

概要

複数の人がそれぞれ自由に会話している環境で、それぞれの発話内容を認識するシステムが望まれている。複数の話者が同時に話した時に、各話者の音声の認識を行う際、複数のマイクロフォンを用いる手法が一般的である [1]。しかし、マイクロフォンの数を超える数の音声が発声された場合、認識精度が低下する。

過去の研究では、男女2話者が別々の単語を同時に発声した場合に、単一のマイクロフォンにより認識することを想定し、同時発話認識率の調査が行われた [2]。

認識手法として単純法とマルチパス法が用いられたが、どちらも認識アルゴリズムがビタビ探索のため、重畳した音声に対し、フレーム単位ではどちらかの話者のHMMで計算することになり、論理的に問題がある。

そこで、本研究では、フレーム単位で両方の話者のHMMを計算するために、男女それぞれのHMMとフレームの3次元モデルにより、1つの状態に両方の話者HMMを用いた認識を行う手法を提案する。

結果は、特徴パラメータ FBANK における提案手法の認識率が 13.0% となり、マルチパス法の認識率 39.5% に比べ低くなった。

目次

1	はじめに	1
2	音声認識	3
2.1	音声の特徴抽出	4
2.2	特徴パラメータ	5
2.2.1	FBANK	5
2.2.2	MFCC	5
2.3	HMM(隠れマルコフモデル)	6
2.3.1	HMM を用いた音声認識	8
2.3.2	HMM 法の利点と問題点	9
2.4	HMM の種類	10
2.4.1	離散型 HMM	10
2.4.2	連続型 HMM	10
2.4.3	半連続型 HMM	11
2.5	認識アルゴリズム	12
2.5.1	Forward アルゴリズム (トレリス法)	13
2.5.2	Viterbi アルゴリズム	14
2.5.3	トレリス法と対数を用いたアルゴリズムの違い	15
2.5.4	対数の加算の計算方法	16
2.6	HMM の学習	17
2.6.1	Baum-Welch アルゴリズム	17
2.6.2	離散 HMM のパラメータ推定	19
2.6.3	連続 HMM のパラメータ推定法	19
2.6.4	出現確率が単一 (多次元) ガウス分布で表される場合	19
2.6.5	出現確率が混合ガウス分布で表される場合	19
2.6.6	半連続 HMM の場合	20
2.7	HMM の連結学習	21
3	クロストーク音声認識	22
3.1	従来の研究	22
3.2	認識手法	23

3.2.1	単純法	23
3.2.2	マルチパス法	24
3.2.3	PMC 法	25
3.3	提案手法	27
3.3.1	従来の認識手法の問題点	27
3.3.2	3次元ビタビ探索	27
3.3.3	本研究のモデル	28
3.3.4	男性 HMM と女性 HMM の加算	28
4	認識実験	29
4.1	実験の目的	29
4.2	実験データ	29
4.3	実験手順	30
4.4	HTK によるモデル作成	31
4.5	実験条件	33
5	実験結果	34
5.1	提案手法	34
5.1.1	同時発話認識率の比較	34
5.1.2	各話者認識率の比較	34
5.2	PMC 法	35
5.2.1	他の認識手法との比較	35
5.2.2	状態数混合分布数別	35
5.3	認識率の詳細	36
5.3.1	提案手法	36
5.3.2	マルチパス法	37
5.3.3	PMC 法	38
5.3.4	単純法	39
5.4	認識例	40
5.4.1	マルチパス法で認識失敗したが，提案手法で認識成功した例 . . .	40
5.4.2	マルチパス法で認識成功したが，提案手法で認識失敗した例 . . .	41
5.4.3	単純法で認識失敗したが，提案手法で認識できた例	42
5.4.4	単純法で認識できたが，提案手法で認識失敗した例	43

6	考察	44
6.1	提案手法の有効性	44
6.1.1	提案手法の精度	44
6.1.2	認識傾向	46
6.1.3	認識できた単語	47
6.2	PMC 法の有効性	47
7	おわりに	48

目 次

1	音声認識課程の確率モデル	3
2	left-to-right モデルの例	6
3	HMM を用いた単語音声認識の方法	12
4	トレリス上の $\alpha_t(i)$ の計算	14
5	Viterbi アルゴリズムの適用例	15
6	連結学習の例	21
7	単純法	23
8	単純法の認識の例	23
9	マルチパス法	24
10	マルチパス法の認識結果	25
11	HMM 合成法 (PMC) の原理	25
12	本研究における 3 次元ビタビ探索	27
13	提案する提案手法のモデル	28
14	スペクトル領域での加算 (lmixture)	28
15	クロストーク音声認識の流れ	30
16	状態数混合分布数による変化	34
17	認識手法の傾向	34
18	認識手法の比較	35
19	PMC 法の認識率	35
20	HMM のパスの考えられるパスの遷移 (3state)	45

表 目 次

1	実験に使用した単語	29
2	クロストーク音声 (悪質と重畳した例)	30
3	実験条件	33
4	認識率:提案手法 (FBANK,1state 1mixture)	36
5	認識率:提案手法 (FBANK,3state 4mixture)	36
6	認識率:提案手法 (MFCC)	36
7	認識率:マルチパス法 (FBANK,1state 1mixture)	37
8	認識率:マルチパス法 (FBANK,3state 4mixture)	37
9	認識率:マルチパス法 (MFCC)	37
10	パラメータ合成での認識 (状態数は1)	38
11	パラメータ合成での認識 (状態数は1, 母音, 撥音, 無音のみ 10mixture)	38
12	認識率:単純法 (FBANK,1state 1mixture)	39
13	認識率:単純法 (FBANK,3state 4mixture)	39
14	認識率:単純法 (MFCC)	39
15	認識結果 (例1)	40
16	認識結果 (例2)	41
17	認識結果 (例3)	42
18	認識結果 (例4)	43
19	認識の割合 (男性)	46
20	認識の割合 (女性)	46

1 はじめに

人々が同時に会話などをするような場面で、全ての話者の音声を聞き分けて、認識できるシステムの実現が望まれる。このようなシステムの基本として、2話者が同時に発声する状況を想定したクロストーク音声認識が挙げられる。クロストーク音声認識は技術的に困難な課題であるが、重要な技術のひとつであると考えている。過去の研究では、複数のマイクロフォンを用いる手法が一般的に用いられている [5]。しかし、話者よりもマイクロフォンの数が少ない時や、同一方向からの重畳された音声を認識する場合には対応できない。そこで、人間が1つの耳だけで複数の音声を聞き分けることができるように、単一のマイクロフォンを使った認識システムの改良が望まれている。

岡本らの研究では、男女2話者が別々の単語を同時に発声した場合に、単一のマイクロフォンを使用した状況を想定し、音声認識ツールの HTK[6] を用いた同時発話認識率の調査が行われた。同時発話認識率とは、クロストーク音声の男性話者と女性話者の発話内容を同時に認識できた場合の認識率である。計算機で重畳したクロストーク音声に対し、男女別々の HMM を用いて認識を行う単純法は、

56%となった。男性話者と女性話者それぞれの HMM を用いたマルチパス法は、45%となった。しかし、この2つの手法は、クロストーク音声において、男性話者の音声と女性話者の音声を個別に考える点から妥当性がない。クロストーク音声は男女それぞれの話者の音声を重畳したにも関わらず、時間によって男性話者、女性話者のどちらか一方の音声を認識するからである。

そこで、クロストーク音声認識に雑音対策の認識手法である Parallel Model Combination 法 (以下、PMC 法) が適用された [2]。この手法の HMM は、無雑音音声の HMM と雑音の HMM を合成して作成する。クロストーク音声認識に適用する場合、クロストーク音声の片側音声を雑音モデルと考える。PMC 法が男女それぞれの音声信号の和として認識を行うため、単純法、マルチパス法の問題点が解決できる。認識率として 10% という結果になった。この原因として、状態数と混合分布数を増やした場合に未対応だった点が挙げられる。

パラメータ合成とは、ケプストラム空間での雑音と音声を、スペクトル領域に変換し加算し元に戻すことである。雑音と音声の出現確率がケプストラム空間での正規分布であるが、雑音と音声は線形スペクトル領域で加算の関係にあるため、両者の分布をスペクトル領域での分布に変換し、加算する必要がある。本研究では、特徴パラメータに MFCC と FBANK を用いるが、MFCC では HMM の平均と分散はケプストラム領域で

あるのでパラメータ合成をする必要があるが、FBANK では HMM の平均と分散がスペクトル領域にあるため、パラメータ合成を行わずに、加算をする。実際には、男女それぞれの HMM の状態ごとに、混合分布の重みを二分の一にし、混合分布数を増やす形になる。

本研究では、HMM の加算を行うために、1 つの状態に、男女の HMM を用いる必要があり、男性 HMM、女性 HMM とフレームの 3 次元モデルを作り認識を行う手法を提案する。FBANK による認識率が 4 つの場合の平均で 11% になり、MFCC では、条件に違いはあるが 10% という結果になった。FBANK に関しては、1state 1mixture の結果に比べ、よくなっていることから認識ができているといえるが、MFCC では、混合分布数を増やしても認識率が上がらない結果となり、合成を見直す必要がある。

以降、2 章で音声認識について説明し、3 章でクロストーク音声認識と認識手法について説明する。4 章で認識実験の説明をして、5 章に実験結果として認識結果を示し、6 章に考察を述べる。

2 音声認識

音声認識とは、音声波に含まれる意味内容に関する情報(言語情報)をコンピュータや電気回路によって抽出し、判定することである。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過程は、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、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。

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

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 $P(\omega)$ 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 $P(\omega)$ を計算するかが重要となる[7]。

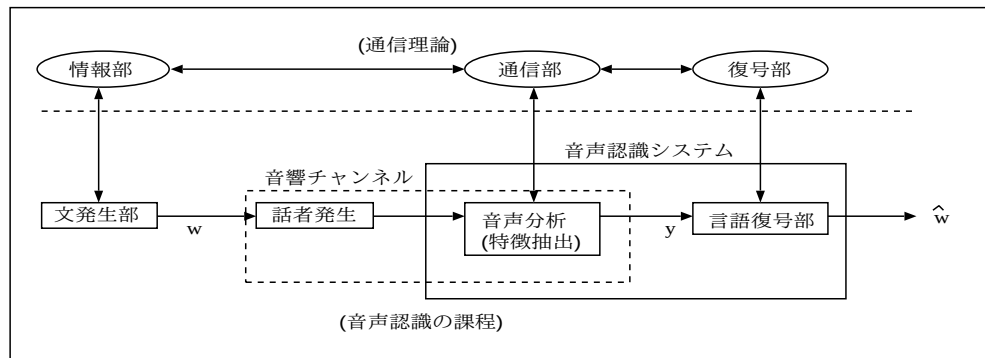


図 1: 音声認識課程の確率モデル

2.1 音声の特徴抽出

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

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

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

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

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

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

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

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

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

なる複素数系列 $S'(k)$ が音声のスペクトル表現として最も一般的に用いられる。音声信号の音素的特徴は主として調音フィルタの振幅伝達特性に含まれている。したがって、音声認識においては、音声信号の振幅スペクトル、あるいはその 2 乗値であるパワースペクトルがスペクトル表現である [8]。

2.2 特徴パラメータ

2.2.1 FBANK

FBANK(フィルタバンク対数パワー) は音声周波数に対して FFT スペクトルを求め、メルスケールに沿って等間隔に配置された三角関数のフィルタをかける。この三角関数の個数がフィルタバンクのチャンネルのチャンネル数 (特徴パラメータにおける次数) を表している。周波数メル分割の式は

$$Mel(f) = 2592 \log_{10}(1 + \frac{f}{700}) \quad (7)$$

となる。そして、フィルタバンクの出力に $\log$ 対数パワーを求めたものが FBANK であり、特徴パラメータにフォルマント成分及びピッチ成分が含まれる。これにより、音声の特徴をより正確に表現できる [8]。

2.2.2 MFCC

MFCC(メル周波数ケプストラム係数) は、まず音声周波数に対して FFT スペクトルを求め、メルスケール上に等間隔に配置された帯域フィルタバンクの出力を抽出する。そして、最終的に離散コサイン変換し得られるケプストラム係数が MFCC である。

高次においてピッチ成分、低次においてフォルマント成分が見られ、通常は扱いやすさの観点から低次のフォルマント成分が使用される [8]。

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos(\frac{\pi i}{N}(j - 0.5)) \quad (8)$$

2.3 HMM(隠れマルコフモデル)

HMM とは Hidden Markov Model(隠れマルコフモデル) のことであり, 出力シンボルによって一意に状態遷移先が定まらないという非決定状態オートマトンとして定義されている. HMM には, ある状態から全ての状態に遷移出来る Ergodic モデルや, 左から右へと状態遷移する left to right モデル等がある.

図 2 に HMM(left to right) の例を示す.

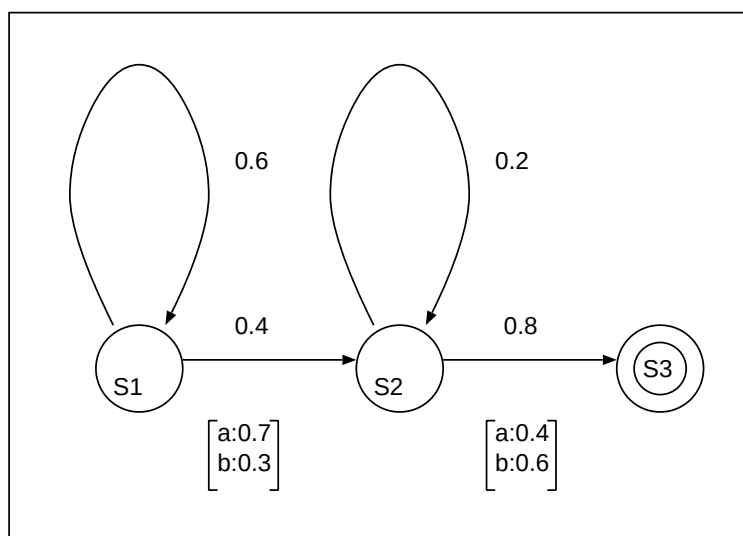


図 2: left-to-right モデルの例

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

$$\pi = \begin{pmatrix} 1.0 & 0 & 0 \end{pmatrix} \quad (9)$$

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

$$A = \begin{pmatrix} 0.6 & 0.4 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \quad (10)$$

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

$$B_a = \begin{pmatrix} 0.3 & 0.3 & 0.0 \\ 0.0 & 0.7 & 0.7 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \quad (11)$$

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

$$B_b = \begin{pmatrix} 0.4 & 0.4 & 0.0 \\ 0.0 & 0.6 & 0.6 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \quad (12)$$

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

例の HMM の出力シンボルが”abb”である場合, 可能な状態遷移系列は $s_1 s_1 s_2 s_3$ と $s_1 s_2 s_2 s_3$ の 2 つで, それぞれの確率は以下のようにして求めることができる.

$$0.6 * 0.7 * 0.4 * 0.3 * 0.8 * 0.6 = 0.024192 \quad (13)$$

$$0.4 * 0.7 * 0.2 * 0.6 * 0.8 * 0.6 = 0.016128 \quad (14)$$

よって, この HMM が”abb”を出力する確率は以下ようになる.

$$0.024192 + 0.016128 = 0.040320 \quad (15)$$

2.3.1 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過程を考え、状態が確率的に遷移するときに対応して確率的にシンボルを出力する。このとき観測できるのはシンボル系列だけであることからHidden(隠れ)マルコフモデルとよばれている。

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

HMMは以下の組から定義される。

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

2.3.2 HMM 法の利点と問題点

HMM が音声認識において有利な点を以下に示す.

- 個人差や調音結合, 発声法 (強さ, 速さ, 明瞭さ) 等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/確率仮定論による理論展開がしやすい.
- 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るため, 両者を統合しやすい.
- 認識時の計算量は比較的少ない.

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いため, 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学習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力に乏しい.
- 時系列パターンの 2 時点におけるパターンの壮観が考慮できない.

2.4 HMMの種類

2.4.1 離散型 HMM

出現するスペクトルパターンを、有限個のシンボルの組合せで表す分布モデルである。スペクトルパターンのベクトル量子化によって、符号ベクトルを生成し、各符号ベクトルの出現確率の組合せによって出現確率を表す。

2.4.2 連続型 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。

本研究では、連続型 HMM のみを使っている。

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
- $\mathcal{N}(\cdot; \mu, \Sigma)$...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。

$$b_{ij}(o_t) = \sum_{m=1}^{M_{ij}} C_{ijm} \mathcal{N}(o_t; \mu_{ijm}, \Sigma_{ijm}) \quad (16)$$

$\mathcal{N}(\cdot; \mu, \Sigma)$ は

- n ...観測行列の次元数
- $(O - \mu)^t \dots (O - \mu)$ の転置行列
- $|\Sigma| \dots \Sigma$ の固有値
- $\Sigma^{-1} \dots \Sigma$ の逆行列

とすると、以下の式で表現される。

$$\mathcal{N}(O; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(O - \mu)^t \Sigma^{-1} (O - \mu)\right) \quad (17)$$

2.4.3 半連続型 HMM

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

$$b_{ij}(x) = \sum_{m=1}^M \lambda_{ijm} b_{ijm}(x) \quad (18)$$

と混合正規分布で表す．ただし，

$$\sum_{m=1}^M \lambda_{ijm} = 1 \quad (19)$$

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 $s_i \rightarrow s_j$ での分布の重み λ_{ijm} のみが変わる．

2.5 認識アルゴリズム

$y = y_1, y_2, \dots, y_T$ 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 [7]. 図 3 に HMM を用いた単語音声認識の方法を示す.

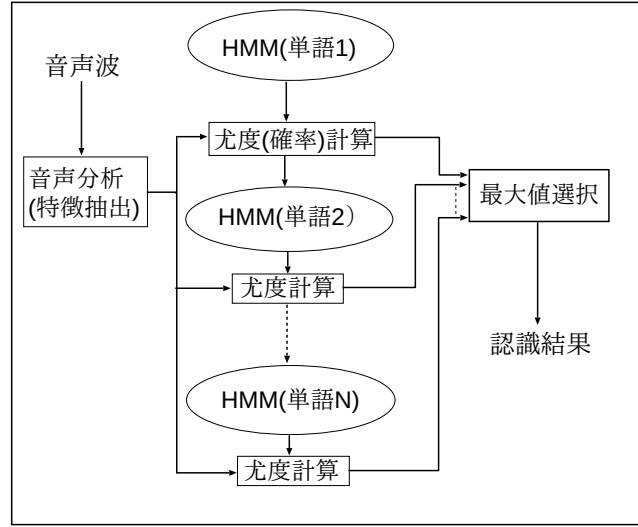


図 3: HMM を用いた単語音声認識の方法

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

$$P(y | M) = \sum_{i_0, i_1, \dots, i_T} P(y | q, M) \cdot P(q | M) \quad (20)$$

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

forward 変数 $\alpha(i, t)$ 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, $i = 1, 2, \dots, S$ とおいて, 以下の式を得る.

$$\alpha(i, t) = \sum_j \alpha(j, t-1) \cdot a_{ji} \cdot b_{ji}(y_t) \quad (t = 1, 2, \dots, T) \cdot \pi_i(t=0) \quad (21)$$

これを計算し, 最後に以下を求めれば良い.

$$P(y | M) = \sum_{i, q \in F} \alpha(i, T) \quad (22)$$

2.5.1 Forward アルゴリズム (トレリス法)

時刻 t の時に $o_1, o_2, \dots, o_t$ という観測系列を出力して、状態 j にいる確率を次のように定義する。

$$\alpha_t(j) = P(o_1, o_2, \dots, o_t, s_t = j \mid \lambda) \quad (23)$$

$P(O \mid \lambda)$ は $\alpha_t(j)$ の漸化式を、次のように計算することによって求めることができる。

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

これを計算して、最後に

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

を求めれば良い。

このアルゴリズムでは、直前のフレームにおける確率 $\alpha_{t-1}(i)$ から $\alpha_t(j)$ $1 \leq j \leq N$ を求める。

図 4 は、前記の図 2 の HMM がラベル系列 abb を出力する例に適応した例である。状態の中の数字は状態確率であり、矢印の横の数字は、状態遷移確率とラベルの出力確率である。

状態 $s1$ の 1 フレーム目を例にとれば、状態確率が 0.42 であり、図 4 より、 $s1$ から $s1$ に遷移する確率は 0.6 であり、 $s1$ から $s2$ に遷移する確率が 0.4 となる。シンボル b を出力すると決めているのでシンボル b を出力する確率は 0.3 である。よって、2 フレーム目の $s1$ と $s2$ の状態確率は以下のように計算できる。

$$s1 : 0.42 * 0.6 * 0.3 = 0.0756 \quad (26)$$

$$s2 : 0.42 * 0.4 * 0.3 + 0.28 * 0.2 * 0.6 = 0.0840 \quad (27)$$

このように出力ラベル系列が対応する時間経過を横軸にして、各状態を縦に並べて状態遷移を示した図で考えると理解しやすい。 $\alpha_t(j)$ は、トレリス上の左上（初期状態）から右下（最終状態）に向かって順次求まる [7]。

これを対数で計算するアルゴリズムを、ここでは Forward アルゴリズムと呼ぶ。

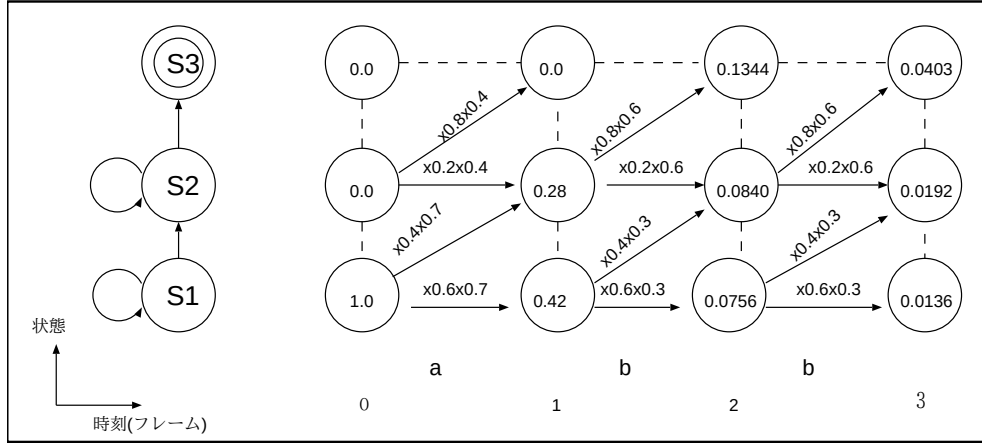


図 4: トレリス上の $\alpha_t(i)$ の計算

2.5.2 Viterbi アルゴリズム

Viterbi アルゴリズムは、モデルの最適な状態遷移行列 (最適経路) と、この経路上での確率を求めるアルゴリズムである。

$P(y | M)$ を厳密に求めないで、近似的に、モデル M が符号ベクトル系列 y を出力するときの最も可能性の高い状態系列上での出現確率を用いることを考える。この出現確率 (尤度) は、各遷移での確率値を対数変換しておくことにより、加算と大小判定のみからなる DP 演算によって高速に求めることができる。このアルゴリズムを以下に示す。 $i = 1, 2, \dots, S$ において

$$f'(i, t) = \begin{cases} \log \pi_i & (t = 0) \\ \max_j \{f'(i, t-1) + \log a_{ji} b_{ji}(y_t)\} & (t = 1, 2, \dots, T) \end{cases} \quad (28)$$

を計算し、対数尤度

$$L = \max_{i, s_i \in F} f'(i, t) \quad (29)$$

を求める [7]。

図 5 は、前記の図 2 の HMM がラベル系列 abb を出力する例に適応した例であるが、最短経路となるのが図の太い矢印で示した経路である。計算方法は、前節における Forward アルゴリズムと状態の計算方法以外同じである。この例では、2 フレーム目の状態 s_2 における繋がれたパスの尤度の比較により、選択された。以下に、2 フレーム目の状態 s_2 に繋がれた 2 つのパスの尤度を以下に示す。

状態 s_2 から状態 s_2 への尤度は、

$$0.28 * 0.2 * 0.6 = 0.0336 \quad (30)$$

であり,

状態 s1 から状態 s2 への尤度は,

$$0.42 * 0.4 * 0.3 = 0.0504 \quad (31)$$

であり, 状態 s1 から状態 s2 への尤度の方が大きいため, s1 からのパスを選択する.

Viterbi アルゴリズムは, 本研究において HMM の初期モデル作成とすべての認識手法に使用されている.

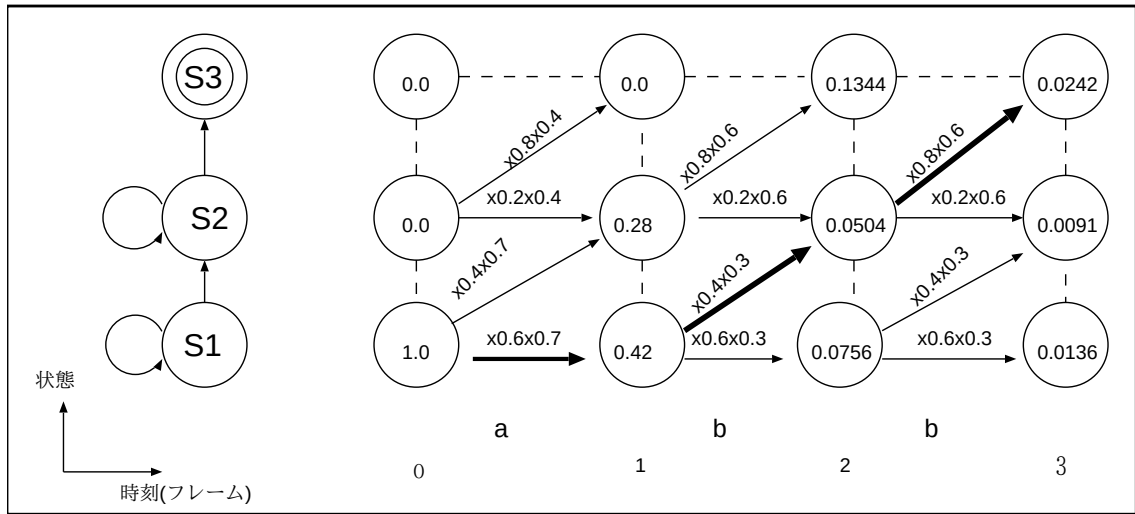


図 5: Viterbi アルゴリズムの適用例

2.5.3 トレリス法と対数を用いたアルゴリズムの違い

トレリス法の計算は小数点の乗算と加算であるが, 音声認識において小数点での尤度の計算は桁が 3, 4 桁以上違う. 対数を用いる場合はトレリス法に比べ以下のような利点がある. プログラム上, 一番最初の利点が重要であり, 尤度の計算が正確にできなくなる欠点を防ぎ, 対数による近似で尤度の計算を行う事が音声認識の基本である.

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

このために対数を用いたアルゴリズムは広く用いられている.

2.5.4 対数の加算の計算方法

本研究の状態の尤度の計算は対数で行うが，グリッドの計算において，状態の尤度の計算が一番の問題である．以下のような (33) から (36) のような場合分けをして，対数の足し算を行わなければならない．このアルゴリズムとして線形補間の方法が報告されている [9]．

$$\begin{aligned} A &= \log(a) \\ B &= \log(b) \\ C &= \log(a + b) \\ &= \log(a) + \log(1 + b/a) \\ &= A + \log(1 + \exp(B - A)) \end{aligned} \tag{32}$$

を利用して，以下のように対数同士の足し算を計算する．

$$\text{if} \quad (A \gg B) \quad A : \tag{33}$$

$$\text{elseif} \quad (B \gg A) \quad B : \tag{34}$$

$$\text{elseif} \quad (A \geq B) \quad A + \log(1 + \exp(B - A)) : \tag{35}$$

$$\text{elseif} \quad (B \geq A) \quad B + \log(1 + \exp(A - B)) : \tag{36}$$

なお，この方法は HTK[6] に使用されている．

2.6 HMMの学習

HMMを学習するためには、HMMを構成する要素のうち状態遷移確率A、出力確率Bの2つを、与えられた学習をもとに定める必要がある。両者の学習規範としては、最尤基準が適用されることが一般的である。HMMの学習における問題点は、学習用に与えられた出力信号系列が、どのような状態系列をたどって、出力されたものかが観測できないことにある。そのため、与えられた出力信号系列の下で状態遷移の期待値を求め、この期待値にもとづいてパラメータの推定を行う [7]。

2.6.1 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 アルゴリズムがある。再推定には、前章で述べた Forward アルゴリズムを出力信号系列において前と後ろから使用する。これを forward-backward アルゴリズムとよぶ。

具体的に示す。モデルには2.5節で述べたMを使用する。学習用音声として、 N 個の観測符号ベクトル系列 $\{y_1^{T(n)} = y_1, y_2, \dots, y_{T(n)}\}_{n=1}^N$ が与えられたとき、尤度である

$$\prod_{n=1}^N P(y_1^{T(n)} | \pi_i, a_{ij}, b_{ij}(k)) \quad (37)$$

を最大化するパラメータセット $\{\hat{\pi}_i, \hat{a}_{ij}, \hat{b}_{ij}(k)\}$ を推定する。以下に手順を示す。

1. モデルMにおいて前章で述べた forward 変数 $\alpha(i, t)$ を用いて、forward アルゴリズムにより尤度 $P(y_1^t | M)$ を出す。
2. backward 変数 $\beta(i, t)$ を出す。

backward 変数とは $[\beta(i, t):]$ 時刻 t に状態 s_i にあって、以後符号ベクトル y_{t+1}^T を出力する確率である。以下に式を示す。

$$\beta(i, T) = \begin{cases} 1 & s_i \in F \\ 0 & s_i \notin F \end{cases} \quad (38)$$

$$\beta(i, t) = \sum_j a_{ij} b_{ij}(y_t) \beta(j, t+1) \quad (t = T, T-1, \dots, 1; i = 1, 2, \dots, S) \quad (39)$$

3. シンボル生成過程で、時刻 t に状態 j いる確率 $\gamma(i, j, t)$ を出す。

$\gamma(i, j, t)$: 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. forward 変数 $\alpha(i, t)$, backward 変数 $\beta(i, t)$, パラメータセットそして forward アルゴリズムにおける尤度 $P(y_1^t|M)$ を用いて表す.

$$\gamma(i, j, t) = \frac{\alpha(i, t-1)a_{ij}b_{ij}(y_t)\beta(j, t)}{P(y_1^t|M)} \quad (40)$$

4. パラメータ p_{ii} , a_{ij} , $b_{ij}(k)$ を、以下のような再推定式によって求める.

$$\hat{\pi}_{ij} = \frac{\sum_j \gamma(i, j, 1)}{\sum_i \sum_j \gamma(i, j, 1)} \quad (41)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^T \alpha(i, t-1)a_{ij}b_{ij}(y_t)\beta(j, t)}{\sum_t \alpha(i, t)\beta(j, t)} = \frac{\sum_t \gamma(i, j, 1)}{\sum_t \sum_j \gamma(i, j, t)} \quad (42)$$

$$\hat{b}_{ij} = \frac{\sum_{t, y_t=k} \gamma(i, j, t)}{\sum_t \gamma(i, j, t)} \quad (43)$$

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.

本研究では、HMM 初期モデルの再推定に使用されている.

2.6.2 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が与えられたとき、尤度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、推定できる。

2.6.3 連続 HMM のパラメータ推定法

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。

2.6.4 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。

$$\hat{\mu}_{ij} = \frac{\sum_{t=1}^T \gamma(i, j, t) y_t}{\sum_{t=1}^T \gamma(i, j, t)} \quad (44)$$

$$\hat{\Sigma}_{ij} = \frac{\sum_{t=1}^T \gamma(i, j, t) (y_t - \mu_{ij})(y_t - \mu_{ij})^t}{\sum_{t=1}^T \gamma(i, j, t)} \quad (45)$$

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。

2.6.5 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)。

$$b_{ij}(y) = \sum_{m=1}^M \lambda_{ijm} b_{ijm}(y) \quad (46)$$

ここで

$$\sum_{m=1}^M \lambda_{ijm} = 1 \quad (47)$$

$$\int b_{ijm}(y) dy = 1 \quad (48)$$

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。

$$\hat{\lambda}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m)}{\sum_{t=1}^T \gamma(i, j, t)} \quad (49)$$

$$\hat{\mu}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m) y_t}{\sum_{t=1}^T \gamma(i, j, t, m)} \quad (50)$$

$$\hat{\Sigma}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m) (y_t - \mu_{ijm}) (y_t - \mu_{ijm})^t}{\sum_{t=1}^T \gamma(i, j, t, m)} \quad (51)$$

ただし,

$$\gamma(i, j, m, t) = \alpha(i, t-1) a_{ij} \lambda_{ijm} b_{ijm}(y_t) \beta(j, t) \quad (52)$$

であり, m 番目の分布関数の遷移 $q_i \rightarrow q_j$ の確率が表され, 値が収束するまで繰り返す.

2.6.6 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.

$$b_{ij}(y) = \sum_{m=1}^M \lambda_{ijm} b_{ijm}(y) \quad (53)$$

ここで,

$$\sum_{m=1}^M \lambda_{ijm} = 1 \quad (54)$$

である. この混合分布のパラメータの内, 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. 従って, 分布の重み λ_{ijm} は, 遷移状態 ($s_i \rightarrow s_j$) ごとに推定する. これらの推定式は,

$$\hat{\lambda}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m)}{\sum_{t=1}^T \gamma(i, j, t)} \quad (55)$$

$$\hat{\mu}_{ijm} = \frac{\sum_{all(s_i \rightarrow s_j)} \sum_{t=1}^T \gamma(i, j, t, m) y_t}{\sum_{all(s_i \rightarrow s_j)} \sum_{t=1}^T \gamma(i, j, t, m)} \quad (56)$$

$$\hat{\Sigma}_{ijm} = \frac{\sum_{all(s_i \rightarrow s_j)} \sum_{t=1}^T \gamma(i, j, t, m) (y_t - \mu_{ijm}) (y_t - \mu_{ijm})^t}{\sum_{all(s_i \rightarrow s_j)} \sum_{t=1}^T \gamma(i, j, t, m)} \quad (57)$$

となる.

2.7 HMMの連結学習

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、大量の音声データを用いる必要があるが、その音声データに、逐一、人手によるラベル付けを行うことは非常に困難である。そこでラベル付けされていない音声データベースを用いて学習を行う方法が連結学習である。ただし、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 6 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。

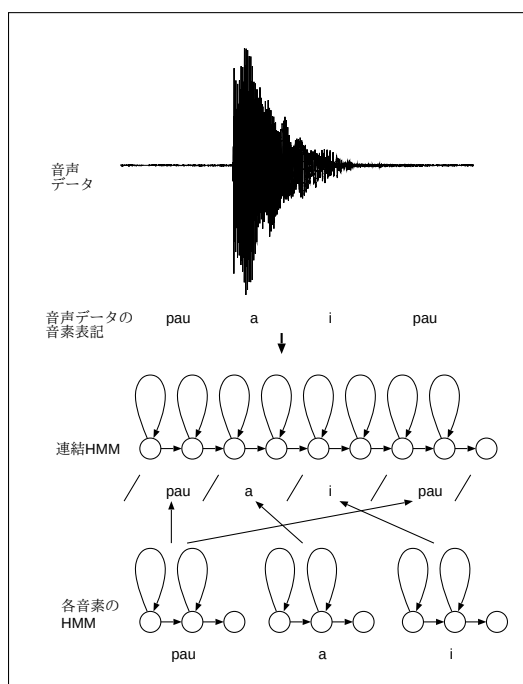


図 6: 連結学習の例

3 クロストーク音声認識

3.1 従来の研究

会議など様々な場面において、人々は同時に会話などをする。このような場面で複数の話者が同時に、違う声の大きさに発話したとき、計算機を用いて全ての話者の音声を認識できるシステムの実現が望まれる。クロストーク音声における従来研究は、複数の話者に対してマイクロフォンを使用し認識を行っている。

複数のマイクロフォンを使用する場合、各マイクロフォンに入力される複数音声の音量の差や、時間差などの情報を利用し、認識を行う。問題点として、複数のマイクロフォンを使用することでコストがかかる点や、マイクロフォンを中心に同角度の場所から、複数の音声が発声された場合に、認識精度が低下する点などが挙げられる。

単一のマイクロフォンを使用する場合、重畳音声を分離する手法 [3] や HMM 合成法を用いた手法 [4] がある。重畳音声を分離する手法は、重畳音声に対し、音源のモデルとして相関関数を用いた重畳音声分解法が提案され、重畳音声を複数の孤立単語音声に分解する性能に関して有効性が示されている。問題点として、重畳音声を複数の孤立単語音声に分解する性能のみが評価されており、重畳音声を分解した後の、音声に対する認識精度は評価されていないことが挙げられる。

HMM 合成法を用いた手法では、音声为重畳している部分において、重畳音声を認識対象の音声と妨害音声为重畳されていると考え、認識対象の音声と妨害音声を合成 HMM を用いて表現し認識を行う。

過去の研究では、クロストーク音声の両方の音声を認識対象として同時発話認識率の調査が行われ、認識手法の単純法では 56%、マルチパス法で 45% の認識率が得られたが、人間の聴覚実験による結果では 77% となり、改善の必要がある。また、クロストーク音声において、男性話者の音声と女性話者の音声を個別に考える点から妥当性がないといわれるため、雑音対策の認識手法である Parallel Model Combination 法（以下、PMC 法）の提案がされたが、まだまだ未完成である [2]。

本研究では、PMC 法の HMM 合成を特徴パラメータ MFCC ではパラメータ合成を行い、FBANK では HMM をそのままスペクトル領域で加算することで行う。また、比較として、単純法とマルチパス法による実験も行う。

3.2 認識手法

従来の認識手法は、単純法、マルチパス法、PMC 法がある。本研究で、3次元ビタビ探索を用いた手法を提案する。

3.2.1 単純法

クロストーク音声に対し、男性話者女性話者 10 単語ずつ別々に HMM を使用し認識を行う。単純法の流れを図 7 に、認識の例を図 8 に示す。

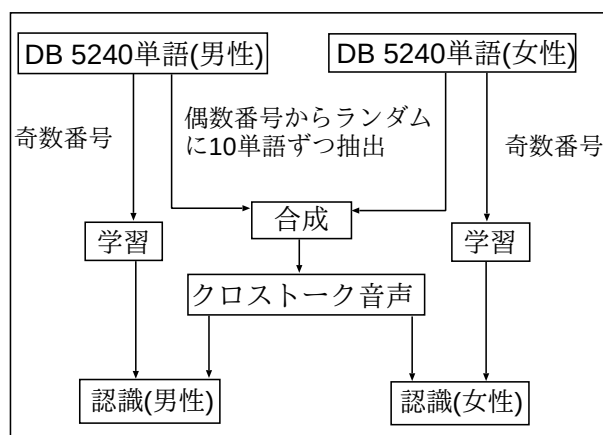


図 7: 単純法

クロストーク音声		男性話者	女性話者
		悪質	対策
単純法 認識結果 それぞれ 10単語認識	尤度順位	男性話者	女性話者
	1	悪質	対策
	2	聞こえる	足元
	3	優れる	欲張る

図 8: 単純法の認識の例

この図では、クロストーク音声の発話内容が「悪質+対策」であり、男性話者認識で尤度順位で一番高いモデルが「悪質」、「対策」であり同時発話認識できている。

3.2.2 マルチパス法

マルチパス法は、男性話者 HMM と女性話者 HMM を平行にしたモデルである。音声 を音素単位で考え、各モーラごとに子音と母音に分け、男性話者の子音と女性話者の子音で、男性話者の母音と女性話者の母音で相互にパスを持つ PMC 法のモデルを構築する。各状態の尤度は、各状態に遷移されるパスの最大値を選択する。従って、モデルの尤度は、モデルの最適経路上での尤度となる。最後に、PMC 法と同様の手法で認識結果を決定する。

マルチパス法のモデルは、HTK を用いて認識を行うため、PMC 法のパスに、相手側 音声の次の音素に遷移するパスを加える。図 8 にマルチパス法のモデルを示す。図 9 に示す。

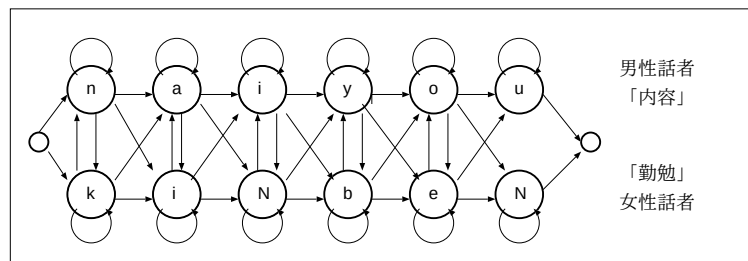


図 9: マルチパス法

3.2.2.1 マルチパス法の認識率の算出方法

1 個の音声に対して複数の候補の単語の尤度を求め、一番高い単語が認識結果になる。具体的には、1 個のクロストーク音声に対して、男性話者 10 単語女性話者 10 単語により、作成した 100 個のモデルの尤度を出し、一番高い組合せのモデルが認識結果になる。100 個のクロストーク音声に対して行い、認識できていた場合を数え認識率とする。PMC 法と提案手法も同じ方法をとる。

図 10 にマルチパス法の認識結果の例を示す。

クロストーク音声		男性話者	女性話者
		悪質	対策
認識結果 ※10単語×10単語の 100個のモデルで認識	尤度順位	男性話者単語	女性話者単語
	1	悪質	対策
	2	悪質	天才
	3	悪質	手拭い
	...	...	...
	99	暴力	細々
	100	中毒	可愛い

図 10: マルチパス法の認識結果

発話内容が「悪質+対策」に対して 100 個のモデルの尤度を出し、順位付けしたのが下の表で、1 位の「悪質」「対策」が認識結果になる。この例は同時発話認識できている。

3.2.3 PMC 法

PMC 法は、Parallel Model Combination 法 [11] の略で、雑音が重畳した音声を認識する一般的な方法である。雑音環境下での PMC 法の HMM は、音声の HMM と雑音の HMM をパラメータ合成して作成される。パラメータ合成は次節で説明する。図 11 に示すように、合成された HMM は 2 つの HMM を掛け合わせた構造になる。状態遷移確率はそれぞれの遷移確率の積になる。

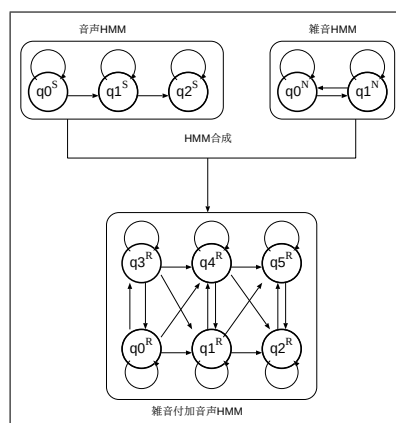


図 11: HMM 合成法 (PMC) の原理

- パラメータ合成

パラメータ合成は、ケプストラムをスペクトル領域にし、足した後、元に戻すことである。具体的には、雑音と音声の出現確率が、ケプストラム空間での正規分布で表されているが、雑音と音声は線形スペクトル領域で加算の関係にあるので、両者の分布をスペクトル領域での分布に置き換える。合成されたスペクトル領域での対数正規分布をケプストラム領域での正規分布に戻せば、雑音の重畳した音声の HMM の出現確率が得られる [7].

1. フーリエ変換

$$\mu_{log} = \Gamma \mu_{cep} \quad (58)$$

$$\Sigma_{log} = \Gamma \Sigma_{cep} \Gamma^T \quad (59)$$

2. 指数変換

$$\mu_{(lin),i} = \exp\left\{\mu_{(log),i} + \frac{\sigma_{(log),ii}^2}{2}\right\} \quad (60)$$

$$\sigma_{(lin),ij}^2 = \mu_{(log),i} \mu_{(log),j} \exp(\sigma_{(log),ij}^2 - 1) \quad (61)$$

3. 加算

$$\mu_{(lin_Y)} = \mu_{(lin_S)} + k_{SNR} \cdot \mu_{(lin_N)} \quad (62)$$

$$\Sigma_{(lin_Y)} = \Sigma_{(lin_S)} + k_{SNR}^2 \cdot \Sigma_{(lin_N)} \quad (63)$$

4. 対数変換

$$\mu_{(log_Y),i} = \log \mu_{(lin_Y),i} - \frac{1}{2} \cdot \left\{ \frac{(\sigma_{(lin_Y),ij})^2}{(\mu_{(lin_Y),i})^2} + 1 \right\} \quad (64)$$

$$\sigma_{((log_Y),ij)}^2 = \log \left\{ \frac{\sigma_{(lin_Y),ij}^2}{\mu_{(lin_Y),i} \mu_{(lin_Y),j}} + 1 \right\} \quad (65)$$

5. 逆フーリエ変換

$$\mu_{cep_Y} = \Gamma^{-1} \mu_{(log_Y)} \quad (66)$$

$$\Sigma_{cep_Y} = \Gamma^{-1} \Sigma_{(log_Y)} (\Gamma^{-1})^T \quad (67)$$

本研究では、状態数 1 混合分布数 1 と状態数 3 混合分布数 4 の HMM を用いるが、パラメータ合成は、状態数 1 のみで行う事ができ、また複数の混合分布数の加算はできないため、状態数 1 で混合分布数 1 と混合分布数 4 の HMM のパラメータ合成となる。

3.3 提案手法

3.3.1 従来の認識手法の問題点

クロストーク音声認識をする上で、重畳した音声を認識する際、フレーム単位で男性女性両方の HMM で計算する必要がある。しかし、従来の認識手法は、どちらか一方の HMM で計算する。単純法は、話者ごとに認識を行うために、その話者の HMM で計算するため、問題がある。また、マルチパス法は、状態の尤度を各状態に遷移されるパスの尤度の最大値とするため、時間単位では、どちらかの HMM を計算する。

そこで、過去の研究では、雑音対策の認識手法である PMC 法の適用が提案されている [2]。音声と雑音がスペクトル領域で加算の関係にあるという前提条件から、雑音を女性音声と考えることで、適用できると考えられた。実際には、特徴パラメータ MFCC において、ケプストラムをスペクトル領域にして加算し元に戻すことで、両方の HMM で計算できる。本研究では、「スペクトル領域で男性の音声と女性の音声は加算の関係である」という前提条件を立て、男女の状態数混合分布数をスペクトル領域で加算する手法を提案する。1 状態に男女の HMM を用いるために、男女の HMM を交差させ、二次元に考え、フレームとあわせて 3 次元ビタビ探索を行う。

3.3.2 3 次元ビタビ探索

従来の 3 次元ビタビ探索は、ハンズフリー音声認識で用いられ、マイクロホンアレーの方向と特徴ベクトルの方向を合わせた 2 次元モデルとフレームを合わせた 3 次元モデルで行われている [10]。本研究の 3 次元ビタビ探索を図 12 に示す。

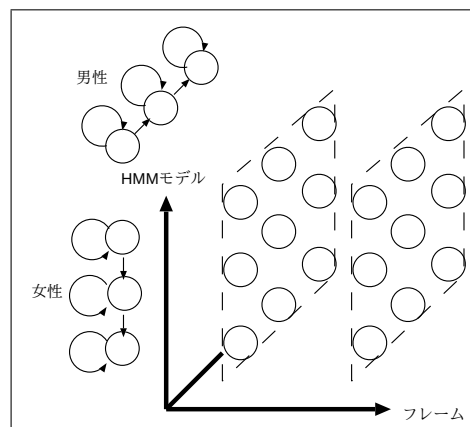


図 12: 本研究における 3 次元ビタビ探索

3.3.3 本研究のモデル

図 13 のように男性話者 HMM と女性話者 HMM を交差させ、モデルを作る．状態の計算方法は，1 つの状態に男性 HMM と女性 HMM を用い，次章の男性 HMM と女性 HMM の加算を行う．

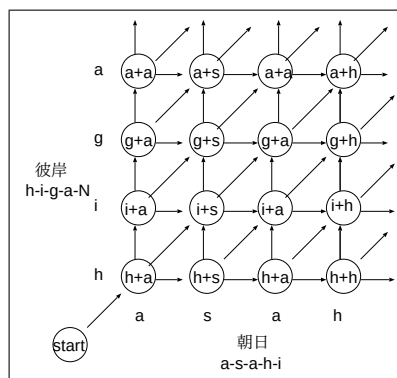


図 13: 提案する提案手法のモデル

3.3.4 男性 HMM と女性 HMM の加算

特徴パラメータ FBANK では，HMM の平均と分散がスペクトル領域であるため，そこで，男性話者 HMM に女性話者 HMM を加える形にして合成 HMM とする．実際には HMM の出力確率のための混合分布を加算し，状態遷移確率を加算する．状態数 1 混合分布数 1 の場合を図 14 に示す．

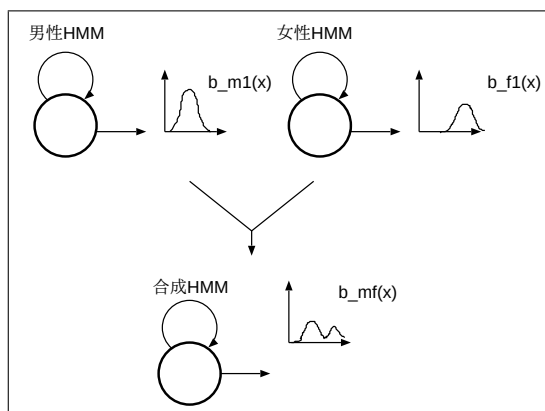


図 14: スペクトル領域での加算 (1mixture)

4 認識実験

4.1 実験の目的

クロストーク音声に対して，提案手法の認識率を調査する．従来の認識手法である単純法とマルチパス法と比較することで，有効性を調べる．また，パラメータ合成を行った PMC 法の認識率を調査する．

4.2 実験データ

音声データベースとして，ATR 単語発話データベース Aset(1 話者につき 5,240 単語)の男性話者 mau,mms 及び女性話者 ftk,fyn の男女各 2 名を使用する．HMM の学習に使用する単語は，5,240 単語の中の奇数番号である．

認識に使用する単語は，偶数番号の中から，4 モーラで発話時間がほぼ同じ単語を 20 個選び，そのうちの 10 単語を男性話者が，残りの 10 単語を女性話者が発声したとする．垂直のモデルを作成時，音素が 8 個では，セグメンテーションエラーがおきるために，7 個以下とした．

本実験で行う認識は，全て音声認識ツール HTK[6] を使用する．

表 1: 実験に使用した単語

男性話者	女性話者
悪質 (akusitsu)	足元 (asimoto)
聞こえる (kikoeru)	可愛い (kawaii)
加える (kuwaeru)	勤勉 (kiNbeN)
失恋 (shitsureN)	答える (kotaeru)
垂直 (suichoku)	すまない (sumanai)
そのうち (sonouchi)	対策 (taisaku)
中毒 (chuudoku)	手拭い (tenugui)
内容 (naiyou)	天才 (teNsai)
暴力 (bouryoku)	微笑む (hohoemu)
論じる (roNjiru)	洋服 (youfuku)

表 2: クロストーク音声 (悪質と重畳した例)

入力音声番号	男性話者	女性話者
1	悪質	足元
2	悪質	可愛い
3	悪質	勤勉
4	悪質	答える
5	悪質	すまない
6	悪質	対策
7	悪質	手拭い
8	悪質	天才
9	悪質	微笑む
10	悪質	洋服

4.3 実験手順

クロストーク音声認識の流れを図 15 に示す.

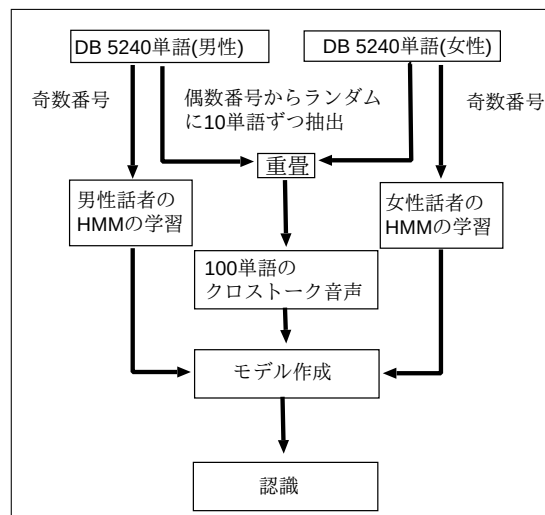


図 15: クロストーク音声認識の流れ

- (1) 認識に使用する単語として選んだ男性話者, 女性話者それぞれ 10 単語を重ね合わせて, 100 単語のクロストーク音声を作成する. なお, 選んだ単語は表 1 に載せる. 作成したクロストーク音声の例を表 2 に載せる.
- (2) HMM を, 「1state 1mixture」と「3state 4mixture」において作成する.
- (3) 男性 HMM と女性 HMM を用いて, 合成 HMM を作る. 合成方法は, 特徴パラメータ FBANK では, 混合分布の加算を, 特徴パラメータ MFCC-E では, パラメータ合成を行う.

- (4) クロストーク音声 100 個のモデルを作る．詳細は，次章の HTK によるモデル作成において説明する．
- (5) クロストーク音声に対して，100 個のモデルを使い，HTK により認識を行う．100 個の入力音声に 100 個のモデルで認識結果を出し，ソートを行う．「悪質+足元」，「悪質+可愛い」，「悪質+勤勉」の 3 例の認識結果を付録 4, 5, 6 にそれぞれ示す．

4.4 HTK によるモデル作成

HTK の認識には，dic ファイルと net ファイルが必要であり，モデルは dic ファイルと net ファイルにより定義される．net ファイルは gram ファイルから作られる．それぞれの意味と本研究で作成した例の一部を以下に示す．

- dic ファイルは音素の定義を行う．モデル作成で扱われる音素（男性音素+女性音素）である．
- gram ファイルはモデル内でどのように音素をつなぐかを BNF(Backus-Naur Form) 法により定義する．本研究では，図 13 のような男女を交差させたモデルを表す．pau は無音部分である．作成した gram ファイルにより，どのような最適な経路をいくかを付録 4, 5, 6 に示す．
- net ファイルは，gram ファイルで示されたパスの経路を実際につないだものである．

— dic ファイルの一部 —

N+N	N+N
N+a	N+a
N+b	N+b
N+d	N+d
N+e	N+e
N+f	N+f
N+g	N+g
N+h	N+h
N+hy	N+hy
N+i	N+i
N+j	N+j
N+k	N+k

gram ファイルの一部 (「akushitsu」と「ashimoto」)

```
(pau+pau (a+a (k+a (u+a (sh+a (i+a (ts+a (u+a (u+sh (u+i (u+m (u+o (u+t
(u+o )))))) —u+sh (u+i (u+m (u+o (u+t (u+o ))))) —ts+sh (u+sh (u+i (u+m
(u+o (u+t (u+o ))))) —u+i (u+m (u+o (u+t (u+o )))) —ts+i (u+i (u+m (u+o
(u+t (u+o )))) —u+m (u+o (u+t (u+o )) —ts+m (u+m (u+o (u+t (u+o ))
—u+o (u+t (u+o )) —ts+o (u+o (u+t (u+o )) —u+t (u+o ) —ts+t (u+t (u+o )
—u+o —ts+o (u+o )))))) — ts+sh (u+sh (u+i (u+m (u+o (u+t (u+o ))))) —u+i
(u+m (u+o (u+t (u+o ))))
```

net ファイルの一部 (「akushitsu」と「ashimoto」)

```
VERSION=1.0
N=24324 L=33311
I=0    W=pau+pau
I=1    W=u+o
I=2    W=!NULL
I=3    W=ts+o
I=4    W=i+o
I=5    W=sh+o
I=6    W=u+o
I=7    W=k+o
I=8    W=a+o
I=9    W=u+o
I=10   W=ts+o
I=11   W=i+o
I=12   W=sh+o
I=13   W=u+o
```

(略)

```
I=24321 W=pau+pau
I=24322 W=!NULL
I=24323 W=!NULL
J=0     S=2     E=0
J=1     S=3     E=1
J=2     S=1     E=2
J=3     S=9     E=2
J=4     S=15    E=2
```

4.5 実験条件

本実験では，HMM の学習に HTK を使用し，実験環境を表 3 にまとめる．特徴パラメータには FBANK を，音素 HMM の共分散行列は Diagonal-covariance を使用する．

表 3: 実験条件

HMM の条件	設定値
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	(1 ループ 2 状態・3 ループ 4 状態) 連続分布型
stream 数	1
混合分布数	4 対数パワー 1・1 対数パワー 1
MFCC	
特徴パラメータ	MFCC(12 次)+対数パワー (計 13 次)
FBANK	
特徴パラメータ	FBANK(24 次)+対数パワー (計 25 次)

5 実験結果

5.1 提案手法

提案手法と他の手法の認識率を示す．認識率は、「mau+ftk」,「mau+fyn」,「mms+ftk」,「mms+fyn」の場合の平均である．

5.1.1 同時発話認識率の比較

1state 1mixture の場合と，3state 4mixture の場合の同時発話認識率を図 16 に示す．

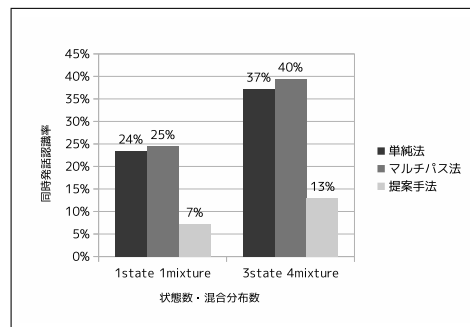


図 16: 状態数混合分布数による変化

提案手法の認識率は，マルチパス法，単純法に比べ低く，13.00%であった．条件を，1state 1mixture から 3state 4mixture にしても，向上が少ない．

5.1.2 各話者認識率の比較

3state 4mixture の場合の各話者認識率と同時発話認識率を認識手法別に図 17 示す．

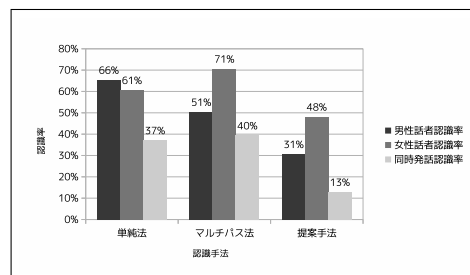


図 17: 認識手法の傾向

提案手法とマルチパス法で，男性話者認識率より女性話者認識率が高い．

5.2 PMC 法

PMC 法と他の認識手法の認識率を示す。認識率は、「mau+ftk」の場合である。

5.2.1 他の認識手法との比較

特徴パラメータ MFCC の場合の認識率を図 18 に示す。

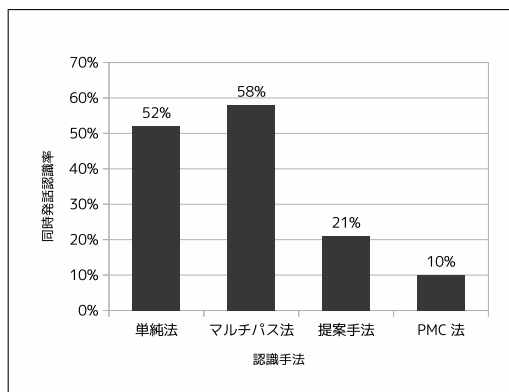


図 18: 認識手法の比較

PMC 法は一番低い認識率となり、10%であった。

5.2.2 状態数混合分布数別

状態数 1 混合分布数 1 の場合を基準に、男性 HMM の混合分布数を 4 にした場合と、女性 HMM の混合分布数を 4 にした場合の認識率を図 19 に示す。混合分布数を mix としている。

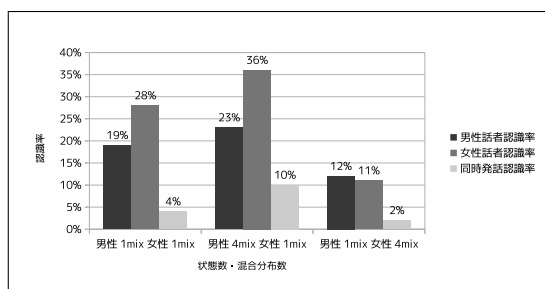


図 19: PMC 法の認識率

男性の混合分布数を増やした場合は、混合分布数 1 の認識率に比べ高い。一方、女性の混合分布数を増やした場合は、混合分布数 1 の認識率に比べ低い。

5.3 認識率の詳細

5.3.1 提案手法

提案手法の詳細な実験結果を示す。特徴パラメータ FBANK で 1state 1mixture の結果を表 4 に、3state 4mixture の結果を表 5 に示す。MFCC の場合は表 6 に示す。

表 4: 認識率:提案手法 (FBANK,1state 1mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	28%	35%	7%
mau+fyn	19%	11%	4%
mms+ftk	28%	26%	7%
mms+fyn	12%	27%	11%
平均	21.75%	24.75%	7.25%

表 5: 認識率:提案手法 (FBANK,3state 4mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	21%	66%	12%
mau+fyn	25%	37%	7%
mms+ftk	47%	42%	15%
mms+fyn	30%	48%	18%
平均	30.75%	48.25%	13.00%

表 6: 認識率:提案手法 (MFCC)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
1state 1mixture			
mau+ftk	29%	36%	10%
3state 4mixture			
mau+ftk	38%	65%	21%

実験結果より以下のことが得られた。

1. 提案手法による同時発話認識率は 13.00%である。
2. 3state 4mixture になると、女性話者認識率が特に高くなる傾向にある。
3. 3state 4mixture になると、男性話者認識率が低下するものがある

5.3.2 マルチパス法

マルチパス法の詳細な実験結果を示す．特徴パラメータ FBANK で 1state 1mixture の結果を表 7 に，3state 4mixture の結果を表 8 に示す．また，MFCC の場合を表 9 に示す．

表 7: 認識率:マルチパス法 (FBANK,1state 1mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	42%	34%	19%
mau+fyn	43%	61%	33%
mms+ftk	44%	55%	21%
mms+fyn	36%	81%	25%
平均	41.25%	57.75%	24.50%

表 8: 認識率:マルチパス法 (FBANK,3state 4mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	60%	41%	26%
mau+fyn	49%	81%	48%
mms+ftk	55%	68%	41%
mms+fyn	38%	93%	37%
平均	50.50%	70.75%	39.50%

表 9: 認識率:マルチパス法 (MFCC)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
1state 1mixture			
mau+ftk	61%	70%	48%
3state 4mixture			
mau+ftk	76%	72%	58%

実験結果より以下のことが得られた．

1. 平均でみると，1state 1mixture，3state 4mixture とともに女性話者認識率が 20%ほど高い
2. 男性話者認識率と同時発話認識率に差がないものがある

5.3.3 PMC 法

PMC 法の詳細な実験結果を示す．認識率は「mau+ftk」の場合である．状態数は1，混合分布数は以下のように組み合わせた．

1. 男性 1mixture 女性 1mixture
2. 男性 1mixture 女性 4mixture
3. 男性 4mixture 女性 1mixture
4. 男性 1mixture 女性 4mixture 母音 10mixture
5. 男性 4mixture 母音 10mixture 女性 1mixture

上3つの場合の認識率を表10に，下2の場合の認識率を表11に示す．

表 10: パラメータ合成での認識 (状態数は1)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
	男性 1mixture+1mixture		
mau+ftk	19%	28%	4%
	男性 4mixture+女性 1mixture		
mau+ftk	23%	36%	10%
	男性 1mixture+女性 4mixture		
mau+ftk	12%	11%	2%

表 11: パラメータ合成での認識 (状態数は1, 母音, 撥音, 無音のみ 10mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
	男性 4mixture+女性 1mixture		
mau+ftk	23%	38%	9%
	男性 1mixture+女性 4mixture		
mau+ftk	12%	18%	3%

実験結果より以下のことが得られた．

1. 1state1mixture による結果を基準に，男性の混合分布数を増やした場合のみ認識率がよくなっている．
2. 母音，撥音，無音による効果が見られない．

5.3.4 単純法

単純法の詳細な実験結果を示す。特徴パラメータ FBANK で 1state 1mixture の場合を表 12 に, 3state 4mixture の場合を表 13 に示す。また, MFCC の場合は表 14 に示す。

表 12: 認識率:単純法 (FBANK,1state 1mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	47%	30%	18%
mau+fyn	58%	51%	25%
mms+ftk	58%	42%	25%
mms+fyn	41%	70%	26%
平均	51.00%	48.25%	23.50%

表 13: 認識率:単純法 (FBANK,3state 4mixture)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	81%	42%	33%
mau+fyn	67%	52%	41%
mms+ftk	60%	61%	32%
mms+fyn	54%	88%	43%
平均	65.50%	60.75%	37.25%

表 14: 認識率:単純法 (MFCC)

組合せ	男性話者認識率	女性話者認識率	同時発話認識率
1state 1mixture			
mau+ftk	62%	49%	30%
3state 4mixture			
mau+ftk	79%	71%	52%

実験結果より以下のことが得られた。

1. 平均でみると男性話者認識率が高いが, 差はない
2. 同時発話認識率は話者ごとにばらつきがない

5.4 認識例

単純法は10単語認識のため、尤度順位は10位までである。また、マルチパス法と提案手法は100モデルを尤度で順位付けし、上位5位までを示す。

5.4.1 マルチパス法で認識失敗したが、提案手法で認識成功した例

男性話者「mau」と女性話者「ftk」の場合で示す。マルチパス法で認識できた数は26であり、提案手法では12である。マルチパス法で認識失敗したが、提案手法で認識できた数は3である。具体的には、男性話者の単語「聞こえる (kikoeru)」の組み合わせ3つである。

表15に、入力音声「聞こえる」、「可愛い」に対し、マルチパス法による認識結果と、提案手法による認識結果を示す。

表 15: 認識結果 (例 1)

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (ftk)
聞こえる (kikoeru)	可愛い (kawaii)

マルチパス法				提案手法			
順位	尤度	男性話者	女性話者	順位	尤度	男性話者	女性話者
1	-2416.95	加える	可愛い	1	-1915.11	聞こえる	可愛い
2	-2496.90	論じる	可愛い	2	-2056.56	加える	可愛い
3	-2496.90	悪質	可愛い	3	-2078.70	聞こえる	天才
4	-2500.59	聞こえる	可愛い	4	-2083.62	聞こえる	手拭い
5	-2509.20	失恋	可愛い	5	-2158.65	聞こえる	すまない

マルチパス法による認識結果では尤度順位で4位となっており、同時発話認識に失敗している。一方、提案手法による認識結果では、尤度順位で1位となっており、同時発話認識が成功している。マルチパス法では女性話者の単語が尤度順位の上位を占めているが、提案手法では男性話者の単語が上位を占める結果となった。

5.4.2 マルチパス法で認識成功したが、提案手法で認識失敗した例

男性話者「mau」と女性話者「ftk」の場合で示す。マルチパス法で認識できた数は26であり、提案手法では12である。マルチパス法で認識成功したが、提案手法で認識失敗した数は14である。表16に、入力音声「内容」、「対策」に対し、マルチパス法による認識結果と、提案手法による認識結果を示す。

表 16: 認識結果 (例 2)

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (ftk)
内容 (naiyou)	対策 (taisaku)

マルチパス法			
順位	尤度	男性話者	女性話者
1	-3103.6	内容	対策
2	-3201.6	内容	答える
3	-3327.18	内容	手拭い
4	-3327.18	内容	天才
5	-3426.54	内容	可愛い

提案手法			
順位	尤度	男性話者	女性話者
1	-2223.18	加える	手拭い
2	-2235.6	加える	対策
3	-2257.68	加える	可愛い
4	-2310.12	加える	すまない
5	-2312.88	加える	勤勉
10	-2511.90	内容	対策

マルチパス法による認識結果では、尤度順位が1位となっており、同時発話認識できている。提案手法による認識結果では尤度順位で10位となっており、同時発話認識に失敗している。提案手法の尤度順位上位で男性話者「加える」の組み合わせが多い。

5.4.3 単純法で認識失敗したが、提案手法で認識できた例

男性話者「mau」と女性話者「ftk」の場合で示す。単純法で認識できた数は30であり、提案手法では12である。単純法で認識失敗したが、提案手法で認識できた数は7である。例を以下に示す。

主に、男性話者の単語「聞こえる (kikoeru)」と「加える (kuwaeru)」の組み合わせであった。

表 17 に、入力音声「加える」、「微笑む」に対し、単純法による認識結果と、提案手法による認識結果を示す。

表 17: 認識結果 (例 3)

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (ftk)
加える (kuwaeru)	微笑む (hohoemu)

● 単純法

男性話者認識		
順位	尤度	男性話者
1	-1833.1	加える
2	-2659.5	内容
3	-2663.5	聞こえる
4	-3071.0	論じる
5	-3177.9	暴力

女性話者認識		
順位	尤度	女性話者
1	-3213.8	答える
2	-3418.8	可愛い
3	-3714.9	すまない
4	-3740.4	対策
5	-3769.3	天才
8	-4263.1	微笑む

● 提案手法

提案手法			
順位	尤度	男性話者	女性話者
1	-1645.52	加える	微笑む
2	-1648.80	加える	勤勉
3	-1657.20	加える	手拭い
4	-1754.40	加える	すまない
5	-1797.60	加える	可愛い

単純法による認識結果では、男性話者認識では認識できたが、女性話者では認識失敗した結果となり、同時発話認識に失敗している。ちなみに、女性話者認識では尤度順位 8 位の結果となった。一方、提案手法による認識結果では、尤度順位で 1 位となっており、同時発話認識が成功している。

5.4.4 単純法で認識できたが、提案手法で認識失敗した例

男性話者「mau」と女性話者「ftk」の場合で示す。単純法で認識できた数は30であり、提案手法では12である。単純法で認識失敗したが、提案手法で認識できた数は7である。例を以下に示す。

主に、男性話者の単語「聞こえる (kikoeru)」と「加える (kuwaeru)」の組み合わせであった。

表18に、入力音声「悪質」,「手拭い」に対し、単純法による認識結果と、提案手法による認識結果を示す。

表 18: 認識結果 (例 4)

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (ftk)
悪質 (akushitsu)	手拭い (tenugui)

● 単純法

男性話者認識		
順位	尤度	男性話者
1	-3209.2	悪質
2	-3483.0	聞こえる
3	-3560.6	加える
4	-3613.2	暴力
5	-3646.5	論じる

女性話者認識		
順位	尤度	女性話者
1	-3479.3	手拭い
2	-3500.5	対策
3	-3653.4	洋服
4	-3802.2	足元
5	-3821.7	天才

● 提案手法

提案手法			
順位	尤度	男性話者	女性話者
1	-2058.24	聞こえる	手拭い
2	-2176.16	加える	手拭い
3	-2205.64	聞こえる	天才
4	-2307.48	加える	可愛い
5	-2324.9	加える	天才
14	-2485.7	悪質	手拭い

単純法による認識結果では、男性話者認識でも、女性話者認識でも認識できており、同時発話認識に成功している。一方、提案手法による認識結果では、尤度順位で14位となっており、同時発話認識に失敗している。

6 考察

6.1 提案手法の有効性

6.1.1 提案手法の精度

認識実験において、状態数混合分布数を増やすことにより認識率がよくなっていることから、認識はできていると考える。次に、提案手法の精度が低かった原因を3点考えている。

1点目は、混合分布の加算において、仮定の間違いである。本研究では、「スペクトル領域で男性の音声と女性の音声は加算の関係である」という仮定のもと、スペクトル領域で混合分布の加算を行ったが、実際は、男性の音声と女性の音声はスペクトル領域で加算の関係ではなく、乗算の関係の可能性がある。その可能性を探るために、特徴パラメータをかえ、PMC法の調査を行う必要がある。

2点目は、3次元ビタビ探索を用いた手法の精度の調査不足である。本研究では、男女2話者のHMMのモデルをBNF記法により作成し、HTKに組み込み認識を行ったが、さまざまな条件の加算したHMMにより認識精度を求めた結果、特徴パラメータMFCCにおける提案手法の21%が最高である。HMMの加算を行う上で、男女が公平であるため、妥当性のあるモデルだと考えており、今後の研究を考え、高い精度を得られるように改良する必要がある。モデルを考えると、付録4, 5, 6に示される単語モデルの認識結果より、最適経路がわかることから、パスはつながっているといえる。今後は同じモデルを用い、HMMの条件をかえて認識率の調査を行う。また、原理上問題があるが、出力音声に対し、男女のHMMを評価し出現確率が高い話者のHMMを選択し、計算する手法により3次元ビタビ探索を用いた手法の調査ができる。

3点目は、HMMの状態数3において、男女2話者が同時に遷移する場合しか考慮していない点である。本研究では、1状態同士、2状態同士のように、加算しているが、本来は男性HMMの1状態目と女性HMMの2状態目のように状態のずれを考慮する必要がある。

例を示し説明するために、図20に状態数3の音素を加算する際に考えられるパスを示す。男性話者の音素を「n」とし、女性話者の音素を「k」とし、交差された状態をcとする。図のn_1は、男性音素「n」の1状態目を指す。c1は、男性音素「n」の1状態目と女性音素「k」の1状態目が交差した場合を指す。

本研究のHMMの加算は、男女2話者のHMMの同じ状態でしか加算を行っていない。図で示すと、状態c1と状態c5と状態c9しか使われていない。本来は、状態c1からc9ま

で考慮する必要があるため、混合分布の加算は、1 音素の加算で9つ必要になる。状態遷移確率は、両方自己ループ、片方自己ループし片方遷移、両方遷移の3つの場合があるため、男女の HMM の状態遷移確率の積をする必要がある。一方で、状態数1 混合分布数1 の場合では、状態 c1 しか使われないので、混合分布の加算は1つのみとなる。そのため、本研究における 3state 4mixture の精度が 1state 1mixture の精度と差がなかったと考えている。今後は、1 音素の加算において9つの状態の加算を行う予定であるが、音声認識ツール HTK では、3 状態の音素を表現するために、音素を分けて定義する必要があり、セグメンテーションエラーを起こす可能性がある。

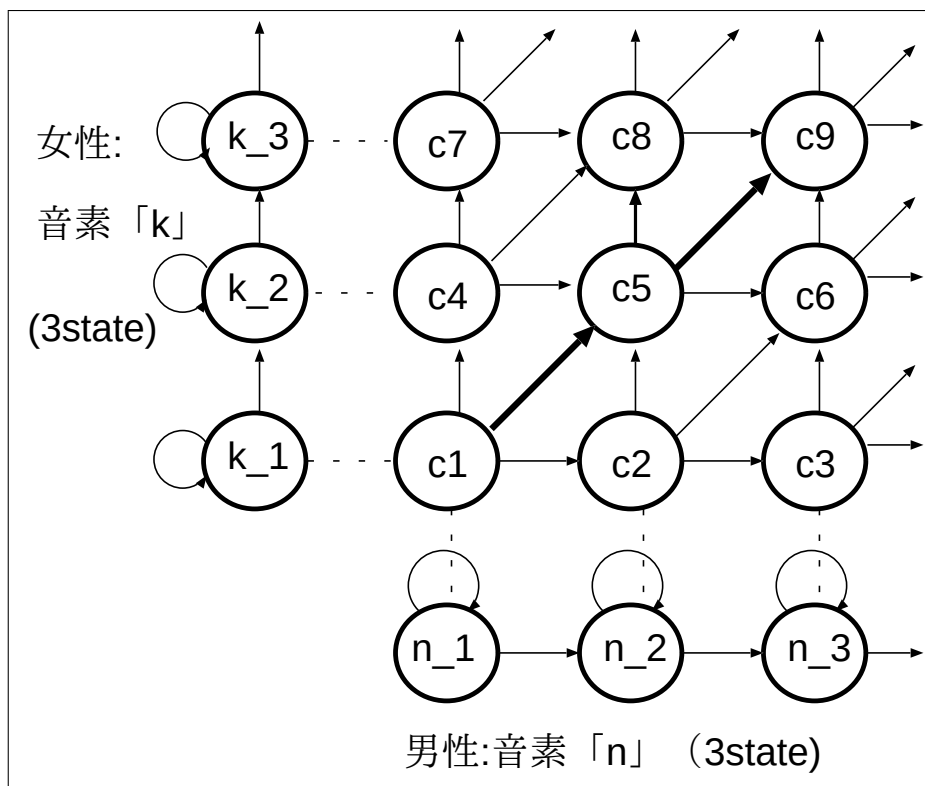


図 20: HMM のパスの考えられるパスの遷移 (3state)

6.1.2 認識傾向

提案手法の認識結果を見て、提案手法の有効性を調べる。まず、同時発話認識できた単語の数を調べ、マルチパス法と単純法と比較し、傾向を調べる。単語ごとに、同時発話認識できた割合を調べた。10単語中、何単語認識できたかを示す。男性話者の単語は表19、女性話者の単語は表20である。「mau+ftk」,「mau+fyn」,「mms+ftk」,「mms+fyn」の4つの場合の平均を示す。

例えば、提案手法で「悪質」は10単語の20.0%で2単語認識できたことになる。

表 19: 認識の割合 (男性)

男性話者単語	頻度 (提案手法)	頻度 (マルチパス法)	頻度 (単純法)
悪質 (akusitsu)	20.0%	40.0%	57.5%
聞こえる (kikoeru)	17.5%	35.0%	40.0%
加える (kuwaeru)	37.5%	72.5%	32.5%
失恋 (shitsureN)	5.0%	25.0%	50%
垂直 (suichoku)	0.0%	22.5%	35.0%
そのうち (sonouchi)	10.0%	40.0%	55.0%
中毒 (chuudoku)	5.0%	12.5%	30.0%
内容 (naiyou)	7.5%	55.0%	27.5%
暴力 (bouryoku)	5.0%	37.5%	35.0%
論じる (roNjiru)	25.0%	37.5%	60.0%

表 20: 認識の割合 (女性)

女性話者単語	頻度 (提案手法)	頻度 (マルチパス法)	頻度 (単純法)
足元 (asimoto)	7.5%	27.5%	37.5%
可愛い (kawaii)	25.0%	35.0%	37.5%
勤勉 (kiNbeN)	30.0%	38.0 %	45.0%
答える (kotaeru)	0.0%	35.0 %	40.0%
すまない (sumanai)	7.5%	37.5 %	37.5%
対策 (taisaku)	15.0%	50.0 %	40.0%
手拭い (tenugui)	10.0%	50.0 %	45.0%
天才 (teNsai)	5.0%	30.0 %	30.0%
微笑む (hohoemu)	12.5%	35.0 %	20.0%
洋服 (youfuku)	20.0%	40.0 %	40.0%

結果より以下のことが得られた。

1. 音素「k」からはじまる単語が男女ともに認識されているものが多い。
2. マルチパス法、単純法に比べ、認識できた単語に偏りがある。

6.1.3 認識できた単語

表 19 と表 20 より，単純法，マルチパス法に比べ，認識できた単語に偏りがあり．認識に用いる単語を変更すると，認識率が大きく変化すると考えられる．提案手法の信頼度は低い

6.2 PMC 法の有効性

認識結果より，HMM の条件をかえても認識率に伴わなかったため，見直す必要がある．パラメータ合成は，また，従来の PMC 法は雑音と音声を対象としているので，雑音が重畳した音声を作成し，雑音 HMM と音声 HMM を合成したモデルで認識を行い，，認識精度が改良されることを確認してから行う必要があると考える．

7 おわりに

本研究は、クロストーク音声認識において、3次元ビタビ探索を用いたモデルにより認識を行う手法を提案した。1つの状態に用いる男女のHMMの合成は、スペクトル領域での混合分布の加算を行う方法とパラメータ合成を用いたPMC法を用いて行った。スペクトル領域での混合分布の加算を行う方法による結果は、特徴パラメータFBANKにおける提案手法では13.0%となり、マルチパス法による結果の39.5%に比べ低い結果となった。特徴パラメータMFCCでは21%となり、マルチパス法による結果の48%に比べ低い結果となった。本研究では、男性の音声と女性の音声は加算の関係であるという仮説を立てて研究を行ったが、実際はスペクトル領域で加算の関係にない可能性があり、今後は、特徴パラメータを変更して可能性を探る。

また、PMC法による結果は、最大で10%となり、同じ条件によるマルチパス法の結果の48%に比べ低い結果となった。PMC法も同じ仮説で行われているが、同じく低い結果になったので、今後は、従来のPMC法を行うために、音声と雑音を用いた実験を行う予定である。

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研究室の村田教授と村上准教授と徳久助手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、鳥取大学知能情報工学科知識A研究室の清水 忠昭 助教授に感謝の意を表します。

また、本稿を執筆するにあたり参考にさせていただいた論文、本の著者の方々、計算機C研究室のみなさまの多大なる御協力に感謝の意を表します。

参考文献

- [1] P. Heracleous, S. Nakamura, T. Yamada, and K. Shikano. “ A Microphone Array-Based 3-D N-Best Search Method for Recognizing Multiple Sound Sources. ” IEICE Transactions on Information and Systems Vol.E85-D, No.6, pp.994-1002, 2002.
- [2] 岡本：“PMC 法を用いたクロストーク音声認識” 修士論文, 2007
- [3] 武藤, 杉山. “ 時間拘束条件下での重畳音声分解法. ” 日本音響学会講演論文集, No.3-3-21, pp.135-136, 2001.
- [4] 滝口, 西村. “ HMM 合成法を用いた混合音声の認識. ” 日本音響学会講演論文集, No.2-Q-12, pp.113-114, 2000.
- [5] P. Heracleous, S. Nakamura, T. Yamada, and K. Shikano. “ A Microphone Array-Based 3-D N-Best Search Method for Recognizing Multiple Sound Sources. ” IEICE Transactions on Information and Systems Vol.E85-D, No.6, pp.994-1002, 2002.
- [6] Seve Young, etc:HTK Ver3.2 reference manual,2002 Cambridge University
- [7] 古井 貞熙, 音声情報処理, 森北出版株式会社
- [8] 鹿野 清宏、伊藤 克亘, 河原 達也:” 音声認識システム”, オーム社,2001
- [9] 村上:”確率的言語モデルによる自由発話認識に関する研究” 博士論文, 2001
- [10] T. Yamada, S. Nakamura, and K. Shikano. “ Hands-free Speech Recognition Based on 3-D Viterbi Search Using a Micophone Array. “ ICASSP98,pp.245-248(1998).
- [11] M.J. Gales, and S.J. Young.“ Robust Continuous Speech Recognition Using Parallel Model Combination. ” IEEE Transactions on Speech and Audio Processing, Vol.4, No.5, pp.352-359, 1996.

付録

1. 認識結果 (提案手法, 特徴パラメータ FBANK, 男性 mau 女性 ftk, 3state4mixture)
2. 認識結果 (マルチパス法, 特徴パラメータ FBANK, 男性 mau 女性 ftk, 3state4mixture)
3. 認識結果 (単純法, 特徴パラメータ FBANK, 男性 mau 女性 ftk, 3state4mixture)
4. 提案手法の単語モデル別の認識結果 (男性話者 akushitsu と女性話者 ashimoto)
5. 提案手法の単語モデル別の認識結果 (男性話者 akushitsu と女性話者 kawaii)
6. 提案手法の単語モデル別の認識結果 (男性話者 akushitsu と女性話者 kiNbeN)