TR-IT-0101

不特定話者音声認識の研究
A Study on Speaker-Independent Speech Recognition

小坂 哲夫
Tetsuo KOSAKA

1995年3月

概要

音声認識における不特定話者の認識の問題を解決する方法は2通りあると考えられる。一つは不特定話者音声認識の性能そのものを向上させることであり、一つは話者適応により、適応前は性能の悪い認識システムを、使用していくにつれその話者の個人性を学習し性能を向上させる方法である。本研究では以上の二つの方法により不特定話者に対する音声認識の性能向上を図った。その結果を報告する。

もくじ

第 1 章

序論

1.1 研究の背景と目的

音声認識研究は近年極めて盛んになってきている。コンピュータの普及に伴い、専門家だけでなく一般のユーザーがコンピュータを用いるようになってきた。このためコンピュータの利用をし易くするための、マン・マシーンインタフェースの研究開発が重要になっている。またコンピュータの性能向上により、従来のテキストファイルのみならず、画像・音声といった多種のメディアが利用可能となってきた。このマルチメディアの観点からも音声認識技術への期待は大きい。

また研究の観点では、米国国防総省の組織 DARPA が国内の音声研究機関を動員し精力的に研究を推進したことが近年の音声認識研究の進展の大きな原動力となった。また国内では 1986 年に自動翻訳電話の基礎的な研究を行なう ATR 自動翻訳電話研究所が設立され、多くの企業の研究者を受け入れ研究を推進した。このような背景のもと、技術的には 1980 年代後半ごろから隠れマルコフモデル (HMM)、ニューラルネット (NN)、スペクトルリーディングなどによる研究が盛んになった。この中で現在、HMM が大語彙不特定話者音声認識の本命と目されて研究が進展している。

HMM は音声を確率過程とみなし、話者による揺らぎや発声による揺らぎを確率で表現し吸収する。このため従来行なわれていた動的計画法 (DP) に比べ不特定話者の音声認識に適した方法と言われる。しかしまだ、その不特定話者音声に対する能力は十分ではない。現状では特定話者の認識率にははるかに及

ばない。

この不特定話者の認識の問題を解決する方法は2通りあると考えられる。一つは不特定話者音声認識の性能そのものを向上させることであり、一つは話者適応により、適応前は性能の悪い認識システムを、使用していくにつれその話者の個人性を学習し性能を向上させる方法である。本論文は以上の二つの方法により不特定話者に対する音声認識の性能向上を目指すものである。

1.2 基本的立場

1.2.1 不特定話者モデルからの話者適応

話者適応の問題が話者間のスペクトル写像として捉えられたこと、ベクトル量子化の符合ベクトルなどの代表点の対応により写像する方法がとられたことにより、話者適応技術は大きな進歩を見た[1]。このような特定話者間のVQコードのマッピングによる方法では確かに、元の特定話者のモデルを使用場合に比べ、適応後大幅に認識率が向上する。しかし近年連続分布型HMMにより不特定話者に対する認識率が大幅に向上した結果、適応後の認識率が不特定話者の認識率に及ばない状況が出てきた。そこで現在は不特定話者用の連続分布型HMMからの話者適応に研究の主流が移りつつある。

そこでまず話者適応のもとになる不特定話者モデルの検討が重要と考え、第3章で連続分布型HMMを基本とした不特定話者音素モデルの検討を行なった。特に環境非依存のHMMでは従来あまり検討されていなかった混合数の決定法について論ずる(第3.2節)。また音素環境依存モデルにおいて不特定話者モデルの効率的な生成法を第3.3節と第3.4節で論ずる。

1.2.2 少量サンプル学習問題としての話者適応

話者適応は、(話者B)の少量の学習データしか得られない場合、もし大量の学習データを用いて作られた他話者(話者A)又は不特定話者のモデルがあれば、それをうまく活用する手法とみなすこともできる。話者ごとの学習データ量が十分手に入らない場合が現実には多いので、このような少量サンプル学習問題は重要である[2]。

少量サンプル学習問題を考える良い例としては、MAP 推定 (Maximum A Posteriori Estimation) による話者適応法がある [3]。この方法を用いると適応データが 0 つまり、適応前の性能は初期モデルによる性能と等しく、適応データ量が無限大の場合、特定話者モデルでの認識性能と等しいものが得られる。つまりこの方法ではデータ量に応じて初期モデルから特定話者モデルに向けて性能が除々に向上する。このためこの方法は話者適応のみならず、データが少ない場合のモデルの学習にも利用することができる。

以上のような観点から話者適応を考えると、話者適応の性能は適応後の認識率とそのために要した適応用データの量によって図ることができる。つまりより少ないデータで高い認識率の向上が得られるものが性能の高い適応方法と考えることができる。しかし従来の視点では、適応後の認識率がどれだけ特定話者の認識率に近付いたかが主に議論され、適応用データの量に関する議論は少なかった。

本論文では以上のように話者適応の問題を捉え直し、より少ないデータで認識率の向上を図ることを目指している。

より少ない適応データでの話者適応を言い直すと、データ数に応じた話者適応法ともいうことが出来る。少ないデータでの話者適応といっても限界があり、また適応データ数がおおければやはり認識率の向上は大きい。そこで与えられた適応用のデータ量で最適な話者適応法を考える必要がある。適応データ数と適応時に修正する自由パラメータ数との間には密接な関係がある。パラメータ推定では一般に自由パラメータが多い場合は、推定に必要なデータの量も増加するし、逆にパラメータ数が少ない場合は少ないデータ量で推定が可能となる。

一般に話者適応においても以下に述べるトレード・オフが存在する (第3章参照)。

- 自由パラメータ数が少ない適応手法では、少ないデータで適応が可能である。しかし適応後の認識の向上は少ない。
- 自由パラメータ数の多い適応手法では、適応後の認識率の向上は高い。しかし適応データを多く必要とする。

このことから、データ数が少ない場合はより少ない自由パラメータでの適応法が必要であり、データ数が多い場合は自由パラメータの多い適応法が使用可能

であることがわかる。つまり、適応のための自由パラメータの数を適応サンプルの量に応じて制御する方法が必要となる。本研究ではこのような、自由パラメータの数を制御する話者適応法を目指している。

しかし従来の研究を見ると数十秒〜数分程度の比較的適応データを要する、自由パラメータ数の多い適応法は存在するが、より少ない数秒程度のデータで効果的な話者適応の研究があまりなされていない。そこでまず少ない適応データで効果的な話者適応法の検討を第4章で行なった。

1.2.3 適応データ量に応じた話者適応

全節で述べたように適応データ量に応じて自由パラメータを制御し適応を行なう方法が重要であると考えられる。これを実現する方法として次の2つの方法が考えられる (第5章参照)。

- 複数の話者適応手法の適応サンプル数に応じた自動切替え。
- 適応サンプル数に応じた学習パラメータ数の自動調整。

前者は異なる自由パラメータ数の話者適応をサンプル数により切替える方式である。つまり適応サンプル数が少ないうちは、自由パラメータ数の少ない話者適応法を用い、適応サンプル数が増加したら、自由パラメータ数の多い話者適応法を用いる。この切替えを何らかの基準により自動的に行なう。後者は自由パラメータ数は同じにおくが、適応の初期段階でいくつかのパラメータに対し「結び」の関係にするなど拘束を与え、その拘束を適応サンプル数に応じて除々に緩め実質的に自由パラメータ数を制御する方法である。

第4章の少量データによる話者適応の成果をもとに、前者に基づく方法として、5.2節に複数の話者適応法に基づく動的話者適応を提案する。また後者に基づく方法として「MAP-VFS 話者適応法」を提案し5.3節に述べる。

1.2.4 話者適応を利用した不特定話者音声認識

少ないデータ量で動作する話者適応を応用すれば、不特定話者音声認識の性能向上が期待できる。そこで本研究ではさらに話者適応法を利用した不特定話者音声認識についても検討した。この方法では入力音声をまず認識し認識結果

を出力する。この誤りを含む認識結果を教師信号とし、入力音声を使った話者適応を行なう。これにより適応されたモデルを再認識することにより不特定話者認識を実現する。これを実現するためには1発声という少ないデータで動作する教師なし話者適応法が必要である。そこで第4章で検討した話者適応法を利用した。この不特定話者音声認識について第6章で述べる。

1.2.5 話者適応の発話様式適応への応用

従来音声認識の研究は、書き下した文を読み上げる朗読発話に対する研究が主であったが、自由に対話を行なった音声を認識する自由発話音声認識 (Spontaneous Speech Recognition) の研究が始まっている ([4] など)。朗読発話にたいして自由発話では、言語的な違いだけではなく発声の「なまけ」などが起こり音響的にも異なる。そのため朗読発話でのモデルをそのまま用いて自由発話を認識すると、認識率が低下する。そこで話者適応手法を、朗読発話から自由発話への発話様式適応に応用して、不特定話者発声した自由発話音声の認識の検討を行なった。この結果について第7章に述べる。

1.3 本論文の構成

本論文の構成は以下の通りである。

第1章は序論であり、本研究の背景と目的について述べる。

第2章は準備の章であり、本論文の理解のために必要な基本技術に関する概説を行なう。本論文では音響モデルとして HMM および隠れマルコフ網 (HMnet)[5] を用いる。その二者について概説する。次に評価法として用いる HMM-LR 連続音声認識法 [6] 及び、その音素環境依存モデルへの拡張版である SSS-LR 連続音声認識法 [7] について述べる。さらにモデルのパラメータ推定や話者適応法と関わる、移動ベクトル場平滑化法 (VFS)[8] について述べる。

第3章は話者適応の初期モデルとして用いる不特定話者モデルについて検討を行なった。本研究では話者性による特徴の変動を確率で表現することにより吸収する確率モデルに基礎をおいている。その不特定話者モデルの性能向上及び能率的な作成法について論じる。

第4章では、少量の適応データで動作する話者適応法について検討を行なった。従来の話者適応では適応データとして数十秒〜数分程度の量を必要としたが、ここでは数秒程度で動作する話者適応法の提案を行なった。ここで提案した「話者重み学習」及び「木構造話者クラスタリングによる話者適応」では、このような極めて少量のサンプルで適応が可能である。

第5章では、前章の少量データで動作する話者適応での成果をもとに、少量データでも、データ数が増加した場合でも、データ量に応じて動作する、データ量に応じた話者適応について検討した。ここではこれを実現する方法として「複数の話者適応法に基づく動的話者適応」と「MAP-VFS 話者適応法」を提案する。

第6章では以上による話者適応を利用した、新たな不特定話者認識の枠組を提案する。従来の不特定話者音声認識では特定話者と同じモデル作成のアルゴリズムを用い、モデルパラメータ推定の学習データだけ多数話者を与えるという方法をとっていた。これに対し本提案法では2段階に認識することにより、モデルを話者の観点から絞り込み認識性能の向上を図る方法をとった。これについて説明する。

第7章では、本論文で提案した話者適応法を発話様式適応に応用し自由発話音声 (Spontaneous Speech Data) の認識の検討を行なった。

第8章は結論であり、本研究の成果を要約した。

第 2 章

準備

2.1 まえがき

この章では、本論文の理解のために必要な基本技術に関する概説を行なう。本論文の音響モデリングと密接な関わり合いのある隠れマルコフモデル (HMM) 及び隠れマルコフ網 (HMnet) についてまず述べる。次に評価法と関わる HMM-LR 連続音声認識法及び、その音素環境依存モデルへの拡張版である SSS-LR 連続音声認識法について述べる。さらにモデルのパラメータ推定や話者適応法と関わる、移動ベクトル場平滑化法 (VFS) について述べる。

本論文では音声認識の音響モデルに関する検討を対象としている。音響モデルとしては様々な種類が考えられるが、本論文では現在音声認識の分野で最も有望視されている隠れマルコフモデル (Hidden Markov Model: HMM) に基礎をおく方式を採用した。そこで本章ではまず HMM について概説する。

音響モデルを考える場合、どのような単位でモデリングするかが重要である。HMM が用いられてきた初期の段階では、単語によるモデリングも行なわれていたが、語彙が増加するごとに単語モデルを登録する必要があり、一般に任意の音声の認識には不向きである。そこで任意の音声を認識する場合は、音素や音節のような、言語的な単位でかつ単語より細かい単位、いわゆるサブワード単位を採用するのが一般的である。しかし音素・音節などはあくまで言語的な単位であり、音響的にコンパクトな単位とは直接は対応しない。特に音素の音響的パタンは常に一定ではなく、前後の音素の影響でパタンが変化する調音結合の現象が起きたり、また発話速度などの変化に伴い、早く発声した場合に発

音の「なまけ」の現象が起こり得る。つまり同じ音素カテゴリに分類されるものでも、様々なパタンを持っており、これを異音と呼ぶ。

しかし従来は異音を考慮することなく、音素など直接言語的な単位が音声認識に使われてきた。これは異音を認識の最小単位にした場合、(1) モデル数が増加しマッチングなどに計算量がかかる、(2) 統計的に安定したモデルを設計する場合、学習サンプルの量が多く必要になるが、モデル数が増加することによりデータベースが十分大きくないとデータが必要量揃わない、などの問題があるためである。しかし近年データベースの増大とコンピュータの性能向上により、異音モデルの使用が可能になってきている。

本論文では一部の研究を除き以上のような理由で異音を認識の最小単位として利用している。この異音パタンのカテゴリの設定法には様々な方法が提案されているが、ここではその中の一つである逐次状態分割法 (SSS) を利用している [5]。SSS アルゴリズムは本研究の一部と密接に関連しているので、本章で概説した。

移動ベクトル場平滑化法 (VFS) は話者適応法の一種である。この方法は服部らにより VQ ベースの手法が提案され [9]、その後大倉らにより混合連続分布へ拡張された [8]。さらに鷹見らにより HMnet へ応用されている [10]。この方法では平滑化と補間という 2 つの手法を用いることにより、比較的少ないサンプルで適応が可能である。また本手法は話者適応のみならず、サンプル数が少ない場合のモデルのパラメータ再推定法としても利用できる。本論文では VFS をパラメータ再推定や話者適応に利用しているので、本章で概説した。

2.2 隠れマルコフモデル (HMM)

2.2.1 まえがき

近年のコンピュータの急速な能力向上により、音声認識は大量の音声サンプルから学習により自動にモデルを作成するという、統計的な手法が主流となっている。特に隠れマルコフモデル (Hidden Markov Model: HMM) がよく使われるようになってきている。HMM は確率理論に基づく学習アルゴリズムによりパラメータの自動推定ができ認識性能も高い。このため特に不特定話者音声

認識に用いられるようになっている。本研究もこのHMMに基礎をおいた音響モデルを使用している。本節ではこのHMMについて概説する。

2.2.2 HMMによる音声のモデリング

音声は非定常信号であり、音声スペクトルが時刻と共に変化することにより、情報の伝達をおこなう。しかし短時間に区切った見方をすれば定常信号であり、音声は性質の異なる定常信号の連続したものとも考えることができる。HMMはこのうような見地から音声をモデル化したものである。定常信号源の遷移は現在の状態に依存して次の状態にどれだけの確率で遷移するかが決まる、マルコフ状態遷移によって制御される。

HMMはその出力確率の表現方法の違いにより、離散分布HMM(Discrete HMM: DHMM)、混合連続分布HMM(Contenuous Mixture Density HMM: CMHMM)、半連続分布HMM (Semi-Contenuous Mixture Density HMM: SCHMM) などがある。本研究ではこのうち混合連続分布HMMを採用した。そこでCMHMMを中心に以下に説明する。

離散分布HMMでは、入力音声をベクトル量子化して離散的な符合列に変換したのち、出力分布をヒストグラムの形で表現したものである。離散型では量子化誤差が起こるので、この問題を避けるために連続分布型のHMMが開発された。

CMHMMは、パラメータ空間における特徴パラメータの分布形状を確率密度分布関数を用いて表現し、これを出力確率分布として用いるものである。量子化誤差がないため、一般にパラメータ推定のためのデータ量が十分であれば、性能は離散型に比べ高い。分布関数としては、多次元ガウス分布が用いられることが多い。とくに各状態の出力分布が単一ガウス分布で表せないような複雑な形状をしている場合は、複数の多次元ガウス分布の混合により出力確率分布を表現する混合ガウス分布モデルが有用である。また個々の多次元ガウス分布の表現方法として、パラメータ間の相関を考慮し全共分散成分を使用するものと、パラメータ間の相関は無視し、共分散行列の対角成分のみを用いるものがある。

さらに離散型と連続型の中間の性質を持つものとして半連続分布型HMMが

ある [11]。これは離散型において出力分布をヒストグラムで表していたものを、ガウス分布のような連続分布に置き換えたものである。これは見方を変えれば、混合連続分布の全ての状態において混合分布を共有しているモデルとも考えられる。よって Tied-Mixture モデルと呼ばれることもある。

HMM のパラメータは Baum-Welch アルゴリズムにより推定できる [12]。これは別名 forward-backward アルゴリズムとも呼ばれ、名前のとおり、前向きと後向きの確率計算を繰り返して、尤度が向上するようにパラメータを更新するものである。

2.2.3 HMM による音声認識

HMM による音声認識では、モデルに対して入力音声が生ずる確率を計算することによって行なう。最大の確率を与えるモデルに与えられたラベルが認識結果となる。この確率計算のアルゴリズムは、前向き計算またはトレリス計算と呼ばれる。このアルゴリズムでは起こり得るすべての状態遷移の時間経路について、確率の総和を求めるが、確率が最も高い経路で代用する方法も存在する。これはビタービアルゴリズムと呼ばれるもので、基本的には動的計画法 (DP) による最適経路探索と等しい。

2.3 逐次状態分割法 (SSS) による HMnet の生成

一般に、連続発声された音声においては、前後音素のコンテキストのみならず、発話速度やピッチなども含めた広い意味での「音素環境」によって音素パターンの変形、変動が著しく生じる。このような音声パターンの変動は、音声認識が困難な大きな理由である。

そこで、このような音素パターンの変動の情報を積極的に音声認識に利用して、認識単位として音素環境依存の音素モデルを用いることが、認識性能の高い音声認識のためには有効である。既に、いくつかの連続音声認識システム [13, 14, 15, 16] では、音素環境情報を活用して認識性能を向上させており、その有効性が広く認められている。

この音素環境依存モデルをトップダウン的な知識ではなく、モデルを作成するためのデータの特徴により自動的に生成する方法として SSS アルゴリズム

が提案された [5]。本研究でもこの SSS アルゴリズムによって作成された隠れマルコフ網 (HMnet) を用いている。以下に SSS アルゴリズムについて概説する。

SSS では音素の特徴空間上に割り当てられた確率的定常信号源（状態）の間の確率的な推移を表現した確率モデルに対して、尤度最大化の基準に基づいて個々の状態をコンテキスト方向または時間方向へ分割するといった操作を繰り返す。この操作により逐次的にモデルの精密化が行なわれる。これによりモデルの単位の決定とそのモデルの構造決定、および各状態のパラメータ推定を、共通の評価基準の下で同時に実現することができる。

SSS における処理の流れを図 2.1 に示し、この図に従って SSS の原理を説明する。

まず初期モデルとして、ただ 1 つの状態と、その状態を始端から終端まで結ぶ 1 本のパスから成るモデルをすべての音声サンプルから形成し、この状態を分割することから始める。ある時点における状態の分割は、パスの分割を伴うコンテキスト方向、あるいはパスの分割を伴わない時間方向のうちのいずれか一方に関して行なわれる。特にコンテキスト方向への分割時には、パスの分割にともなってそれぞれのパスに割り当てられるコンテキストクラスも同時に分割される。実際の分割方法としては、コンテキストクラスの分割方法も含めてその時点で可能な全ての分割方法の中から、音声サンプルに適用した場合の尤度の総和が最も大きくなるものを採用する。

このような状態分割を繰り返すことによって、少ない状態数で高い尤度を達成することのできる効率のよいモデルが生成される。以上の方法で得られるモデルをここでは HMnet と呼ぶ。

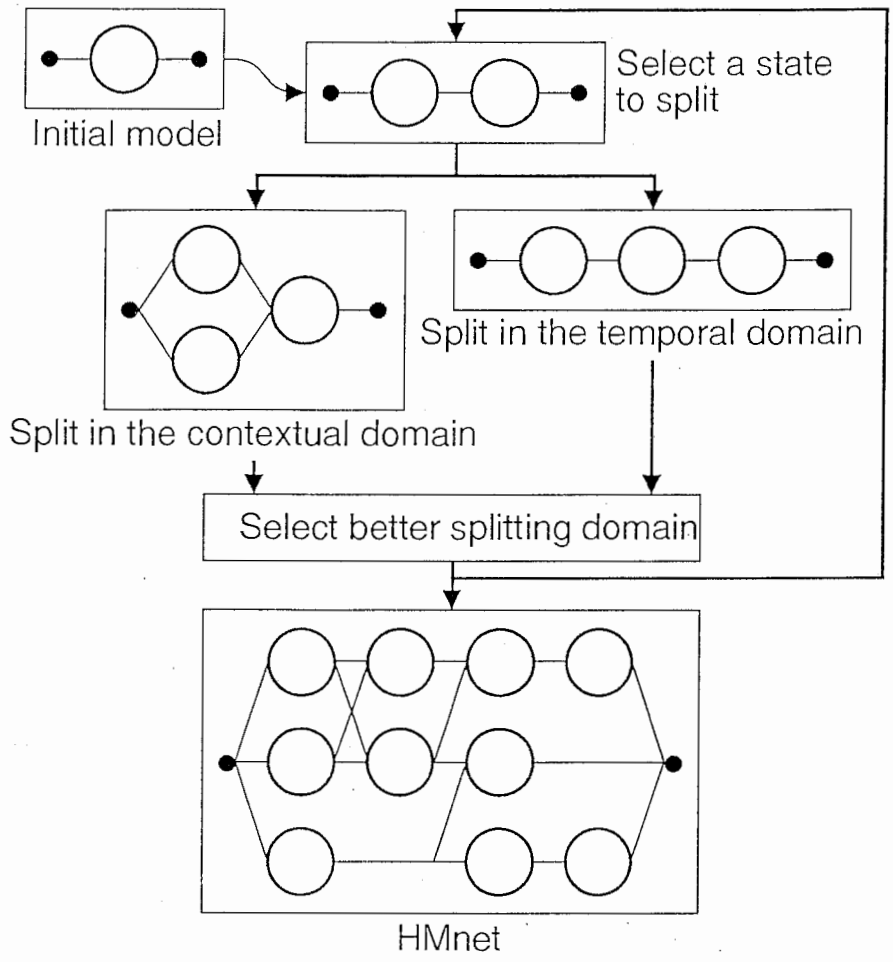

Figure 2.1: SSS の説明図

2.4 HMM-LR

音素ベースの隠れマルコフモデル (HMM) と LR 構文解析アルゴリズムを統合化した音声認識法として HMM-LR がある [6]。HMM-LR 音声認識では、文法から得られる予測的な情報を用いて音声認識の探索空間を縮小し性能の向上を図る。また他の特徴として LR パーザの中から HMM 音素モデルを直接駆動して認識を行なうために、音声認識と言語処理の間に音素ラティス等の中間的なデータを介する必要がなく、高精度でかつ効率的な大語彙連続音声認識シス

テムを構築することができる。

HMM-LR 音声認識の動作概略は以下の通りである。

1. 予測 LR パーザにより、その時点までに既に認識された音素系列から文法上で次に接続しうる (文頭であれば文頭にくることが可能な) 音素を予測する。音素予測の際に、LR 解析表の現在の状態欄に受理動作が指定されていれば、その音素系列は認識候補として残される。

2. 各音素に対して継続時間長の最小値および最大値等の統計情報があらかじめ求められており、これらの値を使って照合する音声区間を決定する。照合する音声区間は現在までに認識された音素系列と予測された音素の最小継続時間長の和を始端、最大継続時間長の和を終端とする区間である。

3. HMM 音素照合部を駆動し、既に認識された音素系列の音素モデルに予測された音素のモデルを連結して音素照合を行ない、照合スコアを求める。ここで照合スコアとは、照合音声区間が音素モデルから生成される尤度であり、forward アルゴリズムあるいは Viterbi アルゴリズムによって計算される。

4. 照合に成功したすべての音素に対して、並行して音素連鎖の枝を伸ばしていく。実際には音素連鎖の枝を伸ばす過程において解析する候補の数が増加してくるので、照合スコアがある一定値以下の場合は枝刈りするというビームサーチを行ない、解析する候補数を削減する。

5. 再び 1 に戻り、認識を続行する。

最終的に、照合がすべて終った段階で照合スコアの高い第 n 候補までを認識結果として出力する。

2.5 SSS-LR

音素コンテキスト依存 LR パーザと、SSS により自動生成された隠れマルコフ網を統合した、SSS-LR 連続音声認識システムについて概説する [7]。

本節ではまず、SSS-LR のシステム構成とその動作の概略について述べ、次に、音素コンテキスト依存 LR パーザのパージングアルゴリズムの説明を行なう。

SSS-LR のシステム構成は、HMM-LR[?] に基づいている。しかし、音素コンテキスト依存の連続音声認識を行なうために、認識及び構文解析のアルゴリズムが HMM-LR とは大きく異なる。

構文解析部 (パーザ) では、予め文脈自由文法 (CFG) から変換された LR テーブルを参照して音素三つ組コンテキストの動的予測を行なうと同時に、その音素コンテキストに合致する継続時間モデルを用いて音素照合区間を設定し、音素照合部に対して照合要求を行なう。構文解析木は、予測された音素コンテキストに従って正確に成長し、単語間の接続部においても前後音素コンテキストに適合するモデルが選択される。

音素照合部では、パーザから要求された音素三つ組コンテキストに適合する HMnet 上の状態パスを 1 つ選択し、異音モデルを形成する。このモデルを用いてパーザで設定された音素照合区間内で尤度計算を行ない、その値をパーザに返す。

次に音素コンテキスト依存 LR パーザについて述べる。先行、中心、後続音素の 3 要因音素コンテキストに依存した構文解析を行なう LR パーザのアルゴリズムとして、以下の 3 方式が考えられる。

1. パーザレベルでの実現 [17]

   音素コンテキストに非依存の LR テーブルを用いて、パージング中に予測音素の先行音素、後続音素を動的に参照し、音素コンテキストを予測する方式。

2. テーブルレベルでの実現 [18]

   LR テーブルを変換して、音素コンテキスト依存の構文解析動作が可能な LR テーブルを生成する方式。

3. 文法レベルでの実現 [19]

   CFG を変換して、音素コンテキスト依存の CFG を生成する方式。LR に限らず CFG 枠組のパーザに適用可能。

このうち、HMnet を駆動するパージングアルゴリズムとして、ここではパーザレベルでの実現方式を採用している。この理由として、HMnet の異音表現効率は高く、非常に多くの異音モデルが表現されるため、テーブルレベルでの実現方式では LR 状態数の増加、及び文法レベルでの実現方式ではルール数の増加が著しく、実現が困難であるという問題が生じるためである。更に、パーザレベルでの実現方式はアルゴリズムが柔軟に変更でき、音素照合部との分離性が高いためにモデルの変更に伴うテーブルや文法の差し替えが不要で実験が容易である。

2.6 移動ベクトル場平滑化 (VFS) 方式の概要

本節ではパラメータ学習に用いられる VFS 方式について概説する [8]。VFS 方式は標準話者の音声パラメータ空間（これを空間Aとする）を連続的な移動伸縮によって、未知話者の音声パラメータ空間（これを空間Bとする）へ変換する手法である。この場合の各モデルの変換は、少数の入力音声サンプルから求められる空間Aと空間Bの間の有限個の対応関係に基づいて空間Aと空間Bの間の変換系の性質を推測し、そのような移動伸縮変換が空間A上の各出力確率密度分布の位置（平均値）および大きさ（分散）に及ぼす影響を実際の分布に反映させることにより行なうことができる。

本手法では、平均値適応・平滑化処理・分散適応の 3 つの処理系により上記の原理を近似的に実現している。しかし特定話者音声認識の実験結果では、分散適応は特に効果が見られなかったため、本実験では分散の適応は行なっていない。以下に平均値適応部及び平滑化処理部について述べる。

まず平均値適応部では、与えられた未知話者の適応用音声サンプルを用いて、空間Aと空間Bの対応関係を求める。それぞれの適応用サンプルに対して、発話内容を表すラベル列に従ってそのサンプルを表す連結モデルを生成し、このモデルを学習することにより、分散および状態遷移確率は固定したまま、各状態 i の平均値ベクトル $\vec{\mu}_i$ のみを $\vec{\mu}_i + \Delta\vec{\mu}_i$ に変更する。

しかし適応用サンプル中に含まれないコンテキストについては出力分布の更新は行なわれない。また更新された場合でも適応用サンプルが少ない場合には、更新された分布について高い精度は期待できない。そこで、平滑化処理部では

空間Aから空間Bへの移動伸縮変換の連続性を前提として、補間と平滑化を行なうことにより、各状態の出力確率分布パラメータの最終的な平均値ベクトルを求める。

以上の手法は話者適応のみならず、少量データによる音素モデルの作成にも有用である。

式 (2.1) による $\Delta\vec{\mu}_i$ の補間と平滑化を行なうことにより、各状態 i の出力確率密度分布パラメータの最終的な平均値ベクトル $\hat{\vec{\mu}}_i$ を求める。

$$\begin{aligned} \hat{\vec{\mu}}_i &= \vec{\mu}_i + \sum_{k=1}^{K} \Delta\vec{\mu}_{c_{(k)}} w_{ic_{(k)}} \\ w_{ij} &= e^{-d_{ij}/\lambda} / \sum_{k=1}^{K} e^{-d_{ic_{(k)}}/\lambda} \\ d_{ij} &= \sum_{l=1}^{L} (\mu_{il} - \mu_{jl})^2 / \sigma_{il}^2 \end{aligned} \tag{2.1}$$

ここで、K は近傍数、L はパラメータ次数、$c_{(k)}$ は状態 i の第 k 近傍の状態番号 (ただし、状態 j の移動ベクトル $\Delta\vec{\mu}_j$ が求められていない場合、j は $c_{(k)}$ の要素から除外する) を表す。また、λ は平滑化の度合いを定める定数であり、この値が大きい程強い平滑化が行なわれる。

第 3 章

確率モデルによる不特定話者音声認識

3.1 まえがき

第1章で述べたように、本論文では (1) 不特定話者音声認識の性能そのものを向上させる、または (2) 話者適応により適応前は性能の悪い認識システムを、使用していくにつれその話者の個人性を学習し性能を向上させる方法により不特定話者に対する音声認識の性能向上を目指している。本章ではこのうち (1) の不特定話者音声認識の性能そのものの向上を目指し、不特定話者音素モデルのモデリング手法について検討した。

本研究は混合分布型の HMM または HMnet に基礎をおく。このような確率モデルではパラメータ推定のアルゴリズムは統計論にベースをおいたアルゴリズムが確立しているが、状態遷移のしかたや状態数などモデルの構造に関する設計の指標の検討は遅れている。そこでまず第 3.2では、特に混合連続型 HMM で問題になる混合数の決定法について検討した。

より精密な混合連続型 HMM を作成す場合、混合数を増加させるという方法も考えられるが、一方で状態数を増加させるという方法もある。状態数を増やすことにより、モデルの種類を増加させることが可能になる。従来は音素によるモデリングが一般的であったが、異音カテゴリのモデルを持つ音素環境依存モデルの研究が近年盛んである。

そこで第 3.3節及び第 3.4節では不特定話者用音素環境依存モデルの作成について述べる。不特定話者の音素環境依存モデルを作成しようとすると、一般に環境に独立な音素モデルに比べ大量の学習データを必要とする。そこでこれら

の節では少ないデータで効率的にモデルを作成する方法を提案する。

第 3.3節では特定話者モデルを混合することにより、不特定話者モデルを作成する方法を提案する。しかし、この方法ではモデルの混合数が話者数と一致するため、任意の混合数のモデルを作成できないという欠点がある。そこで第 3.4節ではさらに話者クラスタリング手法を導入することにより、話者数には依存しない任意の混合数のモデルを作成する方法を提案する。

3.2 混合連続分布 HMM における混合数の自動決定

3.2.1 まえがき

近年混合連続分布 HMM(Continuous Mixture HMMs: CMHMM) の性能向上のために、HMM の構造の検討が行なわれるようになってきた [22][24][5][21]。従来 HMM ではそのパラメータ推定については、確率理論にもとずいた定式化がなされて来たが、構造についてはヒューリスティックに与えられてきた。CMHMM の構造として考えられるのは遷移状態がどのように接続するかというトポロジーの問題、混合数、ステート数などである。

Kamp は統計的な手法により不要な状態を消去することにより最適な状態数を決定する方法を提案している [22]。また何人かの研究者により HMM の最適なトポロジーを決定する方法が提案されている。例えば池田は AIC(Akaike's Information Criterion) [23] により最適なトポロジーを決定する方法を提案している [24]。松尾らは非 left-to-right の HMM トポロジーを学習により求める方法を提案している [21]。また音素環境依存モデルで、どの状態を共有化するかを考慮したトポロジー決定法もいくつか提案されている (例えば [5] など)。

一方 CMHMM の構造のうち混合数の自動決定法を扱ったものは少ない。混合数に関して、山口らは音素モデルごとにサンプル数を考慮して混合数を変えたモデリングを行なっている [25]。しかしヒューリスティックに混合数を定めており再現性がない。また混合数を変えない場合との比較もなく、有効性も確かめられていない。

混合数の自動決定をしたものとしては、Rabiner ら [26] が平均歪みにより混合数を自動決定する方法を試みている。しかし結果は非常に悪くなったと述べ

られているだけで、認識率も示されていない。野村らは [27] クラスタリングを用いた混合数の自動決定について試みている。この方法では単にクラスタリング時の距離の閾値のみによって混合数を決定している。混合数を決定する場合、データの分布の形状のほかデータ量の考慮も必要である。しかしこの方法ではデータ量に対する考慮はしていない。また評価では出力分布が 12 という 1 条件でのみ 0.3% の認識率の向上がある結果があるだけなので、有効性が不明である。以上のように、混合数自動決定に関する有効なアルゴリズムは、ほとんど提案されていないというのが現状である。そこでここでは CMHMM の混合数自動決定に関する検討を行なった。

混合数の自動決定をする場合分布の形状のほか、サンプル数に対する考慮も必要である。一般的に行なわれているように、混合数をすべて均一に与えると、極端に分散の大きさが小さくなる場合がある。これは音素や状態に対する学習データ量にばらつきがあるためである。つまりデータ量が少ないことに起因するパラメータの誤推定が起こっているわけで、このような場合はモデルパラメータ数を減少する必要がある。

ここでは以上のようなサンプル数の影響も考慮した決定法について検討を行った。本節では、混合数を音素モデルごとまたは、状態ごとに決定するアルゴリズムを提案し、その有効性を示した。

3.2.2 混合数の増加における問題点

CMHMM において一律に全ての音素モデルの混合数を増加させるのは、設計に用いるサンプル数のアンバランスや各クラスの正規性の違いを考えると妥当でなく、モデルごとに混合数を変える必要があることが予想される。この問題点について認識率及び HMM の出力値から考察をおこなった。

3.2.3 CMHMM による不特定話者音素認識

混合数の増加に関する問題点を明らかにするためにまず、CMHMM を用いて各音素カテゴリーの混合数を同一にして混合数を増加させる認識実験を行った。音響分析条件を付録 A.3、音声資料等を表 3.1 に示す。本実験は分散行列が対角要素のみの 4 状態 3 ループの CMHMM により、学習とは異なる話者 10

名が発声した 216 単語を対象とした話者オープンの認識実験を行った。HMM は語頭・語中を分けた計 49 音素に対して設計したが、認識実験では語頭・語中の混同を認めた 34 音素を評価対象とした。認識率は文法なしで Viterbi アルゴリズムにより音素系列を求め、これと正解の音素系列との DP によるアライメントをとり、置換、脱落やセグメンテーションを考慮して求めた。認識実験結果を図 3.1に示す。混合数が増加するに従い認識率は向上するが、個々の音素の認識率を検討すると特にサンプル数の少ない音素では認識率が低下する傾向がみられる。このことは混合数をモデルにより変えることの有効性を示唆している。

Table 3.1: 音声資料

音声資料	学習用	男性話者 12 名, 736 単語 736 × 12 = 8,832 データ
	認識実験用	男性話者 10 名, 216 単語 216 × 10 = 2,160 データ
音素 モデル	p1 p2 t1 t2 k1 k2 b1 b2 d1 d2 g1 ng m n N r w y s sh h z ch1 ch2 ts1 ts2 sy hy zy cy py ky by gy1 ngy my ny ry aa ii uu eei oou a i u e o silence (1,2 は語頭、語中を表す)	

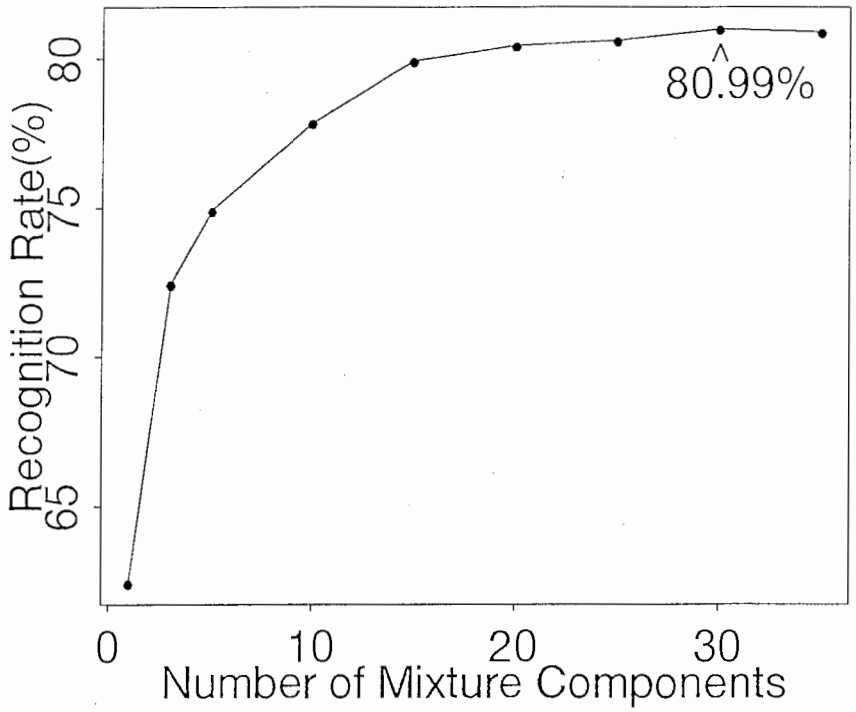

Figure 3.1: 混合数の変化と音素認識率

3.2.4 HMM 出力値による検討

混合数を増加させた場合、限定されたサンプル数でどの程度のモデルの推定精度がえられるかが問題となる。この問題を明らかにするためにまず HMM の出力分布における分散の大きさについて検討した。図 3.2にサンプル数と HMM の共分散行列の大きさの関係をプロットした。各点は HMM の各音素に対応する。分散の大きさは以下のような各分散の行列式の平均で定義した。各音素に対する分散の大きさを f_p、各音素の各状態に対する分散の大きさを f_{pn} とする。

$$f_p = (1/\sum_{n=1}^{N} m_n) \sum_{n=1}^{N} \sum_{m=1}^{m_n} \log |\mathrm{S}_{nm}| \tag{3.1}$$

$$f_{pn} = (1/m_n) \sum_{m=1}^{m_n} \log |\mathrm{S}_{nm}| \tag{3.2}$$

ここで N は状態数、m_n は状態 n における混合数、S_{nm} は状態 n の m 番目の混合分布の共分散行列を表す。図が示すように混合数を上げるに従って全体的に分散は減少するが、特にサンプル数が少ないものでは分散が極端に小さくなりモデルの誤推定がおこなわれている可能性がある。

次に学習データと評価データにおける HMM の出力値を用い誤推定について検討した。図 3.3に学習データ及び評価データにおける HMM の混合数に対する出力値の変化を示した。HMM 出力は対数をとり混合数 1 の値との比を音素毎に表示した。このように学習データにおいては、HMM の出力値は混合数の増加に従って増加するが、評価データでは必ずしも増加するとは限らない。分散の大きさと比較検討すると、分散の極端に小さなものは、学習データにおいては出力値が急上昇するが、評価データにおいては逆に低下し、学習データに対する過学習が起きていることが分かる。

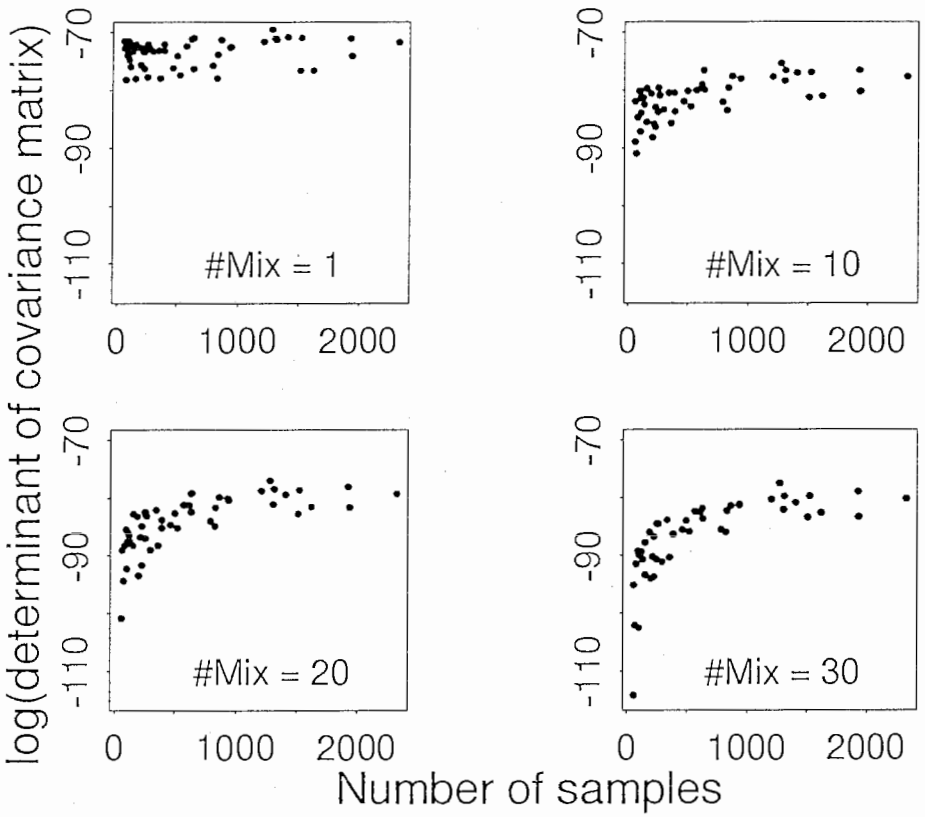

Figure 3.2: 出力分布の共分散行列の行列式の値とサンプル数の関係

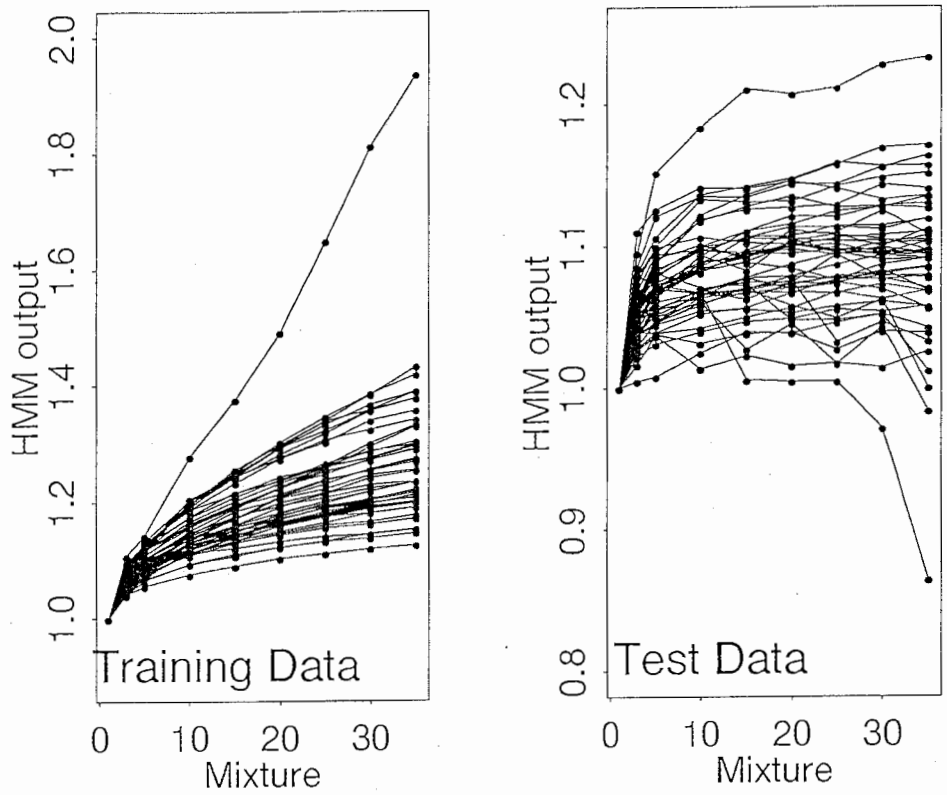

Figure 3.3: HMM 対数出力値（混合数 1 の値で正規化）と混合数の関係

3.2.5 混合数の自動決定アルゴリズム

以上の結果をもとに混合数の自動決定について検討をおこない、不特定話者音素認識実験により評価を行った。

HMM 出力の均一化による方法

HMM による認識において、入力系列に対して正しい音素モデルが照合された場合、HMM の出力値が各音素によらず均一になることが望ましい。そこで HMM の出力値が均一となるよう各音素モデルの混合数を調整する方法を検討した。この場合学習データでは混合数の上昇にともない HMM 出力値も増加するが、それが過学習のためかモデルの推定がよりうまくいっているのか判断が

つかない。そこで話者 10 名の評価データを 2 分割し、5 名で HMM 出力値を求め混合数を決定し、残りの 5 名で評価実験を行なう方法をとる。評価データにおいて HMM の出力値を検討したところ、本方法で混合数を決定すると混合数が極端にばらつき、混合数の決定法としては不適切であることが分かった。

分散の均一化による方法

従来は CHMM の各状態の分布の大きさにかかわらず、一律に等しい混合数を与えていた。本手法では各状態の分布の大きさを各状態に含まれる個々のガウス分布の分散の大きさの平均値で定義し、これを評価基準として各状態の各ガウス分布の分散の値が均一となるよう混合数を決定する。これにより分布の大きな状態では多くの混合数が、分布の小さな状態では少ない混合数が割り当てられる。これは、状態の分布が大きい場合は、より詳細に分布を表現する必要があると考えられるためである。また学習サンプル数が少ないモデルでは分布の大きさが極端に減少するので、評価基準として分散の大きさを用いると自動的に混合数が少なく割り当てられ、モデルの頑健性の向上が期待できる。本実験では個々のガウス分布の分散の大きさを、共分散行列の行列式の値で与えた。

混合数を決定するために、まずいくつかの種類の混合数の音素モデルを作成する。このモデルをもとに全音素の全状態について、混合数と分散の大きさの関係 ($f_p(m_n^p)$,p: 音素番号,n: 状態番号,m: 混合数) を図 3.4 にプロットした。図から混合数を増加するとほぼ単調に分散の大きさが減少することがわかる。いくつかの状態で急激に分散が減少するのが観察されるが、これはサンプル数が少なく過学習が起きているためである。

このグラフから、ある分散の大きさを定めると、そのときの各音素の各状態ごとの混合数が求められるので、これを用いた混合数の自動決定が可能である。表 3.2.5 に全ガウス分布数を与えた場合の混合数自動決定のアルゴリズムを示す。このアルゴリズムでは、音素の各状態別々に混合数を決定するが、音素ごとに混合数を決定することも同様に可能である。実験ではこの 2 つの方法について検討した。

以上のアルゴリズムを用いることにより、分散が単調に減少する場合、混合

数が一意に決定される。またこのアルゴリズムでは、混合数を決定する過程では CMHMM の再学習をする必要がないため高速に混合数が求められる。但し認識時の計算量を考慮して、最大混合数を 35 とした。

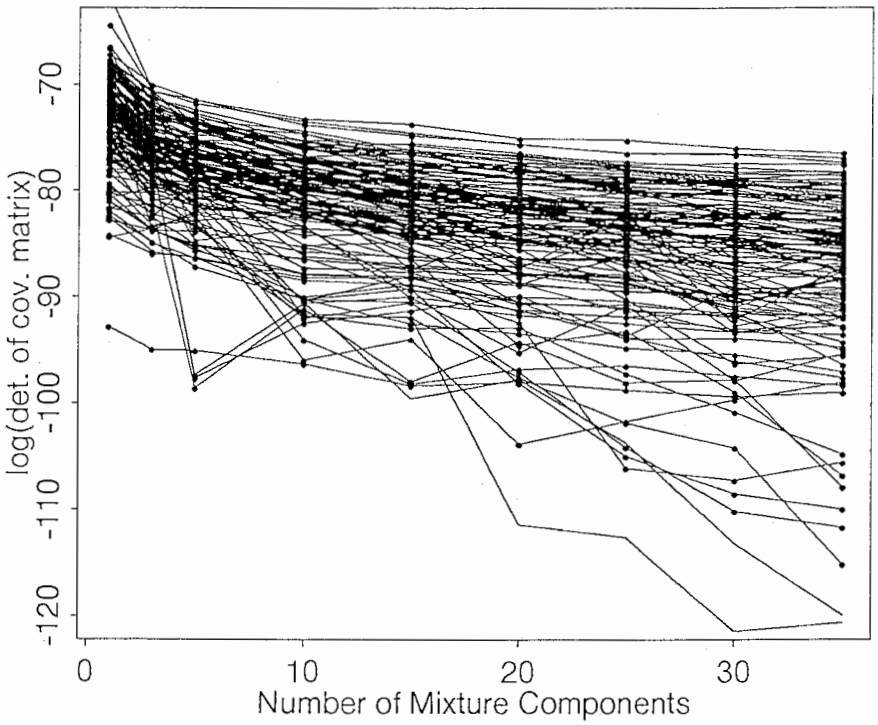

Figure 3.4: HMM 出力分布の分散の大きさと混合数の関係

混合数自動決定アルゴリズム

記号の定義

P: 音素数 (=49)　N: 状態数 (=3)

M: 現在のガウス分布の総数

M_o: 目標とするガウス分布の総数

m_n^p: 音素 p 状態 n の混合数　k: 繰り返し回数

t_k: k 時点での分散の大きさ　α： 更新係数

t_0: t_k の初期値

アルゴリズム

1. 初期値設定

$$t_1 = t_0$$

2. 混合数の計算

$$k = k + 1$$

$$m_n^p = f_p^{-1}(t_k)$$

$$M = \sum_{p=1}^{P}\sum_{n=1}^{N} m_n^p$$

3. t_k の更新

if $M = M_o$ then goto STEP4.

else if $M < M_o$ then $t_{k+1} = t_k - \alpha$

else if $M > M_o$ then $t_{k+1} = t_k + \alpha$

goto STEP2.

4. 求められた混合数で CMHMM の再学習

3.2.6 混合数自動決定の評価実験

不特定話者音素認識実験により、混合数が均一の場合、音素モデルごとに混合数を自動決定した場合、さらに今回提案した手法により、全状態の混合数を自動決定した場合の 3 種類の比較を、平均混合数が 1、3、5、10 の場合について行なった。

音響分析条件、音声資料等を表 3.1に示す。用いた CMHMM は 4 状態 3 ループ、共分散行列が対角要素のみのものである。この CMHMM を用い、学習とは異なる話者 10 名が発声した 216 単語中の音素を対象とした話者オープンの認識実験を行った。CMHMM は語頭・語中を分けた計 49 音素に対して設計したが、認識実験では語頭・語中の同一音素の混同を認めた 34 音素を評価対象とした。認識率は文法なしで Viterbi アルゴリズムにより音素系列を求め、これと正解の音素系列との DP によるアライメントをとり、置換、脱落やセグメンテーションを考慮して求めた。評価式を以下に示す。

$$Cor(\%) = (N - S - D)/N \times 100\%$$

但し、 N: 全評価データ数、 S: 置換数、 D: 脱落数

実験結果を表 3.2に示す。いずれの混合数の場合も、音素モデルごとや状態ごとに混合数を決定したほうが認識率が向上する。また今回提案した状態ごとの混合数決定を音素モデルごとの混合数決定と比較すると、平均混合数 5 や 10 の場合若干認識率が向上する。

図 3.5は各音素モデルごとに混合数を決定した場合の、平均混合数が 10 における各音素モデルごとの HMM の混合数である。合わせてサンプル数も表示した。混合数が多く割り当てられるカテゴリは、母音、鼻音、破裂音の /k/、流音など、不特定話者の場合、話者による変動が大きいと考えられる音素である。逆に /h/ を除く摩擦音・歯擦音では混合数が少なくなる。またサンプルが少ないカテゴリでは混合数も少なくなる傾向が見られるが、サンプル数の多い /s/、 /sh/ でもそれぞれ 7 混合、2 混合と少ない混合数となっており、単にサンプル数によって混合数が決まるわけではないことが分かる。

表 3.3に各状態ごとに混合数を決定した場合の、各音素の各状態の混合数を示す (平均混合数 5)。母音や鼻音で混合数が多くなっており、このような音素ではスペクトルの変動が大きいことがうかがえる。特に母音の場合は、 HMM

の 3 状態のうち前後の状態で混合数が多くなっており、音素環境の影響が大きいことがわかる。無声破裂音では閉鎖区間に当たると考えられる先頭の状態で混合数が少なくなっている。但しこれらの混合数は HMM の学習サンプル数によっても影響受けるので、そのまま音素のスペクトル変動の度合を表しているわけではない。

Table 3.2: 音素認識実験結果 (%)

平均混合数 = 全ガウス分布数 /(音素モデル数×状態数)

平均混合数	混合数均一	音素ごと混合数決定	状態ごと混合数決定
1	62.46		
3	72.46	74.58	73.76
5	74.92	76.93	77.47
1 0	77.84	78.49	79.62

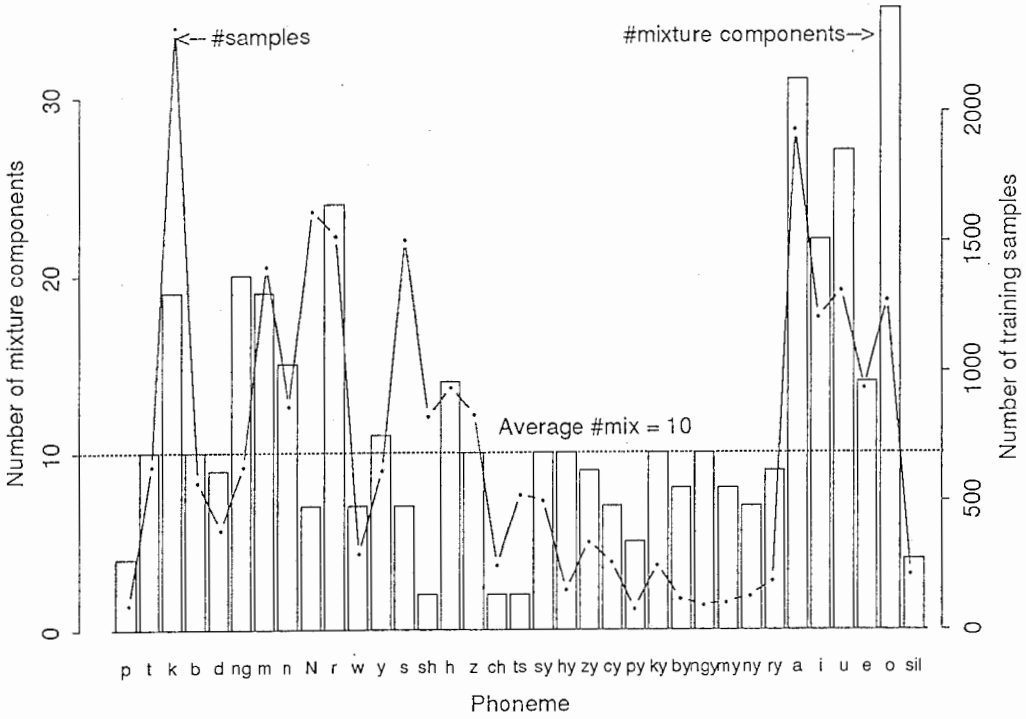

Figure 3.5: 分散の均一化法による音素ごとの混合数及び学習サンプル数

Table 3.3: 平均混合数 5 に自動決定後の各音素・各状態の混合数

音素	各状態の混合数	音素	各状態の混合数	音素	各状態の混合数
p1	2 - 4 - 4	p2	1 - 2 - 7	t1	3 - 5 - 7
t2	1 - 1 - 3	k1	3 - 5 - 11	k2	3 - 3 - 10
b1	1 - 1 - 9	b2	3 - 3 - 13	d1	1 - 1 - 6
d2	4 - 2 - 4	g1	1 - 4 - 8	ng	11 - 7 - 11
m	7 - 5 - 14	n	7 - 6 - 13	N	3 - 1 - 1
r	7 - 8 - 11	w	5 - 3 - 2	y	5 - 5 - 4
s	2 - 1 - 12	sh	2 - 1 - 3	h	5 - 2 - 11
z	5 - 1 - 5	ch1	1 - 1 - 3	ch2	1 - 1 - 2
ts1	1 - 1 - 3	ts2	1 - 1 - 3	sy	1 - 1 - 12
hy	1 - 3 - 9	zy	3 - 2 - 6	cy	2 - 4 - 9
py	1 - 3 - 3	ky	2 - 3 - 7	by	4 - 3 - 3
gy1	2 - 3 - 7	ngy	4 - 5 - 3	my	3 - 4 - 3
ny	4 - 3 - 2	ry	4 - 3 - 5	aa	3 - 1 - 6
ii	2 - 1 - 1	uu	1 - 1 - 5	eei	4 - 1 - 3
oou	13 - 1 - 14	a	10 - 4 - 23	i	9 - 3 - 17
u	14 - 4 - 19	e	9 - 1 - 15	o	23 - 3 - 31
silence	5 - 1 - 9				

3.2.7 まとめ

以上 CMHMM における混合数について検討した。この結果総混合数が限られた条件で、分散の大きさにもとづいて混合数を決定することが有効であることを示した。また音素モデルごとに混合数を決定する方法と比較すると平均混合数 5、10 の場合若干の認識率の向上が認められた。

3.3 話者混合逐次状態分割法による不特定話者音素環境依存モデルの作成

3.3.1 まえがき

これまでに音素コンテキスト依存モデルを用いたいくつかの不特定話者音声認識方式が提案 [13, 14, 15, 16] されている。音素コンテキスト依存モデルを用いることにより、コンテキストに独立な音素モデルを用いる方法に比べ、調音結合による音声の変動を精密に表現することができ、より高性能なシステムを構築できるという利点がある。

しかしながら不特定話者の音素コンテキスト依存モデルを作成しようとすると、一般にコンテキストに独立な音素モデルに比べ大量の学習データを必要とする。そこで本研究では比較的少ないデータで不特定話者を対象とした音素コンテキスト依存モデルを作成する手法として話者混合法 (Speaker-Mixture Method) を提案する。

話者混合法を音素コンテキスト依存モデルに適用するに当たって、コンテキスト依存モデルの作成に、我々がこれまで提案してきた逐次状態分割法 (SSS: Successive State Splitting) を利用する [5]。SSS はコンテキスト依存モデルを作成する場合問題となる音素コンテキストクラスを、知識により人為的に決めるのではなく、モデルの単位や構造を自動的に決定するアルゴリズムである。これにより HMnet(Hidden Markov Network) と呼ばれる詳細かつ頑健なモデルが得られる。

以上話者混合法と SSS を組み合わせた方法をここでは話者混合 SSS 法 (Speaker-Mixture SSS) と呼び、その有効性を不特定話者音声認識実験で検討した。

3.3.2 本方式の概要

本節ではここで提案する「話者混合法」のアルゴリズムの流れを図 3.6に示し、各節に先がけて概説する。本方式の処理の流れを以下に示す。

1. 初期 HMnet の作成

   出力分布が単一ガウス分布であるような初期 HMnet を、話者 1 名の大量のデータから SSS アルゴリズムを用いて生成する。

2. パラメータ学習

次に初期 HMnet に対し、話者適応法を利用して複数の話者の比較的少量のデータによりそれぞれ適応をおこない、複数話者分の HMnet を作成する。

3. 話者混合化

この複数話者分の HMnet の対応する状態を 1 つの混合出力分布として表現することにより、混合連続出力分布 HMnet を作成する。このモデルにより不特定話者音声認識を行なう。

以下の節でこれらの各ステップについて説明する。

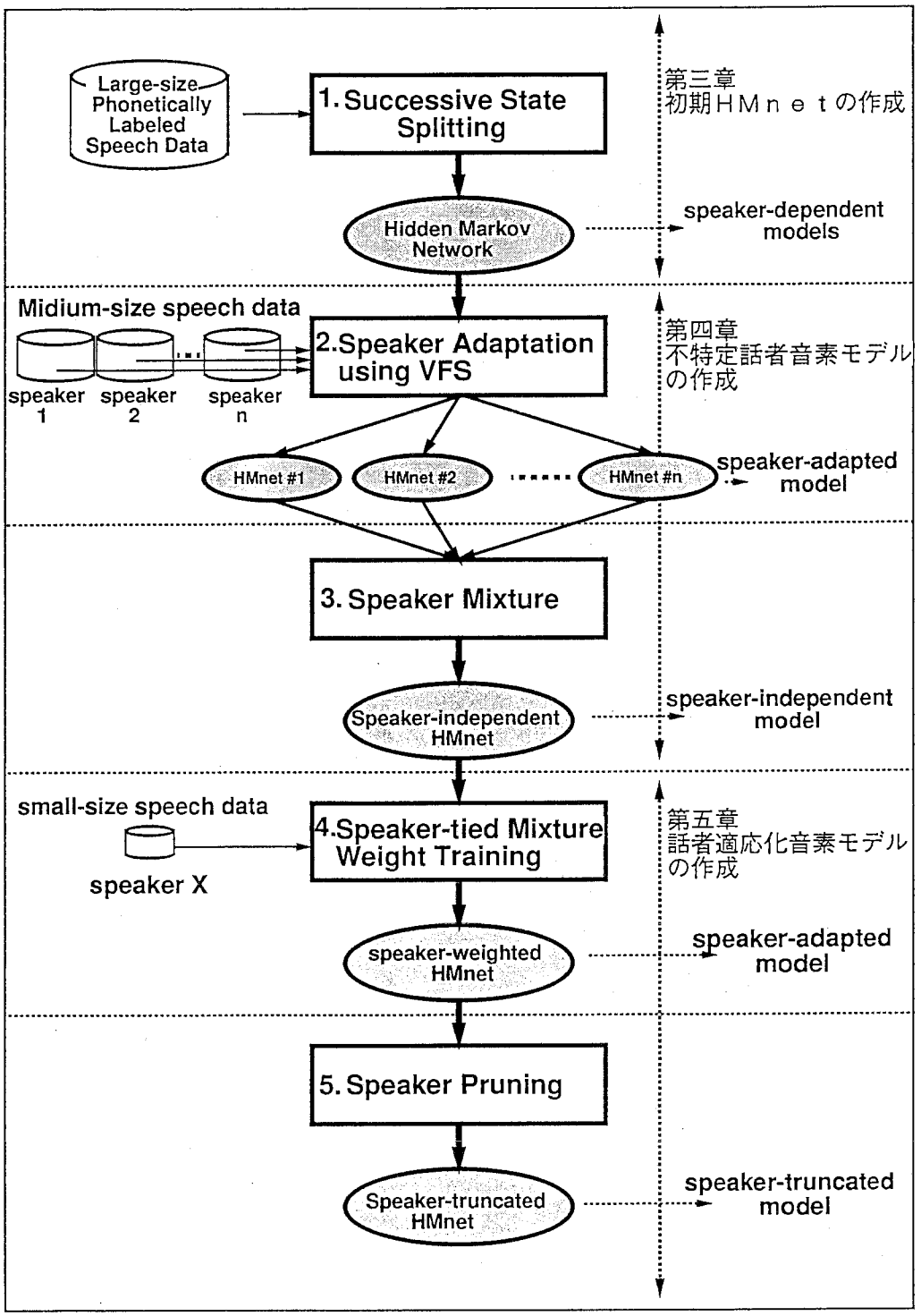

Figure 3.6: 話者混合 SSS 法, 話者重み学習法, 話者プルーニング法による音素モデル作成の流れ図

3.3.3 逐次状態分割法 (SSS) による初期 HMnet の作成

本研究では音素コンテキストによる音声の変動は話者間で共通なものと考え、1名の話者のさまざまな音素コンテキストを含むデータからSSSアルゴリズムを用いて初期HMnetを作成する。

3.3.4 話者混合SSSによる不特定話者音声認識

前章で概説したSSSでも多数話者のデータを用いることにより混合出力分布HMnet作成は原理的に可能ではあるが、以下の問題がある。

- 話者間変動要因が加わるため、話者に共通した音素コンテキストを表現するHMnetの構造が得にくい。
- HMnetを作成するためには、学習データとして多数の音素コンテキストを含むデータを用いなければならないが、これを多数話者分揃えるのは困難。
- HMnetの作成には単一の話者でも現状では膨大な計算量がかかり、これをさらに多数話者で行なうのは困難。

第一点の問題を回避するためには、構造決定と話者のスペクトルの変動の問題を分離して考える必要がある。そこで本アルゴリズムではHMnetの構造は話者間で共通と考え、話者1名の大量のデータからSSSアルゴリズムにより構造を決定し、そのHMnetの構造を用いてパラメータの学習を多数話者で行なう。

第二点のように環境依存モデルでしかも混合出力分布の学習をするには非常に多くのデータを必要とするが、この問題を解決するためパラメータの学習に移動ベクトル場平滑化方式 (Vector Field Smoothing: VFS) による話者適応手法 [8][10] を利用する。

また以上に述べた2つの対策は同時に、第三の問題点であるHMnet作成時の計算量の削減につながる。

SSSアルゴリズムによりHMnetの構造を決定した後に、比較的少量の複数話者の音声データより話者ごとのHMnetのパラメータを求める。

原理的にはBaum-Welchアルゴリズムによりパラメータ学習は可能ではあるが、一般に多数話者で大量の音声データを集めるのは困難である。特に学習データが少量の場合には以下の問題が生じる。

まず状態が共用されているものを除いてデータにHMnetで表現されうる全ての音素環境が出現しない限り、いくつかのパラメータが未学習のままとなる。また学習できたとしても、音素コンテキスト依存モデルでは一般の音素モデルに比べ、個々のパラメータの学習に使われるデータは少なくなり学習が不安定になる。

そこでBaum-Welchアルゴリズムの代わりとして、先に述べたVFS法を用いる。複数話者の音声データを用意し、話者ごとにHMnetのパラメータの学習をVFS法を用いて行なうことにより、複数話者分のHMnetが生成される。

3.3.5 話者混合化による不特定話者HMnetの作成

以上により求められた複数のHMnetを、話者混合化により1つの混合連続出力分布HMnetとして表現する。話者混合化は、同一の構造をもつHMnetにおいて、構造中同一の位置にある状態が持つ出力分布をひとつにまとめ混合連続出力分布として表すことにより行なわれる。

分岐確率は以下の理由でここでは等確率とした。まず適当な初期値を与えた後にBaum-Welchアルゴリズムによって分岐確率のみ再学習して与える方法が考えられるが、予備実験の結果、分岐確率の学習は特に効果が見られなかった。また先に述べた平均値の学習と同様、音声サンプルに多くの音素環境を含んでいないと、全ての分岐確率を学習するのが不可能である。さらに特に同一話者から生成されたガウス分布の分岐確率を全て同一に与えると、後で述べる話者重み学習(Speaker-Tied Weight Training)が可能となるという利点がある。

以上の話者混合の説明を図3.7(a)に示す。

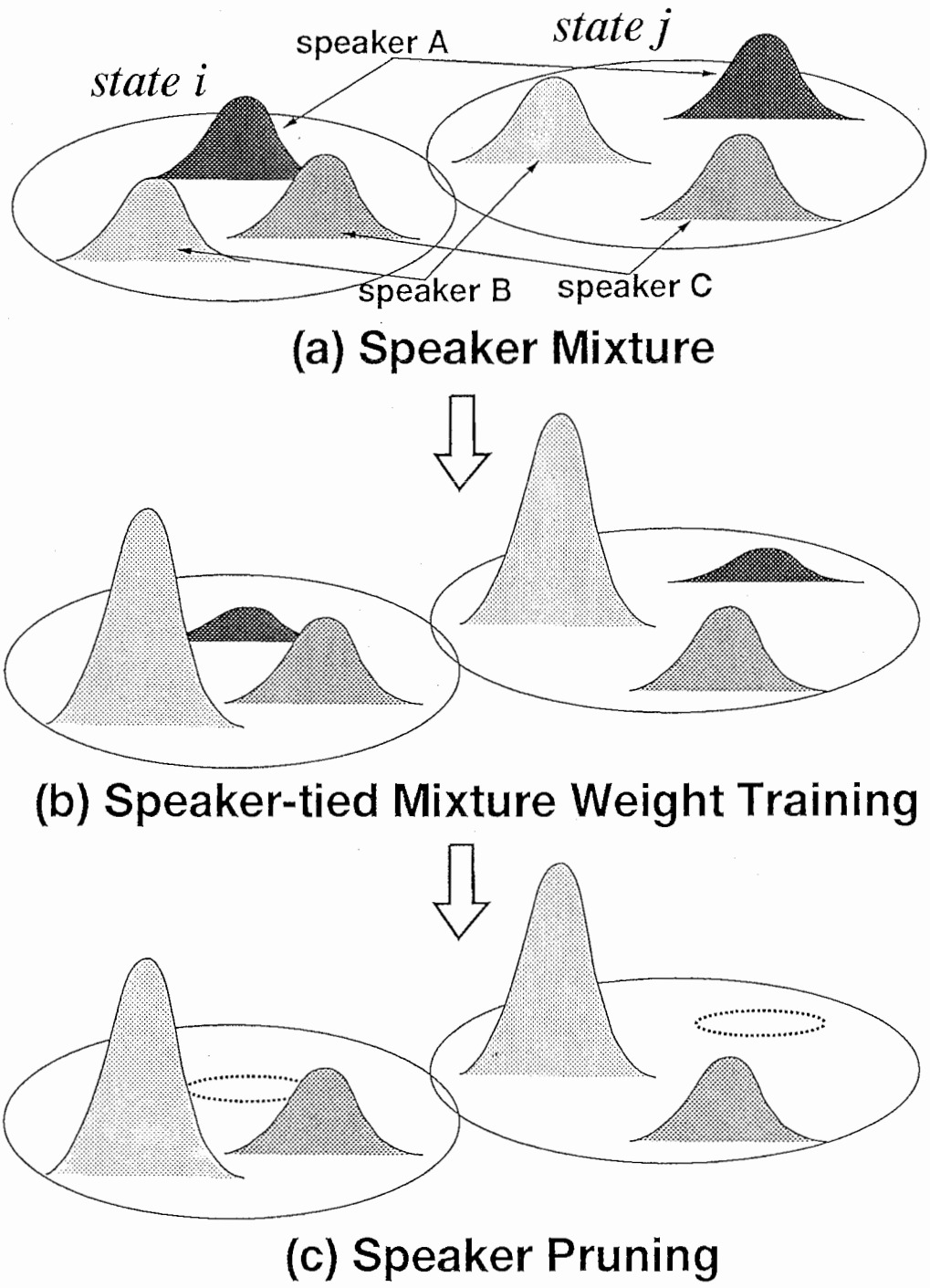

Figure 3.7: 話者混合 SSS の説明図

3.3.6 不特定話者音声認識実験

以上の手法の有効性を検討するために、以下の実験を行なった。実験条件を表 3.4に、使用データを表 3.5に、音響分析条件を付録 A.3に示す。

- 音素認識実験
  不特定話者を対象とした話者クローズド／オープンの認識実験を HMnet の状態数を変えて行なった。HMnet は話者 12 名から作成された混合数 12 の出力分布を持つものであり、話者適応まで含めた以下の実験ではすべてこの 12 混合の HMnet を使用した。

- 文節認識実験
  話者混合 SSS から生成された混合連続分布 HMnet と LR パーザ [7] を組み合わせた話者混合 SSS-LR により不特定話者文節認識実験を行った。LR のビーム幅は 256 で、一般文法 (1407 規則,1035 語) を使用した。

  比較実験として 4 状態 3 ループで共分散行列は対角要素のみの HMM を用いた HMM-LR の結果も求めた。HMM は 52 音素カテゴリについて求めた。混合連続分布 HMM を用いると、混合数を増加させた場合一般に認識率も向上するが、予備実験として混合数を変化させて認識率を調べたところ、ほぼ 20 混合で認識率が頭打ちとなったため 20 混合とした。ビーム幅は 1200 で SSS-LR と同じ一般文法を使用した。

図 3.8に音素認識実験結果を、また図 3.9に文節認識実験結果を示す。いずれの場合も状態数が増え表現できる音素コンテキストクラスの数が増加すると認識性能が向上する傾向が見られる。また文節認識実験において話者混合によるコンテキスト依存モデルを用いた場合には、通常の混合連続 HMM に比較して認識性能が向上しており、コンテキスト依存モデルを用いることの有効性が確認できた。

Table 3.4: 実験条件

	HMnet 作成		音素認識実験		SSS-LR 文節認識実験	HMM-LR 文節認識実験			話者重み学習 (+ 話者ブルーニング)	
	構造決定	パラメータ学習	closed	open		学習		認識	適応用	認識
話者	MAU	C	C	A	A	C		A	MMS,MMY,MSH	
タスク	LW	B	WI†	LW†	SB3	B	B+WI	SB3	B‡	SB3

† 最大 100 音素. ‡ 単語リストの先頭から使用.

Table 3.5: データリスト

使用単語 (文節) セット	
LW 単語セット	5240 単語
B 単語セット	音素バランスを考慮した 216 単語
WI 単語セット	520 単語 (LW 単語セットのサブセット)
SB3 文節セット	279 文節
話者セット	
A 話者セット	男性 9 名 (MHT,MMS,MMY,MNM,MSH, MTK,MTM,MTT,MXM)
C 話者セット	男性 12 名 (M001,M002,M003,M004,M101,M102, M103,M104,M401,M402,M403,M404)

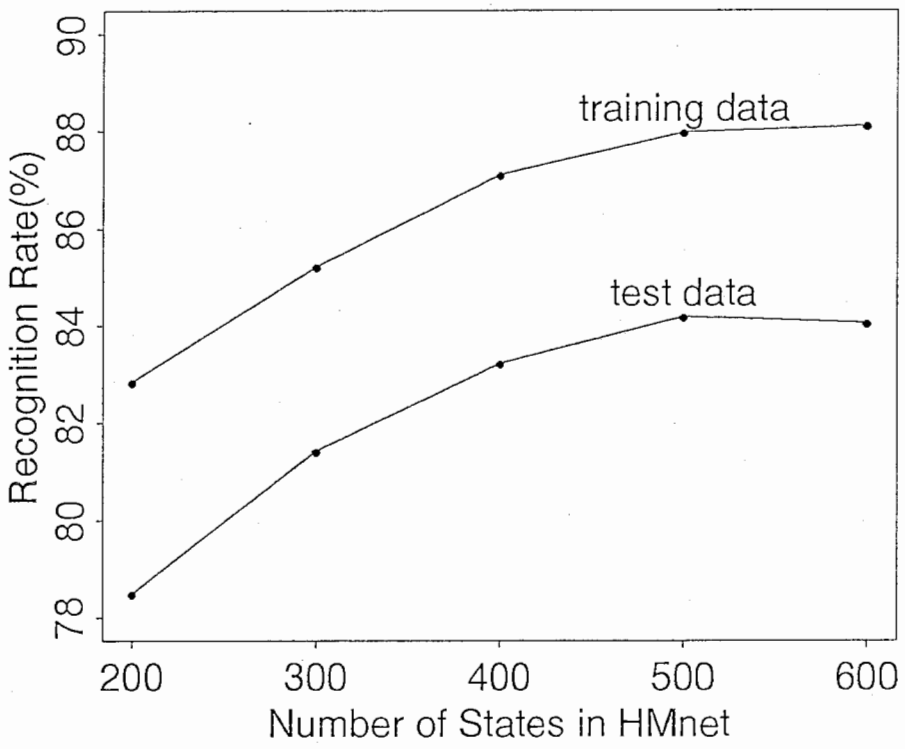

Figure 3.8: 不特定話者音素認識実験結果

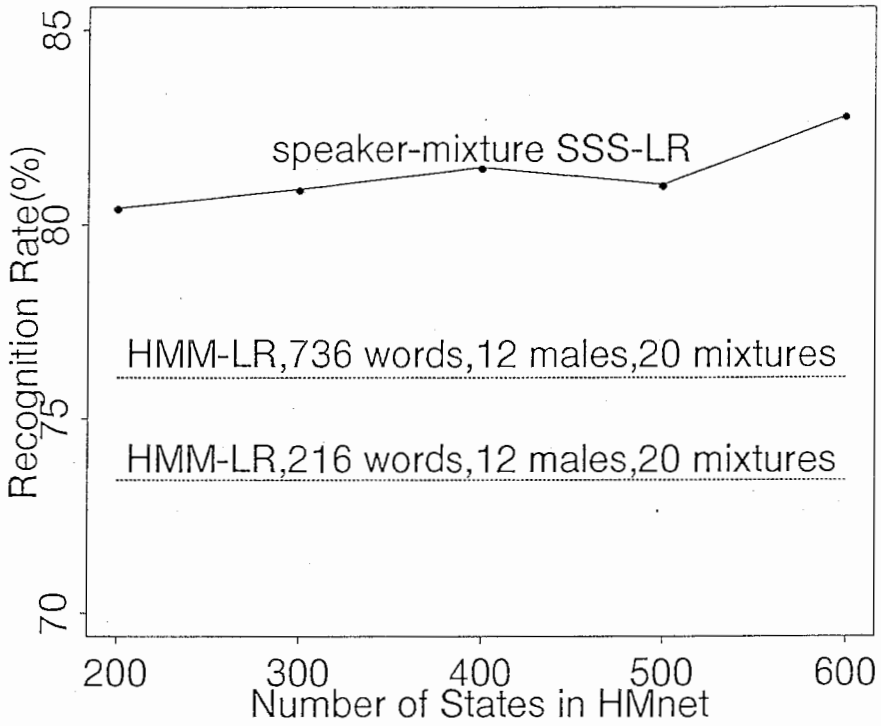

Figure 3.9: 不特定話者文節認識実験結果

3.3.7 まとめ

本節では不特定話者を対象とした環境依存音素モデルを作成するためのアルゴリズムとして話者混合 SSS を提案し、その有効性を示した。

今後の検討課題としては、まず HMnet の構造が 1 名の話者のみから作成されているが、この構造が不特定話者に対して普遍なものがどうか検討した上で、多数話者から求めた話者共通 HMnet[28] の構造を用いた認識実験を行なう必要がある。また 12 名の話者で不特定話者のモデルを作成しているが、何名まで話者を増やす必要があるか調べた上で、必要ならば話者クラスタリング手法を併用した方法を検討したい。

3.4 話者クラスタリング手法を用いた不特定話者音素モデル作成法

3.4.1 まえがき

近年 HMM(Hidden Markov model)、特に連続分布型 HMM(CMHMM: Continuous mixture density HMM) を用いた音声認識の研究が盛んである。CMHMM はいくつかの不特定話者音声認識システムに用いられ、その有効性が確かめられている (例えば [25][29])。CMHMM の学習には一般に Baum-Welch アルゴリズム [12] が用いられる。このアルゴリズムを用いて、特定話者や不特定話者用のモデルを作成する場合、両者で本質的に異なるのはデータの与え方のみである。つまり特定話者モデルを作成する場合は、単一の話者のデータだけを与えパラメータ推定をし、不特定話者のモデルを作成する場合は、多数話者のデータを与えパラメータ推定をする。

以上のように、従来の不特定話者のモデリングでは、話者個人の情報は一切保存しないのが一般的であった。これに対し、話者情報を利用した不特定話者モデルの作成も可能である。話者の情報をモデリングに利用することにより、従来行なわれているように話者の構造を考えずにモデルを作成するのに比較して、能率良くモデルが作成できる可能性がある。また得られたモデルは話者情報を持っているので、認識時における話者適応を行ない易いという利点を持つ。

筆者らは、不特定話者のモデルを作成する場合、話者の情報を用いたモデリングとして話者混合法を提案した (第3.3節参照)。この方法では、特定話者の単一ガウス分布の CMHMM を話者ごとに作成し、それを合成することにより、混合分布の不特定話者用 CMHMM を作成する。この方法で作成された CMHMM では、混合要素の各ガウス分布が単一話者のモデルを表すという特徴がある。そのため、「話者重み学習」により話者適応をおこなったり、「話者プルーニング」により認識時の計算量削減することが可能となった [30]。

しかし従来の話者混合法では、話者数が CMHMM の混合数と等しくなるため、話者が増えるとモデルパラメータが増大し、実際的には多数の話者からモデルを作成できないという欠点があった。不特定話者のモデルを作成する場合、多数の話者のデータの利用が重要であり [31]、それが不可能なのはモデル作成

の上で重大な欠点であった。

そこで、この問題を解決するために、話者モデルのクラスタリングを利用した不特定話者モデルの作成法 (HMMs' Composition and CLustering: CCL) を提案する。この方法では、特定話者モデルをクラスタリングし、音響的特徴が類似の話者を集め、類似の話者のモデルをまず合成する。こうして出来上がった各話者クラスタの特徴を表すモデルを、さらに話者混合法により合成する。このように2段階に合成してモデルを作成することにより、話者情報を保存し不特定話者のモデルに反映する手法をとった。この場合クラスタ数は任意に定めることができるので、任意の混合数のモデルを作成することができる。

以上の方法を用いることで、以下の効果が得られる。まず先に述べたように、作成されたモデルは話者の構造を持つため、「話者重み学習」による話者適応や「話者プルーニング」による認識時の計算量の削減が可能である。また従来、HMM の学習のための Baum-Welch では学習話者数を変更する場合は、最初から学習をやり直す必要があった。しかし本手法によれば、特定話者のモデルを合成するので、特定話者のモデルを用意すれば、簡便に話者の追加・変更が可能である。たとえばデータベースの増加に伴い、話者をモデルに追加したい場合、全データで学習のやり直しをせずに、追加話者についてのみ特定話者モデルを作成すれば、簡便に不特定話者モデルに合成することが可能である。さらに本手法により認識時の性能を低下させることなく、大幅にモデル作成のための時間を削減することが可能となった。実際特定話者モデルがあらかじめ用意されている場合は、2～3分程度で不特定話者モデルを作成できることがわかった。

本稿では、提案したモデル作成法の説明と評価のほか、本手法で不特定話者モデルを作成した場合の、話者の種類と話者数についての検討を行なった。従来話者数については検討した報告があるが [31]、どのような話者がモデル作成には有効かについての検討はあまり行なわれていない。クラスタリングにより話者空間を代表するような話者を少数選べば、多数話者から作成した不特定話者のモデルに匹敵するモデルが作成できる可能性がある。そこでクラスタリング手法により、代表話者を選択した場合と、ランダムに話者を選択した場合の比較についても述べる。

3.4.2 モデル合成法による不特定話者モデルの作成 (CCL)

CCL アルゴリズムの概要

複数の特定話者モデルをクラスタリングした後に、各クラスタに属する話者モデルを選択し、それらを合成することにより不特定話者モデルを作成する方法 (CCL) について述べる。実験ではモデルとして音素環境依存 HMM の一種である HMnet(Hidden Markov network) [5] を使用した。まず以下に CCL のアルゴリズムの概要を示す。

STEP 1 (HMnet 構造の作成) 単一話者の多数データ（ここでは 2620 単語）から HMnet の構造を作成する。

STEP 2 (特定話者 HMnet のパラメータ再推定) N 人の話者から N 個の特定話者用単一ガウス分布 HMnet を作成。この場合の HMnet の構造は、全ての話者で共通とし、STEP 1 で得られた HMnet の構造は変えずに、パラメータのみ再推定して特定話者 HMnet を求める。

STEP 3 (HMnet のクラスタリング) 上記 N 個の HMnet を K クラスにクラスタリング。K は作成される不特定話者モデルの混合数となる。

STEP 4 (話者クラスタモデルの作成) 各クラスに属する HMnet を単一ガウス分布 HMnet に合成する。

STEP 5 (話者混合法による不特定話者モデルの作成) STEP 4 により求められた K 個の単一ガウス分布 HMnet を話者混合法を用いて合成して、K 混合分布 HMnet を作成。

HMnet 構造の作成

本稿では、音素コンテキストによる音声の変動は話者間で共通なものと考え、1名の話者のさまざまな音素コンテキストを含むデータから、SSS アルゴリズムを用いて初期 HMnet を作成する。本アルゴリズムでは、この HMnet の構造を、特定話者・不特定話者モデルに共通なものとして利用した。話者間で HMnet の構造が本当に共通であるかどうかについては十分に検討すべき問題

ではあるが、数種類の音素においての検討結果 [28] からは特に顕著な HMnet 構造の話者依存性は見られなかったため、このように仮定した。

特定話者 HMnet のパラメータ再推定

本節ではアルゴリズムの STEP 2 で行なわれる、特定話者 HMnet のパラメータ再推定について述べる。

STEP 2 では STEP 1 で得られた HMnet の構造は話者で共通して使用することが可能であることを仮定し、パラメータ再推定のみをおこなって、特定話者 HMnet を複数個作成する。このパラメータ再推定には、データ数削減及び学習時間短縮のため移動ベクトル場平滑化法（VFS）[8] を用い平均値のみ変更した。

VFS は少量の適応データにより話者適応を行なうための手法である。これは見方を変えれば、データ量が少ない場合のパラメータ再推定にも利用できる。VFS では標準話者の音声パラメータ空間を連続的な移動伸縮によって、未知話者の音声パラメータ空間へ変換する。本手法では、平滑化と補間を行なっているので、比較的少量のデータで効率良く特定話者モデルのパラメータ再推定を行なうことが可能である。

HMnet のクラスタリング

本稿では、SPLIT 法 [32] で用いられたクラスタリングアルゴリズムに基づく方法によって HMnet のクラスタリングを行なった。SPLIT 法では、2 のべき乗のクラスタを作成する一般的な LBG アルゴリズム [33] とは異なり、歪みが最大となるクラスタを順次分割するため任意の数のクラスタを作成できる。またクラスタリングを行なう前に、あらかじめ要素間の距離テーブルを作成する。これにより、クラスタ中心の初期値をヒューリスティックに与えなくとも良いという利点がある。結局あらかじめ与える必要があるのは距離に対する閾値、又はクラスタ数のみで、この値さえ与えれば、完全に自動的にクラスタリングの結果が得られる。

SPLIT で用いられた方法では、距離としてユークリッド距離による DP マッチングスコアを用いていた。本研究では確率モデルを用いるため、以下で定義

するHMnet間の確率的距離尺度を用いる。

構造の等しい2つのHMnet、$M^{(1)}$と$M^{(2)}$間の距離を以下のように定義する。この場合、文献[34]を参考にし、初期状態確率や状態遷移確率を無視し、HMM間の類似度を求める場合に重要なパラメータと考えられる出力確率のみの距離$d(b_j^{(1)}, b_{g(j)}^{(2)})$により定義した。

$$D(M^{(1)}, M^{(2)}) \triangleq \frac{1}{N}\sum_{j=1}^{N} d(b_j^{(1)}, b_{g(j)}^{(2)}) \tag{3.3}$$

但し$b_j^{(i)}$は$M^{(i)}$の状態jにおける出力確率分布を、Nは$M^{(i)}$の状態数をそれぞれ表す。

ここで、$g(j)$はこの式を最小にする状態間の写像関数である。共通の構造を持つ2つのHMnetについて、以下の仮定をする。

$$g(j) = j \tag{3.4}$$

$d(b^{(1)}, b^{(2)})$はKullback情報量[34][35]や、Chernoff距離、Bhattacharyya距離[36]などで計算できる。ここではBhattacharyya距離と、従来一般に用いられているKullback情報量の比較検討を予備実験により行なったが、その結果性能は大差なかったため、計算量の少ないBhattacharyya距離を用いた。またKullback情報量による尺度の定義に対し、Bhattacharyya距離では対称性、正値性を満たすという特徴がある。

出力分布が単一ガウス分布で表せる場合、$d(b^{(1)}, b^{(2)})$は以下のようにBhattacharyya距離を用いて計算できる。

$$\begin{aligned} d(b^{(1)}, b^{(2)}) = & \\ & \frac{1}{8}(\mu^{(1)} - \mu^{(2)})^t \left(\frac{\Sigma^{(1)} + \Sigma^{(2)}}{2}\right)^{-1}(\mu^{(1)} - \mu^{(2)}) + \\ & \frac{1}{2}\ln\frac{|(\Sigma^{(1)} + \Sigma^{(2)})/2|}{|\Sigma^{(1)}|^{1/2}|\Sigma^{(2)}|^{1/2}}. \end{aligned} \tag{3.5}$$

ここで$\mu^{(i)}, \Sigma^{(i)}$はそれぞれ平均ベクトル、共分散行列を表す。

話者クラスタモデルの作成

本節ではアルゴリズムのSTEP 4で行なわれる、各クラスタに属する複数の単一ガウス分布HMnetを単一ガウス分布HMnetに合成し、話者クラスタモ

デルを作成する方法について述べる。

各クラスタごとに特定話者のモデルを合成する際は、各クラスタのデータによりパラメータを再推定する方法(CCL1)と、再推定なしに合成する方法(CCL2)の2通りを比較した。CCL2の場合はパラメータ再推定をしないので、高速にモデルの合成が可能である。実際話者15人のモデルを合成する場合は1分程度、話者285人のモデルを合成する場合でも2～3分程度で終了する（HP社、HP9000/735を使用）。それに対しCCL1は、パラメータの再推定を行なうため計算時間はかかるが、精度よくパラメータが求まることが期待できる。

CCL1およびCCL2のアルゴリズムを以下に示す。

CCL1 各クラスに属する話者のデータでパラメータ再推定を行ない、各クラスごと単一ガウス分布HMnetを作成する。パラメータ再推定には、データ数削減及び学習時間短縮のためVFSを用いた。

CCL2 以下の式により各クラスに属するHMnetを単一ガウス分布HMnetに合成する。この式は複数のガウス分布を単一ガウス分布と見なして求めた場合の平均値、分散を表す[37]。

ここで、$\mu_j^{(i)}, S_j^{(i)}$ は i 番目のHMnetの状態 j における出力分布（ここでは単一ガウス分布）の平均値及び分散を、$n_j^{(i)}$ は i 番目のHMnetの状態 j におけるサンプル数を、$\hat{\mu}_j$, $\hat{S}_j$ は合成後のHMnetの各状態の出力分布の平均値及び分散を表す。

$$\hat{\mu}_j = \sum_i w_j^{(i)} \mu_j^{(i)} \tag{3.6}$$

$$\hat{S}_j = \sum_i w_j^{(i)} S_j^{(i)} + \sum_i w_j^{(i)} (\mu_j^{(i)} - \hat{\mu}_j)^2 \tag{3.7}$$

このとき、

$$w_j^{(i)} = n_j^{(i)} / \sum_i n_j^{(i)} \tag{3.8}$$

話者混合法による不特定話者モデルの作成

本節ではアルゴリズムの STEP 5 で作成される、話者クラスタモデルを話者混合法により混合分布 HMnet へ合成する方法について述べる。

まず、構造及び遷移確率が等しく、出力分布のみが異なる単一ガウス分布 HMnet が複数個存在すると仮定する。これらの HMnet を、構造中で同一の位置にある状態が持つガウス分布に混合重みを与え混合連続出力分布として表す。これにより話者構造を持つ HMnet、具体的には各混合要素に話者クラスタのラベルが付いた図 3.7(a) のような、不特定話者認識用 HMnet が得られる。

話者混合法で作成したモデルはその性質から、「話者重み学習」による話者適応、及び「話者プルーニング」による認識時における計算量の削減が可能である。「話者重み学習」では話者重みを学習することにより話者適応を行なう (図 3.7(b))。また「話者プルーニング」では、話者重みを学習した結果、重みの値の小さくなった話者をモデルから外して認識時の計算量を削減する (図 3.7(c))。

3.4.3 認識実験

実験条件

本手法の有効性を確認するために、不特定話者音声認識の実験を音素認識および文節認識により行なった。音響分析条件は付録 A.3に示す。学習データは 285 人が発声した、文節発声による音素バランスを考慮した 50 文である。評価データは、男性 5 人、女性 5 人の計 10 人の発声した、国際会議予約に関する文節発声による会話データ、およびそこから視察により切り出された音素である。

HMnet の構造は 1 人の話者が発声した 2620 単語より作成し、状態数は 200 状態とした。さらに 1 状態 10 混合の無音のモデルを別に学習し 200 状態の HMnet と合成した。このため合計の状態数は 201 状態になっている。この HMnet の構造は以下の実験に共通して用いた。

実験のため 285 個の特定話者用 HMnet をあらかじめ作成したが、データ数削減及び学習時間短縮のため VFS を利用しパラメータ推定を行なった。文節認識実験は SSS-LR 連続音声認識法 [7] により行なった。この時のビーム幅は 1200、規則数 1407、語彙数 1035 の文節内文法を用いた。音韻パープレキシティ

は5.9である。

不特定話者モデル作成における話者の影響の検討

不特定話者モデルを作成する場合の、話者の選択及び話者数による影響について検討した。不特定話者モデルを作成する場合、データにおける話者のバリエーションがどの程度モデルに影響するか検討することが重要である。話者数については南ら[31]によって検討されている。この報告から人数が多いほど認識率は向上するが、ある程度で飽和する傾向が見られる。この実験では話者はランダムに選択されており、選択された話者が話者空間を代表したものであるとはいえない。そこで本稿では話者数のみの検討だけではなく、話者を話者空間を代表するように選択した場合どのような効果が見られるかについても検討を行なった。

実験では(1)クラスタリングにより多数話者(285人)からN人を選択する、又は(2)多数話者からランダムにN人を選択する、の比較を行なった。(2)では5回ランダムで選択して実験を行ない、その平均値を結果とした。話者の選択は、まず多数話者を2.3節で述べたクラスタリング法によりNクラスタに分割し、その後各クラスタの中心のモデルの話者を被選択話者とすることにより行なった。モデルの作成は2.1節のCCLにより行なった。この方法では、HMnetの混合数は標準話者数以上にはできないので、例えば話者として15人用いた場合は、15混合以下のHMnetにより実験を行なった。

音素認識実験結果を図3.10に示す。ランダムに話者を選択した場合に比べ、クラスタリングで選択した話者が概して高い認識率を示す。混合数が増加すると認識率の向上も飽和するが、その場合でもクラスタリングにより話者を選択した方が高い認識率を示す。この結果から、話者空間を代表する話者を選択することは、不特定話者モデルの作成に有効であることが分かった。以上の結果に基づき、次節の実験では話者285人中から15人をクラスタリングにより選択して用いた。

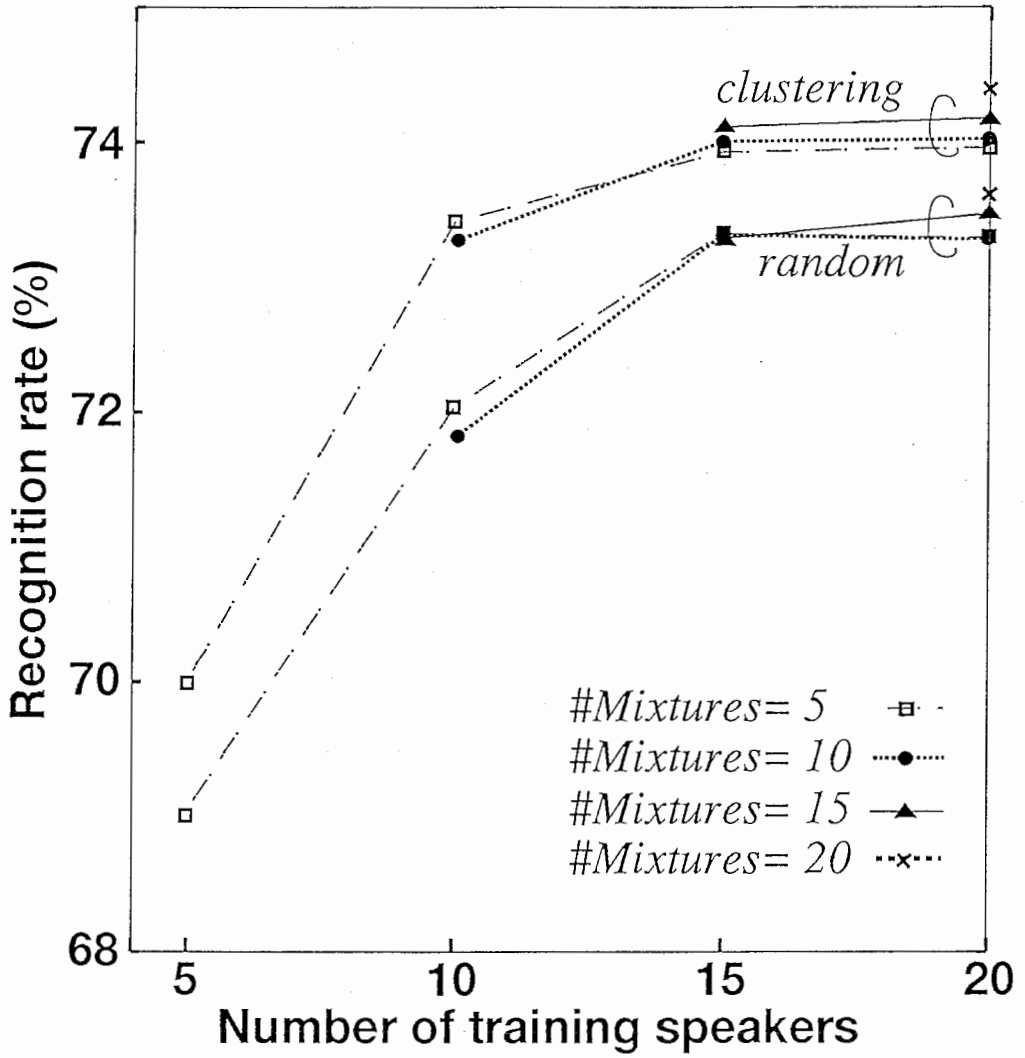

Figure 3.10: 285 人から学習話者をランダムに選んだ場合と、クラスタリングで選んだ場合の学習人数による音素認識率の比較。(CCL2 によりモデルを作成)

モデル合成法 (CCL) の有効性の検討

CCL で作成したモデルの有効性を音素認識による比較実験で検討した。比較実験の流れ図を図 3.11に示す。学習話者からクラスタリングにより 15 名を選択し、それぞれの話者用の特定話者 HMnet 15 個を合成して 5、10、15 混

合の HMnet を作成する。

実験内容を以下に示す。

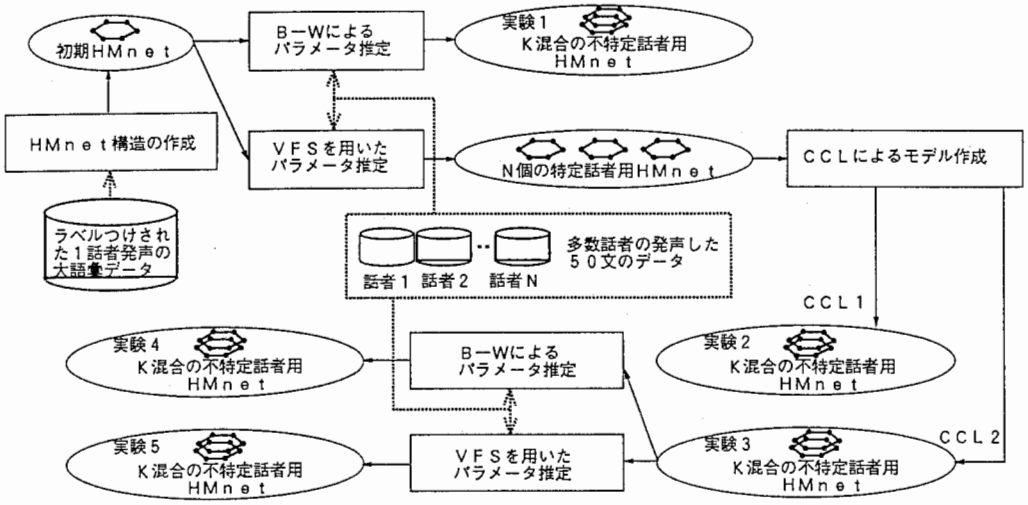

Figure 3.11: 認識実験の流れ図

CCL1 と CCL2 の比較

提案した 2 種類の CCL の比較を実験 2(CCL1) と実験 3(CCL2) により行なった。先に述べたように CCL1 は合成の際パラメータ再推定を行なう方法、CCL2 はパラメータ再推定を行なわずに合成する方法である。

Baum-Welch によるパラメータ再推定と CCL の比較

Bawm-Welch と CCL を比較するため、Baum-Welch によるパラメータ再推定の実験を行なった。Baum-Welch によるパラメータ再推定では、初期値によって性能が変わるので、初期値の与え方の異なる 2 種類の Baum-Welch によって実験を行なった。一つは CCL2 により合成して作成した HMnet を初期モデルとして通常の Baum-Welch で学習した場合（実験 4）である。もう一つは各学習データを時間方向に状態数分だけ分割し、それぞれを混合数と同数にクラスタリングして与え初期出力確率値を計算し、これを初期値として Baum-Welch で学習した場合についての実験である（実験 1）。

視察によるラベルが与えられない場合、実験1のように、データを時間分割して与えるのが一般的である。またラベリングデータが大量に得られる場合は、ラベリングデータにより学習した後に連結学習をすると高い認識率が期待できる[38]。これは連結学習のための初期モデルとして高い性能のものが与えられるためである。実験4は初期モデルが既に高い認識率を示すものを与えるという点で、大量なラベリングデータが与えられた場合に近い効果が期待できる。渡辺らは[39]この初期値について、(1)視察によるラベルにより与える方法、(2)状態の数だけ時間方向にデータを分割する方法、(3)特定話者モデルを初期値とする方法、について比較検討している。認識率は(1) > (3) > (2)であるがその差は大きくない。一般にBaum-Welchの初期値は以上のいずれかの方法で与えられることが多い。このためCCLによるモデルの性能を検討する場合、実験1と実験4のような初期値が異なる場合のBaum-Welchによるパラメータ再推定と比較する必要があると考えた。

実験1と実験4のBaum-Welchにおける繰り返し回数は21回である。この回数は予備実験によりある程度収束する値に定めた。

VFSによるパラメータ再推定とCCLの比較

さらにパラメータ再推定の異なる方法として、初期モデルをCCL2により作成し、VFSでパラメータ再推定した場合（実験5）も検討した。

実験結果

結果を表3.6に示す。表において括弧はHP9000/735による計算時間を表す。但しHMnetの初期構造を決めるための計算時間は各実験とも共通なため、ここでは除いている。

実験2 (CCL1)及び3 (CCL2)で得られたモデルは、話者の構造を保持しているため、このモデルを元に話者適応する場合、「話者重み学習」のような話者適応手法が使えるという利点がある。CCL1とCCL2を比較すると、CCL2の方が高い認識率が得られている。パラメータ推定をしなおしたという点から考えると、CCL1の方が有利なように考えられるが結果は逆である。これはVFSでは平滑化を行なっているため、データ数が少ない場合はこの平滑化が有効に

働くが、パラメータ数が多い場合、逆にパラメータ推定に制限が加わり悪影響があるためと考えられる。

CCL1 及び CCL2 による認識率は、初期値が異なる場合の Baum-Welch によるパラメータ再推定の結果である実験 1 と実験 4 の間にくる。このことから少なくともラベルなしで行なう Baum-Welch パラメータ再推定よりも良い結果が得られることが分かる。また実験 4 や 5 のように、合成法によるモデルを初期モデルとしてさらにパラメータを再推定すると高い認識率が得られる。このことから、本手法は高精度のモデルを作成する場合の初期モデル作成としても役立つ。但しこの場合、モデル中の話者の構造は保持されない。

計算量は Baum-Welch を使用する場合は混合数に応じて非常に多くなる。実験 1 と実験 4 は初期モデルが異なるだけでパラメータ数は同じなので、多少計算時間は違っているが、理論的には同じ計算量になる。CCL1 及び CCL2 は計算時間が非常に少ない。特に CCL2 ではモデル合成時にパラメータ推定を行なわないので計算量が特に少なくて済む。実験 1 の Baum-Welch による方法と比較すると、計算量は約 1/20 ～ 1/60 程度になる。混合数が増加するに伴いその差は拡大する。また CCL2 における計算時間の大部分は特定話者モデルの作成にかかる時間で、モデル合成自体は 1 分前後で終了する。つまり特定話者のモデルをあらかじめ用意しておけば、非常に簡便に話者の変更、追加が可能である。3.2 節の結果を考慮すると、まず 3.2 節の方法により代表話者を選択し、その後 CCL でモデルを合成する方法が能率のよい不特定話者モデルの作成法と思われる。

次に、CCL2 による不特定話者モデルを用いて文節認識を行なった結果を表 3.7に示す。この実験では 285 人全てのモデルをクラスタリングして、5、10、15 混合の HMnet を作成した。結果は 5 位候補までで約 95% と高い認識率が得られた。この結果 CCL によるモデルは文節認識でも有効なことが確認できた。

Table 3.6: CCL、VFS、Baum-Welch による HMnet を用いた音素認識率 (%)。括弧内はモデル作成の時間 (単位：時間)。

実験番号	手法 / 混合数	5	10	15
1	Baum-Welch	64.6 (102.0)	67.4 (188.5)	70.2 (276.1)
2	CCL1	72.5 (14.5)	72.3 (14.4)	72.5 (14.3)
3	CCL2	73.6 (4.4)	74.0 (4.4)	74.1 (4.4)
4	CCL2 + Baum-Welch	77.2 (100.7)	77.8 (179.1)	78.1 (259.6)
5	CCL2 + VFS	74.5 (34.3)	76.0 (62.5)	76.1 (88.9)

Table 3.7: CCL を用いた不特定話者文節認識率 (%)

候補 / 混合数	5	10	15
1 位	76.6	77.4	77.6
5 位	94.9	95.2	95.1

3.4.4 まとめ

本稿では複数の特定話者モデルを合成することにより、話者の構造を持つ不特定話者音素モデルを作成する方法 (CCL) を提案した。話者構造を保ちながらモデルを作成することにより、従来の Baum-Welch 学習によって得られたモデルの性能を損なうことなく、短時間でモデルが作成できることを示した。

計算時間を比較すると約1/20 ～ 1/60程度となり、この差は混合数が増すほど増加する。

またこの方法により、Baum-Welchに比較してはるかに簡便にモデルに対する話者の追加・変更が可能となる。Baum-Welchでは話者を追加・変更する場合は学習をやり直す必要がある。本方法によれは追加話者の特定話者モデルが作成されていれば、単時間で不特定話者に追加することができるし、削除も簡便にできる。実際合成法2によれば、話者15人の特定話者モデルから、不特定話者モデルを作成する時間は1分程度、話者285人の場合でも2～3分程度である。

CCLによって作成されたモデルの性能は、従来のBaum-Welchでクラスタリングにより初期値を与えた場合に比較して、高い認識率が得られる。またCCLによって作成されたモデルを初期モデルとして、パラメータの再推定を行なえばさらに高い認識率が得られる。このことからCCLによるモデルは精密な不特定話者モデルを得る場合の初期モデルとしても有用である。さらにCCLによって得られたモデルは話者の構造を保持しているために、「話者重み学習」のような話者適応法を用いることができ、話者適応する場合有利になると考えられる。

以上のCCLに対する検討のほか、不特定話者音素モデル作成における話者の選択と話者数について検討を行なった。その結果、話者をクラスタリングにより選択することが有効であることが分かった。

今後はさらに、多数の話者をモデル作成に用いた場合の、混合数と認識率の関係について検討を行いたい。また本手法で作成されたモデルをもとにした話者適応、特に「話者重み学習」のような話者構造を利用した話者適応の有効性を検討する予定である。

第 4 章

少量のデータによる話者適応

4.1 まえがき

本論文では第1章で述べたようにデータ量に応じた話者適応法の開発を目指している。しかし従来の研究を見ると数十秒〜数分程度の比較的適応データを要する、自由パラメータ数の多い適応法は存在するが、より少ない数秒程度のデータで効果的な話者適応の研究があまりなされていない。そこで本章では少ない適応データで効果的な話者適応法の検討を行なった。

第3.3節では話者混合SSS法による不特定話者音素モデルの作成について述べた。この方法で作成された話者混合モデルでは、各混合分布に話者のラベルが付与されている。この特性を利用した話者適応が可能である。そこでまず第4.2節では話者混合モデルの話者重みを制御することにより話者適応する方法についてのべる。

しかしこの方法では話者の数を増加すると適応時の計算量が増大し、実際上モデルに用いる話者の数が制限される。そこで話者を木構造に構成し各ノードに位置する話者または話者クラスタモデルを選択することにより能率的に大量話者からの選択を可能にする話者適応法について第4.3節で述べる。

4.2 話者重み学習による話者適応と話者プルーニング

4.2.1 まえがき

従来の研究で多数話者から作成した不特定話者用の音素モデルを用いることにより、比較的良好な認識率が得られることが分かっているが [40] [41][42]、特定話者を対象とした音声の認識率と比較するとやはりまだ認識率は低い。そのため不特定話者モデルをベースとした話者適応を行なうことにより認識率の向上を図る。さらに不特定話者を対象とした混合出力分布モデルを作成した場合、混合数の増加に従って認識時の計算量が多くなるため、これの削減法が必要となる。

そこで本報告では話者混合法の特徴を利用した話者重み学習 (Speaker-Tied Weight Training) による不特定話者モデルからの高速な話者適応法、及び話者適応しつつ混合出力分布の混合数を削減する手法である話者プルーニング法を提案し、そのアルゴリズムの説明及び認識実験結果について述べる。

4.2.2 話者重み学習を用いた話者適応による音声認識

話者重み学習

話者混合 SSS の特徴を利用した、話者重み学習による教師付きの話者適応法を提案する。これまでも VQ 型 HMM における話者重みに関する研究が行なわれてきたが [43][44]、本研究では混合連続分布 HMM[3] の一種である混合連続分布 HMnet の話者適応法であり、話者重みを混合係数として扱った点に特徴がある。

話者混合 SSS で作成された混合連続出力分布 HMnet では、各混合出力分布を構成する混合成分はどの話者のデータから生成されたものであるかという由来が分かっている。従って、各混合成分への分岐確率は各話者への重み係数と理解できる。このため同一話者に由来する混合成分にかかる分岐確率、つまり話者重み係数を「結び (tied)」として扱うことが可能である。

この話者混合 SSS の性質を利用して話者重み学習による話者適応法を提案する。以下にそのアルゴリズムを示す。出力分布の平均値・分散・遷移確率は更新せず重み係数のみを話者間で「結び」として Baum-Welch アルゴリズム

を用いて更新する。この学習には連結学習を用いるので音素位置のラベルは必要としない。

この話者重み学習では学習対象パラメータ数はモデル作成に使用した話者数と一致する。このため従来提案されている重み学習による話者適応 [45] や、一般的に用いられている Baum-Welch アルゴリズムによる全パラメータの学習に比較して学習対象パラメータ数が非常に少ないため、少量の適応用データにより高速に話者適応が行なわれる可能性がある。例えば全ての重みを独立に学習する方法では HMnet の状態数が 200 の場合、$12(mixtures) \times 200(states) = 2400$ のパラメータの学習を要する。これに対し、本実験では混合連続分布 HMnet を 12 名の音声データから作成しているので、話者重み学習での学習対象パラメータ数は 12 のみである。

この話者重み学習による話者適応は一種の確率的話者選択アルゴリズムと見ることもできる [46]。従来から複数話者からの話者モデル選択による話者適応が提案され [47] その有効性が明らかにされている。しかしながら用意された話者モデルのいずれにも適合しない音声が入力された場合、認識の精度が低下する恐れがある。

今回提案する手法はただ 1 名の話者モデルを選択するのではなく、話者重み学習をすることによって、話者選択を重み付けによる表現で行なう点が異なる。特に話者重みを計算するに当たっては、話者間距離などのような認識に用いる距離とは異なる評価基準によらずに、Baum-Welch によるパラメータ推定アルゴリズムを用いることにより、適応時の基準と認識時の基準が一致しているという特徴がある。

以上話者重み学習の説明を図 3.7(b) に示す。

話者プルーニング

不特定話者を対象とした音声認識で話者変動を精密に表現しようとした場合、一般に多くのガウス分布でモデルを表現する必要があり、そのため認識時の計算量の増加が問題となる。そこで本節では話者適応をしつつモデルの単純化をおこなう、話者プルーニング法を提案する。

アルゴリズムを以下に示す。HMnet の混合出力分布のうち、話者重み学習

により重み係数があらかじめ設定された確率以下になった場合、その重み係数を 0 におきかえる。その後混合出力分布の重みの和が 1 となるよう重みを再配分する。

以上の方法では同一話者に由来する「結び」の関係になっている混合分布を全て削減するため大幅にモデルのパラメータを減らすことができ、認識に必要な計算量を大きく減少できる。

また話者プルーニング法によって混合出力分布の混合数を削減した場合、最適な重み係数が異なってくる可能性があるため、混合数を削減した後に話者重みを再学習する方法についても検討を行なった。

以上話者プルーニングの原理を図 3.7 (c) に示す。図に示すように同一話者に由来する話者重みの小さい混合成分を全て削除することによりモデルの単純化を行なう。

話者プルーニングは話者重み学習によって得られる重み係数によって話者を決定する方法と見なすこともできる。このため Baum-Welch アルゴリズムを利用した話者認識など、話者の類似度の判定が必要な手法への応用も考えられる。

話者適応実験

話者重み学習による話者適応の認識性能を検討するため不特定話者モデルからの話者適応実験を、また話者プルーニングの有効性を確認するため話者重み学習＋話者プルーニング手法を用いた場合の、認識時の計算量削減について検討した。

話者適応実験は単語で話者適応した後、同一話者の文節を認識することにより行なう。話者プルーニングは重みが 0.05 以下の話者を切り捨てることにより行なった。さらに話者プルーニングをした後に話者重みの再学習も行なったが、特に再学習の効果は見られなかった。比較として、すべての状態のすべての重みを学習する方法についても実験した。

認識実験の簡単化のため認識率の低い話者 3 名を選んで実験した。実験条件を表 3.4に示す。また音響分析条件は不特定話者音声認識実験と同一とした（付録 A.3）。

図4.1に話者重み学習による話者適応、話者重み学習＋話者プルーニング、全重み学習の3種類の実験の結果を示す。話者適応の結果をみると、いずれの話者も1～5単語程度の非常に少ないサンプルで認識率の向上が得られることがわかった。

表4.1に話者適応用の単語リストを示す。本来ならば適応用の単語はいろいろ入れ換えて実験するべきであるが、認識実験の簡単化のため、このリストの先頭から使用した。例えば適応単語数が1の場合は「勢い」という単語で話者適応を行なっている。話者3名の単語「勢い」の平均発話時間は0.64secなので、/i/・/k/・/o/の3音素のみ含まれる0.64secの音声で話者適応の効果が出ていることが分かる。

また全ての重みを独立に学習する方法では、学習対象パラメータが多いために学習単語が少ないと逆に認識率が低下している。

また話者プルーニングを行なうと、表4.2に示すように各話者で出力分布数が1/2～1/12程度に減少するが、特に認識率の低下は見られず、話者適応をしつつ混合数の削減が可能であることが分かった。また1～5単語で混合数が大幅に減少し、その後学習単語数を増やしてもほとんど混合数が変化しないことから、話者適応用データは5単語程度で十分であることが分かる。

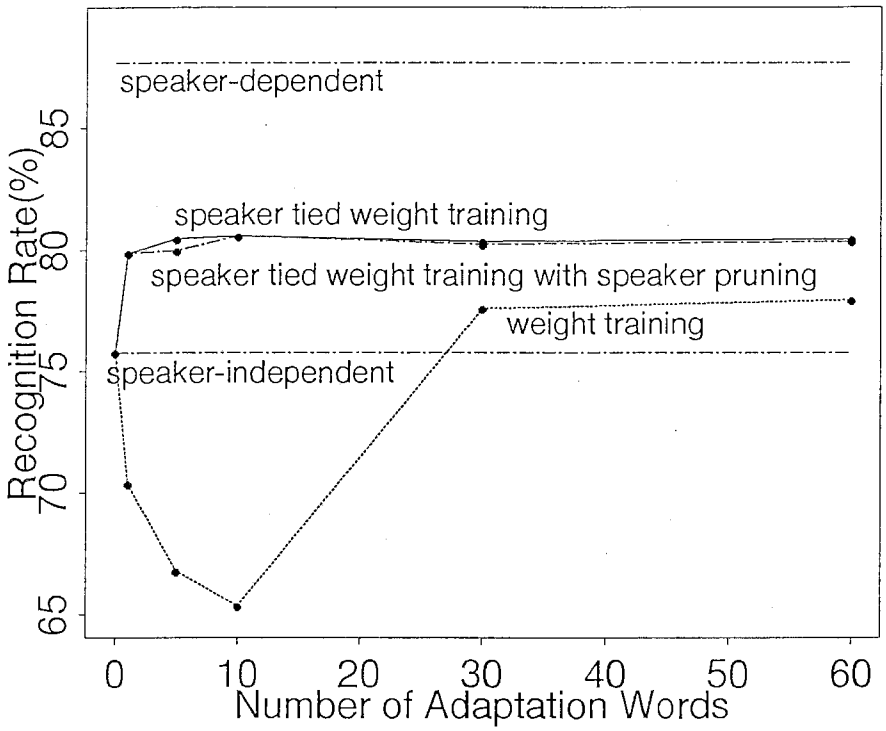

Figure 4.1: 話者適応後の文節認識実験結果 (200 状態)

Table 4.1: 話者適応用単語リスト（先頭から 10 番目まで）

1. ikioi （勢い）
2. iyoiyo （いよいよ）
3. urayamashi: （羨ましい）
4. omoshiroi （面白い）
5. guai （具合）
6. zairyo: （材料）
7. zyu:ichigatsu （十一月）
8. syu:kyo: （宗教）
9. zyuNban: （順番）
10. suichoku （垂直）

Table 4.2: 話者プルーニング後の混合数

話者	学習単語数					
	0	1	5	10	30	60
MMS	12	4	3	5	6	5
MMY	12	2	1	1	1	1
MSH	12	2	2	1	1	1

4.2.3 まとめ

本節では話者重み学習を提案し、少数の学習データで話者適応が可能であることを示した。さらに話者プルーニングにより、認識率の低下なしに話者適応と同時に混合数の削減ができ、認識時の計算量を減少することが可能であることを示した。

4.3 木構造話者クラスタリングを用いた話者適応

4.3.1 まえがき

本論文では話者特徴の同定を基本とした、極めて少ない適応音声サンプルによる話者適応法について述べる。少数サンプルによる適応を実現する方法として、話者モデルを木構造で表現し適応に用いる「木構造話者クラスタリング」を提案し、評価実験によりその有効性を示す。

話者適応法においては、手法や学習パラメータ数によって最適な適応サンプル数は異なる。例えば、学習パラメータ数が少ない適応手法は、一般に少ない適応サンプル数で話者適応の効果がみられるが、最終的な話者適応後の認識率はそれほど高いものは得られない。これに対し、学習パラメータが多い適応手法では、適応のための大量のサンプルを必要とするが、最終的な認識率は高い。従来の話者適応法では一般に、適応サンプル数によらず学習するパラメータの種類や数は不変であった。このため適応サンプル数によって話者適応の効果は、ばらつきがあると考えられる。以上の問題を解決する方法として、複数の話者適応手法の適応サンプル数に応じた自動的な切替え [48] が提案されている。つまり (a) 極少数の適応サンプルで、ある程度の認識率が得られる話者適応法と、(b) 大量の適応サンプルを必要とするが高い認識率の得られる話者適応法、を組合せ、適応サンプル数の多少に関わらず、それぞれの適応数に応じた話者適応を行なう枠組が重要であると思われる。

従来後者の手法は種々提案されているが（例えば [49] [50][8] など）、前者における有効な手法の検討が十分であるとは言えない。そこで本論文では前者を実現する手法として、話者または話者クラスタの選択に基づく話者適応手法を検討した。

このような話者適応では、話者（クラスタ）選択のみにより適応を行なうため、パラメータを調整する他の話者適応法と比較して、非常に少ないデータでの適応が可能である。複数の標準話者からの話者選択による高速話者適応法は、従来からいくつか提案されている。例えば杉山は、最小歪み原理に基づく最適な母音テンプレートの選択による教師なし話者適応法 [47] を提案している。また Niyogi と Zue は尤度による話者間の相関を利用したいくつかの標準話者選

択法を提案 [51] し、英語母音の識別実験により評価を行なっている。これらの研究はいずれも複数の標準話者のモデルから、特定の話者のモデルを選択し話者適応する方法の開発を目的としている。

このような話者選択による話者適応では、入力音声の話者の特徴が、あらかじめ用意された標準話者の特徴とは類似していない場合、性能が低下するという欠点がある。そこで解決策として、単純に標準話者の数を増加させるという方法も考えられる。しかし、話者の増加に伴い選択すべき話者クラス数が増加し現実的とは言えない。

クラス数の増大という問題に対して、クラスタリングによる対処法が考えられる。この例としては標準話者を男女の2クラスに分割し、男女識別の後に認識を行なうという方法が提案されている（例えば [52] [53] など)。しかしこの方法では、(1) クラスの分類がヒューリスティックな知識によっているし、(2) クラス数の決定もヒューリスティックに行なわれている。男女識別の例では、前者については男女というヒューリスティックに決定されたクラス分けがされており、後者については2クラスのみということになる。

クラス数の設定の問題を考えたとき、本来いくつのクラス数が最適であるか不明である。クラス数が異なると、ひとつのクラスが持つ分布の大きさが異なってくると考えられる。頑健性の点からデータ数が少ない場合はクラス数も少なく、データ数が多い場合はクラス数も多くすると良いと考えられる。以上のように、話者適応に最適なクラス数を決める場合は何らかの基準が必要である。

本研究では以上のクラス分類及びクラス数設定の問題を解決するため、階層的な話者クラスタリングによる話者適応法を提案する。この方法では話者特性を階層的に逐次分割することにより、話者モデルの木構造を作成する。木構造で話者特性を表らわすことにより、木構造の上層では話者特性の大局的な特徴、例えば男女の差などを表現するモデルが作成できる。また下層では局所的な特徴を表現するモデルを得ることが期待できる。

提案する方法では自動的にクラスタリングを行なうため、まず (1) のクラス設定の問題が解決できる。また何らかの方法で、木構造の階層のうち入力音声の識別に最適な階層を選択することができれば、(2) のクラス数に関連する、クラスの分布の大きさの決定に関する問題が解決できる。つまり話者クラスの分布の大きさの決定が自動的に行なわれることになる。そこでさらに適応用音

声に対するモデルの尤度により、最適な階層を選択する手法を提案する。

以上の2点が解決でき、更に話者数が増加した場合問題となる計算時間に関して、話者選択に要する計算量の増大を防ぐという効果も得られる。これは話者モデルを木構造で表現することにより、全話者のモデルと照合する場合に比べ、話者モデルの照合の回数が減るためである。

評価実験として、170人の話者の音素モデルセットをクラスタリングして木構造を作成する。ここでは音素モデルセットとしてHidden Markov Network(HMnet)を使用した[5]。この木構造を利用して話者適応し、適応後にSSS-LR[7]を利用した文節認識実験を行ない、本手法の有効性を検討する。

4.3.2 話者適応の原理

提案する話者適応法は話者のモデルの木構造により表現される階層的な話者モデルのクラスタリングに基づく方法である。木構造の上層に属するモデルは多数の話者特徴を包含し、下層に属するモデルは少数または特定話者の特徴を持つ。この木構造を上層から下層に辿り最適なモデルを選択することにより、話者適応が可能となる。入力音声の特徴が木構造を構成する標準話者の一人と似た特徴を持つ場合、下層のモデルが選択されることが期待される。またどの標準話者の特徴とも似ていない場合は、上層のモデルが選択されると予想される。上層のモデルが選択された場合、複数話者の特徴からの内挿の効果が得られると期待される。

木構造を作成するために、まず複数話者のデータからそれぞれの話者用の特定話者音素モデルセットが作成される。複数のモデルセットは、クラスタリングアルゴリズムによりクラスタ化される。生成された個々のクラスタはさらにクラスタリングされサブクラスタが作成される。一つのクラスタが1名の話者になるまでこれを繰り返し、木構造を作成する。木構造が作成された後、個々のクラスタに属する話者の音声データにより統計的音素モデルセットを作成する。統計的モデルを用いるため、最適モデルの選択の基準として、モデルの出力する尤度が利用できる。

4.3.3 アルゴリズム

アルゴリズムの概要

本節では、木構造話者クラスタリングによる話者適応アルゴリズムについて述べる。アルゴリズムの流れ図を図4.2に示し、以下に各節に先がけて概説する。図中にはそれぞれの処理が本論文中のどこに述べられているかも併記した。

本方式の処理の流れを以下に示す。

1. 初期 HMnet の作成

   出力確率分布が単一ガウス分布の初期 HMnet を、話者1名の大量のデータから SSS アルゴリズムを用いて生成し、HMnet のトポロジーを決定する。

2. 複数 HMnet の作成

   次に初期 HMnet に対し、話者適応法を利用して複数の学習話者の比較的少量のデータによりそれぞれ平均値及び分散の適応をおこない、複数話者分（ここでは 170 人分）の HMnet を作成する。

3. 木構造話者クラスタの作成

   本論文で提案する手法により木構造話者クラスタを作成し、木構造の各ノードに属する HMnet を作成する。

4. 話者適応

   木構造話者クラスタを利用し、話者または話者クラスタ選択による話者適応を行なう。

5. 認識

   選択された話者または話者クラスタに属する HMnet を用いて、話者混合法を用いた認識を行なう。

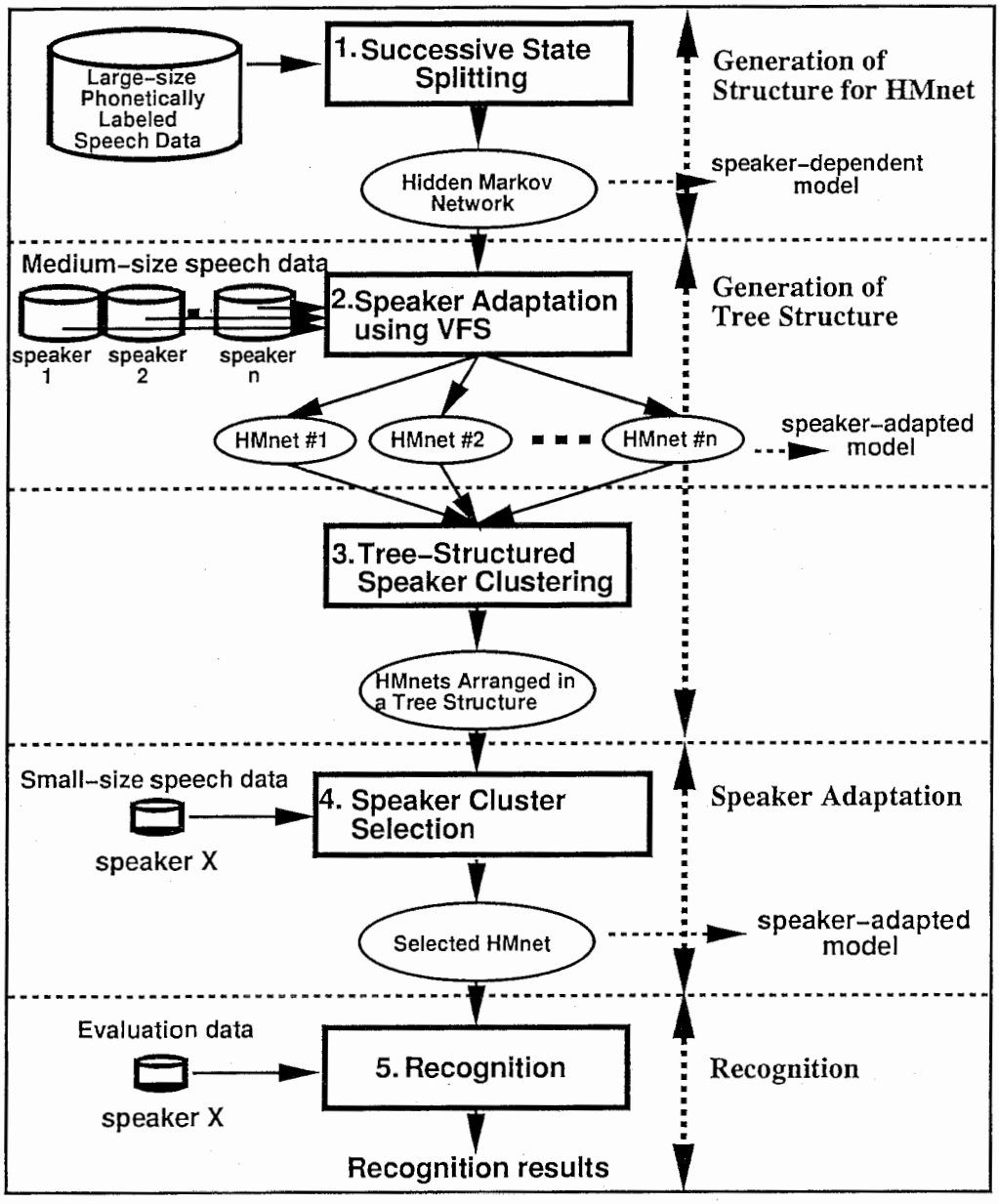

Figure 4.2: 木構造話者クラスタリングを用いた話者適応の流れ図.

4.3.4 特定話者用 HMnet の作成

実験では統計的音素モデルセットとして、隠れマルコフ網 (HMnet) を使用した。

使い方としては先ず、1名の話者のさまざまな音素コンテキストを含む学習データからSSSアルゴリズムを用いて初期HMnetを作成する。ここではHMnetの構造は話者間で共通と仮定する。話者間でHMnetの構造が本当に共通であるかどうかについては十分に検討すべき問題ではあるが、数種類の音素においての検討結果からは特に顕著なHMnetの話者依存性は見られなかったため[28]、このように仮定した。

次に構造は変更せずに出力分布のみ学習することにより、複数の話者について特定話者用HMnetを作成する。出力分布の学習にはBaum-Welchアルゴリズムに加え移動ベクトル場平滑化法(VFS)を併用した。VFSアルゴリズムでは内挿処理と平滑化処理を行なっている。このためBaum-Welch単独に比べ少ないデータでHMnetの学習が可能になる。

以上により構造は同一で出力分布のみ異なる、各話者のHMnetが得られる。

クラスタリングアルゴリズム

木構造を作成する場合、各話者から作成された特定話者用HMnet間のクラスタリングが必要となる。このクラスタリングは第3.4で提案した方法を用いた。この方法では、2のべき乗のクラスタを作成する一般的なLBGアルゴリズムとは異なり、歪みが最大となるクラスタを順次分割する。またHMnet間の距離は比較するHMnetが同一の構造であることを仮定し、確率的距離尺度であるBhattacharyya距離を使用した。

木構造話者クラスタの作成アルゴリズム

前節で述べたクラスタリング法を用いて、話者クラスタの木構造を作成する方法について述べる。ここで提案する木構造作成アルゴリズムでは、各ノードにおけるクラスタ数Kを与えるのみで、自動的にクラスタの作成を行なう。以下にアルゴリズムを示す。

STEP 1 初期化。 $j = 0, l = 0$

STEP 2 $0 \leq n \leq l$について、n番目のクラスタ（但しSTEP 5で終了したクラスタを除く）$S_n(j)$に属する話者をクラスタリングし、K個のサブ

クラスタを作成する。ここで $S_n(j)$ は階層 j における n 番目のクラスタを表す。

STEP 3 以上により求めたサブクラスタに属する話者のデータにより、HMnet を再学習し K 個の HMnet、$\{m_{nK}(j+1),\ldots,m_{nK+(K-1)}(j+1)\}$ を作成する。

STEP 4 $j \leftarrow j+1$. $l \leftarrow l+1$.

STEP 5 $0 \leq n \leq l$ について、$S_n(j)$ に属する話者の数が K 以下となった時、クラスタ n のクラスタリングを終了する。

STEP 6 STEP 2 へ戻る。

話者適応アルゴリズム

以上により作成された木構造話者クラスタを利用した、話者適応アルゴリズムについて述べる。入力音声に対する HMnet の尤度により話者モデルまたは話者クラスタモデルを選択することにより、話者適応を行なう。

話者適応の概念図を図 4.3に示す。HMnet の出力尤度に基づいて、クラスタを選択し、その選択されたクラスタに属するサブクラスタで同様の手続きによりクラスタ選択を行なうことで、木構造を探索する。木構造の探索中に各ノードに属するモデルに対する尤度を記憶しておき、探索の終了後最大尤度を与えるノードのモデルを用いて認識を行なう。

アルゴリズムを以下に示す。STEP 2 におけるモデルに対する尤度の計算法として Type 1・Type2 の 2 種類の実験を行なった。Type 1 はクラスタに属する話者（クラスタ）モデルのうち、最大尤度を与えるモデルの尤度により選択する方法である。Type 2 は以下の節で述べる話者混合法により混合分布モデルを作成し、その混合分布モデルの出力尤度により選択する方法である。Type 2 は認識と同じ評価基準による選択であるのに対し、Type 1 は評価基準は異なるが STEP 2 の結果をそのまま用いることができるので、計算量が少ないという特徴がある。

STEP 1 初期化。 $j=0, l=0$

STEP 2 クラスタ $S_l(j)$ において、$Kl \leq n \leq Kl+(K-1)$ を満たす n について、以下のいずれかの計算をおこなう。但し $A = \{a_0, \ldots, a_{I-1}\}$ は適応用データ、$L(m, a)$ は入力 a に対するモデル m の出力尤度を表す。

Type 1:

$$f^n(j) = \sum_{i=0}^{I-1} L(m_{nK+k}(j+1), a_i) \tag{4.1}$$

Type 2:

K 個の HMnet $\{m_{nK}(j+1), \ldots, m_{nK+(K-1)}(j+1)\}$ から、後述の話者混合法を用いて作成した K 混合の HMnet を $M^n(j+1)$ とする。

$$f^n(j) = \sum_{i=0}^{I-1} L(M^n(j+1), a_i) \tag{4.2}$$

STEP 3 $Kl \leq n \leq Kl+(K-1)$ において以下を計算。

$$l = \underset{n}{\operatorname{argmax}}\, f^n(j) \tag{4.3}$$

$$f(j) = f^l(j) \tag{4.4}$$

STEP 4 $j \leftarrow j+1$

STEP 5 $S_l(j)$ に対するサブクラスタが存在しない場合 STEP 6 へ。それ以外は STEP 2 に戻る。

STEP 6 $f(j)$ の値が最大となる階層、クラスタに属する HMnet を選択し認識に用いる。

このアルゴリズムによる木構造話者クラスタリングの実行例を図 4.4 に示す。この図では F で始まる話者名 (例えば F306) は女性話者を表し、M で始まる話者名は男性話者を表す。この例は 20 話者についてクラスタ数 $K = 4$ としてクラスタリングした。図から分かるように、男女で完全にクラスタが分割されている。また人数を変えて何通りかの実験を行なったが、いずれの場合も同様な傾向が見られた。

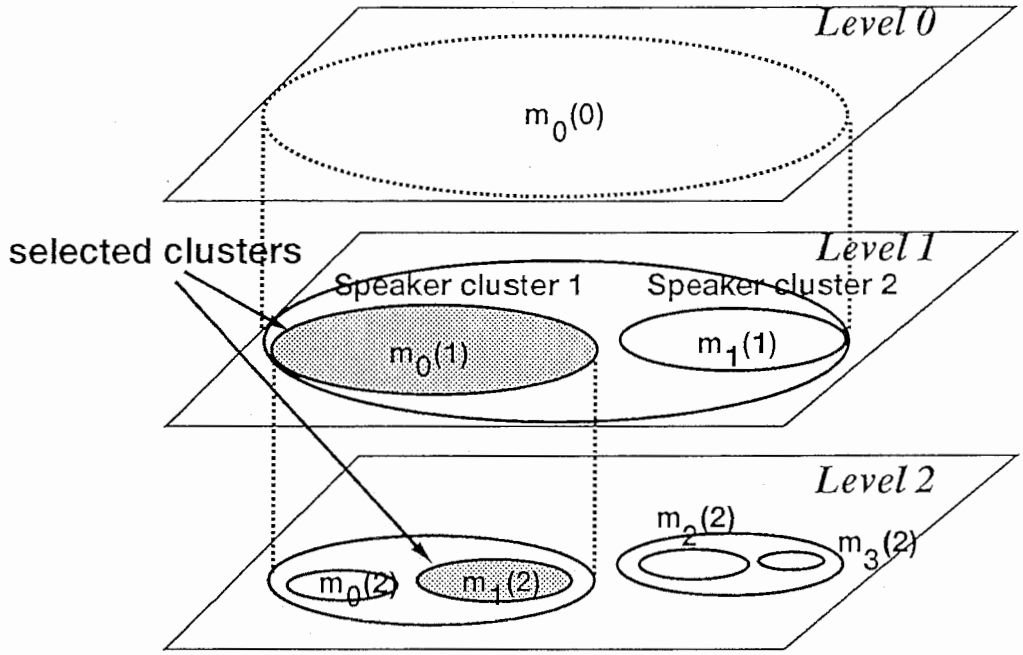

Figure 4.3: 木構造話者クラスタリングによる話者適応の概念図.

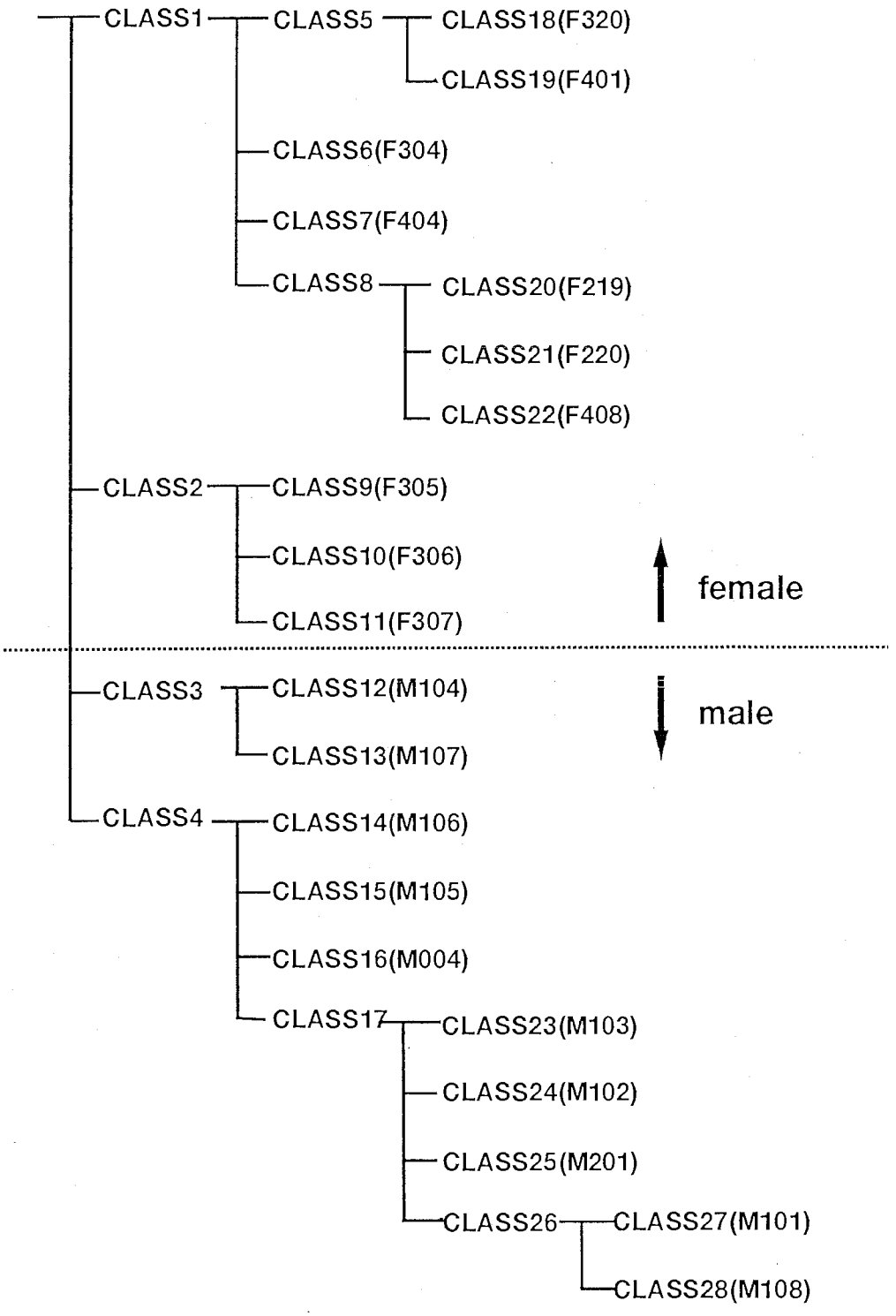

Figure 4.4: 木構造話者クラスタの例.

話者混合法による認識

木構造話者クラスタの各ノードにおける認識には話者混合法を用いる。話者混合法で木構造の各ノードに属する HMnet 簡便に合成して認識に用いることができる。木構造クラスタでは、前節のアルゴリズムにより各ノードには K 個またはそれ以下の HMnet が存在する。話者適応によりノードを選択した時点で、そのノードに属する HMnet を混合して認識に用いる。

4.3.5 認識実験

実験条件

本論文で提案した話者適応法を日本語音素認識実験及び、文節認識実験により評価を行なう。音響分析条件を付録 A.3に、実験条件を表 4.3に示す。

170 人の話者それぞれについて、単一ガウス分布、200 状態の HMnet を作成する。学習には日本語 50 文 (約 310 文節) を用いた。

以上のようにして作成された 170 個の HMnet を最大クラスタ数 $4(K = 4)$ とおき木構造話者クラスタリングした。本方式ではクラスタ数が認識時の話者混合数と等しくなる。このため認識時に使用する HMnet は最大 4 混合で 200 状態のものである。通常 HMnet による不特定話者音声認識では 10 混合以上を用い、状態数もさらに増やすと性能が向上するが、ここでは話者適応の効果を検討するのが主な目的であるので、簡単なモデルを使用した。

2 つの日本語文節発声データセットを話者適応と評価に用いた。音素認識実験では文節数約 256 の SB1 データセットを話者適応に、文節数 279 の SB3 データセットを評価に用いた。音素認識実験に用いる音素は文節発声データから視察により切り出したものである。

話者適応の文節認識実験による評価では特に、適応データの種類による影響を避けるため以下の方法で評価した。まず SB3 データセットからランダムに N 個の文節 $(N = 3, 5)$ を抜き出して適応に用い、適応に用いたデータを除いた $279 - N$ 個の文節で評価を行なう実験を 5 回繰り返し、その平均を認識結果とした。

また比較として特定話者文節認識実験も行なった。HMnet は 200 状態 5 混合を用いた。これは予備実験の結果、5 混合で性能の向上が飽和したためであ

る。

文節認識にはLR文法(1407規則,1035語)によるSSS-LR認識システムを用い、この時のビーム幅は1,200とした。タスクの音素パープレキシティーは5.9、単語パープレキシティーは約100である。

Table 4.3: 実験条件

木構造話者クラスタリング学習データ	
話者	男性85名 + 女性85名
学習データ	日本語50文 (学習にはVFSを使用)
特定話者文節認識実験データ	
話者	男性9名 + 女性9名
学習データ	1,310単語 + 256文節(SB1タスク)
認識データ	279文節(SB3タスク)
話者適応及び音素認識実験データ	
話者	男性9名 + 女性9名
適応データ	256文節中先頭からN個の文節を選択 (SB1タスク, $N = 1, 3, 5$)
認識データ	279文節(SB3タスク)中の音素
話者適応及び文節認識データ	
話者数	男性3名 + 女性3名
適応データ	279文節(SB3タスク)からランダムにN文節を選択。($N = 3, 5$)
認識データ	279文節(SB3タスク)から適応に用いたN文節を除いたデータ

認識実験結果

話者18名に対する音素認識実験の結果を表4.4に示す。"Type 1"及び"Type 2"は3.6節の話者適応アルゴリズムに示した2つの尤度計算の方法を表す。また「単一話者」は本論文で提案した自動的な話者クラスタの選択を用いずに、

木構造の最下層に属する1名の話者を尤度により選択した場合の結果である。“Type 1” 及び “Type 2” の2つの方法の比較では大差ない結果が得られた。また1名の話者の選択では不特定話者と同等かそれを下廻る結果しか得られず、単なる話者選択による方法は有効でないことが分かる。

適応文節数3の場合を例に、話者ごとに木構造の各階層における認識率を調べた。その結果、最下層、つまり1名の話者が選択された場合認識率が最大となる話者は18名中2名、また最上層、つまり不特定話者が選択された場合認識率が最大となる話者はなかった。大部分の話者では、ある程度の大きさにクラスタリングされたモデルを選択した場合、認識率が向上する。これは学習に用いた話者だけでは、全ての話者空間を埋めるにはスパースで、入力話者が、話者と話者の中間点に位置するような場合が多いのではないかと考えられる。

図4.5, 4.6, 4.7に個々の話者における適応文節数3の場合の音素認識実験結果を示す。

図4.5は、話者FMSに対する話者適応後の音素認識実験結果である。この図における横軸は木構造の階層を表す。階層0では、全ての話者の特徴を使ったモデルでの認識実験、つまり不特定話者音声認識に当たる。この場合は提案した話者適応法が不特定話者音声認識に比べ有効であることが分かる。また階層1では4つのHMnetのうち1種類を選択して認識を行なっている（最大クラスタ数$K = 4$とおいて実験しているため）。4つのHMnetのうち2つはほぼ男性音声からなるHMnet、2つはほぼ女性音声からなるHMnetである（170名の話者のうち、男女のクラスが混同している話者は2名のみ）。この階層1に比べ階層2で認識率が上回っているので、性別にクラス分けして作成したモデルよりも高い認識率が得られることが予想される。またこの場合は階層4が木構造における最下層になっている。つまり階層4に属するモデルは1名の話者から作成された特定話者モデルである。このことから話者FMSでは1名の話者からなるモデルよりも、ある程度クラスタ化され、話者空間上で特定話者モデルより広い空間を持つモデルが認識には有効であることが分かる。また同図に “Type 1” の方法で求めたHMnetの対数尤度出力も示した。尤度最大基準により階層を選択するため、この場合階層2を選択する。つまり認識率が最大の階層が選ばれることになる。このように尤度最大基準は階層選択のある程度の目安として使用できると考えられる。

次に図4.6に話者FKNについて示す。この話者の場合、階層5が木構造の最下層である。つまり階層5に特定話者モデルが属し、その場合が一番認識率が高くなっている。また図4.7は話者FAFの場合だが、ここでは階層1で認識率が最も高くなっている。階層1は先に述べたように、男女それぞれ2クラスからなり、性別判定した場合と類似した効果が得られる階層である。このように話者により認識率の高くなる階層は相当異なる。このことから本手法で提案した階層の自動選択が必要であることがわかる。

次に“Type 1”の尤度計算法を用いた話者適応の文節認識による評価を行なった。適応文節数が3及び5の場合について、結果を図4.8に示す。横軸が適応前、縦軸が適応後の認識率を示す。各点が話者を表す。認識対象話者は6名である。また比較のため行なった特定話者文節認識実験では、話者6名の認識率の平均は92.0%であった。話者によって適応の効果にはばらつきがあるが、いずれの話者も適応の効果が出ている。適応文節数が3及び5ではそれほど性能に差はなく、5文節程度で認識率は飽和する。1文節の平均発話継続時間長は約0.9秒である。このように極く少量の適応サンプルでも適応の効果が得られる。例えば一番効果のあった話者では5サンプルで認識率が74.3%から85.1%へ向上(42.0%の誤り率の減少)している。

Table 4.4: 話者18名における音素認識実験結果(%)

	不特定話者	適応用文節数		
		1	3	5
Type 1	71.1	73.1	74.5	74.4
Type 2		72.9	73.5	74.0
単一話者		69.9	71.1	69.5

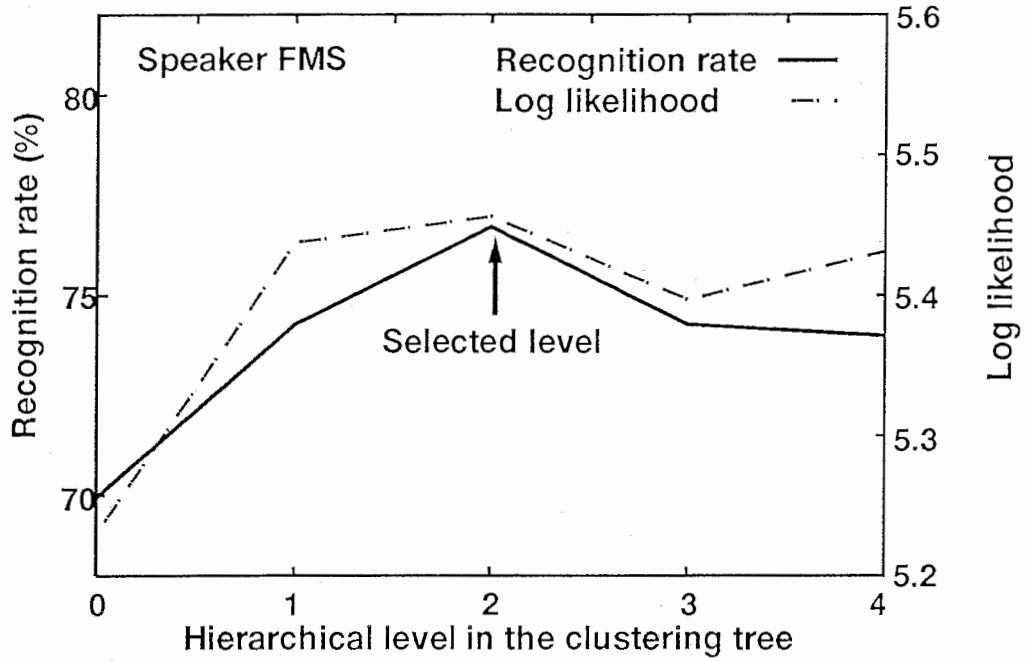

Figure 4.5: 木構造の各階層における音素認識率及び出力尤度 (話者 FMS, 適応文節数 3).

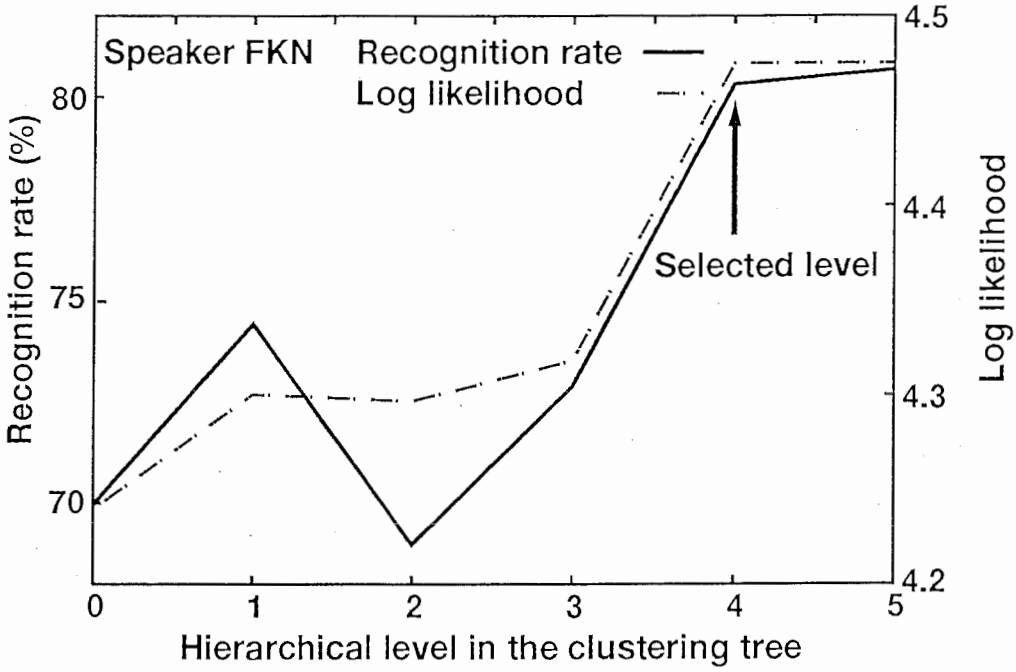

Figure 4.6: 木構造の各階層における音素認識率及び出力尤度 (話者 FKN, 適応文節数 3).

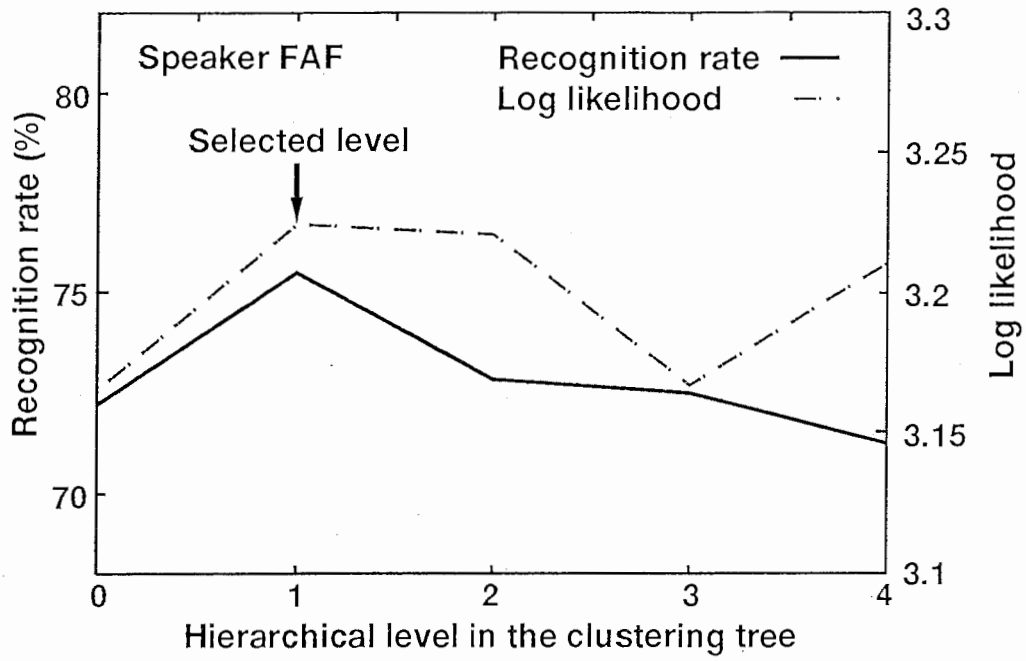

Figure 4.7: 木構造の各階層における音素認識率及び出力尤度 (話者 FAF, 適応文節数 3).

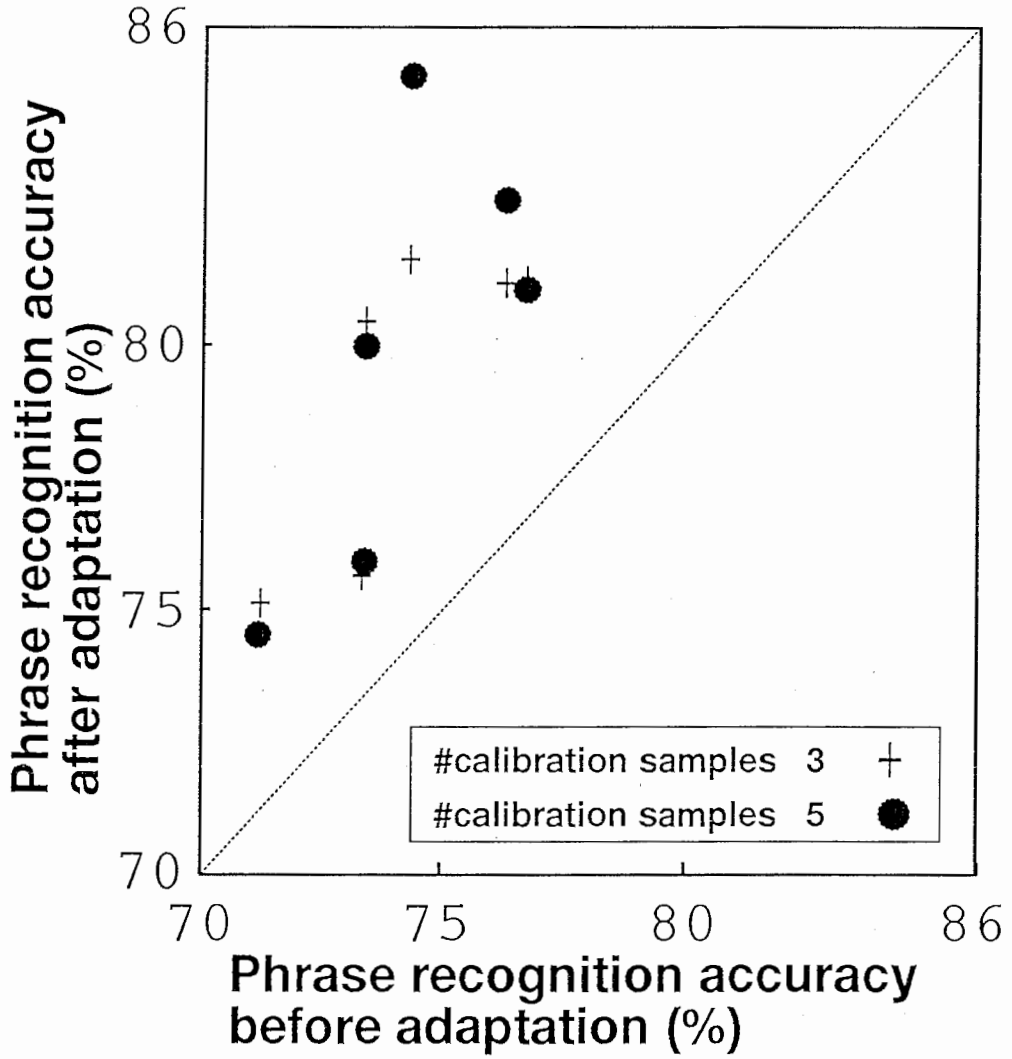

Figure 4.8: 話者 6 名における、話者適応前後の文節認識率の変化.

4.3.6 まとめ

本稿では木構造話者クラスタリングアルゴリズムと、そのアルゴリズムの話者適応への応用について述べた。有効性を検討するため音素及び文節認識実験による評価を行なった。その結果 3 〜 5 秒程度の極く少量の適応サンプルで話者適応の効果が得られ、本手法の有効性が確認できた。また木構造の各階層での認識率の検討により、話者クラスタサイズの自動選択が重要であること、ま

た選択の尺度としてHMnetの尤度が利用できることが分かった。さらに、単一話者を選択した場合との比較から、単一話者のモデルという話者空間上で狭い範囲をカバーするモデルよりも、話者クラスタのモデル、つまり話者空間上である程度の広い範囲をカバーするモデルを選択することが有効であることが明らかになった。

今後の検討課題としては、話者数の増加、教師なし話者適応への発展、話者クラスタ選択後の、平均値などモデルパラメータの適応などについて検討する予定である。特に話者数の増加については、単に標準話者を増加させるだけでなく、複数の話者モデルの内挿により、仮想的な話者を作成した場合の比較などについて検討したい。

第 5 章

データ量に応じた話者適応

5.1 まえがき

本論文では適応データ量に応じて自由パラメータを制御し、より少ないデータで話者適応する方法を目指している。本章ではこれを、(1) 複数の話者適応手法の適応サンプル数に応じた自動切替え、もしくは (2) 適応サンプル数に応じた学習パラメータ数の自動調整により実現する方法について述べる。

前者は異なる自由パラメータ数の話者適応をサンプル数により切替える方式である。つまり適応サンプル数が少ないうちは、自由パラメータ数の少ない話者適応法を用い、適応サンプル数が増加したら、自由パラメータ数の多い話者適応法を用いる。この切替えによる方法を第5.2節で取り上げる。ここでは HMM の尤度により、3 種類の自由パラメータが異なる適応手法を自動切替えする方法を提案する。

後者は自由パラメータ数は同じにおくが、適応の初期段階でいくつかのパラメータに対し「結び」の関係にするなど拘束を与え、その拘束を適応サンプル数に応じて除々に緩め実質的に自由パラメータ数を制御する方法である。これを実現する方法として MAP-VFS 法を提案し第 5.3節で述べる。最大事後確率推定法 (MAP 推定法) では、音素環境依存モデルのようなモデルパラメータ数の多いモデルで適応する場合データ量を多く必要とするという問題があった。これに対し補間・平滑化により少量データでも適応が可能とした方法が MAP-VFS である。さらに平滑化の度合をデータ数に応じて自動的に制御し、データ量に応じた適応が可能となる方法を提案する。

5.2 複数の話者適応法に基づく動的話者適応

5.2.1 まえがき

前章で我々は少量の学習サンプルで話者適応を行なう手法として、話者重み学習法を提案し有効性を示してきた。また比較的少ない学習サンプルで話者適応する手法としてベクトル場平滑化方式（VFS）について検討してきた。これらの話者適応手法はパラメータの自由度が異なり、適応に必要な学習サンプル数も異なる。話者重み学習では数秒程度の発声で話者適応が可能である。しかし数秒の発声による適応後は認識率の向上が飽和する傾向が見られる。これに対しVFSでは数十秒〜1分程度の適応用音声を必要とする。しかし得られる認識率は高い。このような違いは適応における自由パラメータの数の違いによって起こるものと思われる。たとえは話者重み学習では適応における被調整パラメータは混合重みのみである。しかもこれは全状態で「結び」の関係になっているため、混合数が被調整パラメータ数となる。例えば混合数が12のモデルでは、適応するパラメータ数はたかだか12である。これに対しVFSでは全ての分布の平均値を更新する。例えば200状態5混合で34次元ベクトルのHM-netの場合、200 X 5 X 34 のパラメータを調整することになる。

一般に適応法は以下に述べるトレード・オフが存在する。

- 自由パラメータ数が少ない適応手法では、少ないデータで適応が可能である。しかし適応後の認識の向上は少ない。
- 自由パラメータ数の多い適応手法では、適応後の認識率の向上は高い。しかし適応データを多く必要とする。

これらのトレード・オフを考慮すればわかるように、どのようなサンプル数でも最適な適応をするためには、サンプル数に応じて適応対象の自由パラメータ数を制御する必要がある。しかし従来の話者適応では一般にこの自由パラメータ数は固定されていた。

そこで本章ではサンプル数に応じた適応対象の自由パラメータ数を制御する適応手法を提案する。具体的には自由パラメータ数の異なる複数の話者適応方式を組み合わせ、手法の切替えを入力音声により自動的に行なうことにより、

入力語数に応じ動的に話者適応をおこなう手法を提案し、その有効性を認識実験により示す。

5.2.2 動的話者適応の原理

話者適応法においては、手法や学習パラメータ数によって最適な適応サンプル数は異なると考えられる。例えば、学習パラメータ数が少ない適応手法は、一般に少ない適応サンプル数で話者適応の効果がみられるが、最終的な話者適応後の認識率はそれほど高いものは得られない。これに対し、学習パラメータが多い適応手法では、大量の適応のためのサンプルを必要とするが、最終的な認識率は高い。従来の話者適応法では一般に、適応サンプル数によらず学習するパラメータの種類や数は不変であった。このため適応サンプル数によって話者適応の効果は、ばらつきがあると考えられる。以上の問題を解決する方法として、以下の2手法が考えられる [54]。

- 複数の話者適応手法の適応サンプル数に応じた自動切替え。
- 適応サンプル数に応じた学習パラメータ数の自動調整。

前者は異なる自由パラメータ数の話者適応をサンプル数により切替える方式である。つまり適応サンプル数が少ないうちは、自由パラメータ数の少ない話者適応法を用い、適応サンプル数が増加したら、自由パラメータ数の多い話者適応法を用いる。この切替えを何らかの基準により自動的に行なう。

後者は自由パラメータ数は同じにおくが、適応の初期段階でいくつかのパラメータに対し「結び」の関係にするなど拘束を与え、その拘束を適応サンプル数に応じて除々に緩め実質的に自由パラメータ数を制御する方法である。

本章ではこのうち、前者の手法に当たる話者適応手法の自動切替えのアルゴリズムについて提案する。本研究で自動切替えの対象となる話者適応手法は以下の3つである。

- 話者混合重み学習法 (STWT)
- 混合重み学習法 (SFWT)
- 移動ベクトル場平滑化方式 (VFS)

以上3種類の話者適応では学習対象の自由パラメータ数が異なり、大小関係は以下の通りとなる。

$$STWT < SFWT < VFS$$

よって大量の適応サンプルを用いると認識率もこの順番になる傾向がある。また話者適応の速さは逆順となり、概して $STWT$ が少数サンプルで話者適応の効果が表れる。これらを組み合わせることにより、適応サンプルに応じた話者適応が可能となる。

5.2.3 アルゴリズム

以上述べたように適応語数に応じて認識手法を切替えることができれば、効率よく適応することができると考えられる。切替えの手法としては、ヒューリスティックなやりかたとして、切替え時点の単語数をあらかじめ実験により定めておく方法も考えられるが、話者によって切替え時点は異なると考えられるので良い方法とはいえない。そこで提案する手法ではこの切替えを HMnet の対数尤度出力を用いて自動的に行なう。複数の話者適応手法を用いてそれぞれのモデルに対する尤度を計算し、尤度の高い手法を自動選択する。この場合適応に用いたサンプルに対する尤度では、サンプルにチューニングしているため、常に自由パラメータが多い手法が高い尤度を出力し、正確な判断が下せない。そこで認識時に各適応手法で得られた複数のモデルを用い、それぞれに対する尤度を求め、最大の値が得られるものを認識結果とする。以下にアルゴリズムを示す。

n 次元の入力音声系列を $\mathbf{X} = \{x_1, \ldots, x_K\}$ とする。話者適応法 i によって、入力音声の一部の系列 $\{x_1, \ldots, x_k\}(1 \leq k \leq K)$ を用いて適応したあとの HMnet を m_{ik} と示すことにする。以下の3段階により話者適応を行なう。

話者適応 入力音声 $\{x_1, \ldots, x_{k-1}\}$ を用いて初期 HMnet を複数の話者適応法で適応する。適応後 HMnet $\{m_{1k-1}, \ldots, m_{Ik-1}\}$ が得られる。I は適応手法の数である。ここでは初期 HMnet として話者混合法で得られる不特定話者モデルを用いた。

モデルの自動選択 適応後の HMnet のうちの 1 つが最大尤度基準により選択される。

$$M_{k-1} = \underset{i}{\operatorname{argmax}}\, L(x_k, m_{ik-1}) \tag{5.1}$$

ここで M_{k-1} は選択された HMnet、$L(X_k, m_{ik-1})$ は入力系列 x_k に対して HMnetm_{ik-1} から得られた出力尤度である。

認識 音声入力 x_k を選択した HMnetM_{k-1} で認識する。

5.2.4 被選択話者適応法の概説

本方式では、選択に用いる話者適応法として、話者混合重み学習法 (STWT)、混合重み学習法 (SFWT)、移動ベクトル場平滑化方式 (VFS) の 3 種類を用いる。以下に、話者適応前の音素モデルの作成法及び、この 3 種類の話者適応法について概説する。

不特定話者認識用音素モデル

話者適応前の不特定話者音素モデルとして、混合連続分布 HMnet を用いた。モデルの作成手順を以下に示す。まずある標準話者に関して、SSS アルゴリズムにより HMnet のトポロジーを決定し、かつ標準のモデルを学習して作成する。さらに比較的少量の多数話者の発声データから VFS を利用して話者個別の HMnet を作成する。この複数の HMnet が得られた後、話者混合法により混合連続分布 HMnet を作成する。このようにして得られた話者混合モデルでは、話者重み学習が可能となる。

話者混合重み学習法

話者混合重み学習法 (Speaker-Tied Mixture Weight Training: STWT) [30] では、話者混合法により得られた混合連続分布 HMnet の話者重みのみを、連結学習により再学習する。話者重みは全状態を通じて共通だから、パラメータ数が少なく学習が速い。

混合重み学習法

混合重み学習法 (Speaker-Free Mixture Weight Training: SFWT) [45] では、混合分布の重みを全状態について連結学習により再学習する。「話者結び」にはしないため、学習対象パラメータ数は話者重み学習に比べ多い。

移動ベクトル場平滑化方式

移動ベクトル場平滑化方式 (VFS) [10] ではまず、連結学習により標準話者と未知話者間のモデルの平均ベクトルの差分ベクトルを求め、これを移動ベクトルとする。データ数が足りなく学習されなかった平均ベクトルの移動ベクトルは内挿により推定し求める。以上により得られたモデルは十分な適応語数が得られていない場合に推定誤差を含んでいる。そこでさらに移動ベクトルに連続性の拘束条件を入れ、平滑化を行なうことにより推定誤差の吸収を行なう。以上によりモデルの平均ベクトルの適応を行なう。詳細は第2章を参照のこと。

5.2.5 認識実験

以上の手法の有効性を確認するために、認識実験を行なった。実験条件を表5.1に示す。また評価システムの構成図を図5.1に示す。HMnet の学習は孤立単語で行ない、学習と異なる話者により話者適応と文節認識実験を行なった。話者適応と認識とでは異なる文節セットを用いた。適応用には ATR の SB1 データセットを、認識用には ATR の SB3 データセットを用いた。いずれもタスクは国際会議の予約に関するもので文節発声である。HMnet の状態数は 200、混合数は 12 のものを用いた。また文節認識には LR パーザーを用いた。LR のビーム幅は 256 で、一般文法 (1407rules,1035 語) を使用した。実験では動的話者適応のほか、比較として $STWT$、$SFWT$、VFS を単独で用いた場合の結果も求めた。

図5.2に各話者適応手法における文節認識実験結果を示す。話者 MSH において $STWT$、$SFWT$、VFS を単独で用いた場合の結果を見ると学習パラメータの自由度の少ない $STWT$ が、少ない適応語数で話者適応の効果が表れる。これに対し VFS は適応語数が他の手法に比べ多く要するが、語数が多くなると認識率は一番高い。

以上に対し動的話者適応では、ほぼいずれの語数であっても、3手法に比べ高い認識率が得られる。最良の認識率を与える手法より高い認識率が得られるのは、適応サンプルが入力されるごとに選択を行なうためである。例えば入力サンプル数が50の場合は、3手法のうちではVFSが一番高い認識率を与えるが、VFSで誤認識となるものを他の2手法で補う傾向が見られる。ヒューリスティックにあらかじめ定められた語数で切替える方法では、認識率は3手法のうち最良のものを越えることはない。

話者MTMはウエート学習では効果がなく、VFSによる話者適応が有効な話者である。この場合最良の認識率を与えるVFSを常には選択しなかったためVFSの認識率は越えなかったが、VFSに近い認識性能が得られた。

図5.3は各適応語数に対して、どの手法を選択したかの割合を表した図である。この図から適応語数に応じて手法が徐々に切替えられていることが確認できる。

Table 5.1: 実験条件

学習データ		
構造決定用	男性1名	5240単語
パラメータ学習用	男性12名	216単語
話者適応 / 認識データ		
話者	男性2名	
適応データ	256文節(SB1タスク)中から選択	
認識データ	279文節(SB3タスク)	

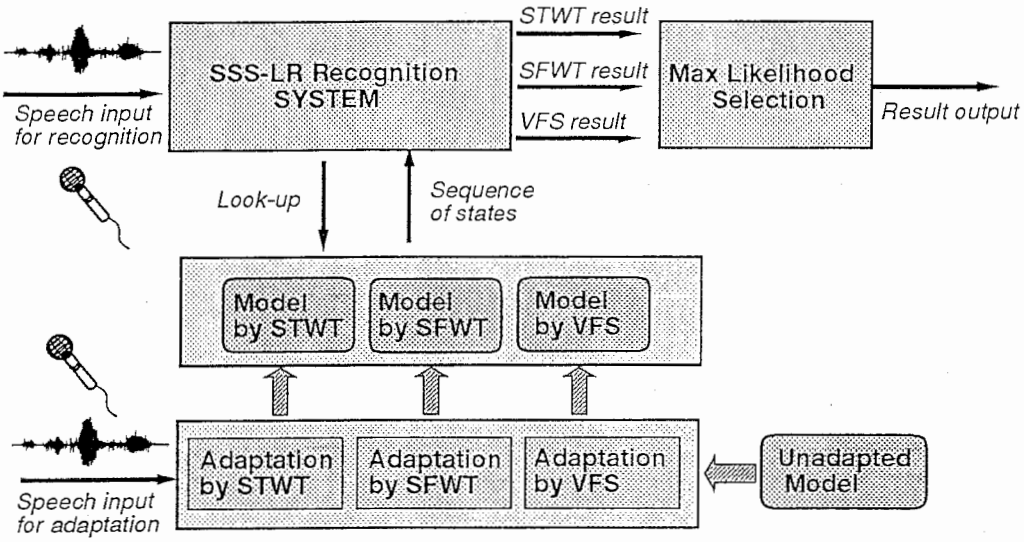

Figure 5.1: 評価システムの構成図

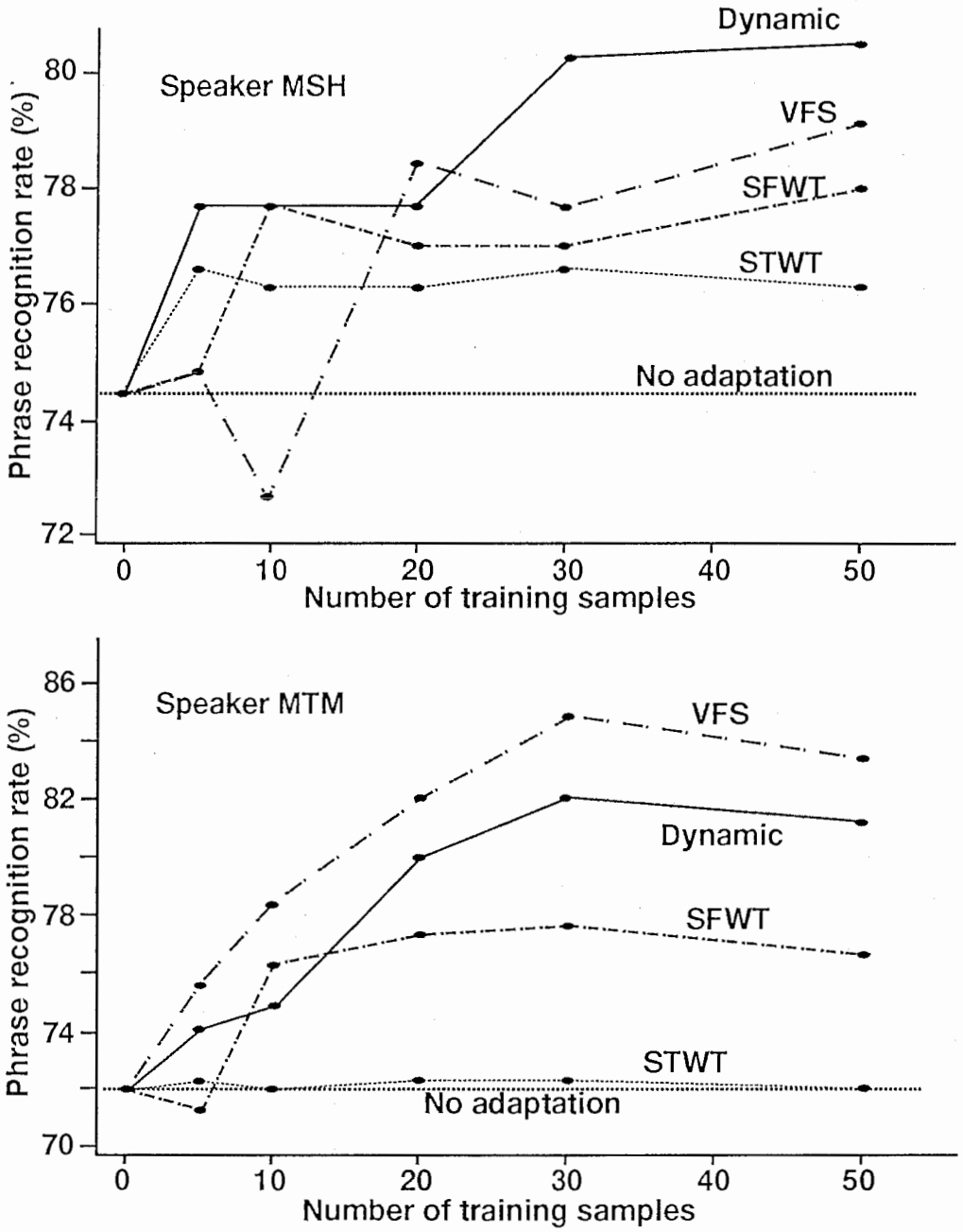

Figure 5.2: 話者適応手法の性能比較

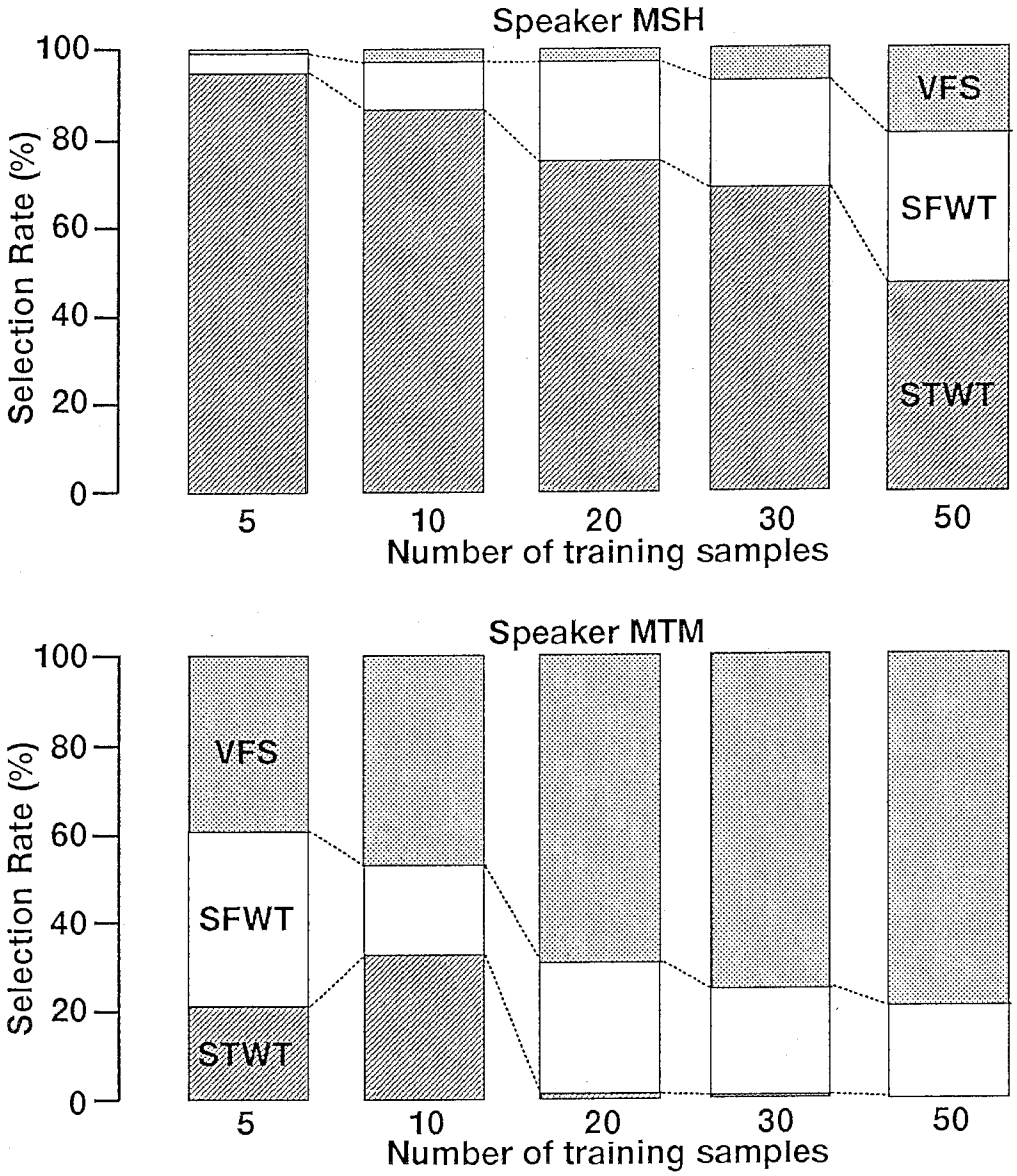

Figure 5.3: 各話者適応法選択の割合

5.2.6 まとめ

本報告では複数の話者適応方式を組み合わせ、手法の切替えを入力音声により自動的に行なうことにより、入力語数に応じ動的に話者適応をおこなう手法を提案し、その有効性を認識実験により示した。今後は適応語数に応じた話者

適応パラメータ数の自動決定について検討する予定である。

5.3 MAP-VFS 話者適応法

5.3.1 まえがき

話者適応を行なう場合には一般に適応データ量は未知の場合が多く、適応データ量によらず常に最適な適応を行なえる手法が望まれる。データ量に応じた適応手法としては最大事後確率推定法 (MAP 推定法)[3] が提案されている。本節では MAP 推定における少量の適応データでの性能を向上させるためにパラメータの補間、平滑化に基づく適応手法である移動ベクトル場平滑化法 (VFS)[8] を組み合わせた MAP-VFS 話者適応法 [68] を提案する。

しかし、MAP-VFS では VFS と同様に適応データ量が増えた場合に平滑化による制約がかえって適応性能の向上を阻害するという問題がある。そこで、さらに適応データ量に応じて平滑化の強度を制御する MAP-VFS 話者適応法を提案し、適応データ量によらず有効に適応できることを示す [60]。

5.3.2 MAP 推定における問題点

MAP 推定では最尤推定法 (ML 推定法) と比較して、事前確率分布による事前知識を使うところが異なる。このような事前知識を用いた話者適応の問題点としては

- 適応話者のモデル推定に有効な事前知識、即ち初期モデルを得ること
- 事前知識と適応データの情報をバランスよく利用したモデル推定
- 適応データの偏りにより発生する未学習パラメータの補間

があげられる。

初期モデルに関しては多数の話者の情報を含んだモデルである不特定話者モデルや適応話者との類似性をもとに選択した標準話者のモデルが一般的に使われている。本研究では適応話者の特徴に関する事前知識を得るために木構造話者クラスタリング法 [55] を用い多数の特定話者モデルから話者の類似性を基に選択した複数の特定話者モデルを合成することにより初期モデルを作成した。

MAP 推定は、既に話者適応に応用した研究がいくつか報告されている [3][56][57][58]。しかしながら、この手法では適応データが存在しないパラメータは学習されないため、これらの未学習パラメータを補間することが必要となる。

少量の適応データによるモデルの推定誤差と未学習パラメータの補間の問題を扱った話者適応手法としては、移動ベクトル場平滑化方式（VFS）が提案されている [8]。VFS では連続分布型 HMM のガウス分布の平均ベクトルの適応を効果的に行なうことができる。VFS のアルゴリズムでは移動ベクトルの補間・平滑化を行なっている。しかし、一般に適応データ量が少量である場合には個々の移動ベクトルに対する適応データ量のばらつきは大きくなる。推定に用いられるデータ量にばらつきが大きいと当然推定された移動ベクトルの信頼性のばらつきも大きくなる。従って、効果的な補間、平滑化を実現するためには各処理において各移動ベクトルの信頼性を考慮することが必要となる。

本研究では MAP 推定法及び、移動ベクトル場平滑化話者適応方式の個々の問題点を克服するとともに、先に述べた事前知識を用いた話者適応の問題点に総合的に対処するためにこれらの二つの手法を統合したアルゴリズム (MAP-VFS アルゴリズム) を提案する。

5.3.3 MAP-VFS アルゴリズム

本論文で提案する MAP-VFS アルゴリズムは 1) MAP 推定法による移動ベクトルの推定、及びその移動ベクトルを用いた 2) 移動ベクトルの補間、3) 移動ベクトルの平滑化、の 3 ステップより構成される MAP 推定法と VFS の統合されたアルゴリズムである。

MAP 推定法による移動ベクトルの推定

初期モデルのガウス分布の平均ベクトルを適応データを用いて連結学習により再学習する。c_p^I と c_p^R をそれぞれ初期モデルと再学習後のモデルの p 番めのガウス分布の平均ベクトルとすると、それに対応する移動ベクトル v_p は c_p^I と c_p^R の差分であり、

$$v_p = c_p^R - c_p^I \tag{6}$$

となる ($p \in K_1$、K_1 は各ガウス分布のうち適応データの存在したものの集合)。

この適応学習において、初期モデルの各平均ベクトル c_p^I を事前分布の平均値として用いると各平均ベクトルの MAP 推定値 c_p^R は式 (5) より以下のようになる。

$$c_p^R = \frac{n_p}{n_p + \tau} m_p + \frac{\tau}{n_p + \tau} c_p^I \tag{7}$$

ここで、n_p, m_p はそれぞれ p 番めのガウス分布に対して観測される学習データの総数及び、その平均値である。また、τ に関しては全てのガウス分布に対して同じ値を用いた。

式 (7) で得られた MAP 推定値を式 (6) に代入することにより移動ベクトル v_p は式 (8) のようになる。

$$v_p = c_p^R - c_p^I = \frac{n_p}{n_p + \tau}(m_p - c_p^I) \tag{8}$$

従って、従来の最尤推定法により求めた移動ベクトル v_p^{ML} $(= m_p - c_p^I)$ を用いて MAP 推定法による移動ベクトルを表すと以下のようになる。

$$v_p = \frac{n_p}{n_p + \tau} v_p^{ML} \tag{9}$$

式 (9) より、MAP 推定法により求めた移動ベクトル v_p が最尤推定法による移動ベクトル v_p^{ML} を適応データ量 n_p の関数で重み付けしたものであることがわかる。以下の移動ベクトルの補間、平滑化の処理に式 (9) で表される MAP 推定法により求めた移動ベクトルを用いることにより移動ベクトルの信頼性を考慮した補間、平滑化を行なうことができる。

移動ベクトルの補間

少量の適応データによって適応学習を行なうため全ての平均ベクトルが学習できるわけではない。このような未学習のガウス分布の平均ベクトルを c_q^I ($q \in K_2$、K_2 は各ガウス分布のうち適応データの存在しなかったものの集合) で表すと、c_q^I の移動ベクトル v_q は c_q^I の近傍にある適応データが存在し、適応学習が行なわれた平均ベクトル c_p^I の移動ベクトル v_p を用いて次式のように補間さ

れる。

$$v_q = \sum_{k \in N(q)} \lambda_{q,k} v_k \; / \sum_{k \in N(q)} \lambda_{q,k} \tag{10}$$

ここで、$N(q)$ は c_q^I の k- 近傍にある平均ベクトル群を表しており、$\lambda_{q,k}$ は c_q^I と c_k^I の距離によって決まる重み係数である。

そして、この補間された移動ベクトル v_q によって c_q^I は c_q^R に変換される。

$$c_q^R = c_q^I + v_q \tag{11}$$

移動ベクトルの平滑化

全ての学習された移動ベクトル $v_p (p \in K1)$ に対して式 (12) により平滑化を行なう。この平滑化の操作は全ての移動ベクトルが連続性の拘束条件で結びつけられるという仮定にもとづいている。即ち、ある話者の音響的特徴空間は滑らかな移動ベクトル場に沿って別の話者に連続的に変換可能であると仮定している。

$$v_p^S = \sum_{k \in N(p)} \lambda_{p,k} v_k \; / \sum_{k \in N(p)} \lambda_{p,k} \tag{12}$$

従って、ガウス分布の平均値は平滑化された移動ベクトルを用いて次式によって修正される。

$$c_p^S = c_p^I + v_p^S \tag{13}$$

本報告では、補間、平滑化の両操作において k- 近傍数として 6 を使用した。重み係数 $\lambda_{a,b}$ は式 (14) により計算した。

$$\lambda_{a,b} = \exp\left(\frac{-d_{a,b}}{f}\right) \tag{14}$$

$d_{a,b}$ は c_a^I と c_b^I の距離であり、f は重みの制御係数である。

5.3.4 MAP-VFS 話者適応の評価実験

評価は日本語の 26 音素による音素認識実験により行なった。適応／認識データの実験条件の詳細を表 5.2に示す。また音響分析条件を付録 A.3に示す。実験には 200 状態の隠れマルコフ網 (HMnet) を使用した。HMnet の各状態のガウス分布の混合数は不特定話者モデルの場合は 5 混合、話者クラスタモデルの場合は最大 4 混合である。また、共分散行列には対角共分散行列を用いた。

実験では適応データの選択方法が話者適応性能に影響を及ぼす可能性を考慮し各適応文節数 N に対して選択文節を変えた評価をそれぞれ5回繰り返して平均の音素認識率を求めた。

Table 5.2: 実験条件

適応／認識データ	
話者	男性4名 (MAU,MMY,MSH,MTM) 女性3名 (FAF,FMS,FYM)
適応データ	256文節 (SB1タスク) からランダムに 選んだ N 個の文節 ($N = 1,3,5,7$)
認識データ	279文節 (SB3タスク)

285人の特定話者モデルを合成することにより作成した不特定話者モデルに対する話者適応実験を行なった [30]。

図5.4は、次の4つの手法 1) 最尤推定法による適応データの存在するパラメータのみの再学習 (ML)、2)MAP 推定法による再学習 (MAP)、3) 最尤推定法による移動ベクトル場平滑化方式 (VFS)、及び本報告で提案した 4)MAP 推定法による移動ベクトル場平滑化方式 (MAP-VFS)、による話者適応結果である。MAP-VFS による話者適応が最も高い認識率を示しており、また MAP 及び、VFS もそれぞれ ML に比べてよい性能を示している。この結果より、MAP-VFS 法が MAP 及び VFS を効果的に統合していることがわかる。さらに表5.3に全ての評価話者に対する MAP-VFS 法による話者適応結果を示す。本表により、MAP-VFS が全話者に対して適応文節数の増加に応じて性能が向上する安定した話者適応を実現していることがわかる。7文節による話者適応結果を見ると MAP-VFS 法により音素認識率は不特定話者モデルより 1.0 ～ 6.5%(全話者平均 2.9%) 向上している。

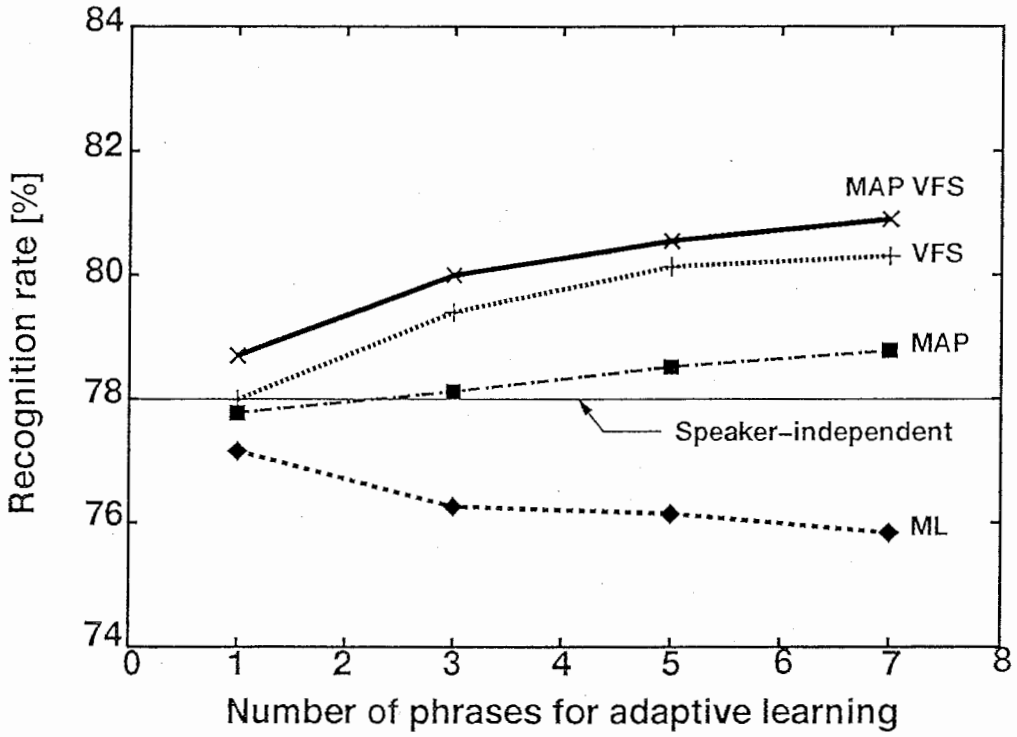

Figure 5.4: 話者適応手法による音素認識率の比較 (男性話者 4 名、女性話者 3 名の平均値)

Table 5.3: 不特定話者モデルに対する MAP-VFS 法による話者適応結果－音素認識率 (%)

話者名	適応文節数				
	適応前	1	3	5	7
MAU	81.8	81.7	82.5	83.3	83.6
MMY	81.0	80.8	81.6	82.0	82.0
MSH	75.9	76.3	77.9	78.4	79.0
MTM	82.5	83.4	84.1	84.4	85.0
FAF	78.5	79.6	81.1	82.1	81.9
FMS	78.9	79.5	80.4	80.7	81.0
FYM	67.4	69.6	72.4	72.9	73.9
平均	78.0	78.7	80.0	80.5	80.9

5.3.5 木構造話者クラスタモデルの適応実験

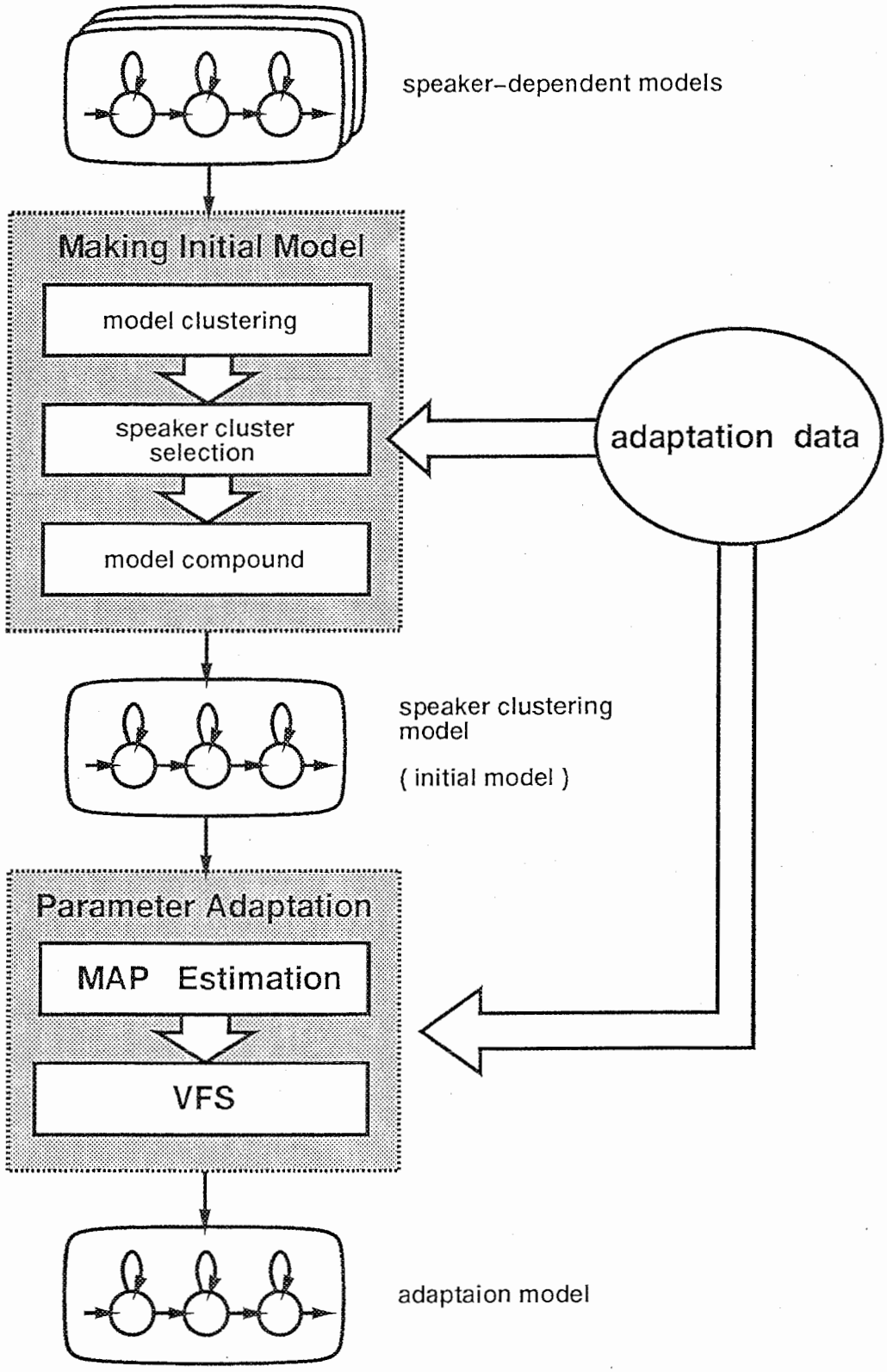

Figure 5.5: 話者クラスタモデルに対する話者適応

図 5.5に初期モデルの作成と適応の処理の流れを示す。まず、木構造話者クラスタリングアルゴリズムを用いて 170 人の特定話者のモデルをその類似性に基づいてクラスタリングし、さらに適応データを用いて適応話者の特徴と類似したクラスタを選択している。このように適応話者との類似性を基に選ばれたクラスタに含まれる話者モデルを合成して初期モデルを作成することにより適応話者の特徴に関する情報を多く含んだ事前知識を得ることができる。ここでは、上記の方法により作成した初期モデルに対して MAP-VFS 法により話者適応を行ないその効果を評価した。表 5.4は全ての評価話者の話者クラスタモデルに対する話者適応結果である。また、図 5.6には全話者の平均値をグラフで示す。7 文節での適応結果を見ると音素認識率は不特定話者モデルより 3.9 ～ 10.9%(全話者平均 6.2%) 向上している。これには話者クラスタの選択によって得られた事前知識による効果とその事前知識を利用した MAP-VFS 法による話者適応の効果の両方が含まれている。この話者クラスタによる効果と MAP-VFS 法による効果を分けて見ると全話者平均で前者が 3.7%、後者が 2.5% である。これを前項の実験により得られた不特定話者モデルから直接 MAP-VFS 法により適応した場合の効果 2.9% と比較することにより話者クラスタと MAP-VFS 法の効果が僅かな損失で有効に加算されていることがわかる。この結果は話者クラスタ法により得られた適応話者に類似した初期モデルが MAP-VFS 法と効果的に結びつけられて、高い話者適応性能を達成していることを示している。

Table 5.4: 話者クラスタモデルに対するMAP-VFS法による話者適応結果－音素認識率(%) (下段の括弧内は木構造話者クラスタリング法により作成された初期モデルでの認識率、適応前のモデルは170人分の特定話者モデルを合成して作成したモデルで表2の適応前モデルとは異なる)

話者名	適応文節数				
	適応前	1	3	5	7
MAU	81.1	83.9 (83.2)	85.0 (83.8)	86.1 (83.8)	85.9 (83.6)
MMY	77.4	80.5 (79.7)	80.7 (80.3)	81.5 (80.2)	81.3 (78.6)
MSH	74.7	77.7 (76.9)	77.2 (75.4)	79.5 (76.4)	79.0 (75.8)
MTM	80.8	84.2 (83.4)	86.4 (85.2)	87.3 (85.2)	86.9 (85.4)
FAF	76.1	78.4 (78.5)	81.3 (79.5)	82.4 (80.5)	81.8 (79.6)
FMS	74.5	80.4 (80.8)	81.4 (80.5)	82.0 (80.5)	82.1 (80.5)
FYM	66.7	71.5 (72.1)	74.9 (73.3)	77.1 (74.1)	77.6 (73.7)
平均	75.9	79.5 (79.2)	81.0 (79.7)	82.3 (80.1)	82.1 (79.6)

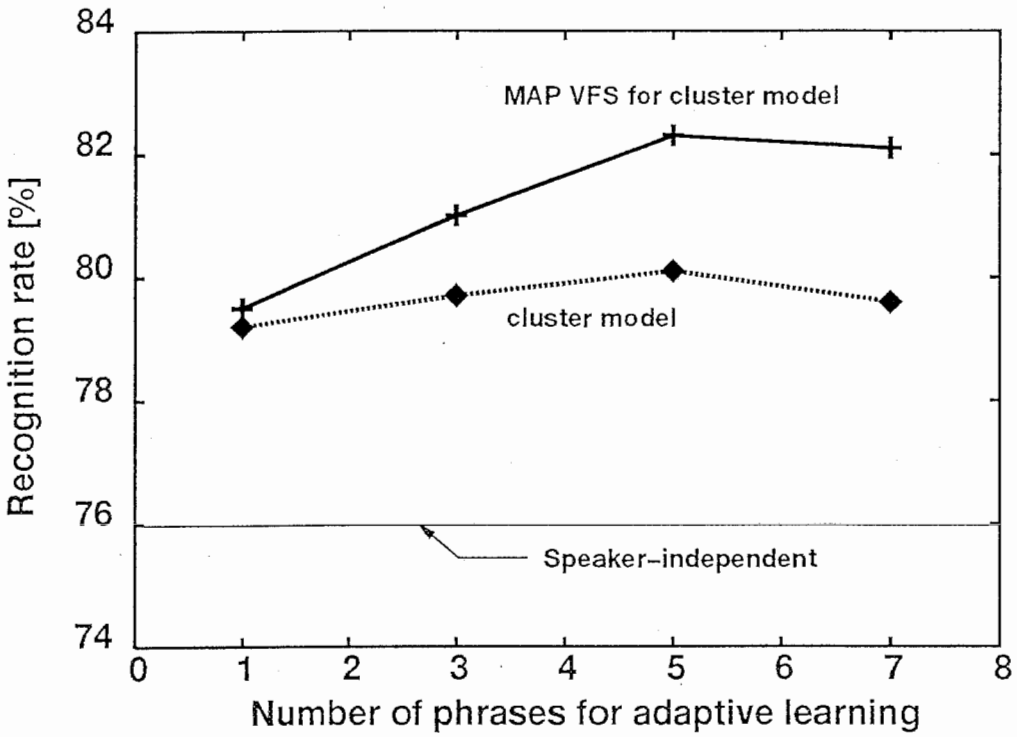

Figure 5.6: 話者クラスタモデルに対する話者適応結果 (男性話者 4 名、女性話者 3 名の平均値)

5.3.6 平滑化係数制御による MAP-VFS

MAP-VFS では適応データ量が増えた場合に平滑化による制約がかえって適応性能の向上を阻害するという問題がある。そこで、さらに適応データ量に応じて平滑化の強度を制御する MAP-VFS 話者適応法を提案し、適応データ量によらず有効に適応できることを示す。

VFS ではデータ量が少ない場合における移動ベクトルの信頼性低下を防ぐために、近傍の移動ベクトルの情報を使って、以下のように平滑化を行ない推定誤差の吸収を行なっている。

$$v_p^S = \sum_{k \in N(p)} \lambda_{p,k} v_k \; / \; \sum_{k \in N(p)} \lambda_{p,k}$$

$$\lambda_{p,k} = \exp\left(\frac{-d_{p,k}}{f}\right)$$

ここで$\lambda_{p,k}$は近傍の移動ベクトルの情報を使う場合の重みを決める係数であり、平均ベクトルμ_p^I, μ_k^I間の距離$d_{p,k}$と平滑化の強度を決める係数fに依存して決まる。このようにして平滑化された移動ベクトルv_p^Sを用いて話者適応後のガウス分布の平均ベクトルμ_p^Sが次式により求められる。

$$\mu_p^S = \mu_p^I + v_p^S$$

一般に平滑化される移動ベクトルv_pの推定に用いた学習データ量が少ない時は、その推定誤差が大きいと考えられるため平滑化によりその近傍の移動ベクトルの情報を使用することによりその推定誤差の吸収が効果的になされる。しかし、適応データ量が十分にある場合は推定誤差は小さくなり平滑化の必要がなくなるばかりか逆に平滑化により各パラメータの情報表現能力の低下をまねくことになる。実際にこの問題により、適応データが増加した場合には平滑化を行なった方が平滑化を行なわない場合より適応性能が劣化することが知られている。このような絶対的な適応データ量に関する問題を解決し、適応データ量によらず常に高い適応性能を得るために、本節では適応データ量をもとに平滑化の強度を制御することを提案しその効果を報告する。

平滑化係数の制御

平滑化係数の制御においては適応データの内容によって各パラメータに対する適応データ量に偏りがあることを考慮し、また状態数、混合数等のモデルの構造に依存しない基準で制御を行なうために各パラメータ（ガウス分布の平均値）ごとに独立に行なった。以下に本論文で用いたp番めのガウス分布に対する平滑化係数f_pの制御式を示す。

$$f_p = f\frac{\alpha}{n_p + \alpha}$$

ここで、fは全てのパラメータに対して共通に与えられる平滑化係数の初期値であり、n_pはp番めのガウス分布の適応データ量を表している。この式により各パラメータに対する平滑化の強さは適応データ量の増加に従って弱められ

ていき $n_p \to \infty$ では平滑化を行なわない場合と同様の状態に収束することがわかる。また、この時の収束の速さは係数 α によって決められるが本稿では α は実験的に求めた値を使用した。

5.3.7 平滑化係数制御による MAP-VFS の評価実験

実験には 200 状態の隠れマルコフ網 (HMnet) を使用した。初期モデルには不特定話者モデル（285 人分の特定話者モデルから合成することにより作成）[71] を用い、HMnet の各状態のガウス分布の混合数は 5 混合とした。平滑化係数の初期値は $f = 30$ とし、収束の速さを決める係数 α は実験的に求め $\alpha = 0.3$ とした。分析条件を付録 A.3に示す。適応／認識データを表 5.5に示す。

表 5.6に MAP、MAP-VFS、及び平滑化係数制御を行なった MAP-VFS による話者適応結果を示す。またこの表における平均値を図に示した (図 5.7)。まず、MAP と MAP-VFS を比較すると適応文節数が 3,7 文節のように少ない場合はほとんどの話者で MAP より MAP-VFS の方が認識誤り率は低く VFS が有効であることがわかる。しかし、適応文節数が 20 文節まで増えると MAP と MAP-VFS の効果が 4 名の話者で逆転しており 47 文節以上では全ての話者において MAP-VFS より MAP の方が認識誤り率が低くなっている。この結果より平滑化は適応データ量が少ない場合には有効であるが適応データ量が多い場合には逆に話者適応性能を劣化させていることが確認できる。次に、上記の MAP 及び MAP-VFS による話者適応結果を本稿で提案している適応データ量に応じた平滑化係数制御を行なった MAP-VFS による話者適応結果と比較する。適応文節数が 20 文節以下の場合には多くの場合で平滑化係数制御を行なった MAP-VFS が通常の MAP-VFS と同程度、或はより高い話者適応性能を示しており、また同時にこの範囲では常に MAP より高い認識性能を示している。一方、47 文節以上では平滑化係数を制御することによって通常の MAP-VFS より高い認識性能が得られ、また MAP に近い性能が得られている。以上の結果より MAP-VFS において平滑化係数を適応データ量に応じて制御することが効果的な平滑化を行なう上で有効であることがわかる。

Table 5.5: 実験条件

学習データ	
男性 146 名 + 女性 139 名 (各話者 50 文章)	
適応／認識データ	
話者	男性 4 名 (MAU,MMY,MSH,MTM) 女性 3 名 (FAF,FMS,FYM)
適応データ	256 文節 (SB1 タスク) の先頭から順に 取り出した N 個の文節
認識データ	279 文節 (SB3 タスク)

Table 5.6: 話者適応結果－音素認識誤り率 (%)

上段：MAP、中段：MAP-VFS

下段：MAP-VFS (平滑化係数制御)

話者名	適応文節数					
	適応前	3	7	20	47	256
MAU	19.2	19.2	16.1	16.4	11.9	6.9
		18.1	16.0	15.7	14.3	11.0
		18.1	15.2	15.7	12.3	6.9
MMY	19.0	19.4	18.0	16.2	13.7	9.1
		18.7	17.5	16.9	15.4	13.1
		19.2	17.0	16.5	14.2	9.3
MSH	24.1	24.2	21.4	18.3	14.2	9.9
		23.2	20.5	19.5	17.7	15.7
		23.6	20.3	17.9	14.1	10.3
MTM	17.5	17.5	14.9	13.0	9.3	4.9
		15.2	14.3	14.0	13.2	10.1
		15.3	13.9	12.7	10.2	5.2
FAF	21.5	20.3	17.7	17.2	13.0	7.7
		18.8	16.2	15.5	15.3	14.0
		18.7	17.0	15.5	12.1	7.7
FMS	21.1	21.4	20.4	19.4	14.7	9.0
		20.7	18.4	17.5	16.0	15.5
		20.5	18.3	18.0	14.3	9.4
FYM	32.6	30.5	27.2	22.3	19.0	13.8
		28.6	27.3	24.7	23.0	20.2
		28.6	25.9	21.0	19.2	14.0
平均	22.0	21.8	19.4	17.5	13.7	8.8
		20.5	18.6	17.7	16.4	14.5
		20.6	18.2	16.8	13.8	9.0

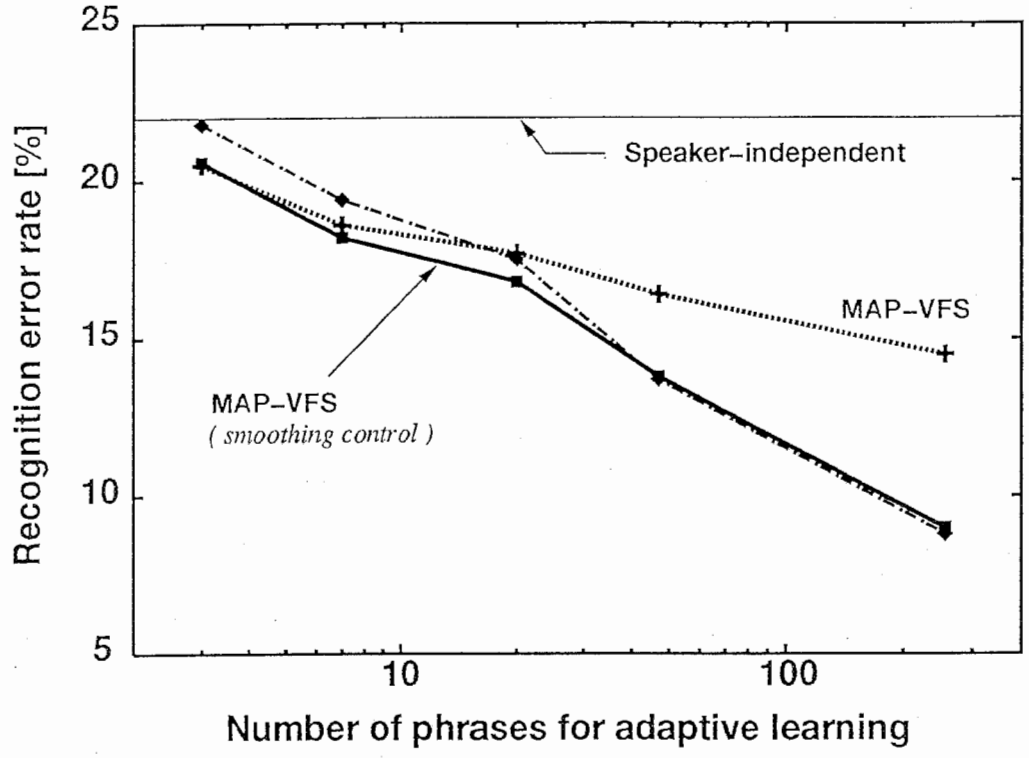

Figure 5.7: MAP, MAP-VFS, 平滑化制御 MAP-VFS による話者適応結果

5.3.8 まとめ

本節では少量データによる話者適応法として、MAP 推定法と VFS の統合アルゴリズム (MAP-VFS 法) を提案し、この MAP-VFS 法が両者の欠点を補間する手法であることを示した。本手法の有効性を日本語音素認識実験により評価し MAP 及び VFS をそれぞれ単独で用いた場合よりも高い性能が得られた。さらに MAP-VFS 法における事前知識の重要性を考慮し適応話者にとってよい初期モデルを得るために木構造話者クラスタリング法を用いた。そして、この初期モデルを用いることにより不特定話者モデルを初期モデルに用いた場合より高い音素認識率を得た。

さらに学習データ量に依存せず常に最適な話者適応性能を得ることを目的として、MAP-VFS 法に適応データ量に応じた平滑化係数の制御を組み込むことを提案しその有効性を確認した。

第 6 章

教師なし話者適応を利用した不特定話者音声認識

6.1 まえがき

これまでに我々は木構造話者クラスタリング法を提案し、教師つき話者適応への応用を図ってきた [55]。この方法は話者（クラスタ）選択による話者適応で、極少数の適応用サンプルで話者適応が行なえる。今回は、この手法を教師なし話者適応へと拡張をおこなった。また極少数のサンプルで適応できる特徴を利用して、木構造話者クラスタを用いた不特定話者音声認識について検討をおこなった。以上を SSS-LR による文節認識実験により評価した結果について報告する。

6.2 教師なし話者適応の原理

適応用の音声の入力に対し、一旦認識系により音声認識をおこない、その結果出力される音素系列をフィードバックし、話者適応時の教師信号として用いることにより、教師なし話者適応を実現する [58][61]。ここで用いる木構造話者クラスタリングによる話者適応では、木構造の枝の選択のみを行ない、平均値や分散などのパラメータの変更は行なわないため、少ないデータで教師なし学習ができると期待できる。

6.3 木構造話者クラスタリングによる不特定話者音声認識の原理

木構造話者クラスタリングを用いた不特定話者音声認識の原理について述べる。1発話のみの評価データで教師なし話者適応を行なうことでこれを実現する。アルゴリズムを以下に示す。

STEP 1 入力音声を不特定話者音素モデルを用いて認識する。

STEP 2 認識結果の音素系列をフィードバックし、STEP 1で用いた入力音声と、この音素系列を入力として話者選択を行なう。

STEP 3 選択後の音素モデルを用いて入力音声を再認識。

以上のような2パスで認識を行なう。本方法で認識率を向上するためには、誤認識するデータの認識率を改善する必要がある。このため誤認識をフィードバックしても、正しい方向へ学習をすすめる必要があるという本質的な問題がある。しかし、認識結果は文法などの知識によりある程度修正されたものであり、さらに文節で評価した場合誤っているだけで、すべての音素系列が誤っているわけではない。実際認識誤りデータを調べると、助詞の部分だけ誤ったものが多い。このことから誤認識結果のフィードバックでも話者適応は十分可能と考えられる。

6.4 認識実験

6.4.1 実験条件

以上の手法の有効性を確認するために、教師なし話者適応及び、木構造話者クラスタリングを用いた不特定話者音声認識の評価を文節認識実験により行なった。音響分析条件を付録A.3に示す。使用データを表6.1に示す。データとしては、279文節中からランダムにN個$(N = 1,3,5,7)$の文節を選んで適応を行ない、適応に用いなかった$279 - N$個の文節で評価を行なった。適応データに話者適応の性能が依存する可能性を考え、以上の評価を5回繰り返し平均の認識率を求めた。各話者について文節発声データにより単一ガウス分布・200状態のHMnetを作成した。学習には学習データ量削減のため移動ベクトル場

平滑化法 (VFS) を用いた。以上のようにして作成された 170 個の HMnet を最大クラスタ数 4 とおき木構造話者クラスタリングした。本方式ではクラスタ数が認識時の混合数と等しくなる。このため認識時に使用する HMnet は最大 4 混合で 200 状態のものである。通常 HMnet による不特定話者音声認識では 10 混合以上を用い、状態数もさらに増やすと性能が向上するが、ここでは話者適応の効果を検討するのが主な目的であるので、簡単なモデルを使用した。文節認識には LR 文法 (1407 規則,1035 語) を使用し、ビーム幅は 600 とした。

Table 6.1: 学習・評価データ

学習用データ			
話者	男性 85 人 + 女性 85 人	サンプル数	50 文
話者適応用及び認識用データ			
話者	男性 3 人 + 女性 3 人		
適応用データ	279 文節 (SB3 タスク) からランダムに選んだ N 個の文節 (N=1,3,5,7)		
認識用データ	279 文節 (SB3 タスク) から適応用に用いた文節を除いた $279 - N$ 個の文節		
不特定話者音声認識用データ			
話者	男性 3 人 + 女性 3 人		
認識用	279 文節 (SB3 タスク)		

6.4.2 教師なし話者適応実験結果

表 6.2に教師なし及び教師つき話者適応の結果を示す。教師なし適応後の平均認識率を見ると、適応データが 1 〜 7 文節で認識率は大差なく、極少数の適応データで適応が完了することが分かる。しかし、個々の認識率を見るとその値は変動していることから、適応サンプルに適応性能が依存していると考えられる。教師なしと教師つきの場合の適応の差を図 6.1に示す。図は適応文節数 5 の場合について、教師なしと教師つきのそれぞれの場合について、横軸に適

応前の縦軸に適応後の認識率を話者ごとにプロットしたものである。この図の点線のラインよりも上にプロットされた場合、適応の効果があったことになる。図から教師つきと教師なしでは明確な差が見られない。このように話者選択という枠組では、教師なしでも教師つきに近い性能が得られることが分かった。

Table 6.2: 各適応文節数における教師なし及び教師つき話者適応結果 (%) (200 状態で最大 4 混合の HMnet 使用, 上段教師なし、下段教師つき)

話者名	適応用文節数				
	適応前	1	3	5	7
MHT (男性)	61.6	70.8	74.5	68.3	71.6
		70.9	74.7	74.2	74.6
MMS (男性)	68.7	74.8	73.5	76.1	75.6
		73.9	71.2	74.8	73.0
MMY (男性)	62.6	69.5	67.9	67.6	66.4
		68.4	65.7	69.0	67.2
FAF (女性)	66.0	67.7	69.1	70.1	67.2
		68.9	67.2	71.1	70.7
FKN (女性)	55.8	68.4	67.0	67.3	67.6
		67.7	69.0	66.6	65.8
FMS (女性)	57.2	62.1	64.0	64.0	63.8
		66.6	64.1	63.9	64.1
平均	62.0	68.9	69.3	68.9	68.7
		69.4	68.7	69.9	69.2

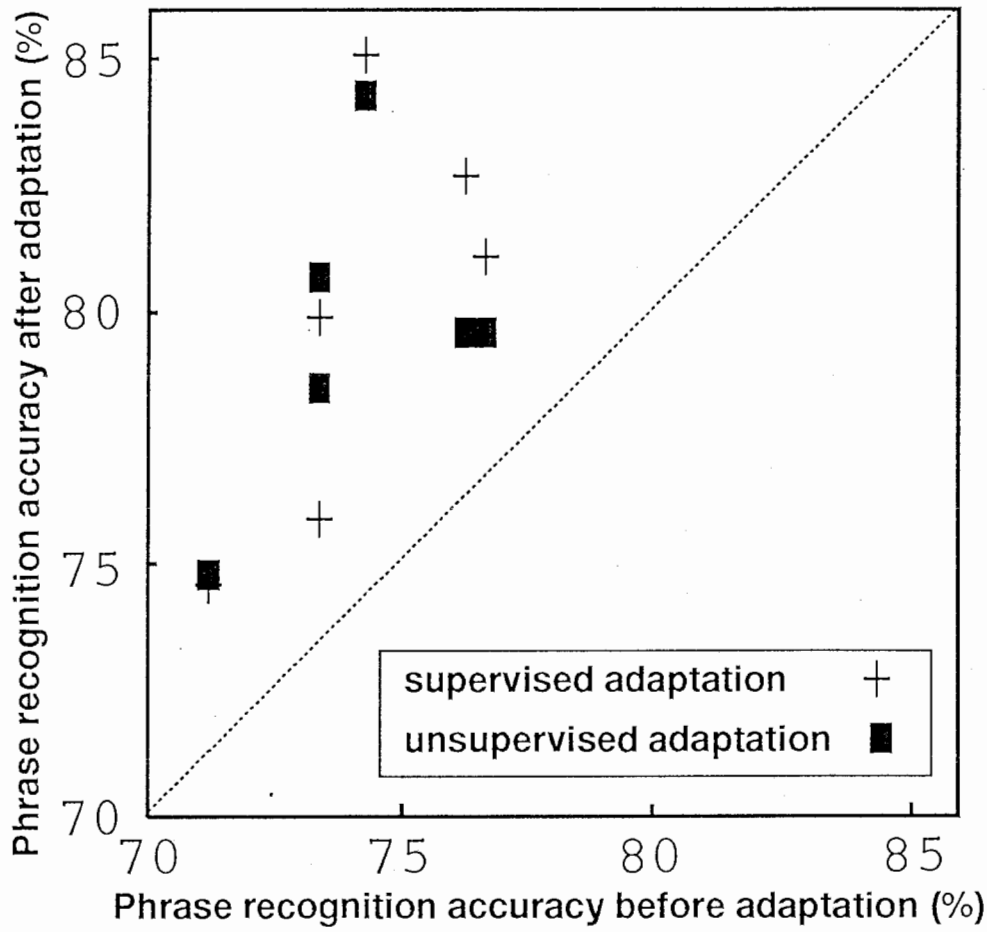

Figure 6.1: 教師つきと教師なし話者適応の比較 (横軸は話者適応なしの文節音声認識率, 縦軸は 5 文節音声を用いた話者適応後の認識率)

6.4.3 不特定話者認識実験結果

認識実験結果を表 6.3に示す。話者によってばらつきはあるが、いずれの話者でも木構造話者クラスタリングを用いた方式が話者混合法による不特定話者音声認識 [30] に比べ認識率が高くなっている。ここでは評価対象として文節を用いているが、文節の平均時間長は約 0.9(sec) であり、この程度の長さの入力音声で教師なし話者適応の効果が出なければ、不特定話者モードでの認識率の向上は期待できない。実験によると認識率は向上しているため、1 sec 以下の、

文節中に誤りを含む情報をフィードバックしても話者適応の効果が出ていると考えられる。

また表 6.2の適応文節数 1 と表 6.3を比べると同じ適応文節数でも差があり、表 6.3の方が低い。これは適応の場合は、適応サンプルに対しオープンデータで評価しているのに対し、不特定話者認識の場合はクローズのデータで評価しているからである。フィードバック型の教師なし適応の場合、フィードバック情報に誤りを含むためクローズでの評価の方が低くなると考えられる。

表 6.4は話者 FKN について、1 pass 目で誤認識を起こし、その情報をフィードバックしたところ 2pass 目では正解となった例を示す。

Table 6.3: 木構造話者クラスタリングを用いた不特定話者文節音声認識率 (%) (200 状態, 4 混合 HMnet 使用)

話者名	従来法	クラスタリング使用
MHT （男性）	61.6	63.4
MMS （男性）	68.7	72.7
MMY （男性）	62.6	68.0
FAF （女性）	66.0	67.4
FKN （女性）	55.8	66.2
FMS （女性）	57.2	63.3
平均	62.0	66.8

Table 6.4: 誤認識情報をフィードバックして正解となった例 (話者 FKN)

発声内容	1pass 目の結果	2pass 目の結果
事務局から	zhjuukjukakara	zhimukjokukara
大阪市	orasetarashii	oosakashi
5 日から	itakaqtara	itsukakara

6.5 まとめ

本報告では木構造話者クラスタリングに基づいた教師なし話者適応法を提案した。この手法を文節音声認識実験により評価しごく少ない適応データにより適応が可能であることを示した。さらに少数の適応データで動作する教師なし適応の応用例として、木構造話者クラスタリングを用いた不特定話者音声認識の検討を行なった。今後は話者モデル選択後のパラメータの適応、話者数の増加について検討する予定である。

第 7 章

話者適応の発話様式適応への応用

7.1 まえがき

近年自由発話音声 (Spontaneous Speech) 認識の研究が除々に開始されている。従来音声認識システムは書いた文章を読み上げる、朗読調発話により主に評価されていた。しかし音声対話システムの研究の進展と共に、話者の発声に制約をかけない自由発話での評価が始まっている。米国では ARPA プロジェクトの一貫として ATIS(Air Travel Information System) タスクによる音声理解の研究が進展しており、CMU、MIT、BBN などを中心に行なわれている [62]。また国内でもいくつかの機関が音声対話システムの評価として自由発話音声に対するタスク達成率などにより評価している (例えば、地理情報の案内 [63]、電話番号案内 [64][65]、品物の注文 [66] など)。しかし音響モデルに関しては朗読文などから学習したものがほとんどである。自由発話からの学習は、村上 [67] が試みているが、これは特定話者に関してのみである。

以上のように自由音声に対する音響モデルの検討はほとんど行なわれていないのが現状である。そこで不特定話者の自由発話音声認識に向けた第一段階として、音素認識実験による検討を行なった。性能のよいモデルを作るためには、大量のトランスクリプションやラベルが付けられた自由発話データが必要である。しかしこれを大量に整備するのは作業負担が大きい。そこで朗読調の文または文節発声データで学習されたモデルをもとに、少量の自由発話音声データで発話適応してモデルを作成することを試みた。少量データによる学習という点では発話適応と話者適応は共通性がある。そこでここまでの章で検討してき

た話者適応技術が発話適応として使用できないか検討した。

ここではまず文節発声で学習された不特定話者用 HMnet を初期モデルとし、MAP-VFS 話者適応法 [68] などを用いて、話者、発話様式、話者＋発話様式の 3 種類で適応した場合の比較を行なった。その検討結果を述べる。

7.2 自由発話音声データベースの概要

構築中の自由発話音声データベースのタスクは旅行会話に関する対話で、1) 通訳を介した日本語対英語による会話、及び 2) 日本語対日本語による会話である。会話は非対面で行ない、発話者は役割を書いたプロットを基に会話をすすめ、両話者の発話が重ならない等の制約が課せられている。本稿で評価対象とするのは、1) の会話で、トピックがホテル予約に関する日本語による申込者側の発声である。通訳者が間に入っているため、2) と比較して冗長語、言い淀みが少ない傾向にある。

7.3 認識実験

パラメータ及び評価方法

音響分析条件を付録 A.3に示す。

音素認識実験には One-Pass DP 法 [69] に基づく音素タイプライタにより [70]、音素の Accuracy を求めた。評価した音素の種類は 26 で長母音と単母音の区別は行なっていない。また断りがない限り日本語中に存在する音節並びによる接続の制限を加えている。この時の音素パープレキシティーは 10.2 である。音素 Accuracy は以下の式により求める。

$$Acc = \frac{N_{phones} - N_{subst} - N_{ins} - N_{del}}{N_{phones}} \tag{7.1}$$

ここで、N_{phone} は正解音素系列中の総音素数、N_{ins} は音素挿入誤り数、N_{del} は音素脱落誤り数、N_{sub} は音素置換誤り数を表すものとする。

7.3.1 不特定話者文節認識用音素モデルの作成及び評価実験

自由発話データ認識用の音素モデルを作成する場合、通常は Baum-Welch アルゴリズムにより作成するが、発話様式の適応を行なえば少量のデータでもモデルが作成可能であると考えられる。そこで、ここでは MAP-VFS、MAP、VFS の各話者適応方式を発話様式適応に利用してモデルを作成した。

まず文節発声データで音素モデルを作成し、それを初期モデルとして少量の自由発話データを使って適応し、自由発話認識用のモデルを作成する。さらに初期音素モデルを用いて、発話様式の異なるデータを認識し、発話様式の違いによる認識率の違いを比較検討した。

ここでは音素モデルとして、HMnet を使用した。HMnet の構造は話者 1 名が発声した 2620 単語により 5 混合、200 状態のモデルを作成することにより得た。さらに 10 混合 1 状態の無音モデルを付加し、総状態数 201 のモデルを作成した。この HMnet の構造は以下の実験ですべて共通とした。次に 285 名 (各話者ごと文節発声の音素バランス 50 文) 中からクラスタリングにより 15 名を選択し [71]、パラメータを求めた。求め方としてはまずモデル合成法 (CCL)[71] により不特定話者用の HMnet を作成した後、さらに Baum-Welch でパラメータの再推定を行なった。

この文節発声データでパラメータを求めた HMnet を用いて、発話様式の異なる以下の 3 種類のデータの認識を行なった。

朗読発話文節発声 国際会議予約に関する、文節に区切ってテキストを読み上げ発話したデータ (ATR SB3) で、評価話者は男性 4 名、女性 3 名。評価音素総数は 18,835 個。

朗読発話文発声 国際会議予約に関する、文単位にテキストを読み上げ発話したデータ (ATR SC3) で、評価話者は文節発声と同じ 7 名。評価音素総数は 15,818 個。

自由発話 旅行会話に関する対話で、評価話者は文節発声と異なる男性 3 名、女性 4 名。評価音素数は 3,400 個。

音素認識結果を表 7.1 に示す。

Table 7.1: 発話様式の異なるデータに対する認識率 (%)

発話	朗読発話 文節発声	朗読発話 文発声	自由発話
認識率	81.6	72.5	63.3

文節データでパラメータ推定をしたモデルを用いているため、文節発声のデータに対する認識率は高く 81.6% が得られる。これに対し発話様式及びタスクの異なる発声に対しては認識率が低く、文節発声 > 文発声 > 自由発話 の順になる。このように発話様式が異なる場合認識率は低く、発話様式適応の必要があると考えられる。

7.3.2 話者適応法を利用した自由発話データへの適応

以上に述べた文節で学習した HMnet を初期モデルとして、話者適応法を利用して自由発話データへ適応を行なった。また比較として Baum-Welch によるパラメータ推定も行なった。実験は男性話者 2 名 (MMAKO, MNOSA)、女性話者 2 名 (FYUYO, FYOMA) の計 4 名で行なった。適応 / 学習及び評価データ数は対話データのため、話者ごとに異なる。データ数を表 7.2に示す。以下の実験を行なった。

方法 1 VFS により適応を行なう。分散適応あり (表 7.3)。

方法 2 MAP により適応を行なう。分散適応なし (表 7.4)。

方法 3 MAP-VFS により適応を行なう。分散適応なし (表 7.5)。

方法 4 パラメータ制御つき MAP-VFS により適応を行なう。分散適応なし (表 7.6)。

方法 5 VFS により適応を行なう。分散適応あり。日本語による拘束なし (表 7.7)。

方法6 Baum-Welchによりパラメータ推定を行なう。この場合データ数の不足による頑健性の低下を防ぐため、アルゴリズムを変更した。まず各混合分布に対するサンプル数が1以下の場合は、パラメータの更新はしない。この場合のサンプル数の数え方は混合重みを考慮してカウントするので、実際は複数個のサンプルが入力されないとパラメータの更新は行なわない。また分散は拡大する方向には更新するが、縮小する場合は更新しない(表7.8)。

各方法について以下の5項目の実験を行なった(最終項目の発話適応後話者適応は全ての方法で行なってはいない)。

(a) 適応なし 文節学習のモデルを適応なしに用いて自由発話音声を認識。

(b) 話者適応 単一話者が発声した、朗読文によりモデルの適応を行ない、同一話者の対話を認識。これは話者適応のみを行なったことになる。適応のためのデータは音素バランス50文である。

(c) 発話適応 複数話者が発声した、自由発話音声でモデルの適応を行ない、別話者の対話で認識。これは発話適応のみを行なったことになる。つまりこれが不特定話者自由発話モデルを用いた音素認識の結果と考えられる。適応のためのデータは話者6名の自由発話各1対話である。

(d) 話者+発話適応 単一話者が発声した、自由発話音声でモデルの適応を行ない、同一話者の別の対話を認識。これは話者及び発話様式適応を同時に行なったことになる。適応のためのデータは自由発話1対話分である。

(e) 発話適応後話者適応 発話適応と話者適応を同時に行なわずに、発話適応した後に話者適応を行なう。同一話者の別の対話を認識。適応のためのデータは適応条件(c)と(d)のデータの和になる。

Table 7.2: 学習及び評価に用いたデータ文数。括弧内は音素数。

適応条件	話者			
	MMAKO	MNOSA	FYUYO	FYOMA
(a)	0	0	0	0
(b)	50(2953)	50(3411)	50(2943)	50(3177)
(c)	80(3224)	81(3371)	86(3528)	78(3108)
(d)	12(614)	11(467)	6(310)	14(730)
(e)	(c) と (d) のデータの和			
評価データ数	12(552)	12(449)	13(475)	8(275)

Table 7.3: VFS による適応結果 (%)

実験番号	話者				平均
	MMAKO	MNOSA	FYUYO	FYOMA	
(a)	65.0	70.6	60.6	62.5	64.7
(b)	75.9	80.6	73.5	75.6	76.4
(c)	72.1	75.1	70.9	74.2	73.1
(d)	83.9	84.9	81.1	78.2	82.0
(e)	82.2	83.3	80.4	75.6	80.4

Table 7.4: MAP による適応結果 (%)

実験番号	話者				平均
	MMAKO	MNOSA	FYUYO	FYOMA	
(a)	65.0	70.6	60.6	62.5	64.7
(b)	79.3	79.3	75.6	72.0	76.6
(c)	68.8	73.3	64.2	64.7	67.8
(d)	83.5	84.2	78.1	76.0	80.5
(e)	81.3	84.0	80.8	77.5	80.9

Table 7.5: MAP-VFS による適応結果 (%)

適応条件	話者				平均
	MMAKO	MNOSA	FYUYO	FYOMA	
(a)	65.0	70.6	60.6	62.5	64.7
(b)	76.3	80.6	71.4	75.6	76.0
(c)	72.6	72.4	69.9	74.9	72.5
(d)	81.7	82.6	79.6	80.7	81.2
(e)	81.0	83.3	79.8	78.9	80.8

Table 7.6: MAP-VFS 適応 (パラメータ制御有り) による適応結果 (%)

実験番号	話者				平均
	MMAKO	MNOSA	FYUYO	FYOMA	
(a)	65.0	70.6	60.6	62.5	64.7
(b)	78.8	78.6	75.4	72.4	76.3
(c)	68.1	72.6	64.2	65.1	67.5
(d)	83.3	83.3	78.9	76.7	80.6
(e)	81.9	84.0	79.4	77.5	80.7

Table 7.7: VFS による適応結果。日本語拘束なし (%)

実験番号	話者				平均
	MMAKO	MNOSA	FYUYO	FYOMA	
(a)	57.6	62.1	50.3	56.0	56.5
(b)	70.7	71.5	62.5	64.7	67.4
(c)	65.2	66.8	59.4	64.4	64.0
(d)	77.7	77.7	71.2	66.5	73.3

Table 7.8: Baum-Welch によるパラメータ再推定 (%)

適応条件	話者				平均
	MMAKO	MNOSA	FYUYO	FYOMA	
(a)	65.0	70.6	60.6	62.5	64.7
(b)	79.9	81.3	75.2	70.5	76.7
(c)	70.7	72.6	65.7	67.6	69.2
(d)	81.9	82.6	76.8	77.1	79.6

7.3.3 結果の考察

日本語の拘束有りと無しとでは認識率で 10% 程度異なるが、同じような傾向を示す。各話者適応法と Baum-Welch ではほぼ同じような傾向を示す。ただし VFS のみ分散適応も行なっている。またパラメータ制御つき MAP-VFS ではスムージングレートがデータ数によって自動的に制御されるため、VFS と比較して、最適なスムージングレートを選択したくともよいという利点がある。

話者＋発話適応を行なったものが認識率が一番よい。適応条件 (d) と (e) を比較すれば分かるように、(d) のデータのみで十分適応が可能である。話者適応のみと発話適応のみの比較をすると、話者適応のみの方が効果がある。このことから朗読音声と同様に、自由発話データでも話者変動の問題の解決が重要であることが分かった。

7.4 まとめ

自由発声中の音素認識による発話／話者適応の効果を検討した。文節発声で学習された HMnet を MAP-VFS により話者、発話様式、話者＋発話様式の 3 種類のデータで適応してモデルを作成した。その結果 MAP-VFS での発話適応で良好な結果が得られることが分かった。また自由発話音声の認識においても、個人性が大きな問題となることが分かった。今後は学習及び評価話者を増加させての検討、さらに自由発声用音素モデルにおける話者適応について検討

する必要がある。

第 8 章

結論

本論文では音声認識における不特定話者の問題を扱った。話者変動の対処としてはおおまかに分けて2通り存在する。一つは不特定話者音声認識の性能そのものを向上させることであり、一つは話者適応により、適応前は性能の悪い認識システムを、使用していくにつれその話者の個人性を学習し性能を向上させる方法である。本論文は以上の二つの方法により不特定話者に対する音声認識の性能向上を目指し、その実現方法について述べた。

第1章はまえがきであり、音声認識研究の現状を述べその問題点を明らかにすると共に、本研究の必要性を述べた。

第2章は準備の章であり、本論文の理解のために必要な基本技術に関する概説を行った。本論文では音響モデルとしてHMMおよび隠れマルコフ網(HM-net)を用いたためその2者の概説を行なった。さらにこれらのモデルを用いた連続音声認識法であるHMM-LR連続音声認識法及び、その音素環境依存モデルへの拡張版であるSSS-LR連続音声認識法について述べた。さらにモデルのパラメータ推定や話者適応法と関わる、移動ベクトル場平滑化法(VFS)について述べた。

第3章は話者適応の初期モデルとして用いる不特定話者モデルについて検討を行なった。まず第3.2節では、混合連続分布型HMMで問題となる混合数の自動決定法について論じた。分散の行列式の値を基準として、音素モデルごと、または分布ごとに混合数を決定する方法を提案した。その結果、分布ごと、音素モデルごと、混合数がすべて均一の順で認識率が得られ、混合数を自動決定することの有効性を示した。

第 3.3節では、不特定話者用の音素環境依存モデルの効率的な作成法「話者混合法」について提案した。不特定話者の文節認識で認識実験を行なった結果、環境非依存の HMM (3 状態、20 混合) で 75.79% の認識率だったのに対し、600 状態、12 混合の音素環境依存モデルで 82.80% の認識率が得られ有効であることが分かった。

第 3.4節ではさらに話者混合法に話者クラスタリング法を取り入れることにより任意の混合数でモデルを合成できる CCL 法を提案した。その結果従来の Baum-Welch 学習に比較してモデルの性能を損なうことなく短時間でモデルが作成できることを示した。計算時間を比較すると混合数 5 の場合で約 1/20、混合数 15 の場合で約 1/60 で、この差は混合数が増すほど増加する。また混合数 5 で比較した場合、クラスタリングにより初期値を与えたのち Baum-Welch で学習する方法で認識率が 64.6% だったのに対し、CCL では 73.6% と高い認識率が得られた。さらに CCL で作成したモデルを初期値として Baum-Welch 学習をすることにより認識率として 77.2% が得られ、高性能なモデルを得る場合の初期値としても有効であることが分かった。

第 4章では、少量の適応データで動作する話者適応法について検討を行なった。従来の話者適応では適応データとして数十秒〜数分程度の量を必要としたが、ここでは数秒程度で動作する話者適応法の提案を行なった。

まず第 4.2節では話者重み学習による話者適応法について述べた。この方法では話者ラベルの付与された HMM の各混合分布において、同一のラベルがふられているものを「結び」の関係として、話者重みを学習する方法である。さらに重みがある閾値以下になった場合その混合分布を削除し認識時の計算量の削減を行なう話者プルーニングについても検討を行なった。その結果 5 単語で適応した場合、認識率が 75.78% から 80.46% に向上し、非常に少ないデータでも話者適応が可能であることが分かった。さらに話者プルーニングをおこなうことにより、認識率の低下なしに混合数を 1/2 〜 1/12 程度に削減できることを示した。

つぎに第 4.3節では木構造話者クラスタリングによる話者適応法について述べた。基本的には話者選択型の話者適応であるが、話者モデルを木構造に構成しモデルを選択することにより、大量話者を扱うことが可能になった。実験では 170 名の話者を用いて木構造を構成し話者適応をおこなった。SSS-LR によ

り文節認識実験を行なったところ適応文節数5の場合で、74.2% から 79.9% へ認識率が向上 (22.1% の誤り率の減少) した。また一番効果のある話者で見ると 74.3% から 85.1% へ向上 (42.0% の誤り率の減少) し本手法が有効であることが分かった。

第5章では、前章の少量データで動作する話者適応での成果をもとに、少量データでも、データ数が増加した場合でも、データ量に応じて動作する、データ量に応じた話者適応について検討した。ここではこれを実現する方法として「複数の話者適応法に基づく動的話者適応」と「MAP-VFS 話者適応法」を提案した。

まず第5.2節では複数の話者適応手法を尤度基準により切替え、データに応じた話者適応を行なう「複数の話者適応法に基づく動的話者適応」について述べた。実験の結果データ数が増加するに従い自由パラメータの少ない適応法からパラメータの多い適応法へと自動切替えが行なわれることが確認され、複数の話者適応の自動切替えに対する見通しが得られた。

第5.3節では、適応データ量に応じてモデルパラメータの自由度を制御する話者適応法として MAP-VFS 法を提案した。この方法により音素認識実験した結果、7文節による適応で不特定話者モデルより 1.0 ～ 6.5%(全話者平均 2.9%) の認識率の向上が得られた。さらに第4.3節の木構造話者クラスタリングと併用することにより、7文節での適応後で音素認識率は不特定話者モデルより 3.9 ～ 10.9%(全話者平均 6.2%) 向上する結果が得られた。以上より MAP-VFS 法では事前確率分布として適当なもの選択すれば、効果が高いことが分かった。

第6章では以上による話者適応を利用した、新たな不特定話者認識の枠組を提案した。本提案法では予備認識の結果を用いて話者適応をし、再認識するという2段階の認識を行なうことによりモデルを話者の観点から絞り込み認識性能の向上を図る方法をとった。この結果不特定話者文節認識率が 74.2% から 76.4% の認識率の向上が得られた。

第7章では、本論文で提案した話者適応法を発話様式適応に応用し自由発話音声 (Spontaneous Speech Data) の認識の検討を行なった。文節データで学習されたモデルを初期モデルとして、少量の自由発話データで発話様式適応した結果、72.5% の音素認識率 (Accuracy) が得られた。これは ARPA の各研究機関が、音素認識率が 70% 前後得られる音響モデルを用いて高性能なディクテー

ション・システムを構築していることを考えると、十分に高い認識率であると言える。

第8章は結論であり、本研究の成果を要約した。

謝辞

研究の機会を与えて頂いた、ATR自動翻訳電話研究所樽松明社長(現電気通信大学教授)、及びATR音声翻訳通信研究所山崎泰弘社長に感謝いたします。

現NTTヒューマンインタフェース研究所嵯峨山茂樹氏にはATR自動翻訳電話研究所音声情報処理研究室室長在職時に、HMnetをベースとした不特定話者音声認識および話者適応の研究における著者の指導者として、お世話頂いた。

ATR音声翻訳電話研究所第一研究室室長匂坂芳典博士には、研究室室長として貴重な御指導と御助言を頂いた。

音声翻訳通信研究所松永昭一博士には話者クラスタリング手法を用いた話者適応および不特定話者モデル作成、さらに発話様式適応に関して主任研究員として御指導、御助言頂いた。

現ビクターの鷹見淳一氏は話者混合逐次状態分割法の研究に関して共同研究者としてお世話頂いた。またHMnet作成に関するソフトウエアを提供して頂いた。現三洋電機(株)の大倉計美氏、現日本電気の服部浩明博士、現エプソンの宮沢康永氏には話者適応や音響モデルに関して熱心な御議論を頂いた。現豊橋技術科学大学の倉岡幹雄氏には話者クラスタリング手法を用いた不特定話者モデルの作成に関する共同研究者として御協力頂いた。ATR音声翻訳通信研究所の外村政啓氏にはMAP-VFS話者適応に関して共同研究者として御協力頂いた。また多くのATRの研究員の皆様に御助言、御討論頂いた。

現シャープ(株)の山口耕一氏には混合分布かたHMM-LRのソフトウエアを提供して頂いた。また現三菱電機の永井明人にはSSS-LRのソフトウエアを提供して頂いた。

このほか学会、研究会を通じて多くの研究者の方々に御助言を頂いた。また数多くのATRの職員の皆様、データベースの整備を行なって頂いた方々、研究環境の整備を行なって頂いた方々にお世話になった。

以上記して感謝します。

付録 A

音声特徴の抽出

A.1 まえがき

音声認識などでは、音声波形そのものを扱うことはほとんどなく、多くの場合スペクトルパラメータに変換されて表現される。その大きな理由は、波形のままでは情報が冗長過ぎて扱いにくいこと、および人間の聴覚に関与するのはスペクトルであること、などである。このため、冗長な情報である音声信号波形に、音声分析を行なって、情報を圧縮し、効率の良い音響パラメータに変換して扱う。

ここでは、そのための音声分析のアルゴリズムについて説明する。

A.2 LPC 音声分析

本研究ではLPC ケプストラム分析に基ずく音声パラメータを使用した。具体的には16 次のLPC ケプストラム、16 次のデルタケプストラム、log パワー、デルタ log パワーの計34 次元ベクトルである。

そこで以下LPC ケプストラム分析およびデルタケプストラムの求め方を具体的に述べる。これによって、音声波形から音声の特徴パラメータベクトル時系列が得られる。

1. サンプリングと高域強調

   マイクロフォンなどから入力されたアナログ音声信号は、まずディジタル信号に変換される。このためには、サンプリング定理に従って、音声

信号を (たとえばカットオフ周波数 6kHz の) 低域通過フィルタ (LPF) を通して帯域制限し、その帯域の倍の周波数 (Nyquist 周波数, たとえば 12kHz) でサンプリングし、AD 変換器により (たとえば 16 ビットで) 量子化して、ディジタル信号を得る。

音声信号 $\{X_t\}$(t は整数の時刻) は、通常低域のほうにエネルギー成分が偏っているので、その後の音声分析精度を高くするために、一次差分フィルタにより高域強調する。その計算は、高域強調係数 $\alpha(=0.98)$ を用いて、$X'_t = X_t - \alpha X_{t-1} \quad (t = 1, 2, 3, \cdots, N)$ のように行なう。

2. データ窓

音声信号は時々刻々変化する非定常な信号であるが、20ms などの短時間については定常信号と同様のスペクトル分析ができる。このような短時間の信号を、元の音声信号に矩形窓を掛けて切り出すと、両端の効果が現れてスペクトル推定に害がある。そこで、Hamming 窓:

$$h_t = 0.54 - 0.46 \cos \frac{2\pi t}{N} \qquad t = 1, 2, 3, \cdots, N \tag{A.1}$$

を掛けて切り出す。

これを、音声信号に、$x_t = h_t X'_t \quad (t = 1, 2, \cdots, N)$ のように掛ける。N は窓長で、12kHz サンプリング、20ms 窓の場合は、$N = 240$ である。このような窓掛けを短時間 (たとえば 5m 秒) ずつずらして行ない、1 フレームごとの音声データを得る。

3. 自己相関分析とラグ窓かけ

LPC 分析に先立って自己相関関数を計算する。1 フレームの窓長を N サンプルとすると、音声信号サンプル値 $\{x_1, x_2, \cdots, x_N\}$ から、

$$v_\tau = \frac{1}{N} \sum_{t=1}^{N-\tau} x_t x_{t+\tau} \qquad \tau = 0, 1, 2, 3, \ldots \tag{A.2}$$

により自己相関関数を求める。

4. PARCOR アルゴリズム (Levinson-Durbin-Itakura 法)

線形予測分析 (LPC) を行なう。音声信号 x_t を過去の p 個の値から

$$\hat{x}_t = -a_1 x_{t-1} - a_2 x_{t-2} - a_3 x_{t-3} - \cdots - a_p x_{t-p} \tag{A.3}$$

により予測するモデルで、定常区間内で予測値の二乗誤差が最小になるような係数 $\{a_i\}$ を求めるものが、p 次 (= 16 次) 線形予測分析 (Linear Predictive Coding: LPC) と呼ばれるものである。その解は、次のような正規方程式 (Yule-Walker 方程式とも呼ばれる) を解くことにより得られる。

$$\begin{pmatrix} v_0 & v_1 & v_2 & \cdots & v_{p-1} \\ v_1 & v_0 & v_1 & \cdots & v_{p-2} \\ v_2 & v_1 & v_0 & \cdots & v_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{p-1} & v_{p-2} & v_{p-3} & \cdots & v_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = - \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_p \end{pmatrix} \tag{A.4}$$

この行列は特殊な形 (Toeplitz 行列と呼ばれる) をしているため Levinson-Durbin アルゴリズムあるいは 板倉・斉藤の PARCOR アルゴリズムと呼ばれている効率の良い解法がある。

初期値として、

$$k_1 = \frac{v_1}{v_0} \tag{A.5}$$

$$a_1^{(1)} = -k_1 \tag{A.6}$$

$$u_1 = v_0 - k_1 v_1 \tag{A.7}$$

$i = 2, ..., p$ について、

$$w_{i-1} = \sum_{j=1}^{i-1} a_j^{(i-1)} v_{i-j} + v_i \tag{A.8}$$

$$k_i = \frac{w_{i-1}}{u_{i-1}} \tag{A.9}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \quad (j = 1, 2, \cdots, i-1) \tag{A.10}$$

$$a_i^{(i)} = -k_i \tag{A.11}$$

$$u_i = u_{i-1} - k_i w_{i-1} \tag{A.12}$$

このアルゴリズムで得られた、$\{a_i^{(p)}, i = 1, 2, \cdots, p\}$ が、求める線形予測係数である。このアルゴリズムは、通常の消去法などに比べて計算量を大きく減らすことができ、乗算は $p^2 - p + 1$ 回、除算は p 回で済む。

5. LPC ケプストラム計算

ケプストラムは、対数パワースペクトル $f(\omega)$ のフーリエ展開係数で、

$$\log f(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega} \tag{A.13}$$

である。線形系の縦続接続は、時間領域では畳み込み (convolution) だが、フーリエ領域では積になり、対数スペクトル領域では和となり、従ってケプストラム領域でも和となるため扱いやすい。音声認識の特徴パラメータとして盛んに用いられる。スペクトル線形予測係数 $\{a_i\}$ とケプストラムの間には、

$$c_m = -a_m - \sum_{i=1}^{m-1} \frac{m-i}{m} a_i c_{m-i} \qquad (m = 0, 1, 2, \cdots, n) \tag{A.14}$$

のような Oppenheim の漸化式と呼ばれる簡単な関係がある。これにより、LPC ケプストラムは容易に求められる。

6. 動的特徴パラメータ計算

音声信号の知覚においては、瞬時ごとのスペクトルだけでなく、スペクトルの変化も大きく寄与していると言われる。そこで、音声スペクトルがダイナミックに変化する様子を、音声スペクトルパラメータ (具体的には LPC ケプストラムなど) のベクトル時系列に部分的に重みをつけた窓で最小自乗近似直線を求め、その傾き (回帰係数) を特徴量とする方法が用いられる。LPC ケプストラムの場合のこの特徴量は、俗にデルタケプストラムと呼ばれる。

原点を時刻 t (フレーム番号) に取れば、局所的な変化は重み付き最小二乗法で直線近似できる。すなわち、

$$\theta(i) = ai + b \tag{A.15}$$

をモデルとして、誤差

$$E = \sum_{i=-n}^{n} w_i(\theta(i) - ai - b)^2 \tag{A.16}$$

を最小にするように、係数 a と b を決定する。ただし、$\{w_i\}$ $(w_i \geq 0)$ は、重み係数である。

もし、w_i が対称 $(w_i = w_{-i})$ ならば、a と b は簡単に求められ、

$$a = \frac{\sum_{i=-n}^{n} iw_i\theta(i)}{\sum_{i=-n}^{n} i^2 w_i} \tag{A.17}$$

$$b = \frac{\sum_{i=-n}^{n} w_i\theta(i)}{\sum_{i=-n}^{n} w_i} \tag{A.18}$$

により得られる。$\theta(i)$ がケプストラムの場合、係数 a は、最近は、デルタケプストラムと呼ばれ、近年、非常によく用いられる。

以上のようにして、音声信号から、フレームごとにLPCケプストラムベクトルが求められる。

A.3 音響分析条件

本研究で用いられた音響分析条件を以下に示す。

Table A.1: 音響分析条件

分析条件	サンプリング周波数 12kHz
	ハミング窓 (20ms)
	プリエンファシス 0.98
	分析周期 5ms
	16次LPCケプストラム + ログパワー +
	16次 Δ ケプストラム + Δ パワー

付録 B

音声データベース

ＡＴＲでは、音声認識、音声合成および音声知覚等の研究に用いる、大語彙の高品質な研究用日本語音声データベースを構築している。

B.1 構成

B.1.1 セットＡ (大語彙単語音声および文節・文章音声)

セットＡは各話者ごとに 9 種類のデータセットからなり、内容は以下のとおりです。

- 重要語 5240 語
- 音素バランス単語 216 語
- 単音節 101 個
- 数字音声 25 個
- アルファベット 35 個
- 外来語 9 個
- 国際会議申し込み会話 (文節発声 DSA) 115 文
- 国際会議申し込み会話 (文節発声 DSB) 115 文
- 国際会議申し込み会話 (文章発声 DSC) 115 文

Table B.1: 研究用日本語音声データベース (セットA)

NO	話者	内容	種類	標準規格	備考
1	MAU	男性アナ			
2	MHT	男性ナレータ			
3	FKN	女性ナレータ			
4	FSU	女性アナ			
5	FKS	女性アナ	単語音声データ	unix システム	話者毎に販
6	FYN	女性アナ	(8,500 語)	tar コマンド格納	販売します
7	MTK	男性ナレータ			
8	MMY	男性アナ		1 ブロック 512	
9	MMS	男性アナ	12kHz 1 回発声のみ	バイト	
10	MNM	男性アナ	または	20 ブロック化	
11	MXM	男性アナ	20kHz 1 回発声		
12	FFS	女性アナ	(2 回発声)		サンプリン
13	FYM	女性ナレータ		MT 記録密度	グ周波数
14	FMS	女性アナ	1 回発声のみ	6250bpi	
15	FKM	女性アナ	ラベルデータ	オープンリール	12kHz 又は
16	FAF	女性アナ		または	20kHz の、
17	FTK	女性アナ		12,500ftpi	どちらかを
18	MTT	男性アナ		150M データ	ご指定下さ
19	MTM	男性アナ		カートリッジ	い
20	MSH	男性アナ			

B.1.2 セットB (音素バランス文章音声)

Table B.2: 研究用日本語音声データベース (セットB)

51	MYI	基本周波・言語韻律情報付男性アナ	503文 (10,000語)
52	MTK	男性ナレータ	12kHz連続発声のみまたは20kHz連続発声および文節区切発声

B.1.3 セットC (多数話者)

セットCは6種類のデータセット (C1 ～ C6) からなり、内容については以下のとおりです。

- C1 : 最重要語 520
- C2 : バランスリスト 216 +数字 A 15 +連続発声 A 50
- C3 : 連続発声 B 50 + C 50
- C4 : 文節発声 A 50
- C5 : 文節発声 B 50
- C6 : 文節発声 C 50

注1. A～ Cは、音素バランスを考慮して選んだ50文章からなるセットで内容が異なる。

Table B.3: 研究用日本語音声データベース Set C

NO (*は1～6)	話者グループ	備考
C*_ M01	男性20話者	話者グループ毎にC*単位に販売
M02	男性14話者	します。
F01	女性20話者	20kHz
F02	女性11話者	(標準規格はA、B¿/D に同じ)

B.2 Set A

データのあらまし

Aセット ５２４０単語『新明解国語辞典より抜粋』（５２４０個）

音素連鎖バランス単語リスト （２１６個）
１０１音節『単音節』 （１０１個）
数字A・B （１５＋１０ ２５個）

アルファベット （３５個）
外来音 （９個）

文章『国際会議の申し込み』（１１５文 x ３）

発声方法 ： 文章データ
自由発声『区切りなし』
文節A発声『長単位』
文節B発声『短単位』

話者　　：アナウンサー　１６名　（男性８名　女性８名）
　　ナレータ　　４名　（男性２名　女性２名）

話者名　　：アナウンサー　(MAU MNM MMS MMY MXM MTT MTM MSH)
　　(FSU FKS FFS FMS FYN FKM FAF FTK)
　　ナレータ　(MHT MTK　FKN FYM)

発声回数　：２回

ラベリング　：１回目発声のみ

音声切り出し　：全社、１、２回目とも

Down_sampling：１回目発声のみ

B.3　Set B

データのあらまし

Ｂセット　５０３文『Ａ～Ｊ』（５０３文ｘ２）

発声方法　：連続発声『区切りなし』
　　文節発声『区切りあり』

話者　　：ナレータ　８名　（男性４名 女性４名）
　　アナウンサー　３名（男性２名 女性１名）

話者名　　：ナレータ　(MHT MYI MHO MTK　FKN FYM FKS (FKN))
　　アナウンサー（MMY MSH　FTK)

発声回数　：　２回＊

ラベリング　：　ＮＥＣ殿納品データ…自由発声のみ（６名）
ＭＹＩ・ＭＨＯ・ＭＴＫ・ＦＹＭ　・ＦＫＳ・ＦＫＮ＊＊

三菱殿納品データ…両発声（２名）
ＭＨＴ・ＦＫＮ

ＡＴＲデータ…両発声（３名）
ＭＭＹ・ＭＳＨ・ＦＴＫ

音声切り出し　：　全社、１回目発声のみ

Down_sampling：　連続発声のみ（１回目発声）

＊ … 三菱・ＡＴＲは２回目発声をおこなっているが、ＮＥＣは１回のみの発声となっている。（92.11.16　確認済　ＡＴＲ担当：藤原氏）

＊＊ … 三菱納品のＦＫＮと同一人物であるため、外部販売はしない方針である。

B.4 Set C

データのあらまし

Ｃセット　５２０単語『５２４０単語からの抜粋』（５２０個）

音素連鎖バランス単語リスト（２１６個）
数字Ａ（１５個）

文章『Ｂセットからの抜粋・Ａ～Ｃ』（１５０文ｘ２）

発声方法　：文章データ
連続発声『区切りなし』
文節発声『区切りあり』

話者　：素人

発声回数　：１回

ラベリング　：文章データは、連続発声のみ付与
（ＡＴＲでは、両発声とも付与。他社の一部に両発声付与あり）

音声切り出し　：全社、全データ

Down_sampling：全データ

付録 C

ATRにおける研究概要および著者業績

C.1 ATRにおける研究概要

連続出力分布HMMを用いた不特定話者音声認識 (平成3年9月～平成3年11月)

従来離散型HMMのVQマッピング型の話者適応化法により、不特定話者への対応を行ないているが、自由度の少ないマッピング法には適応にも限界があると考えられる。特に1対1の話者間でうまくマッピングできるかは明らかではない。そこで初めから複数話者によってトレーニングされた連続型HMMを用い、この不特定話者用のHMMを更に話者適応することにより認識率の向上を図ることを目的とする。まず不特定話者HMMの混合数について検討を行なった。混合数を上げることで認識率は向上し、２０ミクスチャーのとき３４音素の認識で８０．４４％の認識率を得た。この程度のミクスチャーで認識率はほぼ頭打ちとなるが、データ数の少ない音素では認識率が低下しており、データ数によるミクスチャーの制限をする必要があると考えられる。

CMHMMにおける混合数の自動決定 (平成3年10月～平成4年3月)

混合連続HMMを用いる場合、その混合数の決定法が問題となる。そこで出力分布の分散の大きさを基準した混合数の自動決定法を提案した。分散を基準とすることにより、サンプル数が少なくその結果分散が小さくなっているカテゴリでは混合数が少なく割り当てられ過学習を避けることができる。また分散が大きなカテゴリではより詳細に出力分布を求めることができると考えられる。混合数の自動決定は音素カテゴリ別の場合と、さらにHMMの状態ごと別に決

定する方法の2種類の実験をおこなった。その結果音素認識実験において、状態毎の自動決定、カテゴリ毎の自動決定、混合数均一の順で高い認識率が得られた。

次に混合連続HMMを用いた不特定話者の文節認識の検討を行なった。HMMの混合数は自動決定の場合と、従来どおりの混合数均一の場合とで比較した。実験はオープン話者で、２７９文節を一般文法を用いて行なった。この結果、両者で81.09%の認識率が得られたが、混合数自動決定の場合では混合数15で、混合数均一の場合は混合数20でこの値が得られたため、混合数自動決定の場合はパラメータの削減が可能であることが分かった。

話者混合逐次状態分割法を用いた不特定話者音声認識 (平成４年４月〜平成４年９月)

不特定話者音声認識性能向上のため、従来用いていた音素モデルに代わり、音素コンテキスト依存モデルを用いたシステムの構築を行なう。またこの不特定話者用のモデルを初期モデルとした高速な話者適応法の開発をおこなう。

本研究では比較的少量の学習データによって不特定話者音素モデルを生成するための原理として話者混合法を提案し、この原理を逐次状態分割法 (SSS) で生成された音素コンテキスト依存モデルと組み合わせることにより連続音声認識を行なった。アルゴリズムの有効性を検討するために不特定話者音素認識実験及び、文節認識実験をおこなった。文節認識実験の結果従来法の不特定話者HMM-LR法と比較して7.0%の認識率の向上を得た。

話者混合原理に基づいて1秒以下の非常に短い発話で動作する話者適応方式として話者重み学習法を開発した。この方式では学習パラメータが話者重みで、これをBawm-Welchアルゴリズムにより学習する。学習パラメータが少ないため非常に高速に話者適応が行なわれる。本手法を用いて0.6秒の単語発声で4.1%の認識率の向上を得た。

以上の話者適応法において、認識率の低下なしに計算量の削減をする方法として話者プルーニング法を提案する。この方法では話者重み学習の結果、重みがある確率以下になった話者モデルを枝刈りする。本手法を用いることにより、認識率の低下なしに混合連続出力分布の混合数を50〜92%削減することができた。

動的話者適応の研究 (平成４年10月〜平成５年３月)

本研究は音声が入力されるごとに性能が向上する動的話者適応を目指したものであり、その手始めとして、入力音声のデータ量に応じて話者適応手法を自動的に切替え、そのデータ量に対し最適な手法を選択する方法について検討した。

話者混合法によって作成された混合連続HMnet(CM-HMnet)をベースに、話者混合重み学習法・混合重み学習法・ベクトル場平滑化方式の3種類の話者適応法を組み合わせ適応語数に応じて、手法を切替える方法について検討した。この方法をとることにより、ベクトル場平滑化方式で見られた少数適応語数における認識率の減少を防げることが分かったが、一方話者により切替え時点が異なるという問題が生じた。

ヒューリスティックな適応手法の切替えによって生じる問題点を解決するため、HMnetの尤度を用いた適応手法の選択について検討した。この結果尤度を用いて適応手法を自動的に選択できることが分かった。またこのような選択手法を用いることにより、3種類のいずれの適応手法よりも高い認識率が得られることが分かった。

木構造話者クラスタリングによる話者適応 (平成5年4月〜平成5年9月)

本研究は少数の適応データで性能が向上する話者適応を目指したものであり、多数の話者の音素モデルセットを木構造にクラスタリングし、クラスタリング木を探索し、最大尤度を出力する話者モデルセットを選択することによる高速な話者適応法について検討した。

話者ごとの音素モデルセットとして、ここではATRで提案されたHidden Markov Network(HMnet)を使用した。クラスタリングは階層的に繰り返し行なうことで、木構造を作成した。ここでモデルのクラスタリングを行なう場合重要なモデル間の距離の定義について検討した。従来Kullback情報量による確率モデル間の距離の定義が提案されているが、ここではさらにBhattacharrya距離との比較を行なった。その結果性能は同等で計算量の点でBhattacharrya距離が優れていることがわかった。

話者適応は以上の方法で作成された木構造を探索することによって行なう。探索はモデルの出力尤度を基準としておこなった。尤度計算の方法としては、話者混合法による尤度計算と、混合分布のうち最大尤度を用いる方法とを検討したが、それほど差がないことがわかった。

木構造話者クラスタリングによる不特定話者音声認識及び合成による不特定話者音素モデル作成法 (平成5年10月～平成6年3月)

これまで行なってきた木構造話者クラスタリング法の、教師なし話者適応への応用について検討した。教師なし話者適応は、入力音声を一度認識系により認識を行ない、その認識結果をフィードバックすることにより行なう。評価の結果教師つきに匹敵する認識率が得られることが分かった。

さらに、木構造話者クラスタリングを不特定話者音声認識に適応した。これは1発話のみの音声データで教師なし話者適応を行なうことにより実現する。評価の結果、従来の手法に比べ高い認識率が得られた。

従来話者数が増加した場合モデル作成に計算コストがかかるという問題があった。この問題に対し、パラメータの再推定をすることなく、特定話者音素モデルを合成し不特定話者音素モデルを作成する方法を提案した。

MITにおける研究 (平成6年4月～平成6年9月)

4月下旬より9月上旬までMITのSLS(Spoken Language Systems) Groupに滞在し、MITのシステム関連の研究及び、話者適応の研究を行なった。MITはGALAXYと呼ばれる音声理解システムを構築しており、そのマルチリンガル化を担当した。また話者適応は少ないサンプルで動作する適応手法について検討し、TIMIT英語音声データベース上での評価を行なった。

MITのSLS Groupでは、現在GALAXYシステムと呼ばれる音声理解システムを構築中である。このシステムの特徴としては、マルチドメインに対応し現在のところ、NYNEXイエローページによる市内案内、天気情報、米国航空会社の航空チケット予約ができるようになっている。音声の理解部は各国語共通になっており、現在英語およびスペイン語の入出力が可能である。今回はさらに日本語への対応を目指し日本語の出力部分を担当した。このために日本語の辞書および変換テーブルを用意し、音声理解した結果であるセマンティックフレームを日本語に変換する。その出力を日本語合成器に渡し日本語の発声を得る。助詞などの扱いに多少の難があるが、おおむねセマンティックフレーム出力は言語に依らない情報を出力していることが分かった。

以前提案した話者重み学習法による話者適応では、入力話者が標準話者から遠い場合効果に限界があることが分かった。そこで適応データにより学習した話者重みに従い、出力分布を内挿し、新たな出力分布を求めることにより話者

適応する方法について検討した。その結果従来法より効果的であることが分かった。しかし認識率の向上が少ないため原因を調べたところ、音素により効果にばらつきがあるためだと分かった。

自由発話音声データの発話様式適応による評価 (平成 6 年 9 月〜平成 7 年 3 月)

不特定話者の自由発話音声認識に向けた第一段階として、音素認識実験による検討を行なった。まず文節発声で学習された不特定話者用 HMnet を初期モデルとし、MAP-VFS 話者適応法を用いて、話者、発話様式、話者＋発話様式の 3 種類で適応した場合の比較を行なった。

その結果 MAP-VFS での発話適応で良好な結果が得られることが分かった。不特定話者の自由音声データでの音素認識率として 72.5% が得られた。また自由発話音声の認識においても、個人性が大きな問題となることが分かった。

C.2 著者業績

音響学会

Kosaka92ASJ03 小坂 哲夫, 嵯峨山 茂樹: “不特定話者を対象とした混合連続分布 HMM 音声認識における混合数の検討,” 日本音響学会平成 4 年度春季研究発表会講演論文集, 2-Q-20, pp. 197–198 (1992.03).

Kosaka92ASJ10a 小坂 哲夫, 嵯峨山 茂樹: “混合連続分布 HMM 音素モデルの構造自動決定法の検討,” 日本音響学会平成 4 年度秋季研究発表会講演論文集, 2-1-1, pp. 79–80 (1992.10).

Kosaka92ASJ10b 小坂 哲夫, 鷹見 淳一, 嵯峨山 茂樹: “話者混合 SSS による不特定話者音声認識,” 日本音響学会平成 4 年度秋季研究発表会講演論文集, 2-5-9, pp. 135–136 (1992.10).

Takami92ASJ10b 鷹見 淳一, 小坂 哲夫, 嵯峨山 茂樹, “話者方向を加えた逐次状態分割法 (SSS) による話者共通隠れマルコフ網の生成,” 音講論集, 3-1-8 (1992.10).

Yamaguchi92ASJ10b 山口耕市, 永井明人, 鷹見淳一, 大倉計美, 小坂哲夫, 福沢圭二, 加藤喜永, 北研二, Harald Singer, 村上仁一, 杉山雅英, 嵯峨山茂樹, 服部浩明, 小森康弘, 沢井秀文, 花沢利行, 中村哲, 甲斐充彦, 南泰浩, 川端豪, 鹿野清宏: "ATREUS: ATR における連続音声認識諸方式の比較," 日本音響学会平成 4 年度秋季研究発表会講演論文集, 2-Q-5, pp. 181–182 (1992.10).

Kosaka93ASJ03 小坂 哲夫, エドワード・ウイレムス, 鷹見 淳一, 嵯峨山 茂樹, "複数の話者適応に基づく動的話者適応," 音講論集, 2-4-9, pp.35-36 (1993-03).

Kosaka93ASJ10 小坂 哲夫, 松永 昭一, 嵯峨山 茂樹, "木構造話者クラスタリングを用いた話者適応," 音講論集, 2-7-14, pp.97-98 (1993-10).

Kosaka94ASJ03 小坂 哲夫, 松永 昭一, 嵯峨山 茂樹, "不特定話者連続音声認識における木構造話者クラスタリング," 音講論集, 3-7-7, pp.101-102 (1994-03).

Kosaka94ASJ10 小坂 哲夫, 松永 昭一, 倉岡 幹雄, "クラスタリング手法を用いた不特定話者モデル作成法," 音講論集, 1-R-12, pp.215-216 (1994-10).

Tonomura94ASJ10 外村 政啓, 小坂 哲夫, 松永 昭一, "最大事後確率推定法を用いた移動ベクトル場平滑化話者適応方式," 音講論集, 2-8-20, pp.77-78 (1994-10).

Kosaka95ASJ03 小坂 哲夫, 松永 昭一, "発話／話者適応による自由発話音声中の音素認識," 音講論集, 2-5-4, pp.37-38 (1995-03).

Tonomura95ASJ03 外村 政啓, 小坂 哲夫, 松永 昭一, 門田 暁人, "MAP-VFS 話者適応法における平滑化係数制御の効果," 音講論集, 2-5-6, pp.41-42 (1995-03).

音声研究会

Kosaka92SP09 小坂 哲夫, 鷹見 淳一, 嵯峨山 茂樹: "話者混合 SSS による不特定話者音声認識と話者適応," 電子情報通信学会技術研究報告, SP92-52, pp. 17–24 (1992.09).

Willems92SP12 ウィレムス, 小坂 哲夫, 鷹見 淳一, 嵯峨山 茂樹: "音声認識のための隠れマルコフ網の動的話者適応法," 信学技報, SP92-102, pp. 57–64, (Dec. 1992).

Nagai93SP1 永井明人, 山口耕市, 鷹見淳一, 大倉計美, 小坂哲夫, 福沢圭一, 加藤喜永, Harald Singer, 村上仁一, 杉山雅英, 嵯峨山茂樹, 保坂順子, 森元逞, 北研二 (徳島大学), 服部浩明 (日本電気), 小森康弘 (キヤノン), 沢井秀文 (リコー), 花沢利行 (三菱電機), 中村哲 (シャープ), 甲斐充彦 (豊橋技科大), 南泰浩, 川端豪, 鹿野清宏 (以上 NTT), 樽松明, "ATR における連続音声認識システム 'ATREUS' の諸方式と性能," 信学技報, SP92-122, pp. 51–58, (1993.1).

Kosaka93SP12 小坂 哲夫, 松永 昭一, 嵯峨山 茂樹: "話者適応のための木構造話者クラスタリング," 電子情報通信学会技術研究報告, SP93-110, pp. 49–54 (1993.12).

Tonomura94SP10 外村 政啓, 小坂 哲夫, 松永 昭一: "最大事後確率推定法と移動ベクトル場平滑化法を統合した話者適応方式," 電子情報通信学会技術研究報告, SP94-51, pp. 25–30 (1994.10).

Kosaka95SP01 小坂 哲夫, 松永 昭一, 倉岡 幹雄: "話者クラスタリング手法を用いた不特定話者音素モデルの作成," 電子情報通信学会技術研究報告, SP94-80, pp. 9–16 (1995.01).

ATR Workshop

Matusnaga93ATR11 S. Matsunaga, H. Singer, H. Lucke, H. Sakamoto, J. Murakami, J. Takami, K. Yamaguchi, R. Isotani, T. Kosaka, Y. Miyazawa and Y. Sagisaka: "Speech Recognition for Spontaneous Language Translation at ATR," Proc. of ATR International Workshop on Speech Translation, (1993. 11).

国際会議

Kosaka92SST12 Tetsuo Kosaka, Shigeki Sagayama: "An Algorithm for Automatic HMM Structure Generation in Speech Recognition," Proc. of Fourth Australian International Conference on Speech Science and Technology pp. 104–109, (1992.12).

Sagayama92SST12 S. Sagayama, M. Sugiyama, K. Ohkura, J. Takami, A. Nagai, H. Singer, H. Hattori, K. Fukuzawa, Y. Kato, K. Yamaguti, T. Kosaka and A. Kurematsu: "ATREUS: Continuous Speech Recognition Systems at ATR Interpreting Telephony Research Laboratiories, " Proc. of Fourth Australian International Conference on Speech Science and Technology, pp. 324–329, (1992.12).

Kosaka93ICASSP Tetsuo Kosaka, Jun-Ichi Takami, Shigeki Sagayama: "Rapid Speaker Adaptation Using Speaker-Mixtue Allophone Models Applied to Speaker-Independent Speech Recognition," Proc. of ICASSP93, pp. 570–573 (1993.04).

Kosaka94ICASSP Tetsuo Kosaka, Shigeki Sagayama: "Tree-Structured Speaker Clustering for Fast Speaker Adaptation," Proc. of ICASSP94, pp. 245–248 (1994.04).

Kosaka95ICASSP Tetsuo Kosaka, Shoichi Matsunaga, Mikio Kuraoka: "Speaker-Independent Phone Modeling Based on Speaker-Dependent HMMs' Composition and Clustering," Proc. of ICASSP95 (1995.04 掲載予定).

Tonomura95ICASSP Masahiro Tonomura, Tetsuo Kosaka, Shoichi Matsunaga: "Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Probability Estimation," Proc. of ICASSP95 (1995.04 掲載予定).

Kosaka93Eurpspeech Tetsuo Kosaka, Edward Willems, Jun-Ichi Takami, Shigeki Sagayama: "A Dynamic Approach to Speaker Adaptation of

Hidden markov Networks for Speech Recognition," Proc. of Eurospeech93, pp. 363–366 (1993.09).

Kosaka94ICSLP Tetsuo Kosaka, Shoichi Matsunaga, Shigeki Sagayama: "Tree-Structured Speaker Clustering for Speaker-Independent Continuous Speech Recognition," Proc. of ICSLP94, pp. 1375–1378 (1994.09).

論文

Kosaka94IEICE02 小坂 哲夫, 鷹見 淳一, 嵯峨山 茂樹: "話者混合逐次状態分割法による不特定話者音声認識と話者適応," 電子情報通信学会論文誌 A, Vol. J77-A, No. 2, pp. 103–111 (1994.02).

Kosaka95IEICE01 小坂 哲夫, 松永 昭一, 嵯峨山 茂樹: "木構造話者クラスタリングを用いた話者適応," 電子情報通信学会論文誌 D-II, Vol. J78-D-II, No. 1, pp. 1-9 (1995.01).

Kosaka95IEICE06 Tetsuo Kosaka, Shigeki Sagayama: "Automatic Determination of the Number of Mixture Components for Continuous HMMs Based on a Uniform Variance Criterion," IEICE Transactions Inf. and Syst., Vol. E78-D, No. 6 (1995.06 掲載予定).

参考文献

[1] 中村哲：“音声認識における話者適応”， 信学技報 SP94-3,pp.17–24 (1994.5).

[2] 樽松明他：“自動翻訳電話”, ATR国際電気通信基礎技術研究所編 (1994.1).

[3] C.-H. Lee, C.-H. Lin and B.-H. Juang: “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” IEEE Trans. on Signal Processing, Vol. 39, No. 4, pp. 806-814 (1991.04).

[4] 村上仁一, 嵯峨山茂樹：“自由発話音声認識における音響的および言語的な問題点の検討”, 信学技報 SP91-100, pp.71–78 (1991).

[5] 鷹見淳一， 嵯峨山茂樹：“音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成”, 信学技報 SP91-88,pp.57–64 (1991.12).

[6] Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K.: “ATR HMM-LR Continuous Speech Recognition System”, Proc. of ICASSP90, S 2.4, pp.53–56 (1990).

[7] 永井明人, 鷹見淳一, 嵯峨山茂樹：“逐次状態分割法 (SSS) と音素コンテキスト依存 LR パーザを統合した SSS-LR 連続音声認識システム,” 信学技報, SP92-33(1992-06).

[8] 大倉計美, 杉山雅英, 嵯峨山茂樹: “混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式,” 信学技報, SP92-16, pp. 23-28 (1992-06).

[9] 服部浩明, 嵯峨山茂樹: “少数語彙による移動ベクトル場平滑化話者適応方式の文節認識による評価,” 音響講論集, 2-Q-15(1992.03)

[10] 鷹見淳一, 嵯峨山茂樹：“隠れマルコフ網 (HM-Net) を用いた話者適応,” 音響講論集, 1-1-8, pp. 15-16 (1992.03).

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

[12] L. E. Baum and J. A. Eagon: “An inequality with applications to statistical prediction for functions of Markov processes and to a model fr ecology,” Bull. Am. Math. Soc., 73, (1967).

[13] Y.L.Chow, M.O.Dunham, O.A.Kimball, M.A. Krasner, G.F.Kubala, J.Makhoul, P.J.Price, S.Roucos and R.M.Schwartz, “BYBLOS: The BBN Continuous Speech Recognition System”, ICASSP'87, pp.89-92 (1987).

[14] K.F.Lee, H.W.Hon, M.Y.Hwang, S.Mahajan and R.Reddy, “The SPHINX Speech Recognition System”, ICASSP'89, pp.445-448 (1989).

[15] X.D.Huang, K.F.Lee, H.W.Hon, M.Y.Hwang, “Improved Acoustic Modeling with the SPHINX Speech Recognition System”, ICASSP'91, 10.S5.24 (1991.5).

[16] 伊藤克亘, 田中穂積, 速水悟, “音素文脈依存音韻 HMM を用いた連続音声認識”, 音響学会講演論文集, 1-5-21, pp.41-42 (1991.10).

[17] 永井明人, 北研二, 嵯峨山茂樹, ”HMM-LR 法における音素文脈依存型 LR パーザの検討”, 音響学会講演論文集, 3-8-16, pp.125-126 (1990.9).

[18] 永井明人, 嵯峨山茂樹, 北研二, ” 音素コンテキスト依存型 LR テーブルの生成アルゴリズム”, 音響学会講演論文集, 3-5-2, pp.91-92 (1991.3).

[19] 永井明人, 菊池英明, 嵯峨山茂樹, 北研二, “文脈自由文法から音素コンテキスト依存文法への変換アルゴリズム”, 音響学会講演論文集, 3-1-6, pp.81-82 (1992.3).

[20] 武田, 黒岩, 山本：“音素イベントHMMによる単語及び連続音声の認識”, 音講論 2-P-1,pp.139–140 (1991.10).

[21] 松尾, 石亀：“トポロジーの学習を考慮したHMMによる音素認識の検討”, 音講論 2-P-2,pp.141–142 (1991.10).

[22] Kamp, Y., “State Reduction in Hidden Markov Chains Used for Speech Recognition,” IEEE Trans. on ASSP, Vol. 33, No. 4, pp.1138-1145, Oct. 1985.

[23] 坂元、石黒、北川：“情報量統計学”, 共立出版, (1983.1).

[24] 池田思朗：“HMMの構造探索による音素モデルの生成”, 信学技報, SP93-26 (1992.06).

[25] 山口, 嵯峨山：“混合連続分布型HMMを用いたHMM-LR連続音声認識,” 音響学会講論集, 1-P-5 (1992.3).

[26] Rabiner, L. R., Juang, B.-H., Levinson, S. E. and Sondhi, M. M., “Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities,” AT&T Technical Journal, vol. 54, 6, pp. 1211-1234, Jul. 1985.

[27] 野村俊之, 板倉文忠：“連続数字音声認識におけるHMMの状態数及び混合数について”, 音響学会講論集, 3-P-22 (1991.3).

[28] 鷹見淳一, 小坂哲夫, 嵯峨山茂樹：“話者方向を加えた逐次状態分割法(SSS)による話者共通隠れマルコフ網の生成,” 音響講論集, 3-1-8, pp.155-156 (1992-10).

[29] 南, 山田, 鹿野, 松岡：“番号案内を対象とした大語い連続音声認識アルゴリズム,” 信学論 A, Vol. J77-A, No. 2, pp. 190-197 (1994.2).

[30] 小坂 哲夫, 鷹見 淳一, 嵯峨山 茂樹: “話者混合逐次状態分割法による不特定話者音声認識と話者適応,” 信学論 A, Vol. J77-A, No. 2, pp. 103-111 (1994.2).

[31] 南, 松岡, 鹿野：“音韻ラベルを用いない HMM 評価法とそれを用いた連続音声用 HMM の評価,” 信学論 A, Vol. J77-A, No. 2, pp. 267-273 (1994.2).

[32] 菅村, 相川, 鹿野, 好田: “SPLIT, マルチテンプレート法による不特定話者単語音声認識,” 音声研資, S82-64 (1982.12).

[33] Y. Linde, A. Buzo and R. M. Gray: “An algorithm for vector quantizer design”, IEEE Trans. Commun., COM-28, 1, pp. 84–95 (1980.1).

[34] B.-H. Juang and L. R. Rabinaer: “A Probabilistic Distance Measure for Hidden Markov Models,” AT&T Technical Journal Vol. 64, No. 2, (1985.2).

[35] P. D'Orta, M. Ferretti and S. Scarci: “Phoneme Classification for Real Time Speech Recognition of Italian,” Proc. of ICASSP87, pp. 81-84 (1989).

[36] Keinosuke Fukunaga: “Introduction to Statistical Pattern Recognition (Second Edition),” Academic Press, Inc., San Diego, (1990).

[37] 石川: “新統計学,” 槙書店 (1991.4).

[38] Cambridge Univ. Eng. Dept. Speech Group and Entropic Res. Labs. Inc. “HTK: Hidden Markov Model Toolkit V1.5,” Entropic Res. Labs. Inc. (1993.9)

[39] 渡辺, 吉田, 古賀：“半音節を単位とした HMM を用いた大語い音声認識,” 信学論 D-II, Vol. J72-D-II, No. 8, pp. 1264-1269 (1989.8).

[40] K.-F. Lee and H.-W. Hon: “Large-Vocabulary Speaker Independent Continuous Speech Recognition Using HMM,” Proc. ICASSP88, pp. 123-126 (1988)

[41] 渡辺隆夫, 磯谷亮輔, 塚田聡：“半音節を単位とする HMM を用いた不特定話者音声認識,” 信学論 (D-II), J75-D-II, 8, pp. 1281-1289 (1992.08).

[42] 南泰浩, 松岡達雄, 鹿野清宏：“不特定話者連続音声データベースを用いたHMMの連結学習,” 音響講論集, pp. 9-10 (1992.03).

[43] 服部浩明, 中村哲, 鹿野清宏：“話者適応における複数話者への重み付け,” 音響講論集, 2-3-1, pp.51-52 (1990.03).

[44] R.Schwartz, Y.-L. Chow and F.Kuala: “Rapid Speaker Adaptation using a Probabilistic Spectral Mapping,” Proc. ICASSP87, pp. 633-636 (1987.04).

[45] 松岡, 鹿野：“混合ガウス分布不特定話者HMMをベースとした重み係数による話者適応化法,” 音講論, 1-1-6(1992.3).

[46] 今村明弘: “統計的話者分類による話者適応形HMM音声認識,” 信学技報, SP91-17, pp. 87-93 (1991.06).

[47] 杉山雅英, 鹿野清宏：“母音標準パタンの教師なし学習法,” 音響学会音声研資, S83-48, pp.373-379 (1983-12)

[48] 小坂哲夫, エドワード・ウイレムス, 鷹見淳一, 嵯峨山茂樹: “複数の話者適応に基づく動的話者適応,” 音響講論集, 2-4-9, pp.35-36 (1993-03).

[49] 篠田浩一, 磯健一, 渡辺隆夫: “音声認識のためのスペクトル内挿を用いた話者適応化,” 信学論 (A), J77-A, 2, pp. 120-127 (1994-02).

[50] 平田好充, 中川聖一: “連続出力分布型HMMにおける話者適応化の日本語音韻認識による評価,” 信学技報, SP90-16 (1990-06).

[51] P. Niyogi and V. W. Zue: “Correlation Analysis of Vowels and Their Application to Speech Recognition,” Proc. of Eurospeech91, pp. 1253-1256 (1991).

[52] 小坂哲夫, 牧野正三, 城戸健一: “男女声自動識別を用いた母音認識の検討,” 音響講論集, 1-9-12, pp. 23-24 (1984-10).

[53] 松尾広, 牧野正三, 城戸健一: “自動性別判定を用いた母音・子音定常部の認識に関する検討,” 音響講論集, 1-5-14, pp. 27-28 (1987-10).

[54] ウィレムス, 小坂, 鷹見, 嵯峨山：“音声認識のための隠れマルコフ網の動的話者適応法,” 信学技報, SP92-102, (1992.12).

[55] 小坂, 松永, 嵯峨山: “話者適応のための木構造話者クラスタリング,” 信学技報, SP93-110 (1993.12).

[56] J.-L. Gauvain and C.-H. Lee: “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2 (1994)

[57] 越川、中川: “最大事後確率推定法を用いた連続出力分布型 HMM の適応化,” 音響学会誌,49,10,pp. 721-728(1993)

[58] 松岡、 C.-H. Lee: “最大事後確率推定法 (MAP 推定法) によるオンライン話者適応化,” 信学技報,SP93-133(1994.1).

[59] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973

[60] 外村 政啓, 小坂 哲夫, 松永 昭一, 門田 暁人: “MAP-VFS 話者適応法における平滑化係数制御の効果,” 音講論集, 2-5-6, pp.41-42 (1995-03).

[61] 鶴見, 中川: “最大事後確率推定法による連続出力分布型 HMM の教師なし話者適応化,” 信学技報, SP93-104, (1993.12).

[62] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund and M. A. Pryzbocki: “1993 Benchmark Tests for the ARPA Spoken Language Program,” Proc. of ARPA Workshop on Human Language Technology (1994.3).

[63] 速水, 伊藤, 田中：“音声対話システムの構築とそれを用いた会話音声収集,” 信学技報, SP91-101, pp. 79–86 (1991.12).

[64] 吉岡, 南, 鹿野：“電話番号案内を対象としたマルチモーダル対話システムの作成と音声入力の評価,” 信学技報, SP93-128, pp. 1–8 (1994).

[65] 武田, 黒岩, 井ノ上, 野垣内, 山本, 庄境, 尾和, 高橋, 松本: “連続音声認識に基づく内線番号案内システムの試作,” 音講論集, pp. 79–80 (1993.3).

[66] 竹林, 坪井, 金沢, 貞本, 山下, 瀬戸, 永田, 新居, 橋本, 新地: “不特定話者音声対話システム TOSBURG の開発,” 音講論集, pp. 135–136 (1992.3).

[67] 村上: ” 自由発話音声認識における音響的および言語的な問題点の検討,” ATR Technical Report, TR-IT-0032 (1993.12).

[68] 外村, 小坂, 松永: “最大事後確率推定法と移動ベクトル場平滑化法を統合した話者適応方式,” 信学技報, SP94-51, pp. 25–30 (1994.10).

[69] I. S. Bridle, et al.: “An algorithm for connected word recognition,” Proc. of ICASSP82, pp. 899-902 (1982).

[70] 大脇, ハラルド, 鷹見, 樽松: “音素配列構造の制約を用いた音素タイプライタ,” 信学技報, SP93-113 (1993.12).

[71] 小坂, 松永, 倉岡: “クラスタリング手法を用いた不特定話者モデル作成法,” 音講論集, 1-R-12, pp.215-216 (1994-10).