TR-I-0262

N 信号源分解問題

N-Source Decomposition Problem

渡辺 秀行　村上仁一　杉山 雅英
Hideyuki WATANABE　Jin'ichi MURAKAMI　Masahide SUGIYAMA

内容梗概

入力信号系列を複数個の信号源に分解する問題は、音声情報処理において応用範囲が極めて広い。ここではこの問題の定式化、解法を述べる。解法として、Universal VQ 符号帳とその出現頻度による方法、ergodic HMM による方法を提案する。前者は各区間の長さ (segmentation) が既知の場合における出現頻度分布のクラスタリングに基づく手法、後者はカテゴリ遷移を ergodic HMM でモデリングし segmentation とカテゴリ識別の同時推定を得るものである。実験では、応用例の一つとして複数話者発話の分解問題を検討する。実験により、提案する手法の有効性と問題点を示す。さらに、今後検討すべき課題について論ずる。

目次

表目次

図目次

1 はじめに

N 信号源分解問題は、与えられた信号系列に対して区間（ここではセグメントと呼ぶ）を検出し（segmentation）、各セグメントを複数個の信号源（カテゴリ）に識別（discrimination）する問題である。これは、多くの音声情報処理分野に対し基本的でありかつ重要な問題である。例えば、複数話者発話の識別問題、複数言語識別問題、音響単位の自動決定問題（音響言語モデル）等が応用例として挙げられる。

ここでは、N 信号源分解問題の解法を考察する。N 信号源分解問題は大別して3つの重要な問題から成る。即ち、1）入力系列のカテゴリ遷移の位置（セグメントの切れ目）を探す問題（segmentation 問題）、2）各セグメントを N 個のカテゴリに識別する問題、および3）カテゴリ数 N を推定する問題である。この3種の問題を一挙に解くことは極めて困難であり、問題がある程度既知であると仮定して解法を検討し、徐々に制限を緩めていく方向で考察を進めていくことが望ましい。

まず、カテゴリ数 N 及び segmentation が既知の場合において、Universal VQ 符号帳とその出現頻度による解法を提案する。この手法では、全セグメントのベクトルから作成された1個の符号帳を用意し、これを用いて各セグメントにおける量子化コードの出現頻度分布を算出することによりカテゴリ性を抽出する [4]。この複数個の頻度分布を N 個のカテゴリにクラスタリングすることにより識別が行なわれる。

次に、segmentation が未知の場合において、ergodic HMM による解法を提案する。この手法は、カテゴリ遷移を ergodic HMM[9, 10, 11]（各時刻において全状態へ遷移可能な HMM）でモデリングするものである。ここでは、Baum-Welch アルゴリズムで HMM パラメータを推定し、Viterbi アルゴリズムで最適状態遷移系列を推定することにより、segmentation とカテゴリ遷移の同時推定が得られる。

さらに、カテゴリ数 N を推定する問題を検討する。頻度分布クラスタリングの手法においては、クラスタリングの際に算出される歪み（distortion）を基準とする方法を、ergodic HMM による手法においては、Baum-Welch アルゴリズムの際に算出される尤度を基準にする方法を考察する。

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。実験により、話者数 N と segmentation が既知の場合、頻度分布のクラスタリングにより良好な話者識別が行なえることを示す。また、クラスタリングの際の歪みを基準として、話者数 N がある程度決定できることを示す。更に、segmentation が未知の場合において、ergodic HMM による解法では、パラメータ推定アルゴリズムにおける収束計算のための初期値設定が重要であること、及び話者性（カテゴリ性）を有するような特徴ベクトルを抽出する必要があることを示す。最後に、残された課題とその解決の展望について論ずる。

2 N 信号源分解問題の定式化

2.1 問題設定

図 1 に示すように、今、n 次元実ベクトルで特徴付けられる信号の系列 $X = \{x_t\}(t = 1, 2, ..., T)$ が与えられているとする。この系列は、隣同士で性質の異なる K 個のセグメント X_k $(k = 1, 2, ..., K)$ からなるとし、各セグメントが $N(\leq K)$ 個（有限個）のカテゴリ（状態）C_j $(j = 1, 2, ..., N)$ のいずれかから生じた系列であると考える。ここで、セグメント X_k の構成要素が $\{x_{t_{k-1}+1} ... x_{t_k}\}$（ただし、$t_0 = 0$, $t_K = T$）、要素数が M_k $(= t_k - t_{k-1})$ であるとする。

N 信号源分解問題は、与えられたベクトル信号系列に対し、セグメントの切れ目の位置 t_k $(k = 1, 2, ..., K-1)$ を探し（segmentation）、この K 個のセグメントを $N(\leq K)$ 個のカテゴリ（状態）に識別 (discrimination) することである。この際、カテゴリ数 N を推定することも必要である。

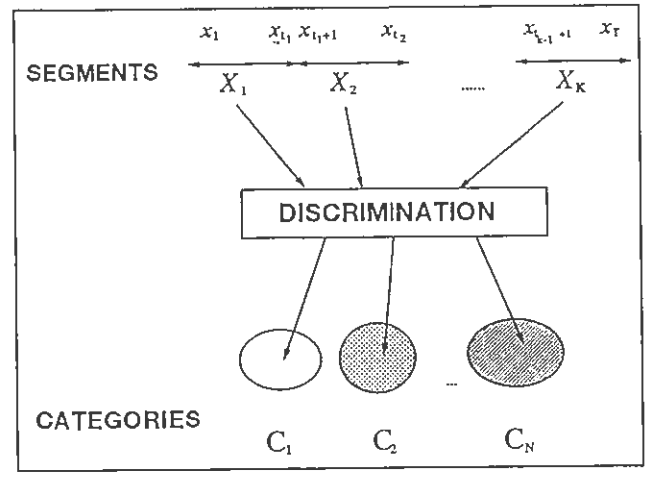

図 1: N 信号源分解問題

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声特徴ベクトル系列に対応し、各カテゴリは各話者に、セグメントの切れ目は話者の遷移に対応する。また、音響単位の自動決定問題（音響言語モデル）がある。この場合、カテゴリを音素のような音響単位に対応付ける。これは、言語モデルの構築において、従来の音素にかわる

新しい音韻概念の獲得が目的である [12]。その他、言語識別問題、発話様式（単語・文発声・連続発声等）の識別問題、及び音声・非音声識別問題など、信号系列を分解してカテゴライズする問題に帰着するタスクに応用できる。

2.3 Universal VQ 符号帳とその出現頻度による解法

まず、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 信号源分解問題は、各セグメント X_k $(k = 1, 2, ..., K)$ を、N 個のカテゴリ C_j $(j = 1, 2, ..., N)$ にクラスタリングする問題に帰着する。

2.3.1 VQ コードの出現頻度分布によるカテゴリ性の抽出

クラスタリングに際し、各セグメントのカテゴリ性を表現する量を算出することが重要である。その方法として、共通ベクトル量子化 (Universal VQ) 符号帳 とその VQ コードの出現頻度による方法が提案されており、既にテキスト独立型の話者認識や言語識別に対して有力であることが報告されている [4, 5]。ここで、この手法について簡潔に述べる。

Universal 符号帳 U は、各セグメントからベクトルを取りだし、その全てのベクトルに対し 1 個の符号帳を作成することにより得られる。識別用入力信号系列の各セグメント X_k $(k = 1, 2, ..., K)$ はそれぞれ U により量子化されるが、U の各 VQ コード u_l $(l = 1, 2, ..., L)$ （L はコードブックサイズ）に量子化される頻度分布は、ベクトル数がある程度多ければ各カテゴリ固有の特徴となり [4]、入力系列の各セグメントのカテゴリ性を表現すると考えられる。ここで、k 番目セグメントの M_k 個のベクトルのうち u_l に量子化されたベクトル数を K_{kl} とすると、出現頻度 P_{kl} 及び頻度分布 q_k は次式で与えられる（図 2）。

$$P_{kl} = \frac{K_{kl}}{M_l} \tag{1}$$

$$q_k = [P_{k1}\ P_{k2}\ ...\ P_{kL}]^T \tag{2}$$

尚、肩付 T はベクトルの転置を表す。各 q_k の成分について $\sum_{l=1}^{L} P_{kl} = 1$ が成立する。以後、この頻度分布 q_k （L 次元ベクトル）を k 番目セグメントのカテゴリ性を表現する量として扱う。

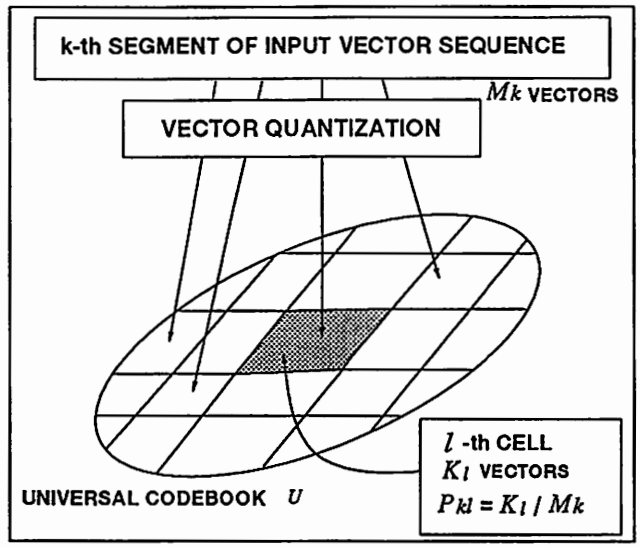

図 2: k 番目セグメントにおける VQ コード出現頻度分布の計算

2.3.2 カテゴリ識別のためのクラスタリング手法

次に、q_k $(k = 1, 2, ..., K)$ を N 個のカテゴリにクラスタリングする問題を考える。クラスタリング手法として種々の方法が提案されているが、ここでは、特に有力な手法として知られる k-means アルゴリズム [1, 6]、及びそれを骨子とする手法であり音声のベクトル量子化器の作成でよく用いられる LBG アルゴリズム [1, 2] の適用を考える。k-means、LBG 両アルゴリズムをそれぞれ表 6、7 に示す。k-means アルゴリズムは、まず各入力ベクトルに対し最も近い centroid のラベルを与えることにより新しいセルを形成し、次に各セルの centroid を更新する。これらを全体の歪み (distortion) がある程度小さくなるまで続ける。k-means アルゴリズムは centroid の初期値により様々な収束点を与える。LBG アルゴリズムは、k-means アルゴリズムで求まったセルを 2 分割 (splitting) することにより次ステップの k-means アルゴリズムの初期値を自動的に与える。しかし LBG では、セルの個数が 2 の巾乗となり、このことは量子化器の作成に対しては効果的であるが、任意カテゴリ数のクラスタリングを考える本問題には適していない。また比較的歪みの小さいセルまでも splitting されてしまうことは、クラスタリングの観点から適切ではないと思われる。よってここでは、LBG アルゴリズムの splitting を若干変更した、表 8 の手法を提案する（以降、変形 LBG アルゴリズムと呼ぶ）。すなわち、通常の LBG アルゴリズムでは全ての centroid が splitting されるのに対し、変形 LBG では、歪みが最も大きいセルの centroid のみが splitting される。これにより、2 の巾乗以外のカテゴリ数の場合でもクラスタリングが可能となり、かつ歪みの小さいセルが split-

ting されることがなくなる。LBG と変形 LBG の比較実験は、4.2.1 で示す。

またクラスタリングにおける距離尺度 $d(q_k, z_j)$ として、ここでは、次式で表される通常の Euclid 距離及び二つのタイプの Kullback-Leibler (KL) 情報量を適用する。

Euclid 距離：

$$d(q_k, z_j) = \sqrt{\sum_{l=1}^{L}(P_{kl} - z_{jl})^2} \tag{3}$$

KL 情報量 type1 ：

$$d(q_k, z_j) = \sum_{l=1}^{L} z_{jl} \log \frac{z_{jl}}{P_{kl}} \tag{4}$$

KL 情報量 type2 ：

$$d(q_k, z_j) = \sum_{l=1}^{L} P_{kl} \log \frac{P_{kl}}{z_{jl}} \tag{5}$$

ただし z_j は j 番目セルの centroid であり、$z_j = [z_{j1} \ldots z_{jL}]^T$ である。KL 情報量 [7, 8] は、二つの確率分布間の「近さ」を測る客観的な基準として考え出されたものであり、Euclid 距離よりも、二つのベクトルが確率分布であることを考慮している。KL 情報量type1 は各セグメントの頻度分布が centroid から見てどれだけ「逸脱」しているかを測る尺度、KL 情報量 type2 は逆に centroid が各頻度分布からどれだけ「逸脱」しているかを測る尺度と考えることができる [7]。式 (4)、(5) の値は一般的に異なる。

以上三つの尺度を適用した場合、centroid 計算はそれぞれ次式によって行なわれる（付録 A）。

Euclid 距離：

$$z_j^{(m+1)} = \frac{1}{K_j^{(m+1)}} \sum_{q_k \in C_j^{(m+1)}} q_k \tag{6}$$

KL 情報量 type1 ：

$$z_{jl}^{(m+1)} = \frac{Q_{jl}^{(m+1)}}{\sum_{l=1}^{L} Q_{jl}^{(m+1)}} \tag{7}$$

$$Q_{jl}^{(m+1)} = \left(\prod_{q_k \in C_j^{(m+1)}} P_{kl} \right)^{\frac{1}{K_j^{(m+1)}}}$$

KL 情報量 type2 ： Euclid 距離と同様。

尚、$K_j^{(m+1)}$ は、セル $C_j^{(m+1)}$ に属するベクトル数を表し、$(m+1)$ は、k-means アルゴリズムの $m+1$ 回目のイテレーションを意味する。以上述べた解法の概略を、図 3 に示す。

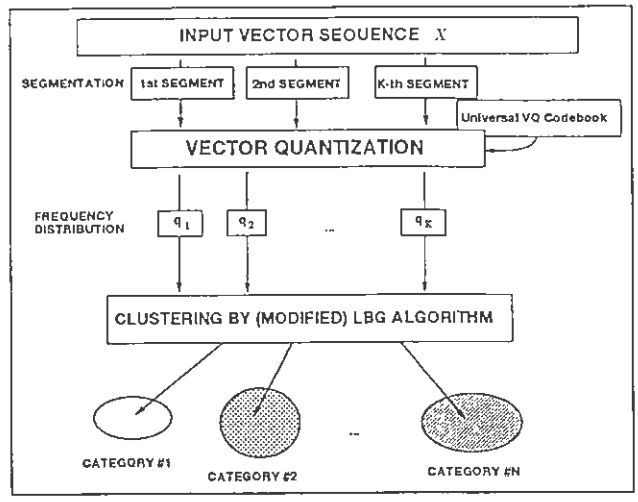

図 3: Universal 符号帳とその出現頻度による解法

2.4 ergodic HMM による解法

前述の解法は、入力系列の segmentation が未知である場合に困難となる。ここでは、segmentation 未知の場合におけるカテゴリ識別を考察する。これへ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。

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るものである [1, 2, 3]。すなわち音声認識では、オートマトン制御の下で確率的定常信号源（状態）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 が適用される。このことから、N 信号源分解問題に対しても、オートマトン制御によりカテゴリ（状態）を切替えることで入力系列の状態遷移を表現するモデルとして、HMM が有力であると期待される。

一般に音声認識では、音声の特性を考慮して、例えば図 4 に示すような LR (left-to-right) 型の HMM が用いられる。ところが N 信号源分解問題では明らかに、以前に通過した状態に戻るような遷移も考えられる。よって、図 5 に示すようなエルゴディック (ergodic) HMM が有効である。すなわち ergodic HMM は各時刻において全状態へ遷移可能なモデルである。尚 ergodic HMM は、スペクトル遷移の動的な特徴を考慮できるという特性から、話者や言語のモデリングに有力であることが報告されている [9, 10, 11]。

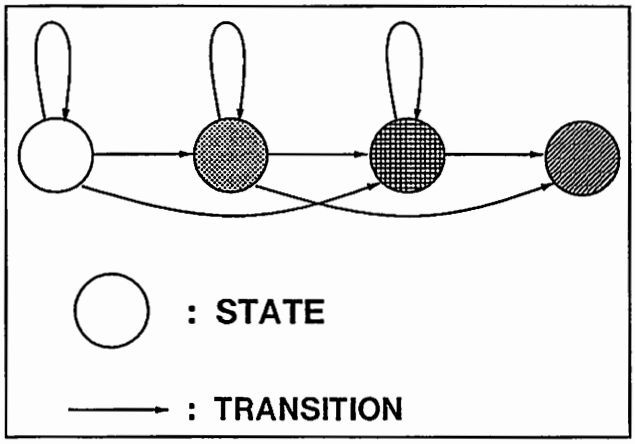

図 4: left-to-right HMM (4 states, 8 transitions)

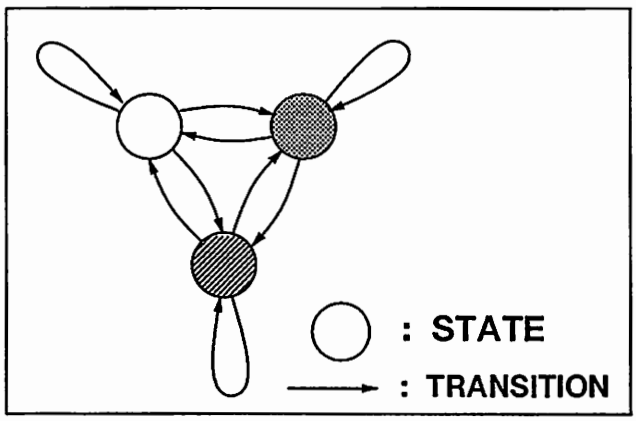

図 5: ergodic HMM (3 states)

HMM では、状態遷移系列は直接観測できない。よって、全ての可能な状態遷移に対する HMM の出力シンボル（入力系列に対応する）の尤度を最大化するような HMM のパラメータを推定し、その推定パラメータから、シンボル系列を生み出す確率の最も高い状態遷移系列を推定する手順がとられる。すなわち、ergodic HMM によるカテゴリ識別では、次の二つの問題が重要である。

1. 入力系列 X の尤度を最大にする HMM のパラメータ M を推定する（モデルの推定、または学習）。M は、初期状態確率、状態遷移確率、シンボル出力確率から成る。

2. M が 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（最適状態遷移系列の推定）。

上の 2 により求まる系列が、入力系列の segmentation とカテゴリ識別の同時推定を与える。

HMM による解法の概略を、図 6 に示す。

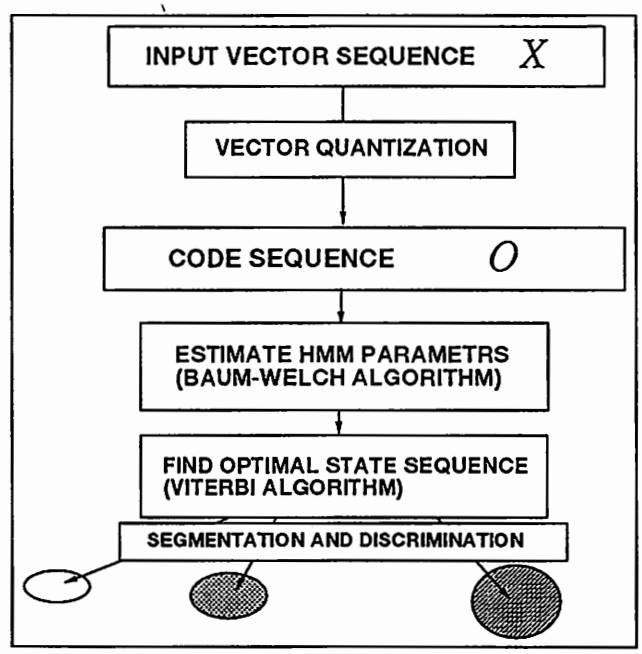

図 6: 離散型 ergodic HMM による解法

ここでは、離散型 (discrete) HMM を考える。離散型 HMM では、入力系列は VQ コードブックにより量子化され、パラメータ推定、状態系列推定共にシンボルの出力確率は VQ コードの出現頻度として与えられる。シンボル出力確率を複数種類のガウス分布の混合と考える連続型 (continuous) HMM の適用については、今後の課題である。

2.4.1 HMM のパラメータ推定について

離散型 HMM は、パラメータ $M = (\pi, A, B)$ で特徴付けられる。ここで、

$$\begin{aligned} \pi &= \{\pi_i\}_{i=1}^{N} \\ A &= \{a_{ij}\}_{i,j=1}^{N} \\ B &= \{b_j(l)\} \quad (j = 1...N, l = 1...L) \end{aligned} \tag{8}$$

であり、π_i、a_{ij} 及び $b_j(l)$ はそれぞれ、初期状態が i である確率、状態 i から状態 j に遷移する確率及び状態 j がシンボル（VQ コード）u_l を出力する確率である。また N は状態数、L は VQ コードブックサイズである。尚ここでは問題の性質上、各状態がシンボルを出力すると考える状態出力タイプ (Moore 型) の HMM を考える。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 [1, 2] を、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。Baum-Welch アルゴリズムを付録 B に示す。

N 信号源分解問題において Baum-Welch アルゴリズムを適用する場合、2 つの重要な問題が生ずる。1 つは、入

力系列の長さがかなり長いために、シンボルの出現確率を表すαパラメータの値が小さくなり過ぎ、アンダフローが起こる問題である。この一解決策として、スケーリング法が知られている。スケーリングを適用したBaum-Welchアルゴリズムを付録Bに示す。これにより、αパラメータが全ての$t \in \{1,2,...,T\}$で$\sum_{i=1}^{N} \alpha_t(i) = 1$となるため、アンダーフローが回避される。またもう一つの問題は、Baum-Welchアルゴリズムが必ずしも尤度関数の大域的最大点に収束するとは限らず、パラメータの初期値により様々な局所的最大点に収束することである。すなわち、所望のカテゴリ識別を行なうためには初期値の設定が重要となる。これについては、実験で考察する。

2.4.2 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、入力系列を生み出す確率の最も高い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アルゴリズムが知られている。Viterbiアルゴリズムを付録Cに示す。これにより、入力系列のsegmentation及びカテゴリ識別の同時推定が与えられる。

Viterbiアルゴリズムは、確率の計算を全て対数の加算で行なえるため、入力系列の長さがかなり長くなる場合でもアンダーフローの心配がない。

3 カテゴリ数Nの推定について

前節で提案した手法に対しては、カテゴリ数Nを事前に知る必要がある。ここでは、Nの推定について考察する。

Universal符号帳と出現頻度による解法の場合、クラスタリングにおいてセル全体の平均distortion Dが計算される（表6）。セル個数が真のカテゴリ数と等しい場合におけるDの値が、話者数や性別などが変わってもあまり変化しないならば、Dの絶対的な値を基にNをある程度推定できると思われる。これについては、4.2.4で定量的に考察する。

ergodic HMMの場合、Baum-Welchのパラメータ推定アルゴリズムにおいて、シンボル系列に対するモデルの尤度が計算される。出現頻度による解法と同様、尤度の絶対的な値を基にNを推定できるのではないかと考えられる。また、Nの推定値を客観的に与える方法として、AIC基準やMDL基準等の情報量基準の適用も考えられる（付録D）。これらについては今後の課題として残しておく。

4 複数話者識別実験

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、先に提案した解法について実験により考察する。

4.1 実験条件

4.1.1 音声分析条件及び扱う信号系列

音声分析条件を表1に示す。ここでは簡単化のため、特徴量としてLPCケプストラム（1次〜16次）を採用する。Δケプストラムの採用による識別性能の向上が考えられるが、これについては今後の課題とする。

表1: 音声分析条件

音声特徴量	LPCケプストラム
自己相関分析	15次
LPC分析	14次
打ち切り次数	16次
標本化周波数	12 kHz
フレーム長	21.3 ms
フレーム周期	10.7 ms
高域強調	$(1-0.97z^{-1})$

ここで扱う話者名と発声の順序を表2に示す。例えば男性2名、女性2名の会話では、MAU, FKN, MNM, FKS, MAU, FKN, MNM, FKS, ... の順で発話されるものとする。識別用信号系列を図7に示す。一発話が1セグメントに対応し、話者遷移の一順したものが10個分で、一系列が成り立っている。よって、2人の会話では全部で20セグメント、3人では30セグメント、8人では80セグメントになる。この図に示されている系列は1回試行分に対応し、識別評価は10回試行の平均値で行なわれる。各試行のデータ（MAUの場合）を表3に示す。他の話者でも同様である。すなわち、各試行の系列は表4のようになる。なおここでは、各単語に無音区間が含まれている。

表2: 話者名（話者の遷移もこの順）

2名	男性1名女性1名	MAU, FKN
	男性2名	MAU, MNM
3名	男性2名女性1名	MAU, FKN, MNM
	男性3名	MAU, MNM, MHT
4名	男性2名女性2名	MAU, FKN, MNM, FKS
	男性4名	MAU, MNM, MHT, MMS
8名	男性4名女性4名	MAU, FKN, MNM, FKS, MHT, FFS, MMS, FKM
	男性8名	MAU, MNM, MHT, MMS, MSH, MMY, MTK, MTM

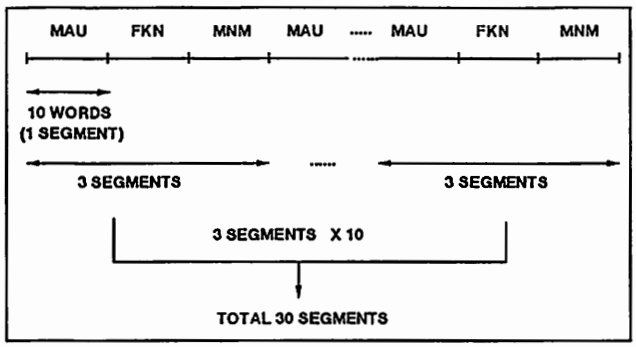

図 7: 信号系列（各話者 10 単語の発声による、男性 2 名、女性 1 名の会話の場合）

表 3: 識別用データ（MAU の場合）

MAU-1-0001...MAU-1-0010 MAU-1-0051...MAU-1-0060 ... MAU-1-0451...MAU-1-0460	（1 回目の試行用）
MAU-1-0501...MAU-1-0510 MAU-1-0551...MAU-1-0560 ... MAU-1-0951...MAU-1-0960	（2 回目の試行用）
...	...
MAU-1-4501...MAU-1-4560 MAU-1-4551...MAU-1-4560 ... MAU-1-4951...MAU-1-4960	(10 回目の試行用)

表 4: 入力信号系列（MAU・FKN・MNM の会話、4 単語の例）

1 回目 試行	MAU-1-0001...MAU-1-0004 FKN-1-0001...FKN-1-0004 MNM-1-0001...MNM-1-0004
	MAU-1-0051...MAU-1-0054 FKN-1-0051...FKN-1-0054 MNM-1-0051...MNM-1-0054
	...
	MAU-1-0451...MAU-1-0454 FKN-1-0451...FKN-1-0454 MNM-1-0451...MNM-1-0454
...	...
10 回目 試行	MAU-1-4501...MAU-1-4504 FKN-1-4501...FKN-1-4504 MNM-1-4501...MNM-1-4504
	MAU-1-4551...MAU-1-4554 FKN-1-4551...FKN-1-4554 MNM-1-4551...MNM-1-4554
	...
	MAU-1-4951...MAU-1-4954 FKN-1-4951...FKN-1-4954 MNM-1-4951...MNM-1-4954

4.1.2 識別評価方法

信号系列を複数の状態（カテゴリ）に分類した場合、各状態がどの話者に対応しているのかまでは同定できない。すなわち、求まる系列は状態の番号であり、番号と話者の対応には $N!$ 個の組合せが存在する。そのため識別評価は容易ではない。ここでは、表 9 の評価方法を適用する。すなわち、各話者に符号を割り当て、$N!$ 通りの割り当て方について正解率を算出し、その中の最大値を識別率とする。この表の場合は識別率が 70 パーセントということになる。最終的な識別率は、各試行の識別率の 10 回分を平均して算出される。この評価は、出現頻度分布による方法ではセグメント単位で、ergodic HMM の方法ではフレーム単位で行なわれる。

この評価は、特に ergodic HMM の場合ではかなりの計算量を要する。分岐限定法などの組合せ最適化法の適用による評価演算の高速化については、今後の課題である。

4.1.3 Universal 符号帳の作成手順

Universal 符号帳作成用のデータを、表 5 に示す。他の話者でも同様である。ここでは、符号帳を 1 人 100 単語で作成する。コードサイズは 256 とし、距離尺度として Eu-

clid 距離を採用する。話者8人の場合では、フレーム数が約80000フレームとなり計算量が膨大となるため、符号帳を2段階で作成する。すなわち、まず話者毎に、100単語（約40000フレーム）から512個のコードを求め、次に8名分のコード（計4096個）から、256個のコードを求める。話者2、3、4人の場合は、1段階で符号帳を作成する。

表 5: 符号帳作成用データ（MAU の場合）

MAU-1-0041...MAU-1-0050 MAU-1-0091...MAU-1-0100 ... MAU-1-0491...MAU-1-0500	（1人100単語）

4.2 Universal 符号帳とその出現頻度による解法

4.2.1 LBG アルゴリズムと変形 LBG アルゴリズムの比較実験

LBG アルゴリズムと変形 LBG アルゴリズムの性能を比較した実験結果を図 8 に示す。これは男性2人と女性2人の計4人の発話の場合である。グラフの横軸は各セグメントの単語数、縦軸は識別率を表す。図を見て明らかなとおり、いずれの尺度を用いても、変形 LBG アルゴリズムは LBG アルゴリズムよりも良い識別率を与えている。これは、全てのセルを splitting するより、distortion 最大のセルのみ splitting してゆく方が、クラスタリングの性能を上げるという観点で優れていることを示している。

この実験結果より、以降の実験ではクラスタリング手法として変形 LBG アルゴリズムを用いることにする。

4.2.2 話者数と識別率との関係

話者数と識別率との関係について調べた実験結果を、図9（男女同数）、図 10 （男性のみ）に示す。話者数が増加するに従い識別率が下降する傾向が示されている。話者2人の発話の場合では、男女の場合でも男性のみの場合でも、1単語の発話でほぼ完全に識別できることがわかる。10単語の場合、尺度として KL 情報量 type2 を適用すれば、男女同数の場合でほぼ100%、男性のみの場合では、3人でほぼ100%、4人で約95%、8人で約90%の識別率が得られる。1単語の場合、話者が4人を越えるといずれの尺度を用いても識別が困難になることがわかる。また、特に話者8人の場合において、Euclid 距離では単語数が増加しても識別率があまり上昇しないのに対し、KL 情報量では識別率の上昇が明確である。これは、単語数の増加に従い同一話者の頻度分布のばらつきが小さくなることが、KL 情報量に基づいたクラスタリングに強く反映するためであると考えられる。逆に音声長が短い場合は、Euclid 距離が最も良い識別を与えていると言える。

4.2.3 単語数と識別率との関係

各セグメント（発話）の単語数と識別率との関係について調べた実験結果を図 11 （話者4人の場合）、図 12 （話者8人の場合）に示す。単語数が増加するに従い識別率が上昇する傾向が示されている。男女同数の場合と男性のみの場合とでは、後者の方が識別率が低いこともわかる。話者4人・男女同数の場合、Euclid 距離では10単語でも識別率が90%強であるのに対し、KL 情報量 type2 では5単語でほぼ完全に識別できることがわかる。話者8人で男女同数の場合、Euclid 距離や KL 情報量 type1 では10単語未満で良好な識別率が得られないのに対し、KL 情報量 type2 では6単語で95%の識別率が得られている。話者8人で男性のみの場合、状況はかなり厳しくなり、KL 情報量 type2 の場合において10単語で約90%であるのが最高である。また、Euclid 距離では男女同数の場合と男性のみの場合とで識別率があまり違わないのに対し、KL 情報量では、特に単語数が多い場合、男女同数の場合の方が識別率が大きく上昇している。これは、男女の違いが、Euclid 距離よりも情報量に強く反映するためであると考えられ、大変興味深い特性である。話者8人の場合において、3種の尺度による識別率の関係は、音声長が5単語より短い場合、Euclid > KL type1 > KL type2、音声長が5単語以上の場合、Euclid < KL type1 < KL type2 である。このことは、各セグメントの音声長により尺度を使い分ける必要があることを示している。

4.2.4 distortion の値を基準とした話者数の推定について

変形 LBG アルゴリズムの際に算出される平均 distortion D を基準として、話者数 N を推定する問題を考える。良好な推定のためには、真の話者数と一致したセル個数の場合における D の値（以降、最適 distortion と称する）が、話者数や話者、性別、及び単語数によらず一定であることが望ましい。

まず、単語数を固定して、話者数や性別を変化させた場合の最適 distortion の変化を図 13 に示す。この図は、5単語、KL 情報量 type2 の場合である。図を見てわかる通り、多少のばらつきはあるものの、話者数や性別が異なっても最適 distortion の値はあまり変化しないことがわかる。他の単語数及び他の尺度の場合も同様である。

次に、話者を固定して、単語数を変化させた場合の最適 distortion の変化を図 14 に示す。この図は、男性2名・女性2名、KL 情報量 type2 の場合である。図を見ると、単語数の増加に従い最適 distortion の値が減少することがわかる。これは、単語数の増加により、同一話者におけるコー

ド出現頻度分布のばらつきが減少するためであると考えられる。

以上により、各発話（セグメント）の長さ（フレーム数）がある程度わかれば、distortion の値からある程度話者数を推定することができると考えられる。また、図 14 を見てわかる通り、単語数がある程度多くなれば最適 distortion の減少度が小さくなることから、各発話の長さがある程度長ければ、長さが未知であっても distortion の値からある程度話者数を推定することができると思われる。

4.2.5 distortion の変化を基準とした話者数の推定について

変形 LBG アルゴリズムにおいて、セル個数の増加による distortion の変化を基準として、話者数 N を推定することを試みる。

男性 2 名・女性 1 名、5 単語、KL 情報量 type2 の場合における、セルの個数と distortion の関係を図 15 に示す。グラフでは、3 回試行の結果を重ね書きしている。この設定は、識別率が 100% の場合を取り上げたものである。図を見てわかるように、セルを増加させた場合、セル個数が真の話者数になるまでは distortion の減少が急激であるが、それ以降は減少度が小さいことがわかる。これは次のように解釈できる。すなわち、図 17 のようにベクトルのクラスタ形成が明確である場合、セル個数が真のクラスタ数と一致するまでは、全セルについての centroid からのばらつきが大きく減少するが、セル個数をそれより多くしても、centroid からのばらつきの総和があまり変化しなくなることが原因であると推測される。

逆に図 18 に示すように、クラスタ形成が明確ではない場合は、distortion の減少度が終始緩やかであると考えられる。図 16 は、男性 4 名、5 単語、KL 情報量 type2 の場合であるが、distortion 変化のカーブが鈍くなっており、話者数推定が図 15 に比べて困難であると思われる。これは、頻度分布ベクトルが全て男性のものであることにより、クラスタ形成が明確でないためであると考えられる。なおこの場合、識別率も約 90% と低い。

よって、出現頻度分布ベクトルのクラスタ形成が明確であり、1 つのクラスタが 1 人の話者に対応している場合、distortion の変化を見ることにより話者数をある程度推定できると考えられる。この問題に対しては、頻度分布が話者性をいかに良く表現できているかが重要なポイントとなるであろう。具体的な推定実験は今後の課題である。

4.3 ergodic HMM による解法

次に、ergodic HMM による解法についての実験結果を示し、考察を行なう。ここでは、行なった実験の経過を中心に述べることにする。結果として、Baum-Welch アルゴリズムにおけるパラメータの初期値設定が重要であること、及び話者識別をターゲットとする場合はベクトル（フレーム）単位に話者性を特徴付けることが必要であることがわかった。

4.3.1 ランダムなパラメータ初期値を与えた場合

まず、HMM パラメータ $M = (\pi, A, B)$ の初期値として、初期状態確率 $\pi^{(0)}$ を等確率、状態遷移確率 $A^{(0)}$ も等確率に、シンボル出力確率 $B^{(0)}$ をランダムに与えた場合で実験を行なった。なお、$B^{(0)}$ のみランダムに与えるのは、等確率に与えた場合に B の学習が進行しないためである。この結果、男性 1 名・女性 1 名の場合に識別率 70 ～ 80%、男性 2 名・女性 2 名では 40 ～ 50% しか得られない。この率は各セグメントの単語数が異なってもあまり変化しない。このように識別率が低いのは、初期値をランダムに与えていることが原因と考えた。

4.3.2 出現頻度分布による手法の結果を用いた場合

次に、$\pi^{(0)}$、$A^{(0)}$ を等確率に与え、$B^{(0)}$ を Universal 符号帳のコード出現頻度分布による手法を基に算出した場合の実験を行なう。ここでは、$B^{(0)}$ を次式で与える。

$$b_j^{(0)}(l) = \frac{1}{K_j} \sum_{k \in \hat{s}_k = j} P_{kl} \qquad (9)$$
$$(j = 1...N; l = 1...L)$$

ここで、K_j は K 個のセグメントのうち推定されたカテゴリが j であるセグメント数、$\hat{s}_k$ は k 番目セグメントの推定カテゴリ、P_{kl} は 2 節で述べた通り、k 番目セグメントにおけるコード l の出現頻度である。

この結果、男性 1 人・女性 1 人で識別率 97 ～ 99%、男性 2 人で約 95% の識別率が得られ、ランダムに初期値を与えた場合に比べ識別率の大幅な向上が見られた。しかし、出現頻度分布のクラスタリングによる方法ではいずれの場合でも識別率が 100% であり、ergodic HMM の適用による識別率の改善は見られなかった。さらに、男性 2 名・女性 2 名の場合では、識別率が 70 ～ 80% であり、ランダムに初期値を与えた場合と比較して識別率の向上は見られたものの、頻度分布のクラスタリングの方法では 90% 以上が得られていることを考えると、以前識別率は良くない。これは、Viterbi アルゴリズムの推定結果において、例えば図 19 のように、同一の状態が集中しているところに他の状態が割り込むような形で出現するのが原因である。

4.3.3 Garbage Model の導入

図 19 に示すような現象が起こるのは、各単語が無音区間などの Garbage を含むことが原因の一つとして考えられる。ここでは、男性 2 名・女性 2 名の場合において 5-state HMM で実験を行なう。すなわち、5 番目状態を Garbage Model に対応付けることにより、識別率の向上を計る（以

降、Garbage Model に対応付けられる状態を Garbage state と呼ぶ）。初期値として、$\pi^{(0)}$ と $A^{(0)}$ は等確率に与え、1～4番目状態の $B^{(0)}$ は出現頻度による手法の結果を基に算出し、Garbage state の $B^{(0)}$ はランダムに与える。識別評価は、推定状態が Garbage state であるフレームを除外して行なう。この結果、識別率は75%前後で、4.3.2の結果に比べ識別率の向上はほとんど見られなかった。すなわち、図19に示すような割り込み状態が、うまく Garbage state に吸収されていないことが言える。これについては、Garbage state のパラメータ初期値の与え方等、改善の余地が充分あると思われる。

4.3.4 Baum-Welch アルゴリズムの学習能力について

ここでは、図19に示すような現象が起こるのは、Baum-Welch アルゴリズムの学習能力に問題があるのではないかと考え、イテレーション回数と識別率との関係について実験を行なうことにより、この問題を考察する。

男性2名・女性2名、10単語の場合の結果を図 20 に示す。図を見てわかる通り、$\pi^{(0)}$、$A^{(0)}$、$B^{(0)}$ を全て真値に与えた場合、パラメータ更新をしなくてもほぼ100%の識別率を与えている。つまり、HMM の表現能力自体は問題がないことがわかる。$\pi^{(0)}$、$A^{(0)}$ を等確率に与え、$B^{(0)}$ のみ真値に与えた場合は、イテレーション回数2回で識別率が80%強であるのが最高で、さらにイテレーションを進めると率が下がっていく。この時の π、A の推定値を見てみると、π は1つの状態のみ1.0で他状態が0.0と完全に学習されるのに対し、A は、各状態における他状態への遷移確率が真値に比べ大きくなっている。このことから、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になり、これが原因で図19のような現象がおきてしまうと考えられる。$\pi^{(0)}$、$A^{(0)}$ を真値とし、$B^{(0)}$ をランダムとした時もやはり学習がうまく行われていない。

以上により、$\pi^{(0)}$、$A^{(0)}$、$B^{(0)}$ を全て真値に与えるとほぼ100%の識別率が得られるが、$B^{(0)}$ のみを真に与えた場合は、A がうまく学習されないため、識別率が低いことがわかった。また、$\pi^{(0)}$、$A^{(0)}$ のみを真値として与えた場合も、B がうまく学習されていない。つまり、Baum-Welch アルゴリズムの学習に対して初期値の設定が重要であることが示されたことになる。

4.3.5 音声分析のフレーム長を長くした場合

ここでは、各フレームに話者性をより特徴付けることにより、識別率の向上を計る。音声分析の際のフレーム長及びフレーム周期を10倍に長くする（即ち、フレーム長213ms、フレーム周期107ms）。男性4名・10単語の場合で、$\pi^{(0)}$、$A^{(0)}$ を等確率に、$B^{(0)}$ を真値に与えたときの結果を図21に示す。同時に、変更前のフレーム長・フレーム周期（21.3ms, 10.7ms）の結果も示している。図を見てわかる通り、変更前の短いフレーム長では最大85%の識別率であるのに対し、フレーム長を長くした場合、5回のイテレーションで100%近い識別率が得られるのがわかる。男性8人の場合を図22に示す。イテレーション10回で98%の識別率が得られている。

よって、フレーム長を長くして、フレーム単位に話者性をより特徴付けることにより、$B^{(0)}$ を真値に与えれば、$A^{(0)}$ もよく学習され、良好な識別が行なわれることがわかった。これは結果的に、HMM がフレーム単位で状態遷移をモデリングしていることを考えると、当然のことと言える。

フレーム長が長い場合で、$\pi^{(0)}$、$A^{(0)}$ を等確率、$B^{(0)}$ をランダムに与えたときの結果を図23に示す。男性4人では95%の識別率が得られているが、男性8人では40%とかなり低い。ここでも、初期値の設定が重要であることが示されている。

なお、イテレーション回数と識別率との関係を調べた実験では、簡単化のため1番目試行用のデータのみ行なった（つまり試行回数1回である）。従って、他のデータで実験を行なえば、多少異なる結果が得られる可能性がある。

5 むすび

入力信号系列を複数個の信号源に分解する問題を取り上げ、その解法を示した。実験では、応用例として複数話者発話の分解問題を検討した。それにより、以下の結論を得た。

1. segmentation が既知の場合における、Universal 符号帳とその出現頻度による解法では、クラスタリングの距離尺度として Kullback-Leibler の情報量を用いることにより、良好な識別が行なえる。

2. 出現頻度による解法において、話者数 N は、クラスタリングの際に算出される歪み（distortion）を基準にして、ある程度推定できると考えられる。

3. 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値設定が重要である。また、音声分析の際のフレーム長を長くして、各フレームに話者性をより特徴付けることにより、より良好な識別が行なえる。結局、話者識別に限らず、ergodic HMM によりカテゴリ識別を行なうためには、ベクトル単位にカテゴリ性を持たせることが必要であると考えられる。

今後の課題として、以下の事項が挙げられよう。

1. 今回は、出現頻度分布のクラスタリングを k-means アルゴリズムに基づいて行なったが、k-means アルゴリズムは全体の歪みを小さくする基準で行なわれるものであり、このことはベクトル量子化器の作成には有効であるが、クラスタリングの観点で最良であるかどうかは疑問である。クラスタリング手法としては k-means

アルゴリズム以外にも種々の手法が提案されており [6]、それらの適用についても検討する価値があると思われる。

2. Universal 符号帳とその出現頻度による解法における、distortion を基にした話者数 N の推定法については、さらに検証実験を進めて有効性を確認する必要がある。

3. ergodic HMM による解法において、

   (a) VQ コードサイズについて検討を行なう必要がある。サイズが大き過ぎると、異なる話者で共有するコードが多くなり、Viterbi アルゴリズムが誤った状態遷移を推定する可能性が大きくなる。逆にサイズが小さ過ぎると、各フレームが話者性を表現しずらくなる。即ち、最適なコードサイズが存在するものと思われる。

   (b) 話者性を特徴付けるような音響特徴量の抽出について検討する必要がある。今回は、フレーム長を長くすることにより話者性を特徴付けたが、他にも良い方法があると思われる。その一つとして、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が考えられる。この方法は、今回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。

   (c) Fuzzy VQ の導入や連続型 HMM の適用について検討することも重要である。

   (d) Baum-Welch アルゴリズムにおける収束計算の初期値設定について考察する必要がある。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、所望のカテゴリ識別を得るための初期値設定に関する理論的考察が重要である。

   (e) 通常の単純マルコフモデルでは、自己遷移と他状態への遷移を区別するような表現力が弱いと思われる。よって、2重マルコフ過程を考慮した second-order HMM の考察も重要であると考えられる。

   (f) 尤度を基準にした話者数推定法について検討する必要がある。

4. 検証実験の観点から、今回扱ったデータに問題があると思われる。まず、異なる話者で同じ単語が使われているのは問題である。話者識別の検証を行なうためには、異なる話者で異なる単語を用いる必要がある。また、今回扱った信号系列では、話者の遷移が ergodic の形式ではない。すなわち、例えば男性2名・女性2名の会話では、MAU から MNM の遷移が存在しない。よって、ergodic HMM の検証を行なうためには、全ての話者遷移が起こるようなデータを用いなければならない。

5. 単語間の無音区間を取り除いた実験を行なう必要がある。

6. 音声特徴量として、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、識別率が向上すると考えられ、これについての検討も必要である。

7. 特に ergodic HMM の場合や話者数が多くなる場合において、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れる。

8. N 信号源分解問題の他の応用例について検討する必要がある。

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、貴重な御助言、御検討を頂いた音声情報処理研究室 嵯峨山茂樹研究室長、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] X.D. Huang, Y. Ariki and M.A. Jack : "Hidden Marcov Models for Speech Recognition," Edinburgh University Press, Edinburgh (1990).

[2] 中川聖一："確率モデルによる音声認識", 電子情報通信学会 (1988).

[3] 嵯峨山茂樹："数理統計モデルによる音声認識の現状と将来", 音響学会誌, 48, pp.26-32 (1992).

[4] 石毛俊一, 佐野浩司, 間野一則, 白井克彦："スペクトルの量子化分布による話者認識", 信学技報 SP86-69, pp.59-64 (1986).

[5] M. Sugiyama : "Automatic Language Recognition Using Acoustic Features," Proc. ICASSP91, pp.813-816 (1991).

[6] 長尾真："パターン情報処理", 電子情報通信学会編, コロナ社 (1983).

[7] 堀部安一："情報エントロピー論", 森北出版 (1989).

[8] 坂元慶行, 石黒真木夫, 北川源四郎："情報量統計学", 情報科学講座 A.5.4, 共立出版 (1983).

[9] 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).

[10] 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).

[11] 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).

[12] 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).

[13] 中溝高好："信号解析とシステム同定", コロナ社 (1988).

[14] 森田啓嚢："算術符号から MDL 基準へ", 数理科学, No.290, Aug., pp.25-31 (1987).

[15] J. Rissanen : "A Universal Prior for Integers and Estimation by Minimum Description Length," Ann. Stat., vol.11, pp.416-431 (1983).

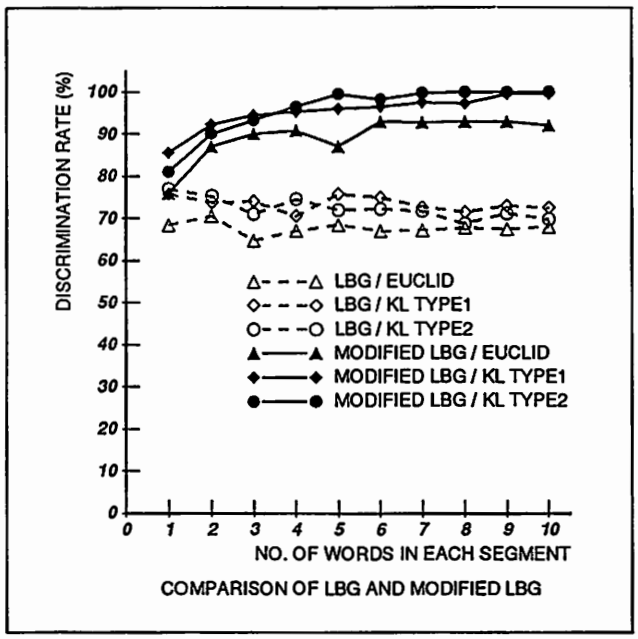

図 8: LBG アルゴリズムと変形 LBG アルゴリズムの比較（男性 2 名・女性 2 名）

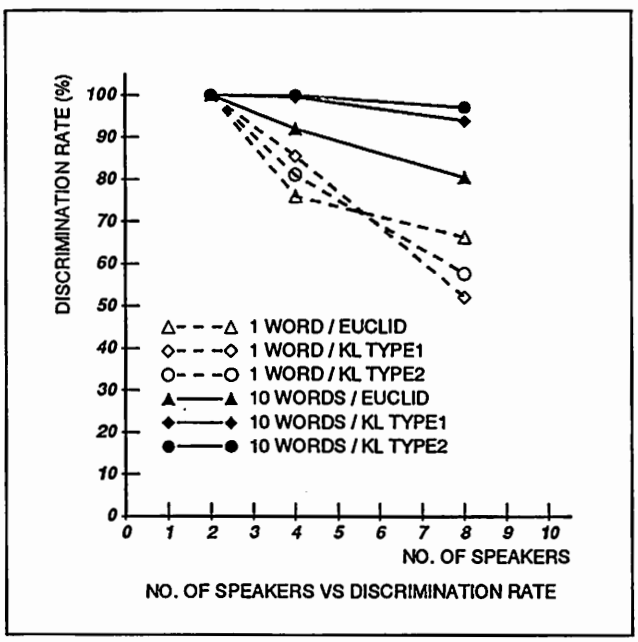

図 9: 出現頻度による解法における話者数と識別率との 関係（男女同数）

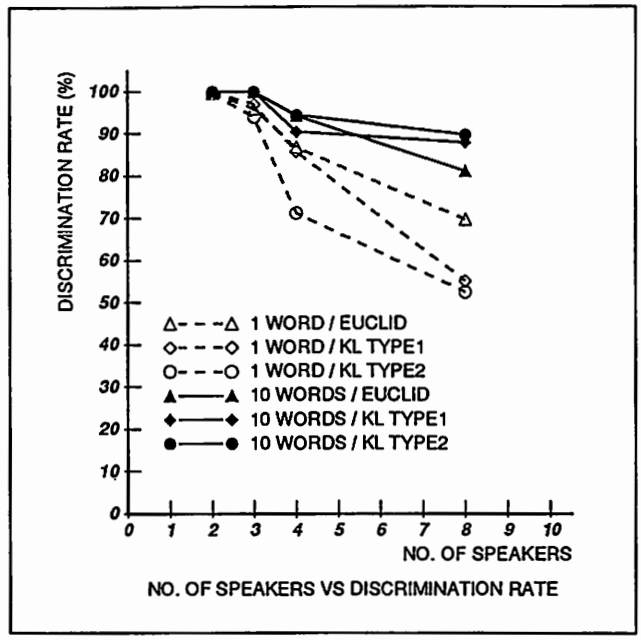

図 10: 出現頻度による解法における話者数と識別率との 関係（男性のみ）

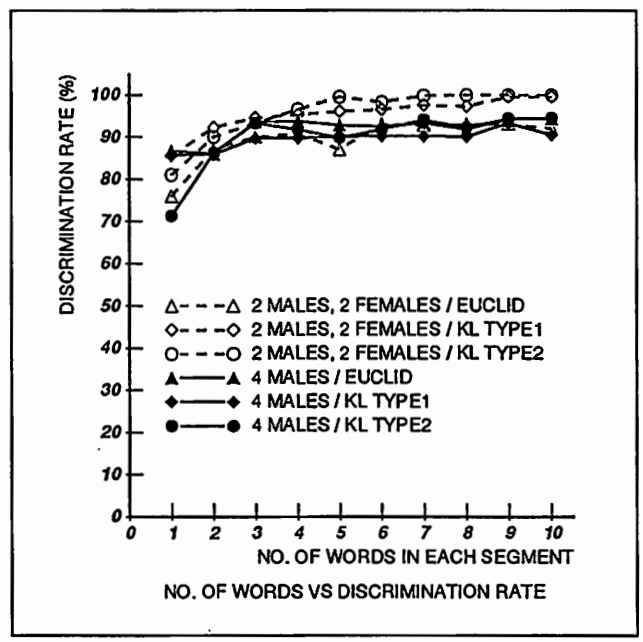

図 11: 出現頻度による解法における単語数と識別率との 関係（話者 4 人）

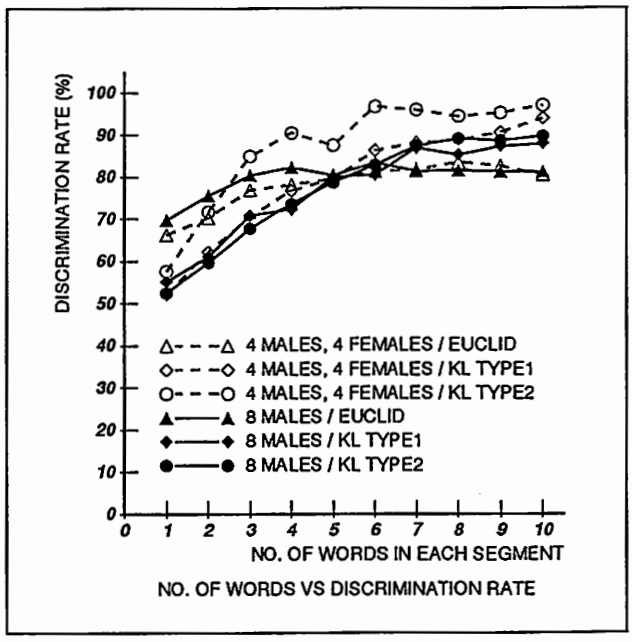

図 12: 出現頻度による解法における単語数と識別率との 関係（話者 8 人）

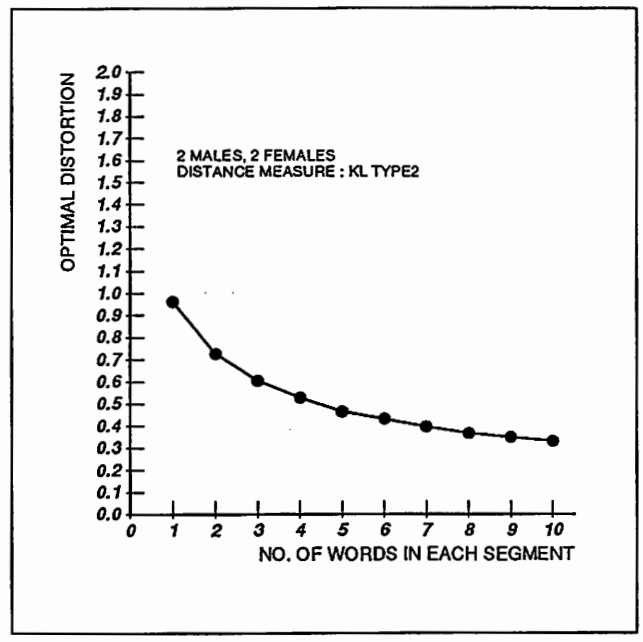

図 14: 最適 distortion と単語数との関係 （男性 2 名・女性 2 名、尺度： KL 情報量 type2）

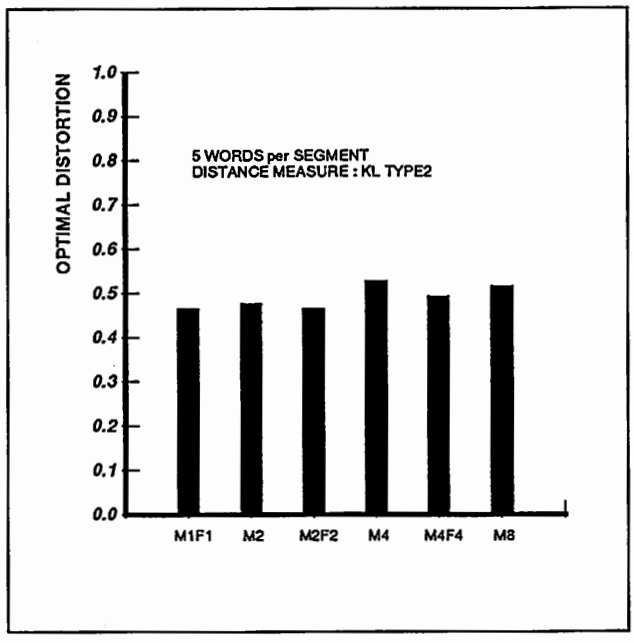

図 13: 最適 distortion と話者との関係（5 単語、尺度： KL 情報量 type2；左から、男 1 女 1、男 2、男 2 女 2、男 4、男 4 女 4、男 8）

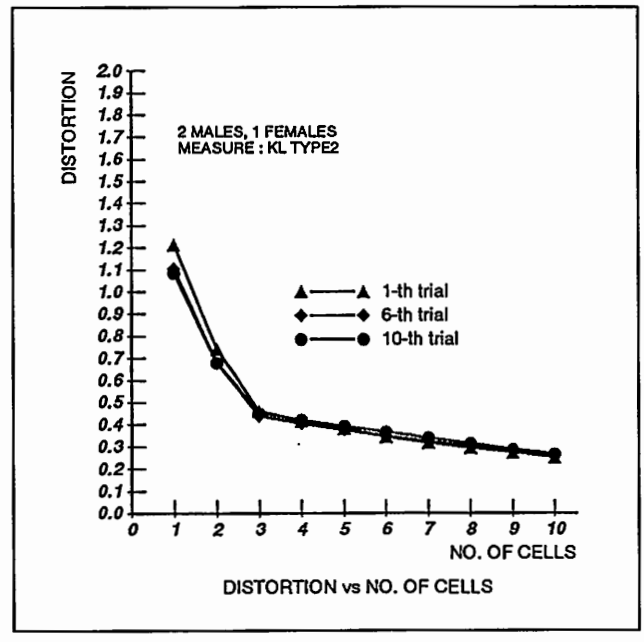

図 15: distortion とセル個数との関係 （男性 2 名・女性 1 名、尺度： KL 情報量 type2）

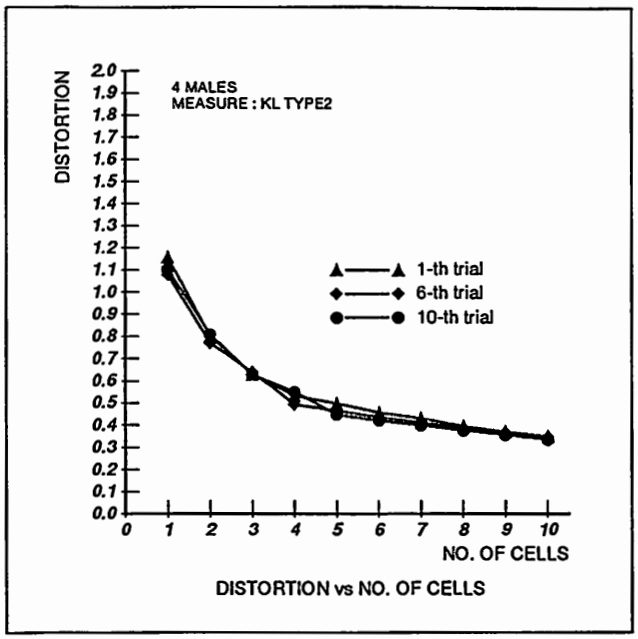

図 16: distortion とセル個数との関係 （男性 4 名、尺度：KL 情報量 type2）

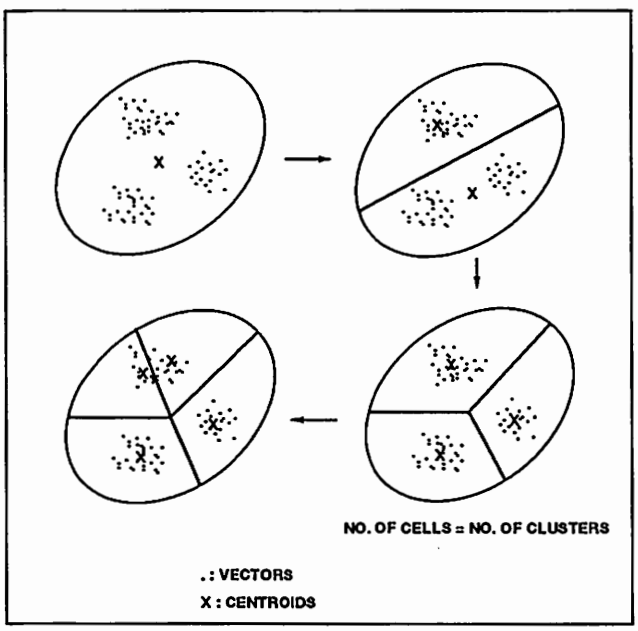

図 17: セル個数の増加による、ベクトル群の各 centroid からの ばらつきの変化（クラスタ形成が明確な場合）

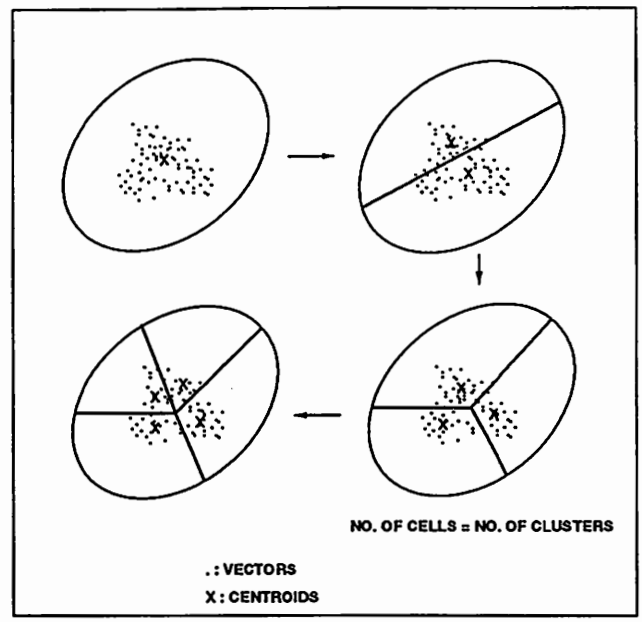

図 18: セル個数の増加による、ベクトル群の各 centroid からの ばらつきの変化（クラスタ形成が不明確な場合）

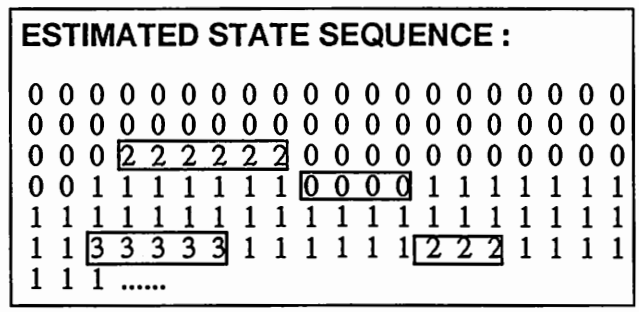

図 19: Viterbi アルゴリズムによる最適状態遷移系列の 推定結果の例（四角で囲んだ部分に他の状態が割り込んでいる）

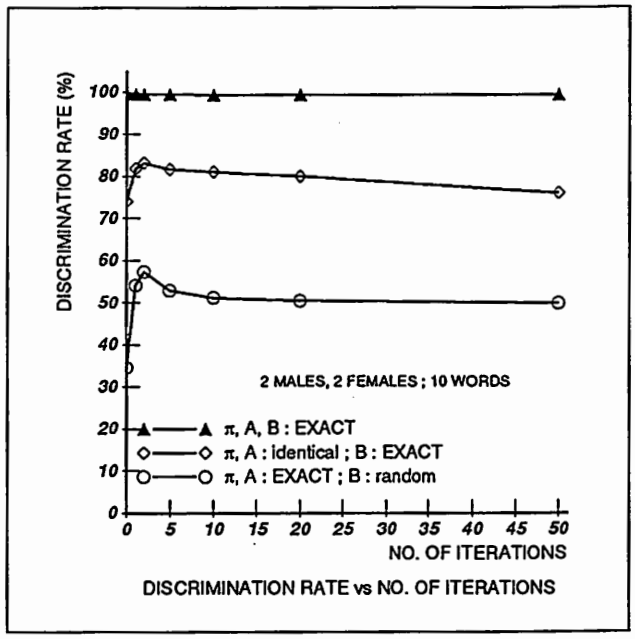

図 20: イテレーション回数と識別率との関係 （男性 2 名・女性 2 名、 10 単語）

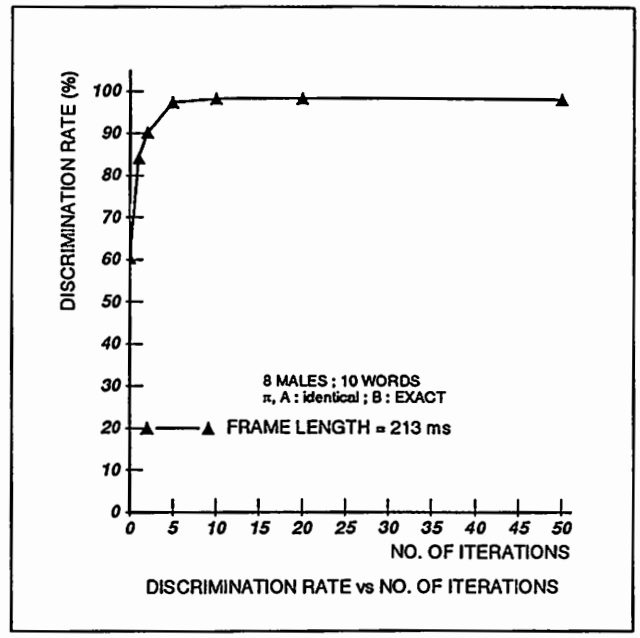

図 22: イテレーション回数と識別率との関係 （男性 8 名、 10 単語）

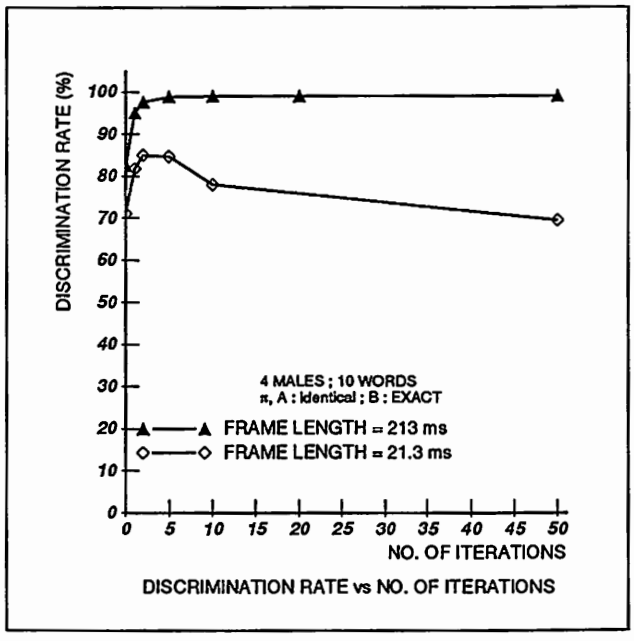

図 21: イテレーション回数と識別率との関係 （男性 4 名、 10 単語）

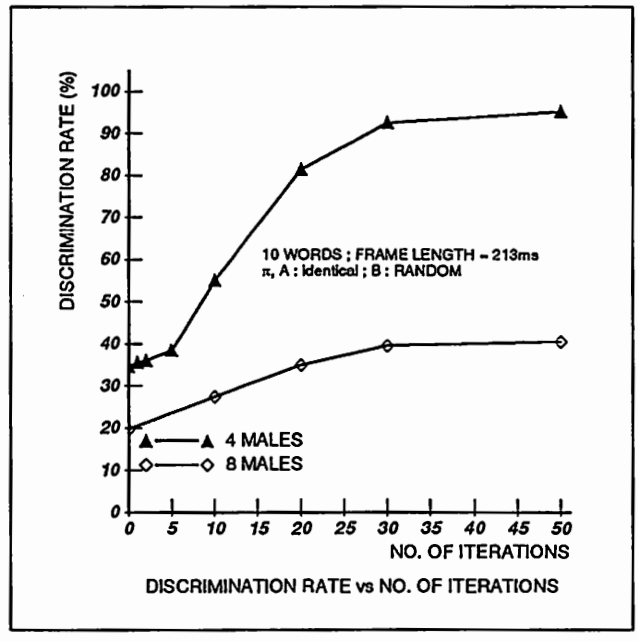

図 23: イテレーション回数と識別率との関係 （10 単語）

表 6: k-means アルゴリズム

1.	(初期設定) centroid 初期値 $z_j^{(0)}$ $(j=1,2,...,N)$ を与える。$m=0$ とする。
2.	(クラスタ更新) 各セグメント $k=1,2,...,K$ に対し、全ての $i=1,2,...,L$ について $d(q_k,z_j^{(m)}) < d(q_k,z_i^{(m)})$ $(j=1...N, i \neq j)$ ならば、$q_k \in C_j^{(m+1)}$ とする。ここで $d(a,b)$ は、a、b 間の距離 (歪み) 尺度を表す。
3.	(centroid 更新) 各セル $C_j^{(m+1)}$ の centroid $z_j^{(m+1)}$ $(j=1,2,...,N)$ を計算する。 ここで $z_j^{(m+1)}$ は次式で与えられる。 $z_j^{(m+1)} = \arg\min_z D_j(z)$ $D_j(z) = (1/K_j^{(m+1)}) \sum_{q_k \in C_j^{(m+1)}} d(q_k,z)$ (j 番目セルの distortion)
4.	$D=(1/K)\sum_{j=1}^{N} \sum_{q_k \in C_j^{(m+1)}} d(q_k,z_j^{(m+1)})$ (全体の平均 distortion) の減少度が与えられたしきい値より小さければ stop、 さもなくば $m=m+1$ として step2 へ行く。

表 7: LBG アルゴリズム

1.	(初期設定) 全ての q_k $(k=1,2,...,K)$ の centroid z_1 を計算し、$M=1$ とする。 ここで z_1 は次式で与えられる。 $z_1 = \arg\min_z \sum_{k=1}^{K} d(q_k,z)$
2.	$M=N$ ならば stop。
3.	各 z_j $(j=1,2,...,M)$ を z_j と z_{M+j} に splitting し、$M=2M$ とする。
4.	z_j $(j=1,2,...,M)$ を初期値として k-means アルゴリズムを実行する。 求まった centroid を改めて z_j $(j=1,2,...,M)$ とする。
5.	2 へ行く。

表 8: 変形 LBG アルゴリズム

1.	(初期設定) 全ての q_k $(k=1,2,...,K)$ の centroid z_1 を計算し、$M=1$ とする。
2.	$M=N$ ならば stop。
3.	distortion 最大のセルの centroid z_{j^*} を z_{j^*} と z_{M+1} に splitting し、$M=M+1$ とする。
4.	z_j $(j=1,2,...,M)$ を初期値として k-means アルゴリズムを実行する。 求まった centroid を改めて z_j $(j=1,2,...,M)$ とする。
5.	2 へ行く。

表 9: 識別評価方法（話者 3 人・10 系列の場合）

真の話者	A	A	A	B	B	B	C	C	C	C	
推定系列	1	1	0	0	2	0	0	2	2	2	
A=0,B=1,	0	0	0	1	1	1	2	2	2	2	
C=2	×	×	○	×	×	×	×	○	○	○	4/10 = 40 percent
A=0,B=2,	0	0	0	2	2	2	1	1	1	1	
C=1	×	×	○	×	○	×	×	×	×	×	2/10 = 20 percent
A=1,B=0,	1	1	1	0	0	0	2	2	2	2	
C=2	○	○	×	○	×	○	×	○	○	○	7/10 = 70 percent
A=1,B=2,	1	1	1	2	2	2	0	0	0	0	
C=0	○	○	×	×	○	×	○	×	×	×	4/10 = 40 percent
A=2,B=0,	2	2	2	0	0	0	1	1	1	1	
C=1	×	×	×	○	×	○	×	×	×	×	2/10 = 20 percent
A=2,B=1,	2	2	2	1	1	1	0	0	0	0	
C=0	×	×	×	×	×	×	○	×	×	×	1/10 = 10 percent
											⇓ 最大値= 70percent =識別率

6 付録

A 量子化頻度分布のクラスタリングにおける centroid 計算

k-means アルゴリズムにおいて、j 番目セルの centroid は、次式の最小化問題を解くことにより求まる。

$$minimize_{z_j} \sum_{q_k \in C_j} d(q_k, z_j)$$

$$condition : \sum_{l=1}^{L} z_{jl} = 1 \tag{10}$$

ただし

$$q_k = [P_{k1}\ P_{k2}\ ...\ P_{kL}]^T$$

$$z_j = [z_{j1}\ z_{j2}\ ...\ z_{jL}]^T$$

である。これは次に示すラグランジェの未定乗数法により解かれる。

$$minimize_{z_j,\lambda} J(z_j, \lambda)$$

$$J(z_j, \lambda) = \sum_{q_k \in C_j} d(q_k, z_j) - \lambda(\sum_{l=1}^{L} z_{jl} - 1) \tag{11}$$

A.1 Euclid 距離の場合

J は次式で表される。ただし、Euclid 距離の二乗を用いている。

$$J(z_j, \lambda) = \sum_{q_k \in C_j} (q_k - z_j)^T (q_k - z_j) - \lambda(\sum_{l=1}^{L} z_{jl} - 1) \tag{12}$$

J の z_{jl} による偏微分は

$$\begin{aligned} \frac{\partial J}{\partial z_{jl}} &= \sum_{q_k \in C_j} (-2P_{kl} + 2z_{jl}) - \lambda \\ &= -2 \sum_{q_k \in C_j} P_{kl} + 2K_j z_{jl} - \lambda \end{aligned} \tag{13}$$

となり、上式をゼロにする z_{jl} は次式で与えられる。

$$z^o_{jl} = \frac{1}{K_j} \sum_{q_k \in C_j} P_{kl} + \frac{\lambda}{2K_j} \tag{14}$$

また、$\partial J/\partial \lambda = 0$ より $\sum_{l=1}^{L} z^o_{jl} = 1$ となり、上式の両辺を $\sum_{l=1}^{L}$ で総和をとることにより、$\lambda = 0$ が得られる。よって centroid は、セルに属するベクトルの相加平均で与えられる。

A.2 KL 情報量 type1 の場合

J は次式で表される。

$$\begin{aligned} J(z_j, \lambda) &= \sum_{q_k \in C_j} \sum_{l=1}^{L} z_{jl} \log \frac{z_{jl}}{P_{kl}} - \lambda(\sum_{l=1}^{L} z_{jl} - 1) \\ &= K_j \sum_{l=1}^{L} z_{jl} \log z_{jl} - \sum_{l=1}^{L} z_{jl} \sum_{q_k \in C_j} \log P_{kl} \\ &- \lambda(\sum_{l=1}^{L} z_{jl} - 1) \end{aligned} \tag{15}$$

J の z_{jl} による偏微分は

$$\frac{\partial J}{\partial z_{jl}} = K_j \log z_{jl} + K_j - \sum_{q_k \in C_j} \log P_{kl} - \lambda \tag{16}$$

であり、上式をゼロにする z_{jl} は、

$$\begin{aligned} \log z^o_{jl} &= \frac{1}{K_j} \sum_{q_k \in C_j} \log P_{kl} + \frac{\lambda}{K_j} - 1 \\ &= \log \left[\prod_{q_k \in C_j} P_{kl} \right]^{\frac{1}{K_j}} + \log e^{\lambda/K_j - 1} \end{aligned} \tag{17}$$

となることにより、次式で与えられる。

$$z^o_{jl} = Q_{jl} \cdot \mu \tag{18}$$

$$Q_{jl} = \left[\prod_{q_k \in C_j} P_{kl} \right]^{\frac{1}{K_j}}$$

$$\mu = e^{\lambda/K_j - 1}$$

また、$\sum_{l=1}^{L} z^o_{jl} = 1$ となることから、上式の両辺について総和をとり、

$$\mu = \frac{1}{\sum_{l=1}^{L} Q_{jl}} \tag{19}$$

が得られる。よって centroid は次式で与えられる。

$$z^o_{jl} = \frac{Q_{jl}}{\sum_{l=1}^{L} Q_{jl}} \tag{20}$$

A.3 KL 情報量 type2 の場合

J は次式で表される。

$$J(z_j, \lambda) = \sum_{q_k \in C_j} \sum_{l=1}^{L} P_{kl} \log \frac{P_{kl}}{z_{jl}} - \lambda(\sum_{l=1}^{L} z_{jl} - 1)$$

$$= \sum_{q_k \in C_j} \sum_{l=1}^{L} P_{kl} \log P_{kl}$$
$$- \sum_{l=1}^{L} \left(\sum_{q_k \in C_j} P_{kl} \right) \log z_{jl}$$
$$- \lambda (\sum_{l=1}^{L} z_{jl} - 1) \quad (21)$$

J の z_{jl} による偏微分は

$$\frac{\partial J}{\partial z_{jl}} = -\frac{\sum_{q_k \in C_j} P_{kl}}{z_{jl}} - \lambda \quad (22)$$

であり、上式をゼロにする z_{jl} は、

$$z^o_{jl} = -\frac{\sum_{q_k \in C_j} P_{kl}}{\lambda} \quad (23)$$

となる。また、$\sum_{l=1}^{L} z^o_{jl} = 1$ となることから、上式の両辺について総和をとり、

$$\lambda = -\sum_{q_k \in C_j} \sum_{l=1}^{L} P_{kl} = -K_j \quad (24)$$

が得られる。故に、centroid は次式で与えられる。

$$z^o_{jl} = \frac{1}{K_j} \sum_{q_k \in C_j} P_{kl} \quad (25)$$

よって、Euclid 距離の場合と同様、centroid はベクトルの相加平均で与えられる。

B 離散型 HMM の場合の Baum-Welch アルゴリズム

B.1 基本的手続き

HMM パラメータ M の最尤推定値は、次式で与えられる。

$$\hat{M} = \arg\max_{M} P(O \mid M) \quad (26)$$

ここで $P(O \mid M)$ は、入力信号の VQ コード系列

$$O = \{o_1 o_2 ... o_T\}$$

の尤度であり、次式で与えられる。

$$P(O \mid M) = \sum_{allS} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(o_t) \quad (27)$$
$$S = \{s_1 ... s_T\}$$
$$s_t \in \{1, 2, ..., N\} \quad (t = 1, 2, ..., T)$$
$$o_t \in \{u_1, ..., u_L\} \quad (t = 1, 2, ..., T)$$
$$\pi_{s_1} = a_{s_0 s_1}$$

Baum-Welch アルゴリズムによるパラメータ M の最尤推定は、状態遷移系列 S が非観測データであるとして EM アルゴリズムを適用することにより行なわれる。結果として [1]Baum-Welch アルゴリズムは、次に示される繰り返し手続きにより行なわれる。

Baum-Welch アルゴリズム：

1. パラメータ $M = (\pi, A, B)$ の初期値 $M^{(0)}$ を与える。$m = 0$ とおく。
2. 次式で与えられる α、β 両パラメータを計算する（後述）。

$$\alpha_t(i) = P(o_1 ... o_t, s_t = i \mid M^{(m)}) \quad (28)$$
$$\beta_t(i) = P(o_{t+1} ... o_T \mid s_t = i, M^{(m)}) \quad (29)$$
$$(t = 1...T,\ i = 1...N)$$

3. 次式により、M を更新する。

$$\pi_i^{(m+1)} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{N} \alpha_T(j)} \quad (30)$$

$$a_{ij}^{(m+1)} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij}^{(m)} b_j(o_{t+1})^{(m)} \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \quad (31)$$

$$b_j(l)^{(m+1)} = \frac{\sum_{t:o_t=u_l} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \quad (32)$$

4. $P(O \mid M^{(m)}) = \sum_{j=1}^{N} \alpha_T(j)$ の変化があるしきい値より小さくなればストップ、さもなくば $m = m + 1$ として 2. へ行く。

上の 2. における α 及び β は、それぞれ次に示す Forward アルゴリズム及び Backword アルゴリズムで効率良く計算される。

Forward アルゴリズム：

1. 全ての $i \in \{1...N\}$ に対して、$\alpha_1(i) = \pi_i b_i(o_1)$ とする。
2. 時間軸 $(t = 2, ..., T)$ に沿って、全ての $j \in \{1...N\}$ に対し、次式により $\alpha()$ を計算する。
   $\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t)$

Backward アルゴリズム：

1. 全ての $i \in \{1...N\}$ に対して、$\beta_T(i) = 1$ とする。
2. 時間軸 $(t = T-1, ..., 1)$ に沿って、全ての $j \in \{1...N\}$ に対し、次式により $\beta()$ を計算する。
   $\beta_t(j) = \sum_{i=1}^{N} a_{ji} b_i(o_{t+1}) \beta_{t+1}(i)$

B.2 スケーリングを行なう場合の手続き

スケーリングは、全ての $t \in \{1...T\}$ において α、β 両パラメータがアンダーフローを起こさないよう、それらに

適当な係数を掛けることにより行なわれる。スケーリングを行なうForward、Backwardアルゴリズムは、以下のようになる。

Forwardアルゴリズム：

1. 全ての$i \in \{1...N\}$に対して、 $\alpha^*{}_1(i) = \pi_i b_i(o_1)$ とする。
2. 全ての$i \in \{1...N\}$に対して、 $\alpha'_1(i) = c_1 \alpha^*{}_1(i)$。
3. 時間軸$(t = 2, ..., T)$に沿って、全ての$j \in \{1...N\}$に対し、次式により$\alpha'()$を計算する。
   $\alpha^*{}_t(j) = \sum_{i=1}^{N} \alpha'_{t-1}(i) a_{ij} b_j(o_t)$
   $\alpha'_t(i) = c_t \alpha^*{}_t(i)$

Backwardアルゴリズム：

1. 全ての$i \in \{1...N\}$に対して、 $\beta^*{}_T(i) = 1$ とする。
2. 全ての$i \in \{1...N\}$に対して、 $\beta'_T(i) = c_T \beta^*{}_T(i)$。
3. 時間軸$(t = T-1, ..., 1)$に沿って、全ての$j \in \{1...N\}$に対し、次式により$\beta'()$を計算する。
   $\beta^*{}_t(j) = \sum_{i=1}^{N} a_{ji} b_i(o_{t+1}) \beta'_{t+1}(i)$
   $\beta'_t(i) = c_t \beta^*{}_t(i)$

ここでt時刻目のスケーリング係数c_tを次式で与える。

$$c_t = \left[\sum_{i=1}^{N} \alpha^*{}_t(i) \right]^{-1} \tag{33}$$

これにより、全時刻において$\sum_i \alpha'_t(i) = 1$となり、アンダーフローが回避される。

スケーリングを行なったα'、β'とスケーリングをしないα、βとの間に次式の関係がある。

$$\alpha'_t(i) = C_t \alpha_t(i) \tag{34}$$

$$\beta'_t(i) = D_t \beta_t(i) \tag{35}$$

$$C_t = \prod_{\tau=1}^{t} c_\tau \tag{36}$$

$$D_t = \prod_{\tau=t}^{T} c_\tau \tag{37}$$

これにより、スケーリングを行なうBaum-WelchアルゴリズムにおけるMの更新式と$P(O \mid M)$の計算式は次式となる。

$$\pi_i^{(m+1)} = \frac{1}{c_1} \frac{\alpha'_1(i)\beta'_1(i)}{\sum_{j=1}^{N} \alpha'_T(j)} \tag{38}$$

$$a_{ij}^{(m+1)} = \frac{\sum_{t=1}^{T-1} \alpha'_t(i) a_{ij}^{(m)} b_j(o_{t+1})^{(m)} \beta'_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha'_t(i)\beta'_t(i)/c_t} \tag{39}$$

$$b_j(l)^{(m+1)} = \frac{\sum_{t:o_t=u_l} \alpha'_t(j)\beta'_t(j)/c_t}{\sum_{t=1}^{T} \alpha'_t(j)\beta'_t(j)/c_t} \tag{40}$$

$$P(O \mid M) = \frac{1}{C_T} \sum_{i=1}^{N} \alpha'_t(i) \tag{41}$$

$$C_T = \prod_{t=1}^{T} c_t$$

（文献[1]では、Mの更新式がスケーリングなしの場合とありの場合とで変わらないと述べられているが、正確には上式で示されるように両者間で異なる。これは、$\alpha'_t(i)\beta'_t(i) = c_t \prod_{t=1}^{T} \alpha_t(i)\beta_t(i)$ となることより容易にわかる。）

C Viterbiアルゴリズム

推定されたHMMMがコード系列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。これは動的計画法を適用するものであり、対数演算によるViterbi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、
   $\delta_1(i) = \log \pi_i + \log b_i(o_1)$
   $\phi_1(i) = 0$
   とおく。
2. 時間軸$t = 2...T$に沿って、全ての$j \in \{1...N\}$に対し
   $\delta_t(j) = \max_i[\delta_{t-1}(i) + \log a_{ij} + \log b_j(o_t)]$
   $\phi_t(j) = \arg\max_i[\delta_{t-1}(i) + \log a_{ij}]$。
3. 最適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。
   $\max_S P(O, S \mid M) = \max_j \delta_T(j)$
   $s_T^* = \arg\max_j \delta_T(j)$
4. 時間軸$t = T - 1...1$に沿って、次式により最適状態遷移系列を得る。
   $s_t^* = \phi_{t+1}(s_{t+1}^*)$

求まった$S^* = \{s_1^* \ ... \ s_T^*\}$から、segmentationとカテゴリ識別の同時推定が得られる。

D 情報量基準の適用によるNの推定に関する一考察

データをモデリングする際に、最適なパラメータ数を客観的に与える基準として、AIC及びMDLの情報量基準が提案されている[13, 14, 15]。これらは、それぞれ次式を最小にするkを最適なパラメータ数とするものである。

$$AIC(k) = -2 \log f(X \mid \hat{\Theta}^{(k)}) + 2k \tag{42}$$

$$MDL(k) = -\log f(X \mid \hat{\Theta}^{(k)}) + \frac{1}{2} k \log T \tag{43}$$

ここで $f(X \mid \hat{\Theta}^{(k)})$ はパラメータ数 k のモデルによるデータ X の最大尤度、T はデータ長である。

ergodic HMM による話者数（状態数）N の推定問題に、これらの情報量基準を適用することを考える。最大尤度は、Baum-Welch アルゴリズムで求まった M を基に Forward アルゴリズムを行なうことにより求まる。また、初期状態確率が自由度 N、状態遷移確率が N^2、シンボル出力確率が NL であることより、モデル全体の自由度は計 $N(1+N+L)$ となる（ただし L はコードサイズ）。よって、AIC、MDL を適用した話者数の推定値は、次式を最小にする N で与えられる。

$$AIC_{HMM}(N) = -2\log P(O \mid \hat{M}^{(N)}) + 2N(1+N+L) \tag{44}$$

$$MDL_{HMM}(N) = -\log P(O \mid \hat{M}^{(N)}) + \frac{1}{2}N(1+N+L)\log T \tag{45}$$

ここで、$\hat{M}^{(N)}$ は状態数 N の ergodic HMM における シンボル系列 O の最尤推定値である。一般に N 信号源分解問題では、系列長 T がかなり大きくなるため、対数尤度の値が極端に小さくなる（即ち、絶対値が極度に大きくなる）ことが予想される。よって、評価式第 2 項に T を有している MDL の方が AIC に比べ N の良好な推定を与えると思われる。

実験で使用したソフトウェア

～/research/CorrCep/run.CorrCep：波形から相関列を作り、その相関列からLPCケプストラム列を作るためのシェルスクリプト（C-SHELL）。相関列は一時的に～/research/WaveCorr/cor_seq/の下に書き込まれ、ケプストラム列は～/research/CorrCep/cep_seq/の下に保存される。ディスク容量の関係で、相関列はすぐにdeleteされるようになっている。この方法は明らかに効率が悪いので、処理速度を早めるためには改良が必要であろう。

～/research/WaveCorr/WaveCorr.c：波形から相関列を作成するためのソースプログラム（C言語）。

～/research/WaveCorr/ham2.c：ハミング窓のサブプログラム（窓長最大4096サンプル）。

～/research/CorrCep/CorrCep.c：相関列からケプストラム列を作成するためのソースプログラム。

～/research/VQ/run.make_codebook：VQコードブック作成のためのシェルスクリプト。コードブックは、～/research/VQ/Codebook/の下に書き込まれる。

～/research/VQ/src/make_codebook.c：コードブック作成用ソースプログラム。サブプログラムとして、centroid_cep_ecld.c（セントロイド計算）、label_cep_ecld.c（ラベリング）、split_cep_ecld.c（LBGアルゴリズムの際のsplitting）（3者共、尺度はEuclid距離）を同ディレクトリに持つ。

～/research/QFD/run.qfd.cluster：単語名リスト（./name_seq/の下）を使って入力ケプストラム系列を読み込み、各セグメントのコード出現頻度分布を算出し（qfd.c）、それらをクラスタリングする（cluster.c）ためのシェルスクリプト。

～/research/QFD/name_seq/：入力信号の単語名リストのファイルが保管されているディレクトリ。例えば、MAUFKNMNMFKS.40.10.seq0は、MAU,FKN,MNM,FKSの順の発話で、全部で40セグメント（発話）、各セグメントが10単語、0番目試行用データ、の意味である。8人以上の場合は、ファイル名が長くなるので、MFMFMFMF.80.10.seq0、MMMMMMMM.80.10.seq0等としている。

～/research/QFD/qfd.c：各セグメントのコード出現頻度分布（Quantization Frequency Distribution）を計算するためのソースプログラム。上に示した単語名リストファイルから単語名を一つ一つ読み込み（catで）、ケプストラム系列を読み込む。各セグメント毎にvoid型関数qfd()で出現頻度を計算する。計算結果は./qfd_result/の下に保存される。

～/research/QFD/cluster.c：頻度分布をクラスタリングする。変数measure_typeは、0:euclid、1:KL type1、2:KL type2である。以下をサブプログラムとして持つ。

- ～/research/QFD/centroid_pr.c：k-meansアルゴリズムにおけるcentroid計算。
- ～/research/QFD/label_pr.c：k-meansアルゴリズムにおけるラベリング（セル形成）。
- ～/research/QFD/split_pr.c：LBGアルゴリズムにおけるsplitting。
- ～/research/QFD/split_pr_maxdm.c：変形LBGアルゴリズムにおけるsplitting（distortion最大のセルのみsplitting）。

～/research/QFD/cluster_result/：クラスタリングの結果が保存されるディレクトリ。ASCII形式で書き込まれる。例えば、MAUFKNMNMFKS.40.10.clstr.2.0は、MAU,FKN,MNM,FKSの順の発話で、全部で40セグメント（発話）、各セグメントが10単語の場合のクラスタリング結果で、適用した尺度はKL type2、データは0番目試行用、の意味である。

～/research/QFD/foo.c：上記のディレクトリに書き込まれたクラスタリング結果を読み込み、識別評価をするためのプログラム。順列をリストするルーチンが含まれている。

～/research/HMM/MAIN/run.discrete_HMM：HMMの実験用のシェルスクリプト（VQ、Baum-Welch、Viterbi、識別評価を含む）。

～/research/HMM/VQ/vq.c：入力ケプストラム系列をラベリング（量子化）するためのプログラム（結果は、～/research/HMM/BW/dataにASCII形式で書き込まれる）。同時に、各話者の単語発声終了時刻（フレーム）を、～/research/HMM/Evaluate/word_frame.outに書き込む（識別評価の際に参照される）。

～ /research/HMM/VQ/b_init_qfd.c : シンボル出力確率の初期値を、出現頻度分布のクラスタリングの結果を基に算出する。クラスタリングの結果は ～ /research/qfd/qfd_result/ から読み込む。算出結果は ～ /research/HMM/BW/init_sym に binary 形式で書き込まれる。

～ /research/HMM/VQ/init_exact.c : HMM のパラメータ初期値として、全て真値があたえられる。算出結果は、$\pi^{(0)}$ は～ /research/HMM/BW/init_for に、$A^{(0)}$ は～ /research/HMM/BW/init_trs に、$B^{(0)}$ は ～ /research/HMM/BW/init_sym にそれぞれ binary 形式で書き込まれる。

～ /research/HMM/BW/main.c : Baum-Welch アルゴリズムのメインルーチン。入力変数は、NN : 状態数、max_iteration : 繰り返し回数の上限、converged : 収束判定の基準値（対数尤度の差で判定する）、init_type : パラメータ初期値のタイプ — 'eee' : $\pi^{(0)}$, $A^{(0)}$, $B^{(0)}$ 全て真値、'tte' : $\pi^{(0)}$, $A^{(0)}$ が等確率 (Toukaku-ritsu) で $B^{(0)}$ が真値、'ttr' : $\pi^{(0)}$, $A^{(0)}$ が等確率で $B^{(0)}$ がランダム (Random) 。π 、A 、B の収束値は、それぞれ./test.pi、./test.a、./test.b に binary 形式で書き込まれる。

～ /research/HMM/BW/init_value.c : 引数 init_type[0] = 't' ならば、$\pi^{(0)}$ として等確率の値が ./init_for に上書きされる。さもなくば、init_for には何も上書きされない。引数 init_type[1] = 't' ならば、$A^{(0)}$ として等確率の値が ./init_trs に上書きされる。さもなくば、init_trs には何も上書きされない。引数 init_type[2] = 'r' ならば、$B^{(0)}$ としてランダムな値が ./init_sym に上書きされる。さもなくば、init_sym には何も上書きされない。

～ /research/HMM/Viterbi/viterbi.c : ～ /research/HMM/BW/ の下の test.pi, test.a, test.b を読んで、Viterbi アルゴリズムを行なう。結果は標準出力に出される。

～ /research/HMM/Evaluate/foo.c : viterbi の結果と word_frame.out を読み込んで、識別評価を行なう。順列をリストするルーチンを含んでいる。