概要

本研究におけるクロストークとは，男性話者と女性話者の 2 話者が別々の孤立単語を同時に発声する状況を想定している．

クロストーク音声認識は技術的に困難な課題であり，従来，研究例が少ないが，現実の音声認識では重要な技術の一つである．

複数の話者が同時に話したときに，各話者ごとに音声の認識を行う場合，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，人間では 1 つの耳だけで複数の音声を聞き分けることが出来る．このように複数話者の重畳音声を認識する場合に単 1 のマイクロフォンで音声認識を行う研究例は少ない．類似した研究として，音声を分離する手法 [2] や，HMM 合成法を用いた手法 [3] が提案されている．

本研究では，男性話者と女性話者の 2 話者が同時に発話した場合に，単 1 のマイクロフォンを使用した状況を想定し，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の調査を行う．まず男女個別のモデルを利用して，単純な方法で認識実験を行う．また，雑音が重畳した音声を認識する方法である Parallel Model Combination 法とマルチパス法を用いて認識実験を行う．

Parallel Model Combination 法 [4] は，雑音が重畳した音声を認識する一般的な方法である．無雑音音声の HMM と雑音の HMM から目的の雑音環境の音声モデルを合成し，認識を行う手法である．

本研究では，Parallel Model Combination 法をクロストーク音声認識に適応させる．

具体的には，雑音モデルをクロストーク音声の片側音声だと考え，男性話者の音素 HMM と女性話者の音素 HMM から PMC 法のモデルを合成し認識を行う．

マルチパス法に関しても基本的なアルゴリズムは Parallel Model Combination 法と同様である．

実験の結果，単純法で 56 %，Parallel Model Combination 法で 10 %，マルチパス法で 45 %の認識精度が得られた．

目 次

1 はじめに **1**

2 音響分析 **2**

2.1 特徴抽出 2

2.2 ケプストラム分析 4

2.3 MFCC 5

2.4 FBANK 5

3 HMM による音声認識 **6**

3.1 HMM とは 6

3.2 HMM を用いた音声認識 7

3.2.1 HMM の例 7

3.3 HMM の種類 10

3.3.1 離散型 HMM(Descrete HMM) 10

3.3.2 連続型 HMM(Continuous HMM) 10

3.3.3 半連続分布型 HMM(Semi-continuous HMM) 11

3.4 HMM 法の利点と問題点 12

3.5 認識アルゴリズム 13

3.5.1 Viterbi アルゴリズム 14

3.5.2 Forward アルゴリズム 15

3.6 離散 HMM のパラメータ推定 17

3.7 連続 HMM のパラメータ推定法 18

3.7.1 出現確率が単一 (多次元) ガウス分布で表される場合 18

3.7.2 出現確率が混合ガウス分布で表される場合 18

3.7.3 半連続 HMM の場合 19

3.8 連結学習 20

4 クロストーク音声における従来の研究 **21**

5 本研究での認識手法 **22**

5.1 単純法 22

5.2 Parallel Model Combination 法 23

5.3 マルチパス法 25

6 評価実験 **27**

6.1 評価データと学習データ 27

6.2 実験条件 30

7 実験結果 **33**

7.1 単純法の実験結果 34

7.1.1 単純法において認識成功の例 36

7.1.2 単純法において認識失敗の例 37

7.2 マルチパス法の実験結果 38

7.2.1 マルチパス法において認識成功の例 40

7.2.2 マルチパス法において認識失敗の例 42

7.3 PMC 法の実験結果 43

7.3.1 PMC 法において認識成功の例 44

7.3.2 PMC 法において認識失敗の例 45

8 考察 **47**

8.1 人手による聴覚実験 47

8.2 誤認識に対する考察 48

8.3 PMC 法の精度 49

9 おわりに **50**

10 謝辞 **51**

図 目 次

1 left-to-right モデルの例 8
2 単語 HMM を用いた単語音声認識の方法 13
3 トレリス上の $\alpha_t(i)$ の計算 16
4 連結学習の例 20
5 単純法での認識の様子 22
6 本研究での PMC 法のモデル 23
7 認識結果の求め方 24
8 本研究でのマルチパス法のモデル 25
9 クロストーク音声認識の手順 28
10 音素 HMM の作成と学習及び認識の流れ 29
11 認識結果のマルチパス法のモデル 40
12 認識結果のマルチパス法のモデル 41
13 認識結果のマルチパス法のモデル 42
14 認識結果の PMC 法のモデル 44
15 認識結果の PMC 法のモデル 45
16 認識結果の PMC 法のモデル 46

表 目 次

1 dic ファイルの例 26
2 gram ファイルの例 26
3 実験に使用した単語 27
4 実験条件 (単純法, マルチパス法) 31
5 実験条件 (PMC 法) 32
6 実験結果 (単純法) 33
7 実験結果 (マルチパス法) 33
8 実験結果 (PMC 法) 34
9 単純法の認識結果 35
10 36
11 36

12 37
13 37
14 マルチパス法の認識結果 39
15 40
16 41
17 42
18 PMC 法の認識結果 43
19 44
20 45
21 46
22 聴取実験の認識率 47
23 聴取実験の認識率 47
24 例：人間で認識， 計算機で誤認識 48
25 例：計算機で認識， 人手で誤認識 48

1　はじめに

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，違う声の大きさで発話したとき，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．このようなシステムの初歩として，クロストーク音声認識があげられる．このクロストーク音声とは，2 話者が同時に発声する状況を想定している．しかし，クロストーク音声認識は技術的に困難な課題であり，従来，研究例が少ないが，現実の音声認識では重要な技術の一つである．

複数の話者が同時に話したときに，各話者ごとに音声の認識を行う場合，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，人間では 1 つの耳だけで複数の音声を聞き分けることが出来る．このように複数話者の重畳音声を認識する場合に単 1 のマイクロフォンで音声認識を行う研究例は少ない．類似した研究として，重畳音声を分離する手法 [2] や，HMM 合成法を用いた手法 [3] が提案されている．

過去の研究では，男女 2 話者の単独同時発話を対象に，現状の技術を用いた認識率の実験的評価が行われているが，実験対象とする単語数が多いこともあって，低い認識率にとどまっている．また，実験では，片側音声を対象とした単独認識率のみが評価されており，両側音声の同時発話認識率は不明であった [5]．

以前の研究では，認識対象単語数と認識率の関係を調べるため，10 単語を対象とした認識実験を行い，同時発話認識率についても評価した [6]．

本研究では，男性話者と女性話者の 2 話者が同時に発話した場合に，単 1 のマイクロフォンを使用した状況を想定し，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の調査を行う．まず男女個別のモデルを利用して，単純な方法で認識実験を行う．また，雑音が重畳した音声を認識する方法である Parallel Model Combination 法とマルチパス法を用いて認識実験を行う．

結果として，最も認識精度が高かった実験は，単純法 MFCC Full-covariance において，認識率が 56%であった．また，人間による聴覚実験と比較すると誤り率で 2 倍程度の認識率が得られることがわかった．

2　音響分析

音声は音声生成のモデルのパラメータによって効率よく表現され，音声の音響音韻的性質はこのパラメータによって特徴づけられる．音声生成のモデルのパラメータのように先生の音響音韻的な性質を持つパラメータを音響パラメータと呼ぶ.

音声の音響音韻的な性質は音響パラメータによって特徴づけられるが，音声を構成する言語音の音韻識別には，音響パラメータの全データが必要になるわけではない．音韻識別に必要な部分を特徴パラメータと呼ぶ.

従来経験的に，音声情報はそのスペクトルによって特徴づけられることが知られてきた. その 1 つの表現がフォルマント構造である．音声波形のスペクトルが複数個の共振周波数の存在によって特徴づけられることは古くから知られており，その共振を周波数の低い方から順番に「第 1 フォルマント」,「第 2 フォルマント」，…と名付けられている．

波形とスペクトルの関係は原理的にはフーリエ変換で記述できる．従来その処理は，帯域フィルタ群による周波数分析によって近似的に実現されてきた. 最近になって, 計算機による高速フーリエ変換 (FFT) の技術が実用化され，ディジタル化された波形からそのスペクトルを高速フーリエ変換によって直接求めることがスペクトル分析の主流になっている．

2.1　特徴抽出

音声は, 様々な音素に対応する言語音から構成されており，信号の性質が常に変化している非定常信号であるが，100 分の 1 秒程度の短時間区間では一応定常的な信号とみなすことが出来るので，音声信号のスペクトル分析において定常過程に対するスペクトル推定の方法を利用することが出来る．

人間も音声を聞きとる際に，スペクトル分析を行っていると考えられている．認識においても短時間スペクトル分析が重要であると考えられる．音声認識を行うためには，まず， 音声区間の検出を行うことが必要であり，音声は，声帯による音源（有声音源，無声音源）の成分に喉から口にかけての声道の形状によって調音されることによって生成される．このため，音声の短時間スペクトルは，音源に対応する，周波数方向に細かく変化する成分（微細構造）と，声道の形状による調音に対応する，緩やかに変化する成分（スペクトル包絡）の積となる．

音声の認識において重要な音韻性の識別に必要な情報は，スペクトル包絡に集中している．このため，短時間スペクトルからスペクトル包絡を抽出する方法が重要となる．

スペクトル分析の手法としては，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，短時間スペクトル (密度) を抽出して用いる．切り出された信号が定常確率過程に従うと仮定して与えられた信号 $s(n)$ 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．

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

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っ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 $s_w(n)(n = 0, \cdots, N-1)$ 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．

$$\text{ハミング窓} : w(n) = 0.54 - 0.46\cos(\frac{2n\pi}{N-1})(n = 0, \cdots, N-1) \tag{2}$$

$$\text{ハニング窓} : w(n) = 0.5 - 0.5\cos(\frac{2n\pi}{N-1})(n = 0, \cdots, N-1) \tag{3}$$

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．

$$S(e^j\omega) = \sum_{n=0}^{N-1} s_w(n)e^{-j\omega n} \tag{4}$$

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 t が一般的である．すなわち

$$S'(k) = S(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} s_w(n)e^{-j\frac{2\pi}{N}kn}(k = 0, \cdots, N-1) \tag{5}$$

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．FFT の結果に対して，各周波数の大きさ成分を二乗してパワースペクトルに変換することが多い．

2.2　ケプストラム分析

もし，音声の言語情報が声道の形状による共振特性によって担われていると仮定すれば，分析によって抽出する特性は，まずそのスペクトル包絡である．

時間 (波形) 的には，音声波形は音源波形と声道共振系のインパルス応答との畳込みで表現される．したがって，周波数次元では両者の特性の積で表される．

音源と共振系の特性を分離して抽出する方法は逆畳込みと呼ばれる．

その方法の 1 つがケプストラム (cepstrum) 分析である．

ケプストラム $c(\tau)$ は, 波形の短時間振幅スペクトル $|S(e^{j\omega})|$ の対数の逆フーリエ変換として定義される．音源のスペクトラムを $G(e^{j\omega})$ ，声道共振系のインパルス応答の伝達特性を $H(e^{j\omega})$ とすると次の関係が得られる．

$$S(e^{j\omega}) = G(e^{j\omega})H(e^{j\omega}) \tag{6}$$

この対数を取ると，

$$log|S(e^{j\omega})| = log|G(e^{j\omega})|log|H(e^{j\omega})| \tag{7}$$

となる．次にこれをフーリエ逆変換すると，

$$c(\tau) = \mathcal{F}^{-1}log|S(e^{j\omega})| = \mathcal{F}^{-1}log|G(e^{j\omega})| + \mathcal{F}^{-1}log|H(e^{j\omega})| \tag{8}$$

となり，これがケプストラムである．

離散フーリエ変換 (DFT) で求めると，

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} log|S(k)|e^{j2\pi kn/N}(0 \leq n \leq N-1) \tag{9}$$

となる．

従来の音声認識では, 特徴パラメータとしてケプストラムが使われてきた，ケプストラムは低次にフォルマント情報を高次にピッチ情報を含んでいる．しかしピッチ情報は正確なピッチ周波数の抽出が困難であるため，音声認識ではフォルント情報しか用いられていない．

2.3 MFCC

ケプストラムパラメータには, 多様な計算方法がある． その中には MFCC(メル周波数ケプストラム係数) がある．MFCC 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．これは，言い換えれば声道特性のみを用いていることになる．

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j - 0.5)) \tag{10}$$

N はフィルタバンクチャンネの数を表し，m_j は対数フィルタバンクの振幅を表す.

本研究では MFCC12 次+対数パワーの形で用いる．

2.4 FBANK

人の聴覚は，音の高さに関して，メル尺度に近い非線型の特性を示し，低い周波数では細かく，高い周波数では荒い周波数分解能力を持つ.

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，パワーケプストラムの全域に, 人間の聴覚の特性にあわせて低周波部分は細かく，高周波部分は大まかに調べるた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

$$Mel(f) = 2592 \log_{10}(1 + \frac{f}{700}) \tag{11}$$

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.

FBANK は混合ガウス分布に Full-covariance を用いた場合に MFCC よりも認識率が高いことが知られている [7].

本研究において，基本周波数 16KHz の音に対して FBANK24 次+対数パワーの形で用いる.

3　HMMによる音声認識

人間は，日常のコミュニケーションの大半を音声を介して行う．人と計算機のインターフェースを人にとって容易かつ自然なものにするには音声メディアの利用と情報処理技術にかかっている．特に，計算機による音声認識技術が中核を担うことになる．

単語ごとに区切って発声した音声を認識することを孤立単語音声認識といい，通常，これを簡単に単語音声認識と呼ぶ.

単語音声の認識には,DP マッチング (Dynamic Programming Matching) による方法，セグメンテーションと音素ラベリングに基づく方法，HMMM(Hidden Markov Model) による方法，ニューラルネットワークによる方法などが利用される．

HMM による方法は，米国で統計確率的手法である HMM の研究が行われ，1980 年代には単語音声認識の標準的手法となった．HMM は，その統計的アルゴリズムの高い学習能力と認識性能により，現在では広く使われるようになってきている．

3.1　HMMとは

HMM(隠れマルコフモデル) とは，外から観測できるものがモデルによって生成された出力データ系列だけであって，一般にモデルの内部の状態とその遷移の様子は外から見られないことから付けられた呼称である．

音声パターンは時系列の形で表され，様々な原因により変動がある． 音響パラメータの時系列は変動分を含み，このようなパターンの確率的な性質は HMM によって精密に表現できると考えられている． HMM は非定常信号源を定常信号源の連結で表す．

3.2 HMM を用いた音声認識

音声認識は, パターン認識の一分野である．音声波形から認識に有効な特徴パラメータが抽出された後は，通常のパターン認識の技術と本質的に変わりはない．通常のパターン認識との違いは，音声パターンが時系列パターンであることと言語情報の制約を受けることである．パターン認識には構造的･構文的パターン認識法と統計的･確率的パターン認識法が存在する。最近になって，音声パターンの時系列パターンに対しての統計的･確率的パターン認識法が HMM(Hidden Markov Model; 隠れマルコフモデル) による手法である．

HMM は, 出力シンボルによって一意に状態遷移先が決まらないという意味での非決定状態オートマトンとして定義される．このモデルでは, 状態と出力シンボルの 2 課程を考え，状態が確率的に遷移するときに対応して確率的にシンボルを出力する．このとき観測できるのはシンボル系列だけであることから Hidden(隠れ) マルコフモデルとよばれている．

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 $\sum_j a_i j = 1$.
- 出力確率の集合;$B = \{b_{ij}(k)\}$; $b_i(k)$ は状態 s_i から においてシンボル k を出力する確率.
- 初期状態確率の集合;$\pi = \{\pi_i\}$;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

3.2.1 HMM の例

音声認識に用いられる HMM は，left-to-right モデルと呼ばれるものである．left-to-right モデルの例を図 1 に示す．

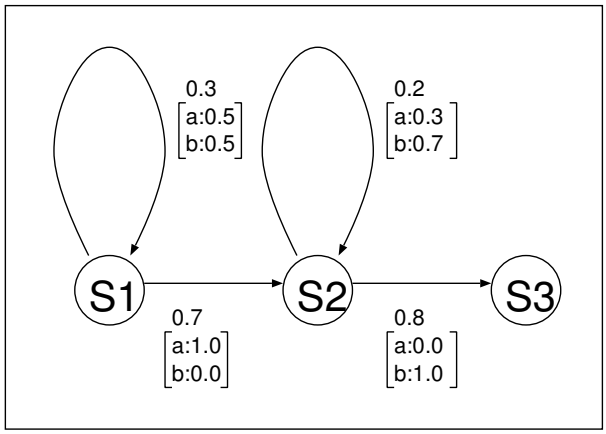

図 1: left-to-right モデルの例

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．

$$\pi = \begin{pmatrix} 1.0 & 0 & 0 \end{pmatrix} \tag{12}$$

状態遷移確率の集合 A は以下であり，図では [] 上部の数字で示される．

$$A = \begin{pmatrix} 0.3 & 0.7 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \tag{13}$$

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

$$B_a = \begin{pmatrix} 0.5 & 1.0 & 0.0 \\ 0.0 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \tag{14}$$

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．

$$B_b = \begin{pmatrix} 0.5 & 0.0 & 0.0 \\ 0.0 & 0.7 & 1.0 \\ 0.0 & 0.0 & 0.0 \end{pmatrix} \tag{15}$$

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”aab” である場合，可能な状態遷移系列は $s_1s_1s_2s_3$ と $s_1s_2s_2s_3$ 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．

$$0.3 * 0.5 * 0.7 * 1.0 * 0.8 * 1.0 = 0.084 \tag{16}$$

$$0.7 * 1.0 * 0.2 * 0.3 * 0.8 * 1.0 = 0.0336 \tag{17}$$

よって, この HMM が”aab” を出力する確率は以下のようになる.

$$0.084 + 0.0336 = 0.1176 \tag{18}$$

3.3 HMM の種類

HMM にはスペクトルパターンの表現方法により，離散型 HMM，連続分布型 HMM，半連続分布型 HMM に大別される．以下にそれぞれの特徴を述べる．

3.3.1 離散型 HMM(Descrete HMM)

出現するスペクトルパターンは，有限個のシンボルの組み合わせで表現される．出力確立は，スペクトルパターンのクラスタ化 (ベクトル量子化) によって，代表スペクトルパターン (符号ベクトル) を生成し，各符号ベクトルの出現確立の組み合わせによって表現する．

3.3.2 連続型 HMM(Continuous HMM)

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

出現するスペクトルパターンを連続値として表す分布モデルである．出現確率を表す方法としては単一ガウス分布や混合ガウス分布が用いられる．パラメータの自由度を減らすために無相関ガウス分布を用いることが多い．

出現確率 $b_ij(o_t)$ が混合ガウス分布に従う場合は，

- M_{ij}...状態 i から状態 j の遷移における混合数
- C_{ijm}...状態 i から状態 j の遷移における混合数のときの重み
- $\mathcal{N}(;\mu,\Sigma)$...平均ベクトルμ, 共分散行列Σをもつ混合ガウス分布

とすると，以下のように計算される．

$$b_{ij}(o_t) = \sum_{m=1}^{M_{ij}} C_{ijm}\mathcal{N}(o_t;\mu_{ijm},\Sigma_{ijm}) \tag{19}$$

$\mathcal{N}(;\mu,\Sigma)$ は

- n...観測行列の次元数
- $(O-\mu)^t$...$(O-\mu)$ の天地行列

- $|\Sigma|$ ...Σの固有値
- Σ^{-1}...Σの逆行列

とすると，以下の式で表現される．

$$\mathcal{N}(O;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n \mid \Sigma \mid}} \exp(-\frac{1}{2}(O-\mu)^t \Sigma^{-1}(O-\mu)) \tag{20}$$

3.3.3 半連続分布型 HMM(Semi-continuous HMM)

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

$$b_{ij}(x) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(x) \tag{21}$$

と混合正規分布で表す．ただし，

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{22}$$

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

3.4 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.

3.5　認識アルゴリズム

$y = y_1, y_2, \cdots, y_T$ を観測 (出力) 系列とする．具体的には，スペクトルやケプストラムの時系列である．このとき，各 HMM モデルによって y が生起する確率 (尤度)P(y | M)(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出しこれを認識結果とする．図 2 に単語 HMM を用いた認識方法を示す．

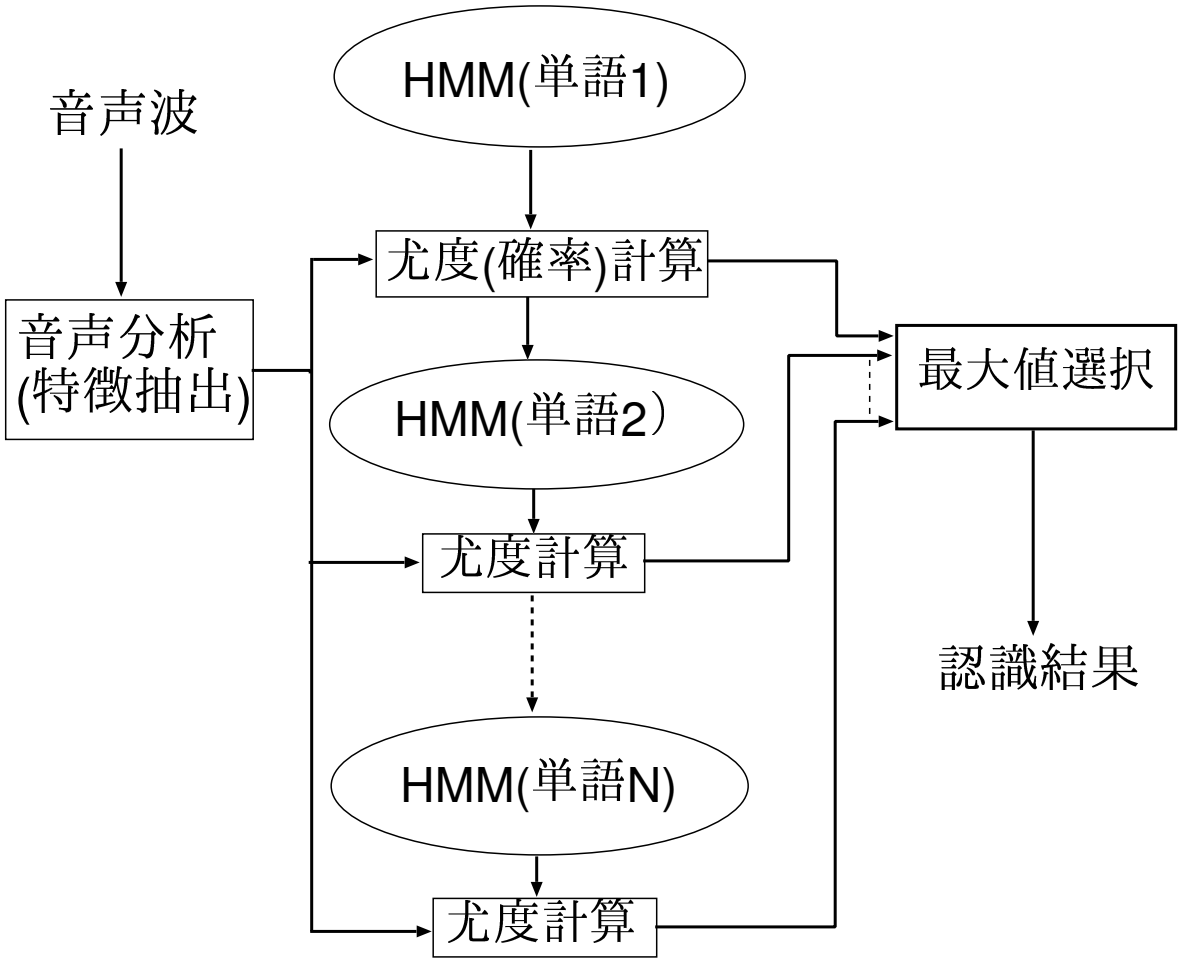

図 2: 単語 HMM を用いた単語音声認識の方法

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

$$P(y \mid M) = \sum_{i_0, i_1, \cdots, i_T} P(y \mid q, M) \cdot P(q \mid M) \tag{23}$$

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

フォワード変数 $\alpha\ (i, t)$ 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，$i = 1, 2, \cdots, S$ とおいて，以下の式を得る．

$$\alpha\ (i, t) = \sum_{j} \alpha\ (j, t-1) \cdot \alpha_{ji} \cdot b_{ji}(y_t)(t-1, 2, \cdots, T) \cdot \pi_i(t=0) \tag{24}$$

これを計算し，最後に以下を求めれば良い．

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

3.5.1 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．

$P(y \mid M)$ 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．$i = 1, 2, \cdots, S$ において，

$$f'(i, t) = \begin{cases} \log \pi_i & (t = 0) \\ \max_j \{f'(i, t-1) + \log a_{ji} b_{ji}(y_t)\} & (t = 1, 2, \cdots, T) \end{cases} . \tag{26}$$

を計算し，対数尤度

$$L = \max_{i,\ s_i \in F} f'(i, t) \tag{27}$$

を求める．

この Viterbi アルゴリズムは対数を用いた計算なので，trellis 法を用いる場合に比べ以下のような利点がある．

- 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に用意に適用できる．

このために Viterbi アルゴリズムは広く用いられている．

Viterbi アルゴリズムは，本実験において HMM の初期モデル作成と認識に使用されている．

3.5.2 Forward アルゴリズム

時刻 t の時に $o_1, o_2, ..., o_t$ 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。

$$\alpha_t(j) = P(o_1, o_2, ..., o_t, s_t = j \mid \lambda) \tag{28}$$

$P(O \mid \lambda)$ 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。

1. 初期化

   全ての状態 $j(1 \leq j \leq N)$ に対して，

   $$\alpha_0(j) = \pi_j \tag{29}$$

   とする．

2. 導出過程

   時間軸 $(t = 1, \cdots, T)$ に沿って，全ての状態 $j(1 \leq j \leq N)$ に対し，$\alpha_t(j)$ を次のように計算する．

   $$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_{ij}(o_t) \tag{30}$$

3. 結果

   $$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{31}$$

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 $\alpha_t(j)$ $1 \leq j \leq N)$ を求めている。

図 3 は，前記の図 1 の HMM がラベル系列 aab を出力する例に適応した例である．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．この方法での計算量は $\mathrm{O}(N^2T)$ である。

また，$P(a, a, b \mid \lambda) = \alpha_3(3) = 0.1176$ となる．

Forward アルゴリズムは，本実験において認識に使用されている．

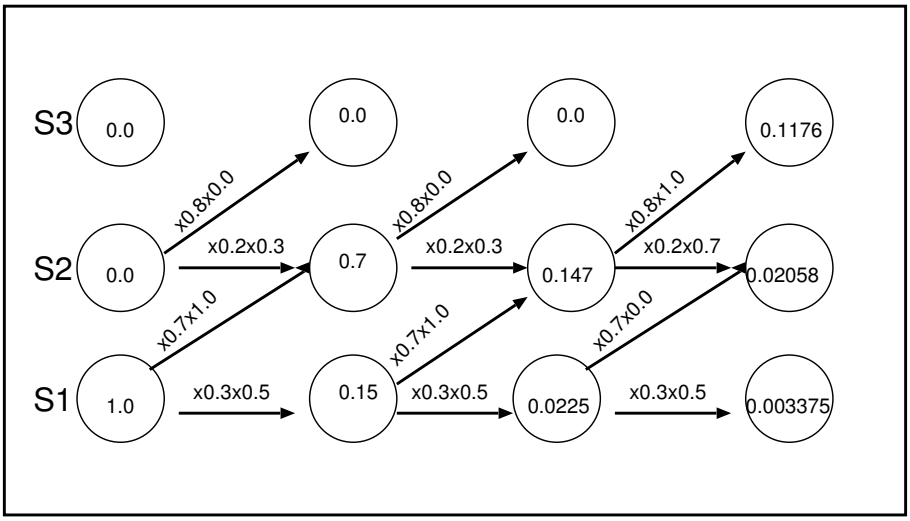

図 3: トレリス上の $\alpha_t(i)$ の計算

3.6 離散 HMM のパラメータ推定

学習用音声として，N 個の観測符号ベクトル系列 $\{y_1^{T(n)} = y_1, y_2, \cdots, y_{T(n)}\}_{n=1}^{N}$ が与えられたとき，

$$\prod_{n=1}^{N} P(y_1^T(n)|\pi_i, a_{ij}, b_{ij}(k)) \tag{32}$$

を最大化するパラメータセット $\{\hat{\pi}_i, a_{ij}, b_{ij}(k)\}$ は,Baum-Welch アルゴリズムによって，次のように推定できる．

まず以下のような変数 $\beta(i,t)$,$\gamma(i,j,t)$ を定義する．

$\beta(i,t)$: 時刻 t に状態 s_i にあって，以後符号ベクトル y_{t+1}^T を出力する確率

$\gamma(i,j,t)$: モデル M が y_1^T を出力する場合において，時刻 t に状態 s_i から状態 s_j へ遷移し符号ベクトル y_t を出力する確率

このとき，以下の関係が得られる．

$$\beta(i,T) = \begin{cases} 1 & s_i \in F \\ 0 & s_i \notin F \end{cases} \tag{33}$$

$$\beta(i,t) = \sum_j a_{ij} b_{ij}(y_t) \beta(j, t+1) \quad (t = T, T-1, \cdots, 1; i = 1, 2, \cdots, S) \tag{34}$$

$$\gamma(i,j,t) = \frac{\alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{P(y_1^t|M)} \tag{35}$$

以上を用いて，パラメータ pi_i, a_ij, $b_ij(k)$ を，以下の再推定によって求める．

$$\hat{\pi_{ij}} = \frac{\sum_j \gamma(i,j,1)}{\sum_i \sum_j \gamma(i,j,1)} \tag{36}$$

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T} \alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{\sum_t \alpha(i,t) \beta(j,t)} = \frac{\sum_t \gamma(i,j,1)}{\sum_t \sum_j \gamma(i,j,t)} \tag{37}$$

$$\hat{b_{ij}} = \frac{\sum_{t, y_t = k} \gamma(i,j,t)}{\sum_t \gamma(i,j,t)} \tag{38}$$

実際は，すべての学習サンプルに対してこの計算を行ってから 1 回パラメータを更新するというサイクルを，値が収束するまで繰り返す．

3.7 連続 HMM のパラメータ推定法

連続 HMM のパラメータ推定においては，初期確率 π_i と遷移確率 a_ij の推定式は離散 HMM の場合と同じである。

3.7.1 出現確率が単一 (多次元) ガウス分布で表される場合

出現確率のガウス分布 $N(\mu_{ij}, \Sigma_{ij})$ は次式のように最尤推定できる.

$$\hat{\mu_{ij}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t) y_t}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{39}$$

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

離散 HMM の場合と同様に，この推定を値が収束するまで繰り返す.

3.7.2 出現確率が混合ガウス分布で表される場合

混合ガウス分布の場の出現確率は，次のように表される (ガウス分布の数を M とする).

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{41}$$

ここで

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{42}$$

$$\int b_{ijm}(y) dy = 1 \tag{43}$$

である．混合ガウス分布の出現確率は，単一ガウス分布の場合と同様に次式で表せる.

$$\hat{\lambda_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{44}$$

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

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

ただし，

$$\gamma(i,j,m,t) = \alpha(i,t-1)a_i j \lambda_{ijm} b_{ijm}(y_t)\beta(j,t) \tag{47}$$

で，m 番目の分布関数の遷移 $q_i \to q_j$ の確率 (遷移回数) を表している．これらの推定も値が収束するまで繰り返す.

3.7.3 半連続 HMM の場合

符号張の中の分布数を M として，出現確率は次のようになる.

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \tag{48}$$

ここで,

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{49}$$

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 $\sum_m$ 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 $(s_i \to s_j)$ ごとに推定する．これらの推定式は,

$$\hat{\lambda_{ijm}} = \frac{\sum_{t=1}^{T} \gamma(i,j,t,m)}{\sum_{t=1}^{T} \gamma(i,j,t)} \tag{50}$$

$$\hat{\mu_{ijm}} = \frac{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m) y_t}{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{51}$$

$$\hat{\Sigma_{ijm}} = \frac{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)(y_t - \mu_{ijm})(y_t - \mu_{ijm})^t}{\sum_{all(s_i \to s_j)} \sum_{t=1}^{T} \gamma(i,j,t,m)} \tag{52}$$

となる.

3.8 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 4 に示す．音声データの音素表記 “pau a i pau” を元にして各音素 HMM を連結し，連結した HMM のパラメータを音声データから推定する．

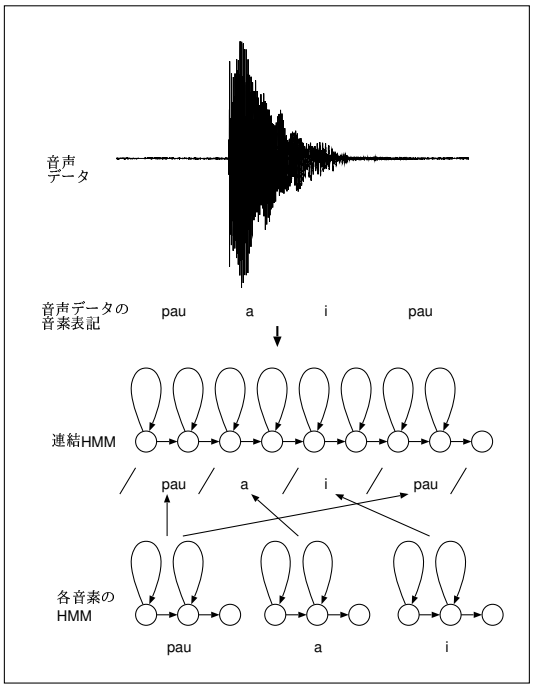

図 4: 連結学習の例

4　クロストーク音声における従来の研究

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，違う声の大きさで発話したとき，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複数のマイクロフォンを使用する手法が一般的である [1]．複数のマイクロフォンを使用することで，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．この手法は有効性が示されている．しかし，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などが挙げられる．

また，単 1 のマイクロフォンを利用する手法では，重畳音声を分離する手法 [2] や，HMM 合成法を用いた手法 [3] が提案されている．

重畳音声を分離する手法では，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．しかし，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，音声が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が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．しかし，問題点として，重畳音声の認識対象音声だけのみの認識精度しか調査されていないことが挙げられる．

本研究では，クロストーク音声の両方の音声を認識対象として，次の章で述べる 3 種類の手法で認識率の調査を行う．

5　本研究での認識手法

本研究で用いた認識手法を以下で述べる．全ての手法において男性話者と女性話者の音素 HMM の学習は，それぞれ独立に学習する．

5.1　単純法

単純法では，クロストーク音声に対し，男性話者の HMM を利用して男性話者の認識結果を求め，女性話者の HMM を利用して女性話者の認識結果を求める．最後に男性話者の認識結果と女性話者の認識結果から男性話者と女性話者が同時に認識できた場合の認識率を調査する．

本研究ではこの手法を単純法と呼ぶ．

単純法では認識に HTK[8] に用意されているプログラムを用いる．

また図 5 に単純法での認識の様子を示す．

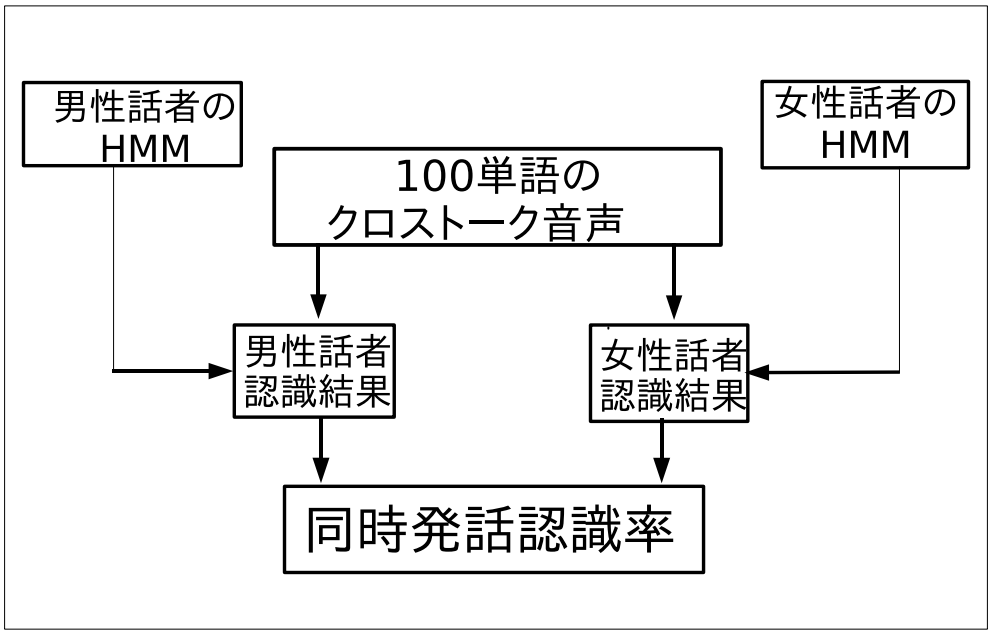

図 5: 単純法での認識の様子

5.2 Parallel Model Combination 法

Parallel Model Combination 法 [4](以下 PMC 法) は，雑音が重畳した音声を認識する一般的な方法である．無雑音音声の HMM と雑音の HMM から目的の雑音環境の音声 HMM を合成し，認識を行う手法である．

本研究では，雑音モデルをクロストーク音声の片側音声だと考え，男性話者の音素 HMM と女性話者の音素 HMM から PMC 法のモデルを合成し認識を行う．

PMC 法の各状態の尤度は，各状態に遷移されるパスの尤度の和をとる．モデルの尤度は，最終状態の尤度とする．最後に，最も尤度が高かった PMC 法のモデルを認識結果とする．

図 7 にその様子を示す．

本研究では，PMC 法のモデルは，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．

図 6 に本研究での PMC 法のモデルを示す．

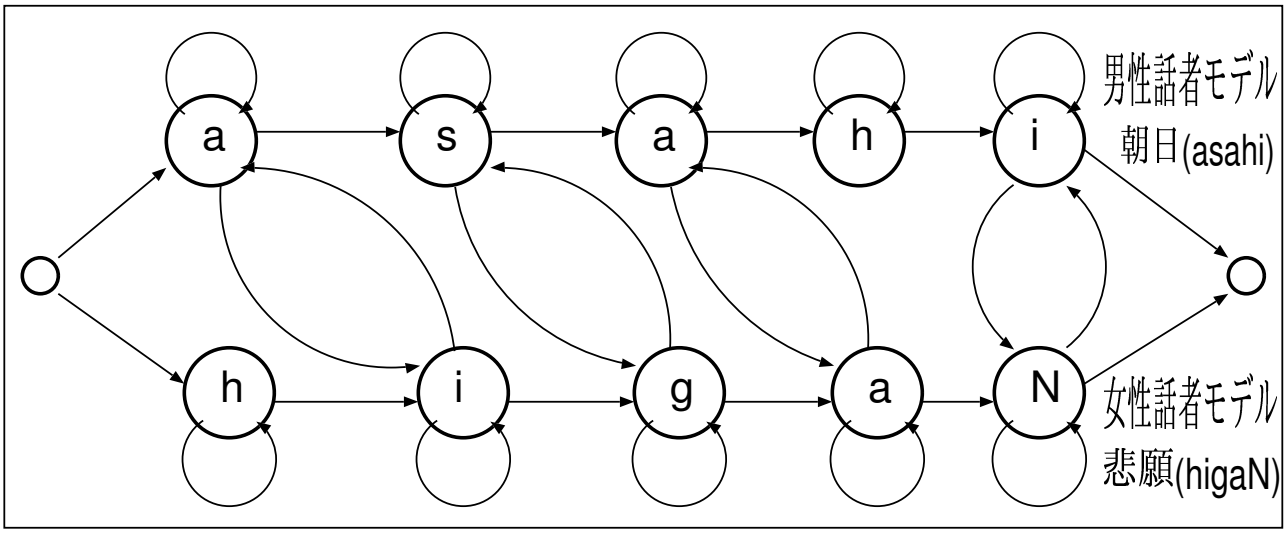

図 6: 本研究での PMC 法のモデル

図 6 は男性話者の音声に「朝日 (asahi)」，女性話者の音声に「悲願 (higaN)」を使用した場合の PMC 法のモデルである．

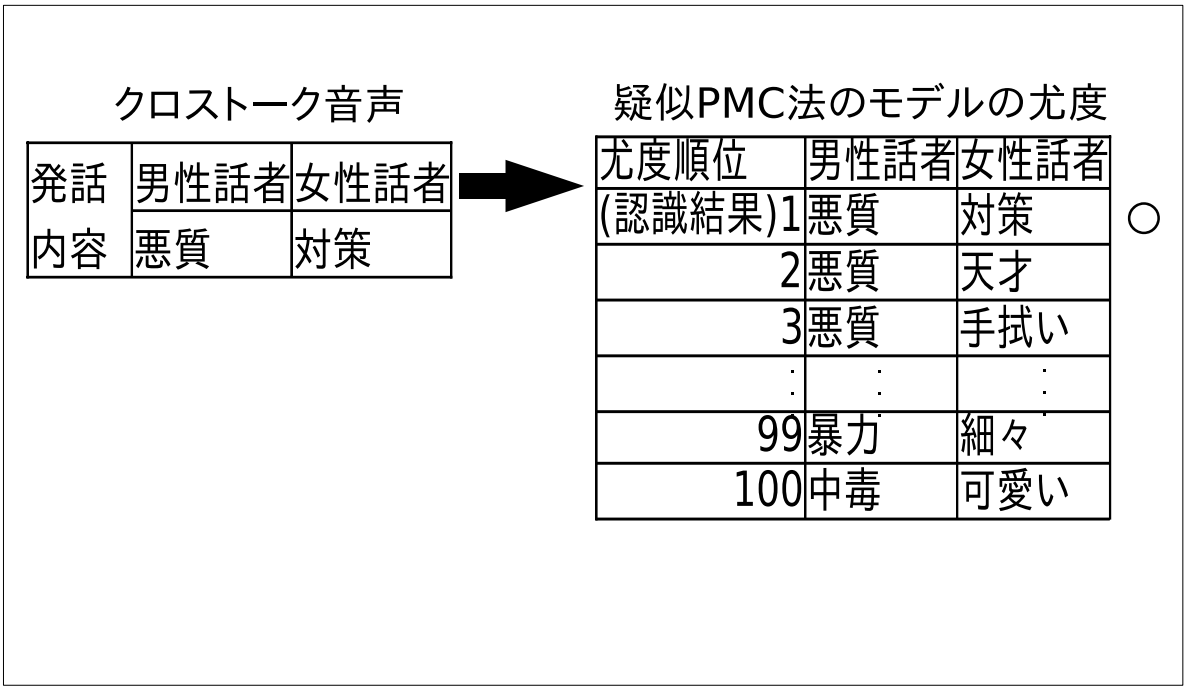

図 7: 認識結果の求め方

図 7 の例は, クロストーク音声の発話内容が, 男性話者が「悪質 (akusitsu)」, 女性話者が「対策 (taisaku)」だった場合, マルチパス法のモデルの尤度が最大のモデルも, 男性話者が「悪質 (akusitsu)」, 女性話者が「対策 (taisaku)」のモデルであり, クロストーク音声の発話内容と一致しており, この場合は正しく同時発話認識ができている.

5.3 マルチパス法

マルチパス法は，PMC 法とほぼ同様のアルゴリズムである．しかし，マルチパス法の各状態の尤度は，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．

本研究では，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 に本研究におけるマルチパス法のモデルを示す．

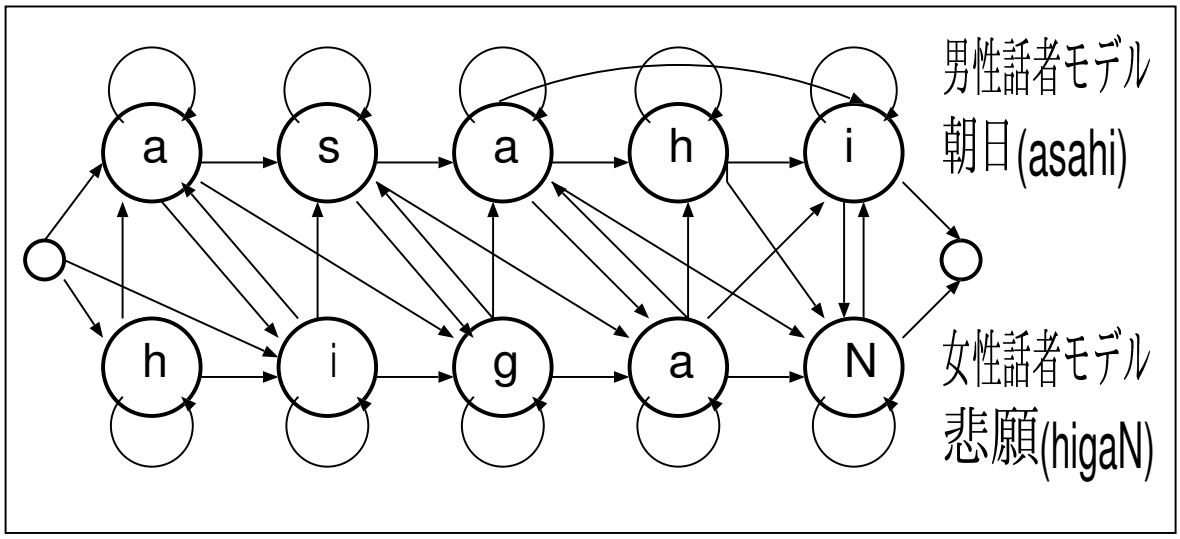

図 8: 本研究でのマルチパス法のモデル

図 8 は男性話者の音声に「朝日 (asahi)」，女性話者の音声に「悲願 (higaN)」を使用した場合のマルチパス法のモデルである．

例として，図 1 と図 2 に男性話者の発話内容が「朝日 (asahi)」，女性話者の発話内容が「悲願 (higaN)」である場合のマルチパス法のモデル場合の実際に HTK で使用した dic ファイル及び gram ファイルの内容を示す．男性話者の音素と女性話者の音素を混同させないために，男性話者の音素の前には m- を，女性話者の音素の前には f- を追加する．図 8 は，図 2 の gram ファイルで表現される．

表 1: dic ファイルの例

m-a	m-a
m-s	m-s
m-a	m-a
m-h	m-h
m-i	m-i
m-pau	m-pau
f-h	f-h
f-i	f-i
f-g	f-g
f-a	f-a
f-N	f-N

表 2: gram ファイルの例

(	m-pau	f-h	(| m-a | f-i |)	(| m-s | f-g |)	(| m-a | f-a |)	m-h	(| m-i | f-N |)	m-pau	)

6 評価実験

6.1 評価データと学習データ

図 9 にクロストーク音声認識の手順を示す．本研究では，実際に男女 2 話者が同時に発話した音声を使用せず，各々の話者の音声を重畳した音声を作成する．具体的には，ATR 単語発話データベース Aset 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．

偶数番号の音声の中から 4 モーラで発話時間がほぼ同じ語を，ランダムに 10 単語ずつ抽出する．それぞれを相互に重畳した音声 (クロストーク音声) を 100 個作成する．1 セットにつき 100 単語のクロストーク音声を 4 セット作成し，評価データとして利用する．表 3 に実験に使用した単語を示す．

ATR 単語発話データベース Aset の奇数番号の 2620 単語の音声は HMM の学習データとして使用する．

図 10 に音素 HMM の作成と学習及び認識の流れを示す．音素 HMM の作成と学習は，HTK で行う．まず最初に，男性話者と女性話者の初期モデルの作成を行う．次に，学習データを使用して，男性話者と女性話者 HMM の学習及び連結学習を行う．最後に学習された男性話者と女性話者 HMM を利用して，Viterbi アルゴリズムと Forward アルゴリズムを使用し認識を行う．

表 3: 実験に使用した単語

	男性話者	女性話者
1	悪質 (akushitsu)	足元 (ashimoto)
2	聞こえる (kikoeru)	可愛い (kawaii)
3	加える (kuwaeru)	勤勉 (kiNbeN)
4	失恋 (shitsureN)	細々 (komagoma)
5	優れる (sugureru)	すまない (sumanai)
6	そのうち (sonouchi)	対策 (taisaku)
7	中毒 (chuudoku)	手拭い (tenugui)
8	内容 (naiyou)	天才 (teNsai)
9	暴力 (bouryoku)	滅ぼす (horobosu)
10	わざわざ (wazawaza)	欲張る (yokubaru)

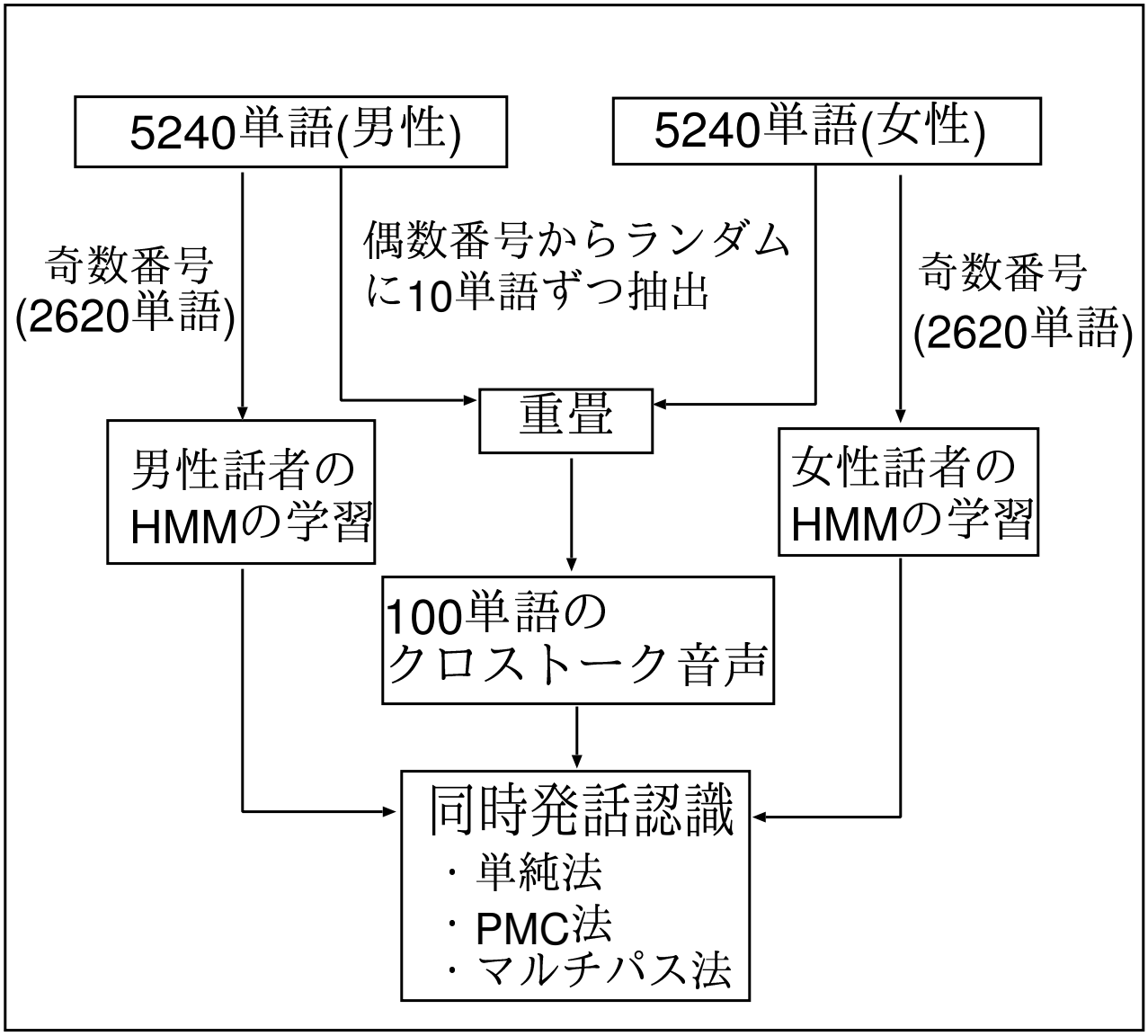

図 9: クロストーク音声認識の手順

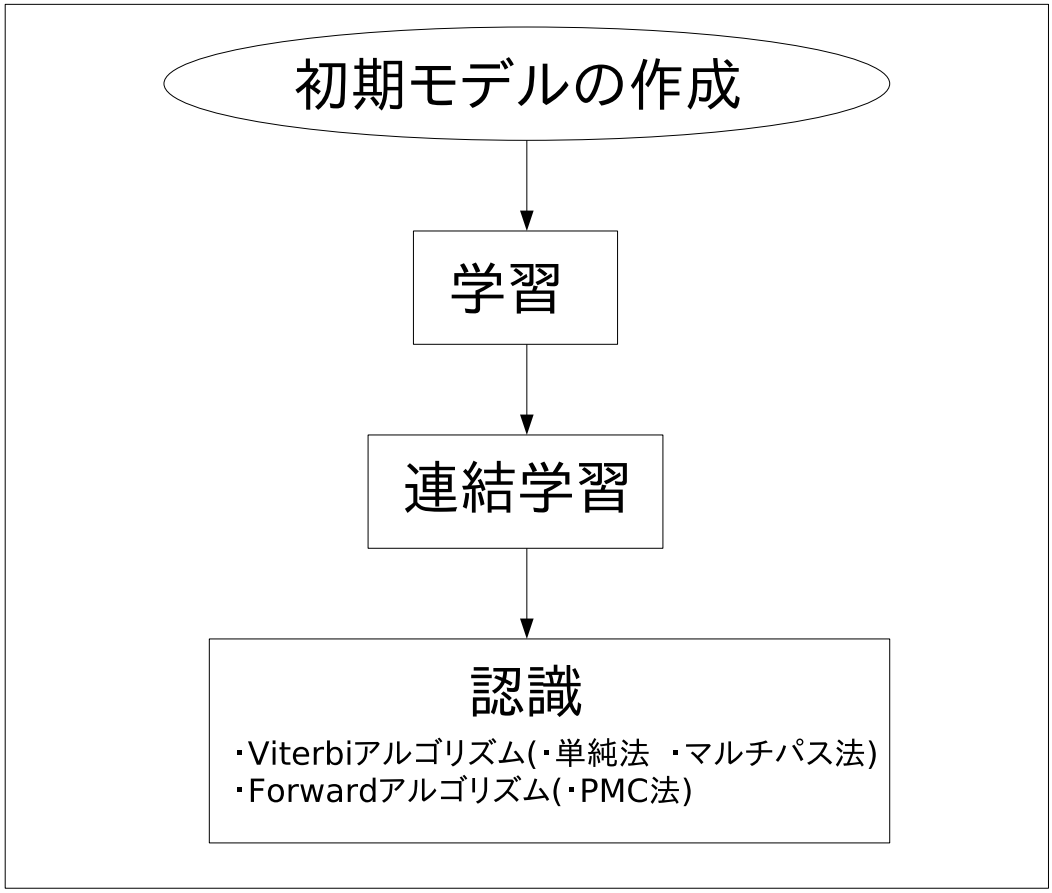

図 10: 音素 HMM の作成と学習及び認識の流れ

6.2 実験条件

本実験では単語音声認識ツールの HTK を使用する．

また，特徴パラメータには MFCC と FBANK と MELSPEC を使用する．単純法とマルチパス法のその他の実験環境は表 2 にまとめる．また，PMC 法のその他の実験環境を表 3 にまとめる．

MFCC を用いた実験では MFCC，対数パワーを，FBANK を用いた実験では FBANK，対数パワーを,MELSPEC を用いた実験では MELSPEC，対数パワーをそれぞれ別の多次元ガウス分布で表現する．

実験条件は MFCC と FBANK と MELSPEC で同一になるように混合分布数を決定している．なお，特徴パラメータの次数は同一にするのが困難であるので同じではない．

実験でのパラメータの再推定において，データ不足により作成できない音素 HMM が存在する場合，混合分布数が MFCC 4，対数パワー 2 で作成できない音素 HMM は MFCC 2，対数パワー 1 にする．混合分布数が MFCC 2，対数パワー 1 で作成できない音素 HMM は MFCC 1，対数パワー 1 にする．混合数を MFCC 1，対数パワーにしても作成できない音素 HMM は実験には用いない．FBANK と MELSPEC も同様にして作成できない音素の混合分布数を減らしていく．

表 4: 実験条件 (単純法, マルチパス法)

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態 連続分布型
stream 数	2

MFCC

特徴パラメータ	MFCC(12 次) +対数パワー (計 13 次)
連続型 HMM の初期モデルの混合分布数	(母音・撥音・無音)MFCC 4 ，対数パワー 1 (その他の音素)MFCC 2，対数パワー 1
音素 HMM の共分散行列	Diagonal-covariance Full-covariance

FBANK

特徴パラメータ	FBANK(24 次) +対数パワー (計 25 次)
連続型 HMM の初期モデルの混合分布数	(母音・撥音・無音)FBANK 4 ，対数パワー 1 (その他の音素)FBANK 2，対数パワー 1
音素 HMM の共分散行列	Diagonal-covariance Full-covariance

MELSPEC

特徴パラメータ	MELSPEC(24 次) +対数パワー (計 25 次)
連続型 HMM の初期モデルの混合分布数	(母音・撥音・無音)MELSPEC 4 ，対数パワー 1 (その他の音素)MELSPEC 2，対数パワー 1
音素 HMM の共分散行列	Diagonal-covariance

表 5: 実験条件 (PMC 法)

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 連続分布型
stream 数	2

MFCC

特徴パラメータ	MFCC(12 次) +対数パワー (計 13 次)
連続型 HMM の 初期モデルの 混合分布数	MFCC 1，対数パワー 1
音素 HMM の 共分散行列	Diagonal-covariance

FBANK

特徴パラメータ	FBANK(24 次) +対数パワー (計 25 次)
連続型 HMM の 初期モデルの 混合分布数	FBANK 4，対数パワー 1
音素 HMM の 共分散行列	Diagonal-covariance

MELSPEC

特徴パラメータ	MELSPEC(24 次) +対数パワー (計 25 次)
連続型 HMM の 初期モデルの 混合分布数	MELSPEC 4，対数パワー 1
音素 HMM の 共分散行列	Diagonal-covariance

7　実験結果

単純法の結果を表 6 に，マルチパス法の結果を表 7 に，PMC 法の結果を表 8 に示す．

表中の括弧内の分母は評価データの総数を，分子は認識できたクロストーク音声数を示す．また，話者の行には使用したクロストーク音声の話者の組合せを示す．

表 6: 実験結果 (単純法)

話者	mau+ftk	mau+fyn	mms+ftk	mms+fyn	平均
MFCC Diagonal	47% (47/100)	44% (44/100)	41% (41/100)	48% (48/100)	45% (180/400)
MFCC Full	57% (57/100)	62% (62/100)	54% (54/100)	50% (50/100)	56% (223/400)
FBANK Diagonal	27% (27/100)	40% (40/100)	29% (29/100)	38% (38/100)	34% (134/400)
FBANK Full	44% (44/100)	48% (48/100)	50% (50/100)	44% (44/100)	47% (186/400)
MELSPEC Diagonal	8% (8/100)	8% (8/100)	8% (8/100)	17% (17/100)	10% (41/400)

表 7: 実験結果 (マルチパス法)

話者	mau+ftk	mau+fyn	mms+ftk	mms+fyn	平均
MFCC Diagonal	46% (46/100)	55% (55/100)	34% (34/100)	44% (44/100)	45% (179/400)
MFCC Full	52% (52/100)	49% (49/100)	38% (38/100)	42% (42/100)	45% (181/400)
FBANK Diagonal	28% (28/100)	51% (51/100)	30% (30/100)	40% (40/100)	37% (149/400)
FBANK Full	41% (41/100)	40% (40/100)	34% (34/100)	35% (35/100)	38% (150/400)
MELSPEC Diagonal	30% (30/100)	45% (45/100)	24% (24/100)	32% (32/100)	33% (131/400)

表 8: 実験結果 (PMC 法)

話者	mau+ftk	mau+fyn	mms+ftk	mms+fyn	平均
MFCC Diagonal	9% (9/100)	16% (16/100)	7% (7/100)	7% (7/100)	10% (39/400)
FBANK Diagonal	1% (1/100)	1% (1/100)	1% (1/100)	2% (2/100)	1% (5/400)
MELSPEC Diagonal	1% (1/100)	1% (1/100)	1% (1/100)	1% (1/100)	1% (4/400)

実験結果より以下のことが得られた. 実験より以下の結果が得られた.

1. 最も認識精度が高かった実験は，単純法 MFCC Full で，平均 56%の精度が得られた.
2. 特徴パラメータにおいて，MFCC，FBANK，MELSPEC の順で認識精度が高い.
3. マルチパス法において，単純法と比較して FBANK Diagonal と MELSPEC Diagonal で認識精度が改善した.
4. PMC 法は認識精度が低い.

7.1 単純法の実験結果

単純法の詳細な実験結果を表 9 に示す.
実験結果より以下のことが得られた.

1. 単純法において，最も認識精度が高かった実験は，MFCC Full で，平均 56%の精度が得られた.
2. 特徴パラメータにおいて，MFCC ＞ FBANK ＞ MELSPEC の順で認識精度が高い.
3. 音素 HMM の共分散行列において，Full は Diagonal より認識精度が高い.
4. 女性話者のほうが男性話者より認識精度が高いものが多い.
5. 特徴パラメータにおいて，MELSPEC の認識精度が低い.

表 9: 単純法の認識結果

MFCC Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	83%	60%	47%
mau+fyn	54%	82%	44%
mms+ftk	61%	72%	41%
mms+fyn	52%	93%	48 %
平均	63%	77%	45%
MFCC Full			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	89%	67%	57%
mau+fyn	69%	91%	62%
mms+ftk	75%	74%	54%
mms+fyn	53%	94%	50%
平均	72%	82%	56%
FBANK Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	76%	39%	27%
mau+fyn	63%	67%	40%
mms+ftk	63%	57%	29%
mms+fyn	49%	86%	38%
平均	63%	63%	34%
FBANK Full			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	73%	58%	44%
mau+fyn	55%	87%	48%
mms+ftk	69%	75%	50%
mms+fyn	48%	94%	44%
平均	61%	79%	47%
MELSPEC Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	61%	14%	8%
mau+fyn	33%	30%	8%
mms+ftk	42%	27%	8%
mms+fyn	36%	61%	17%
平均	43%	33%	10%

7.1.1 単純法において認識成功の例

単純法において同時発話認識に成功した例を表 10，表 11 に示す．

表 10:

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (fyn)
悪質 (akushitsu)	足元 (ashimoto)

認識結果

尤度順位	男性話者 (尤度)	女性話者 (尤度)
(認識結果) 1	悪質 (-7263.5)	足元 (-6044.3)
2	聞こえる (-7333.7)	対策 (-6405.7)
3	優れる (-7359.7)	欲張る (-6501.0)

表 10 の例は，男性話者 mau が「悪質 (akushitsu)」，女性話者 fyn が「足元 (ashimoto)」と発話した場合の認識結果の例である．

この例では，男性話者の認識結果「悪質 (-7263.5)」と女性話者の認識結果「足元 (-6044.3)」ともにクロストーク音声の発話内容と一致しており，正しく同時発話認識できている．

表 11:

クロストーク音声の発話内容	
男性話者 (mms)	女性話者 (ftk)
そのうち (sonouchi)	すまない (sumanai)

認識結果

尤度順位	男性話者 (尤度)	女性話者 (尤度)
(認識結果) 1	そのうち (-4641.5)	すまない (-4737.5)
2	優れる (-5036.8)	滅ぼす (-5009.8)
3	中毒 (-5067.5)	天才 (-5029.0)

表 11 の例は，男性話者 mms が「そのうち (sonouchi)」，女性話者 ftk が「すまない (sumanai)」と発話した場合の認識結果の例である．

この例では，男性話者の認識結果「そのうち (-4641.5)」と女性話者の認識結果「すまない (-4737.5)」ともにクロストーク音声の発話内容と一致しており正しく同時発話認識できている．

7.1.2 単純法において認識失敗の例

単純法において同時発話認識に失敗した例を表 12，表 13 に示す．

表 12:

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

認識結果

尤度順位	男性話者 (尤度)	女性話者 (尤度)
(認識結果) 1	優れる (-6214.0)	可愛い (-5141.0)
2	聞こえる (-6243.4)	手拭い (-5495.9)
3	加える (-6295.8)	すまない (-5521.3)

表 12 の例は，男性話者 mau が「聞こえる (kikoeru)」，女性話者 fyn が「可愛い (kawaii)」と発話した場合の認識結果の例である．

この例では，男性話者の認識結果は「優れる (-6214.0)」であり，女性話者の認識結果「可愛い (-5141.0)」である．女性話者は正しく認識できているが，男性話者が誤認識となっており，同時発話認識に失敗している．

表 13:

クロストーク音声の発話内容	
男性話者 (mms)	女性話者 (ftk)
失恋 (shitsureN)	細々 (komagoma)

認識結果

尤度順位	男性話者 (尤度)	女性話者 (尤度)
(認識結果) 1	中毒 (-6375.1)	細々 (-5041.0)
2	聞こえる (-6530.2)	滅ぼす (-5489.9)
3	優れる (-6576.1)	足元 (-5500.3)

表 13 の例は，男性話者 mms が「失恋 (shitsureN)」，女性話者 ftk が「細々 (komagoma)」と発話した場合の認識結果の例である．

この例では，男性話者の認識結果は「中毒 (-6375.1)」であり，女性話者の認識結果「細々 (-5041.0)」である．女性話者は正しく認識できているが，男性話者が誤認識となっており，同時発話認識に失敗している．

7.2 マルチパス法の実験結果

マルチパス法の詳細な実験結果を表 14 に示す．
実験結果より以下のことが得られた．

1. マルチパス法において，最も認識精度が高かった実験は，MFCC Full で，平均 45%の精度が得られた．
2. MFCC ＞ FBANK ＞ MELSPEC の順で認識精度が高い．
3. 音素 HMM の共分散行列において，Full と Diagonal は同程度の認識精度である．
4. 女性話者のほうが男性話者より認識精度が高いものが多い．
5. 単純法と比較して，特徴パラメータにおいて，BANK Diagonal と MELSPEC Diagonal で認識精度が改善した．

表 14: マルチパス法の認識結果

MFCC Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	79%	63%	46%
mau+fyn	67%	86%	55%
mms+ftk	76%	55%	34%
mms+fyn	89%	50%	44%
平均	78%	64%	45%
MFCC Full			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	82%	68%	52%
mau+fyn	61%	87%	49%
mms+ftk	60%	76%	38%
mms+fyn	49%	91%	42%
平均	63%	81%	45%
FBANK Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	69%	37%	28%
mau+fyn	68%	80%	51%
mms+ftk	62%	57%	30%
mms+fyn	49%	88%	40%
平均	62%	66%	37%
FBANK Full			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	70%	59%	41%
mau+fyn	47%	89%	40%
mms+ftk	54%	77%	34%
mms+fyn	42%	89%	35%
平均	53%	79%	38%
MELSPEC Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	68%	43%	30%
mau+fyn	60%	74%	45%
mms+ftk	53%	58%	24%
mms+fyn	42%	83%	32%
平均	56%	65%	33%

7.2.1 マルチパス法において認識成功の例

マルチパス法において同時発話認識に成功している例を表 15，表 16 に示す．

表 15:

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (fyn)
内容 (naiyou)	勤勉 (kiNbeN)

マルチパス法のモデルの尤度による順位

尤度順位	男性話者	女性話者	モデルの尤度
(認識結果) 1	内容	勤勉	-5141.3
2	内容	細々	-5256.9
3	内容	可愛い	-5256.9

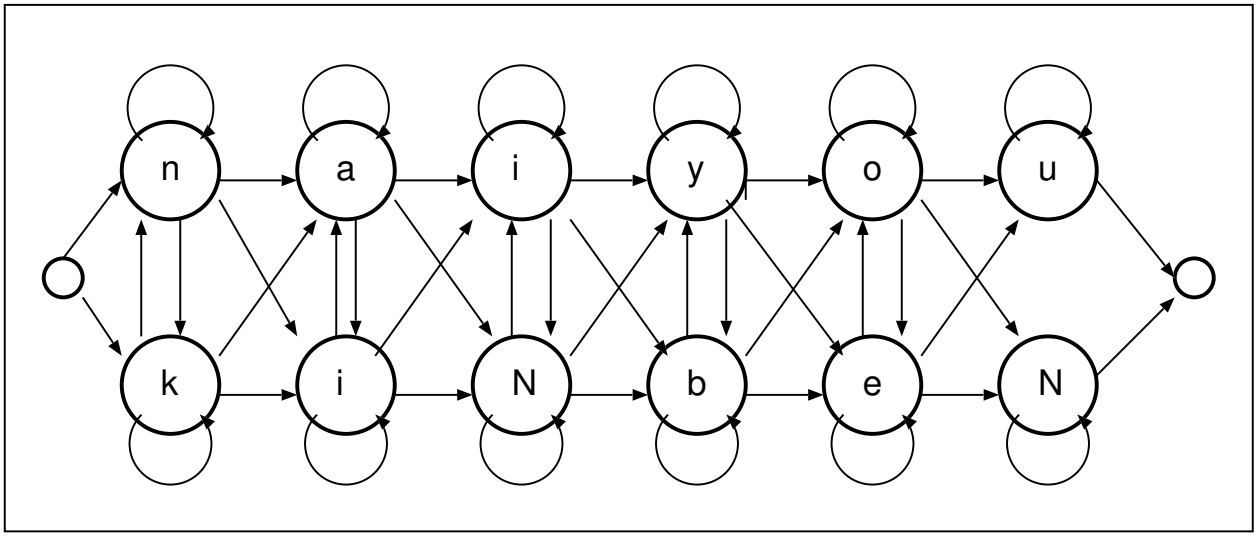

図 11: 認識結果のマルチパス法のモデル

表 15 と図 11 の例は，男性話者 mau が「内容 (naiyou)」，女性話者 fyn が「勤勉 (kiNbeN)」と発話した場合の認識結果の例である．

この例では，クロストーク音声に対して，マルチパス法のモデルの尤度が最大となったモデルが，男性話者が「内容 (naiyou)」，女性話者が「勤勉 (kiNbeN)」のモデルである，このモデルはクロストーク音声の発話内容と一致しており，正しく同時発話認識できている．

表 16:

クロストーク音声の発話内容	
男性話者 (mms)	女性話者 (ftk)
わざわざ (wazawaza)	欲張る (yokubaru)

マルチパス法のモデルの尤度による順位

尤度順位	男性話者	女性話者	モデルの尤度
(認識結果) 1	わざわざ	欲張る	-4501.4
2	加える	欲張る	-4555.1
3	内容	欲張る	-4572.8

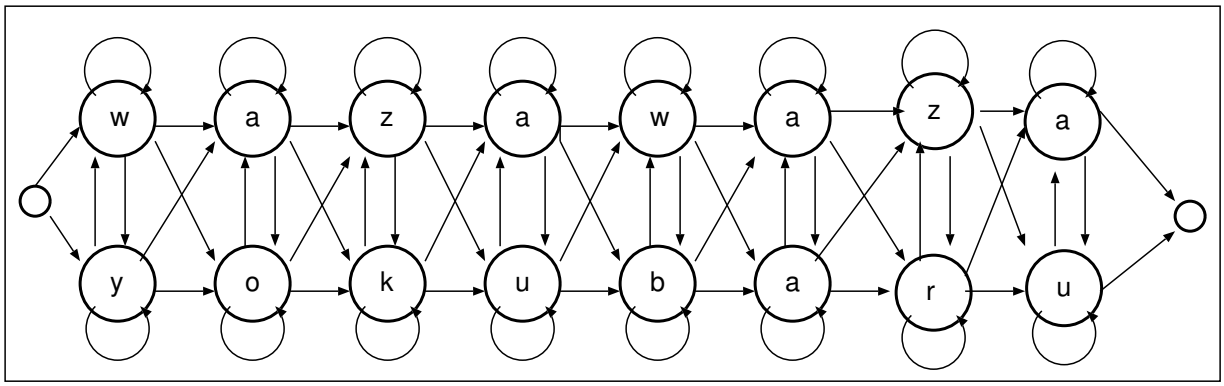

図 12: 認識結果のマルチパス法のモデル

表 16 と図 12 の例は，男性話者 mms が「わざわざ (wazawaza)」，女性話者 ftk が「欲張る (yokubaru)」と発話した場合の認識結果の例である．

この例では，クロストーク音声に対して，マルチパス法のモデルの尤度が最大となったモデルが，男性話者が「わざわざ (wazawaza)」，女性話者が「欲張る (yokubaru)」のモデルである．このモデルはクロストーク音声の発話内容と一致しており，正しく同時発話認識できている．

7.2.2 マルチパス法において認識失敗の例

マルチパス法において同時発話認識に失敗している例を表 17 に示す.

表 17:

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (fyn)
中毒 (chuudoku)	細々 (komagoma)

マルチパス法のモデルの尤度による順位

尤度順位	男性話者	女性話者	モデルの尤度
(認識結果) 1	失恋	細々	-5181.2
2	中毒	細々	-5190.9
3	優れる	細々	-5221.6

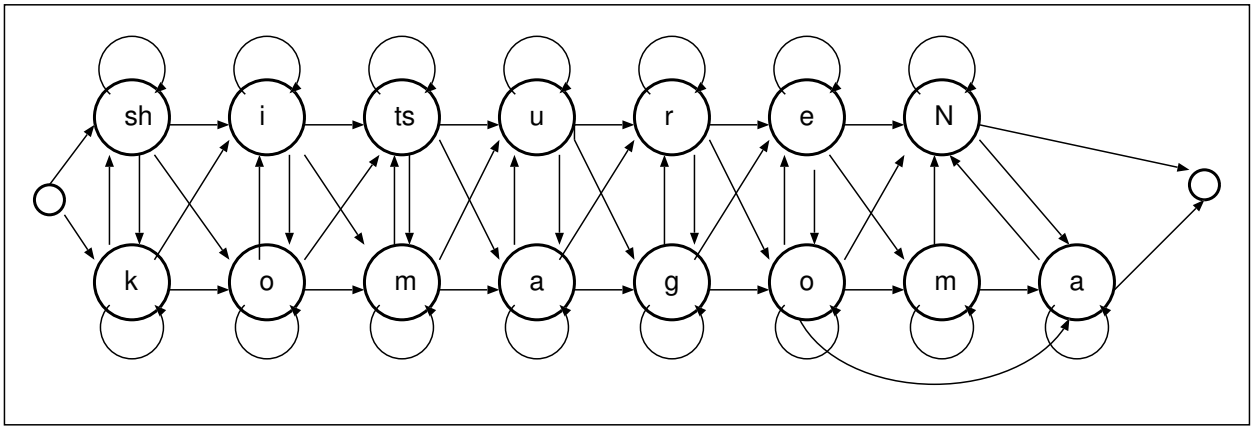

図 13: 認識結果のマルチパス法のモデル

表 17 と図 13 の例は，男性話者 mau が「中毒 (chuudoku)」，女性話者 fyn が「細々 (komagoma)」と発話した場合の認識結果の例である.

この例では，クロストーク音声に対して，マルチパス法のモデルの尤度が最大となったモデルが，男性話者が「失恋 (shitsureN)」，女性話者が「細々 (komagoma)」のモデルである，このモデルはクロストーク音声の発話内容と一致しておらず，同時発話認識に失敗している.

7.3 PMC 法の実験結果

PMC 法の詳細な実験結果を表 18 に示す.

表 18: PMC 法の認識結果

MFCC Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	30%	24%	9%
mau+fyn	26%	48%	16%
mms+ftk	26%	26%	7%
mms+fyn	17%	33%	7%
平均	25%	33%	10%
FBANK Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	10%	10%	1%
mau+fyn	9%	15%	1%
mms+ftk	10%	10%	1%
mms+fyn	10%	20%	2%
平均	10%	14%	1%
MELSPEC Diagonal			
	男性話者認識率	女性話認識率	同時発話認識率
mau+ftk	10%	10%	1%
mau+fyn	11%	10%	1%
mms+ftk	10%	10%	1%
mms+fyn	10%	10%	1%
平均	10%	10%	1%

実験結果より以下のことが得られた.

1. PMC 法において，最も認識精度が高かった実験は，MFCC Diagonal で，平均 10%の精度が得られた.
2. 女性話者のほうが男性話者より認識精度が高いものが多い.
3. MFCC と FBANK 共に認識精度が低い.

7.3.1 PMC 法において認識成功の例

PMC 法において同時発話認識に成功している例を表 19 に示す.

表 19:

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (fyn)
暴力 (bouryoku)	勤勉 (kiNbeN)

PMC 法のモデルの尤度による順位

尤度順位	男性話者	女性話者	モデルの尤度
(認識結果) 1	暴力	勤勉	761.1
2	暴力	細々	528.9
3	暴力	欲張る	470.9

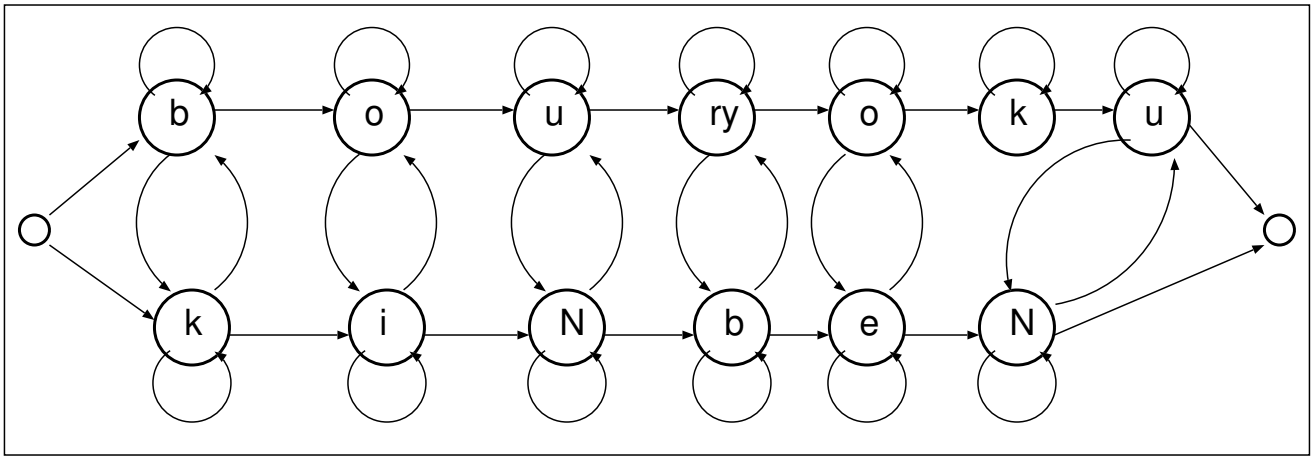

図 14: 認識結果の PMC 法のモデル

表 19 と図 14 の例は，男性話者 mau が「暴力 (bouryoku)」，女性話者 fyn が「勤勉 (kiNbeN)」と発話した場合の認識結果の例である.

この例では，クロストーク音声に対して，PMC 法のモデルの尤度が最大となったモデルが，男性話者が「暴力 (bouryoku)」，女性話者が「勤勉 (kiNbeN)」のモデルである. このモデルはクロストーク音声の発話内容と一致しており，正しく同時発話認識できている.

7.3.2 PMC 法において認識失敗の例

PMC 法において同時発話認識に失敗している例を表 20 ，表 21 に示す.

表 20:

クロストーク音声の発話内容	
男性話者 (mau)	女性話者 (fyn)
加える (kuwaeru)	足元 (ashimoto)

PMC 法のモデルの尤度による順位

尤度順位	男性話者	女性話者	モデルの尤度
(認識結果) 1	中毒	可愛い	783.1
2	中毒	欲張る	771.2
3	暴力	可愛い	770.8

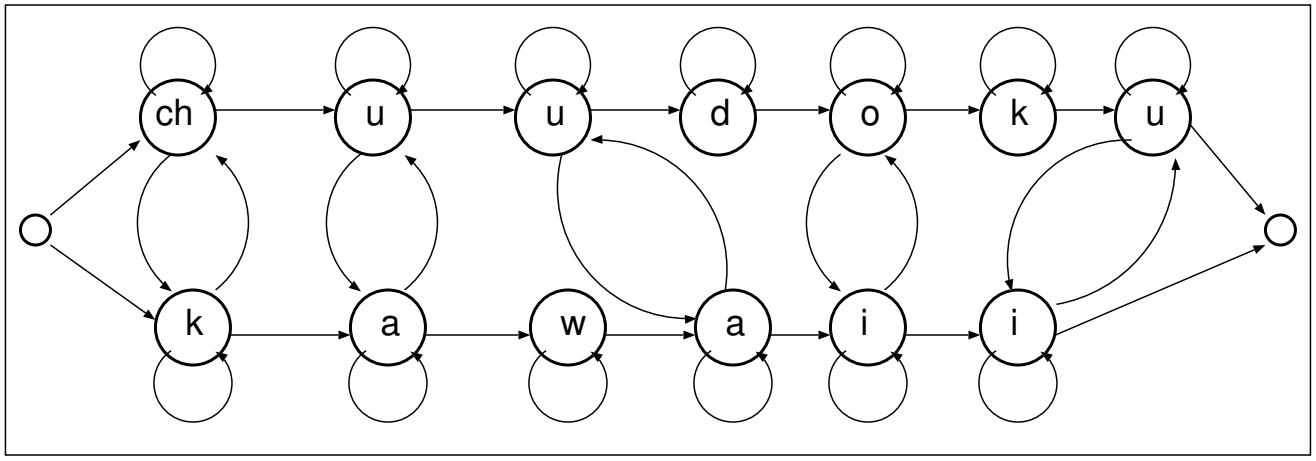

図 15: 認識結果の PMC 法のモデル

表 20 と図 15 の例は，男性話者 mau が「加える (kuwaeru)」，女性話者 fyn が「足元 (ashimoto)」と発話した場合の認識結果の例である.

この例では，クロストーク音声に対して，PMC 法のモデルの尤度が最大となったモデルが，男性話者が「中毒 (chuudoku)」，女性話者が「可愛い (kawaii)」のモデルである. このモデルはクロストーク音声の発話内容と一致しておらず，同時発話認識に失敗している.

表 21:

クロストーク音声の発話内容	
男性話者 (mms)	女性話者 (ftk)
暴力 (bouryoku)	手拭い (tenugui)

PMC 法のモデルの尤度による順位

尤度順位	男性話者	女性話者	モデルの尤度
(認識結果) 1	暴力	可愛い	795.1
2	中毒	手拭い	768.2
3	暴力	手拭い	731.3

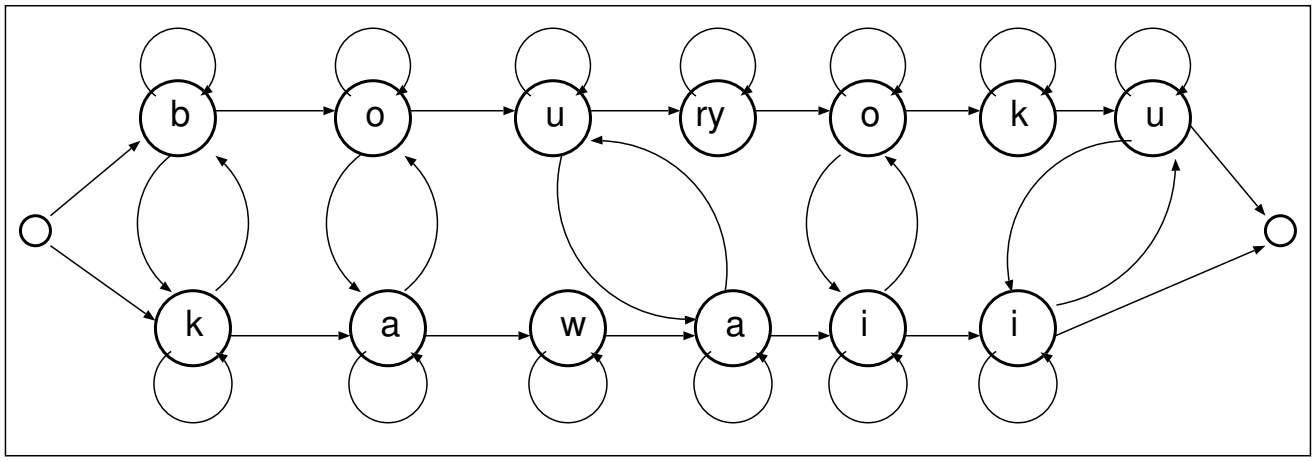

図 16: 認識結果の PMC 法のモデル

表 21 と図 16 の例は，男性話者 mms が「暴力 (bouryoku)」，女性話者 ftk が「手拭い (tenugui)」と発話した場合の認識結果の例である．

この例では，クロストーク音声に対して，PMC 法のモデルの尤度が最大となったモデルが，男性話者が「暴力 (bouryoku)」，女性話者が「可愛い (kawaii)」のモデルである．このモデルはクロストーク音声の発話内容と一致しておらず，同時発話認識に失敗している．

8 考察

8.1 人手による聴覚実験

計算機による認識率と比較するため，実験に使用したクロストーク音声に対して，人間による聴覚実験を行い，同時発話認識率を求めた．なお，被験者は男性 3 名，女性 1 名である．

表 22 に認識結果を示す．

表 22: 聴取実験の認識率

	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	90%	86%	77%
mau+fyn	92%	95%	87%
mms+ftk	85%	85%	74%
mms+fyn	73%	96%	70%
平均	85%	91%	77%

結果より以下のことが得られた．

1. 同時発話認識率の平均が 77%であった．
2. 計算機による実験結果と人間による聴覚実験結果と比較すると，計算機による認識率が低い．
3. 人間による認識での誤り率の改善は男性話者認識, 女性話者認識率, 同時認識率も共に約 50%程度である (単純法 MFCC Full)．

また，比較として実際に実験データに使ったっ単語を提示した場合の聴覚実験も行った．なお，被験者は男性 4 名である．

表 23 に認識結果を示す．

表 23: 聴取実験の認識率

	男性話者認識率	女性話者認識率	同時発話認識率
mau+ftk	98%	100%	98%
mau+fyn	95%	100%	95%
mms+ftk	92%	96%	90%
mms+fyn	94%	96%	92%
平均	95%	98%	94%

8.2 誤認識に対する考察

表 24，及び表 25 に人間による聴覚実験と計算機において誤認識が起こった認識結果の例を示す．

表 24: 例：人間で認識， 計算機で誤認識

正解 (男性/女性)	加える (kuwaeru)	手拭い (tenugui)
計算機の結果	加える (○)	天才 (teNsai)
聴取実験結果	加える (○)	手拭い (○)
正解 (男性/女性)	わざわざ (wazawaza)	すまない (sumanai)
計算機の結果	中毒 (chuudoku)	すまない (○)
聴取実験結果	わざわざ (○)	すまない (○)

人間で認識できたが， 計算機で認識できなかったものには，音声単語データ内に撥音，及び濁音が含まれる単語や，母音 a を含む単語で誤認識が起こりやすい傾向が見られた．

表 25: 例：計算機で認識， 人手で誤認識

正解 (男性/女性)	内容 (naiyou)	対策 (taiaku)
計算機の結果	内容 (○)	対策 (○)
聴取実験結果	概要 (gaiyou)	対策 (○)
正解 (男性/女性)	暴力 (bouryoku)	可愛い (kawaii)
計算機の結果	暴力 (○)	可愛い (○)
聴取実験結果	効力 (kouryoku)	可愛い (○)

計算機で認識できたが， 人間で認識できた例には，実験に使用した音声単語データそのものの子音の撥音が弱いなどといった実験データの品質が悪いもので誤認識が起こりやすい傾向が見られた．また，男性話者と女性話者の音量の差が大きいために，片方の音声が聞き取りにくく誤認識が起こっているものも見られた．

8.3 PMC 法の精度

通常，音声認識では音素 HMM の混合分布数を大きくすることで，認識精度が向上する．

しかし，本研究における PMC 法は，状態数と混合分布数が増加した場合に対応できていないため，PMC 法を改良し，混合分布数が増加した場合に対応することで認識精度が改善する可能性があると考えている．

また，パスの本数や繋ぎ方も検討することで認識精度が改善する可能性があると考えている．

9 おわりに

本研究では，音声認識の分野で困難な課題である，クロストーク音声認識の認識率の調査を行った．認識手法においては，Parallel Model Combination 法をクロストーク音声認識に適応することを提案し，男性話者と女性話者の 2 話者が別々の音声単語を同時に発声した場合に，単 1 のマイクロフォンを利用する状況を想定した音声認識において単純法と Parallel Model Combination 法とマルチパス法の認識率を調査した．

実験結果より以下のことを確認した．

1. 最も認識精度が高かった実験は，単純法 MFCC Full で，平均 56%の精度が得られた．
2. 特徴パラメータにおいて，MFCC，FBANK，MELSPEC の順で認識精度が高い．
3. マルチパス法において，FBANK Diagonal と MELSPEC Diagonal で認識精度が改善した．
4. PMC 法は認識精度が低い．
5. 計算機における認識率は，人間による聴覚実験と比較すると誤り率で 2 倍程度の認識率である．

今後は，PMC 法を改良し認識率の調査を行う．本研究における PMC 法は，状態数と混合分布数が増加した場合に対応できていないため，PMC 法を改良し，混合分布数が増加した場合に対応することで認識精度が改善する可能性がある．

10　謝辞

最後に，本研究において御指導を賜わりました，鳥取大学知能情報工学科計算機 C 研究室の池原 悟 教授，村上 仁一 助教授，徳久 雅人 助手，ならびに，多くの御意見と御助言をいただきました，鳥取大学知能情報工学科知識 A 研究室の清水 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．

発表文献

[1] 岡本, 村上, 池原.“単 1 のマイクロフォンを利用した同時発話音声認識.”日本音響学会講演論文集, No.1-2-12, pp.23-24, 2006.

[2] 岡本, 村上, 池原.“1 本のマイクロフォンを利用した同時発話音声認識.”日本音響学会講演論文集, No.1-P-4, pp.151-152, 2006.

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] 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.

[3] 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.

[4] 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.

[5] 中野, 村上, 池原.“クロストーク孤立単語音声認識.”鳥取大学大学院工学研究科修士論文，2002.

[6] 岡本, 村上, 池原.“クロストーク音声認識における同時発話認識率の調査.”鳥取大学工学部卒業論文，2004.

[7] 谷口, 村上, 池原.“FBANK を用いた孤立単語音声認識.”日本音響学会講演論文集, No.3-Q-3, pp.157-158, 2003

[8] S. Young, P. Woodland, and G. Evermann.“HTK Book.”Cambridge University Engineering Department, 2002.