HMMを用いた未知、複数信号クラスタリング問題の検討

村上仁一　杉山雅英　渡辺秀行

ATR自動翻訳電話研究所

〒619-02 京都府相楽郡精華町光台2-2

あらまし

入力された信号系列を複数個の信号源に分解する問題、いわゆる N 信号分解問題は、音声情報処理のみならず言語処理などの分野にも応用が予想され極めて範囲が広い問題と言える。本稿では、まず、この問題に対して ergodic HMM をもちいた方法を提案する。ergodic HMM では、状態が信号源に対応し、状態から出力されるシンボル系列が信号系列に相当するため、入力された信号系列を、複数の信号源に分解する segmentation の位置、および分解された segmentation がどの信号源に属しているかを判断する、カテゴリ識別を同時に推定することができる。次に、実験により N 信号源分解問題の応用として、複数話者発話の識別問題を調べた。この結果、erogodic-HMM を用いた方法では、学習アルゴリズムにおける初期パラメータの値が重要であること、および話者性を有する特徴ベクトルとして長時間の LPC ケプストラムが有効であることが示された。

キーワード ●隠れマルコフモデル ●複数信号源クラスタリング ●話者クラスタリング ●全遷移型隠れマルコフモデル

Study of Unknown-Multiple Signal Source Clustering Problem using ergodic HMM

Jin'ichi Murakami　Masahide Sugiyama　Hideyuki Watanabe

ATR Interpreting Telephony Research Laboratories

2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02

Abstract

In this paper, we consider signals that have originated from a sequence of sources. The problem of segmenting the signal and identifying segments to their sources is addressed. This problem has wide applications in many fields. This report describes a resolution method using ergodic Hidden Markov Models (HMM). In this model, each HMM state corresponds to a signal source. The signal source sequence can therefore be determined by using Viterbi decoding over the observation sequence. Baum-Welch training can be used to estimate the HMM parameters from training material. As an application of the multiple signal source indentification problem, an experiment was performed on unknown speaker identification. As a result, a recognition rate of 78% for 4 male speakers was obtained. The results further indicated that the model is sensitive to the initial values of the ergodic HMM and that the long distance LPC cepstrum is an effective way of preprocessing the signal.

key words ●Hidden Markov Model ●Multiple Signal Source ●Speaker Clustering ●ergodic HMM

1 まえがき

未知・複数信号源クラスタリング問題 (N 信号源分割同定問題) は、観測された信号系列から異なる N 個の信号源より生成される区間 (ここではブロックと呼ぶことにする) を検出し (segmentation)、各ブロックが各々どの信号源 (ここではカテゴリと呼ぶことにする) から生成されたのかを分類・識別 (discrimination) する問題である。この問題は、音声情報処理分野に限らず基本的かつ重要な問題である。

未知・複数信号源クラスタリング問題は大きくわけて以下の 4 つの部分問題から構成される。

1. カテゴリを特徴付ける特徴量の問題
2. カテゴリ遷移の segmentation 問題
3. 各ブロックを N 個のカテゴリに識別する問題
4. カテゴリ数 N を推定する問題

これらの 4 つの部分問題を同時に解くことは容易ではないので、いずれかの条件を既知であると仮定して解法を検討する。

前回の報告 [1] ではセグメンテーション位置およびカテゴリ数が既知の場合に、Universal VQ 符号帳およびその出現頻度のヒストグラムの Kullback 情報量による各セグメントの N 個のカテゴリへの識別問題を報告した。本稿では、 2 と 3 の問題、即ち、観測された信号系列において、自動的に segmentation を行ない、segmentation された各区間のカテゴリを識別する問題を扱う。手法として ergodic HMM を利用する。

HMM は、非定常信号源の一つのモデルとして、特に音声認識の分野で広く用いられている [2]。音声認識では、オートマトン制御の下で確率的定常信号源を次々に切替えることにより、音声信号を生成する信号源を表現するモデルとして、HMM が利用されている。このことから N 信号源分解問題では、オートマトン制御の下でカテゴリを次々と切替えることで信号系列を生成するモデルとして、HMM が利用できる。

実験では N 信号源分解問題の応用として複数話者発話の識別問題を検討する。実験から、男性話者 4 名の場合、長時間窓分析を用いた LPC ケプストラムを用いることにより、音声資料により異なるが、フレーム単位で平均、約 70%の識別率が得られることを示す。この値は高い値ではないが、学習アルゴリズムにおける初期パラメータの値によって識別率が 90%を越える音声資料があることから、この初期パラメータを事前にある程度正確に推定しておくことにより、今後さらに識別率が向上する可能性がある。最後に残された課題と、その解決の展望について報告する。

なお、文献 [3, 4] において未知話者クラスタリングに関する類似の検討がなされている。しかしながら、文献で扱われた問題は少数の信号源 (pilot, controllers, noise) を扱っており、特定の信号源 (pilot) の同定 (追跡) に主眼がおかれている。また、通常の話者識別問題と本報告における複数話者発話の識別問題の関係は、通常の話者識別問題では事前に多量の音声を用いて話者モデルを作成し、それとは異なる未知の音声に対して識別を行なう問題であるのに対し、本研究は連続した音声中に含まれる複数の信号源をその音声のみを用いて事前学習することなく分割・分類する問題を扱うため、異なっている。しかしながら、この研究の手法は、話者識別の問題・手法と密接に関連している。

2 N 信号源分解問題

2.1 問題の定式化

図 1 に示すように、n 次元ベクトルで与えられる信号の系列を $\boldsymbol{X} = (\boldsymbol{x}_t)(t = 1, 2, ..., T)$ とする。この系列は、K 個のブロック $\boldsymbol{X}_k$ $(k = 1, 2, ..., K)$ からなり、各ブロックが $N(\leq K)$ 個のカテゴリ $\boldsymbol{C}_j$ $(j = 1, 2, ..., N)$ のいずれかから生じた系列であるとする。ここで、ブロック $\boldsymbol{X}_k$ の構成要素を $(\boldsymbol{x}_{t_{k-1}+1} ... \boldsymbol{x}_{t_k})$ (ただし、$t_0 = 0, t_K = T$) 、継続長を M_k $(= t_k - t_{k-1})$ とする。

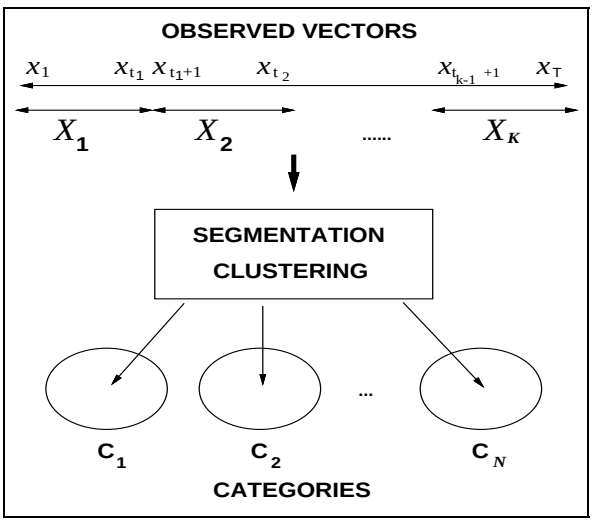

図 1: N 信号源分割同定問題

N 信号源分割同定問題は、与えられたベクトル信号系列に対し、信号源の特性の違いに着目してブロックの切れ目の位置 $t_k(k = 1, 2, ..., K-1)$ を探し (segmentation) この K 個のブロックを $N(\leq K)$ 個のカテゴリに識別しカテゴリ数 N を推定することである。

最も簡単な場合として、カテゴリ数 N 及び segmentation が既知の場合を考える。この場合、N 信号源分割同定は、各ブロック $\boldsymbol{X}_k$ $(k = 1, 2, ..., K)$ を N 個のカテゴリ $\boldsymbol{C}_j$ $(j = 1, 2, ..., N)$ にクラスタリングする問題になる。

2.2 N 信号源分解問題の応用例について

N 信号源分解問題の応用例として、複数話者発話の分解問題が挙げられる。この場合、信号系列は LPC ケプストラム等の音響パラメータに対応し、各カテゴリは各話者に、ブロックの境界は話者の遷移に対応する。また、従来の音素にかわる新しい音韻概念の獲得のための音響単位の自動決定問題 [5] や、話者の言語を識別する問題や、発話様式 (単語・文発声・連続発声等) の識別問題や、音声・非音声の識別問題、さらに言語における品詞ラベルの自動的な作成および付与の問題 [6] などにも応用できる。

3 ergodic HMM を用いた解法

3.1 ergodic HMM

segmentation 未知の場合におけるカテゴリ識別を考える時、この対処の一つとして、HMM (Hidden Markov Model , 隠れマルコフモデル) の適用が考えられる。一般に音声認識では、音声の特性を考慮して、left-to-right 型の HMM が用いられる。しかし、話者認識や言語のモデリングにおいては、全ての状態が全ての状態に接続している ergodic HMM が使用されている [7, 8, 9]。この ergodic HMM を、N 信号源分解問題に利用した時、カテゴリーが状態に相当し、信号系列は状態から出力されるシンボル系列と考えることができる。N 信号源分解問題に ergodic HMM を用いた場合、前述の 2 および 3 の問題は、次の二つの問題のようになる。

1. 信号系列 $\boldsymbol{X}$ の尤度を最大にする HMM のパラメータ $\boldsymbol{M}$ を推定する問題 (モデルの推定、または学習)。なお、パラメータ $\boldsymbol{M}$ は初期状態確率 $\boldsymbol{\pi} = (\pi_i)$、状態遷移確率 $\boldsymbol{A} = (a_{ij})$ 及びシンボル出力確率 $\boldsymbol{B} = (b_j(l))$ で構成される。
2. $\boldsymbol{M}$ が $\boldsymbol{X}$ を出力する時の最も可能性の高い状態遷移系列を推定する問題 (最適状態遷移系列の推定)。

なお、推定された最適状態遷移系列が信号系列の segmentation とカテゴリ識別を同時に与える。
ergodic HMM による解法の概略を、図 2 に示す。

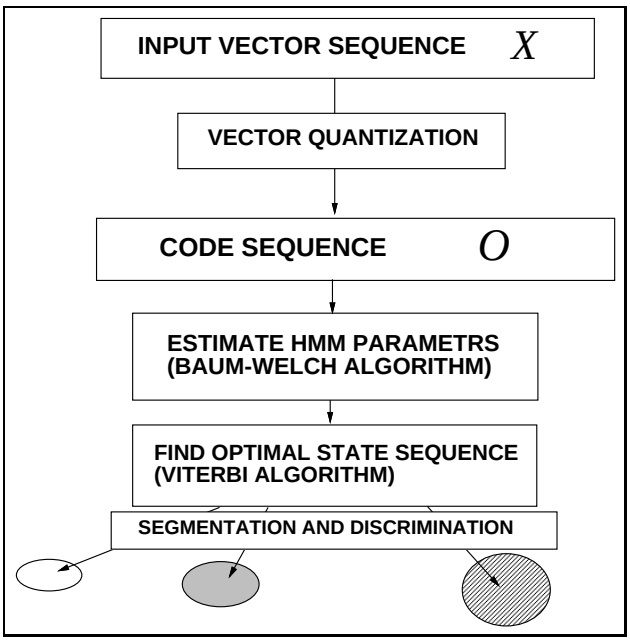

図 2: 離散型 ergodic HMM による解法

なお、ここでは問題の性質上、各状態がシンボルを出力する状態出力タイプ (Moore 型) の離散型 (discrete) HMM を考える。

3.2 HMM のパラメータ推定

HMM のパラメータ $\boldsymbol{M}$ の推定手法として、Baum-Welch のアルゴリズムが知られている。これは、最大尤度推定のための EM アルゴリズム [2] を、$\boldsymbol{M}$ の推定に適用したものであり、推定値は繰り返し計算で求められる。このアルゴリズムを N 信号源分解問題に適用する場合、 2 つの問題点がある。

1. アンダーフロー

   N 信号源分解問題においては、信号系列の長さは、通常の音声認識の場合と比較すると、きわめて長くなるために、Baum-Welch の学習においてアンダフローが起きやすい。この解決方法として、Forward 値の α パラメータを全ての $t \in \{1, 2, ..., T\}$ で $\sum_{i=1}^{N} \alpha_t(i) = 1$ となるようにスケーリング [2] する必要がある。

2. 初期パラメータ

   Baum-Welch のアルゴリズムは、理論的には初期パラメータにより様々な局所的極大点の方向に進む。したがって、所望のカテゴリ識別を行なうためには初期パラメータの設定が重要となる。これについては、実験で考察する。

3.3 最適状態遷移系列の推定について

ここでは、Baum-Welch アルゴリズムで求まったパラメータ推定値から、信号系列を生み出す最適状態遷移系列を推定する問題を考える。この効率的な推定手法として、Viterbi アルゴリズムと Forward アルゴリズムが知られている。このアルゴリズムを次に示す。この両アルゴリズムから計算された状態遷移系列 $\boldsymbol{S}^* = \{s_1^*, ..., s_T^*\}$ から、segmentation とカテゴリ識別が直接得られる。

3.3.1 Viterbi アルゴリズム

推定された HMM のパラメータ $\boldsymbol{M}$ がコード系列 $\boldsymbol{O}$ を出力する時の最も可能性の高い状態遷移系列は、Viterbi アルゴリズムにより効率的に求まる。Viterbi アルゴリズムを次に示す。

1. 全ての $i \in \{1, ..., N\}$ に対し、

$$\delta_1(i) = \log \pi_i + \log b_i(\boldsymbol{o}_1), \phi_1(i) = 0$$

とおく。

2. 時間軸 $t = 2, ..., T$ に沿って、全ての $j \in \{1, ..., N\}$ に対し

$$\delta_t(j) = \max_i [\delta_{t-1}(i) + \log a_{ij} + \log b_j(\boldsymbol{o}_t)],$$

$$\phi_t(j) = \arg\max_i [\delta_{t-1}(i) + \log a_{ij}]$$

3. 状態遷移系列に対する対数尤度及び T 時刻目の最適状態を次式で求める。

$$\max_{\boldsymbol{S}} P(\boldsymbol{O}, \boldsymbol{S} \mid \boldsymbol{M}) = \max_j \delta_T(j),$$

$$s_T^* = \arg\max_j \delta_T(j)$$

4. 時間軸 $t = T-1, ..., 1$ に沿って、次式により状態遷移系列を得る。

$$s_t^* = \phi_{t+1}(s_{t+1}^*)$$

3.3.2 Forward アルゴリズム

Baum-Welch の学習においては、HMM のパラメータ $\boldsymbol{M}$ がコード系列 $\boldsymbol{O}$ を出力する時の尤度は、各状態からの総和で計算される。したがって状態遷移系列を、各時刻における最大の尤度を持つ状態として選択する方法がある。この Forward アルゴリズムは次のようになる。

1. 全ての $i \in \{1, ..., N\}$ に対し、

$$\delta_1(i) = \pi_i \times b_i(\boldsymbol{o}_1)$$

$$s_1^* = \arg\max_i \delta_1(i)$$

2. 時間軸 $t = 2, ..., T$ に沿って、全ての $j \in \{1, ..., N\}$ に対し

$$\delta_t(j) = \sum_i [\delta_{t-1}(i) \times a_{ij} \times b_j(\boldsymbol{o}_t)]$$

$$s_t^* = \arg\max_j \delta_t(j)$$

3. 状態遷移系列に対する尤度及び T 時刻目の最適状態を次式で求める。

$$\max_{\boldsymbol{S}} P(\boldsymbol{O}, \boldsymbol{S} \mid \boldsymbol{M}) = \sum_j \delta_T(j)$$

$$s_T^* = \arg\max_j \delta_T(j)$$

4 複数話者識別実験の実験条件

ここでは、N 信号源分解問題の一応用例として複数話者発話の分解問題を取り上げ、ergodic HMM を用いて識別実験を行なう。

4.1 話者特徴量

本実験では、話者特徴量を表す音響パラメータとして LPC ケプストラム (1 次 ~ 16 次) を用いる。ところで、話者識別においては長時間平均スペクトルが有効であることが知られている [10]。そこで実験では、通常の分析窓 (フレーム窓長 21.3 ms、フレーム更新周期 10.7 ms) で LPC 分析した時と、長時間窓 (フレーム長を 213 ms、フレーム更新周期 107 ms) で分析した時の 2 種類で行なう。その他の音声分析条件を表 1 に示す。

表 1: 音声分析条件

音声特徴量	LPC ケプストラム係数
標本化周波数	12 kHz
LPC 分析	14 次
打ち切り次数	16 次
フレーム長	21.3 ms (256 点)— 213 ms (2560 点)
フレーム更新周期	10.7 ms (128 点)— 107 ms (1280 点)
高域強調	$(1 - 0.97z^{-1})$

4.2 音声資料

実験には、単語音声 (ATR 5240 単語データ) を用いて疑似的な複数話者による対話音声を作成する。話者は男性 4 名で、1 発話 (1 ブロック) につきランダムな単語を 10 個連接することとし、ブロックごとに話者間でランダムに連接して 1 セットの音声資料を作成する。ただし、1 セットにつき 1 話者の音声は 10 ブロックとする。したがって

話者が 4 人の時、1 つの音声資料は 40 ブロックで構成される。さらにこれを 10 セット作成し、評価は、10 回の実験の平均値で行なう。なお、各単語の前後の無音区間は無音スペクトルが話者ごとに異なるため可能性があるため、削除した。音声資料の例を図 3 に示す。

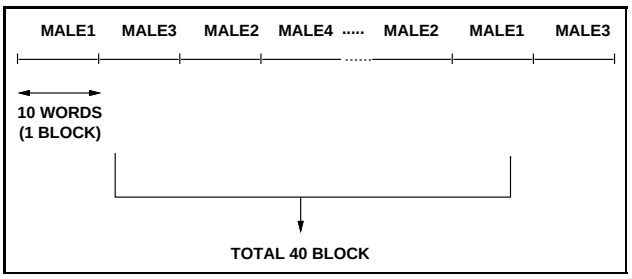

図 3: 疑似音声対話の例 (男性 4 名、各話者 10 単語の発声)

4.3 Universal 符号帳の作成手順

Universal 符号帳は音声資料 $\boldsymbol{X}$ を用いて作成するが、このようにすると音声資料毎にベクトル量子化が必要となるため、ここでは簡単化して $\boldsymbol{X}$ とは独立な評価に用いない単語音声データから作成することにする。1 人あたり 100 単語を用いることにし、話者が 4 人以下の場合は 1 段階で、話者 8 人の場合はフレーム数が多く (約 80000) 計算量が膨大となるため 2 段階で作成する。まず話者毎に、100 単語 (約 40000 フレーム) から 512 個の符号を求め、次に 8 名分の符号 (計 4096 個) から、所定個の符号を求める。実験では符号サイズを $L = 256$ とし、距離尺度として Euclid 距離を用いる。

4.4 識別評価方法

信号信号系列を ergodic HMM によって複数のカテゴリに分類しても、各カテゴリーがどの話者に対応しているのかは同定できない。すなわち、求まる系列はカテゴリーの番号であり、番号と話者の対応には $N!$ 個の組み合せが存在する。そのため以下の式で識別率を算出した。

$$R = \frac{1}{K} \max_{\sigma} \sum_{k=1}^{K} d(\tau(\boldsymbol{X}_k), \sigma(S_k)) \qquad (1)$$

ここで、τ, σ, S_k, d はクラスター番号を与える写像、$(1, 2, \ldots, N)$ の任意の置換、各ブロックの正解カテゴリー番号、値が一致した時のみ 1、それ以外は 0 である関数である。ここでは、各カテゴリー番号を σ で置換し、$N!$ 通りの置換について正解率を算出しその中の最大値を識別率とした。なお、話者が 4 人の場合では $4! = 24$ 通りの組み合せがある。平均識別率 R_{AVE} は各試行の識別率の 10 試行分を平均して算出される。R の算出はフレーム単位で行なわれる。 なお、N が大きくなる時 R の算出に多大の計算量を要するため、分岐限定法などの組み合せ最適化法による高速化が必要である。

5 ergodic HMM による複数話者識別実験の結果

Baum-Welch の学習アルゴリズムは局所的極大点の方向に進むため、初期パラメータによって学習後のパラメータが大きく変化し、そのため、識別率が大きく変動することが予測される。また、HMM において状態の物理的な意味は、初期パラメータに依存すると考えられ、パラメータをランダムに与えた時、話者が状態に対応する場合も考えられるが、例えば母音と子音が状態を表す可能性もある。そこで初期パラメータを 3 種類の方法で与え、識別実験を行なう。

1. 実験 1 : $\boldsymbol{B}^{(0)}$ にランダムな値を与えた場合
   HMM パラメータ $\boldsymbol{M}$ の初期パラメータとして、初期状態確率 $\boldsymbol{\pi}^{(0)}$ と状態遷移確率 $\boldsymbol{A}^{(0)}$ を等確率に、シンボル出力確率 $\boldsymbol{B}^{(0)}$ をランダムに与える。なお、$\boldsymbol{B}^{(0)}$ を等確率に与えた場合、Baum-Welch の学習アルゴリズムにおいて forward および backward の値がすべての状態で同じ値になるため、収束しない。
2. 実験 2 : $\boldsymbol{B}^{(0)}$ に真値を与えた場合
   初期パラメータとして、$\boldsymbol{B}^{(0)}$ は話者ごとに真値を与えるが、$\boldsymbol{\pi}^{(0)}$ と $\boldsymbol{A}^{(0)}$ は等確率にする。
3. 実験 3 : 全パラメータに真値を与えた場合
   初期パラメータとして、$\boldsymbol{\pi}^{(0)}$、$\boldsymbol{A}^{(0)}$ そして $\boldsymbol{B}^{(0)}$ の全てのパラメータに真値に与える。

表 2: Baum-Welch 学習における初期パラメータの値

実験番号	状態初期確率 $\boldsymbol{\pi}^{(0)}$	状態遷移確率 $\boldsymbol{A}^{(0)}$	シンボル出力確率 $\boldsymbol{B}^{(0)}$
実験 1	identical	identical	random
実験 2	identical	identical	true
実験 3	true	true	true

5.1 通常の分析窓長の LPC ケプストラムにおける実験結果

通常の分析窓長 (21.3ms) で LPC ケプストラムを分析した時の実験結果を図 4 に示す。この図において、縦軸が平均識別率で横軸が HMM の学習回数である。○ は実験 1 の、△ は実験 2 の、□ は実験 3 の結果である。また、実線で○ △ □ は Viterbi ア

ルゴリズムによる結果で、破線で● ▲ ■ は Forward アルゴリズムによる平均識別率である。この図から以下のことがわかる。

1. $\boldsymbol{B}^{(0)}$ をランダムにした時 (実験 1) は、平均識別率で 30%から 40%しか得られない。また、この値は学習回数を増加しても向上しない。
2. $\boldsymbol{B}^{(0)}$ のみ真値にした時 (実験 2) は、平均識別率を実験 1 と比較すると大きく向上する。ただし、学習後のパラメータを真値と比較すると初期状態遷移確率 π は真値を持つのに対し、状態遷移確率 $\boldsymbol{A}$ は、が真値に比べ大きくなっていた。これは、推定された HMM では、各時刻において他の状態へ容易に遷移しやすいことを意味する。
3. パラメータの全てを真値にした時 (実験 3) でも平均識別率は最大でも約 70%にしかならず、学習回数を増加させるにしたがい、逆に平均識別率は減少していく。
4. Viterbi アルゴリズムと Forward アルゴリズムでは、Forward アルゴリズムの方が、やや高い平均識別率を得る。

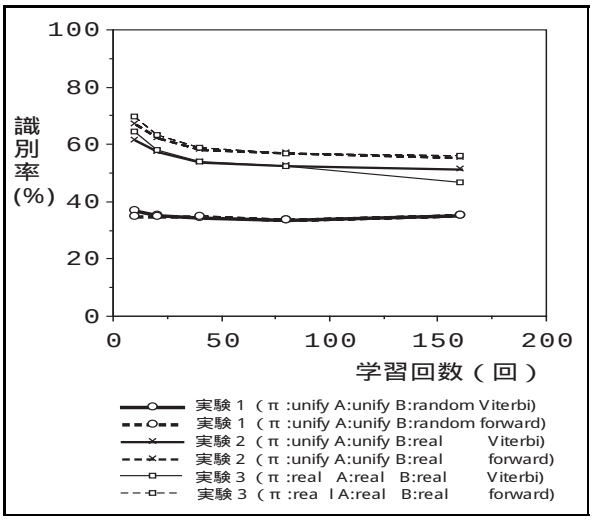

図 4: 通常の分析窓長の LPC ケプストラムを用いた時の学習回数と識別率との関係

5.2 長時間窓 LPC ケプストラムによる実験結果

長時間窓 (213ms) で LPC ケプストラムを計算した場合の実験結果を図 5 に示す。この図から次のことがわかる。

1. $\boldsymbol{B}^{(0)}$ をランダムにした時の実験 (実験 1) では、学習回数が 10 回の場合、平均識別率は約 30%と低いが、学習回数が 160 回の時には 75%近い平均識別率が得られた。この値は通常の分析窓長による実験では 35%であったことから、長時間分析は通常のケプストラムより話者の特徴を良く表現しているパラメータであると予測される。
2. 音声資料 1 セットごとの識別率を図 6 に載せた。実験条件は長時間窓分析で Viterbi アルゴリズムで状態遷移を計算したものである。横軸は学習の繰り返しの回数で、縦軸は識別率である。これから、学習回数が増加するにしたがい、平均の識別率は上昇するが、同じ学習回数でも、音声資料ごとの識別率は大きく変化していることがわかる。例えば、学習回数が 160 回のとき、平均識別率は約 75.0%であるが、個々の識別率は 57.5%から 92.1%と、大きな広がりを持つ。これは、初期パラメータによって個々の音声資料の識別率が大きく変化するためと思われる。なお、学習回数を 1280 回まで増加させた時、平均識別率は 80.4%まで上昇した。
3. $\boldsymbol{B}^{(0)}$ のみ真値にした時 (実験 2)、実験 1 より、かなり高い平均識別率 95%を得た。また、学習回数を重ねるごとに平均識別率はすこしづつ上昇する。
4. 初期パラメータの全てを真値にした実験 (実験 3) では、学習回数を増加させても識別性能に全く変化がなく、約 96% から 97%であった。この値は、$\boldsymbol{B}^{(0)}$ に正しい値を与えて学習させた時 (実験 2) の識別性能より僅かに高いが、差はほとんどない。したがって、識別性能は特に初期パラメータの状態シンボル出力確率の値に大きく依存すると考えられる。
5. Viterbi アルゴリズムと Forward アルゴリズムでは、Forward アルゴリズムの方が、わずかに高い平均識別率を得るが、あまり差がない。

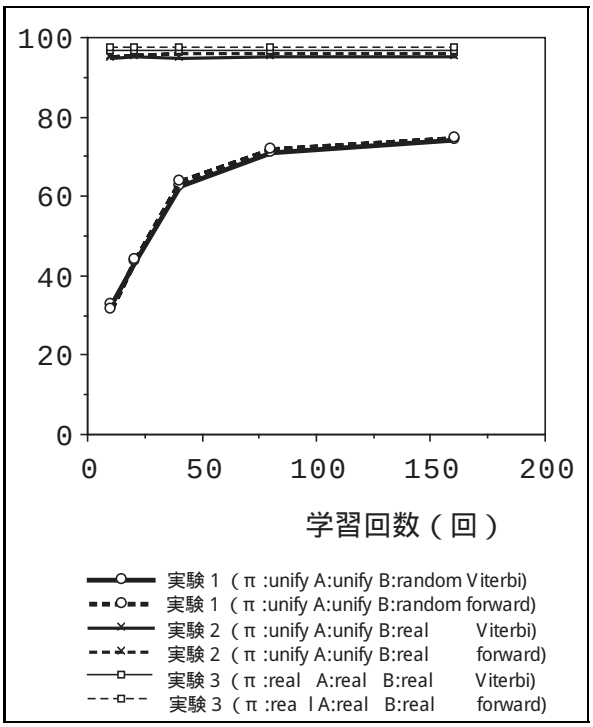

図 5: 長時間窓分析による LPC ケプストラムを用いた時の学習回数と識別率との関係

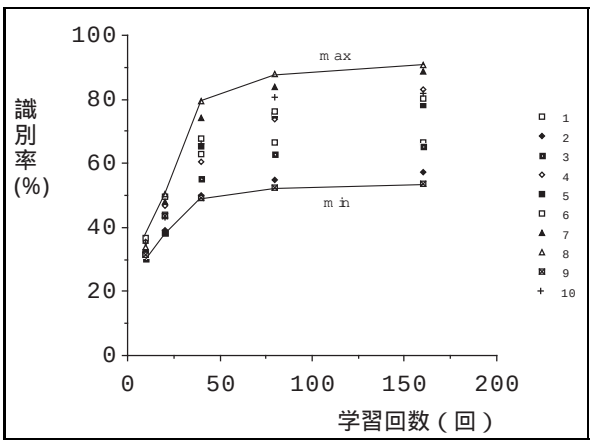

図 6: 音声資料別の識別率

5.3 尤度の変化

Baum-Welch の学習アルゴリズムは局所的極大点の方向に進むため、初期パラメータによって学習後のパラメータが大きく変化し、そのため学習時において尤度が変わっていると考えられる。そこで実験 1、実験 2、実験 3 における学習回数にたいする対数尤度の値を求めた。この結果を図 7 に示す。縦軸は対数尤度の値を示し、横軸は学習回数を示している。ただし、値は１０個の音声資料の平均値である。この結果から、最も高い平均識別率を持つ実験 3 の対数尤度の値が最も高いことがわかる。したがって、実験 1 において、学習回数を増加させる時、尤度が低下するため、平均識別率はさらに向上する可能性がある。

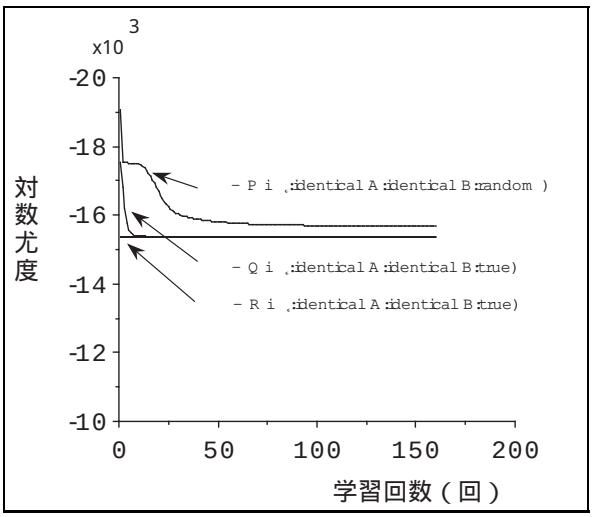

図 7: 学習回数と尤度の変化

5.4 話者 8 名における識別実験

ここでは、話者が 8 人の場合の識別実験結果について報告する。この場合、カテゴリーと話者の組み合わせが $8!=40320$ 通りあるため、識別率評価方法において、計算量が多くなる。そのため、実験は長時間窓分析のみ行なった。実験結果を図 8 に示す。この図から次のようなことがわかる。

1. 実験の結果、$\boldsymbol{B}^{(0)}$ のパラメータをランダムにした時 (実験１)、約 50%の平均識別率しか得られない。
2. パラメータをすべて真値にした時 (実験３)、Forward アルゴリズムでは平均識別率は 87% の値を得たが、Viterbi アルゴリズムでは、平均識別率が 45%しか得られない。この原因は、不明である。
3. いずれの実験でも、男性 8 名の場合は男性４名の時と比較して、識別性能が、大幅に低下する。したがって、話者が多くなるにしたがい、識別性能が急激に低下する傾向があると思われる。

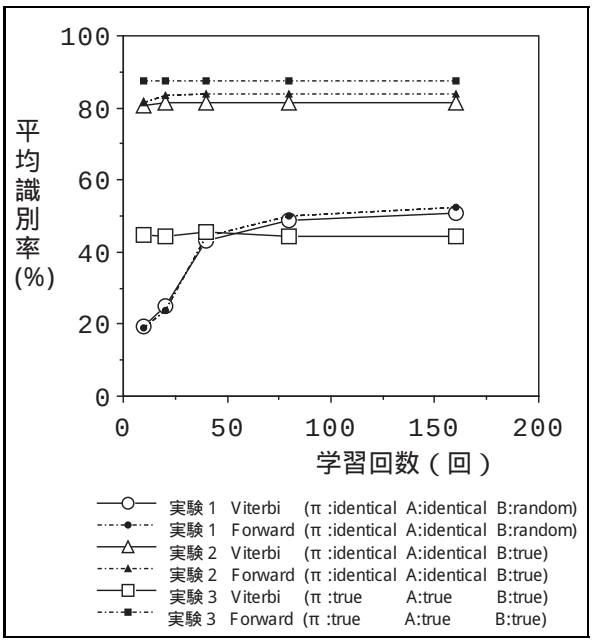

図 8: 男性話者 8 名における学習回数と識別率との関係

これらの実験から、ergodic HMM を用いて複数話者発話の識別問題を扱う時、Baum-Welch アルゴリズムの学習における初期パラメータの設定が重要であること、特にシンボル出力確率の設定が重要であることが示された。また、長時間窓分析の LPC ケプストラムを用いることにより、より良好な識別性能が得られることがわかった。

6 むすび

本稿では、信号系列を複数個の信号源に分解する問題を取り上げ、ergodic HMM を用いた解法を示した。応用例として複数話者発話の分解問題をあげ、実験により ergodic HMM による識別性能を示した。この実験から以下の事柄が示された。

1. segmentation とカテゴリ識別を同時に推定する、ergodic HMM による解法では、パラメータ推定の際の収束計算の初期パラメータの設定が重要である。
2. HMM のパラメータの初期状態確率、状態遷移確率、シンボル出力確率のうち、シンボル出力確率が最も重要である。
3. 音声分析の際、長時間窓分析した LPC ケプストラムを用いることにより、より良好な識別性能が得られる。

今後の課題として、以下の事項が挙げられよう。

1. 話者特徴量の抽出の検討

   話者性を特徴付けるような話者特徴量の抽出について検討する必要がある。今回は、長時間窓分析した LPC ケプストラムが話者性を特徴づけることが示された。この性質を定量的に測定するため、この分析窓長を変えて識別性能を調べる必要がある。なお、この類似の方法として文献 [?] では統計的な segmental model で実験を行なっている。他にも複数個のフレームから VQ コードの出現頻度分布を計算し、これを特徴量とする方法 [11] が考えられる。この方法は、前回の報告 [1] で行なった頻度分布のクラスタリングの結果を見ると、話者性抽出に対してかなり有効であると思われる。また、ケプストラムとともにΔケプストラムを用いることにより、識別性能が向上すると考えられる。

2. 最適化 VQ コードサイズ

   VQ 符号帳について検討を行なう必要がある。複数話者発話の分解問題のためには、コードサイズが 256 では小さい可能性がある。一方、サイズを大きくした場合、学習データにおいて出現するコードが少なくなるため。シンボル出力確率に信頼性が低下する。今後、最適なコードサイズを求める必要があろう。

3. 連続型 HMM の適用

   今回の実験では離散型 HMM のみ実験を行なった。HMM には、その他、Fuzzy VQ HMM や連続型 HMM がある。特に連続型 HMM は、Universal 符号帳が必要でなくなるため、VQ 歪などの問題点がなくなる。これらの HMM について検討することも重要である。

4. Baum-Welch アルゴリズムにおける初期パラメータの設定

   HMM の尤度関数は様々な局所的最適点が存在すると言われているが、それらの局所的最適点がそれぞれ異なった概念のカテゴリ識別を表現していると考えられる。例えば、ある最適点は話者識別を表現し、別のある最適点は言語識別を表現しているという具合である。よって Baum-Welch アルゴリズムにおける収束計算の初期パラメータの設定について考察する必要がある。また、今回の実験からシンボル出力確率を真値にした時、識別性能が大きく向上することがわかった。したがって、逆に、この値の信頼性と識別率の関係を調べる必要がある。

5. Second-order HMM

   通常の単純マルコフモデルでは、現在の状態は 1 つ前の状態しか考慮にいれていないため長時間における他状態への遷移の情報が欠落している。このため、より長時間の推移を考慮する second order HMM の考察も重要であると考えられる。

6. 識別評価方法

話者数を多くした場合の実験において識別評価方法を考える必要がある。今回の実験では全ての可能性を調べて、最も高い値を識別率としたが、この方法では、話者数が多くなるにしたがい、組み合せの数は階乗で多くなる。したがって識別評価手続きの高速化が必要である。これに対しては、分岐限定法などの組合せ最適化の適用が考えらよう。

7. ergodic HMM の状態数

   実際の音声にはポーズ区間がある。この区間は、物理的に話者を特定できない。今回の実験では、ergodic HMM の状態を話者の数と同じにしたため、この音声中にポーズがある場合には問題になる。この解決方法として、状態の数を話者の数より 1 つ多くして、ポーズ区間のみを生成する状態を挿入する方法がある。同様な手法で雑音区間など考えることができる。

謝辞

研究の機会を与えて頂いた北海道大学電子科学研究所 永井信夫教授、三木信弘助教授、ATR 自動翻訳電話研究所 樽松明社長、熱心に御討論していただくと同時に有益な御助言をいただいた音声情報処理研究室の皆様に感謝致します。

参考文献

[1] 渡辺秀行, 杉山雅英, 村上仁一: “未知・複数信号源クラスタリング問題,” 信学技報 SP92-45, pp.47-54 (1992-07).

[2] X.D. Huang, Y. Ariki and M.A. Jack : “Hidden Marcov Models for Speech Recognition,” Edinburgh University Press, Edinburgh (1990).

[3] G.Yu, M.Siu, H.Gish, An Unsupervised, Sequential, Learning Algorithm for the Segmentation of Speech Waveform with Multiple Speakers, Proc. of ICASSP92, 86.10 (Apr.1992).

[4] H.Gish, G.Yu, Clustering of Speakers Engaged in Dialog, Proc. of Speech Research Symposium XII (Jun.1992).

[5] 児島宏明, 田中和世, 速水悟 : “単語音声サンプルからの階層的な音韻概念の獲得”, 音響学会講演論文集,2-P-4 (1991-10).

[6] 村上仁一、嵯峨山茂樹: “HMM を用いた形態素解析,” 情報処理学会講演論文集,4F-01 (1992).

[7] 松井和子, 古井貞熙 : “エルゴード的 HMM による話者認識”, 音響学会講演論文集,3-6-14 (1991-10).

[8] 杉山雅英, 吉田哲也 : “スペクトル遷移の確率モデルによる言語識別の検討”, 音響学会講演論文集,1-P-1 (1992-03).

[9] 上田佳央, 中川聖一 : “隠れマルコフモデルによる自然言語のモデル化と多言語間の識別”, 情報処理学会講演論文集,7N-1 (1992).

[10] 古井貞熙, 板倉文忠, 斉藤修三 : “長時間平均スペクトルによる話者認識”, 信学会論文誌, 56A-11(1973).

[11] 杉山雅英, 村上仁一, 渡辺秀行 : “話者特徴に基づく音声セグメンテーションとクラスタリング,” 電子情報通信学会秋季全国大会, SA-6-1, (1992-09).