なしの話者選択の実験の流れ

3.2 従来手法の不特定話者音声認識

従来の不特定話者認識は，複数の話者の音声を 1 つの HMM に学習する．認識するまでの流れを図 6 に示す．具体的な手順を以下に示す.

1. 先行研究では，話者を 10 名用意し，認識する話者以外の 9 名の話者の音声から 1 つの HMM を作成する.

2. そして，作成した HMM を使用して認識する話者の偶数番号の音声を認識し，認識率を求める.

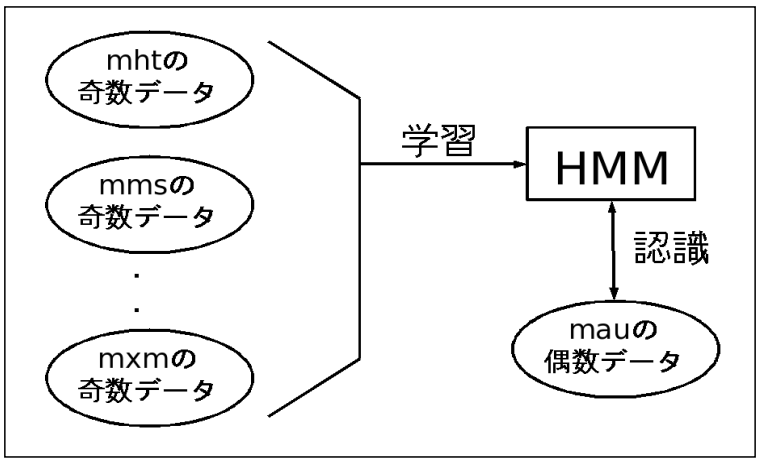

図 6: 従来の不特定話者音声認識の流れ

4　評価実験

4.1　学習データと評価データ

本研究では，データベースとして ATR 単語発話データベース Aset(1 話者 5260 単語) を使用する．男性話者 10 名，女性話者 10 名の計 20 話者を使用する．そして，Aset のデータを奇数番号と偶数番号に別け，奇数番号のデータを学習データ，偶数番号のデータを評価データとする．

4.2　実験条件

本実験では認識に HTK[6] を使用し，実験環境は表 1 にまとめる．HMM の共分散行列には Diagonal-covarience 及び，Full-covariance(以下，Diagonal，Full) の 2 種類を使用する．stream 数は 3 に設定し，FBANK，Δ FBANK，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．

Full の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，混合分布数が FBANK 2，Δ FBANK 2，対数パワー，Δ 対数パワー 1 で作成できない音素は FBANK 1，Δ FBANK 1，対数パワー，Δ 対数パワー 1 にする．この条件にしても作成できない音素 HMM は実験には使用しない．

表 1: 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
FBANK 特徴ベクトル	24 次 FBANK+Δ 24 次 FBANK+ 対数パワー+Δ 対数パワー (計 50 次)
Diagonal-covariance	
連続型 HMM の 初期モデルの 混合分布数	母音・撥音・無音 FBANK 10 Δ FBANK 10 対数パワー，Δ 対数パワー 4
連続型 HMM の 初期モデルの 混合分布数	その他の音素 FBANK 4 Δ FBANK 4 対数パワー，Δ 対数パワー 2
半連続型 HMM の 混合分布数	FBANK 256 Δ FBANK 256 対数パワー，Δ 対数パワー 16
Full-covariance	
連続型 HMM の 初期モデルの 混合分布数	母音・撥音・無音 FBANK 4 Δ FBANK 4 対数パワー，Δ 対数パワー 2
連続型 HMM の 初期モデルの 混合分布数	その他の音素 FBANK 2 Δ FBANK 2 対数パワー，Δ 対数パワー 1
半連続型 HMM の 混合分布数	FBANK 128 Δ FBANK 128 対数パワー，Δ 対数パワー 8

5　実験結果

5.1　教師ありの話者選択の認識精度

5.1.1　認識する話者と各 HMM との認識率

表 2 には，話者を mau とし，各話者の HMM と mau の奇数番号の音声を認識させた認識率を示す．他の話者と各 HMM との認識率は付録に添付する．付録には話者を mau として，mau の HMM の使って認識した特定話者音声認識の認識率も記述してあるが，本研究ではその結果は参考として，直接使うことはない．実験の結果，話者 mtt の HMM

表 2: 話者 mau が場合の各 HMM との認識率

HMM の話者	認識率
mht	75.38% (1975/2620)
mms	76.60% (2007/2620)
mmy	71.76% (1880/2620)
mnm	75.53% (1979/2620)
msh	69.89% (1831/2620)
mtk	67.56% (1770/2620)
mtm	72.40% (1897/2620)
mtt	84.58% (2216/2620)
mxm	69.05% (1809/2620)

を使った場合が最も認識率が高くなっている．よって話者 mau に適した話者として mtt を選択し，mau の偶数番号のデータを認識する．次の節に選択した HMM を使い，認識する話者の偶数番号の音声を認識させた結果を示す.

5.1.2 話者別の教師ありの話者選択の認識精度

Diagonal を使用して行った教師ありの話者選択の認識率と使用した HMM の話者を示す．表 3 に男性話者，表 4 に女性話者を示す.

表 3: 男性話者結果 (教師あり，Diagonal)

認識する話者	認識率	使った HMM
mau	85.15% (2231/2620)	mtt
mht	82.79% (2169/2620)	mtt
mms	78.15% (2057/2620)	mtt
mmy	77.75% (2037/2620)	mtt
mnm	72.86% (1909/2620)	mtm
msh	76.15% (1995/2620)	mms
mtk	75.15% (1969/2620)	msh
mtm	79.73% (2089/2620)	mxm
mtt	83.28% (2182/2620)	mau
mxm	80.27% (2103/2620)	mtm
平均	79.12% (20741/26200)	–

表 4: 女性話者結果 (教師あり，Diagonal)

認識する話者	認識率	使った HMM
faf	78.78% (2064/2620)	fym
ffs	74.95% (1963/2620)	fkn
fkm	72.86% (1909/2620)	fkn
fkn	79.76% (2089/2620)	fym
fks	82.10% (2151/2620)	fms
fms	79.35% (2079/2620)	fks
fsu	77.98% (2043/2620)	fkn
ftk	80.57% (2111/2620)	fyn
fym	84.92% (2225/2620)	fkn
fyn	81.76% (2142/2620)	ftk
平均	79.30% (20776/26200)	–

Full を使用して行った教師ありの話者選択の認識率と使用した HMM の話者を示す．男性話者を表 5，女性話者を表 6 に示す．

表 5: 男性話者結果 (教師あり，Full)

認識する話者	認識率	使った HMM
mau	81.65% (2132/2611)	mtt
mht	79.55% (2077/2611)	mms
mms	78.78% (2057/2611)	mtt
mmy	61.51% (1606/2611)	mtt
mnm	62.47% (1631/2611)	mtm
msh	65.15% (1701/2611)	mms
mtk	66.99% (1749/2611)	msh
mtm	78.70% (2062/2611)	mxm
mtt	77.02% (2011/2611)	mau
mxm	68.21% (1781/2611)	mtm
平均	72.00% (18807/26110)	–

表 6: 女性話者結果 (教師あり，Full)

認識する話者	認識率	使った HMM
faf	73.69% (1924/2611)	fkm
ffs	71.13% (1855/2611)	fsu
fkm	62.81% (1640/2611)	faf
fkn	77.16% (2014/2611)	fym
fks	76.98% (2010/2611)	fms
fms	74.42% (1943/2611)	fks
fsu	75.14% (1962/2611)	ffs
ftk	74.07% (1934/2611)	fyn
fym	83.86% (2177/2611)	fkn
fyn	74.45% (1944/2611)	ftk
平均	74.37% (19403/26110)	–

実験の結果，教師ありの話者選択で最も認識率が高かったのは，共分散行列が Diagonal-covarience の場合の実験で男性話者平均 79.12%，女性話者平均 79.30% となった．

5.2 教師なしの話者選択の認識精度

5.2.1 認識する話者の単語認識における尤度の総和

表 7 には，認識する話者を mau とし，各話者の HMM と mau の奇数番号の音声の単語認識における尤度の総和を示す．他の話者と各 HMM との尤度の総和は付録に添付する．付録には話者を mau として，mau の HMM の使って求めた特定話者音声認識の場合の尤度の総和も記述してあるが，本研究ではその結果は参考として，直接使うことはない．実験の結果，話者 mtt の HMM を使った場合が最も尤度の総和が高くなっている．よって話者 mau に適した話者として mtt を選択し，mau の偶数番号のデータを認識する．次の節に認識する話者と選択した HMM との評価の結果を示す．

表 7: 話者 mau が場合の各 HMM との単語認識における尤度の総和

HMM の話者	尤度の総和
mht	928.6000
mms	1967.4000
mmy	1838.9000
mnm	530.9000
msh	1042.5000
mtk	516.5000
mtm	166.5000
mtt	2439.5000
mxm	110.8000

5.2.2 話者別の教師なしの話者選択の認識精度

Diagonal を使用して行った教師なしの話者選択の認識率と使用した HMM の話者を示す．表 8 に男性話者，表 9 に女性話者を示す.

表 8: 男性話者結果 (教師なし，Diagonal)

認識する話者	認識率	使った HMM
mau	85.15% (2231/2620)	mtt
mht	73.82% (1934/2620)	mtk
mms	77.75% (2037/2620)	msh
mmy	77.75% (2037/2620)	mtt
mnm	64.96% (1702/2620)	mxm
msh	76.15% (1995/2620)	mms
mtk	72.44% (1898/2620)	mms
mtm	79.73% (2089/2620)	mxm
mtt	83.28% (2182/2620)	mau
mxm	80.27% (2103/2620)	mtm
平均	77.13% (20208/26200)	–

表 9: 女性話者結果 (教師なし，Diagonal)

認識する話者	認識率	使った HMM
faf	74.66% (1956/2620)	fms
ffs	71.78% (1880/2620)	fsu
fkm	70.57% (1849/2620)	fsu
fkn	79.76% (2089/2620)	fym
fks	75.92% (1989/2620)	faf
fms	75.73% (1984/2620)	faf
fsu	72.86% (1909/2620)	ffs
ftk	80.57% (2111/2620)	fyn
fym	84.92% (2225/2620)	fkn
fyn	81.76% (2142/2620)	ftk
平均	76.85% (20134/26200)	–

Full を使用して行った教師なしの話者選択の認識率と使用した HMM の話者を示す．男性話者を表 10，女性話者を表 11 に示す．

表 10: 男性話者結果 (教師なし，Full)

認識する話者	認識率	使った HMM
mau	81.65% (2132/2611)	mtt
mht	79.55% (2077/2611)	mms
mms	64.11% (1674/2611)	mht
mmy	56.80% (1483/2611)	msh
mnm	63.35% (1654/2611)	mxm
msh	65.12% (1701/2611)	mms
mtk	58.90% (1538/2611)	mms
mtm	78.70% (2062/2611)	mxm
mtt	77.02% (2011/2611)	mau
mxm	68.21% (1781/2611)	mtm
平均	69.34% (18104/26110)	–

表 11: 女性話者結果 (教師なし，Full)

認識する話者	認識率	使った HMM
faf	70.20% (1833/2611)	fks
ffs	71.13% (1855/2611)	fsu
fkm	51.97% (1357/2611)	fsu
fkn	77.16% (2014/2611)	fym
fks	76.98% (2010/2611)	fms
fms	74.42% (1943/2611)	fks
fsu	75.14% (1962/2611)	ffs
ftk	74.07% (1934/2611)	fyn
fym	83.38% (2177/2611)	fkn
fyn	74.45% (1944/2611)	ftk
平均	72.89% (19029/26110)	–

教師なしの話者選択で最も認識率が高かったのは，共分散行列が Diagonal-covarience の場合の実験で男性話者平均 77.13%，女性話者平均 76.85% となった．

5.3　男女 20 人の平均の認識率

教師ありの話者選択の Diagonal と Full，教師なしの話者選択の Diagonal-covariance と Full-covariance のそれぞれの条件での男女 20 人の平均の認識率を図 7 に示す．教師あり

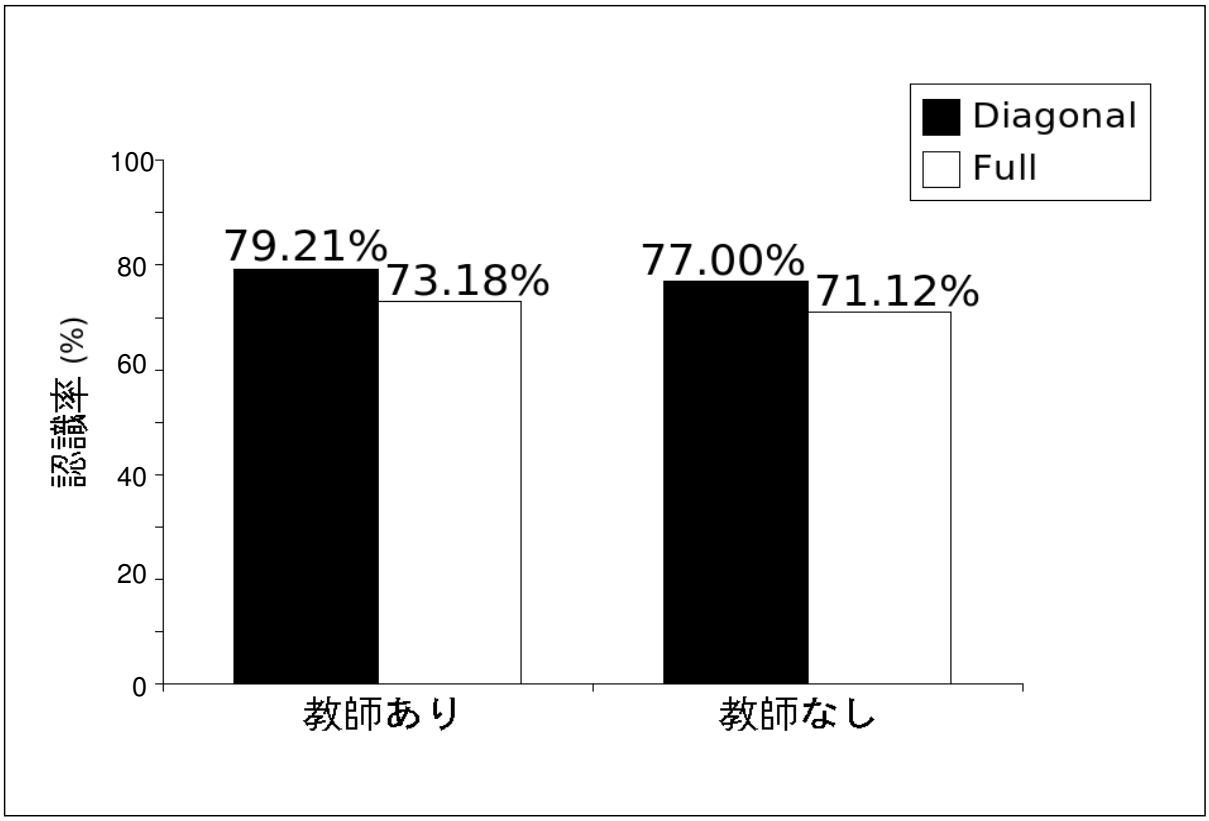

図 7: 男女 20 人の平均の認識率

と教師なしの話者選択のいずれの場合も Diagonal の場合に認識率が高い結果となっており，教師ありの話者選択は 79.21%，教師なしの話者選択は 77.00%となった．話者選択型不特定話者音声認識において Diagonal が有効ということがわかった．また教師ありと教師なしの話者選択を比較したところ，いずれの場合も教師ありの話者選択の方が高く，Diagonal の場合に 2.21%の差であり，Full の場合に 2.06%であった．

6　考察

6.1　従来の手法と比較

実験の結果，男女 20 人の平均の認識率の中で最も高い認識率は，教師ありの話者選択で共分散行列に Diagonal を使った場合において 79.21%となった．従来の手法の Diagonal を使った場合の認識率は 85.77%となっており，本研究の手法の結果は従来の手法より低い結果となった．また，話者ごとに認識率を比較するために，表 12 に話者ごとの従来の手法と本研究の手法の認識率を示す．

表 12: 本研究の手法と従来の手法の認識率 (FBANK，Diagonal)

話者	本研究の手法の 認識率 (教師あり)	従来の手法の 認識率
mau	85.15% (2231/2620)	85.38% (2237/2620)
mmy	77.75% (2037/2620)	85.88% (2250/2620)
mnm	72.86% (1909/2620)	86.11% (2256/2620)
faf	78.78% (2064/2620)	87.79% (2300/2620)
fms	79.35% (2079/2620)	84.69% (2219/2620)
ftk	80.57% (2111/2620)	84.77% (2221/2620)

各話者の認識率を比較しても，本研究の手法が従来の手法より認識率が低い結果となっている．結果より，不特定話者音声認識では話者選択の手法で行った音質の似ている話者の HMM を使って認識するよりも，学習データ量が多い HMM を使って認識する方が有効であるといえる．本研究で行った音質の似ている話者を選択する方法だけでは，従来の手法を超えることは困難と考えられる．よって今後は，話者適応などを組み合わせることにより，認識率を向上させる必要がある．

6.2 教師ありと教師なしの話者選択の比較

教師ありと教師なしの話者選択を比較したところ，いずれの場合も教師ありの話者選択の方が高く，Diagonal の場合に 2.21%の差であり，Full の場合に 2.06%の差だった．これは教師ありと教師なしの話者選択の実験で選択した話者に違いがあり，認識率に影響したためである．表 13，表 14 に Diagonal および Full を使った実験において，教師あり，教師なしの話者選択で選択した話者が違った場合の，選択した話者と認識率を示す．

表 13: 選択した話者の比較 (Diagonal-covariance)

認識する話者	教師ありで使った HMM	教師ありの認識率	教師なしで使った HMM	教師なしの認識率
mht	mtt	83.28%	mtk	73.82%
mms	mtt	78.15%	msh	77.75%
mnm	mtm	72.86%	mxm	64.96%
mtk	msh	76.15%	mms	72.44%
faf	fym	78.78%	fms	74.66%
ffs	fkn	74.96%	fsu	71.78%
fkm	fkn	72.86%	fsu	70.57%
fks	fms	82.10%	faf	75.92%
fms	fks	79.35%	faf	75.92%
fsu	fkm	79.98%	ffs	72.86%

表 14: 選択した話者の比較 (Full-covariance)

認識する話者	教師ありで使った HMM	教師ありの認識率	教師なしで使った HMM	教師なしの認識率
mms	mtt	78.78%	mht	64.11%
mmy	mtt	65.51%	msh	56.80%
mnm	mtm	62.47%	mxm	63.35%
mtk	msh	66.99%	mms	58.90%
faf	fkm	73.69%	fks	70.20%
fkm	faf	62.81%	fsu	51.97%

教師あり，教師なしの話者選択で選択した話者が違う人数は，Diagonal で 20 話者中 10 話者，Full で 20 話者中 6 話者となっている．また，選択した話者が違った場合，教師な

しの話者認識は 16 話者中 15 話者において認識率が下がっている．その差は約 1%から約 13%と幅があるが，話者 20 人の平均を求めると，教師ありと教師なしの話者選択の差は約 2%となる．教師なしの話者選択では適切な話者選択ができない場合があり，教師ありの話者選択と比較して認識率が低くなるが，大きな差ではないといえる．

7　おわりに

本研究では，話者選択型音声話者音声認識の研究を行った．特徴パラメータにFBANKを用いて従来の手法と比較した．また，話者選択において，教師ありの話者選択と教師なしの話者選択の2種類の実験を行い，精度を調査した．

実験の結果，最も認識率が高かったのは教師ありの話者選択，Diagonalの場合に男女20話者平均で79.21%という認識率が得られた．しかし，従来の手法と比較すると低い認識率であり，話者選択の手法では従来の手法より認識率を高くすることは困難であるという結果となった．

また，教師ありと教師なしの話者選択の比較を行ったところ，全ての条件において教師なしの話者選択が低い認識率であったが，差は約2%となっており，大きな差がないという結果となった．

話者選択の手法だけでは認識率の向上が困難であるので，今後の課題として，話者適応の手法を用いて認識率の向上が挙げられる．

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. 音声認識に関して協力して頂いた計算機工学講座博士前期課程 2 年の堀田波星夫さんにお礼を申しあげます.

参考文献

[1] 堀田 波星夫，村上，池原:” 不特定話者における同音異義語音声認識”．日本音響学会春季研究発表会，発表予定，(2006)

[2] 堀田 波星夫，村上，池原:” アクセントを用いた単語音声認識”，鳥取大学知能情報工学科卒業論文，(2004)

[3] 古井　貞熙，音声情報処理，森北出版株式会社

[4] 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)

[5] 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)

[6] Seve Young, et al.:HTK Ver3.2 Reference manual, Cambridge University(2002).

[7] 妹尾 貴宏, 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)

付録

1. 本研究に使用したスクリプトファイル

2. 教師ありの話者選択の実験結果
   各話者の認識率 (Diagonal，Full)

3. 教師なしの話者選択の実験結果
   各話者の尤度 (Diagonal，Full)