修士学位論文

隠れマルコフ網を用いた話者適応手法に関する研究

東北大学大学院工学研究科
電気・通信工学専攻
阿部 俊朗

目 次

図目次

表目次

第1章

序論

1.1 研究の背景

現在音声認識技術はその認識性能の向上により、様々なシステムのインターフェースとして使われている。例えば、カーナビゲーションシステムや駅の券売機、最近では腕時計型の PHS などがある。これらのシステムに音声認識が使われる背景として、ひとつにはシステムの操作を手でおこなうことが困難であるという点がある。カーナビゲーションシステムでは、手での操作は安全面に問題がある。また腕時計型 PHS においては構造上、手で操作しづらい。もうひとつの背景としては、音声が人間にとって訓練などを必要としない非常に使いやすいコミュニケーションの道具であるという点である。以上のような理由から音声認識を使ったシステムは利用者にとって簡単で扱いやすいインターフェースであると言える。

しかし現在実用化されているシステムには話者や語彙などの制約があり、今後様々なシステムに音声入力インターフェースを実現するためには、これらの制約がない音声認識システムの開発が必要不可欠である。

1.1.1 話者適応の必要性

音声認識の研究は前節で述べたような理由から、大語彙、不特定話者、連続音声を対象としたものが現在では中心となっている。その認識性能は高いものが得られているが、更なる性能の向上のためには様々な問題を解決する必要がある。

そのうち不特定話者を対象とした認識システムを構築するためには、話者の個人性による音声の音響的特徴の変動が非常に大きな問題となってくる。話者の個人性が生じる原因には次のようなことが挙げられる。

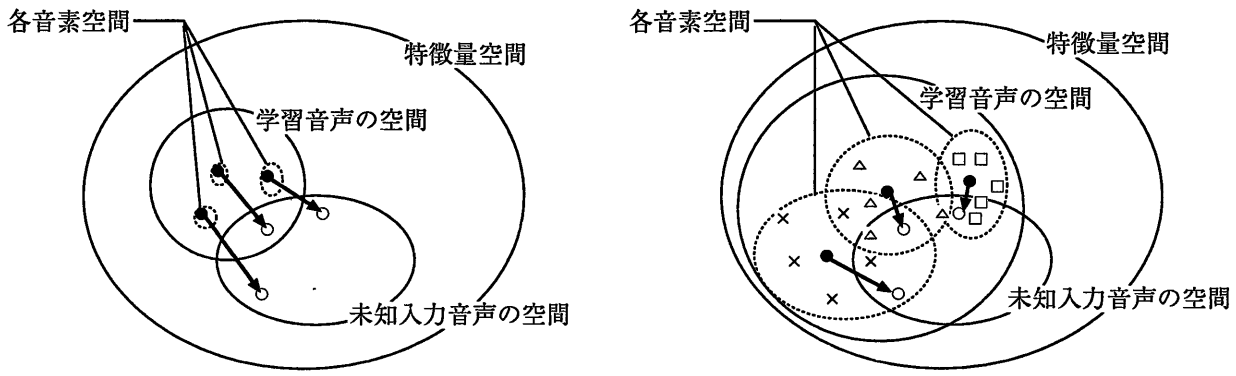

図 1.1: 特定話者モデルと不特定話者モデル

- 発声器官の違い

  音声は声道などの発声器官によって生成される。しかし声道の長さや声帯の形などに個体差があるため音声に個人差が生じる。

- 社会的環境の違い

  生まれた場所や生活環境によってなまりなどが生じ、音声に話者の個人差が生じる。

- 発話環境の違い

  同一人物が発話しても発話の内容によって調音結合が生じたり、イントネーションが異なるために音声に音響的な変動が生じる。また長い時間経過した場合では同一人物でも音声に違いがあると言われる。

このような話者の個人性の問題を解決する1つの方法として、音声認識で用いる音響モデルに、多数の話者が発声したサンプルを用いて学習をおこなう不特定話者モデルを使う方法がある。

これを1人の話者が発声したサンプルで学習した特定話者モデルと比較し、その特徴を考えてゆく。特定話者モデルの場合、学習と同一人物で認識をおこなった場合にかなり高精度な認識ができる。しかし学習と違う話者で認識をおこなった場合、認識対象である話者と学習で用いた話者の音声の特徴が類似しているかどうかで、その性能は大きく異なる。よって認識対象が特定されていないとき、このようなモデルは不向きである。

一方不特定話者モデルの場合は、認識対象の話者が学習で用いた話者であるときの性能は特定話者モデルに比べ劣るものの、認識対象の話者が学習で用いた話者でない場合でも認識性能が大きく変動することはない。なぜなら不特定話者モデルは多数の話者のそれぞれの特徴を学習させたため、話者性による音声パターンの変動もある程度吸収できるから

である。しかしそれは話者による音素の特徴を幅広くとっているのであって、認識対象である人間の個人性を端的に表わしているということではない。

これをイメージするために図1.1を用いて説明する。図中の点線で囲まれた部分が学習音声中で得られた各音素(2.2.1節)の特徴を表わす分布であり、黒丸がその分布の中心である。認識時にサンプルがこの点線の中にあればその音素が認識結果となる。不特定話者モデルでの△や×や□は学習サンプルの各話者の音素の分布の中心を示している。また白丸が認識対象となる話者の音素サンプル、矢印は学習話者と認識対象の話者で対応する音素を表わす。

特定話者モデルの場合、ひとつの音素に対してそれぞれのサンプルの特徴があまり異ならないので、音素空間は小さい。よって各音素空間が重なることが少ないため、認識対象と学習に用いた話者が同じときには高い認識性能を持つ。しかし認識対象となる話者の入力音声との特徴が異なると、入力された音素は実際に対応する音素空間から外れる可能性も高く、認識性能が落ちる。それに対して不特定話者モデルの場合はそれぞれの音素の特徴が話者により異なるので各音素空間は大きい。そのため音素空間が互いに重なることがあるため、特定話者モデルを用いたときより認識性能は劣る。だが認識対象となる話者の入力音声は対応する音素空間内にあることが多いので、認識性能が話者によって左右されることはあまりない。

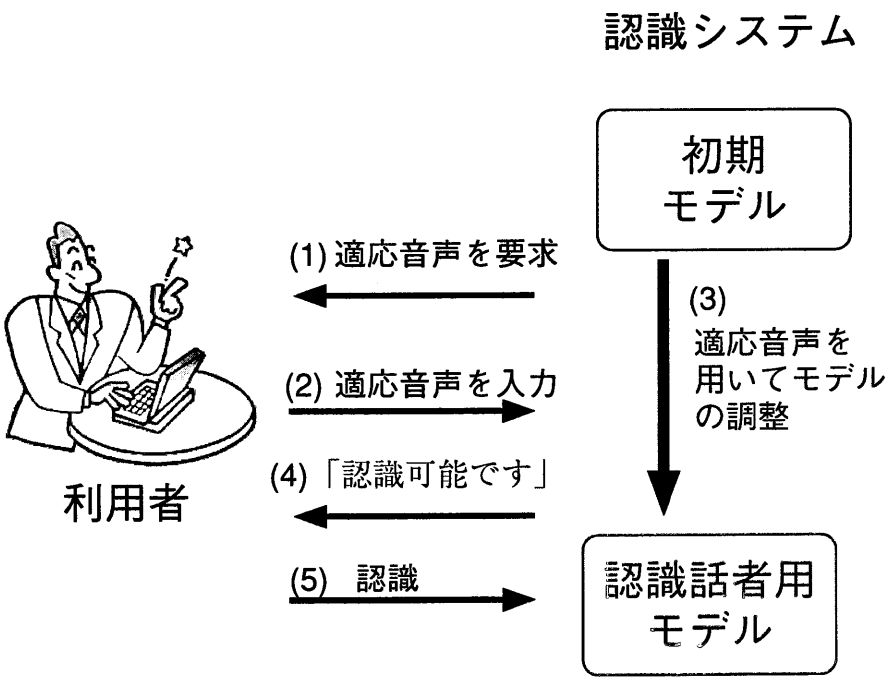

図 1.2: 話者適応の概要

しかし不特定話者モデルでも、話者性を完全に取り除くことはできず、更に高い認識性能を得るためには、話者の個人性を吸収するより、話者の個人性を端的に表わすようにモデルを調整して話者性を陽に考慮する必要がある。

そこで、話者の個人性を陽に考慮し認識性能の低下を解決する手段として話者適応手法が研究されている (例えば [8])。話者適応をシステムに実装した場合の処理の概略は図 1.2 のようになる。まず、多数の話者の多量の音声データで信頼度の高い不特定話者モデルをあらかじめ学習しておく。そして認識前に対象となる話者に決まった言葉を発話してもらう。内容既知の音声が入力されたら、モデルを認識対象の特徴に合うように調整し、そのモデルを用いて認識をおこなう。話者適応に必要な音声サンプルの量は数単語から数文章である。これらの話者適応法は、あらかじめ発話内容の既知の音声を必要とするので教師あり話者適応と言われる。一方、認識前に話者が発声した音声を必要としない話者適応法もある。 これを教師なし話者適応と言う。

利用者にとっては負担の少ない教師なし話者適応を用いることが良いのだが、その認識性能は教師あり話者適応に比べるとかなり悪い。そこで本研究では教師あり話者適応法に主眼をおく。

1.1.2 教師あり話者適応手法の現状

現在、教師あり話者適応の手法には 2 つの大きな流れがある。ひとつは与えられた適応音声によりモデルのパラメータを調整する方法である。もうひとつはあらかじめ用意した不特定話者モデルから認識対象となる話者を選択する方法である。これらの代表的な手法を以下に示す。

- モデルパラメータの調整による方法

  ＜特徴＞

  多量の適応音声が必要

  ＜代表的手法＞

  “移動ベクトル場平滑化法” (大倉ら)[8] [9] 1992

  適応音声に存在するモデルのパラメータを再学習し、存在しなかったモデルのパラメータは再学習したモデルの情報を用いて補間・平滑化することで話者適応をおこなう手法。

  “事後確率推定法を用いた移動ベクトル場平滑化法” [10] [11] 1994

  パラメータの推定に事後確率を用いた移動ベクトル場平滑化法。適応音声が少ない場合には 移動ベクトル場平滑化法よりも話者適応の能力が高い。

- 話者選択による方法

＜特徴＞

少量の適応音声で話者適応が可能

初期モデルに多量の話者サンプルが必要

＜代表的手法＞

“木構造話者クラスタリングを用いた話者適応法” (小坂ら) [16] 1994

上層に不特定話者のモデル、下層に特定話者のモデルを有した木構造で構成される。この木構造からごく少量の適応音声を用いることで最適なモデルを選択し認識をおこなう手法。

1.2 研究の目的

本研究では、話者選択に基づく話者適応手法が高速処理可能であるという点に注目する。話者選択による話者適応法では、話者の個人性をそれぞれの話者に一律に考慮することで話者適応をおこなう。しかしそれぞれの話者において、詳細に話者の個人性を考えることで、更に高性能な話者適応が可能となると考えられる。本研究では、以上のような概念のもとで高速高性能話者適応手法を提案し認識性能の向上を目指す。

1.3 本論文の構成

本論文の構成は以下のとおりである。

第1章 序論

研究の背景や目的、論文の構成を述べる。

第2章 音響類似性に基づく隠れマルコフ網における移動ベクトル場平滑化法

隠れマルコフ網を用いた音声認識の処理の概要ならびに “移動ベクトル場平滑化法 (VFS)”[8][9] を用いた話者適応法の概要を述べる。また隠れマルコフ網生成アルゴリズムのひとつである SSS-free[7] を用いて生成された HMnet へ VFS を適応する手法を提案する。

第3章 音素毎の木構造話者クラスタリングに基づく話者適応法

木構造話者クラスタリングによる話者適応法 [16] における問題点について述べ、問題点を改善するアルゴリズムを提案する。

第4章 音素間の距離を用いた木構造話者クラスタリング

第3章で提案した“音素毎の木構造話者クラスタリングに基づく話者適応法”の問題点について述べ、その改善法を提案する。

第5章 結論

本研究の成果、今後の課題について述べる。

第2章

音響類似性に基づく隠れマルコフ網における移動ベクトル場平滑化法

2.1 はじめに

不特定話者音声認識は隠れマルコフモデル (HMM) またはそれを応用した隠れマルコフ網 (HMnet) によって高い認識性能を得ることができた。しかし不特定話者音声認識において更に高い認識性能を得るためには、話者の個人性の問題を改善する必要がある。この問題を改善するアプローチのひとつに、信頼度高く学習した不特定話者モデルから、認識対象の話者の音声を用いてモデルをその話者に適応化する話者適応がある。そのなかでもSSS [6] による HMnet に移動ベクトル場平滑化法 (VFS)[8] を適用した方法は、適応音声が多量に必要なものの非常に高い性能が示されている。

また VFS では 5 ～ 10 文章程度の音声サンプルによって話者適応が可能であるが、更に多量の音声サンプルを用いることで、擬似的に特定話者モデルの学習ができる。この学習で生成されたモデルは、初期モデルと同じネットワーク構造を持つ。第 3 章で述べる “木構造話者クラスタリングによる話者適応法” では多数の話者の特定話者モデルを用意して、そのモデルを用いてクラスタリングをおこない木構造を作成するが、全ての特定話者モデルのネットワーク構造が同じだとクラスタリングがおこないやすくなるので初期モデルの学習に VFS を用いる。

一方 SSS が学習データをあらかじめ分類しておく必要があるのに対して、そのような分類をせずに同じような効果をもたらす HMnet の生成アルゴリズム SSS-free [7] が提案されている。しかし SSS-free による HMnet においては、音素環境 (先行音素+当該音素+後続音素) と HMnet のパスが 1 対 1 で対応していないために VFS をそのまま適用することはできない。本研究では SSS-free で生成した特定話者 HMnet が SSS で生成した特定話者 HMnet よりも認識性能が高いという理由から SSS-free を用いた HMnet を研究の対象としている。そのため VFS を SSS-free による HMnet に適応することは不可欠である。

そこで本章では SSS-free に VFS を適用する手法を提案し、話者適応の観点から見た SSS-free の能力を論じる。また HMnet を用いた音声認識処理の概要についても述べる。

2.2　音声の特徴抽出

2.2.1　音声認識の認識単位

小語彙の音声認識では音響モデルの認識単位として単語などが使用された。しかし大語彙では単語を認識単位とした場合、モデルの種類が増加し各モデルに与えられる学習サンプルが少なくなるので信頼性の高いモデルを学習することができなくなる。大語彙連続音声認識において音響モデルの認識単位は以下を使用するのが一般的である。

音節

単独で発話の単位となりうる最小の単位。日本語の音節構造は特殊な例外を除き、次のように表わすことができる。

$$/\mathrm{CV}/ \quad /\mathrm{CVV}/ \quad /\mathrm{CV_N}/ \quad /\mathrm{CV_Q}/$$
$$/\mathrm{CSV}/ \quad /\mathrm{CSVV}/ \quad /\mathrm{CSV_N}/ \quad /\mathrm{CSV_Q}/$$

ただし /C/ は子音音素、 /V/ は母音音素、/S/ は半母音音素、/Q/は促音 (日本語の “っ” に対応)、/N/は撥音 (日本語の “ん” に対応) を示す。また /VV/ は同一母音音素の連続に限る。

音素

人間が知覚できる音の最小単位。ほとんどの音素はローマ字のアルファベット 1 文字と対応している。また言語、定義の仕方によって音素数は異なる。本研究で使用する音声データベースは 2 種類あり ATR 日本語音声データベース音韻バランス 503 文章は 47 種類 (表 2.1)、日本音響学会連続音声データベースは 34 種類である (表 2.2)。ATR 連続音声データベースにおける音素の分布を付録 C に示す。このデータベース

表 2.1: ATR 連続音声データベース音素一覧

N	**a**	**b**	by	cch	**ch**	**d**	dd	dy	**e**	**f**	ff
g	gy	**h**	he	hy	**i**	**j**	**k**	kk	kky	ky	**m**
my	**n**	ny	**o**	**p**	pp	ppy	py	**r**	ry	**s**	**sh**
ss	ssh	**t**	**ts**	tt	tts	**u**	**w**	wo	**y**	**z**	

を用いて音素認識実験をおこなう場合、促音等のサンプル数が少なすぎるために学習ができないことから表 2.1 の太字の 24 種類を対象にしている。

表 2.2: 日本音響学会連続音声データベース音素一覧

N	Q	a	a:	b	c	d	d'	
e	e:	f'	g	h	i	i:	ih	
j	k	m	n	ng	o	o:	p	r
s	t	t'	u	u:	uh	w	w'	z

音素環境

音素の前後の音韻環境を考慮したもの。音声は調音結合によって前後の音韻環境に物理的な変動が生じるためこのような定義をする。異音と言うこともあるが、音声学で用いる異音とは若干意味が異なる。

表 2.3: 音節、音素、音素環境の例

単語	学校					
音節	gaQ			koo		
音素	/g/	/a/	/Q/	/k/	/o/	/o/
音素環境	/%-g+a/	/g-a+Q/	/a-Q+k/	/Q-k+o/	/k-o+o/	/o-o+%/

表2.3に音節、音素、音素環境の例を示す。ここで % は無音区間を示す。またここで示す音素は、日本音響学会連続音声データベースの音素表記を用いている。

2.2.2　本研究の音響分析条件

音声認識をおこなう場合、音声波形をそのまま用いることはあまりなく、スペクトル変換などをしたものを用いるのが一般的である。本研究では表2.4の音響分析条件で特徴抽出をおこなっている。

2.2.2.1　ハミング窓

音声の情報をスペクトル変換するために時間窓を使い短時間毎に音声波形を切り出す。このとき矩形窓で切り出すと時間窓の最初と最後の部分がスペクトルに影響を及ぼすので、時間窓の最初と最後の部分が零になるような窓を使う。ハミング窓はそのような窓のひとつである。

$$h_{ハミング}(t) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi t}{T}\right) & (|t| \leq T) \\ 0 & (|t| > T) \end{cases} \tag{2.1}$$

表 2.4: 音響分析条件

分析条件	サンプリング周波数 12kHz
	16Bit 量子化
	20ms ハミング窓
	フレーム周期 5ms
特徴量ベクトル	$\log pow, cep(16), \Delta \log pow$
	$\Delta cep(16)$ からなる
	34 次元ベクトル

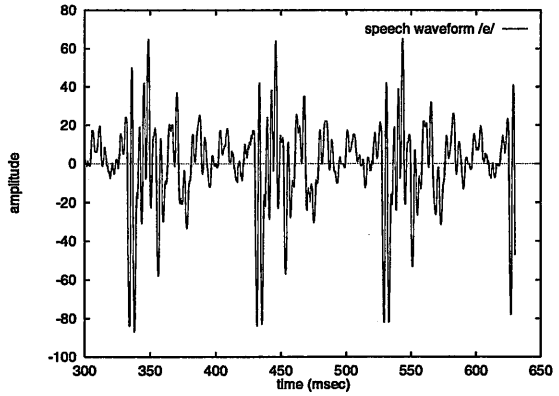

図 2.1: 定常部分の音声波形 / e /

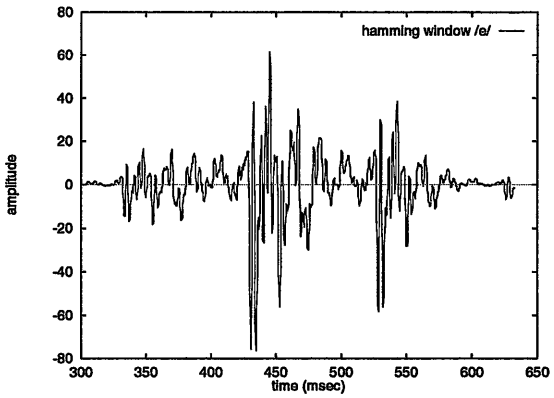

図 2.2: 窓掛け後の音声波形

2.2.2.2　線形予測分析法

音声波形の標本値の間には高い相関があることは実験的によく知られている。そこで、現時点での標本値 x_t とこれに隣接する過去の p 個の標本値との間に以下の式が成立すると仮定する。

$$x_t \cong \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} \tag{2.2}$$

このとき、その線形予測誤差の自乗平均を最小にするという条件で線形予測係数 $\{\alpha_i\}, i = 1, 2, \cdots, p$ を求めることを線形予測分析と言う。

N 個の音声サンプル列を用いた実際の解法については、共分散法と自己相関法の 2 つの方法がある。$\{x_i\}$ の系列が長く、定常である場合には両者の結果は同じであるが、系列が短く、時間的な変化を伴う場合には異なった結果を示す。また、共分散法による計算量は、自己相関法に比べ約 3 倍かかる。

2.2.2.3　ケプストラム

ケプストラム (cepstrum) は波形の短時間振幅スペクトルの対数の逆フーリエ変換として定義され、スペクトル包絡と微細構造を近似的に分離して抽出できる特徴を持つ。分離されたスペクトルの微細構造から音声波形のピッチ周波数 (基本周期) が、またスペクトル包絡からはホルマント周波数を求めることができる。ピッチ周波数、ホルマント周波数ともに音声の特徴を把むのに重要な量である。

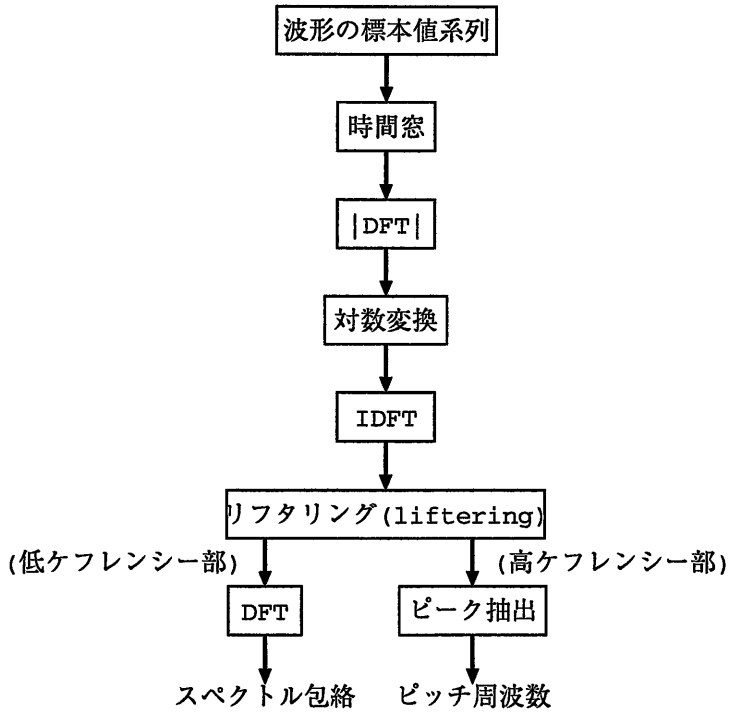

図 2.3: ケプストラム法によるピッチ周波数とスペクトル包絡の抽出

2.2.2.4　LPC ケプストラム

線形予測 (LPC) 分析法によって推定された全極型音声生成システム関数

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-1}} \tag{2.3}$$

を音声信号スペクトル密度と見なしたときのケプストラムを考える。図 2.3 の DFT 、対数変換、 IDFT をそれぞれ、両側 z 変換、複素対数、逆両側 z 変換で置き換えることで、ケプストラムの概念を複素ケプストラムに拡張する。系列 $x(n)$ の複素ケプストラムを $\hat{c}_n$ で表わすと、次の再帰式を得ることができる。

$$\hat{c}_1 = -\alpha_1 \tag{2.4}$$

$$\hat{c}_n = -\alpha_n - \sum_{m=1}^{n-1}\left(1-\frac{m}{n}\right)\alpha_m \hat{c}_{n-m}(1<n\leq p) \tag{2.5}$$

$$= -\sum_{m=1}^{p}\left(1-\frac{m}{n}\right)\alpha_m \hat{c}_{n-m}(p<n)$$

このケプストラムを LPC ケプストラムとよぶ。本研究では、16 次の LPC ケプストラム係数とそのデルタ係数、対数パワとそのデルタ係数からなる 34 次元ベクトルを用いる。

2.3　隠れマルコフ網 (HMnet)

2.3.1　隠れマルコフモデル

隠れマルコフモデル (Hidden Markov Model : HMM) とは、有限個の状態において、一定周期毎に状態を次々と遷移するとともにその遷移の際にラベルを 1 つずつ出力するという時系列パターンの確率有限状態オートマトンである。通常のマルコフモデルとの相異点は出力ラベル系列は観測できるが、どの状態をどのように遷移してきたのかを直接観測できないということである。また HMM は表 2.5 のように定義される。

表 2.5: HMM の定義

HMM $\lambda = (\boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \boldsymbol{F})$

$\boldsymbol{S}$: 状態の有限集合；$\boldsymbol{S}=\{s_i\}$

$\boldsymbol{Y}$: 出力シンボルの集合

$\boldsymbol{A}$: 状態遷移確率の集合；$\boldsymbol{A}=\{a_{ij}\}$; a_{ij}は状態s_iから状態s_jへの遷移確率。ここで$\sum_j a_{ij} = 1$

$\boldsymbol{B}$: 出力確率の集合；$\boldsymbol{b}=\{b_{ij}(k)\}$; $b_{ij}(k)$は状態s_iから状態s_jへの遷移の際に出力シンボルkを出力する確率。

$\boldsymbol{\pi}$: 初期状態確率の集合；$\boldsymbol{\pi}=\{\pi_i\}$; π_iは初期状態がs_iである確率。$\sum_j \pi_j = 1$

$\boldsymbol{F}$: 最終状態の集合

2.3.1.1　HMM の種類

音声認識に用いられる HMM の代表的なものを 図 2.4 に示す。

left-to-right HMM は、状態遷移が1方向だけで逆に戻るような遷移を許さないため、音声の時間的変動を表現するのに適している。しかし音素を単位とした場合、連続音声中で調音結合のために音素は先行音素や後続音素などの影響を受け物理的な音声波形の変形が生じる。このような音韻環境の揺らぎによって音声パターンに変動を起こすため、音素を1つの left-to-right HMM で表わすと出力分布が拡がってしまうという欠点がある。また音素環境を単位とした HMM では、モデルの種類が増えるために1つのモデルに与える学習サンプルが少なくなり過学習がおこる問題が生じる。このため何らかの方法によって音素環境をクラスタリングし、過学習を抑える必要がある。クラスタリングには先験的知識を用いたり、2.3.2節で述べる HMnet を用いる方法がある。

一方 ergodic HMM は、特徴の異なる同一音素でも状態遷移が1通りではないので出力分布が拡がらず柔軟な表現が可能であるが、モデルが小規模の場合には時間的な遷移を表現するのは適さず、またモデルが大規模になるとモデルの自由度が高いためにパラメータの推定が非常に難しくなる。また両者の中間の存在として HMnet があるが詳しくは 2.3.2 節で述べる。

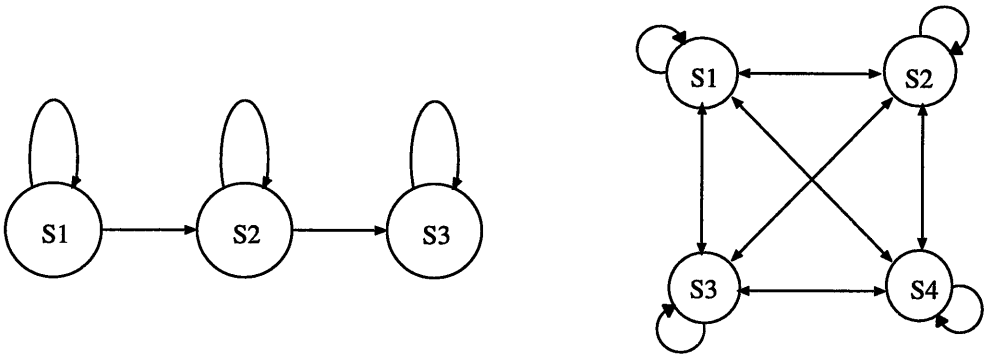

図 2.4: HMM の例

2.3.1.2 HMM の基本問題

HMM を音声認識に用いる場合、以下の3つの基本問題を考えなければならない。

1. 与えられたサンプル $\boldsymbol{O}$ に対して HMM λ の確率 $P(\boldsymbol{O}|\lambda)$ をいかに効率良く計算するか

   HMM からサンプルの生成される確率を単純に計算すると

$$P(\boldsymbol{O}|\lambda) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{O}|\boldsymbol{q},\lambda)P(\boldsymbol{q}\,|\lambda) \tag{2.6}$$

$$= \sum_{q_1,q_2,\cdots,q_T} \pi_{q_1} b_{q_1}(\boldsymbol{o_1}) a_{q_1 q_2} b_{q_2}(\boldsymbol{o_2}) \cdots a_{q_{T-1} q_T} b_{q_T}(\boldsymbol{o_T}) \qquad (2.7)$$

となり $O(2T \cdot n^T)$ の計算時間が必要になる (但し、n^T は与えられたサンプルが時刻 T で受理可能な遷移系列の個数)。これは状態数や可能な状態遷移系列が少しでも多いと計算不可能となる。この問題に対して動的計画法 (DP) を用いて効率的に計算できる Forward アルゴリズムが存在する (付録 B.1)。

2. 与えられたサンプル $\boldsymbol{O}$ に対応する HMM の最適状態遷移系列をどう選ぶか

   あるサンプルに対して HMM の "正しい" 状態遷移系列は存在しない。なぜなら HMM の状態遷移系列は直接観測できず、与えられたサンプルに対して対応する状態遷移系列は、縮退したモデルを除いた場合は一意に決定しないからである。しかし、モデルの構造の学習や連続音声認識の音素区切りの検出や各状態の平均統計量を調べるために、サンプルと状態遷移系列の対応づけが必要になってくる。そのため、"最適な" 状態遷移系列を求めることでこの問題を解決する。代表的な方法としては Viterbi アルゴリズムがある (付録 A)。

3. $P(\boldsymbol{O}|\lambda)$ を最大にするために HMM λ のパラメータをどう調整するか

   サンプルの集合からある最適基準に従ってモデルのパラメータを調整する方法で解析的に直接求めるものは知られていない。しかし、Baum-Welch (付録 B) アルゴリズムのようにパラメータの反復計算をおこなうことで、$P(\boldsymbol{O}|\lambda)$ を局所的に最大にすることはできる。

2.3.2 隠れマルコフ網

音素環境を認識単位とした HMM は割り当てられる学習サンプルが少なくなるため、先験的な知識を用いて音素環境をクラスタリングするなどしてその問題を解決する必要があった。しかし HMnet には、すべての音素環境の HMM のカテゴリや、状態共有構造を統計的に最適化させたネットワーク状のモデルを自動獲得できるアルゴリズムが存在する。この HMnet の各状態には、それぞれ

- 状態番号
- 受理可能な音素環境カテゴリ
- 先行状態および後続状態のリスト
- 出力確率分布のパラメータ

- 自己遷移確率と後続状態への遷移確率

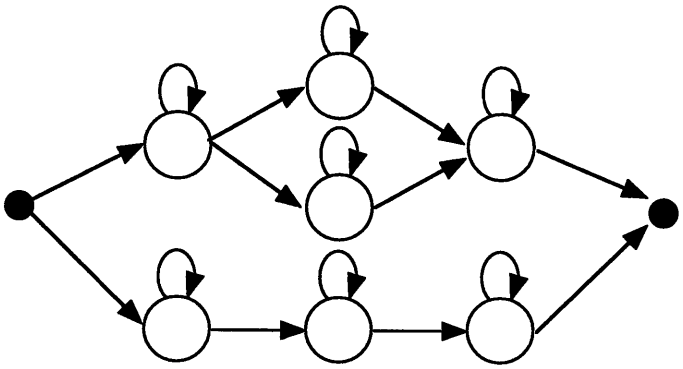

図 2.5: HMnet の例

といった状態固有の情報が割り当てられている。つまり HMnet は音素環境の HMM の尤度を基準に似ている状態をクラスタリングすることで、学習サンプルの少ないものでも信頼度の高い学習ができるモデルである。またネット上でのそれぞれの経路 (パス) は HMM と等価であり、HMM に用いる出力尤度計算やモデルのパラメータの再推定のアルゴリズムをそのまま用いることができる。

2.3.2.1 HMnet 生成アルゴリズム

真に最適な HMnet を生成するためには、音素環境カテゴリの分類や状態共有構造などに関する膨大な組み合わせの問題を解く必要があるが、それを実現するのは難しい。逐次状態分割法 (SSS)[6] や SSS-free[7] は、その近似解法として提案されたアルゴリズムである。

SSS 、 SSS-free のアルゴリズムを以下に示す。

1. 初期モデルの学習

   SSS の場合

   初期モデルとして、1 状態で出力分布が単一ガウス分布 (対角共分散行列) を持つ HMM を用意し、全ての学習サンプルを使って学習する。

   SSS-free の場合

   初期モデルとして、1 状態で出力分布が単一ガウス分布 (対角共分散行列) を持つ HMM を**各音素毎**に用意し、全ての学習サンプルを使って学習する。

2. 分割すべき状態の決定 (SSS、SSS-free 共通)

全ての状態の中で出力分布が最も拡がった状態を選び、分割すべき状態とする。出力分布は単一ガウス分散の値と推定に用いたサンプル数を乗じたものを基準とする。その式を以下に示す。

$$d_i = n_i \times \sum_k^K \frac{\sigma_{ik}^2}{\sigma_{Tk}^2} \tag{2.8}$$

ここで、

K : パラメータ次数

σ_{ik}^2 : 状態 i の出力分布の分散

n_i : 状態 i の推定に用いたサンプル数

σ_{Tk}^2 : 全サンプルの分散 (正規化係数)

3. 状態の分割

選択された状態を2つに分割する。このとき、新しい状態の出力確率分布を以下のようにして求める。

(a) 分割すべき状態を通るすべての学習サンプルについて Viterbi アルゴリズムを使ってこの状態が出力するサンプルの部分系列を切り出してくる。

(b) (a) で切り出された全ての学習サンプルの部分系列を用いて、1状態、2混合の HMM を学習する。

(c) 得られた2つのガウス分布をそれぞれ新しい状態に割り当てる。

このようにして新しい状態の出力確率分布を求めた後、新しい状態の配置を時間方向 (直列) に連結した場合の学習サンプルに対する尤度 P_t と、コンテキスト方向 (並列) に連結した場合の尤度 P_c を計算し、尤度の高いほうを採用する。P_t と P_c は、以下のようにして計算される。

- コンテキスト方向への分割

  コンテキスト方向への分割は、パスが2つに別れるためにそれぞれの学習サンプルがどちらの状態を通るかを決定する必要がある。

SSS の場合

学習サンプルを環境要因 (先行音素や後続音素) 毎にグループ分けし、その集合毎に尤度の高いほうの状態を通るように決定する。

$$P_c = \max_j \sum_l \max(P_m(y_{jl}), P_M(y_{jl})) \tag{2.9}$$

ここで、

j : この状態において分割可能な要因

y_{jl} : 要因 j の値が l 番目の要素である学習サンプルの部分集合

$P_m(y_{jl})$: y_{jl}を状態 m に割り当てた時の尤度

$P_M(y_{jl})$: y_{jl}を状態 M に割り当てた時の尤度

SSS-free の場合

各学習サンプル1つ1つについて、尤度の高いほうの状態を通るように決定する。

$$P_c = \sum_{y_j \in Y_m} \max(P_m(y_j), P_M(y_j)) \tag{2.10}$$

ここで、

Y_m : 分割すべき状態 m を通る学習サンプルの集合

y_j : 状態 m を通る j 番目の学習サンプル

$P_m(y_j)$: y_jを状態 m に割り当てた時の尤度

$P_M(y_j)$: y_jを状態 M に割り当てた時の尤度

- 時間方向への分割 (**SSS、SSS-free 共通**)

  時間方向へ分割する時は、どちらの状態を先に置くかで2とおりの可能性がある。そこで、2つの可能性についてそれぞれ尤度を計算し、その高いほうを P_t とする。

4. 分布の再推定 (**SSS、SSS-free 共通**)

   分割終了後の最適なパラメータを求めるために HMnet 全体を再学習する。その後所定の状態数になるまで、2、3を繰り返す。

2.3.2.2 各アルゴリズムの特徴

SSS と SSS-free のアルゴリズムの違いは次の点である。コンテキスト方向への分割のとき、SSS は学習サンプルの環境要因毎にグループ分けをしそのグループ毎に出力尤度が高いほうの状態を通るようにする。それに対して SSS-free では各学習サンプル毎に出力尤度の高いほうの状態を通るようにする。

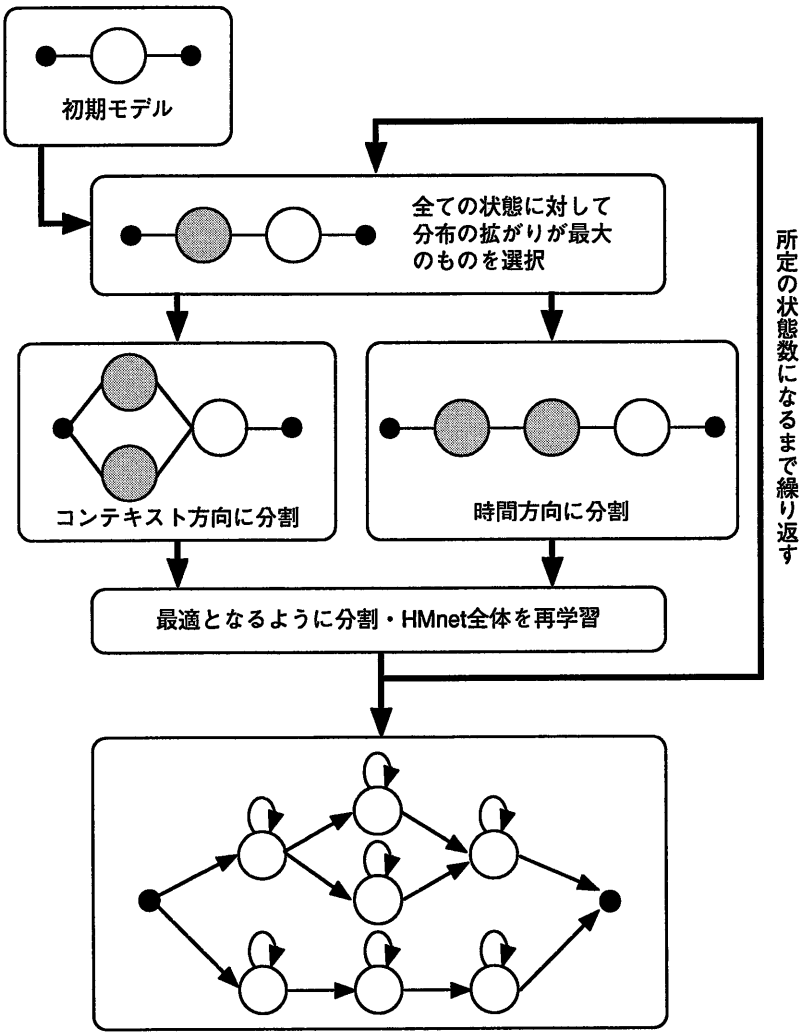

図 2.6: HMnet 生成アルゴリズムの概要

この2つのアルゴリズムによって生成されたHMnetの相異点を述べる。SSSによって環境要因をグループ分けをすることによる利点は、すべての音素環境の直積空間を分割しながらHMnetを構成することによって、例えば /a-k+a/ (/先行音素-音素+後続音素/を示す) という音素環境が学習音声中に現れなかった場合、/a-k+i/ の前半部分と/i-k+a/の後半部分をとってきて、/a-k+a/のモデルにするといった、学習サンプルに含まれなかった環境要素に対する補間作用が期待できるということである。

表 2.6: 与えた環境が不十分の場合の分割

a-m+e		
先々行音素	分布1	分布2
k	231.58	74.05
m	83.74	195.30
	315.32	269.35

a-m+o		
先々行音素	分布1	分布2
k	257.35	47.52
m	97.82	184.50
	355.17	232.02

(数字は各分布での尤度)

しかし、音素の変形に影響している環境要因が全て与えられていない場合、このグループ分けによって問題が生じる。その問題とはグループ中で実際には音響的な特徴が異なっていてもひとつのグループとして一括して扱われることである。そのため状態の分割のとき、そのグループが片方のパスしか通らないので結局状態の分割がおこなわれなくなってしまうことがある。例をひとつ挙げて説明する。いま、与えられた環境要因は前後の音韻環境とし、分割すべき状態を/a-m+e/ 、 /a-m+o/ の2つの音素環境の学習サンプルが通っているとする。また、これらのサンプルは先々行音素が /k/ のものと /m/ のものとがあり、その違いによって音響的な変形を受けているものとする。つまり、音響的には先々行音素が /k/ である /a-m+e/、/a-m+o/ と、先々行音素が /m/ である /a-m+e/、/a-m+o/という2つのグループにクラスタリングされる。このとき2混合のガウス分布が先々行音素が /k/ であるグループの音響的特徴を表現している分布 (分布1とする) と、先々行音素が /m/ であるグループの音響的特徴を表現している分布 (分布2とする) からなっている (表2.6)。このような状況で、ある状態を後続音素について分割することを考える。 SSS では尤度の和が高いほうの状態を経路とするのでこの場合どちらも分布1が高く、結局、状態が分割できなくなってしまう。だが SSS-free においては学習サンプルそれぞれについて尤

度を調べているので、この場合だと先々行音素が /h/ であるグループの音響的特徴を表現している分布 (分布 1) と /k/ であるグループの音響的特徴を表現している分布 (分布 2) に分割されることになる。よって SSS-free で生成された HMnet はより音響的な特徴を表わしたモデルとなっていると言える。

ただ SSS-free の場合学習サンプルである音素環境が存在しないときに環境要因の補間作用が無くなるので、代わりのモデルとして音素環境非依存の HMM を用いなければならないという問題もある。

2.3.3 HMM を用いた音素認識

HMM を用いた音素認識のひとつの手法の概略を示す。HMnet も HMM がネットワーク状になっているだけなので処理過程は同じである。音素区切りが既知の場合、その区間の特徴量ベクトル系列を全ての HMM (HMnet の場合には全てのパス) に与え尤度を計算し、尤度が最大なものを認識結果とする。このときの尤度計算には Forward アルゴリズムもしくは Viterbi アルゴリズムを用いる。

音素区切りが未知の場合、以下の方法で認識をおこなう。

- 全ての HMM の最終状態は全ての HMM の初期状態とナル遷移で連結している。
- 初期確率は各々の HMM の初期状態が 1.0 それ以外は 0.0 とする。
- 時間 T(与えられたサンプルの入力が終了する時間) において各々の HMM の最終状態の中から尤度最大である状態を選び、そこからバックトレースをおこなう。

この条件のもとで Viterbi アルゴリズムをおこなうことで得られた最適系列を認識結果とする。これによって音素区切りの自動決定もおこなうことができる。

図 2.7 が、その処理の概要である。図の四角のマスはトレリスと言われるもので各時間における各状態の尤度が入っている。また矢印は次の状態に遷移が可能であることを表わしている。この図は入力音声が ka の場合を表している。(実際の入力音声は文章を単位としており、無音区間による音声の区切り以外は音声区間の情報は与えられない。) このとき、Viterbi アルゴリズムを用いて全ての時間、状態で尤度を計算する。時刻 T において、各 HMM の最終状態から尤度最大である状態を選び、バックトレースをおこなうことで最適状態遷移系列を求める。この結果 HMM /k/ の最終状態から HMM /a/ の最初の状態に遷移したのが時刻 i なので音素 / k / の区間は時刻 0 から時刻 $i-1$、音素 / a / の区間は時刻 i から時刻 T と自動決定できる。

しかしこの方法は、計算量が非常に大きくなるため枝刈り等の必要がある。また挿入、欠落による認識性能の低下の問題もある。

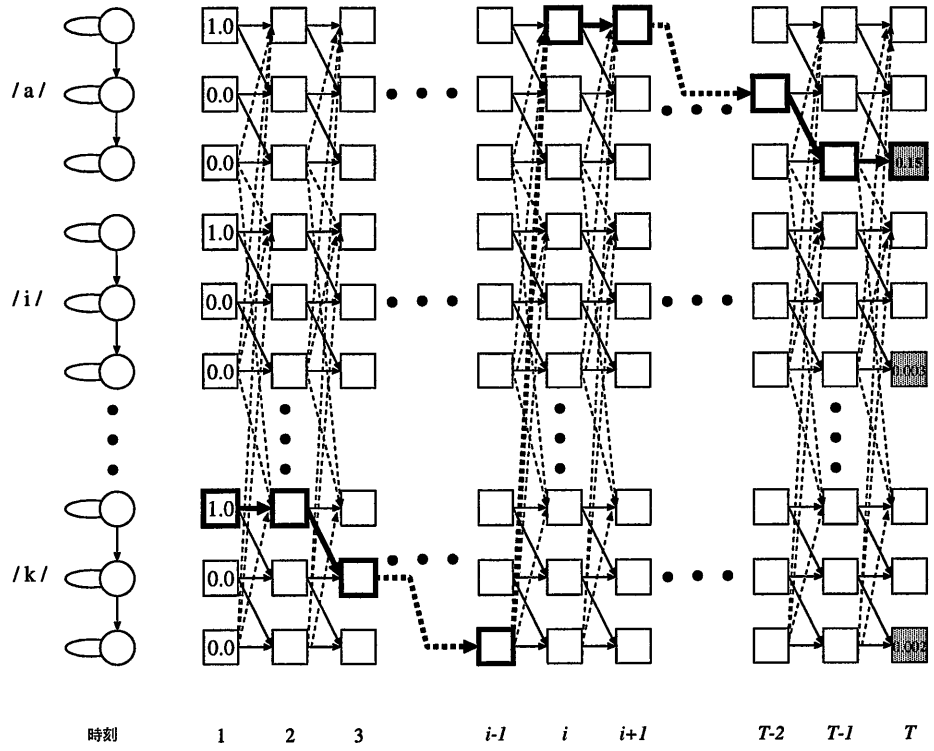

図 2.7: 連続音声認識の概要

2.4 移動ベクトル場平滑化法による話者適応

2.4.1 移動ベクトル場平滑化法のアルゴリズム

連続分布型 HMnet の場合、出力確率密度関数 (ガウス分布) の平均値や分散、遷移確率を話者に適応できるが、 VFS では計算時間の問題や認識の効率性から平均値のみの適応をおこなう。また HMnet で VFS をおこなう場合、話者によって HMnet のネットワーク形状は異ならないということを前提にしている。VFS は、認識対象となる話者の適応用音声の発話内容が既知である条件下において、適応サンプルに対応するモデルのパラメータの再学習と残りのパラメータの再推定という 2 つのステップによって話者適応モデルを求める。

1. 適応サンプルに対応するモデルのパラメータの再学習
   本ステップでは、認識対象となる話者の音声に含まれる音素に関して、学習サンプルによって生成されたモデルの出力確率密度関数の平均値のみを再学習する。

(1) 多量の学習サンプルによって生成した音素モデル (これを標準話者のモデルと呼ぶことにする) を認識対象となる話者の音素モデルとする。

(2) 認識対象となる話者の入力音声のサンプルに1対1で対応するモデルを HMnet から切り出し、そのモデルを音素系列に対応できるように連結する。連結されたモデルについて出力確率密度関数の平均値だけを再学習する。

2. 適応サンプルに存在しなかった音素モデルのパラメータの再推定

本ステップは連結学習前後の HMM の平均ベクトルの差分ベクトルを移動ベクトルとみなし、再学習されなかったモデルの平均ベクトルの移動ベクトルを補間し、補間された移動ベクトルに平滑化をおこなう。

(1) 移動ベクトルの計算

認識対象となる話者の HMnet の各状態の出力確率密度関数の平均ベクトルの組 ($C^I = (c_1^I, \cdots, c_K^I), K$: HMnet の全状態数) のうち再学習された k 番目の平均ベクトル c_k^I ($k \in K_1, K_1$:適応音声中に存在した音素に対応する HMnet の各状態の平均ベクトルの集合) と標準話者の HMnet の各状態の出力確率密度関数の平均ベクトルの組 C^R の中で対応する c_k^R により平均ベクトルの差分ベクトル v_k を求め、これを話者空間の移動ベクトルとする。

$$v_k = c_k^I - c_k^R \quad (k \in K_1) \tag{2.11}$$

(2) 移動ベクトルの補間

C^I のうち再学習されなかった状態の出力確率密度関数の平均ベクトル $c_n^I (n \in K_2, K_2$: 適応音声によって再学習されなかった HMnet の各状態の平均ベクトルの集合) を再学習された k 番目 ($k \in K_1$) の移動ベクトル v_k および、c_n^R と c_k^R 間のファジイ級関数 $\mu_{n,k}$ [12] から求めた移動ベクトル v_n を用いて c_n^I に移動する。

$$v_n = \sum_{k \in K_1} \mu_{n,k} v_k c_n^I = c_n^R + v_n \mu_{n,k} = \frac{1}{\sum_{j \in K_1} \left(\frac{d_{n,k}}{d_{n,j}}\right)^{\frac{1}{m-1}}} \tag{2.12}$$

ここで、$d_{n,k}$は c_n^Rと c_k^R とのユークリッド2乗距離を示す。m はファジネスと呼ばれる重み係数である。

(3) 移動ベクトルの平滑化

上述のステップで得られたモデルは、十分な適応サンプル数が得られていない場合に推定誤差を含んでいる。このような推定誤差を含むものから求められた移動ベクトルの方向は、非連続的な動きをしていると考えられる。そこで、話者空間を移動するための移動ベクトルに連続性の拘束条件を入れ、移動ベクトルの方向性を揃える (平滑化をおこなう) ことで推定誤差を吸収する。

$$v_k^S = \frac{\sum_{m \in N(k)} \alpha_m \mu_{k,m} v_m}{\sum_{m \in N(k)} \alpha_m \mu_{k,m}} c_k^S = c_k^R + v_k^S \tag{2.13}$$

ここで、c_k^S は平滑化をおこなって得られた話者適応後の HMnet の状態 k の出力確率密度関数の平均ベクトル、$N(k)$ は、c_k^R の k-近傍にある平均ベクトルの番号で、本実験では連結学習で求められた平均ベクトル全てを使用する。また α_m は v_m の信頼度を与える定数で、$\alpha_m = \alpha_1 (m \in K_1)$、$\alpha_m = \alpha_2 (m \in K_2)$ とする。

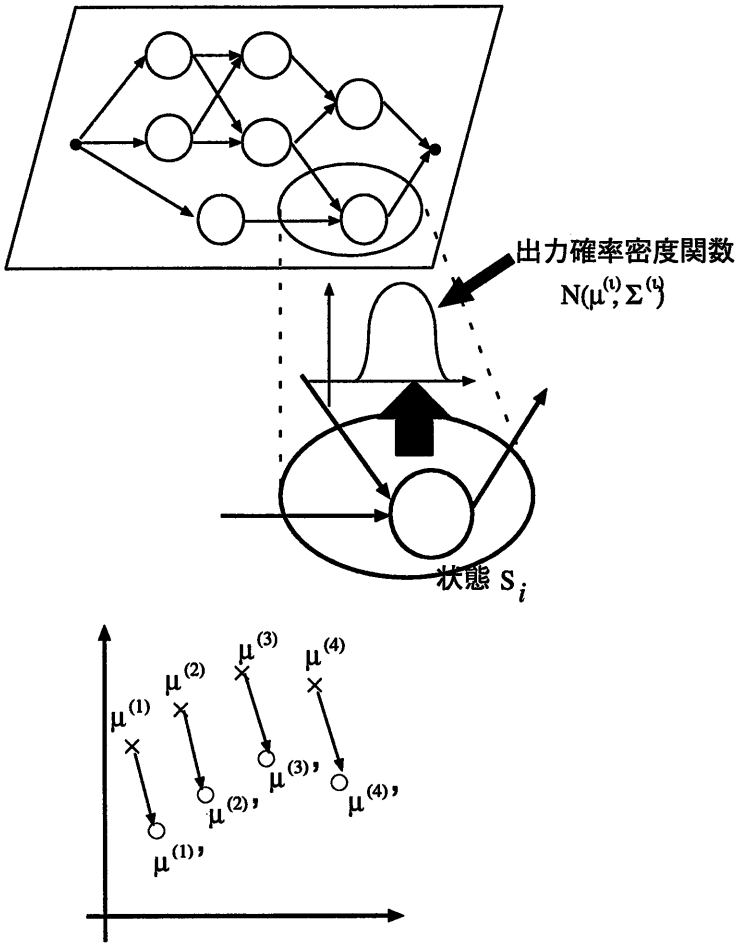

図 2.8: VFS の概念図

ステップ 2 の処理を図 2.9 を用いて説明する。図は HMnet の状態数が 4 である場合について示してある。連結学習によって c_1^R、c_2^R、c_3^R が、c_1^I、c_2^I、c_3^I にそれぞれ移動し、c_n^R は適応音声中で再学習されなかったものとする。この場合 c_n^I は、c_1^R、c_2^R、c_3^R と移動ベクトル v_1、v_2、v_3 および $\mu_{n,1}$、$\mu_{n,2}$、$\mu_{n,3}$ を用いて計算される。またこの c_n^S に対

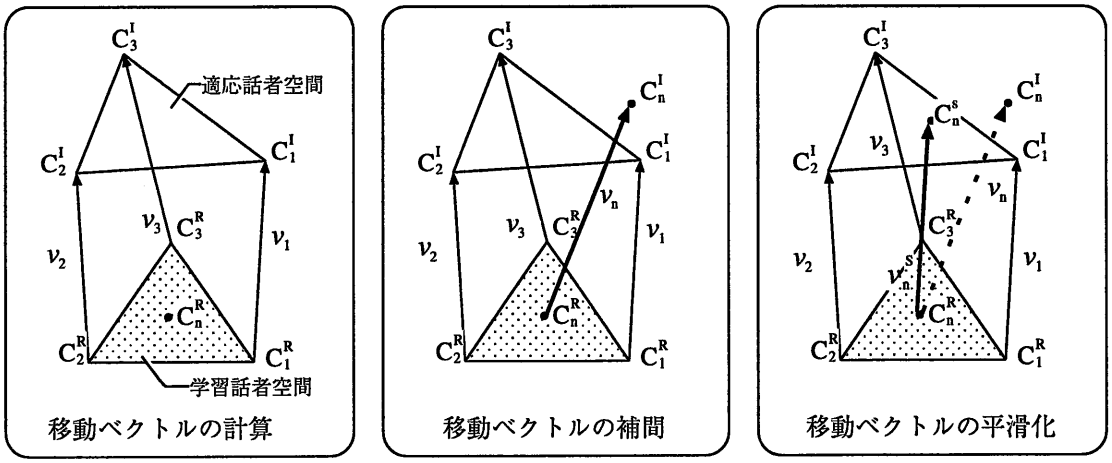

図 2.9: 移動ベクトルの補間・平滑化の概念図

応する移動ベクトル v_n^s は、v_1、v_2、v_3、v_n とそれぞれに対応するファジイ級関数と各移動ベクトルに対する信頼性の重み α_m を用いて平滑化される。

2.4.2 SSS-free に基づく HMnet への VFS の適用

SSS-free に基づいた HMnet に VFS を適用する場合、適応音声に対応する音素環境が HMnet 内に複数存在するため、どのモデルを選択するかが問題になってくる。その問題を解決するために前節のステップ 1 の部分を変更する。

1. 適応サンプルに対応するモデルのパラメータの再学習
   本ステップでは、次の方法で適応サンプルに対応するモデルのパラメータを再学習する。
   (1) 多量の学習サンプルによって生成した標準話者モデルを認識対象となる話者の初期音素モデルとする。
   (2) 認識対象となる話者の入力音声の音素環境の系列にそれぞれ対応するモデルを HMnet から全て切り出してくる。
   (3) 切り出したモデルについて Viterbi アルゴリズムによって尤度が最大となるようなモデルを選択する。
   (4) 選択されたモデル系列に対し、Baum-Welch アルゴリズムによってモデルの出力確率密度関数の平均値のみを再学習する。

これを図 2.10 および 図 2.11 を使って説明する。図 2.10 は標準話者の状態数 7 の HMnet である。認識対象となる話者の適応用音声の各音素についてその音素に対応するようなモデルを標準話者の HMnet からすべて切り出す。このときモデルが同一状態を共有することがあるが、共有している状態の部分をそれぞれのモデルで別々に切り出す。この処理は

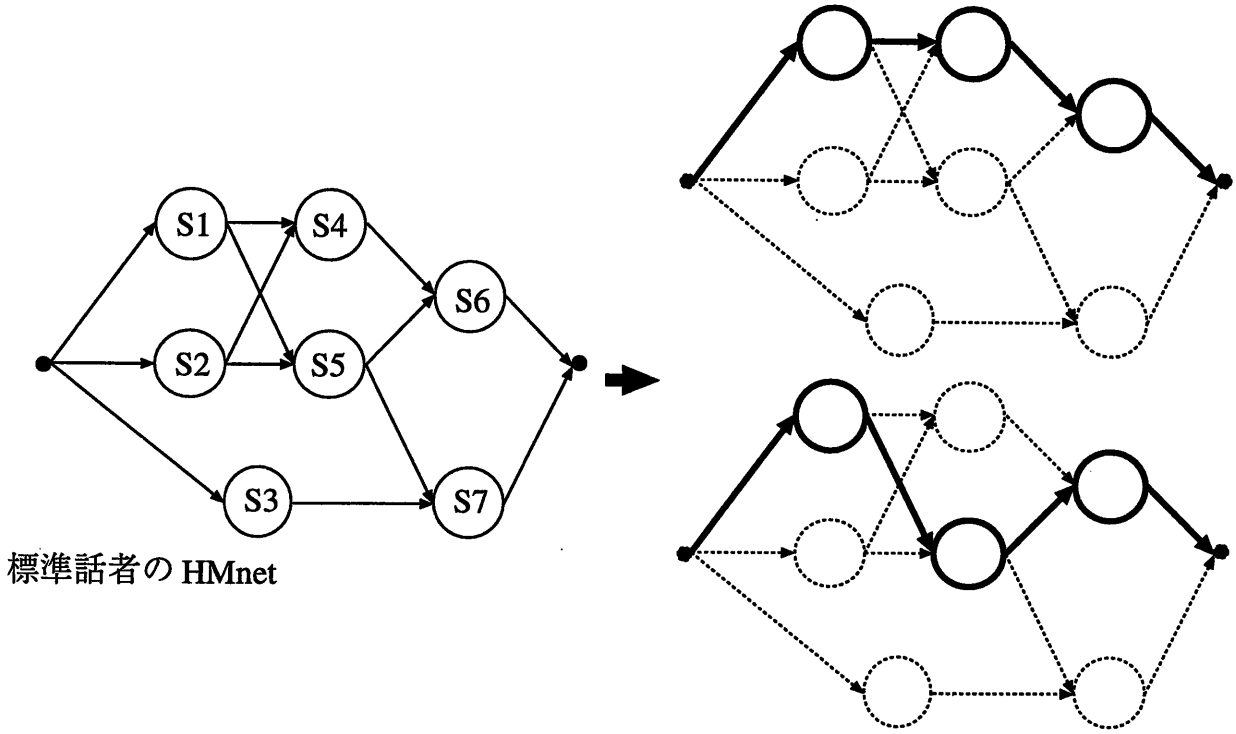

図 2.10: HMnet からのモデルの切り出し

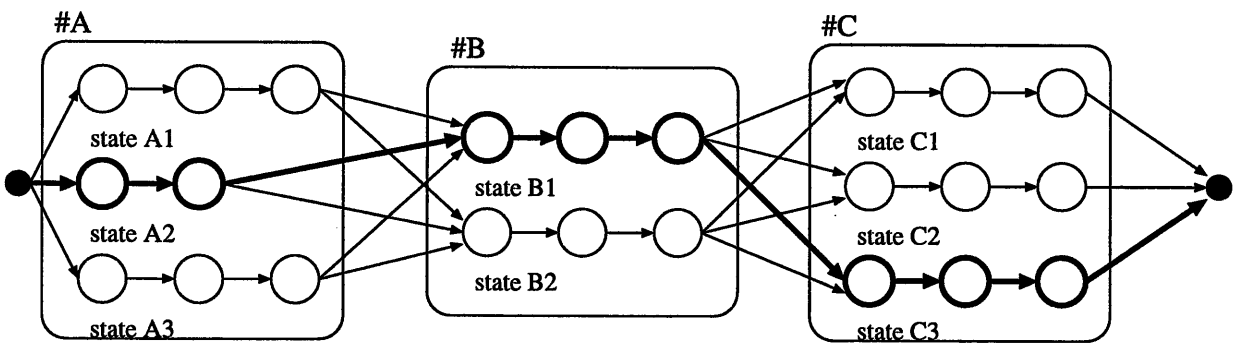

図 2.11: Viterbi アルゴリズムによるモデルの選択

Viterbiアルゴリズムによるモデル選択のときに標準話者のHMnetではその音素に対応していない状態遷移系列を尤度最大遷移系列として選択してしまうことを防ぐためにおこなう。次に図2.11で切り出したモデルの尤度が最大となる状態遷移系列を選出する。このとき図の太線の部分が尤度最大となったとすると、その部分のみをBaum-Welchアルゴリズムによって出力確率密度関数の平均値のみを再学習する。

2.5 性能評価実験及び考察

2.5.1 予備実験

まずパラメータ設定や、適応文章の内容等による話者適応の効果を調べるために予備実験をおこなった。音響分析条件は表2.4であり、標準話者のHMnetは日本音響学会連続音声データベースの36人(男女各18人)が50文章(Aセット)を発話したものをSSS-freeを用いて学習して生成された不特定話者モデルを使用する。また適応用音声としては、HMnetの学習に使用しなかった男性1人に、HMnetで学習した同一文章を使用する。評価用音声としては、適応用音声で使用した人に対しHMnetで学習した同一文章から適応で使用した文章を除いた40文章(Aセット)とHMnetで学習しなかった文章50文章(Iセット)で実験をおこなった。またVFSのパラメータα_1を1.0、α_2を0.0に定めた。

はじめに移動ベクトルの補間・平滑化のところで使用するファジネスを決めるために実験をおこなった。

図2.12の結果から、認識率にさほど上下はないものの最高の認識率だった、1.6にファジネスを定める。

次にHMnetの状態数を一定(本実験では状態数200)にして適応音声の量を変化させたときの認識率の変化を見るための実験をおこなった。図2.13、図2.14の横軸は適応音声の累積音素数である。図2.14で*coverage*と表記してあるが、この*coverage*は以下の様に定義している。

$$coverage(\%) = \frac{\text{適応音声によって再学習された状態数}}{\text{HMnetの状態数}} \times 100 \qquad (2.14)$$

図2.14からみてゆくことにする。適応文章の累積音素数が200くらいでカバレッジが80%になり、それ以上適応文章を増やしてもカバレッジはさほど上昇しない。このことから、それ以上累積音素数を増やした認識結果はVFSの補間や平滑化による誤認識をかなり軽減したものになっていると思われる。

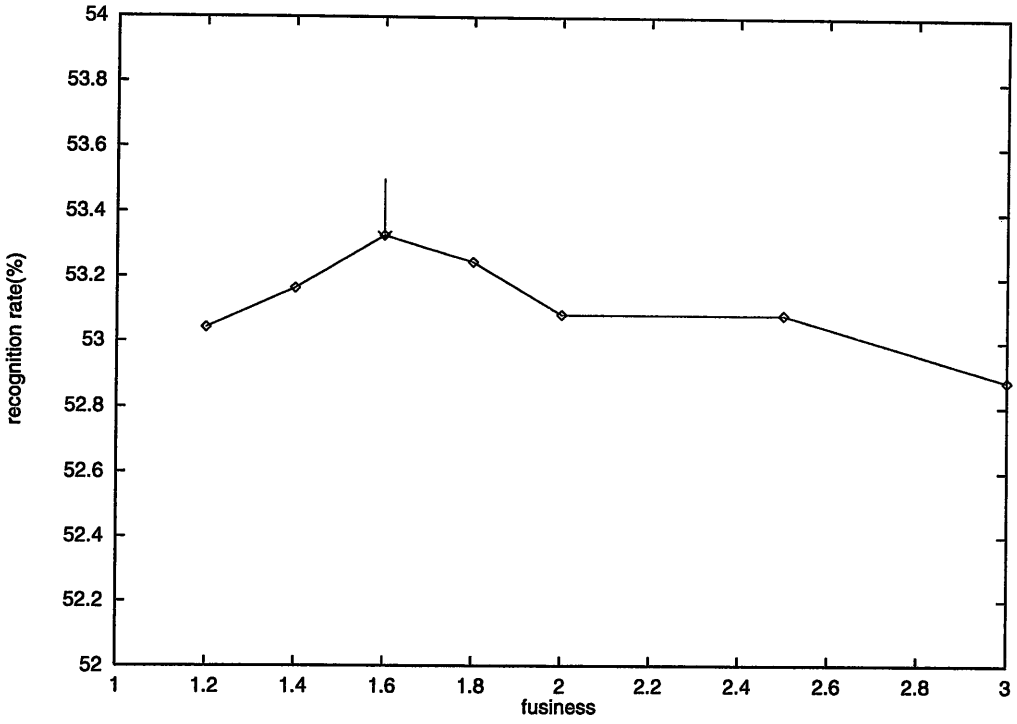

図 2.12: ファジネスによる認識率の変化

そのような推察をふまえ図2.13を見る。ここで音素数0の認識率というのはVFSをかけない場合の音素認識率である。まず学習と同一文章のもの(Aセット)とそうでないもの(Iセット)の結果を比較すると、Aセットのほうが認識率が7～13%良いことがわかる。この原因は、HMnetの学習音声が50文章であるために学習された音素に偏りがあることや、同じ理由により話者適応をした場合も偏った音素のみが十分に適応されているためと思われる。次に適応音声の量による認識率の変化を見てみる。Iセットの適応文章数1のときVFSをかけないものと比較して、認識率が2%程下がっている以外は、Aセット、Iセットともに似たような傾向を示している。累積音素数が300以上になると認識率がほぼ頭打ちになる。(Aセット：約79%、Iセット：約67%)これは先に予想したとおりの結果である。また適応文章が少ないとき認識率があまりよくないのは、適応音声が少なすぎて移動ベクトルが偏った方向にあるために補間、平滑化がうまくいかないことが原因であると思われる。

2.5.2 各アルゴリズムで生成されるHMnetでのVFSの効果

本節では、SSSとSSS-freeによって生成されるHMnetにおいてそれぞれにVFSを適用し、適応文章数を変化させて実験をおこなった。音響分析条件は表2.4、実験条件は表2.7

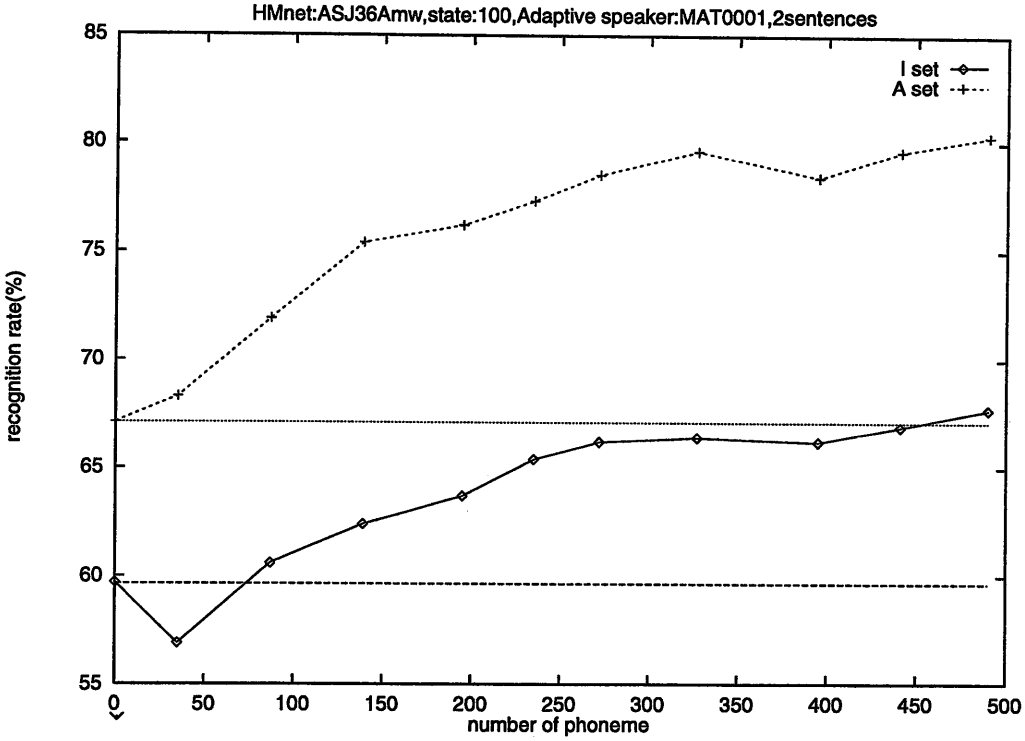

図 2.13: 適応音声の量による認識率の変化

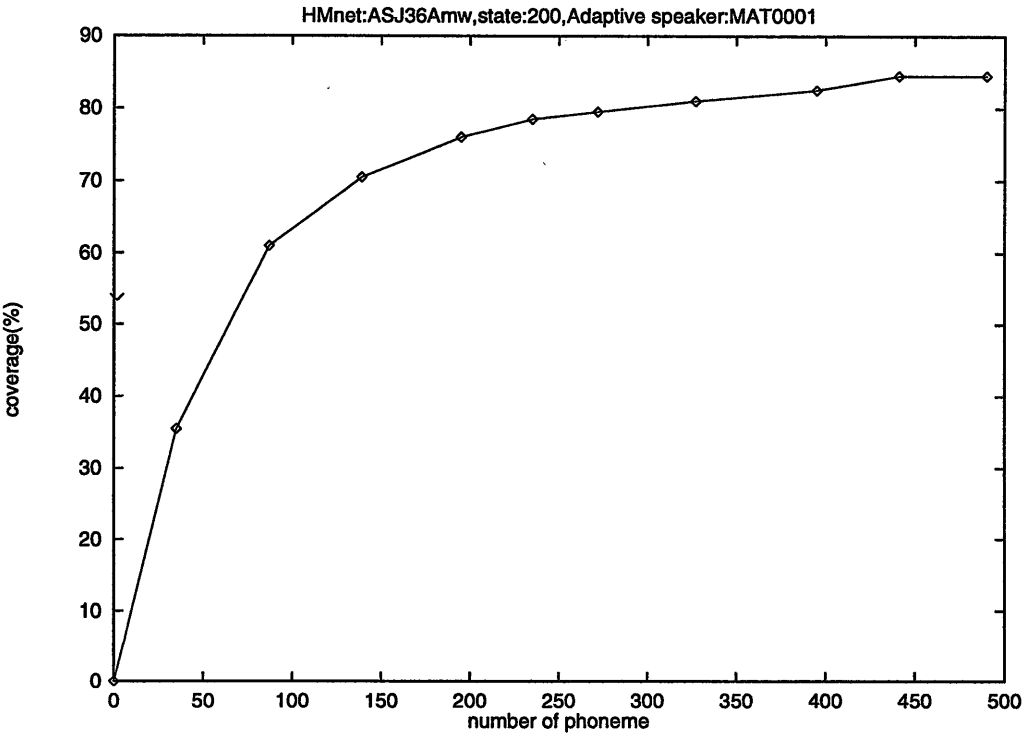

図 2.14: 適応音声の量による HMnet のカバレッジの変化

のとおりである。HMnet は ATR 日本語音声データベース音韻バランス 503 文章から男性1名が発話した 400 文章を用いて学習し、VFS には HMnet の学習に用いなかった男性4人、女性4人について HMnet の学習で使用しなかった文章を使用した。評価には VFS を適応した計8人に対してそれぞれの学習には使用しなかった 50 文章を使って認識をした。また VFS のパラメータのファジネスは先程の実験結果から 1.6 とした。

表 2.7: 実験条件

HMnet	状態数 210
	特定話者モデル
各状態の分布	単一ガウス分布
	対角共分散行列

結果を図 2.15 に示す。グラフの横軸は適応文章数 (1 文章 30 ～ 50 音素) であり、これが 0 のときは VFS を適用しなかったときの認識率を表わす。またこの結果は 8 人の認識率の平均である。(各人の結果は章末に示す。)

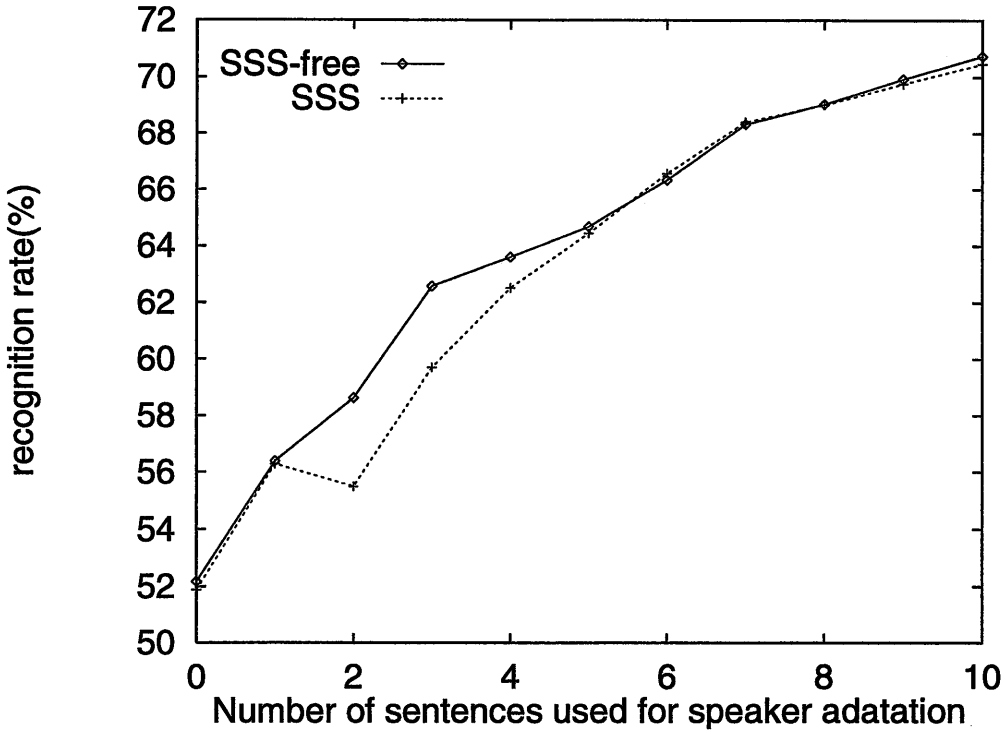

図 2.15: 適応文章数の変化による SSS と SSS-free からなる HMnet での VFS の認識率

結果から、適応文章数 2 ～ 4 文章においては SSS-free で HMnet を生成したもののほう

が SSS によるものよりも効果があることがわかる。認識率の差は最大で 3.1%である (2 文章のとき)。

図 2.16 ～ 図 2.21 各話者の認識結果を示す。結果には 3 通りの傾向があると思われる。1 つは女性に適応したときの結果である。図 2.16、図 2.17 のグラフを見ると適応文章数が 0 のとき (VFS をしないとき) 認識率は、35%前後である。この認識率の原因は話者性によるものである。しかし、適応文章を増やすと認識率が飛躍的に向上し適応文章数 10 文章で認識率はそれぞれ 67%、75%となり VFS にかなりの効果があることがわかる。

男性に適応したときの結果を見ると、図 2.18、図 2.19 では、適応文章数 5 文章までは認識率は適応前よりも良くないが、適応文章数を 10 文章ではそれぞれ 61%、76%と若干ではあるものの VFS の効果が現れたと言える。また図 2.20、図 2.21 においては適応文章をいくら増やしても適応前の認識率を上回ることがなく全く VFS の効果が見られなかった。

この結果を考察する。適応文章が少ない場合には、 VFS の適応サンプルを用いたパラメータ再学習の部分においてそのパラメータに推定誤差を含むことがある。VFS を適応するとき HMnet の学習話者と適応話者の特徴がかなり異なる場合には、その推定誤差はそれ程問題にはならないが、特徴が似ている場合には推定誤差がモデルの精度に大きく影響する可能性がある。

SSS によって生成された HMnet では音素環境とパスが 1 対 1 で対応している。そのため適応サンプルに存在した音素環境については推定誤差を含んだパスしか存在しないのでモデルの精度に影響する可能性がある。

一方 SSS-free による HMnet では 1 つの音素環境に対して複数のパスが対応し、適応サンプル中に存在する音素環境についても適応されたパスとされていないパスが生じる。このため認識時に学習話者と適応話者の特徴が異なる場合は適応されたほうのパスを選択し、特徴が似ている場合には適応されていないほうのパスを選択できる。これが SSS-free によって生成された HMnet のほうが VFS の適応文章の少ない場合に有効である原因と思われる。

また適応文章が増えると、 SSS 、 SSS-free による HMnet のほとんどのパスが再学習されるため、SSS-free による HMnet でパスを選択しても選んだパスが再学習されている可能性が高いので、パスを選択することが適応文章が少ないときに比べて有利に働かず認識率に差がなくなったと考えられる。

2.6 まとめ

本章ではモデルパラメータの調整による話者適応法の移動ベクトル場平滑化による話者適応法を SSS-free によって生成された隠れマルコフ網に適応する手法を提案した。また逐次状態分割法 (SSS) によって生成された隠れマルコフ網に VFS を適応したものと性能比較のための音素認識実験をおこなった。その結果提案手法のほうが、適応文章数が少ない場

合には話者適応に効果があることがわかった。これはSSS-freeによって生成されたHMnetが1つの音素環境に対して複数のパスが存在することが有利に働いたためと考えられる。また適応文章が多い場合でも同程度の話者適応の効果が見られた。この結果からSSS-freeの利点を損なうことなくVFSによる話者適応が可能であることが示された。

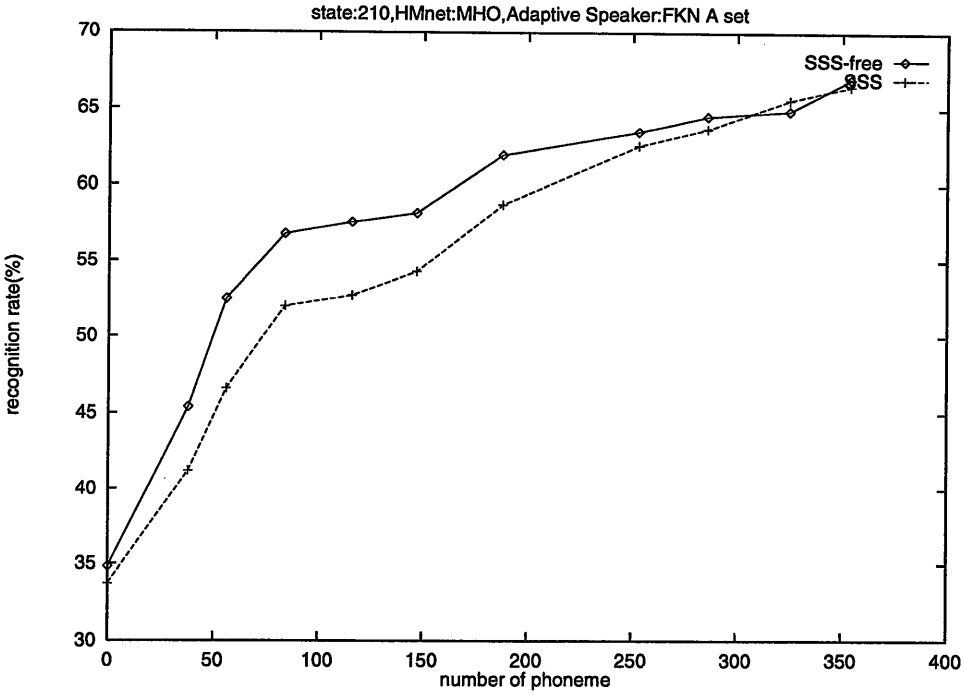

図 2.16: SSS と SSS-free で生成された HMnet の適応の効果の違い (話者オープン・女性 1)

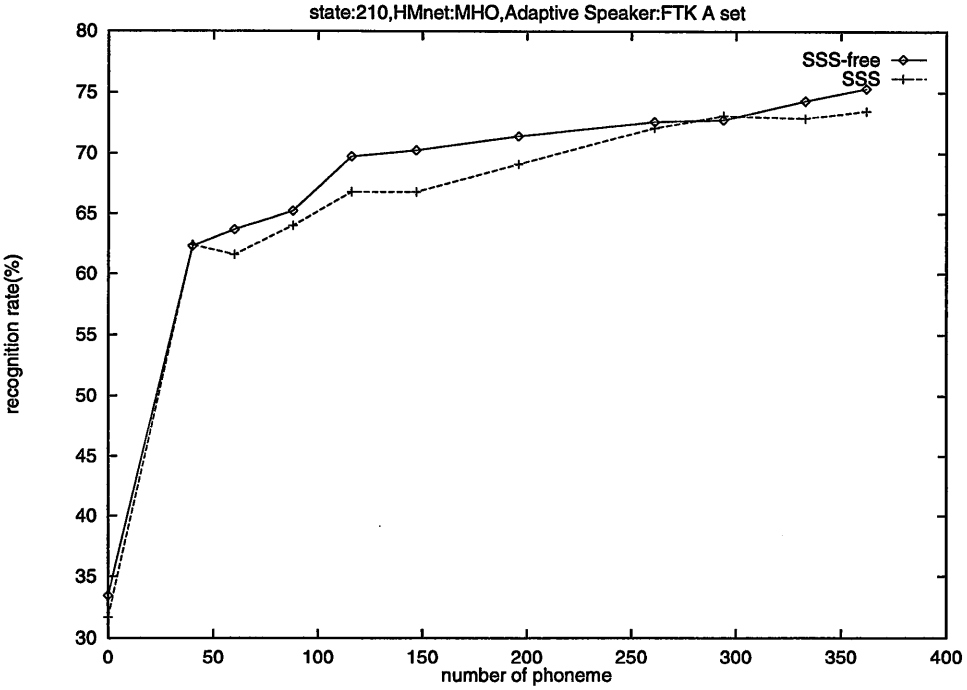

図 2.17: SSS と SSS-free で生成された HMnet の適応の効果の違い (話者オープン・女性 2)

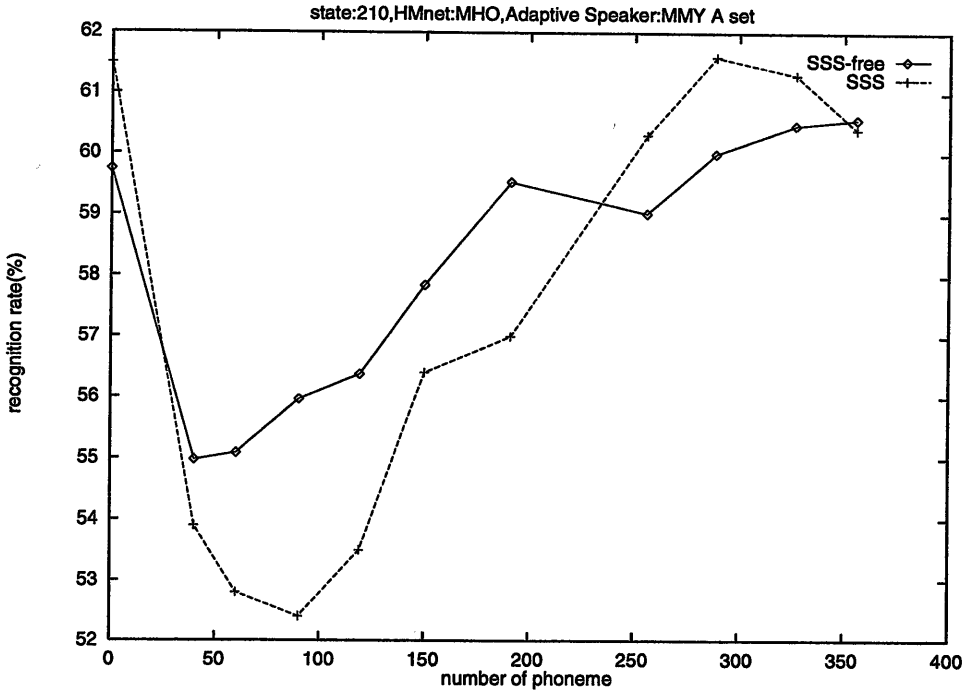

図 2.18: SSS と SSS-free で生成された HMnet の適応の効果の違い (話者オープン・男性 1)

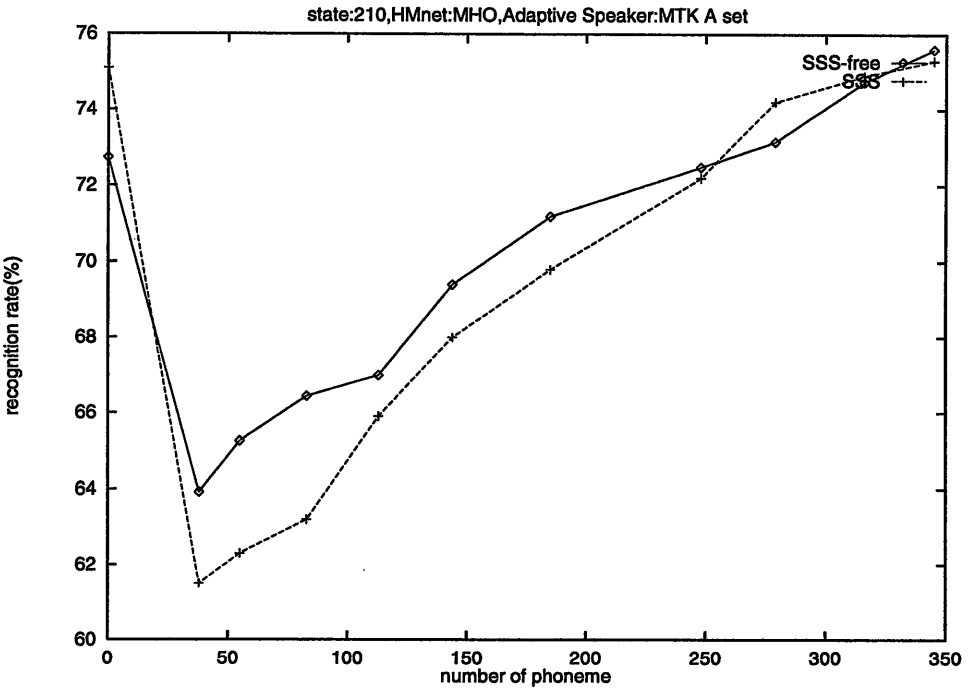

図 2.19: SSS と SSS-free で生成された HMnet の適応の効果の違い (話者オープン・男性 2)

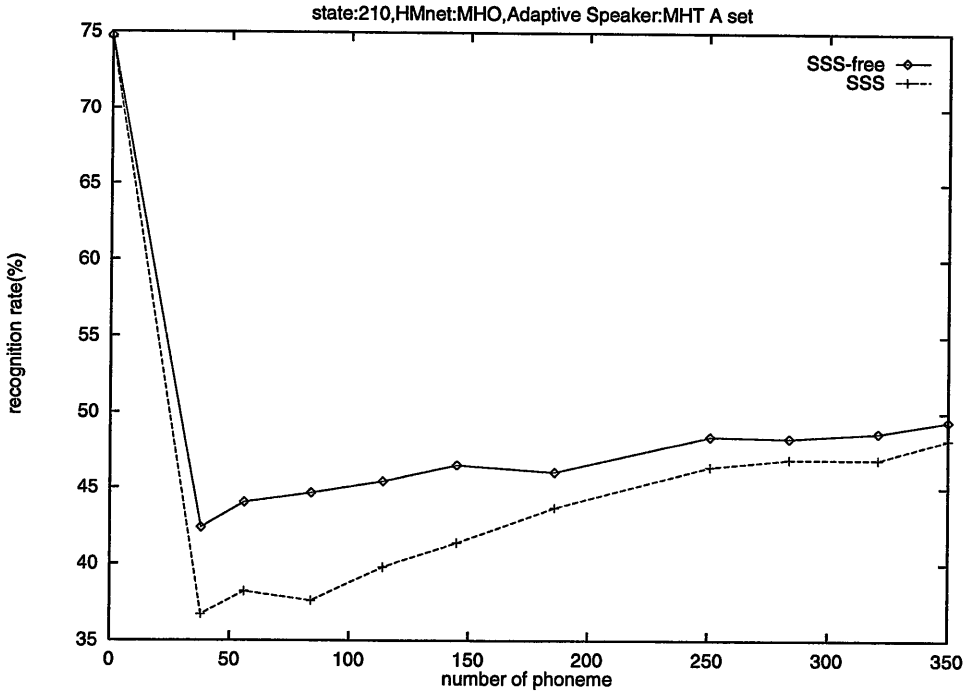

図 2.20: SSS と SSS-free で生成された HMnet の適応の効果の違い (話者オープン・男性 3)

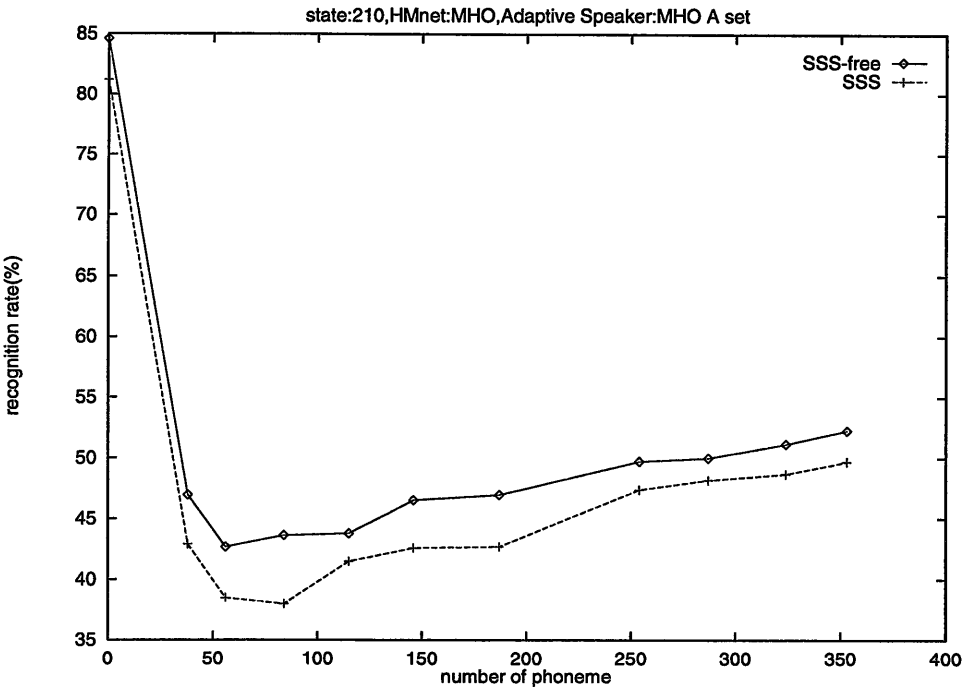

図 2.21: SSS と SSS-free で生成された HMnet の適応の効果の違い (話者クローズ・男性 4)

第3章

音素毎の木構造話者クラスタリングに基づく話者適応法

3.1 はじめに

話者の個人性による認識性能の低下という問題を解決する話者適応にはモデルパラメータの調整による方法と、話者選択による方法の2つの流れがある。話者選択による方法は不特定話者モデルの中から話者の特徴に近いモデルを選択することで話者適応をおこなうものである。これは、少量の適応音声で話者適応が可能であり、利用者に対する負担が少ない話者適応手法である。

なかでも、"木構造話者クラスタリングによる話者適応法" [16](小坂、1994) はごく少量の話者音声で高精度の話者適応が可能であると言われている。このモデルは木構造の各ノードに HMnet が配置されるようになっている。木構造の上層においては不特定話者モデルの特徴を持ち、下層では特定話者モデルの特徴を持っている。また認識時には認識対象が学習で使用した話者の特徴に近ければ下層のモデルを選択し、そうでなければ上層のモデルを選択するので木構造を構成する話者に含まれない話者においても柔軟に適応することが可能である。

しかしこの手法では話者の個人性はどの音素についても同じものとして捉えて話者適応をおこなっている。一方話者の個人性は母音や鼻音にあると言われ、すべての音素の個人性を一律に扱うよりも、音素毎に個人性を考慮した話者適応を考えたほうが精度の高い話者適応が可能であると考えられる。

本章では、まず話者の個人性を音素毎に考慮したほうが話者の特徴を忠実に反映したモデルであることを確かめるために予備実験をおこない、その正当性を確かめる。また音素毎に話者の個人性を考慮した木構造話者クラスタリング手法を提案し、その性能を認識実験をおこなうことで比較、検討する。

3.2 音素毎の話者選択による HMnet 合成法

本節では、音素毎の話者の個人性を考慮した木構造クラスタリングが有効かどうか調べるため以下の2つについて実験をおこなった。

音素毎に個人性を考慮する必要性について

実験1 複数の特定話者 HMnet を用意し、内容既知な音声サンプルを全ての HMnet に与え最尤となる話者の HMnet を音素別に数えあげヒストグラムを作る。

音素毎に個人性を考慮した場合のモデルの認識精度について

実験2 次に示す HMnet の合成法を用いて認識実験をおこなう。(概念図は図 3.2)

1. 1人の話者の多量の音声で HMnet を学習
2. 1. で学習した HMnet から VFS を用いて複数話者の HMnet を生成
3. 2. までで作った全ての HMnet に認識対象の音声を与え最尤となる話者とパスを数えあげる。
4. 3. の結果を使ってモデルを合成する (出力確率のみ)

$$b^{(i)}(x) = \sum_{s \in \text{speaker}} \lambda_s^{(i)} N(\boldsymbol{\mu}_s^{(i)}, \boldsymbol{\Sigma}_s^{(i)}) \tag{3.1}$$

ただし

method 1.

$$\lambda_s^{(i)} = \begin{cases} \dfrac{n_s^{(i)}}{\sum_{s' \in \text{speaker}} n_{s'}^{(i)}} & (\text{if} \sum_{s' \in \text{speaker}} n_{s'}^{(i)} \neq 0) \\ \dfrac{n_s}{\sum_{s' \in \text{speaker}} n_{s'}} & (\text{otherwise}) \end{cases} \tag{3.2}$$

method 2.

$$\lambda_s^{(i)} = \frac{n_s}{\sum_{s' \in \text{speaker}} n_{s'}} \tag{3.3}$$

$N(\boldsymbol{\mu}_s^{(i)}, \boldsymbol{\Sigma}_s^{(i)})$: 話者 s の HMnet の状態 i の出力確率 (ガウス分布)

$n_s^{(i)}$: 話者 s の HMnet の状態 i を通って最尤となるサンプル数

n_s : 話者 s の HMnet を最尤とするサンプル数

method 1 は各音素 (HMnet の各状態) 毎に出力確率の重みを変えたものである。また **method 2** は HMnet の各状態で同じ重みを使用したものである。つまり、**method 1** は

音素毎に木構造話者クラスタを構成したもの、**method 2** は小坂らが提案した手法のモデルの近似したものとして考えることができる。

表 3.1: 実験条件

HMnet	状態数 210、男性 (男性 1)1 人 503 文章で学習
VFS	学習の話者と認識対象の話者を除いた 8 人
	(男性 5 人 + 女性 4 人 − 認識対象)
	にそれぞれ 50 文章を与え適応

この 2 つの重みづけによって合成した HMnet それぞれにおいて認識実験をし比較、検討する。また音響分析条件は表 2.4、実験条件は表 3.1 のとおりである。

実験 1 の結果を見てゆく。表 3.2 は女性 1 の音声サンプル 50 文章 (内容既知) をそれ以外の話者の HMnet に与え最尤であるモデルを音素別に数えあげた結果である。表の縦のアルファベットは音素を示し、表内の数値はそのモデルが選択された割合 (%) を示している。

結果から音素毎に選択されたモデルの割合は、女性 3 が一番高い。つまり女性 1 の特徴は、女性 3 に一番近いということである。しかし、選択されたモデルの割合にばらつきがあり、他の女性にも実験をおこなったところ同様の結果が見られる。この結果から、音素毎に話者の個人性があると考えることができる。

また、男性 1 がかなりの割合で選択されているのもこの結果の大きな特徴である。これは、男性 1 の音声で信頼度の高い HMnet を学習し、それ以外の話者はこのモデルを初期モデルとして VFS を用いて学習したことが原因であると考えられる。例えば、/ z / や / p / などは適応用のサンプルが少ないため初期モデル (男性 1 のモデル) 以外は推定誤差を含んだ学習をしてしまい、比較的信頼度の高い男性 1 のモデルを選択したと考えることができる。

次に実験 2 について考えてゆく。図 3.1 は、それぞれの重みづけによって HMnet を合成したときの認識率である。グラフの横軸は適応文章数 (文) を表わし、横軸は 50 文章を認識させたときの男性 6 人女性 4 人の音素認識率 (%) の平均を示している。また、実験は適応文章数 1、5、10、30、50 文章においてそれぞれおこなった。

その結果、全体的に状態毎に重みをつけたほうが認識率が高く、適応文章数が少ないときにその差が大きいことがわかる。これは、状態毎に重みをつけたほうが忠実に音素毎の個人性を表わしていると言える。

3.3　音素毎の木構造クラスタリングに基づく話者適応手法

前節の結果より、音素毎に話者の個人性を考慮して認識対象のモデルを構成することが有効であることがわかった。そこで本節では音素毎の個人性を考慮した木構造話者クラス

表 3.2: 女性1が選択した特定話者モデルの音素別の割合 (%)

調音様式	音素	女性2	女性3	女性4	男性1	男性2	男性3	男性4	男性5	男性6
(母音)	a	9.32	**36.34**	33.85	1.55	5.59	10.87	0.31	0.00	2.17
	i	13.57	**68.78**	11.76	4.07	0.90	0.45	0.45	0.00	0.00
	u	21.51	**33.14**	13.95	29.65	1.74	0.00	0.00	0.00	0.00
	e	2.38	**86.90**	4.76	5.95	0.00	0.00	0.00	0.00	0.00
	o	10.70	**46.51**	23.26	13.95	4.65	0.93	0.00	0.00	0.00
(半母音)	y	4.88	**51.22**	12.20	12.20	17.07	2.44	0.00	0.00	0.00
	w	**66.67**	0.00	0.00	0.00	16.67	16.67	0.00	0.00	0.00
破裂音	p	0.00	0.00	0.00	**84.62**	7.69	0.00	7.69	0.00	0.00
	b	2.86	8.57	22.86	**65.71**	0.00	0.00	0.00	0.00	0.00
	t	8.24	12.94	11.76	10.59	**54.12**	1.18	0.00	1.18	0.00
	d	4.55	2.27	6.82	**86.36**	0.00	0.00	0.00	0.00	0.00
	k	1.82	27.27	5.45	23.64	**40.91**	0.91	0.00	0.00	0.00
	g	16.13	25.81	14.52	**33.87**	6.45	1.61	1.61	0.00	0.00
鼻音	m	6.94	**31.94**	15.28	19.44	22.22	2.78	1.39	0.00	0.00
	n	33.33	19.79	0.00	**34.38**	4.17	0.00	5.21	1.04	2.08
	N	7.27	**45.45**	0.00	43.64	0.00	1.82	0.00	1.82	0.00
ふるえ音	r	12.50	20.54	12.50	**50.89**	0.00	3.57	0.00	0.00	0.00
摩擦音	s	12.50	26.79	**42.86**	5.36	1.79	0.00	3.57	0.00	7.14
	z	23.08	26.92	7.69	**42.31**	0.00	0.00	0.00	0.00	0.00
	sh	11.11	22.22	**61.11**	2.78	0.00	0.00	0.00	0.00	2.78
	j	0.00	31.71	**56.10**	9.76	2.44	0.00	0.00	0.00	0.00
	h	0.00	9.84	**36.07**	14.75	22.95	1.64	9.84	0.00	4.92
	ch	4.00	12.00	**52.00**	4.00	24.00	0.00	0.00	0.00	4.00
	ts	10.00	**50.00**	15.00	10.00	15.00	0.00	0.00	0.00	0.00
	全体	10.84	**38.49**	18.72	18.96	8.69	2.44	0.86	0.14	0.86

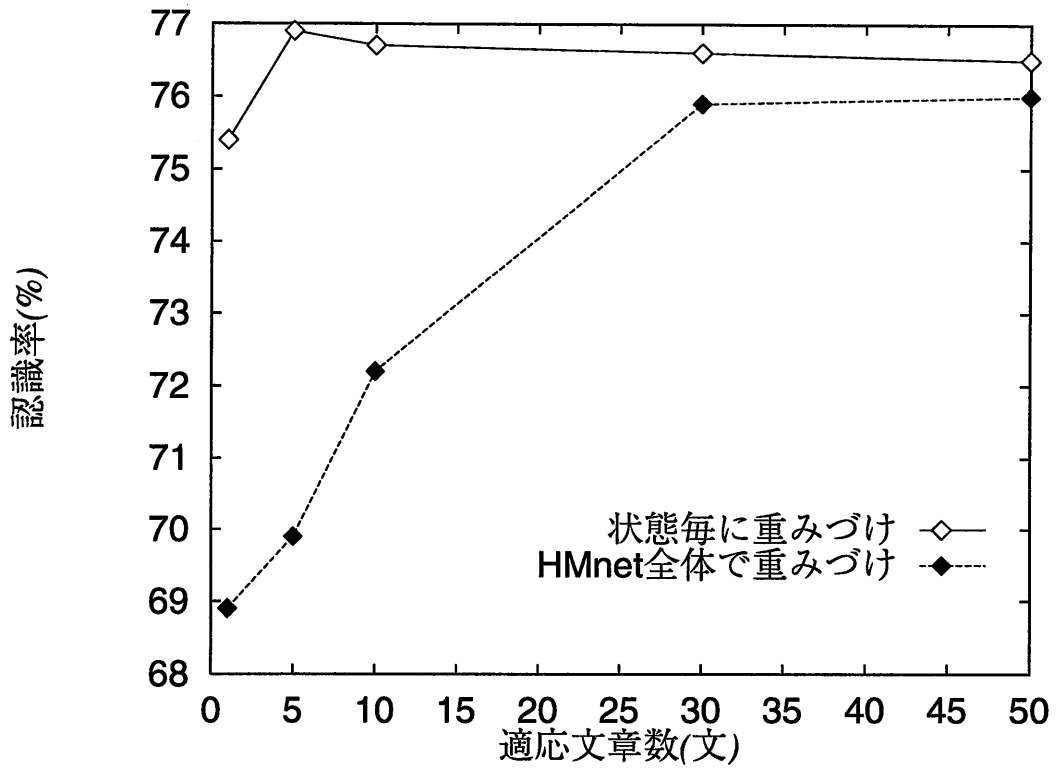

図 3.1: 各々の重みづけで合成した HMnet の認識率

タリングのアルゴリズムを提案する。

3.3.1　学習アルゴリズム

1. **多数の特定話者モデルの作成**

   1 人が発声した多量の学習音声を用いて初期 HMnet を作成する。この HMnet に VFS [8] をおこなうことで、近似的に多数の話者の特定話者 HMnet を生成する。ここで学習に VFS を用いるのは、比較的少量の音声で学習ができるため、また構造の同じ HMnet を得ることができるためである。

2. **各音素毎に木構造話者クラスタの作成**

   HMnet の中の各音素に対応する部分を抜き出す。これを音素 HMnet と呼ぶことにする。HMnet を音素 HMnet と区別するために全音素 HMnet と呼ぶ。各々の音素 HMnet すべてに次の手続きで木構造クラスタを作成する。

   (1) クラスタを構成する特定話者音素 HMnet において距離 $D_p(M_i, M_j)$ が最大となる 2 つのモデル M_i と M_j を求める。ただし $D_p(M_i, M_j)$ は各々の特定話者 HMnet の音素 p に対応する状態の出力分布の Bhattacharyya 距離 [18] の総和である。

   Bhattacharyya 距離とは、2 つの分布間の分類誤差の下限を示すものである。2 つの

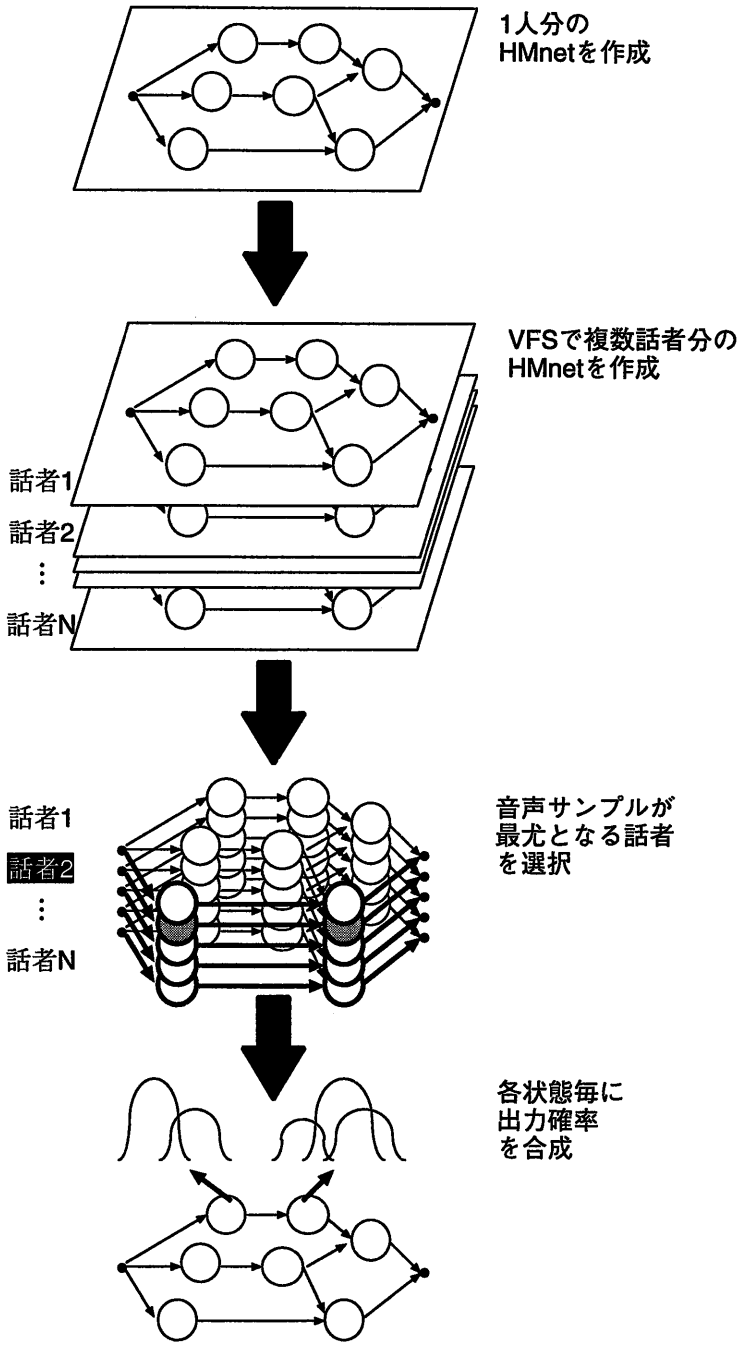

図 3.2: HMnet の合成の概念図

分布を f 、g とすると

$$d = -\log\left\{\int (f(x)g(x))^{\frac{1}{2}} dx\right\}$$

正規分布の場合、次のように表わせる。

$$f = N(\mu_1, \Sigma_1)\,,\; g = N(\mu_2, \Sigma_2)$$

$$d = \frac{1}{8}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^t \left[\frac{\boldsymbol{\Sigma_1} + \boldsymbol{\Sigma_2}}{2}\right]^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma_1} + \boldsymbol{\Sigma_2}|}{2|\boldsymbol{\Sigma_1}|^{\frac{1}{2}} \cdot |\boldsymbol{\Sigma_2}|^{\frac{1}{2}}} \qquad (3.4)$$

(2) 各特定話者モデルを距離 D_p を用いて M_i 、 M_j に近いほうに分類し、クラスタを2つのサブクラスタに分割する。

(3) サブクラスタに含まれる話者情報を用いて、クラスタの代表となる HMnet を作成する。各状態の出力確率は式 (3.5) を用いて合成した。これはクラスタ内の各話者の HMnet の出力確率の重み付き和である。

$$b^{(i)}(x) = \sum_s \frac{n_s^{p(i)}}{\sum_{s'} n_{s'}^{p(i)}} N(\boldsymbol{\mu}_s^{(i)}, \boldsymbol{\Sigma}_s^{(i)}) \qquad (3.5)$$

ただし i は状態番号、s、s' はその話者クラスタに含まれる話者、$n_s^{p(i)}$ は話者 s の HMnet の状態 i に対応する音素 p のサンプル数を表わす。

(4) クラスタに含まれる話者数が1になるまで (2) で分割されたサブクラスタにおいても同様に (1) ～ (3) をおこなう。

また全音素 HMnet においても同様の方法によって木構造話者クラスタを作成する。これは小坂らが提案したものと同じモデルである。

この処理を図 3.3、図 3.4 を用いて説明する。図 3.3 はある HMnet の特徴を表わす空間であり、×、●印は特定話者モデルを表わしている。実際各 HMnet 間の距離は相対的にしかわからないのでこのような表現をすることはできないが、説明をわかりやすくするためにこの図を用いる。

まず各特定話者 HMnet 間の距離テーブルを作る。そのテーブルの中から距離が最大になるふたつの HMnet を選び、クラスタの代表点とする (図 3.3● 1)。それぞれの点において2つの代表点との距離が近いほうのクラスタに分類する。この操作によってできたクラスタは外周のひとつ内側の2つの楕円である。以下各サブクラスタで同様の操作をおこないサブクラスタの要素数が1になるまでおこなう。

このような手順で各音素毎に木構造話者クラスタを作成する。同時に話者クラスタの選択のときに適応音声に存在しない音素のモデルを補間するために全音素 HMnet でも木構造話者クラスタを作成する (図 3.4)。

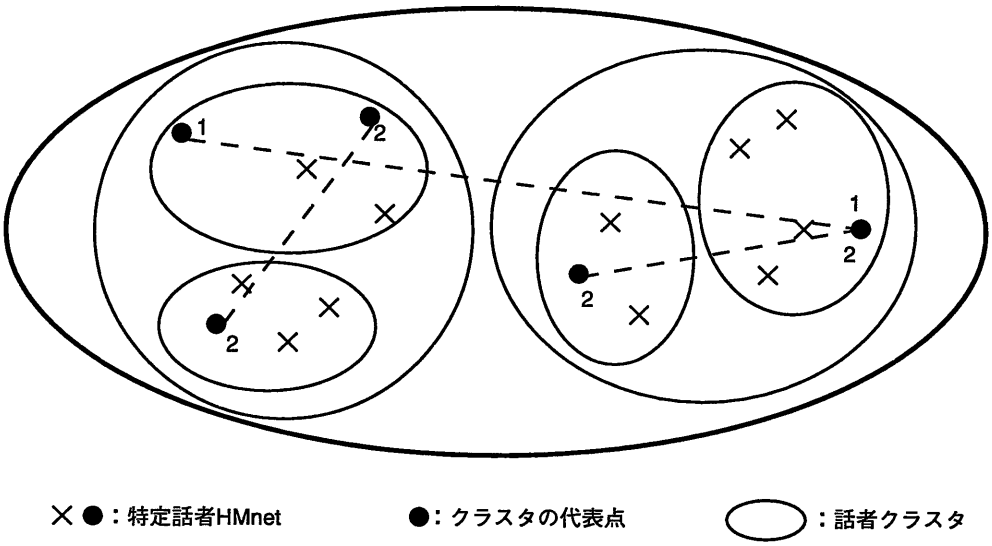

図 3.3: 木構造話者クラスタ構成アルゴリズムの概要

3.3.2　話者適応アルゴリズム

認識対象となる話者の音声を使用して、各音素毎に最適な話者クラスタを選択する。適応音声中に存在する各音素に関して、その音素の木構造話者クラスタを用いて以下の手順で木構造を探索してモデルを選ぶ。

1. クラスタを最上層にマークし、このクラスタのHMnetに適応音声を与え尤度を求める。

2. マークされたクラスタに属するサブクラスタのHMnetに適応音声を与え、最尤となるHMnetを持つサブクラスタを選択する。
3. 選択されたサブクラスタにマークし、クラスタにサブクラスタが存在しなくなるまで2を続ける。
4. 1～3の処理で全ての階層で1つのクラスタが選択される。その全てのクラスタのHMnetの中から更に最尤となるHMnetを選択し、これを認識に用いる。
   ただし適応用音声に存在しない音素に関しては全音素HMnetで構成した木構造クラスタから、適応音声に含まれる全ての音声サンプルを用いて同様の方法で最尤となるHMnetをもつ話者クラスタを選択し、その音素に対応するHMnetの一部を抜きだし認識に用いる。

この処理を図3.5を用いて説明する。今音素 / a /、/ o /、/ p /という音素の木構造話者クラスタがあり適応音声中には音素/ a / と / o / は存在するが、音素 / p / は存在しなかったとする。音素 / a / と / o / については適応音声を用いて最適な話者クラスタを選択する。一方全音素HMnetで作成した木構造話者クラスタから全適応音声を用いて最適

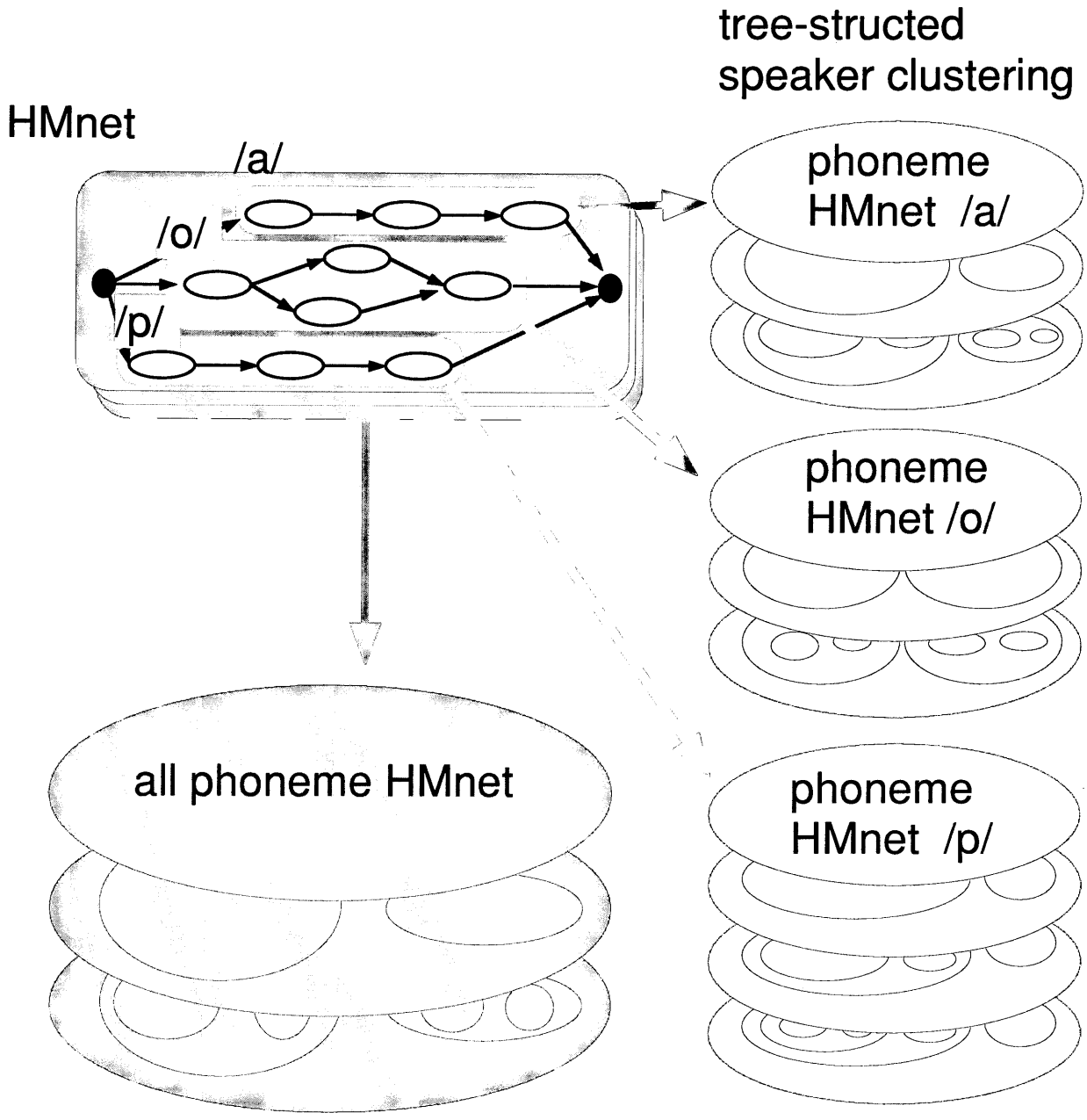

図 3.4: 音素毎の木構造話者クラスタ

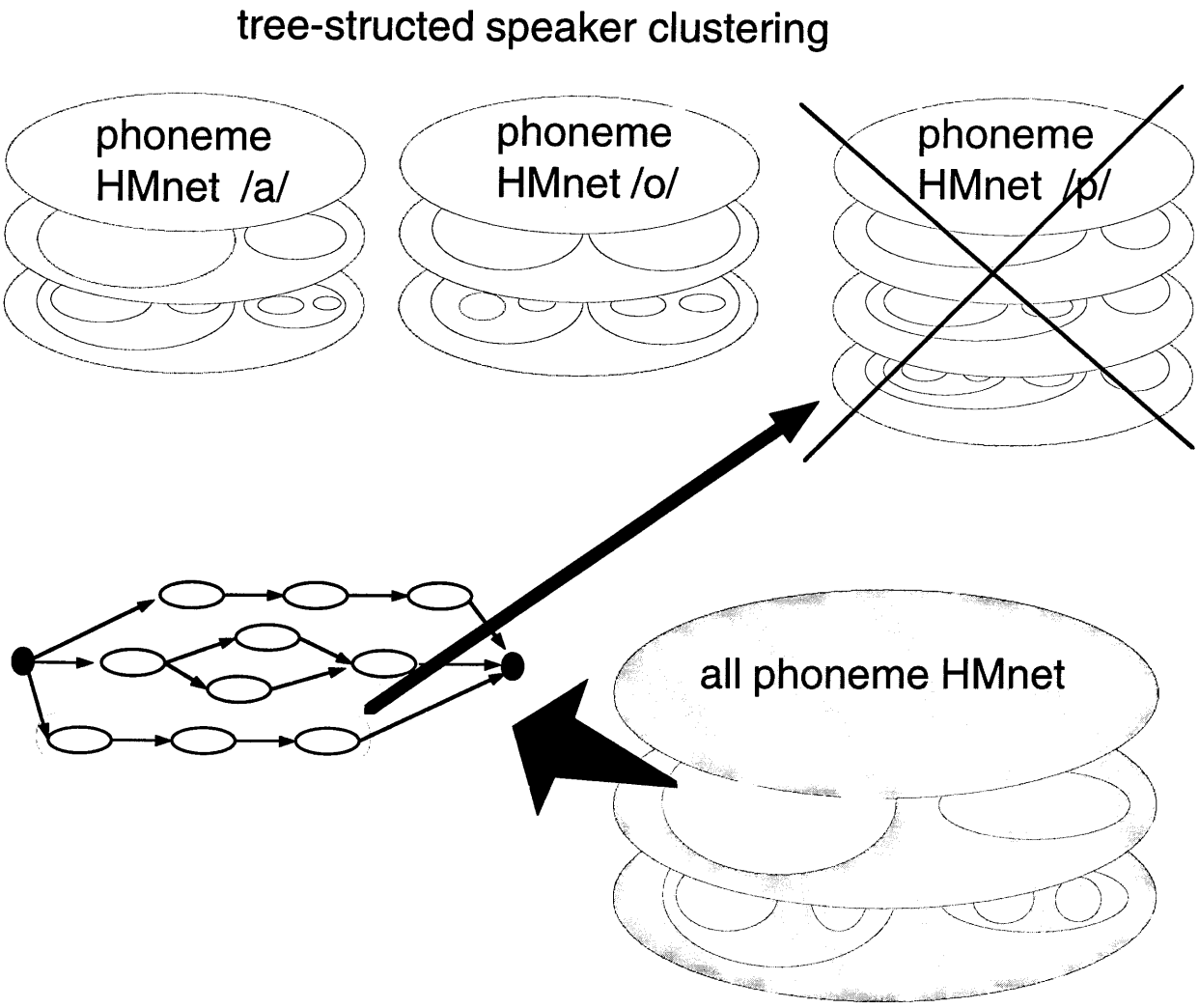

図 3.5: 音素毎の木構造話者クラスタ

な話者クラスタを選択しておく。音素 / p / については全音素 HMnet の木構造話者クラスタで選択された HMnet の音素 / p / に対応する音素 HMnet を抜き出しこれを認識時に利用する。

提案した手法では、木構造を音素毎に作っているため、音素毎に話者の個人性が異なっていても、その個人性を忠実に表現できるような木構造クラスタを構成することができると期待される。また、話者クラスタを各音素毎に選択するために、認識対象となる話者の特徴を忠実に反映するモデルを選択することが期待できる。

3.3.3　性能評価実験及び考察

音素毎に個人性を考慮した木構造話者クラスタと考慮しない木構造話者クラスタを用いた話者適応について木構造を構成する話者が少ない場合と多い場合について認識実験をおこなった。音響分析条件は表2.4、 HMnet の条件は表3.3のとおりである。

表 3.3: 実験条件

初期 HMnet	状態数 210
	特定話者モデル
	SSS-free を用いて学習
各状態の分布	単一ガウス分布
	対角共分散行列

3.3.3.1　木構造を構成する話者が少ない場合の認識実験

まず木構造を構成する話者数が比較的少ない場合に関して実験をおこなった。初期 HMnet は ATR 日本語音声データベース音韻バランス 503 文章から男性1名が発話した 400 文章を用いて学習した。また 50 文章の音声で VFS を用いて 8 人分の HMnet を作成した。認識には男性2人と女性2人について 50 文章を用いた。木構造話者クラスタを選択する適応音声の量は1～5文章と変化させて認識実験をおこなった。

認識結果を図3.6に示す。図中の “音素毎” のデータが音素の個人性を考慮した木構造話者クラスタ、 “HMnet” のデータが考慮していない木構造話者クラスタを用いて話者適応をおこなった場合のそれぞれの子音・母音の認識率を示している。

また音素毎にどのような木構造を構成しているか、どのノードを選択したかをデンドログラムで章末に示す。このとき認識対象の話者は男性 (MSH) であり 1 文章を使って適応をおこなったものである。この図の見方を説明する。まず、青色で書かれた話者は男性、赤色で書かれた話者は女性を示す。緑で囲まれた部分は選択されたクラスタ (話者の集合) で

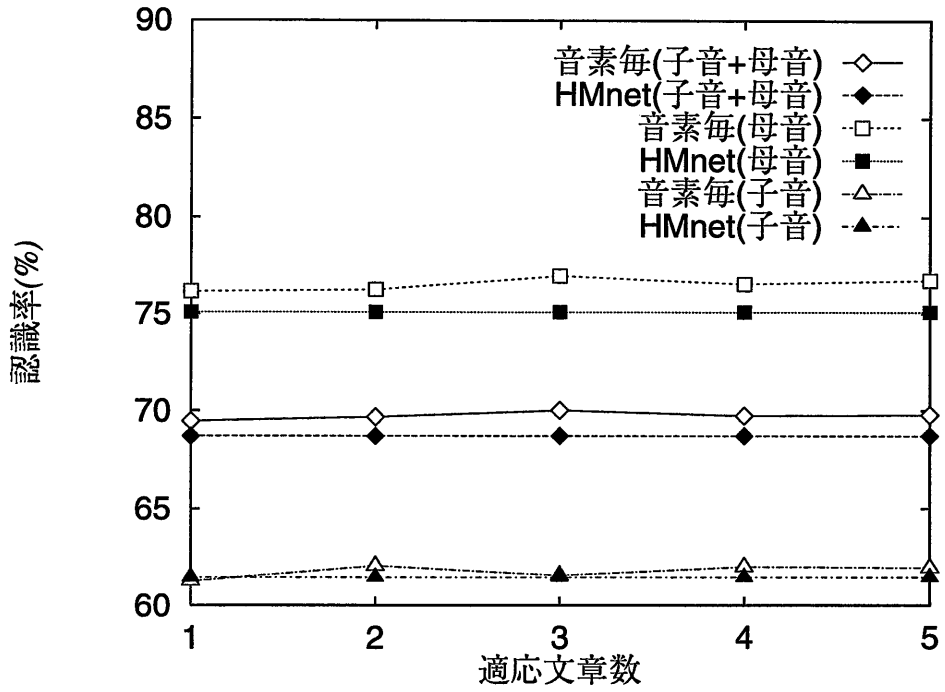

図 3.6: それぞれの木構造クラスタを用いた話者適応の音素認識実験結果

ある。ピンクで囲まれた部分は適応音声に存在しなかったために全音素 HMnet で生成した木構造話者クラスタの情報を用いて補間されたものである。また各枝の長さは、クラスタ間の距離を表わす。例えば全音素 HMnet 木構造話者クラスタの男性のクラスタと女性のクラスタ間の距離は約 460 であり、話者 MTK と MMY 間の距離は約 180 である。

認識結果を見ると母音に関しては、2%程度の認識率の向上が見られた。また子音については認識率にほとんど差が見られなかった。音素毎に木構造話者クラスタの構造を見てみると、ほぼどの音素においても上層のほうは男性と女性への分割のように同じような分割がおこなわれ、下層にいくにしたがって次第に異なる分割をしてゆく。全音素 HMnet で構成した木構造話者クラスタと比較すると 2 層目までは同じ構成であり、男性と女性とにクラスタが分割されている。また 2 層より下の階層では各々異なる分割になっている。

しかし、上層での分割も男女には分かれず、他のものと全く異なる構造を持つ音素 HMnet も得られた。/ s /、/ b /、/ p /がその例である。理由としては、その対象となった音素について、初期 HMnet から VFS で不特定話者モデルを作成するときの学習データが少なかったことが考えられる。その音素認識率を調べると音素の個人性を考慮していない木構造話者クラスタを用いたほうが良い傾向にあることがわかった。これは VFS に与えられた音声の量が少ないため本来の音素の特徴とは異なった音素モデルが生成されたためであると思われる。その結果、木構造話者クラスタは音素の個人性を反映したものとはいい難くなってしまう。これらのことは音素の個人性を考慮せずに木構造を生成すれば起こら

ないことであるので、音素毎に木構造を作成したことが悪く働いたと考えられる。

また、子音には個人性が無いために男女の差が木構造に反映されなかったとも考えられる。

次に、適応時に選択されるクラスタがどのような話者から構成されているかを、全音素HMnetでの木構造話者クラスタと、音素毎の木構造話者クラスタについて比較した。子音の場合、音素毎の木構造から選択された話者クラスタは、全音素での木構造話者クラスタにおいて選択された話者クラスタに類似する場合が大半となっていた。すなわち、双方ともほぼ同じ話者から構成されているクラスタ、あるいは片方がそのようなクラスタと木構造上で隣接したクラスタであった。これが子音の認識率に改善が見られない原因であると考えられる。

一方母音の場合はそれぞれの音素で異なる話者クラスタを選択していて共通性というものはあまり見られなかった。これは音素毎に話者の個人性を考慮して木構造話者クラスタを作成したことに効果があることを示していると思われる。

3.3.3.2　木構造を構成する話者が多い場合の認識実験

次に木構造を構成する話者が多人数の場合について実験をおこなった。音響分析条件は表2.4、実験条件は表3.3であり、初期HMnetも少人数での実験と同じ条件である。また50文章の音声でVFSを用いて50人分のHMnetを作成した。認識は男性2人と女性2人について50文章でおこなった。適応音声の量は1文章である。ただし、認識対象となる話者、木構造を構成する話者は少人数での実験で使ったものと異なるので少人数でおこなった場合の認識率とこの認識率を直接比較することはできない。

認識結果を 図3.7 に示す。この結果から、少人数の場合と同様に多人数で木構造を構成した場合でも音素毎に話者性を考慮したモデルのほうが良い認識結果を得た。少人数の実験と同じように木構造の構造を調べてゆくと、少人数の場合では上層のほうは男性と女性への分割のように同じような分割がおこなわれ、下層にいくにしたがって次第に異なる分割をしていったが、多人数で構成した木構造話者クラスタの場合には、上層のほうにおいても異なる分割がおこなわれているものが増える傾向にある。

適応時に選択された話者クラスタがどのような話者で構成されているかを調べたところ、少人数の場合と同じような傾向が見られた。つまり、子音において選択された話者クラスタは互いに類似する場合が多く、母音の場合はそれぞれの音素で異なる話者クラスタを選択していて共通性というものはあまり見られなかった。

3.4　まとめ

本章では話者選択による話者適応法のひとつである木構造話者クラスタリングによる話者適応法の問題点について述べ、この改善法として音素毎の木構造話者クラスタリングに

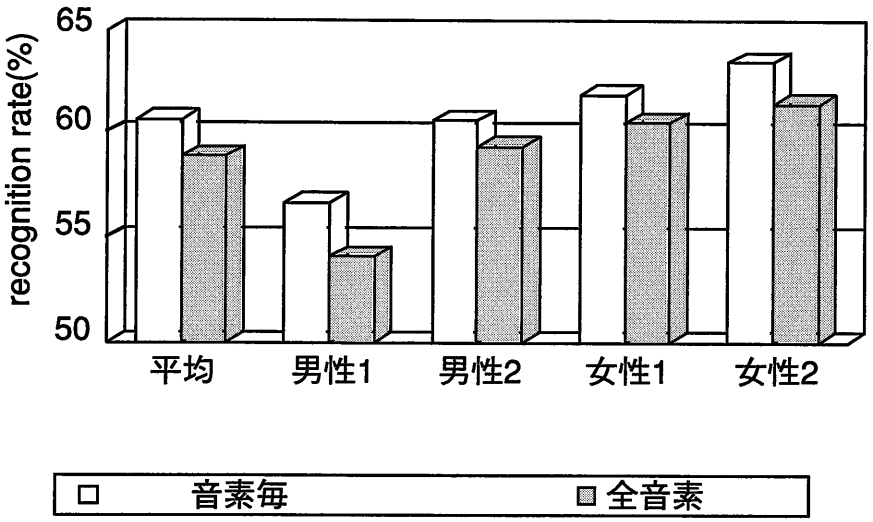

図 3.7: 多人数で構成される木構造話者クラスタによる認識率

基づく話者適応法を提案した。

まず音素毎に木構造を構成することの有効性を調べるために予備実験をおこないその効果を確認した。また、モデルの認識性能を評価するために実験をおこなった結果、木構造を構成する話者が少ない場合でも多い場合でも従来法よりもよい認識性能を得た。これは音素毎に最適な話者クラスタを選択することが有効に働いているためである。

この手法の問題点としては従来法よりも計算量が約 2 倍かかることが挙げられる。また適応音声に存在しない音素に関して、この手法では全音素 HMnet で作成した木構造話者クラスタから全適応音声を用いて選択したクラスタの情報を用いてそのモデルを補間したが、これも改善する必要がある。

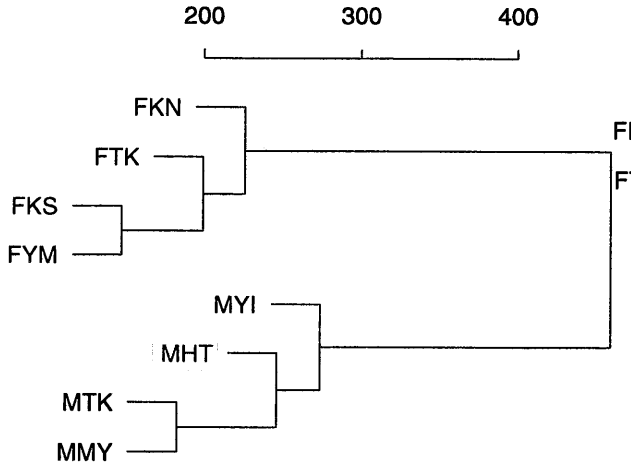

図 3.8: 全音素の木構造

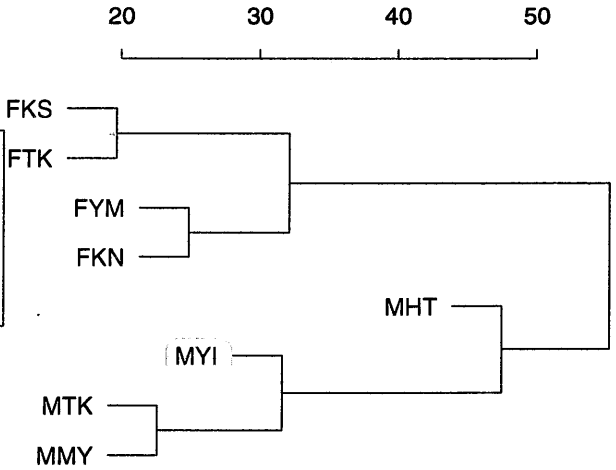

図 3.9: 音素/ a /の木構造

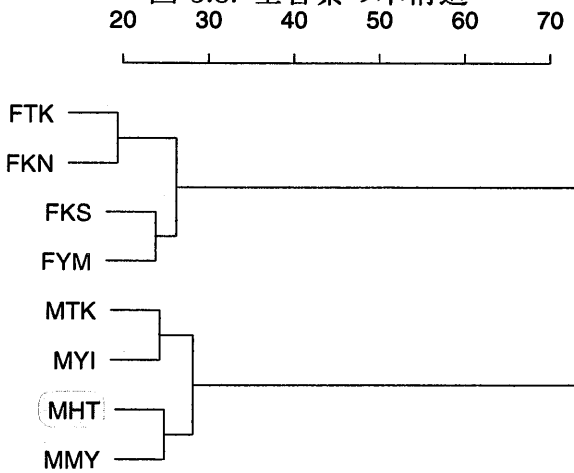

図 3.10: 音素/ i /の木構造

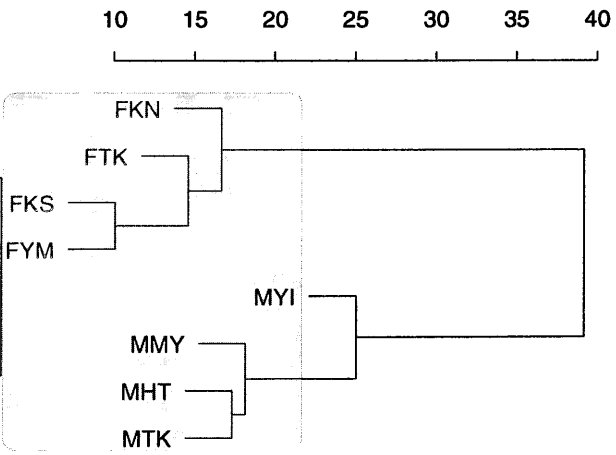

図 3.11: 音素/ u /の木構造

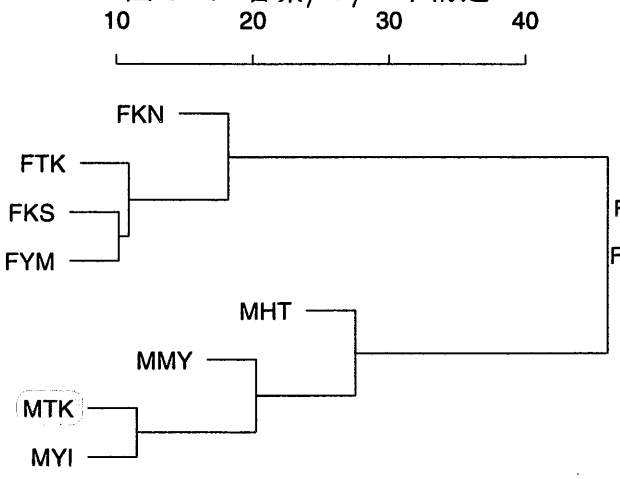

図 3.12: 音素/ e /の木構造

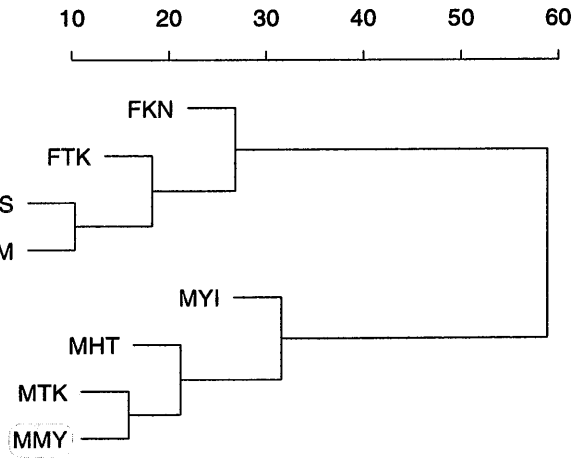

図 3.13: 音素/ o /の木構造

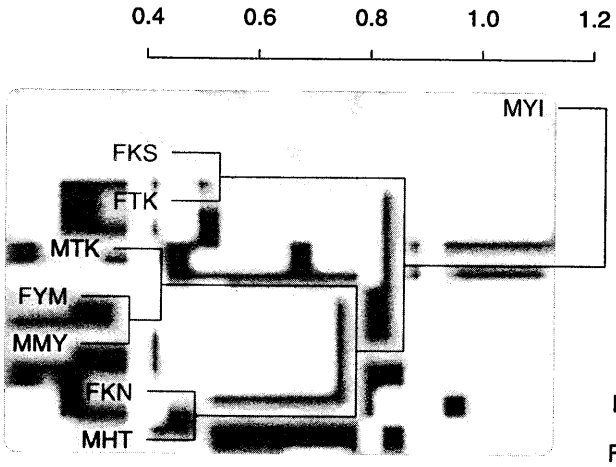

図 3.14: 音素/ p /の木構造

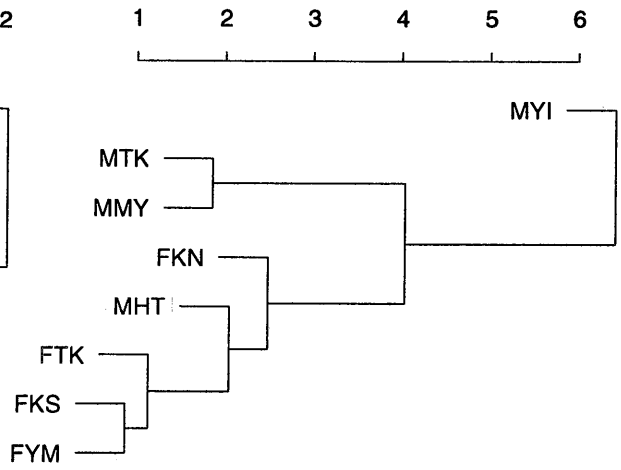

図 3.15: 音素/ b/の木構造

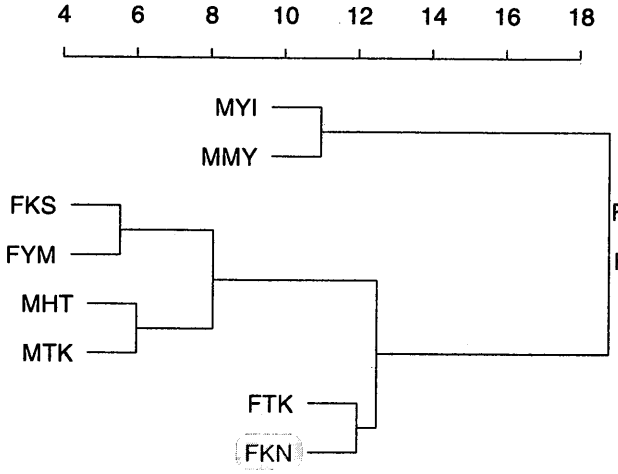

図 3.16: 音素/ t /の木構造

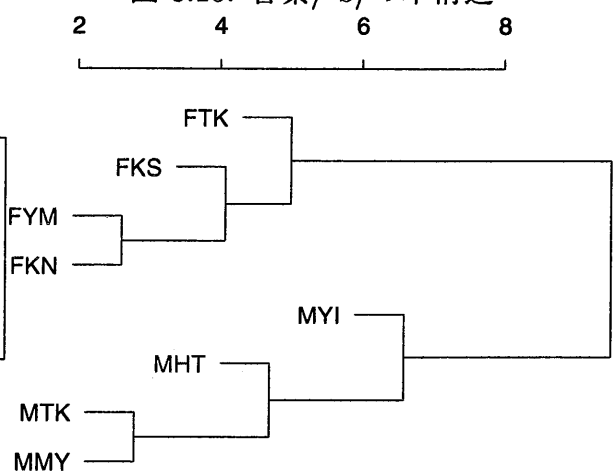

図 3.17: 音素/ d /の木構造

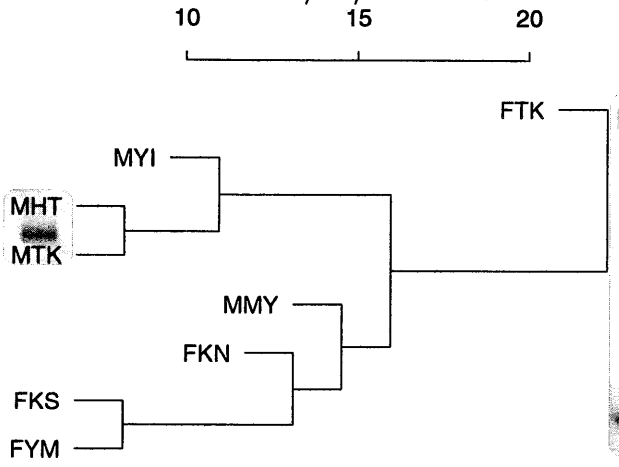

図 3.18: 音素/ k /の木構造

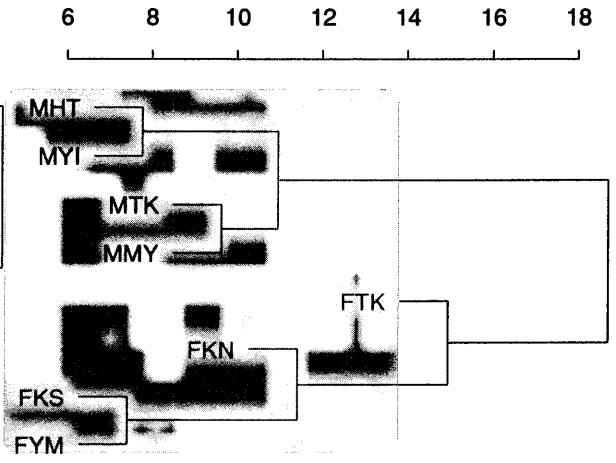

図 3.19: 音素/ g /の木構造

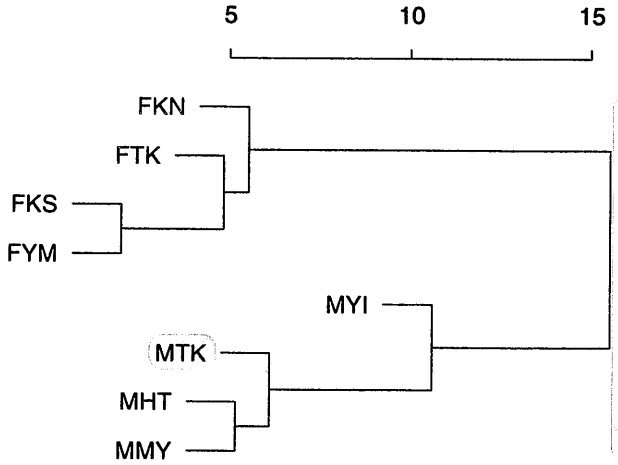

図 3.20: 音素/ m /の木構造

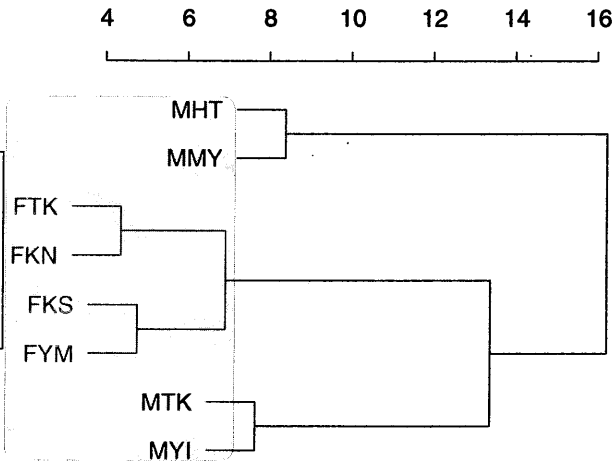

図 3.21: 音素/ N /の木構造

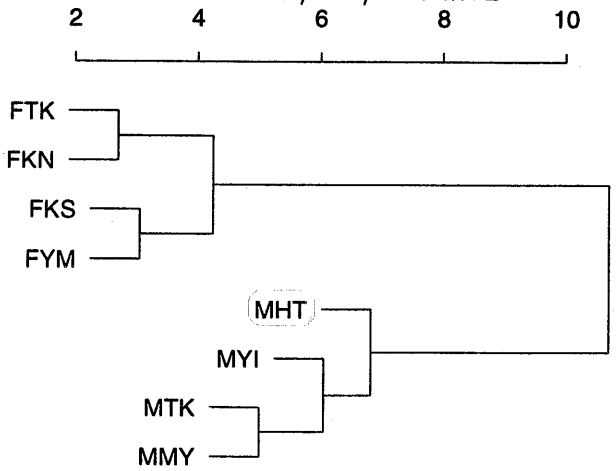

図 3.22: 音素/ r /の木構造

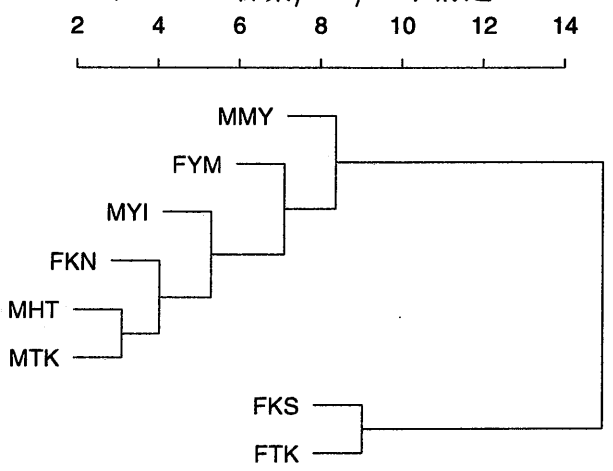

図 3.23: 音素/ s /の木構造

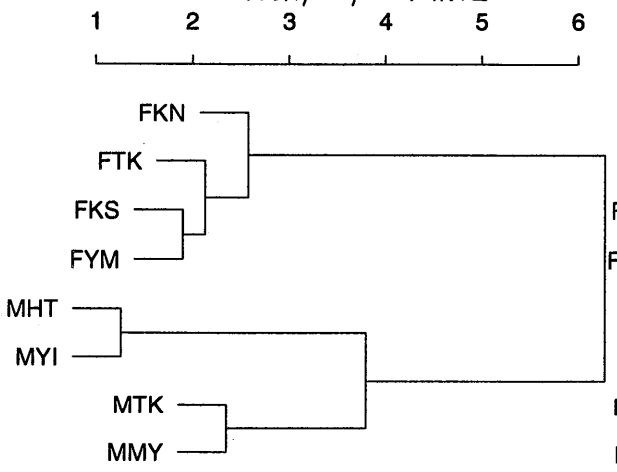

図 3.24: 音素/ z /の木構造

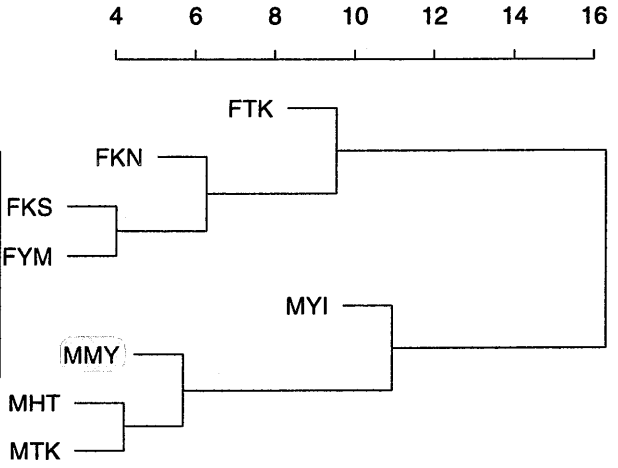

図 3.25: 音素/ sh /の木構造

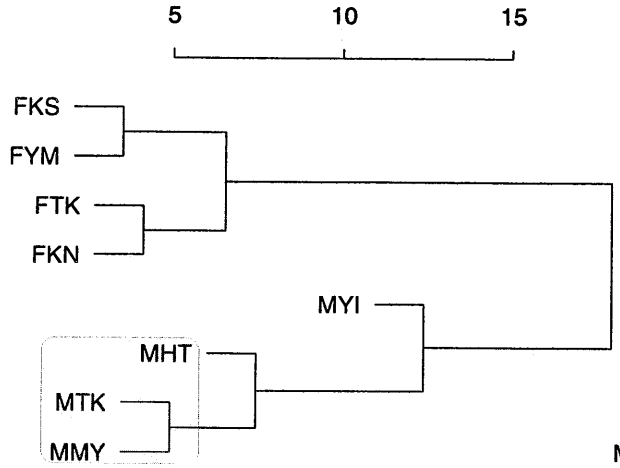

図 3.26: 音素/ j /の木構造

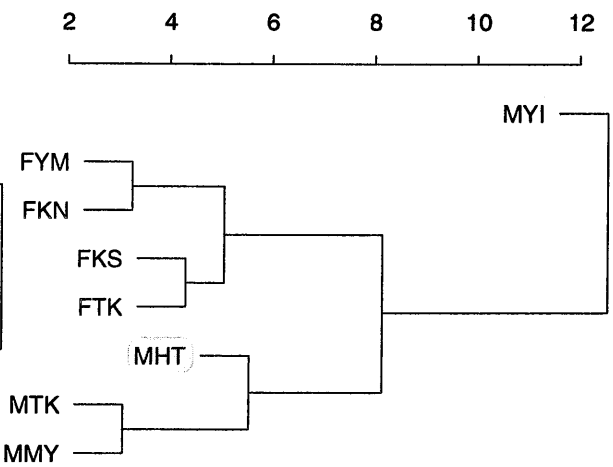

図 3.27: 音素/ h /の木構造

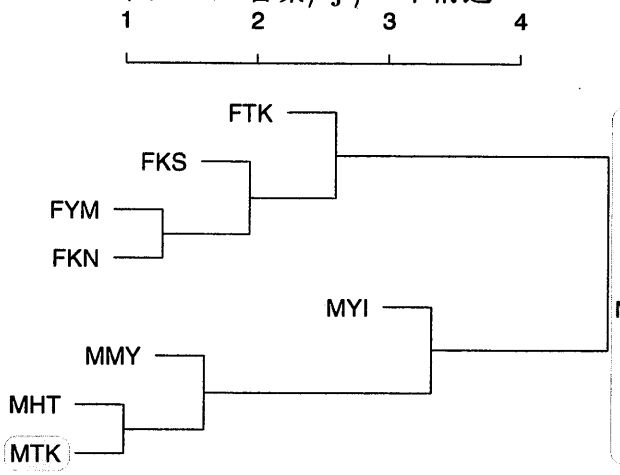

図 3.28: 音素/ ch /の木構造

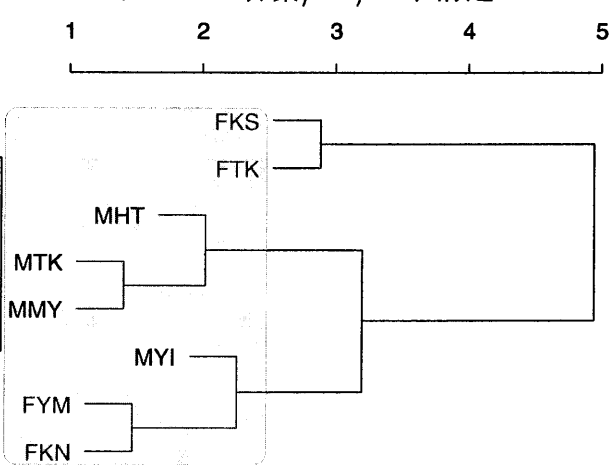

図 3.29: 音素/ ts /の木構造

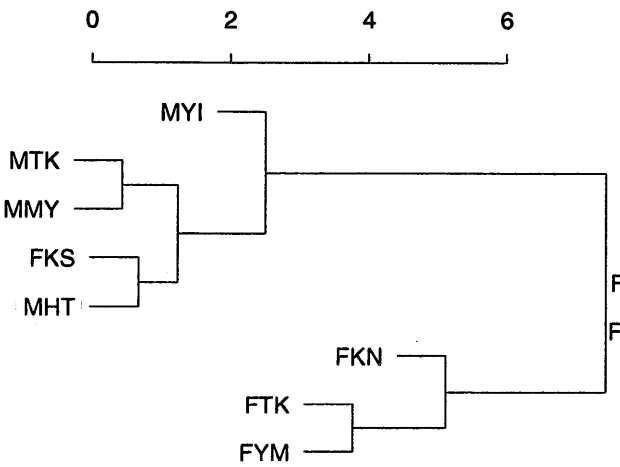

図 3.30: 音素/ w /の木構造

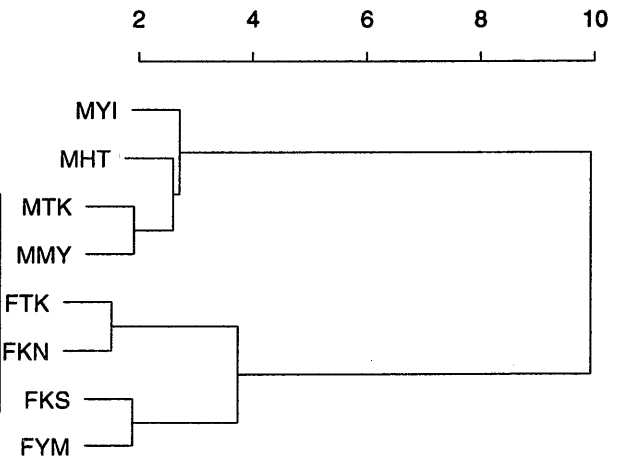

図 3.31: 音素/ y /の木構造

第4章

音素間の距離を用いた木構造話者クラスタリング

4.1 はじめに

前章では話者選択による話者適応法の代表的手法である“木構造話者クラスタリングによる話者適応法”の問題点について述べ、その改善手法として“音素毎の木構造話者クラスタリングに基づく話者適応法”を提案した。この手法は音素毎に木構造話者クラスタを作成することで、従来法よりも話者の個人性を忠実に反映したモデルを構成している。

しかし前章の手法は適応音声中に存在しない音素について、全音素 HMnet での木構造話者クラスタから全ての適応音声によって選択された話者クラスタを用いて、モデルを補間していた。しかし全音素の木構造話者クラスタよりもその音素に近い音素の木構造話者クラスタからモデルを補間したほうが忠実に話者の個人性を反映したモデルを得ることができると考える。

そこで本章では音素間の距離を用いた木構造話者クラスタリング手法を提案し、その話者適応の性能を認識実験をおこない比較・検討してゆく。

4.2 音素間の距離を用いた木構造話者クラスタリング

4.2.1 HMnet を利用した音素間距離の定義

まず最初に各音素間の距離について考える。適応音声に存在しない音素について別の音素のモデルで補間をおこなうのなら“木構造が似ている”ということを近さの基準に考えるのであって、“音響的に似ている”ということを近さの基準に考える必要はない。しかし木構造間の距離を定義することは非常に難しい。

そこで本節では音響的に似ていれば木構造も似ているであろうという予測を取り入れ、各音素間の距離を HMnet のパラメータを用いて定義する。概略は以下の通りである。

1. 各音素 HMnet での中心となるパス (HMM) を選択する。
   本手法では各パスにおいて式 (4.1) を用いて平均歪が最小となるものを音素中心パスとした。

$$\bar{D}(\boldsymbol{M}) = \left(\prod_{i \in \boldsymbol{P}}^{n} m \times D(\boldsymbol{M}, \boldsymbol{M_i}) \right)^{\frac{1}{n}} \tag{4.1}$$

ただし $\boldsymbol{M}$: HMM
$D(\boldsymbol{M_1}, \boldsymbol{M_2})$: HMM $\boldsymbol{M_1}, \boldsymbol{M_2}$ の距離
$\boldsymbol{P}$: 音素 / phone / のパスの集合
n : 音素 / phone / のパスの総数
m : 音素 / phone / のサンプル数

2. 各中心音素 HMM 間の距離を音素間距離とする。

状態数の異なる HMM 間の距離 $D(M_i, M_j)$ は、モンテカルロシミュレーションや HMM の学習で用いた音声サンプルを用いて、尤度計算をすることによって定義することができる。しかし、この方法では非常に計算量が多い。また、今回実験に用いた HMnet の状態数 210 の場合での音素 / a / のパスは 455 であり、 / a / の中心パスを求めるのに $455^2 \approx 200000$ 回の距離計算が必要である。そのため距離計算に要する計算量をなるべく削減したい。

状態数の等しい HMM 間での距離には、この近似式として対応する状態間の Bhattacharyya 距離の総和を取る方法がある。しかし状態数が異なる場合にはこのような近似式は定義されていない。そこで本手法では HMM 間の距離は以下のように DP マッチングを用いて計算をおこなう。

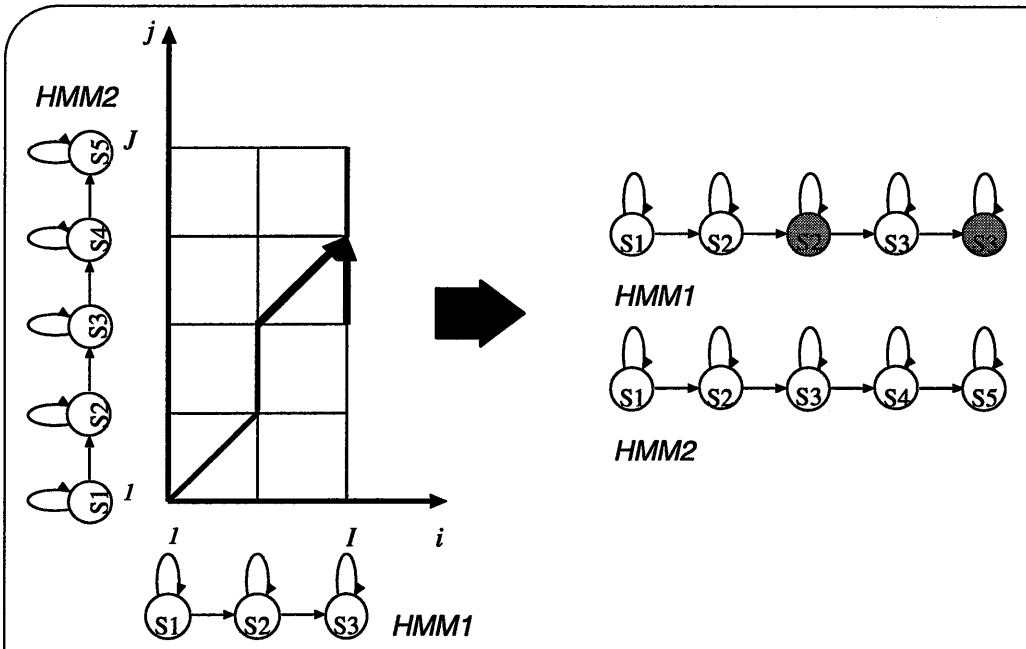

図 4.1: 状態間の異なる距離の定義

$$
\begin{aligned}
d(i,j) &= g(i,j) + \min(d(i,j-1), d(i-1,j-1)) \\
D(\boldsymbol{M_1}, \boldsymbol{M_2}) &= \frac{d(I,J)}{J} \qquad (4.2)
\end{aligned}
$$

ただし

$g(i,j)$: HMM1 の状態 i と HMM2 の状態 j の Bhattacharyya 距離

I : HMM 1 の状態数

J : HMM 2 の状態数 $(J \geq I)$

これは短い状態の HMM を適当に伸長させて長い HMM に対応したものからそれぞれ対応する状態で Bhattacharyya 距離を取ったものである。また式 (4.2) の分母の J は HMM の状態の長さの違いによる距離の違いを正規化するためのものである。

この方法によって定義された音素間の距離を用いて各音素における類似音素を表 4.1 に示す。参考として音韻環境非依存 left-to-right HMM (3 状態 1 混合) から算出した音素間の距離を用いた各音素における類似音素を表 4.2 に示す。本実験で使用した HMnet を学習した音声データは ATR 連続音声データベースであり、対象となる音素は表 2.1 の太字の計 24 種類である。

表 4.1: HMnet を用いた各音素の類似音素

調音様式	音素	1位		2位		3位		4位		5位	
(母音)	a	h	(3.652)	o	(3.707)	w	(3.751)	p	(4.272)	u	(4.848)
	i	y	(2.957)	g	(4.656)	e	(4.701)	N	(4.921)	u	(5.431)
	u	N	(2.389)	o	(2.956)	g	(3.440)	e	(3.751)	b	(3.891)
	e	u	(3.751)	N	(4.141)	g	(4.312)	p	(4.542)	y	(4.656)
	o	u	(2.956)	N	(3.005)	w	(3.088)	a	(3.707)	m	(3.980)
(半母音)	w	o	(3.088)	m	(3.111)	a	(3.751)	h	(3.940)	n	(4.371)
	y	i	(2.957)	e	(4.656)	r	(5.003)	g	(5.013)	N	(5.331)
破裂音	p	r	(1.743)	h	(1.810)	k	(1.928)	g	(1.956)	t	(2.338)
	b	g	(1.129)	d	(1.618)	r	(1.879)	p	(2.349)	h	(2.929)
	t	k	(1.857)	p	(2.338)	r	(2.389)	d	(2.877)	ch	(2.934)
	d	b	(1.618)	r	(1.662)	g	(1.912)	z	(2.578)	p	(2.741)
	k	t	(1.857)	p	(1.928)	h	(2.160)	r	(3.147)	b	(3.236)
	g	b	(1.129)	N	(1.821)	r	(1.845)	d	(1.912)	p	(1.956)
鼻音	m	n	(1.203)	N	(2.443)	g	(2.969)	w	(3.111)	b	(3.294)
	n	m	(1.203)	N	(2.186)	g	(3.180)	b	(3.749)	u	(4.336)
	N	g	(1.821)	n	(2.186)	u	(2.389)	m	(2.443)	o	(3.005)
ふるえ音	r	d	(1.662)	p	(1.743)	g	(1.845)	b	(1.879)	t	(2.389)
摩擦音	s	ts	(1.244)	z	(2.078)	sh	(2.729)	ch	(2.859)	t	(3.793)
	z	ts	(1.292)	s	(2.078)	j	(2.230)	d	(2.578)	ch	(2.825)
	sh	ch	(1.281)	j	(2.729)	s	(2.729)	ts	(2.897)	z	(3.345)
	j	ch	(2.068)	z	(2.230)	sh	(2.729)	d	(2.850)	ts	(3.096)
	h	p	(1.810)	k	(2.160)	b	(2.929)	g	(2.955)	r	(3.637)
	ch	sh	(1.281)	j	(2.068)	ts	(2.125)	z	(2.825)	s	(2.859)
	ts	s	(1.244)	z	(1.292)	ch	(2.125)	sh	(2.897)	j	(3.096)

表 4.2: 音韻環境非依存 HMM を用いた各音素の類似音素

調音様式	音素	1位		2位		3位		4位		5位	
(母音)	a	o	(0.460)	e	(1.101)	u	(1.498)	N	(2.101)	m	(2.357)
	i	u	(1.032)	N	(1.250)	e	(1.422)	o	(1.883)	n	(2.881)
	u	N	(0.763)	o	(0.807)	e	(0.807)	i	(1.032)	a	(1.498)
	e	u	(0.807)	o	(0.877)	a	(1.101)	i	(1.422)	N	(1.529)
	o	a	(0.460)	u	(0.807)	e	(0.877)	N	(1.194)	m	(1.727)
(半母音)	w	h	(1.583)	g	(3.020)	m	(5.079)	b	(5.658)	r	(6.050)
	y	r	(2.406)	g	(2.497)	sh	(3.046)	j	(3.221)	e	(3.313)
破裂音	p	k	(0.338)	b	(0.527)	t	(1.080)	sh	(1.451)	r	(1.872)
	b	p	(0.527)	k	(0.776)	t	(1.131)	r	(1.277)	sh	(1.382)
	t	k	(0.660)	d	(0.905)	sh	(0.941)	p	(1.080)	b	(1.131)
	d	t	(0.905)	sh	(1.094)	r	(1.216)	ch	(1.289)	z	(1.328)
	k	p	(0.338)	t	(0.660)	b	(0.776)	sh	(1.088)	d	(1.412)
	g	h	(1.259)	r	(1.494)	m	(1.510)	n	(2.000)	b	(2.084)
鼻音	m	n	(1.005)	g	(1.510)	r	(1.720)	o	(1.727)	b	(1.914)
	n	m	(1.005)	o	(1.772)	r	(1.956)	g	(2.000)	sh	(2.114)
	N	u	(0.763)	o	(1.194)	i	(1.250)	e	(1.529)	a	(2.101)
ふるえ音	r	sh	(0.915)	d	(1.216)	b	(1.277)	t	(1.370)	g	(1.494)
摩擦音	s	ts	(0.638)	z	(0.824)	t	(1.236)	d	(1.382)	sh	(1.869)
	z	s	(0.824)	ts	(0.944)	d	(1.328)	t	(1.718)	sh	(2.025)
	sh	ch	(0.392)	j	(0.654)	r	(0.915)	t	(0.941)	k	(1.088)
	j	ch	(0.279)	sh	(0.654)	d	(1.787)	r	(1.814)	k	(2.361)
	h	g	(1.259)	w	(1.583)	b	(2.743)	r	(2.773)	m	(2.944)
	ch	j	(0.279)	sh	(0.392)	d	(1.289)	r	(1.630)	k	(1.685)
	ts	s	(0.638)	z	(0.944)	d	(2.029)	ch	(2.067)	sh	(2.465)

この結果を見るとHMnetを用いた各音素の類似音素は、音声スペクトルを用いて求めた音素間の類似度と必ずしも一致しない。むしろ音韻環境非依存のHMMを用いた距離のほうが近い関係にある。

しかしこの距離はHMnetの構成に則した距離、つまり話者適応の結果に効果がでれば良いので一致する必要はない。

4.2.2 音素間の距離を用いた木構造話者クラスタリングに基づく話者適応アルゴリズム

前節で定義した音素間の距離を用いた木構造話者クラスタリングに基づく話者適応アルゴリズムを以下に示す。木構造の構成のほうは、前章で提案した手法とほとんど変わらない。まず、あらかじめ1人の話者の多数の音声サンプルを用いて特定話者HMnetを学習する。その特定話者HMnetに多量の話者のデータでVFSをおこない擬似的に複数の話者の特定話者HMnetを作成する。

次に各音素HMnet毎に前章の方法を用いて木構造話者クラスタを作成する。ただし前章では全音素HMnetの木構造話者クラスタを一緒に作成したが本手法では作成しない。

認識対象となる話者の音声を使用して、各音素毎に最適な話者クラスタを選択する。適応音声中に存在する各音素に関して、その音素の木構造話者クラスタを用いて以下の手順で木構造を探索してモデルを選ぶ。

1. クラスタを最上層にマークし、このクラスタのHMnetに適応音声を与え尤度を求める。
2. マークされたクラスタに属するサブクラスタのHMnetに適応音声を与え、最尤となるHMnetを持つサブクラスタを選択する。
3. 選択されたサブクラスタにマークし、クラスタにサブクラスタが存在しなくなるまで2を続ける。
4. 1～3の処理によって全ての階層で1つのクラスタが選択される。その全てのクラスタのHMnetの中から更に最尤となるHMnetを選択し、これを認識に用いる。
   ただし適応用音声に存在しない音素に関しては、前節で求めた音素間の距離テーブルを用いて、適応音声中に存在する音素の中で一番距離の近いものを選ぶ。一番近い音素で選択された話者クラスタの話者情報を抜き出し、適応音声に存在しなかった音素HMnetの各状態の出力確率密度関数の平均値を再合成する。

$$b^{(i)}(x) = \sum_{s} \frac{n_s^{p(i)}}{\sum_{s'} n_{s'}^{p(i)}} N(\mu_s^{(i)}, \Sigma_s^{(i)}) \tag{4.3}$$

ただしiは状態番号、s、s'は一番近い音素で選択された話者クラスタに含まれる話

者、$n_s^{p(i)}$ は話者 s の HMnet の状態 i に対応する音素 p のサンプル数を表わす。
この処理により、適応音声に存在しなかった音素のモデルの情報を補間し認識をおこなう。

この手法では音素毎に木構造話者クラスタを構成し、音素毎に最適な木構造話者クラスタを選択するので、更に忠実に話者の個人性を反映したモデルを構成することが可能である。また適応音声中に存在しない音素に関しても、その音素モデルに特徴が近いモデルの情報を用いて補間をおこなうので、前章で提案した手法に比べ話者の個人性を反映したモデルを構成することが可能であると考えられる。

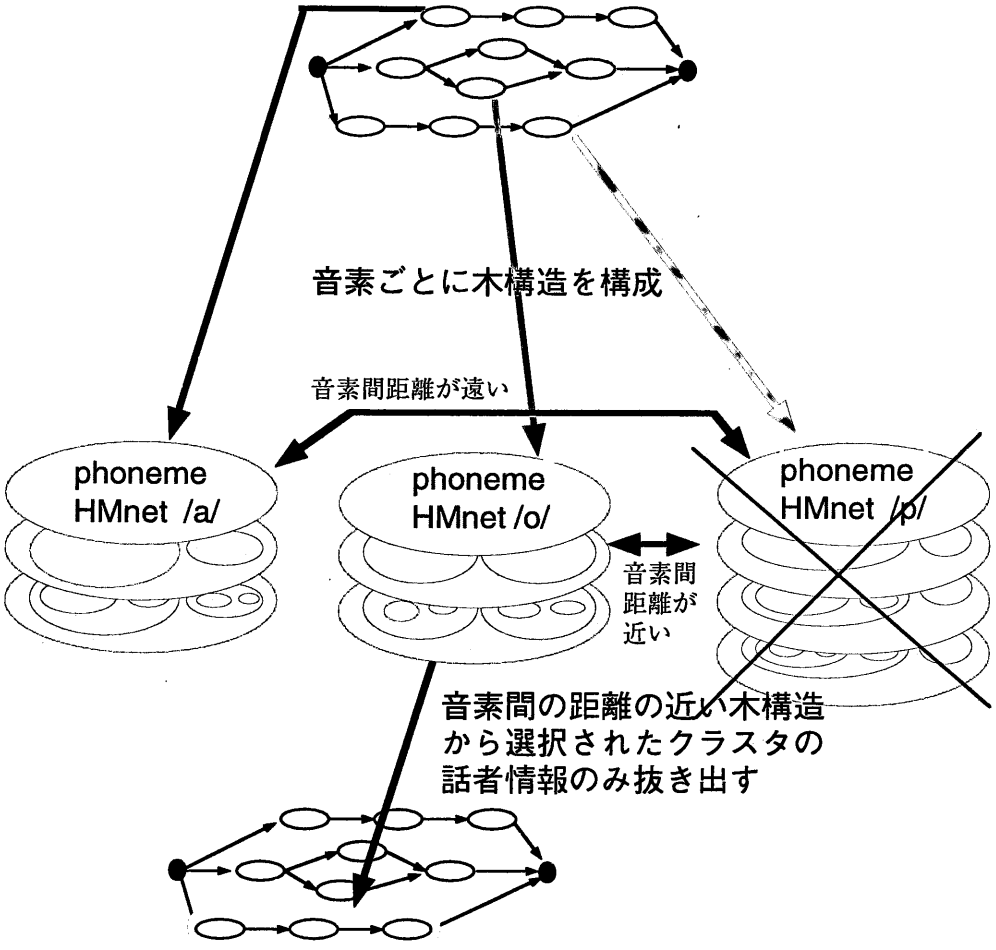

図 4.2: 音素間の距離を用いた木構造話者クラスタリングによる話者適応法の概要

4.3　性能評価実験及び考察

本章で提案した手法の話者適応の性能を調べるために、音素認識実験をおこなった。今回比較したのは、3章で提案した“音素毎の木構造話者クラスタリングに基づく話者適応法”と、小坂らが提案した“木構造話者クラスタリングによる話者適応法”のふたつである。

音響分析条件は表2.4、HMnetの条件は表4.3のとおりである。初期HMnetはATR日本語音声データベース音韻バランス503文章から男性1名が発話した400文章を用いて学習した。また50文章の音声でVFSを用いて8人分のHMnetを作成した。認識には男性4人と女性4人について50文章を用いた。木構造話者クラスタを選択する適応音声の量は1文章である。また不特定話者モデルには、認識対象となる話者と学習に用いた話者のモデルは含まれていない。

表 4.3: 実験条件

初期HMnet	状態数 210
	特定話者モデル
	SSS-freeで学習
各状態の分布	単一ガウス分布
	対角共分散行列

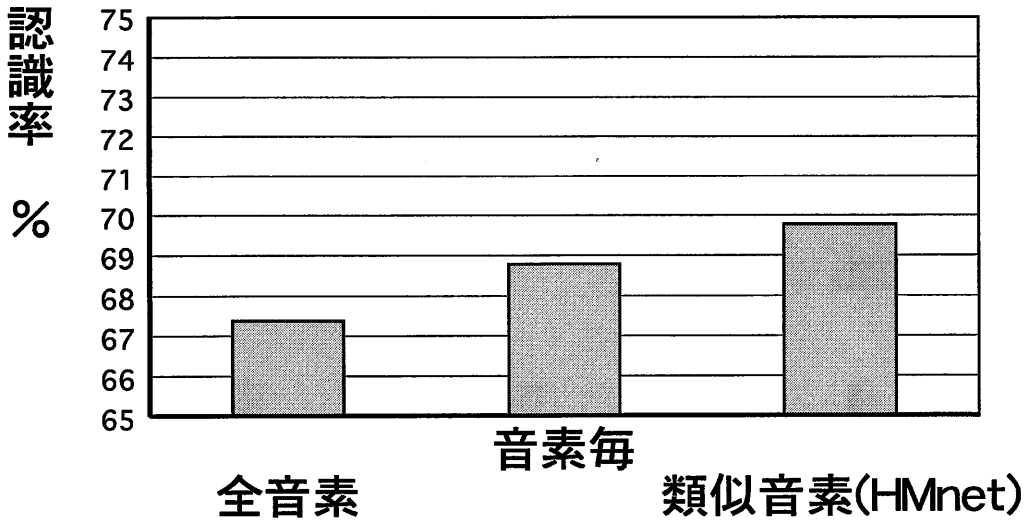

図 4.3: 認識実験結果

認識実験の結果を 図4.3に示す。実験結果は8人の平均である。今回提案した手法は、

従来法に比べ約2.5%、前章の結果に比べ約1%の認識率の向上が見られた。これは、適応音声に存在しなかった音素について、全音素HMnetで補間したモデルに比べ、特徴の近い音素のモデルを用いて補間したほうが話者の個人性を反映した結果であると考えられる。

また適応音声に存在しない音素は全て子音であり、本章で提案した手法によって補間されたモデルは全て子音ということになる。前章では従来法に比べ子音の認識率は改善されなかったが、本章で提案した手法によって子音にも認識率の向上が見られるようになった。

4.4　まとめ

本章ではHMnetの情報を用いて音素間の距離を定義した。またこの距離を用いて音素毎の木構造話者クラスタリングによる話者適応法の改善法を提案した。性能評価実験をおこなった結果、従来法より約2.5%、前章で提案した方法より約1%の認識性能の向上が見られた。前章の手法では、改善できなかった子音の認識率が向上したためである。

第5章

結論

5.1 本研究の成果

本研究において以下の成果が得られた。

第2章では、モデルパラメータ調整による話者適応法の代表的手法である移動ベクトル場平滑化による話者適応法を音響的類似度に基づく HMnet に適応する手法を提案した。逐次状態分割法によって生成された HMnet に VFS を適応したものと音素認識性能を比較した結果、適応文章数が比較的少ない場合に関しては本手法のほうが最大約3%の認識性能の向上が見られた。

これは SSS-free による HMnet では、1つの音素環境に対して複数のパスが存在することが有利に働いたためと考えられる。

第3章では、話者選択による話者適応法の代表的手法である木構造話者クラスタリングに基づく話者適応法が話者の個人性を全ての音素に対して同じものとして捉えているという問題点を改善する手法を提案した。その話者適応の性能を見るために音素認識実験をおこなった結果、母音に対しては約2%の認識率の向上が見られた。

これは、音素毎に木構造話者クラスタを構成することによって、話者適応時に各々の音素について最適な話者クラスタを選択することが有利に働いたためである。

第4章では、第3章で提案した問題点を挙げ、その改善手法として、音素間の距離を用いた音素毎の木構造話者クラスタリングによる話者適応法を提案した。音素認識実験をおこなった結果、従来法に比べ約2.5%、第3章の方法に比べ約1%の認識性能の向上が見られた。また前章の方法では、子音に関しては認識性能の向上が見られなかったが、本章の手法によって子音の認識率も改善することができた。

5.2 今後の課題

今後の課題としては次の点が挙げられる。

第3章、第4章で提案した手法では、小坂らの提案した“木構造話者クラスタリングによる話者適応法”よりも計算量が多いため改善の必要がある。

また話者適応法において、利用者に最も負担をかけないのは教師なしの話者適応である。本手法を拡張した教師なし話者適応法について検討してゆく必要がある。

謝辞

本研究を進めるにあたり、多大な御指導とともにこの研究の機会を与えて下さった東北大学大学院工学研究科教授 阿曽弘具氏 に心から感謝いたします。

また東北大学大学院電気通信研究所教授 矢野雅文氏 におきましては、異なる視点から私の気づかない点を御指摘いただきましたことを深く感謝いたします。

また音声ゼミにおきましては東北大学大型計算機センター教授 牧野正三氏、東北大学大学院工学研究科助教授 木幡稔氏に全面的な御指導をいただきましたことを深く感謝いたします。

大学院ゼミにおきましては東北大学大学院情報科学研究科教授 曽根敏夫氏、助教授 鈴木陽一氏、助手 小沢賢司氏、助手 高根昭一氏に貴重な御意見をいただきましたことを感謝いたします。

日々の研究におきましては昼夜を問わず御指導していただいた東北大学大型計算機センター助手 鈴木基之氏、貴重な御意見、御助言をいただいた東北大学大学院工学研究科助手 大町真一郎氏、東北大学情報処理教育センター助手 後藤英昭氏、広島大学総合科学部助手 黒岩丈介氏、東北大学大学院工学研究科 森大毅氏、日本放送協会 佐藤英氏、東北大学大学院工学研究科 佐藤俊治氏、菅谷至寛氏ならびに阿曽研究室の皆様に感謝いたします。

参考文献

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

[2] 新美康永:“音声認識”, 共立出版.

[3] Lawrence Rabiner, Biing-Hwang Juang 共著, 古井貞熙 監訳:“音声認識の基礎 (上),(下)” NTT アドバンステクノロジ株式会社.

[4] 服部四郎:“音声学”, 岩波書店.

[5] 安居院猛, 中島正之:“コンピュータ音声処理”, 産報出版.

[6] 鷹見淳一, 嵯峨山茂樹: “逐次状態分割法による隠れマルコフ網の自動生成”, 電子情報通信学会論文誌, J76-D-II,No.10,pp. 2155–2164 (1993).

[7] Motoyuki Suzuki, Shozo Makino, Akinori Ito, Hirotomo Aso and Hiroshi Shimodaira : “A New HMnet Construction Algorithm Requiring No Contextual Factors”, IEICE Trans. Inf. & Syst., E78-D, No.6, pp. 662–668 (1995).

[8] Hiroaki HATTORI and Shigeki SAGAYAMA: “Speaker Adaptation based on Vector Field Smoothing”, 信学技報, SP 92-15 (1992).

[9] 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, SP 92-16 (1992).

[10] 高橋淳一, 嵯峨山茂樹: “最大事後確率推定と移動ベクトル場平滑化の組み合わせによる高速話者適応化”, 平成 6 年度日本音響学会秋季講演会論文集, 2-8-19

[11] 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”, 平成 6 年度日本音響学会秋季講演会論文集, 2-8-20

[12] 中村哲、鹿野清宏 : “ファジィベクトル量子化を用いたスペクトログラムの検討” 信学技報, SP 87-123 (1987).

[13] 丸山活輝, 花沢利行, 川端豪, 鹿野清宏: "HMM 音韻連結学習を用いた英単語音声の認識, 信学技報, SP88-119 (1988).

[14] 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).

[15] Tetsuo Kosaka, Shigeki Sagayama:"Automatic Determination of the Number of Mixture Components for Continuous HMMs Based on a Uniform Variance Criterion", IEICE Transactions Inf. & Syst., E78-D, No.6, pp. 642–647 (1995).

[16] 小坂哲夫, 鷹見淳一, 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 A,Vol.J77-(A), No.2, pp. 103–111 (1994).

[17] 管村昇, 相川清明, 鹿野清宏, 好田正紀: "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", 日本音響学会研究会資料, S82-64 (1982).

[18] 高木幹雄, 下田陽久 監修: "画像解析ハンドブック", 東京大学出版会

[19] 粟津辰功, 牧野正三: "認識結果を利用した教師なし話者適応に関する研究", 日本音響学会平成 5 年度秋季研究発表会 2-7-15 pp. 99–100 (1993).

[20] Keisuke Fukunaga:"Introduction to Statistical Pattern Recognition", ACADEMIC PRESS New York and London (1972)

研究業績一覧

1. “SSS-free に基づく HMnet における VFS の効果”,
   阿部俊朗, 鈴木基之, 牧野正三, 阿曽弘具,
   日本音響学会平成 8 年度秋季発表会, 3-3-20 (1996.9).

2. “音素毎の話者クラスに基づく話者適応法”,
   阿部俊朗, 鈴木基之, 牧野正三, 阿曽弘具,
   日本音響学会平成 9 年度秋季発表会, 1-1-14 (1997.9).

3. “音素毎の話者クラスタリングに基づく話者適応法”,
   阿部俊朗, 鈴木基之, 牧野正三, 阿曽弘具,
   電子情報通信学会音声研究会, SP97-74, pp. 41–47 (1997.12).

4. “音素毎の話者クラスタリングに基づく話者適応法”,
   阿部俊朗, 鈴木基之, 牧野正三, 阿曽弘具,
   東北大学通信工学研究所 第 296 回音響工学研究会, (1998.1).

付録 A

Viterbi アルゴリズム

Viterbi アルゴリズムとは、あるラベル系列Oが与えられときそれを出力する際、効率よくその最適状態遷移系列を推定するためのアルゴリズムであり以下のように与えられる。

1. 初期化
   全てに状態iに対して$\boldsymbol{f}(i,0)=\pi_i$
2. $t=1,2,\cdots,T$ に対し 3 , 4 を実行
3. 全ての状態iに対し 4 を実行
4.

$$\begin{aligned}\hat{j} &= \underset{j(\text{ナル遷移を除く})}{\operatorname{argmax}} \boldsymbol{f}(j,t-1)a_{ji}\cdot b_{ji}(y_t)\\ \hat{k} &= \underset{k(\text{ナル遷移})}{\operatorname{argmax}} \boldsymbol{f}(k,t)a_{ki}\\ \boldsymbol{f}(i,t) &= \max\{\boldsymbol{f}(\hat{j},t-1)a_{\hat{j}i}\cdot b_{\hat{j}i}(y_t), \boldsymbol{f}(\hat{k},t)a_{\hat{k}i}\}\end{aligned}$$

   $\boldsymbol{Q}(i,t)=\boldsymbol{Q}(\hat{j},t-1)\otimes\hat{j}; \boldsymbol{f}(i,t)=\boldsymbol{f}(\hat{j},t-1)a_{\hat{j}i}\cdot b_{\hat{j}i}(y_t)$ のとき
   $\boldsymbol{Q}(i,t)=\boldsymbol{Q}(\hat{k},t)\otimes\hat{k}; \boldsymbol{f}(i,t)=\boldsymbol{f}(\hat{k},t)a_{\hat{k}i}$のとき
5.

$$\begin{aligned}\hat{i} &= \underset{i\in F}{\operatorname{argmax}} f(i,t)\\ P''(y_1,y_2,\cdots,y_T) &= P(y_1,y_2,\cdots,y_T,\boldsymbol{Q}(\hat{i},T))=\boldsymbol{f}(\hat{i},T)\end{aligned}$$

ここで$\otimes$は状態遷移系列と状態を連結し、新たな状態遷移系列を作る演算子である。最適状態遷移系列は$\boldsymbol{Q}(\hat{i},T)$で与えられる。また実際の計算時には、$\log P''(y_1,y_2,\cdots,y_T)=\max\log\boldsymbol{f}(i,T)$のように変更すればアンダーフローの問題がなくなる。

付録 B

Baum-Welchアルゴリズム

Baum-Welch アルゴリズムは与えられたラベル系列 O に対して、HMM λ の尤度 $P(\boldsymbol{O}|\lambda)$ を最大にするために、パラメータ (遷移確率、出力確率) に適当な初期値を与えて反復計算することでパラメータを再推定させるアルゴリズムである。またこのアルゴリズムは Forward-Backward アルゴリズムとも言われ、その名のとおり Forward アルゴリズムと Backward アルゴリズムからなる。

B.1 Forwardアルゴリズム

1. 初期化 全ての状態 i に対して、$\alpha(i,0) = \pi_i$
2. $t = 1, 2, \cdots, T$ に対し 3 , 4 を実行
3. 全ての状態 i に対し 4 を実行
4. $\alpha(i,t) = \sum_{j(\text{ナル遷移を除く})} \alpha(j,t-1)a_{ji}b_{ji}(y_t) + \sum_{j(\text{ナル遷移})} \alpha(j,t)a_{ji}$
5. $P(y_1, y_2, \cdots, y_T) = \sum_{i \in F} \alpha(i,T)$

Viterbi アルゴリズムと Forward アルゴリズムの違いは尤度計算のときの漸化式の更新で Viterbi アルゴリズムは前の最大値をとったが、Forward アルゴリズムはそれらの和をとるということである。

B.2 Backwardアルゴリズム

1. 初期化 $\beta(i,T) = 1$ if $i \in F$; $\beta(i,T) = 0$ if $i \notin F$
2. $t = T, T-1, \cdots, 2, 1, 0$ に対して 3, 4 を実行

3. 全ての状態 i に対して 4 を実行

4. $\beta(i,t) = \sum_{j(\text{ナル遷移を除く})} \beta(j,t+1)a_{ij}b_{ij}(y_{t+1}) + \sum_{j(\text{ナル遷移})} \beta(j,t)a_{ij}$

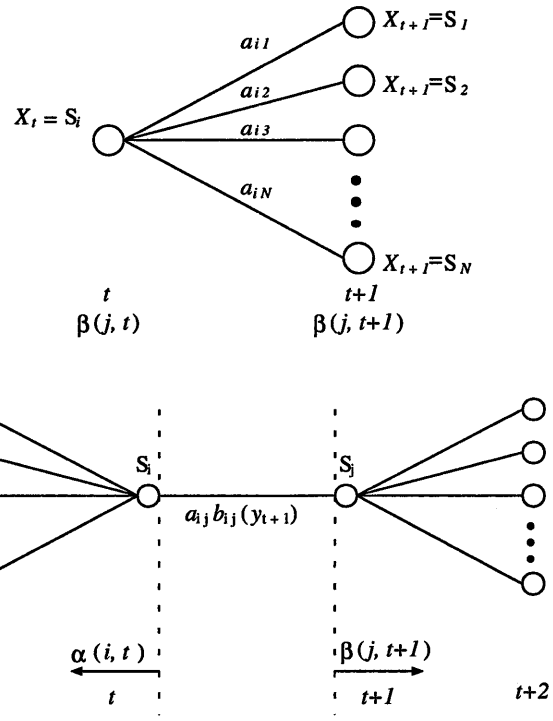

図 B.1: $\beta(i,t)$ の計算手順と $\alpha(i,t)$ と $\beta(j,t+1)$ の関係

図 B.1 は、$\beta(j,t+1)$ の計算手順と $\alpha(i,t)$、$\beta(i,t)$ の関係の概念図を示している。
定義から明らかなように、

$$\sum_{i\in F}\alpha(i,T) = \sum_{i}\beta(i,0)\pi_i \tag{B.1}$$

ここで、π_i は状態 i の初期確率である。このとき、次の関係式が成立する。

$$P(\mathbf{Y}_1^T = \mathbf{y}_1^T, \mathbf{X}_{t-1} = i, \mathbf{X}_t = j) = \alpha(i,t-1)\cdot a_{ij}\cdot b_{ij}(y_t)\cdot\beta(j,t) \tag{B.2}$$

$$\beta(i,t) = \sum_j a_{ij}b_{ij}(y_{t+1})\beta(j,t+1) \tag{B.3}$$

この α、β を用いて HMM のパラメータ推定をおこなう場合には次の式を用いる。ただし、ここでは出力確率を連続分布の多次元無相関正規分布と仮定する。確率密度関数は $\boldsymbol{\theta} = \{\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2\}$ として

$$b_{ij}(\boldsymbol{y}, \boldsymbol{\mu}_{ij_{ij}}, \boldsymbol{\sigma}_{ij}^2) = \prod_k \left(\frac{1}{2\pi\sigma_{ij}^2}\right)^{\frac{1}{2}} e^{-\sum_k \frac{1}{2\pi\sigma_{ijk}^2}(y_k - \mu_{ijk})^2} \tag{B.4}$$

で与える。ここで添字 k はベクトルの第 k 要素を示す。

このとき、最尤推定値を求めるための再推定式は

$$\hat{a}_{ij} = \frac{\sum_t \alpha(i,t-1)a_{ij}\beta(j,t)b_{ij}(\boldsymbol{y}_t, \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2)}{\sum_t \alpha(i,t)\beta(i,t)} \tag{B.5}$$

$$\hat{\boldsymbol{\mu}}_{ij} = \frac{\sum_t \alpha(i,t-1)a_{ij}\beta(j,t)b_{ij}(\boldsymbol{y}_t, \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2)\boldsymbol{y}_t}{\sum_t \alpha(i,t-1)a_{ij}\beta(j,t)b_{ij}(\boldsymbol{y}_t, \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2)} \tag{B.6}$$

$$\hat{\sigma}_{ijk}^2 = \frac{\sum_t \alpha(i,t-1)a_{ij}\beta(j,t)b_{ij}(\boldsymbol{y}_t, \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2)(y_{tk} - \mu_{ijk})^2}{\sum_t \alpha(i,t-1)a_{ij}\beta(j,t)b_{ij}(\boldsymbol{y}_t, \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}^2)} \tag{B.7}$$

で与えられる。

もし、$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{mi} = \boldsymbol{\mu}_{ni}$、$\boldsymbol{\sigma}_i^2 = \boldsymbol{\sigma}_{mi}^2 = \boldsymbol{\sigma}_{ni}^2$ なら、すなわち状態遷移先の状態のみに依存する場合には、

$$\hat{\boldsymbol{\mu}}_{ij} = \frac{\sum_t \alpha(i,t-1)\beta(j,t)\boldsymbol{y}_t}{\sum_t \alpha(i,t-1)\beta(j,t)} \tag{B.8}$$

$$\hat{\sigma}_{ijk}^2 = \frac{\sum_t \alpha(i,t-1)\beta(j,t)(y_{tk} - \mu_{ijk})^2}{\sum_t \alpha(i,t-1)\beta(j,t)} \tag{B.9}$$

となる。

付録 C

音声データベースの音素の分布

ATR 連続音声データベースは A ～ I セットが各 50 文章、J セットが 53 文章の計 503 文章から構成される。表 C.1、表 C.2 にその分布を示す。

表 C.1: ATR 連続音声データベースの音素分布 (その 1)

調音様式	音素	A	B	C	D	E	F	G	H	I	J	合計
母音	a	334	366	369	395	366	426	386	348	338	301	3629
	i	271	295	305	325	304	327	307	274	269	246	2923
	u	307	317	331	342	305	346	324	279	237	201	2991
	e	185	193	200	185	184	201	209	177	181	142	1858
	he	3	4	3	4	2	4	5	3	1	0	29
	o	268	286	284	315	290	334	335	294	261	215	2883
	wo	29	34	37	41	32	31	33	32	30	23	322
半母音	w	7	10	10	18	12	18	14	9	12	12	122
	y	47	46	42	44	44	50	43	39	27	32	414
鼻音	m	74	78	75	90	84	106	88	75	77	72	819
	my	3	0	0	0	0	1	0	1	1	1	7
	n	110	122	129	142	128	154	140	127	121	101	1277
	ny	4	5	3	4	2	2	2	2	3	1	28
	N	101	91	98	108	94	105	96	72	79	56	900

表 C.2: ATR 連続音声データベースの音素分布 (その 2)

調音様式	音素	A	B	C	D	E	F	G	H	I	J	合計
ふるえ音	r	112	118	127	129	117	124	134	111	113	103	1188
破裂音	ry	5	5	4	4	5	5	6	3	0	4	41
	p	13	6	5	8	5	7	7	5	6	4	66
	pp	8	8	8	5	6	8	5	6	3	4	64
	ppy	0	1	1	1	0	1	0	0	0	1	5
	py	1	0	0	0	1	0	0	0	0	0	2
	b	37	37	40	40	37	38	36	32	36	26	359
	by	1	0	1	1	2	1	2	1	1	2	12
	t	85	101	104	113	98	101	102	95	86	72	957
	tt	17	21	21	22	25	28	25	30	27	26	242
	d	54	65	63	62	66	68	71	59	57	47	612
	dd	1	0	0	0	0	0	1	1	1	0	4
	dy	0	0	0	0	0	0	0	0	0	0	0
	k	121	134	134	138	131	154	140	136	130	120	1338
	kk	7	3	7	4	4	6	9	5	1	3	49
	kky	1	2	0	0	0	0	1	0	1	1	6
	ky	12	10	12	11	9	14	10	10	7	5	100
	g	70	72	71	75	74	86	74	61	70	51	704
	gy	5	5	5	3	3	4	2	3	1	2	33
摩擦音	s	68	66	71	64	72	75	80	72	67	51	686
	ss	1	1	0	0	0	1	3	2	1	0	9
	z	26	27	27	24	19	21	23	21	16	11	215
	sh	46	57	63	63	56	67	72	56	50	30	560
	ssh	3	2	6	6	4	3	2	3	1	3	33
	j	44	41	43	45	38	37	35	26	24	17	350
	h	78	69	76	81	69	83	79	74	59	57	726
	hy	5	3	4	4	1	1	3	3	1	3	28
	f	16	17	19	18	15	19	15	12	11	10	152
	ff	0	0	0	0	0	0	0	1	0	0	1
	ch	27	26	27	34	27	32	23	19	20	16	251
	cch	1	2	0	1	1	1	1	3	1	1	12
	ts	26	27	27	27	26	26	28	24	21	20	252
	tts	0	0	0	0	0	0	1	0	0	0	1
(無音)	pau	94	89	93	105	87	103	105	88	82	58	909