博 士 論 文

音声認識における特徴量の非同期性と音素環境依存性のモデル化に関する研究

指導教官　下平 博 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻 知能情報処理学講座

松田 繁樹

2003 年 2 月 10 日

要 旨

本研究では，音響特徴ベクトル時系列における個々の音響特徴量の振舞いに着目した音声認識性能の改善に関する検討を行う．音声認識システムの音響モデルとして広く用いられている隠れマルコフモデル（HMM）は，音声の観測量がベクトルであることを仮定している．本研究は，このような常識を覆し,「個別特徴量の集合」として捉えることにより,従来になかった次に述べる 2 つの仮説について検討を行った.

第 1 の仮説として,「個々の特徴量の時間非同期性」に着目した．音響特徴ベクトル時系列の個々の音響特徴量の値は必ずしも同じタイミングで変化していない．従来型 HMM は,音響特徴ベクトルを構成しているすべての特徴量の値が HMM の状態遷移と同じタイミングで変化することを仮定したモデルである．従って，このような信号を従来型 HMM でモデル化した場合，大量の時間方向状態数が必要となる．しかし，大量の時間方向状態数は,モデルパラメータの多大な増加を招くため，モデルの統計的信頼性の低下に繋がる．そこで本検討では，個々の特徴量の値がお互いに異なるタイミングで変化するモデルの検討を行う．本検討の結果，特定話者音声認識において，本手法の有効性を確認した.

第 2 の仮説として,「個々の特徴量の音素環境依存性」に着目した．従来提案されたパラメータ共有法の幾つかは，音素環境依存性を考慮したクラスタリングを行なっている．この従来のパラメータ共有法は，全ての特徴量に対して共通のパラメータ共有構造を割り当てる手法である．しかし，音声の観測量である音響特徴ベクトルは，お互いに異なった振舞いを持つ音響特徴量の集合であり，お互いに異なった複雑性や音素環境依存性を持つと考えられる．そこで本検討では，個々の特徴量に依存したパラメータ共有構造を，音素環境依存性を考慮して決定することにる，音声認識性能の改善を検証した．本検討の結果，比較的少ないパラメータ数を持つモデルで，その有効性を確認した.

以上のように，従来の音声認識においては「特徴ベクトル」という概念のもとに特徴量が一括して扱われて来たのに対し，本研究では，個別特徴量の扱いを提唱し，時間特性と環境依存性の 2 つの側面から検証し，モデルや学習及び，認識アルゴリズムを提案し，実験を通して有効性を実証した.

Abstract

This thesis presents our proposals to improve speech recognition systems. The state-of-the-art hidden Markov model (HMM) based systems usually treat the acoustic features as a chain of stationary signal sources. The observed values of these features are represented by vectors. We assume that they might be better modeled by individual vector components. We discuss two methods based on this assumption.

In the first method, we try to model asynchronous changes of individual acoustic vector components. Conventional HMM implicitly assumes that individual components change their statistical properties simultaneously. This assumption may not be true. Temporally changing patterns of individual acoustic components do not necessarily synchronize with each other. We propose a new HMM that allows asynchronous state transitions between individual vector components. We demonstrate that this new HMM outperforms the conventional HMM in speaker-dependent speech recognition task.

In the second method, we try to model phoneme context dependency of individual acoustic vector components. Conventional parameter tying techniques provide a common tying structure for all vector components, no matter how different is their individual components complexity and phoneme context dependency. In this discussion, we propose a new parameter tying technique that allows to have distinct tying structures for each component. Our experimental results show that proposed HMM with feature-depended tying works better than conventional HMM with a common tying.

Both proposed methods are based on treating the observed feature vector as vector of individual components. Moreover, we discuss time characteristics and phoneme context dependencies of individual components, and develop the new HMM structures and the new training techniques. All these methods are evaluated in continuous phoneme recognition task.

目 次

図 目 次

第 1 章

序論

1.1 本研究の背景

近年，音声認識技術の発展に伴い，その実用化が進んでいる．キーボードに代わる情報入力手段としての音声認識や，会議の自動議事録システム，外国人との会話の自動翻訳，聴覚障害者のための会話支援など多くの分野への応用研究が行なわれている．しかし，その音声認識性能は人間には遠く及ばないのが現状であり，一層の性能向上のための研究努力が必要である．

現在，主流となっている音声認識システムの構成法は，統計的音声認識手法に基づくものである．入力された音声波形は，まず，音響分析によって音響特徴ベクトル時系列に変換される．次いで，その音響特徴ベクトル時系列に対して，音響的な確からしさと言語的な確からしさの和が最大となる発話内容を，発話候補仮説の集合の中から探索することで音声認識が成される．音響的な確からしさや言語的な確からしさは，言語モデルと音響モデルにより確率として計算される．高性能な音声認識を実現するためには，音響と言語のいずれのモデルも高精度かつ頑健でなければならない．

現在，最も広く用いられている音響モデルは，隠れマルコフモデル（Hidden Markov Model: HMM）[1, 2, 3, 4, 5, 6] である．次章で詳しく述べるが，HMM はマルコフモデルの拡張であり，個々の状態は観測ベクトルに対する確率分布を持ち，それらの状態を接続する状態遷移確率から構成されている．HMM は，音声などの非定常信号を，有限個の定常信号源の連鎖として表現するモデルである．

HMM は，高速な確率計算アルゴリズムや，分布や状態遷移確率などのモデルパラメータに対する学習アルゴリズムが存在し，容易に利用することができることから，現在の音

声認識システムにおける音響モデルの主流となっている．ある仮定した発声内容から，観測された音響特徴ベクトル時系列が生成される確率を，Viterbi アルゴリズムなどの動的計画法を基礎とした手法により高速に計算することができる．また，複数の発声候補仮説から最も尤もらしい発声内容を高速に探索するアルゴリズム [8, 9, 10, 11] が提案されている．その他，分布や状態遷移確率などのモデルパラメータを，学習データから自動的に推定するアルゴリズム [13] が提案されている．

HMM において，入力は「特徴ベクトル」という用語で表されるように，ベクトルであることをほぼ前提としている．本研究では，その前提を覆して新しい HMM の原理を提唱し，その効果を実験を通じて確認し，論じるものである．

1.1.1 HMM の高性能化に関する従来の研究

従来より行われてきた HMM の高性能化に関する研究について，本研究に多少とも関係ある手法及び，その対極にある手法に限定して概観する．

1. 状態出力確率分布の精密化

HMM の個々の状態の定常分布を複数混合分布化することにより，精密な分布形状を表現する手法が提案されている．現在最も広く用いられる手法は，Juang ら (1985) により提案された，複数の多次元ガウス分布の混合化である複数混合分布モデル [25] である．複数混合分布モデルを用いることにより，単一のガウス分布では表現できない詳細な分布形状の表現が可能となる．更に，高橋ら (1996) は，個々の特徴量をスカラー量子化に基づく離散分布の混合を用いる，離散混合分布型 HMM[26] を提案し，正規分布に捕らわれないモデル化で音声認識性能の向上が図れることを示した．

本研究との関連

これらの手法は，多次元ベクトル分布の混合であるため，混合要素分布各々の形状や混合数は全ての特徴量に対して共通である．従って，個々の音響特徴量がお互いに異なる複雑性を持つような特徴ベクトル時系列に対するモデル化の観点からは論じられていない．若干関連のある研究として，マルチストリーム型モデルがある．Bocchieri ら (1997) は，相関の強い特徴量を一つのストリームと考え，個々のストリーム毎に分布を共有化することで，少ない分布数で高速に尤度計算を行なう手法 [19] を

提案している．しかし，ここで述べた個々の特徴量分布の精密化の観点からは論じられていない．

2．状態継続時間の精密化

HMM は，音響特徴ベクトル時系列の時間的な振舞いを，ベクトル分布の連鎖として表現するモデルである．個々のベクトル分布は状態遷移確率により接続されており，この状態遷移確率は個々のベクトル分布の継続時間長（停留時間）を表現している．この状態遷移確率に対する特徴ベクトル時系列の確率は，その時系列の長さと共に指数関数的に減少するため，個々の状態に留まることをうまく表現しているとは言えない．そこで，状態遷移確率の代わりに継続時間分布を用いる手法などが提案されている．Russell ら (1985) は，自己ループのない状態を並べることにより，直接的に継続時間を制御する手法 [27] を提案している．また，Rabiner ら (1985) は，Viterbi アルゴリズムなどにより計算した個々の状態の継続時間に対して，後処理的に継続時間を考慮した確率を計算する手法を提案した．Levinson(1986) は，個々の状態の継続時間分布として，連続分布を用いる手法 [29] を提案している．

本研究との関連

これらの手法は，ベクトル分布を持つ状態の継続時間を精密にモデル化するための手法であり，全ての特徴量の値が同期して変化することを暗に仮定している．従って，個々の特徴量の変化のタイミングがお互いに異なるような音響特徴ベクトル時系列のモデル化の観点からは論じられていない．

3．環境依存音素モデル

個々の音素の特徴ベクトル時系列は，先行，後続の音素の影響を受けて変形する．この点に着目した音素環境依存の音素モデルが Schwartz ら (1985) [30] によって提案され，現在の高性能な音声認識にとって無くはならない技術となっている．音素環境依存モデルは，先行，当該，後続などの音素環境の組毎に別々のモデルを用意する手法であるため，モデル全体のパラメータ数は爆発的に増大する．モデルパラメータ数の増加は，モデルの統計的信頼性の低下に繋がるため，次に述べるパラメータ共有手法により，パラメータ数を削減することが不可欠である．

Lee ら (1988)[31] は，頑健な音素環境依存モデルを生成するため，同一音素の音素環境依存モデル同士を，距離の近いものから共有化し，Generalized Triphone と呼ぶクラスタを生成する手法を提案している．音素環境を考慮した音素環境依存モデルのトップダウンクラスタリング手法として，嵯峨山 (1988) は，音素環境クラスタリング [32] と呼ぶ手法により，先行や当該，後続の音素環境の違いによる音素パターンの変形をクラスタリングする木構造を自動生成し，それにより共有構造を決定するアルゴリズムを提案した．その他，速水ら (1990) は，音素決定木を基礎としたクラスタリング法 [34] を提案している．

HMM の状態共有構造の決定においては，鷹見ら (1993)[33] や Ostendorf ら (1997)[35] は，音素環境クラスタリングを基礎とした手法により，音素環境依存性を考慮した状態共有構造を自動生成するアルゴリズムを提案している．また，音素決定木を基礎とした手法が Young ら (1994)[37] や堀ら (1997)[36] により提案されている．

本研究との関連

一般に，音素環境依存モデルや状態に対するパラメータ共有構造の決定には，音素環境依存性を考慮したクラスタリングが行なわれている．しかし，これらのクラスタリング法は，全ての特徴量のパラメータ共有構造が同一であることを仮定しており，個々の音響特徴量がお互いに異なる共有構造を持つモデルではない．高橋ら (1999) の 4 階層共有構造 [45] で提案されている特徴量分布の共有は，単純に距離の近い特徴量分布同士を単純に共有化する手法であり，個々の特徴量分布の音素環境依存性を考慮したクラスタリング法ではない．

4. マルチバンド音声認識と Audio-Visual 音声認識

本研究の着眼点に比較的近い研究として，マルチバンド音声認識手法と Audio-Visual 音声認識手法がある．これらの手法は，雑音などによる音声認識性能の低下を，複数の個別のストリーム中から利用可能なストリームを用いることで，頑健な音声認識を実現しようとする方法である．

音声認識において一般的に用いられる音響特徴量であるケプストラム係数は，スペクトラムを対数化し逆フーリエ変換することによって得られる．従って，ある狭帯域の雑音の重畳した音声は，ケプストラム係数全体に影響を与えてしまう．そこで，スペクトラムを部分周波数帯域に分割し，各々のバンド毎に音響分析と確率計算を行うことにより，狭帯域雑音の影響を軽減する，マルチバンド音声認識手法が提案されている．その他，音声波形

の観測が非常に困難な環境においても，音声を認識するための手法として，Audio-Visual 音声認識手法が提案されている．この手法は，音声以外の情報として，唇動画像などの雑音の影響を受けにくい情報を用いる手法である．

このようなマルチバンド音声認識や Audio-Visual 音声認識の分野では，個々のストリーム間の非同期性をモデルに組み込んむ研究が行なわれている．複数のサブバンド間の非同期な振舞いを表現するための手法として，Cerisara ら (2000) が提案した，個々のバンドを別々の HMM によりモデル化する手法 [46] や，個々のストリームの非同期な状態遷移に関連を持たせる手法として Mirghafori ら (1999) の手法 [47]，Logan ら (1998) の Factorial HMM[50]，Nock ら (2000) の Loosely Coupled HMM[48] が提案されている．また，唇動画像の情報を用いて音声認識を行う Audio-Visual 音声認識 [51, 52, 53] においても，唇と音声のお互いに異なるストリーム間の非同期性をモデル化する手法が提案されている．更に，非同期遷移に制約を付加する手法 [47, 52, 53] が提案されている．

本研究との関連

これらの手法は，個々のストリームを独立と考え，不自然な入力の影響を他のストリームへ及ぼさないためと言うことが主な目的である．従って，本研究の着眼点である個々の音響特徴量の振舞いのモデル化ではない．

5. セグメントモデル

近年，HMM では表現することのできない，個々の音響特徴ベクトル間の相関をモデル化するための手法として，セグメントモデル [20, 21, 22, 23, 24] が提案されている．従来の HMM は，個々の状態が受け持つ特徴ベクトル時系列の値の変化を，定常分布の連鎖として扱うモデルであり，状態が受け持つ区間内の個々の音響特徴ベクトル間（フレーム間）の相関は考慮されていない．それに対して，セグメントモデルは，状態へ割り当てられた特徴ベクトル間の相関を考慮した確率を計算する手法である．セグメントモデルの例として，時間に依存した平均ベクトルを持つ分布により確率を計算する手法などが提案されている．しかし，セグメントモデルは，音素セグメンテーション能力や計算量の問題などがあり，音響モデルの主流とはなっていない．

本研究との関連

これは，本論文で述べるような特徴量を個別に扱う方向とは正反対の方向であり，個々の音響特徴量ベクトルの部分区間をマトリクスのパターンとして扱う手法である．

1.2 本研究の目的

背景で述べたように，音声認識性能を改善するための方法として，様々な精密化と頑健化による改善手法が提案されてきた．これらの改善手法の大半は，音声の観測量がベクトルであることを仮定している．しかし，音声認識に用いられる音響特徴ベクトルは，複数の音響特徴量から構成（第 1MFCC，第 2MFCC，また時間微分成分など）されており，個々の音響特徴量の振舞いは，後章で述べるように互いに異なる．個々の特徴量の特性を考慮したモデル化や，各特徴量の間の相関を考慮した統合により，更に効果的に学習データの振舞いをモデル化することができると考えられる．

音声の観測量としての「特徴ベクトル」を HMM によりモデル化していた音声認識の常識を覆えし,「個別特徴量の集合」として捉えることにより，従来になかったさまざまな発想が可能になり，そのための新しい定式化と解決アルゴリズムが必要となる．本研究では，次に述べる特に大きな効果が得られると考えられる 2 つの仮説に対し，定式化を与え，アルゴリズムを導き，実験を通して効果を調べ，仮説の検証を行う．

- **個々の特徴量の非同期性**

  音響特徴ベクトル時系列の個々の音響特徴量の値は必ずしも同じタイミングで変化していない．従来型 HMM は，音響特徴ベクトルを構成しているすべての特徴量の値が HMM の状態遷移と同じタイミングで変化することを仮定したモデルである．従って，このような信号を従来型 HMM でモデル化した場合，大量の時間方向状態数が必要となる．しかし，大量の時間方向状態数は，モデルパラメータの多大な増加を招くため，モデルの統計的信頼性の低下に繋がる．個々の特徴量の値がお互いに異なるタイミングで変化するモデルを用いることで，より効果的に音響特徴ベクトル時系列をモデル化できると考えられる．

- **個々の特徴量の音素環境依存性**

  従来提案されたパラメータ共有法の幾つかは，音素環境依存性を考慮したクラスタリングを行なっている．この従来のパラメータ共有法は，全ての特徴量に対して共通のパラメータ共有構造を割り当てる手法である．しかし，音声の観測量である音響特徴ベクトルは，お互いに異なった振舞いを持つ音響特徴量の集合であり，お互いに異なった複雑性や音素環境依存性を持つと考えられる．個々の特徴量に依存したパラメータ共有構造を音素環境依存性を考慮して決定することにより，音声認識性能の改

善に繋がると考えられる．

以下に，これらの 2 つの側面について説明する．

1.2.1 個別特徴量の非同期性のモデル化

HMM は出力確率分布を持つ状態とそれらを繋ぐ状態遷移確率から構成された確率的モデルであり，音声波形から抽出された音響特徴ベクトル時系列などの非定常な信号は，定常分布を持つ状態の連鎖として表現される．そのため，音響特徴ベクトルを構成しているすべての特徴量の値が HMM の状態遷移と同じタイミングで変化することを仮定したモデルと考えることができる．

しかし，実際に観測される個々の音響特徴量の値は必ずしも同じタイミングで変化していない．例えばケプストラムなどの音響特徴量とその時間微分特徴量値は原理的に同じタイミングで変化しない．個々の音響特徴量の時間変化タイミングがお互いに異なる信号を従来型 HMM でモデル化しようとした場合，大量の時間方向状態数が必要となる．しかし，大量の時間方向状態数は，モデルパラメータの多大な増加を引き起こすため，モデルの統計的信頼性の低下に繋がる．

上述の問題は，個々の特徴量の値が同じタイミングで変化することを仮定しているために発生すると考えられる．そこで本議論では，この仮定を見直し，個々の特徴量の値がお互いに異なるタイミングで変化するモデルの有効性を検証する．このようなモデルを用いることにより，同じパラメータ数で，より効果的に音響特徴ベクトル時系列を表現することが可能になると考えられる．以後，本議論では，個々の特徴量の値がお互いに異なるタイミングで変化することを，非同期性と呼ぶこととする．逆に，お互いに同じタイミングで変化することを同期性と呼ぶこととする．

本議論では，個々の特徴量やストリームの値の変化がお互いに非同期な特徴ベクトル時系列をモデル化するための枠組として，非同期遷移型 HMM （Asynchronous Transition HMM: AT-HMM）を提案する．マルチバンド音声認識や Audio-Visual 音声認識の分野でも，個々のストリーム間の非同期性をモデル化する手法を提案しているが，時間非同期遷移構造に対する議論や，個々の音響特徴量間の非同期性に関する議論が十分に行われているとは言えない．本議論では，個々の音響特徴量間やストリーム間の非同期な状態遷移構造（時間非同期遷移構造）に関して分類を行い,「直積 HMM」と呼ぶ一般的な HMM のクラスの観点から，種々の時間非同期遷移構造を統一的に表現できることを示す．

更に本議論では，これまでの非同期性に関する研究において着目されていなかった，個々の音響特徴量の状態遷移に対する順序制約の概念を提案する．個々の音響特徴量の状態遷移に対して順序制約を導入した AT-HMM は,「時間方向共有法（Temporal Tying Technique）」と呼ぶ新しいパラメータ共有法により，現在音声認識の音響モデルとして一般に広く用いられている left-to-right 型 HMM と同様の構造で実現することができる．一般に，音響特徴ベクトルは 20〜30 個の音響特徴量から構成されている．このような大量のストリーム間の非同期性を直積 HMM で表現することは不可能であった．しかし，この順序制約の概念を導入することにより，個々の音響特徴量間の非同期性をモデル化することが初めて可能となる．

本議論で提案した AT-HMM に対して音声認識性能の評価実験を行い，個々の特徴量の非同期性を積極的に用いたモデル化の有効性を検証する．

1.2.2 個別特徴量の音素環境依存性のモデル化

従来提案されたパラメータ共有法の幾つかは，音素環境依存性を考慮したクラスタリングを行なっている．音響特徴ベクトル時系列は，例え同一音素であったとしても，前後の音素環境によって，その振舞いは大きく異なる．音素環境依存性を考慮したパラメータ共有構造の決定は，モデルの統計的信頼性を改善するための有効な手段の一つである．

前節で紹介したパラメータ共有法は，全ての特徴量のパラメータ共有構造が同一であることを仮定している．しかし，音声の観測量である音響特徴ベクトルは，お互いに異なった振舞いを持つ音響特徴量の集合である．これらの音響特徴量は，お互いに異なった複雑性や音素環境依存性を持つと考えられる．高橋ら (1999) の 4 階層共有構造 [45] で提案されている特徴量分布の共有は，単純に距離の近い特徴量分布同士を単純に共有化する手法であり，個々の特徴量分布の音素環境依存性を考慮したクラスタリング法ではない．

本議論では，個々の音響特徴量に依存したパラメータ共有構造を，音素環境依存性を考慮して決定することにる，音声認識性能の改善を検証する．個々の音響特徴量がお互いに異なるパラメータ共有構造を持つモデルを用いることにより，より効果的に音響特徴ベクトル時系列をモデル化することができると考えられる．

本議論では，個々の音響特徴量のパラメータ共有構造を，別々に音素環境クラスタリングを行なうことにより自動的に決定する手法として，特徴量依存音素環境クラスタリング（Feature-Dependent Phoneme Environment Clustering: FD-PEC）を提案する．この手法

は，従来の分布間距離の近い特徴量分布を共有するような単純な方法ではなく，音素環境依存性を考慮した特徴量分布レベルのパラメータ共有構造を自動的に生成するための手法である．また，尤度を基準としてクラスタを分割することにより，音声情報を多く含む特徴量（MFCC の低次など）に対しては，より多くのパラメータを割り当て，あまり含んでいない特徴量（MFCC の高次など）に対して，少ないパラメータ数が自動的に割り当てられると考えられる．この FD-PEC により，少ないパラメータ数で頑健かつ精密に音声信号をモデル化することができると考えられる．

更に，FD-PEC を実現するための手法として，逐次状態分割法（Successive State splitting: SSS）を基礎とした特徴量依存逐次状態分割法（Feature-Dependent SSS: FD-SSS）を提案する．FD-SSS 法を用いることにより，個々の特徴量に依存したスカラー分布共有構造が自動的に生成される．

FD-PEC 法により生成されたパラメータ共有構造を持つモデルに対して，音声認識性能の評価実験を行い，個々の特徴量の音素環境依存性を考慮したパラメータ共有構造の有効性の検証を行う．

1.2.3　その他の観点

以上に述べた議論の他に,「個々の特徴量の出力確率分布はお互いに異なるのではないか」という仮説も立てられる．個々の特徴量の分布形状は，お互いに異なっている．例えば，MFCC の低次と高次，またパワー項ではお互いに分布形状の複雑さは異なっている．このような異なる分布形状を効果的に表現するため，個々の特徴量毎にガウス分布の混合数を最適化する方法や，ポアソン分布やガンマ分布を積極的に用いる方法が考えられる．この事項については本論文では論じないが，このように「特徴ベクトル」を「個別特徴量の集合」と見直すことにより，さまざまな視点から音響モデル改善の端緒が得られる．

1.3　本論文の構成

本論文は 5 つの章から構成されている．

第 2 章では，統計的音声認識システムの概要を述べる．その後，音響モデルとして広く用いられている HMM について詳しく述べ，HMM を用いた高速な確率計算法，モデルパラメータの推定法を述べる．

第 3 章では,「個別特徴量の時間非同期性のモデル化」について論じる．第 3.1 節では音声認識の音響特徴量として用いられる個々の特徴量間の時間非同期性の検討及び，非同期性を考慮することによるモデル化効率改善の可能性について議論する．3.2 節では非同期遷移型 HMM(AT-HMM) を提案し種々の時間非同期遷移構造について述べる．第 3.3 節では，順序制約付き AT-HMM の生成法を述べる．第 3.4 節では，順序制約付き AT-HMM の時間方向状態数の増加による音声認識性能の評価を行なう．第 3.5 節では，順序制約の有無による AT-HMM の音声認識性能の評価を行なう．第 3.6 節及び第 3.7 節では，日本語音素接続制約の付いた特定話者と不特定話者環境における連続音素認識実験により，順序制約付き AT-HMM の音声認識性能の評価を行う．第 3.8 節は，本議論のまとめである．

第 4 章では,「個別特徴量の音素環境依存性のモデル化」について論じる．第 4.1 では，個々の音響特徴量の音素環境依存性について議論する．第 4.2 節では，従来法である音素環境クラスタリングの概念について述べ，その後，提案手法である特徴量依存音素環境クラスタリング (FD-PEC) の概念及び，特徴量依存音素環境クラスタの構造について述べる．第 4.3 節では，FD-PEC を基礎とした，特徴量依存逐次状態分割法 (FD-SSS) を提案する．第 4.4 節は，FD-SSS により得られた特徴量依存隠れマルコフネットワークから生成した AT-HMM の認識性能の評価を行なう．第 4.5 節は，本議論のまとめである．

第 5 章は，本研究のまとめである．

第 2章

統計的音声認識手法

近年，統計的手法を用いた音声認識システムの研究が盛んに行なわれている．本章では，この音声認識システムの全体の構造を述べた後，その構成要素である音響分析，音響モデル，言語モデル，デコーダについて記述する．

その後，統計的手法を用いた音声認識システムの音響モデルとして広く用いられている，隠れマルコフモデル（Hidden Markov Model: HMM）[1, 2, 3, 4, 5, 6] について述べる．HMM の数学的定義及び，音響特徴ベクトル時系列に対する確率計算法，モデルパラメータの学習法について述べる．また，HMM を用いた音声認識システムのための，代表的な性能改善の手法を述べる．

2.1 統計的音声認識システムの構造

図 2.1 は，現在広く一般的に用いられている統計的手法を用いた音声認識システムの構造である．図の様に，入力された音声波形は，音響分析により音響特徴ベクトル時系列（LPC 係数 [14, 15] や LPC ケプストラムなど）が抽出される．その後，音響モデルと言語モデルの情報を用いて，デコーダにより尤もらしい単語列や音素列が検索される．

統計的音声認識とは，式 (2.1) の様に，特徴ベクトル時系列 $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T)$ が観測された時に，その発話された内容が単語列 $\mathbf{W} = (w_1, w_2, \cdots, w_N)$ である確率 $P(\mathbf{W}|\mathbf{O})$ が最大となる単語列 $\hat{\mathbf{W}}$ を，探索することにより音声を認識する手法である．

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{O}) \tag{2.1}$$

式 (2.1) は，ベイズの定理により，式 (2.2) の様に書換えることができる．統計的音声認

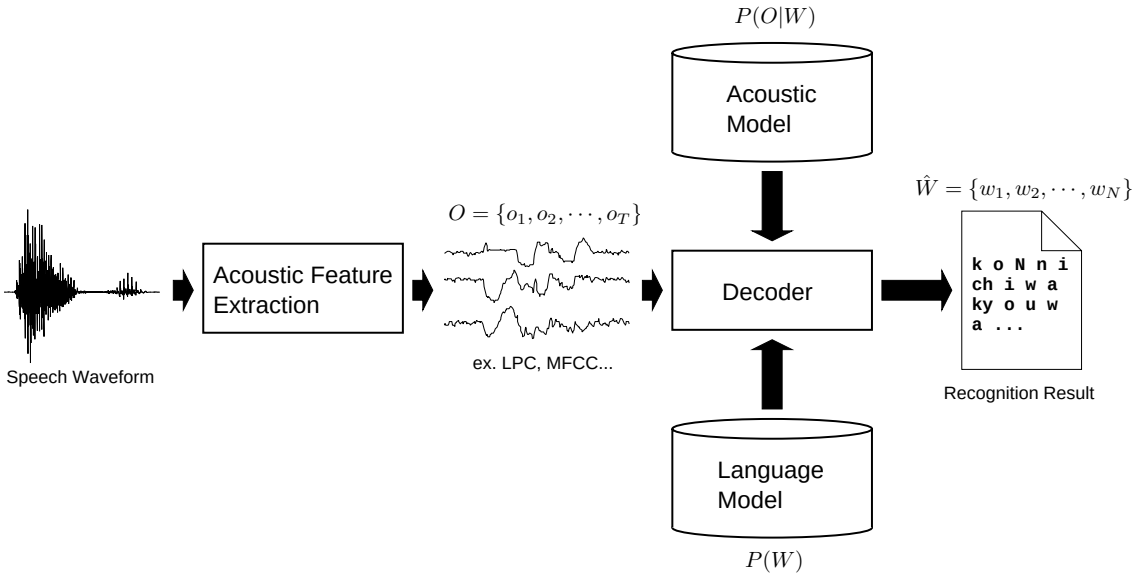


図 2.1: 統計的手法を用いた音声認識システムの構造

識とは言語モデル $P(\mathbf{W})$ により制限された単語列の空間中で，観測された特徴ベクトル時系列に対して最も高い確率の得られる単語列を音響モデル $P(\mathbf{O}|\mathbf{W})$ を用いて検索する手法と考えることができる．

$$\begin{aligned} \hat{\mathbf{W}} &= \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{O}) \\ &= \underset{\mathbf{W}}{\operatorname{argmax}} \frac{P(\mathbf{O}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{O})} \\ &= \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W}) \end{aligned} \tag{2.2}$$

2.1.1 音響分析

収録された音声波形は，例え発話内容が同一であったとしても，個々の話者のピッチや雑音環境，また，マイクの特性などの影響により，お互いに大きく異なっている．音声認識では，話者性や発話環境などにより影響を受けた音声波形から，発話の言語情報を表す音響特徴量を抽出する処理が必要である．

過去においては，フィルタバンクにより音声波形から抽出されたスペクトラムがよく用いられていた．現在では，線形予測係数（Linear Prediction Coefficient: LPC）[14, 15] や，

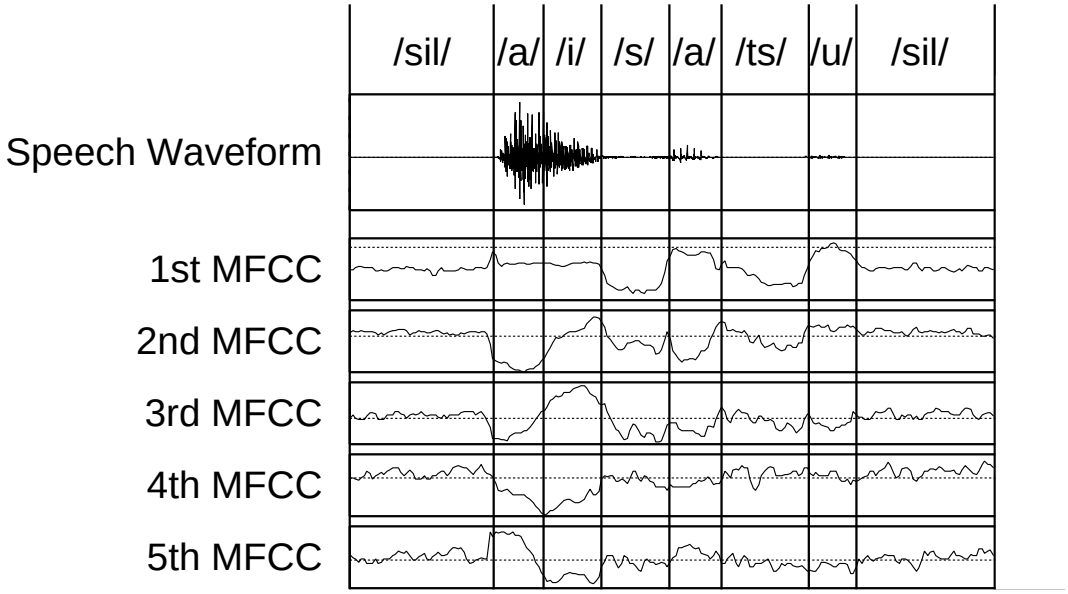


図 2.2: 単語「あいさつ」の音声波形を音響分析することにより得られた，第 1MFCC から第 5MFCC の時間変化

ケプストラム係数 [16, 17] などが広く用いられている．音響特徴ベクトル時系列の例として，図 2.2 に単語「あいさつ」の音声波形から抽出した，第 1MFCC（Mel Frequency Cepstrum Coefficient）から第 5MFCC までの時系列を示す．このケプストラム係数は，短時間スペクトラムの対数の逆フーリエ変換として定義されている．そのため，複数のチャネルの畳み込みを，加算演算として表現することができるため，口や喉の形の変化を抽出するのに適していると考えられる．

2.1.2 音響モデル

音響モデルには，音声波形を音響分析することによって得られる特徴ベクトル時系列の振る舞いがモデル化されている．音響モデルを用いることにより，観測系列 $\mathbf{O}$ に対して，ある単語列や音素列を仮定した場合の確率 $P(\mathbf{O}|\mathbf{W})$ を得ることができる．本論文では，この様な確率計算を行う音響モデルとして，隠れマルコフモデルを用いる．隠れマルコフモデルは，音声のような非定常な信号を定常信号源のマルコフ過程として音声信号を表現する確率モデルである．

認識単位としては，長い方から順に，文 (sentence)，文節 (phrase)，単語 (word)，音節

(syllable)，音素 (phoneme) といった単位が挙げられる．一般に認識単位の長さが長くなるに従って，認識に必要とされるモデル数は増加する．そのため，文などを認識単位とすることは現実的では無く，音節や音素が認識単位として用いられる．本研究では，認識単位として音素を用いた音響モデルの議論を行う．また，先行音素や後続音素などの影響を考慮した，音素環境依存の音響モデルを用いる．

2.1.3 言語モデル

言語モデルには，発話文章の単語接続関係などの文法情報がモデル化されている．言語モデルを用いることにより，ある単語列 $\mathbf{W}$ に対する確率 $P(\mathbf{W})$ を計算することができる．小語彙の音声認識システムでは，単語ラティスなどの単語ネットワークがよく用いられる．大語彙の連続単語音声認識システムなどでは，一般に，N-gram 言語モデル [18] が用いられる．

2.1.4 デコーダ

デコーダは，音響モデルや言語モデルからの情報をまとめ，式 (2.2) の argmax の計算を行い，最終的な認識結果を出力する処理が行なわれる．この様な処理を行う手法として，2 段 DP 法 [8]，Level Building 法 [9]，One Pass DP 法 [10, 11] など，多数のデコーディングアルゴリズムが提案されている．

2.2 隠れマルコフモデル

2.2.1 HMM の構造

統計的手法を用いた音声認識システムの音響モデルとして，HMM が広く用いられている．HMM は出力確率分布を持つ状態と，それらを繋ぐ状態遷移確率から構成された確率モデルである．マルコフモデルでは，個々の状態から出力されるシンボルは一つであった．しかし HMM の状態は，出力シンボルに対する確率分布を持っており，出力されるシンボルは 1 種類とは限らない．隠れマルコフモデルの「隠れ」とは，状態の中に確率分布が隠れていることから付けられた名称である．

図 2.3 に，観測ベクトルを出力しない初期状態と最終状態を持つ，5 状態の HMM の構造

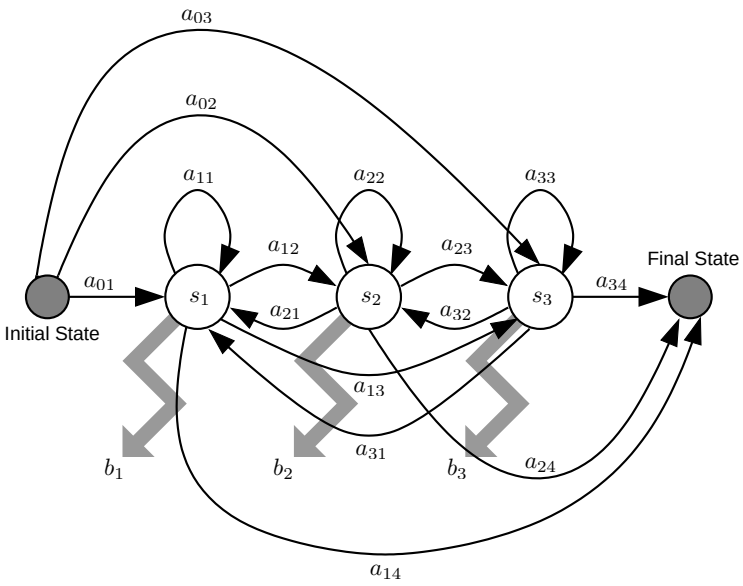


図 2.3: 隠れマルコフモデルの構造

を示す．図中の s_i は第 i 状態，$a_{i,j}$ は第 i 状態から第 j 状態への状態遷移確率，b_i は第 i 状態の出力確率分布であり，出力シンボル y に対する確率 $P(y|b_i) = b_i(y)$ が定義されている．第 1 状態 s_1 から第 3 状態 s_3 はシンボルを出力するが，初期状態 s_0 と終了状態 s_{I+1} はヌル状態であり，シンボルは出力されない．この様に，HMM は，状態の集合 S と状態出力確率分布の集合 B，及び状態遷移確率行列 A から構成されている．本論文では，これらの構成要素全ての集合として，HMM のモデルパラメータを λ と記述する．状態数 I の HMM における，個々の構成要素の定義を次に示す．

$$
\begin{aligned}
\text{モデルパラメータの集合} \quad & \lambda = \{S, B, A\} \\
\text{状態の集合} \quad & S = \{s_0, s_1, s_2, \cdots, s_I\} \\
\text{状態出力確率分布の集合} \quad & B = \{b_1, b_2, \cdots, b_{I-1}\} \\
\text{状態遷移確率行列} \quad & A = \begin{pmatrix}
0 & a_{0,1} & a_{0,2} & \dots & a_{0,I} & 0 \\
0 & a_{1,1} & a_{1,2} & \dots & a_{1,I} & a_{1,I+1} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & a_{I,1} & a_{I,2} & \dots & a_{I,I} & a_{I,I+1} \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\end{aligned}
$$

音声認識では一般に,「Left-to-right 型 HMM」と呼ぶ構造が用いられる．Left-to-right 型

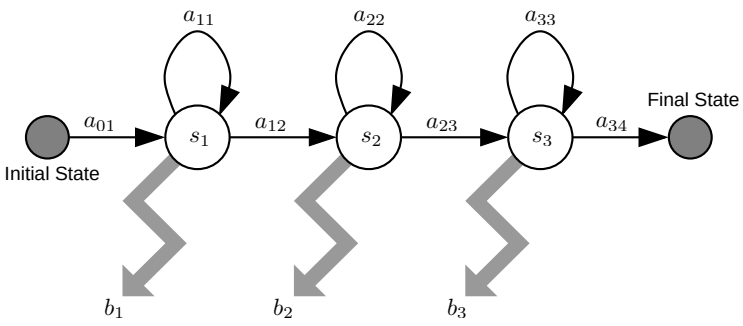


図 2.4: Left-to-right 型隠れマルコフモデルの構造

HMM の構造を図 2.4 に示す．Left-to-right 型 HMM は，第 1 状態から第 2 状態，第 2 状態から第 3 状態へ遷移し，逆方向へは状態遷移しない構造を持っており，音声などの時間的な相関を持つ信号のモデル化に適していると考えられる．

2.2.2 状態出力確率分布

個々の状態の出力確率分布として，音響特徴ベクトルをベクトル量子化することにより得られる離散シンボルに対する離散確率分布や，ガウス分布などの連続確率密度分布を用いる手法が提案されている．しかし，ベクトル量子化は，常に量子化誤差の問題があるため，近年では，連続確率密度分布を用いる手法が主流である．

2.2.3 確率計算法

与えられた単語列（音素列）に対する，観測された特徴ベクトル時系列の確率を HMM を用いて計算するためには，$P(\mathbf{O}|\lambda)$ を計算する必要がある．HMM の個々の状態は，特徴ベクトルを出力するため，入力された特徴ベクトル時系列 $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T)$ を出力可能な状態系列 $\mathbf{q}$ は複数存在する．式 (2.3) は，特徴ベクトル時系列 $\mathbf{O}$ に対して，モデルパラメータ λ と，ある 1 つの状態系列 $\mathbf{q} = (q_1, q_2, \cdots, q_T)$ が与えられた時の確率 $P(\mathbf{O}|\mathbf{q}, \lambda)$ である．また，HMM から特徴ベクトル時系列が出力される確率 $P(\mathbf{O}|\lambda)$ は，式 (2.4) の様に，可能な状態系列各々の場合の確率の和として表される．

$$P(\mathbf{O}|\mathbf{q}, \lambda) = a_{0,q_1}[\prod_{t=1}^{T-1} b_{q_t}(\mathbf{o}_t)a_{q_t,q_{t+1}}]b_{q_T}(\mathbf{o}_T)a_{q_T,I+1} \tag{2.3}$$

$$P(\mathbf{O}|\lambda) = \sum_{all\ \mathbf{q}} P(\mathbf{O}|\mathbf{q}, \lambda) \tag{2.4}$$

可能な状態系列の集合の要素数は，$(I-2)^T$ 個となり膨大である．この様な確率計算を効率的に行なう手法として，動的計画法を基礎としたアルゴリズムが提案されている．Forward アルゴリズムまたは Backward アルゴリズムを用いることにより観測系列長 T に関する多項式オーダーで確率を計算することができる．また，最大の確率の得られる状態系列 $\mathbf{q}$ の時の確率を高速に計算するアルゴリズムとして Viterbi アルゴリズムが提案されている．次に，各々のアルゴリズムについて述べる．

Forward アルゴリズム

Forward アルゴリズムを用いることにより，区間 $[1, t]$ の観測系列を HMM が出力する確率 $\alpha_t(i)$ を帰納的に計算することで，特徴ベクトル時系列全体の確率 $P(\mathbf{O}|\lambda)$ が計算される．式 (2.5) に $\alpha_t(i)$ の定義を示す．

$$\alpha_i(t) = P(\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t, q_t = i|\lambda) \tag{2.5}$$

$\alpha_t(i)$ は次の様に帰納的に計算することができる．

1 初期化

$$\alpha_1(i) = a_{0,i} b_i(\mathbf{o}_1) \qquad (1 \leq i \leq I) \tag{2.6}$$

2 繰り返し処理

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{I} \alpha_t(i) a_{i,j}\right] b_j(\mathbf{o}_{t+1}) \qquad (1 \leq j \leq I)(1 \leq t \leq T-1) \tag{2.7}$$

3 最終処理

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{I} \alpha_T(i) a_{i,I+1} \tag{2.8}$$

Backward アルゴリズム

Backward アルゴリズムを用いることにより，区間 $[t+1, T]$ の観測系列を HMM が出力する確率 $\beta_t(i)$ を帰納的に計算することで，特徴ベクトル時系列全体の確率 $P(\mathbf{O}|\lambda)$ が計算される．式 (2.9) に $\beta_t(i)$ の定義を示す．

$$\beta_t(i) = P(\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, \cdots, \mathbf{o}_T | q_t = i, \lambda) \tag{2.9}$$

$\beta_t(i)$ は次の様に帰納的に計算することができる．

1 初期化

$$\beta_T(i) = a_{i,I+1} \qquad (1 \leq i \leq I) \tag{2.10}$$

2 繰り返し処理

$$\beta_t(i) = \sum_{j=1}^{I} a_{i,j} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j) \qquad (1 \leq i \leq I)(1 \leq t \leq T-1) \tag{2.11}$$

3 最終処理

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{I} a_{0,i} b_i(\mathbf{o}_1) \beta_1(i) \tag{2.12}$$

Viterbi アルゴリズム

Viterbi アルゴリズムは，Forward や Backward アルゴリズムとは異なり，確率 $P(\mathbf{O}|\mathbf{q}, \lambda)$ の最大値と，その最大値を与える状態系列 (最適状態系列)$\mathbf{q}^*$ を求める．すなわち，次式で与えられる最適化問題を解くための効率的アルゴリズムである．

$$P(\mathbf{O}|\mathbf{q}^*, \lambda) = \max_{all\ \mathbf{q}} P(\mathbf{O}|\mathbf{q}, \lambda) \tag{2.13}$$

時刻 t において状態 i から観測シンボルが出力される時，区間 $[0, t]$ の観測時系列を最も高い確率で出力することのできる，最適状態系列の確率 $\delta_t(i)$ を，帰納的に計算することにより，観測系列全体の最適状態系列の時の確率 $P(\mathbf{O}|\mathbf{q}^*, \lambda)$ を計算するアルゴリズムである．式 (2.14) に，$\delta_t(i)$ の定義を示す．

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} P(q_1, q_2, \cdots, q_t = i, o_1, o_2, \cdots, o_t|\lambda) \tag{2.14}$$

1 初期化

$$\begin{aligned} \delta_1(i) &= a_{0,i} b_i(o_1) \qquad (1 \leq i \leq I) && (2.15)\\ \psi_1(i) &= 0 && (2.16)\\ & && (2.17) \end{aligned}$$

2 繰り返し処理

$$\begin{aligned} \delta_t(j) &= \left[\max_{1\leq i\leq I} \delta_{t-1}(i) a_{i,j}\right] b_j(o_t) \qquad (1 \leq j \leq I)(1 \leq t \leq T-1) && (2.18)\\ \psi_t(j) &= \operatorname*{argmax}_{1\leq i\leq I} \delta_{t-1}(i) a_{i,j} && (2.19) \end{aligned}$$

3 最終処理

$$\begin{aligned} P(\mathbf{O}|\mathbf{q}^*, \lambda) &= \sum_{i=1}^{I} \delta_T(i) a_{i,I+1} && (2.20)\\ &= \operatorname*{argmax}_{1\leq i\leq I} \delta_T(i) a_{i,I+1} && (2.21) \end{aligned}$$

4 トレースバック処理

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \tag{2.22}$$

2.2.4 パラメータ推定法

学習データに対して，最大の尤度が得られる HMM のモデルパラメータ λ を解析的に決定するアルゴリズムは知られていない．本節では，モデルパラメータを繰り返し推定し，除々に尤度を増加させることにより，準最適なモデルパラメータ $\hat{\lambda}$ を推定するアルゴリズムとして，Baum-Welch トレーニング法と Viterbi トレーニング法について述べる．Baum-Welch トレーニング法は，HMM の確率計算における全ての状態系列の確率の総和が最大となるパラメータを推定する手法であるのに対して，Viterbi トレーニング法は，最適状態系列に対する確率が最大となるモデルパラメータを推定する手法である．

Baum-Welch トレーニング法

Baum-Welch トレーニング法 (以下，BW 法)[13] は，EM 法 (Expectation-Maximization method)[12] を HMM に適用したパラメータ推定法である．BW 法は，学習データ $\Theta = \{\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_N\}$ が与えられた時，現在のモデルパラメータ λ に対して，式 (2.23) に示す尤度関数の値が最大となるような $\hat{\lambda}$ を反復更新することにより，学習データに対するモデルの尤度を除々に増加させる手法である．

$$L(\lambda) = \prod_{\mathbf{O} \in \Theta} P(\mathbf{O}|\lambda) \tag{2.23}$$

あるモデルパラメータ λ の時に，BW 法により再推定されたモデルパラメータ $\hat{\lambda}$ の推定式を次に述べる．

準備として，与えられた特徴ベクトル時系列 $\mathbf{O}$ に対して，時刻 t で状態 i，時刻 $t+1$ で状態 j に遷移する確率 $\xi_t(i,j)$ を，式 (2.24) に示す．

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|\mathbf{O}, \lambda) = \frac{\alpha_t(i) a_{i,j} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}{P(\mathbf{O}|\lambda)} \tag{2.24}$$

式 (2.24) より，時刻 t に状態 i から特徴ベクトルが出力される確率 $\gamma_t(i)$ は，式 (2.25) により計算される．

$$\gamma_t(i) = P(q_t = i|\mathbf{O}, \lambda) = \sum_{j=1}^{I} \xi_t(i,j) \tag{2.25}$$

状態遷移確率の推定式

式 (2.24) と式 (2.25) から，再推定された状態遷移確率は式 (2.26) により計算される．

$$\hat{a_{i,j}} = \frac{\sum_{\mathbf{O} \in \Theta} \sum_{t=1}^{T} \xi_t(i,j)}{\sum_{\mathbf{O} \in \Theta} \sum_{t=1}^{T} \gamma_t(i)} \tag{2.26}$$

状態出力確率分布の推定式

状態出力確率分布を，ガウス分布とした場合の推定式を次に示す．式 (2.27) はガウス分布であり，μ_i は，状態 i の平均ベクトル，$\mathbf{\Sigma}_i$ は，状態 i の分散ベクトルである．

$$b_i(\mathbf{o}) = \mathcal{N}(\mu_i, \mathbf{\Sigma}_i, \mathbf{o}) = \frac{1}{\sqrt{(2\pi)^2 |\mathbf{\Sigma}_i|}} e^{-\frac{1}{2}(\mathbf{o}-\mu_i)^t \mathbf{\Sigma}_i^{-1} (\mathbf{o}-\mu_i)} \tag{2.27}$$

ガウス分布の平均及び，分散ベクトルの再推定値は，式 (2.28) と式 (2.29) により計算される．

$$\hat{\mu_i} = \frac{\sum_{\mathbf{o}\in\Theta}\sum_{t=1}^{T}\gamma_i(t)\mathbf{o}_t}{\sum_{\mathbf{o}\in\Theta}\sum_{t=1}^{T}\gamma_i(t)} \tag{2.28}$$

$$\hat{\mathbf{\Sigma}}_i = \frac{\sum_{\mathbf{o}\in\Theta}\sum_{t=1}^{T}\gamma_i(t)(\hat{\mu_i}-\mathbf{o}_t)^2}{\sum_{\mathbf{o}\in\Theta}\sum_{t=1}^{T}\gamma_i(t)} \tag{2.29}$$

Viterbi トレーニング

Viterbi トレーニング法は，Viterbi アルゴリズムを用いたパラメータ推定法である．前述の BW 法は，Forward アルゴリズムと Backward アルゴリズムにより計算される $\gamma_i(t)$ を用いてパラメータを再推定していたのに対して，Viterbi トレーニングは，Viterbi アルゴリズムにより得られる最適状態系列 $\mathbf{q}^*$ から計算される．

Viterbi トレーニング法では，BW 法における各パラメータ推定式中の $\xi_t(i,j)$ と $\gamma_t(i)$ は，各々式 (2.30) と式 (2.31) のように書き換えられる．

$$\xi_t(i,j) = \begin{cases} 1, & q_t = i \text{ and } q_{t+1} = j \\ 0, & \text{otherwise} \end{cases} \tag{2.30}$$

$$\gamma_t(i) = \begin{cases} 1, & q_t = i \\ 0, & \text{otherwise} \end{cases} \tag{2.31}$$

近年，学習データは大量化しており，BW 法よりも高速にモデルパラメータを推定することができる Viterbi 法は，大変有効な手法である．

2.2.5 HMM の音声認識性能を改善するための手法

HMM を基礎とした音響モデルを用いた音声認識システムの音声認識性能を改善するための研究には，状態出力確率分布の精密化と，パラメータ共有による頑健化が代表的である．本節では，状態出力確率分布の精密化法として，複数混合ガウス分布を用いた手法と，種々のレベルのパラメータ共有による頑健化手法を述べる．

分布形状の精密化による音声認識性能の改善

- **複数混合ガウス分布**

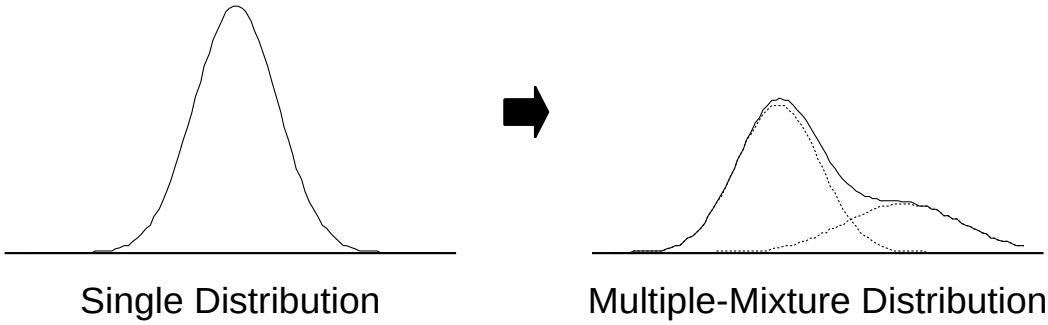


図 2.5: 複数混合分布化による状態出力確率分布の精密化

HMM の状態出力確率分布を複数混合分布化することにより，音響特徴ベクトル時系列の分布形状を，より精密に表現する手法 [25] がある．この複数混合化により，単一のガウス分布では表現できない，複雑な分布形状を表現することができる．図 2.5 に単一ガウス分布と，複数混合ガウス分布の分布形状を示す．式 (2.27) のガウス分布を複数混合化した時の，状態出力確率分布 $b_i(\mathbf{o})$ を式 (2.32) に示す．式中の M は混合数である．また，$\mu_{i,m}$ と $\mathbf{\Sigma}_{i,m}$ は，状態 i，混合成分 m の平均ベクトルと分散ベクトルである．c_m は混合成分 m への分岐確率である．

$$b_i(\mathbf{o}) = \sum_{m=1}^{M} c_m \mathcal{N}(\mu_{i,m}, \mathbf{\Sigma}_{i,m}, \mathbf{o}) \tag{2.32}$$

$$\text{ただし} \quad \sum_{m=1}^{M} c_m = 1$$

パラメータ共有化を用いた頑健化による音声認識性能の改善

出力確率分布のモデルを精密化するに従って，自由パラメータ数は一般に増加する．自由パラメータの増加はモデルの汎化性能の低下に繋がるため，類似した特性を持つパラメータ同士を共有化することにより，モデルの統計的信頼性を改善する手法が多数提案されている．これらのパラメータ共有法は，大きく 4 つのレベルに分類することができる．

1. **環境依存音素レベル** [32, 31]

   類似した環境依存音素のモデル同士を共有化する手法である．

2. **状態レベル** [33, 35, 38, 39]

$_{a}$/k/$_{a}$ と $_{a}$/k/$_{i}$ の環境依存音素の振る舞いを 3 状態の HMM で学習した場合，先行音素が母音の/a/であるため，第 1 状態同士は類似した分布形状を持つと考えられる．この様な状態共有構造を決定することにより，効果的に学習データをモデル化することができると考えられる．

3. **分布レベル** [40, 41, 42, 43, 44]

   個々の状態が複数混合分布により表現されている HMM において，各々の混合分布を共有化する手法である．分布レベルの共有により，状態レベルよりも更に少ないパラメータ数で音響空間を覆うことができると考えられる．

4. **特徴量分布レベル** [45]

   対角共分散型の多次元ガウス分布は，個々の特徴量のスカラー分布（1 次元分布）の積として計算される．この様なスカラー分布間で距離の近い分布同士を共有化することにより，分布レベルよりも更に効率的に音響空間を表現することができると考えられる．

無限の学習データが用意可能な場合，上述のようなパラメータ共有法は基本的に必要ではない．しかしながら，学習データは常に有限である．高性能な音声認識を実現するためには，実際の分布を限られた学習データのみで，可能な限り詳細かつ統計的信頼性を保ったまま推定しなければならない．しかし，パラメータ共有を行わずにモデルを単純に詳細化した場合，個々のパラメータを推定するために用いられる学習サンプル数が極度に減少するため，学習データへ過剰に適応され，結果として実際の分布形状とは大きく異なったモデルが推定される．このような過学習モデルは，音声認識性能の低下に繋がる．上述のパラメータ共有法により,（分布間距離の近いもの同士の共有などの単純なものだけではなく）類似した特性を持つパラメータ同士を共有化することにより，個々のパラメータを推定するために使用される学習サンプル数を増加させ，統計的信頼性を改善することができる．ただし，パラメータ共有を極度に行った場合，モデルの詳細さは失われるため，高性能な音声認識を行うためには，学習データ量に応じた共有を行なわなければならない．

第 3章

個別特徴量の非同期性のモデル化

本章では，個々の特徴量の値が非同期なタイミングで変化する特徴ベクトル時系列の効果的なモデル化法について議論する．本議論の目的は，このような特徴ベクトル時系列を，より詳細かつ頑健に表現することのできるモデルを構築し，音声認識性能を改善させることである．

3.1　個々の特徴量の時間非同期性

一般に音声認識に用いられる音響特徴ベクトルは，ケプストラムなどの音響特徴量だけでなく，それらの時間微分特徴量や 2 次時間微分特徴量など複数の特徴量から構成されている．これらの特徴量はお互いに非同期なタイミングで値が変化していると考えられる．本節では，個々の特徴量間の時間非同期性について議論する．その後，時間非同期性を積極的に考慮することにより期待されるモデル化効率の改善について議論する．

3.1.1　個々の音響特徴量間の同期と非同期

冒頭で述べたように，音声認識に用いられる音響特徴ベクトルは，様々な特徴量から構成されている．これらの特徴量はお互いに非同期なタイミングで値が変化している．例えば，音響特徴量の値が一定の割合で変化している区間では，定義上，その時間微分特徴量の値は一定となり，これら 2 つの特徴量の値は，お互いに非同期なタイミングで変化することが考えられる．また，同様の理由により，時間微分特徴量の値が一定の割合で変化している区間では，2 次時間微分特徴量の値は一定となる．

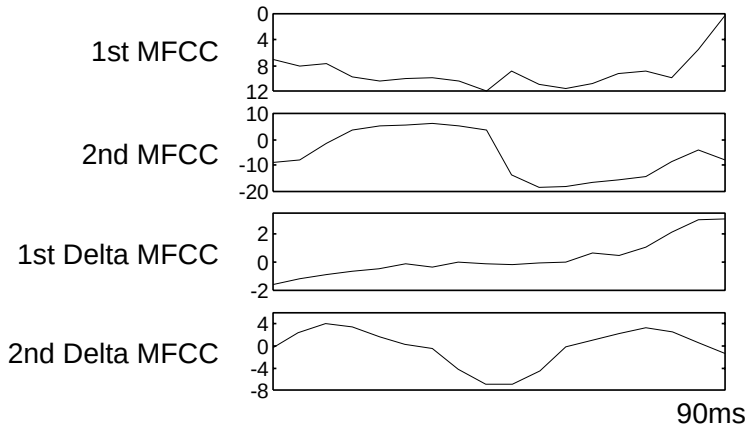
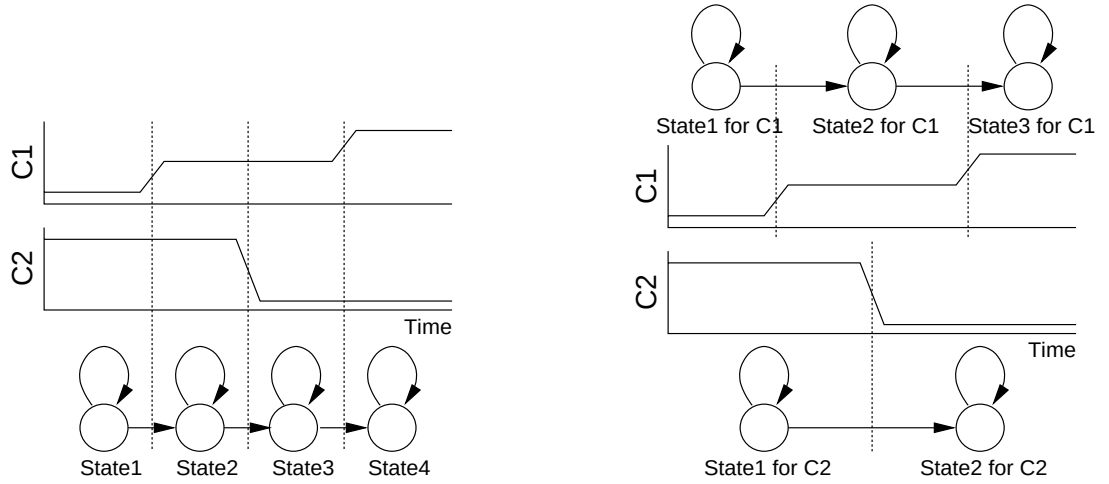


図 3.1: 個々の特徴量の値がお互いに非同期なタイミングで変化している環境依存音素 ₐ/k/ₐ の例

個々の特徴量の時間変化が非同期な例として，図 3.1 に環境依存音素 ₐ/k/ₐ（中心音素は/k/，先行および後続音素は/a/）のサンプルを示す．第 2 MFCC は中央付近で値が変化しているのに対して，第 2 ΔMFCC は中央付近で値は一定となっている．また同様に，第 1 MFCC は中央付近で値は変化せず，後半で値が変化している．従来の HMM は，多次元ベクトル分布を持つ状態の連鎖であるため，ベクトルを構成するすべての特徴量の値が HMM の状態遷移に同期して変化することを暗に仮定したモデルであり，このような個々の特徴量間の非同期性を考慮したモデルとはなっていない．個々の特徴量間の非同期な値の変化を考慮することにより，より効果的な音響特徴ベクトル時系列のモデル化が可能になると考えられる．

3.1.2 非同期な値の変化のモデル化

図 3.2 に，個々の特徴量の値が非同期なタイミングで変化する特徴ベクトル時系列の模式図を示す．この特徴ベクトル時系列は，前半で第 1 特徴量，中央付近で第 2 特徴量，後半で第 1 特徴量の値が変化している．従来型 HMM を用いてモデル化する場合，この特徴ベクトル時系列は 4 つの定常区間から構成されているため，全体として 4 つの状態が必要であった．また，従来型 HMM の個々の状態が対角共分散型のガウス分布によりモデル化されている場合，この特徴ベクトル時系列は 2 次元であるため，各状態は 2 つのスカラー



(A) Modeling of conventional HMM

(B) Modeling of individual features Independently

図 3.2: 個々の特徴量が状態遷移に同期して変化する従来の HMM と，個々の特徴量が非同期に状態遷移するモデル

分布を必要とする．図 3.2-(A) に，従来型 HMM を用いてモデルした場合の，Viterbi セグメンテーションにおける各状態の受け持ち区間を示す．図の様に，この特徴ベクトル時系列は，全体として 8 つのスカラー分布（4 状態 $\times$2 次元）によりモデル化される．

しかしこのモデル化の場合，第 1 特徴量の第 2 状態と第 3 状態は等しい値であるにも拘らず異なる状態に分割され，また，第 2 特徴量の第 1 状態と第 2 状態，第 3 状態と第 4 状態も等しい値であるにも拘らず異なる状態に分割されている．もし，第 1 特徴量は前半と後半で分布が変化し，第 2 特徴量は中央付近で分布が変化するならば，このベクトル時系列は 5 つのスカラー分布（第 1 特徴量は 3 つのスカラー分布，第 2 特徴量は 2 つのスカラー分布を使用）によりモデル化することが可能と考えられる．図 3.2-(B) は，各特徴量の状態を非同期に遷移させるため，個々の特徴量をスカラー量を出力する HMM（スカラー HMM）によりモデル化した例である．図の様に，個々の特徴量間で非同期に値が変化している時系列信号は，すべての特徴量が状態遷移に同期して変化する従来型 HMM より，少ないスカラー分布数でモデル化することができ，モデルの統計的信頼性が改善すると考えられる．

3.2 非同期遷移型 HMM

本章では，個々の特徴量がお互いに非同期に状態遷移する新しいモデルとして，非同期遷移型 HMM（Asynchronous Transition HMM: ATHMM）を提案する．AT-HMM を用いることにより，少ないパラメータ数で従来型 HMM よりも複雑な特徴ベクトル時系列を表現することができ，モデルの統計的信頼性の改善に繋がると考えられる．本節では，個々の特徴量間の時間非同期遷移構造を分類した後，それらの統一的な実現法として直積型 HMM を用いた方法を述べる．

3.2.1 時間非同期遷移構造の分類

同期非同期構造の分類

個々の特徴量の値の変化のタイミングは，大きく 3 つに分類することができる．図 3.3 に，種々の同期非同期構造を示す．

(A) 完全同期 (Synchronous)

全ての特徴量の状態が同期して遷移するタイプであり，従来の HMM はこれに該当する．

(B) 部分非同期 (Partially asynchronous)

同期点と非同期点を混在させる方法である．ここでは，代表的な混在方法として 3 つの構造を説明する．これらの混在方法を組み合わせることにより，より複雑な部分非同期を考えることができる．

(B-1) 特徴グルーピング (Feature grouping)

時間変化が類似した複数の特徴量をグループ化し，グループ内では同期，グループ間では非同期に遷移するタイプである．音声波形と唇の動画像では，唇の動きの後に発声が開始される傾向にある．この特徴グルーピングは，このような異なる特徴量の間の非同期性を表現する時に有効と考えられる．

(B-2) モデル同期 (Model synchronous)

音素や音韻などのモデル境界で同期し，モデル内部では非同期に状態遷移するタイプである．このような部分的な同期を導入することにより，ある音声区間に

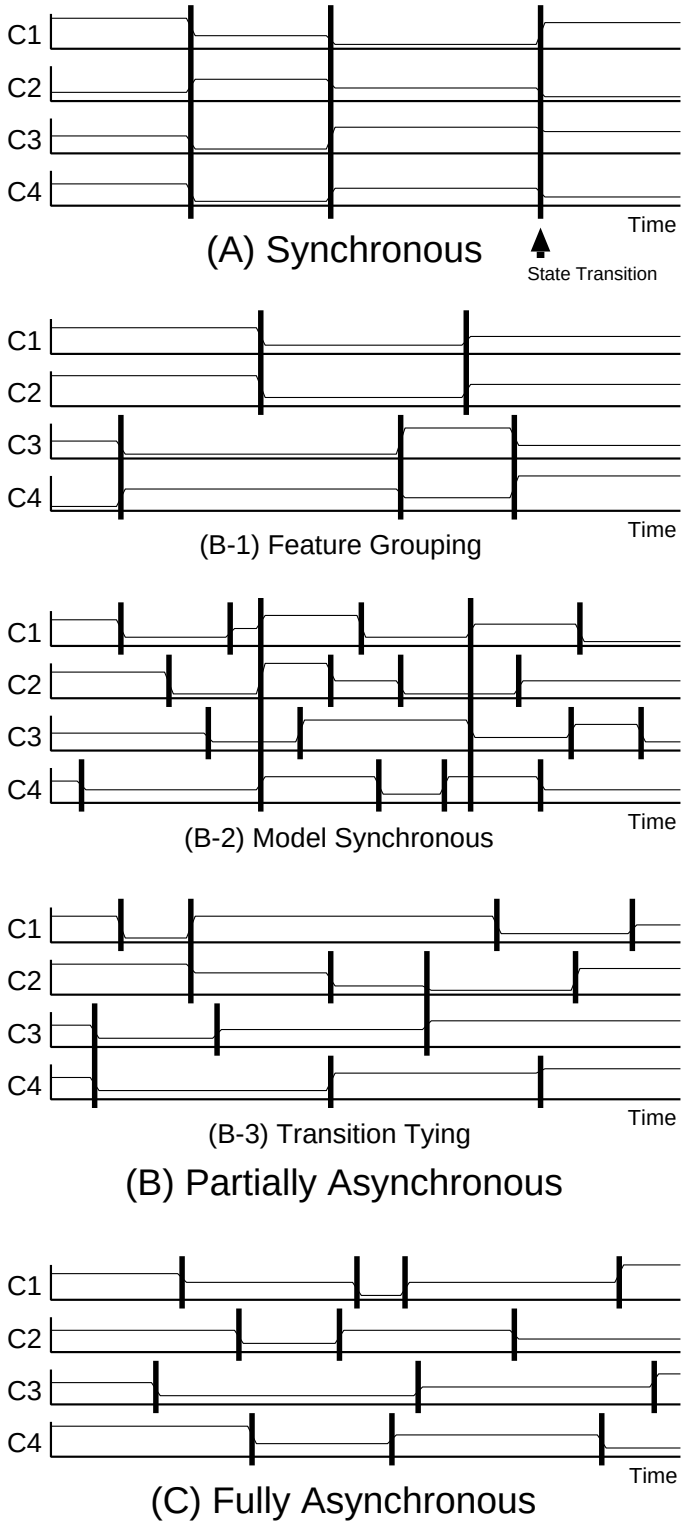


図 3.3: 同期非同期構造の分類

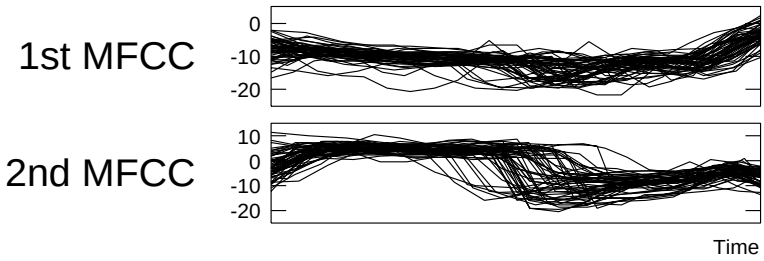


図 3.4: 個々の特徴量の値の変化に順序関係がある音素サンプルの例

おいて，個々の特徴量の特徴ベクトル時系列が異なる音素としてセグメンテーションされることを防ぐ効果があると考えられる．

(B-3) 遷移部分共有 (Transition tying)

以上に述べた 2 種類の同期の導入原理とは異なり，より一般的に同期と非同期が混在するタイプである．同期拘束を持つ方がモデルとして効率的である状態遷移の組合せは同期し，同期拘束を持たない方が効率的である組合せには非同期であるように拘束を用いることにより，効率的なモデル化を行うものである．

(C) 完全非同期 (Fully asynchronous)

個々の特徴量間の時間変化は完全に非同期と考え，特徴量毎に別々に状態遷移するタイプである．個々の特徴量の値の変化が，他の特徴量の値の変化と全く独立の場合に有効と考えられる．

非同期な遷移に対する順序制約の分類

以上の同期非同期構造の違いの他に，個々の特徴量の状態遷移に対して順序の制約を付加することができる．図 3.4 に ATR 研究用日本語音声データベース set-A 中に含まれている環境依存音素 ₐ/k/ₐ のサンプルを示す．これらのサンプルは全体として，第 2 MFCC，第 1 MFCC の順序で状態が遷移していると見ることができる．順序関係の有無による分類を次に示す．この分類では，(I)，(II)，(III) の順に，除々に順序の制約が厳しくなる．

(I) 順序制約無し (Non-sequenced)

状態遷移に順序関係は無く，個々の特徴量の状態遷移はお互いに独立に生起する．2 つの特徴量を別々にモデル化した 2 つの HMM の状態遷移は，お互いに独立であり，

全く関連無く発生する．

(II) 準順序制約付き (Quasi-sequenced)

(I) 順序制約無しと，(III) 順序制約付きの中間に位置する制約である．2 つの特徴量を別々にモデル化した 2 つの HMM において，第 1 特徴量の HMM が第 i 状態から観測信号を出力する場合，第 2 特徴量の HMM からは，第 i から第 $i \pm a$ まで（a は固定値）の状態からのみ観測信号が出力されるなどの制約を持つモデルである．この様な「時間ずれ」は，唇と音声の間の非同期性を扱った Audio-Visual 音声認識で用いられている．

(III) 順序制約付き (Sequenced)

各特徴量の状態遷移に順序関係を持つ手法である．2 つの特徴量を別々にモデル化した 2 つの HMM の状態遷移は，完全に順序付けられている．例として，第 1 特徴量，第 2 特徴量，第 1 特徴量，第 2 特徴量などの順序以外では，状態遷移しないモデルである．

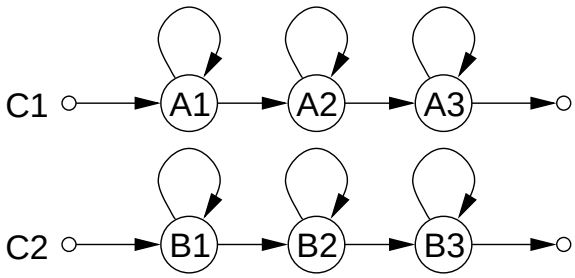
3.2.2 AT-HMM の実現法

スカラー HMM による AT-HMM の実現

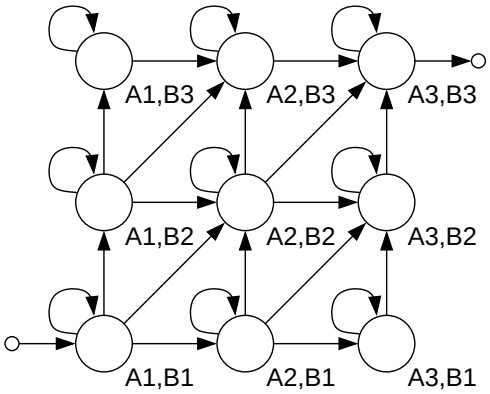
前述の種々の時間非同期遷移構造の中の，(C) 完全非同期な (I) 順序制約無し AT-HMM は，個々の特徴量の状態がお互いに独立に遷移するモデルである．このような AT-HMM は，図 3.5-(1) に示すように各特徴量の振舞いをモデル化したスカラー HMM の状態遷移を別々に Viterbi アルゴリズムにより計算することで実現することができる．与えられた特徴ベクトル時系列に対するモデルの尤度は，個々のスカラー HMM の尤度の積により計算される．また，(B-2) モデル同期な (I) 順序制約無しの状態遷移は，2 段 DP 法を基礎とする方法で計算することができる．

直積 HMM による統一的な AT-HMM の実現

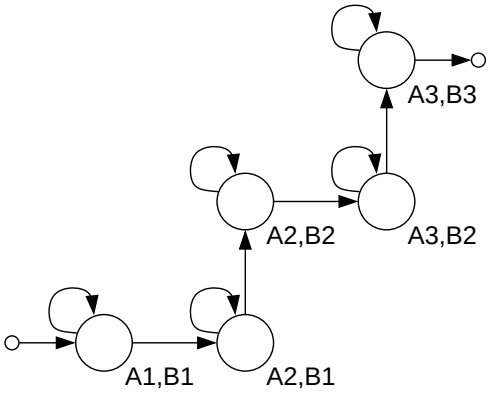
種々の時間非同期遷移構造や順序制約の有無を統一的に表現する手法として，本章中で「直積 HMM」と呼ぶ構造を用いた手法を述べる．直積 HMM とはスカラー値やサブベクトル値を出力する HMM を並列に動作させ，それらの直積としてベクトルを観測するモデル



(1) Non−sequenced AT−HMM implemented using scalar HMMs.

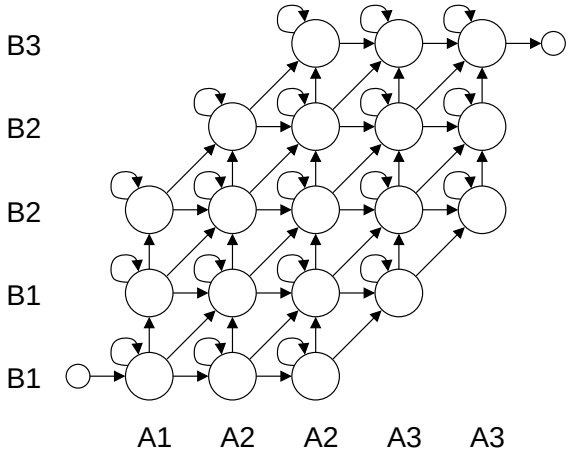


(2) Non−sequenced AT−HMM implemented using direct−product HMM.

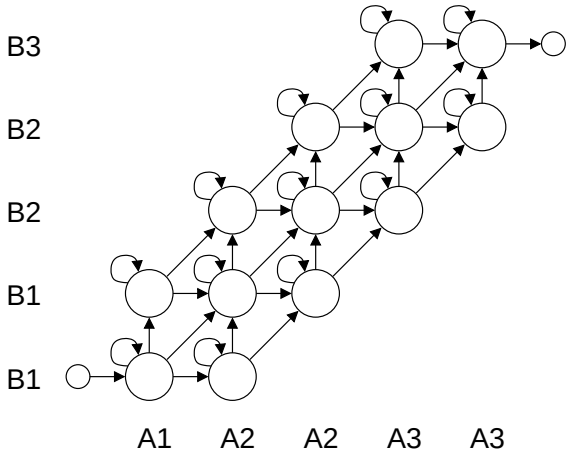


(3) Sequenced AT−HMM implemented using direct−product HMM.

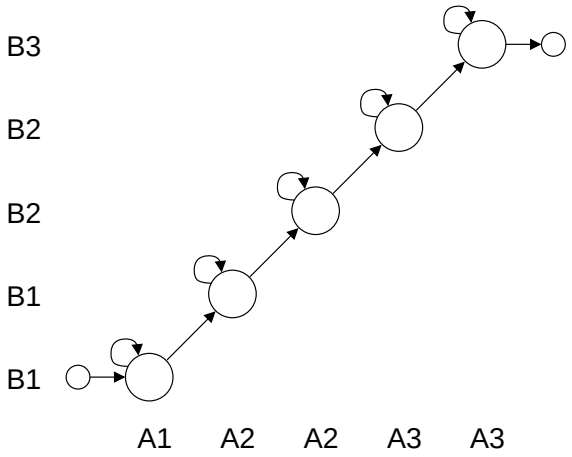
図 3.5: スカラー HMM を基礎とした順序制約無し AT-HMM と直積 HMM を基礎とした順序制約無し及び順序制約付き AT-HMM の実現



(a=2) Limited state asynchrony



(a=1) Limited state asynchrony



(a=0) Limited state asynchrony

図 3.6: 順順序制約付き AT-HMM の実現（a は時間ずれ状態数を表す）

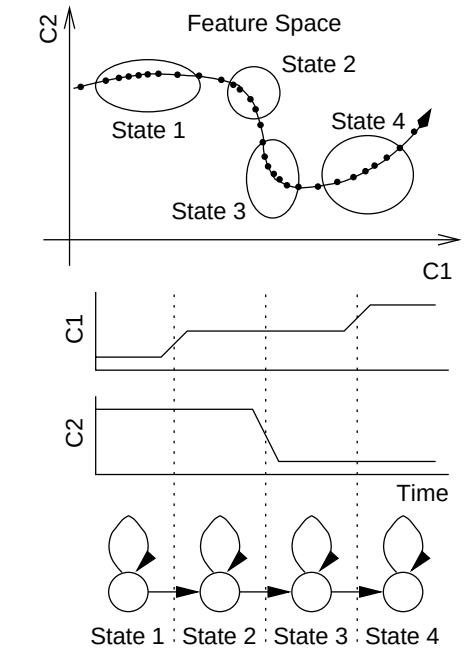
である．これを 1 つの HMM として見ると，個々のストリームに対応する HMM の状態列の直積により表現される．図 3.5-(2) に，図 3.5-(1) の 2 つのスカラー HMM を直積化した，2 次元の直積 HMM の例を示す．図のように直積 HMM の個々の状態は，第 1 特徴量のスカラー HMM における第 a 状態と第 2 特徴量における第 b 状態の直積により表現され，各々の状態の 1 次元分布から構成された 2 次元の分布を持つ．前節で述べたような種々の同期非同期構造や順序制約の有無は，個々の直積状態間の遷移を制限することにより実現することができる．

まず，直積 HMM により実現された (C) 完全非同期な (I) 順序制約無し AT-HMM の構造を図 3.5-(2) に示す．(B-1) 特徴量グルーピングは，多次元ベクトル HMM の直積 HMM により実現することができる．また (B-2) モデル同期は，図 3.5-(2) の構造を複数個，時間方向に接続することにより，接続点で同期する状態遷移が実現できる．次に，(C) 完全非同期な (III) 順序制約付き AT-HMM は，図 3.5-(3) に示す構造により実現することができる．図のように，開始状態から終了状態への状態列を 1 つに絞ることにより順序制約の付いた状態遷移が実現される．また，図 3.5-(3) の (III) 順序制約付き AT-HMM に対して，時間ずれを考慮した (II) 準順序制約付き AT-HMM は，図 3.6 に示す構造により実現される．図の例では，時間ずれ状態数 a を 0, 1, 2 とした時の構造である．(B-1) 特徴量グルーピングや (B-2) モデル同期は，順序制約無しと同様の方法により実現することができる．最後に，(A) 完全同期などの従来型 HMM は，開始状態から最終状態への最短状態列として表現される．このように状態遷移を適当に制限することにより，(B-3) 遷移部分共有などの複雑な時間非同期遷移構造を実現することができる．

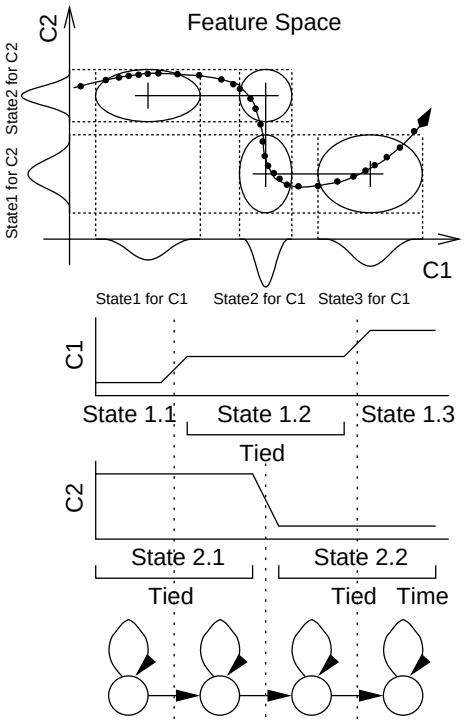
音声認識では，一般に 20 から 30 次元程度の音響特徴ベクトルが用いられている．従って，直積 HMM を基礎とした完全非同期な順序制約無し AT-HMM は，個々のスカラー HMM の状態数を I，特徴ベクトルの次元数が D の場合，状態数は I^D となり，尤度の計算に必要な計算量は膨大である．それに対して，順序制約付き AT-HMM は，HMM を基礎とした音響モデルにおいて最も広く用いられている left-to-right 型 HMM と同じ構造を持つため，状態列内の状態数を適度に抑えることで高速に尤度計算が行えると考えられる．

3.2.3 時間方向共有法による順序制約付き AT-HMM の実現

直積 HMM を基礎とした順序制約付き AT-HMM は，隣り合った状態間のスカラー分布の共有構造を持つ．図 3.5-(3) の順序制約付き AT-HMM では，第 1 特徴量における第 2 状



(A) Modeling by conventional HMM



(B) Modeling by the temporal tying technique

図 3.7: 従来型 HMM と時間方向共有構造を用いて実現した AT-HMM のパラメータ共有構造

態と第3状態及び，第4状態と第5状態のスカラー分布が各々共有化され，また第2特徴量における第1状態と第2状態及び，第3状態と第4状態のスカラー分布が各々共有化されている．本論文では，このように時間的に隣り合った状態間のスカラー分布を共有化する手法を，時間方向共有法（Temporal Tying Technique）として提案する．図3.7は，非同期に変化するベクトル時系列を従来型HMMによりモデル化した場合と，時間方向共有法を用いてモデル化した場合のパラメータ共有構造を模式的に示している．時間方向共有法により，第1特徴量の第2状態と第3状態，第2特徴量の第1状態と第2状態，第3状態と第4状態が各々共有化される．従来のHMMでは8つ必要であったスカラー分布が，時間方向共有法により5つのスカラー分布で表現することができる．この共有構造により，第1特徴量，第2特徴量，第1特徴量の順序でのみ状態が遷移する制約が直接表現される．

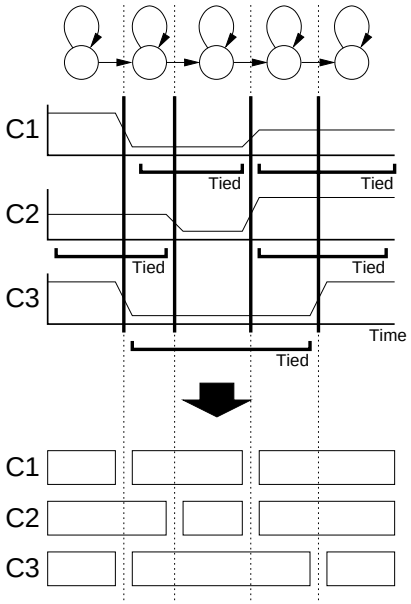
本章では，時間非同期遷移構造を持つ音響モデルとして，時間方向共有法により実現された順序制約付きAT-HMMの議論を主に行なう．

3.3 順序制約付きAT-HMMの生成法

時間方向共有構造により実現される順序制約付きAT-HMMを用いて，個々の特徴量がお互いに非同期に変化する特徴ベクトル時系列を効果的にモデル化するためには，その特徴ベクトル時系列の時間非同期遷移構造を反映した時間方向共有構造を決定する必要がある．本節では，効果的な時間方向共有構造の決定法について議論する．

3.3.1 1つの状態列を決定する近似的手法

順序制約付きAT-HMMの時間方向共有構造を決定する問題は，直積HMMにより実現された順序制約無しAT-HMMの状態遷移を，一つの状態列となるように制限する問題と等価である．現在，高性能な音響モデルを生成するため最尤基準を用いたモデル生成法が広く用いられている．最尤基準を用いて一つの状態列を決定する場合，この問題は，直積HMMに含まれるあらゆる状態列に対して，各々モデルパラメータの再推定を行い，最も大きな尤度の得られる状態列を探索する問題となる．しかし，直積HMMに含まれる状態数はI^D個となるため，膨大な組合せの状態列に対する尤度を計算し比較する必要があると考えられ大変困難である．これらの理由により，近似的に状態列（時間方向共有構造）を決定する手法を検討する必要がある．

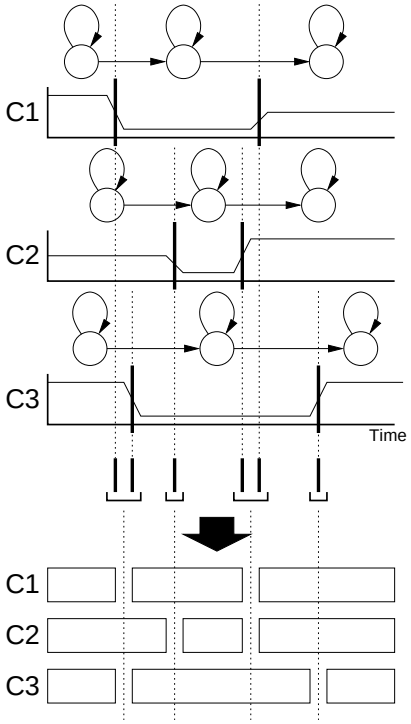


1) Creation of a conventional HMM

2) Tying of scalar distributions that are close to each other.

3) Determine a temporal tying structure.

(I) Generating temporal tying structure using conventional HMM



1) Estimation of a scalar HMM for each of the individual features and calculation of state tran− sition timings from individual transition probabilities.

2) Clustering of state transition timings.

3) Determine a temporal tying structure.

(II) Generating temporal tying structure using scalar HMMs

図 3.8: 従来型 HMM を用いた生成法とスカラー HMM を用いた生成法

(I) 従来型 HMM を用いた生成法

図 3.8-(I) に，時間的に隣合った状態間のスカラー分布を除々に共有化する生成法を示す．この手法は最初に，状態遷移に同期して全ての特徴量の値が変化する従来型 HMM を生成し，その後，時間的に隣合った状態間において，お互いに距離の小さなスカラー分布から除々に共有化する手法である．

(II) スカラー HMM を用いた生成法

図 3.8-(II) に，個々の特徴量の学習データを用いて学習されたスカラー HMM の状態遷移確率から時間方向共有構造を決定する生成法を示す．個々の状態遷移確率は，各状態の平均停留時間（の逆数）を表している．この平均停留時間を時間的に並べて全体を正規化することにより，個々の特徴量における状態遷移の時間的なタイミングから，時間方向共有構造（時間非同期遷移構造）を決定することができると考えられる．

本議論では，これら 2 つの生成法の内，(II) スカラー HMM を用いた手法により生成される順序制約付き AT-HMM を中心に評価実験を行う．

3.3.2 スカラー HMM を用いた生成法の処理の流れ

(II) スカラー HMM を用いた生成法において，個々の特徴量のスカラー HMM の状態遷移の時間的なタイミングは $D \times (I+1)$ 個（音響特徴ベクトルの次元数が 26，音素 HMM の時間方向状態数が 3 の場合，計算される状態遷移タイミング数は 104 である）だけ計算される．この状態遷移タイミングの集合から単純に時間方向共有構造を決定した場合，$D \times I - (D-1)$ 個（特徴ベクトルの次元数が 26 及び状態数が 3 の場合，時間方向状態数は 53 である）の時間方向状態数が必要となる．このような時間方向に大量の状態を持つモデルは，尤度計算に多大なトレリス計算が必要である．また計算時間の問題だけでなく，与えられた音響特徴ベクトル時系列のサンプル数よりも時間方向状態数の方が多くなってしまい，正しい尤度を計算することができず，音声認識性能の低下に繋がると考えられる．そこで，本生成法では，状態遷移タイミングの集合を $N+1$ 個の代表点にクラスタリングすることにより，時間方向状態数 N の順序制約付き AT-HMM を生成する手法を提案する．このクラスタリング処理により，計算時間の削減及び，発声時間の短い音素に対する認識性能の改善を考慮した．また，極端に少ないサンプル数の環境依存音素に対する時間非同

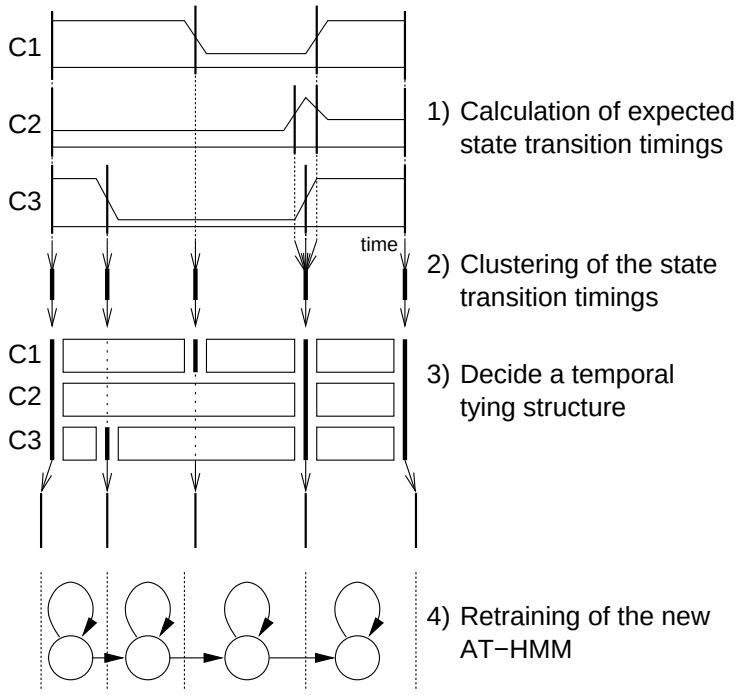


図 3.9: 状態遷移タイミングからの時間方向共有構造の生成

期遷移構造の推定精度を改善するため，類似した時間非同期遷移構造を持つ環境依存音素モデルの状態遷移確率行列を M 個に共有化した．本生成法の処理の流れを次に示す．

Step 1: 状態遷移タイミングベクトルの計算

従来型 HMM と同一の状態共有構造を持つスカラー HMM を用意する．このスカラー HMM の状態遷移確率行列は，学習データ中に含まれた環境依存音素毎に別々に持つ．次に，個々の特徴量の学習データによりスカラー HMM を音素境界既知の条件下で学習し，得られた個々の音素環境 e のモデルから，状態遷移タイミングベクトル v^e を計算する．式 (3.1) に状態遷移タイミングベクトルを示す．$l_{i,d}$ は，特徴量 d の状態 i からの状態遷移タイミングを表し，式 (3.2) により計算される．また，この状態遷移タイミングは，各状態の平均停留時間 $\tau(i,d)$ から計算される．平均停留時間は，式 (3.3) に示すように，各状態の自己遷移確率 $a_{i,d}$ から計算される．

$$\mathrm{v}^e = (l_{0,1}, l_{1,1}, l_{2,1}, \cdots, l_{i,d}, \cdots, l_{I,D}) \tag{3.1}$$

$$l_{i,d} = \frac{\sum_{j=1}^{i} \tau(j,d)}{\sum_{j=1}^{I} \tau(j,d)} \tag{3.2}$$

$$\tau(i,d) = \frac{1}{1 - a_{i,d}} \tag{3.3}$$

Step 2: 状態遷移タイミングベクトルのクラスタリング

Step 1 で得られた個々の環境の状態遷移タイミングベクトル集合 $\{\mathrm{v}^e\}$ を M 個にクラスタリングする．その後，クラスタリング結果に基づき個々の特徴量のスカラー HMM の状態遷移確率行列を共有化し再学習する．

Step 3: 時間方向共有構造の決定

M 個のクラスタ各々に対して，時間方向共有構造を決定する．時間方向共有構造の決定処理の流れを図 3.9 に示す．まず，各クラスタの状態遷移確率行列から，状態遷移タイミングベクトルを計算する．次に，得られた状態遷移タイミングの集合 $\{l_{i,d}\}$ を $N+1$ 個の代表状態遷移タイミングへクラスタリングする．この集合には，開始状態への遷移タイミングと最終状態からの遷移タイミングが含まれている．最後に，クラスタリング結果に基づき，個々の特徴量の状態遷移と同じタイミングで特徴量の値が変化するように，時間方向共有構造を決定する．

図 3.9 は，時間方向状態数を 4 としたときのクラスタリングの様子を模擬的に図示したものである．すべての特徴量において，開始状態への遷移タイミングと，最終状態からの遷移タイミングが，各々一つの代表点にクラスタリングされている．また，第 1 特徴量の二つ目，第 3 特徴量の二つ目及び，第 2 特徴量の一つ目と二つ目の，合計 4 つの状態遷移タイミングが一つの代表点にクラスタリングされている．第 2 特徴量において二つ目と三つ目が一つの代表点にクラスタリングされ，その間のスカラー分布は消滅している．時間方向共有構造としては，第 1 特徴量において第 1 状態と第 2 状態のスカラー分布が共有化され，第 2 特徴量においては第 1 状態，第 2 状態，第 3 状態が共有化されている．遷移タイミングを計算する段階において 3 つの状態を持っていた第 2 特徴量は，このクラスタリングによって 2 つの状態になる．また，第 3 特徴量では第 2 状態と第 3 状態が共有化されている．これらの処理により個々の特徴量をモデル化したスカラー HMM から，4 状態の順序制約付き AT-HMM が生成さ

れる．HMM の状態数は増加しているが，分布を表現するために使用したパラメータ数はほぼ等しい．

この生成法は，個々の環境依存音素は音素境界既知の条件下で学習を行うことにより非同期遷移構造を推定し，また **Step 3** において状態遷移タイミングの集合をクラスタリングしているため，(B-2) モデル同期と (B-3) 遷移共有が混在した時間非同期遷移構造が生成される．

3.4 時間方向状態数に対する評価実験

時間方向共有構造により実現された順序制約付き AT-HMM は，時間方向状態数を増加させたとしても，モデル全体のパラメータ数を保ったまま，時間的な分解能が除々に精密化され音声認識性能の改善が期待される．本節では，順序制約付き AT-HMM の時間方向状態数の増加による音声認識性能の改善を検証する．

3.4.1 実験条件

特定話者における日本語音素接続制約付きの連続音素認識実験により評価を行なう．順序制約付き AT-HMM の時間方向共有構造は，前節で提案したスカラー HMM を用いた手法により生成した．個々の特徴量のスカラー HMM の状態共有構造は，(C) 完全同期の従来型 HMM の状態共有構造を HTK[7] により音素分類木を基礎とした状態のトップダウンクラスタリングを用いて生成し，個々の特徴量の状態共有構造として用いた．そのため，全ての特徴の状態共有構造は同一である．スカラー HMM の時間方向状態数は 3，状態数は 800 状態（単一ガウス分布を持つ AT-HMM でスカラー分布数 20800）である．AT-HMM の時間方向状態数 N は 3 から 10 の 8 種類について評価した．時間方向共有構造のクラスタ数 M は 1989 を用いた．これは，学習データ中に含まれていた環境依存音素（1989 種類）各々に対して別々の時間方向共有構造を決定したモデルである．

学習データには，ATR 研究用日本語音声データベース set-A の，男性話者 mht の重要語 5240 単語中の奇数番目と音素バランス単語 216 単語を使用し，評価データには重要語 5240 単語中偶数番目の 4 分の 1 を使用した．音素ラベルは，付録 A の計 26 音素を使用した．サンプリング周波数 12kHz の波形データをフレーム長 25ms，フレーム周期 5ms，ハミング窓を掛けて分析した．特徴パラメータは，対数パワー，12 次 MFCC，Δ 対数パワー，12 次

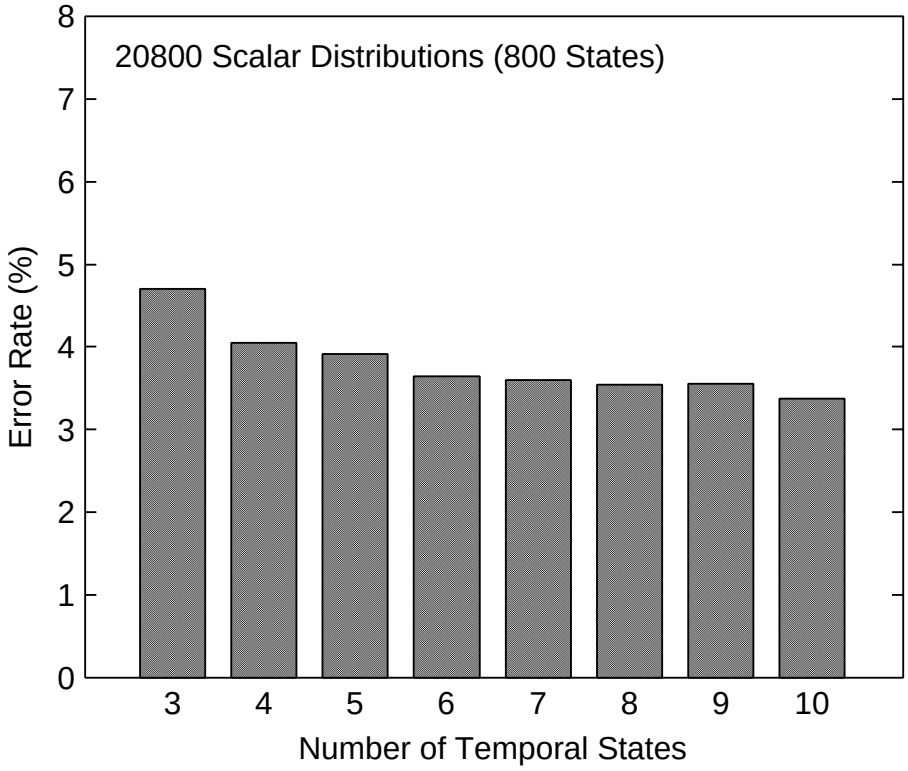


図 3.10: 種々の時間方向状態数を持つ順序制約付き AT-HMM の音素誤り率

ΔMFCC の計 26 次元を使用した.

3.4.2 実験結果

図 3.10 に，時間方向状態数を 3 から 10 まで変化させた場合における，順序制約付き AT-HMM の音素誤り率を示す．図のように，時間方向状態数の増加に従い，音素認識率が除々に改善している.

本実験により，時間方向状態数の増やすことで，時間非同期遷移構造が精密化し，音声認識性能が改善することを確認した.

3.5 順序制約の有無に対する評価実験

本節では，AT-HMM の順序制約の有無による音声認識性能の違いを検証する．図 3.4 に示すように，音声認識に用いられる音響特徴量には，順序の制約が存在すると考えられる.

しかし，このような順序関係が全ての音素環境において全般的に観測される現象であるのか不明である．もし，個々の音響特徴量の値の変化タイミングが，お互いに全く無関係であるば，本議論で述べた順序制約は音声認識性能に悪影響を与えている恐れがある．本節では，順序制約の有無による音声認識性能の評価を行ない，どちらの時間非同期遷移構造がより特徴ベクトル時系列の構造を表すことに適しているか検証を行なう．

3.5.1 実験条件

特定話者の切り出し音素認識実験により評価を行なう．実験に用いたモデルの構造は，完全同期の構造として従来 HMM，順序制約無しの時間非同期遷移構造として，スカラー HMM により実現された AT-HMM，順序制約付きの時間非同期遷移構造として，時間方向共有構造により実現された AT-HMM を用いた．完全同期な従来型 HMM は，HTK[7] により音素分類木を基礎とした状態のトップダウンクラスタリングを用いて生成した．順序制約無し/付き AT-HMM を生成する際に用いた個々の特徴量のスカラー HMM の状態共有構造は，従来型 HMM の状態共有構造を，個々の特徴量のスカラー HMM の状態共有構造として用いた．完全同期な従来型 HMM 及び順序制約無し AT-HMM の時間方向状態数は，3 である．順序制約付き AT-HMM の時間方向状態数は，3, 5, 7 である．時間方向共有構造のクラスタ数 M は 1989 を用いた．これは，学習データ中に含まれていた環境依存音素（1989 種類）各々に対して別々の時間方向共有構造を決定したモデルである．各モデルのスカラー分布数は，5200（従来型 HMM で 200 状態）から 20800 （従来型 HMM で 800 状態）まで 5200 分布毎に 4 種類である．

学習データには，ATR 研究用日本語音声データベース set-A 中，男性 2 話者（mht, mau）女性 2 話者（fms, ffs）の重要語 5240 単語中の奇数番目と音素バランス単語 216 単語（計 2836 単語 $\times$4 話者）を使用し，評価データには重要語 5240 単語中偶数番目の 4 分の 1（計 665 単語 $\times$4 話者）を使用した．音素ラベルは，付録 A の計 26 音素を使用した．サンプリング周波数 12kHz の波形データをフレーム長 25ms，フレーム周期 5ms，ハミング窓を掛けて分析した．特徴パラメータは，対数パワー，12 次 MFCC，Δ 対数パワー，12 次 ΔMFCC の計 26 次元を使用した．

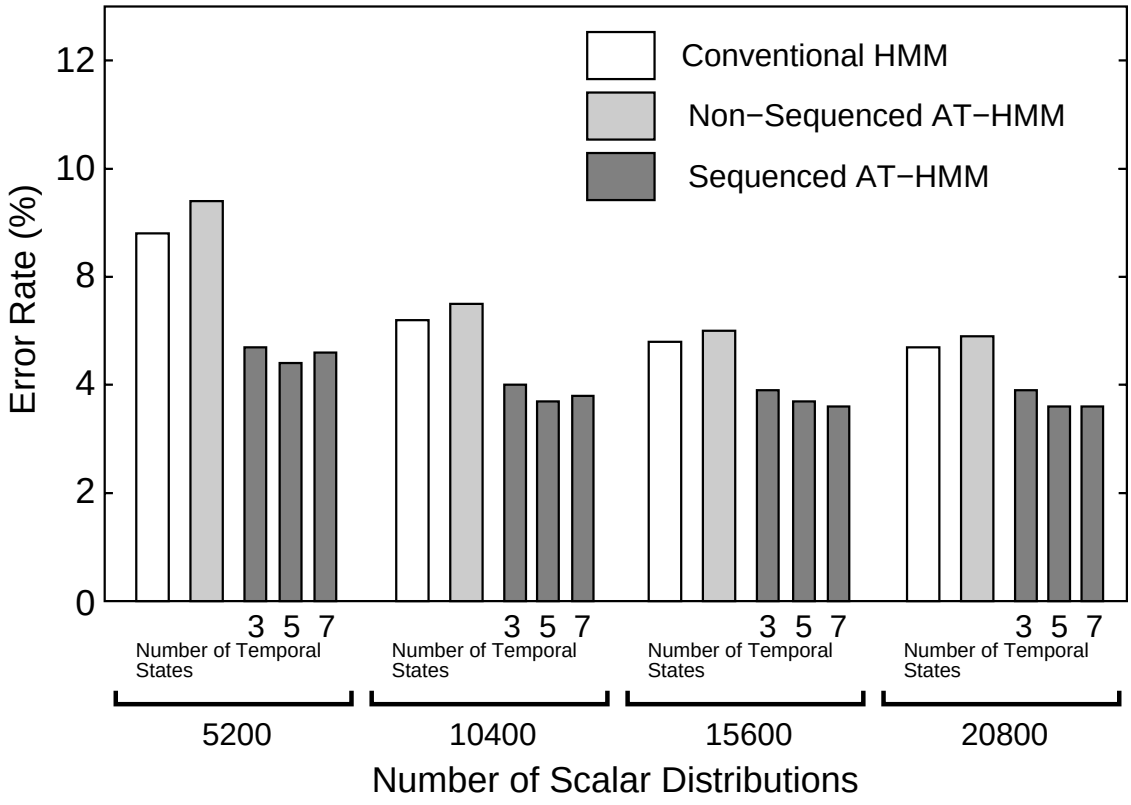


図 3.11: 完全同期な従来型 HMM と順序制約無し/付き AT-HMM の音素誤り率

3.5.2 実験結果

図 3.11 に，各スカラー分布数における完全同期な従来型 HMM と，順序制約無し及び順序制約付き AT-HMM の音素誤り率を示す．図のように，全てのスカラー分布数のモデルで順序制約付き AT-HMM は，完全同期な従来型 HMM より高い音素認識率が得られた．しかし，順序制約無し AT-HMM は，従来型 HMM よりも低い音素認識率となった．

この実験結果から，音響特徴ベクトル時系列の個々の特徴量の値の変化タイミングは，全くの無関係ではないことがわかる．完全に無相関な順序制約無し AT-HMM では，正解モデル以外のモデルに対しても不当に高い確率が計算され，認識性能の低下に繋がったと考えられる．一方，個々の特徴量の値の変化に順序制約を持つ AT-HMM は，個々の特徴量間に相関を持つモデルであり，音声の観測量である MFCC 時系列のモデル化にとって，有効であったと考えられ，音響特徴ベクトル時系列の識別において順序関係は重要であることがわかる．

以上の実験結果は MFCC 系列に関するものであるが，その他の音響特徴量や音声認識以

外のパターン認識システム一般にも共通する可能性がある．

3.6 特定話者連続音素認識実験

本節では，時間方向状態数を増加させた完全同期な従来型 HMM と，順序制約付き AT-HMM の認識性能の比較を行なう．従来型 HMM は，時間方向状態数を増加させた場合，モデルパラメータの多大な増加を招き，その音声認識性能は低下すると考えられる．

更に，順序制約付き AT-HMM の生成法における，時間方向共有構造数 M に対する検証を行なう．もし，時間非同期遷移構造が環境依存音素に依存しているならば，時間方向共有構造数 M が少なくなるに従って，完全同期な従来型 HMM の認識性能へ近付くと考えられる．また，時間方向共有構造自体の共有化により，個々の環境依存音素に対して別々の時間方向共有構造を割り当てた AT-HMM よりも高い認識性能が得られると考えられる．

3.6.1 実験条件

特定話者の連続音素認識実験により評価を行なう．順序制約付き AT-HMM を生成する際に用いた個々の特徴量のスカラー HMM の状態共有構造は，(C) 完全同期の従来型 HMM の状態共有構造を HTK[7] により音素分類木を基礎とした状態のトップダウンクラスタリングを用いて生成し，個々の特徴量の状態共有構造として用いた．そのため，全ての特徴の状態共有構造は同一である．スカラー HMM の時間方向状態数は 3，状態数は 200 状態（単一ガウス分布を持つ AT-HMM でスカラー分布数 5200）から 1200 状態（単一ガウス分布を持つ AT-HMM でスカラー分布数 31200）まで 200 状態毎に 6 種類変化させた．AT-HMM の時間方向状態数 N は 3, 5, 7 について検証した．時間方向共有構造のクラスタ数 M は，300 から 1500 まで 300 ごとに 5 種類と，学習データ中に含まれていた環境依存音素（1989 種類）各々に対して別々の時間方向共有構造を決定したモデルを用いた．

比較実験として，従来型 HMM（200 状態から 1200 状態のモデル），時間方向状態数 3, 5, 7 について認識を行った．状態遷移確率行列の共有化による認識性能への影響を取り除くため，AT-HMM と従来型 HMM の状態遷移確率行列を各音素一つに共有化し再学習した．この処理により AT-HMM と従来型 HMM の分布と状態遷移確率行列のパラメータ数は等しくなる．実験に使用したモデルの各状態は，単一の対角共分散行列型のガウス分布を持つ．

学習データには，ATR 研究用日本語音声データベース set-A 中，男性 2 話者（mht, mau）女性 2 話者（fms, ffs）の重要語 5240 単語中の奇数番目と音素バランス単語 216 単語（計 2836 単語 ×4 話者）を使用し，評価データには重要語 5240 単語中偶数番目の 4 分の 1（計 665 単語 ×4 話者）を使用した．音素ラベルは，付録 A の計 26 音素を使用した．サンプリング周波数 12kHz の波形データをフレーム長 25ms，フレーム周期 5ms，ハミング窓を掛けて分析した．特徴パラメータは，対数パワー，12 次 MFCC，Δ 対数パワー，12 次 ΔMFCC の計 26 次元を使用した．

3.6.2 AT-HMM の特定話者連続音素認識性能

図 3.12，3.13，3.14 に，AT-HMM の時間方向状態数を 3, 5, 7 とした時の，種々のスカラー分布数及び時間非同期遷移構造数における音素誤り率を示す．図には，従来型 HMM の時間方向状態数 3, 5, 7 の音素誤り率も併せて示す．

総スカラー分布数 5200 のモデルを除いて，時間方向共有構造をクラスタリングすることにより，環境依存音素毎に別々に時間方向共有構造を決定したモデル（時間方向共有構造のクラスタ数 1989）よりも高い認識率が得られた．時間方向共有構造数 M を適当に設定することにより，サンプル数の少ない環境依存音素に対する時間非同期遷移構造の推定精度の改善効果があったと考えられる．逆に，時間方向共有構造数 M を少なくするに従い，従来型 HMM の音素認識率へ近付いていった．これは，時間非同期遷移構造が環境依存音素に依存していることを示していると考えられる．また，各スカラー分布数のモデルにおける最小音素誤り率を図 3.15 に示す．図のように，AT-HMM は従来型 HMM と比較して 10%から 40%の誤り削減率が得られた．また AT-HMM は，時間方向状態数の増加に従って認識率が改善した．一方，従来型 HMM はスカラー分布数 5200 から 10400 のモデルにおいて，時間方向状態数の増加に従って認識率が低下した．その原因として，少ない総状態数の従来型 HMM は，各音素の環境依存性を表現するための分布が不足するためと考えられる．AT-HMM は時間方向の状態数を増加させたとしても，分布パラメータ数は一定のまま時間非同期遷移構造が精密化されるため，更に高い認識率が得られたと考えられる．

また，音素誤り率約 5%を得るためには，従来型 HMM はスカラー分布数約 20000 が必要なのに対して，AT-HMM は約 10000 である．このことから，AT-HMM は少ない分布パラメータで従来型 HMM より複雑な音響特徴ベクトル時系列を表現することができると考えられる．

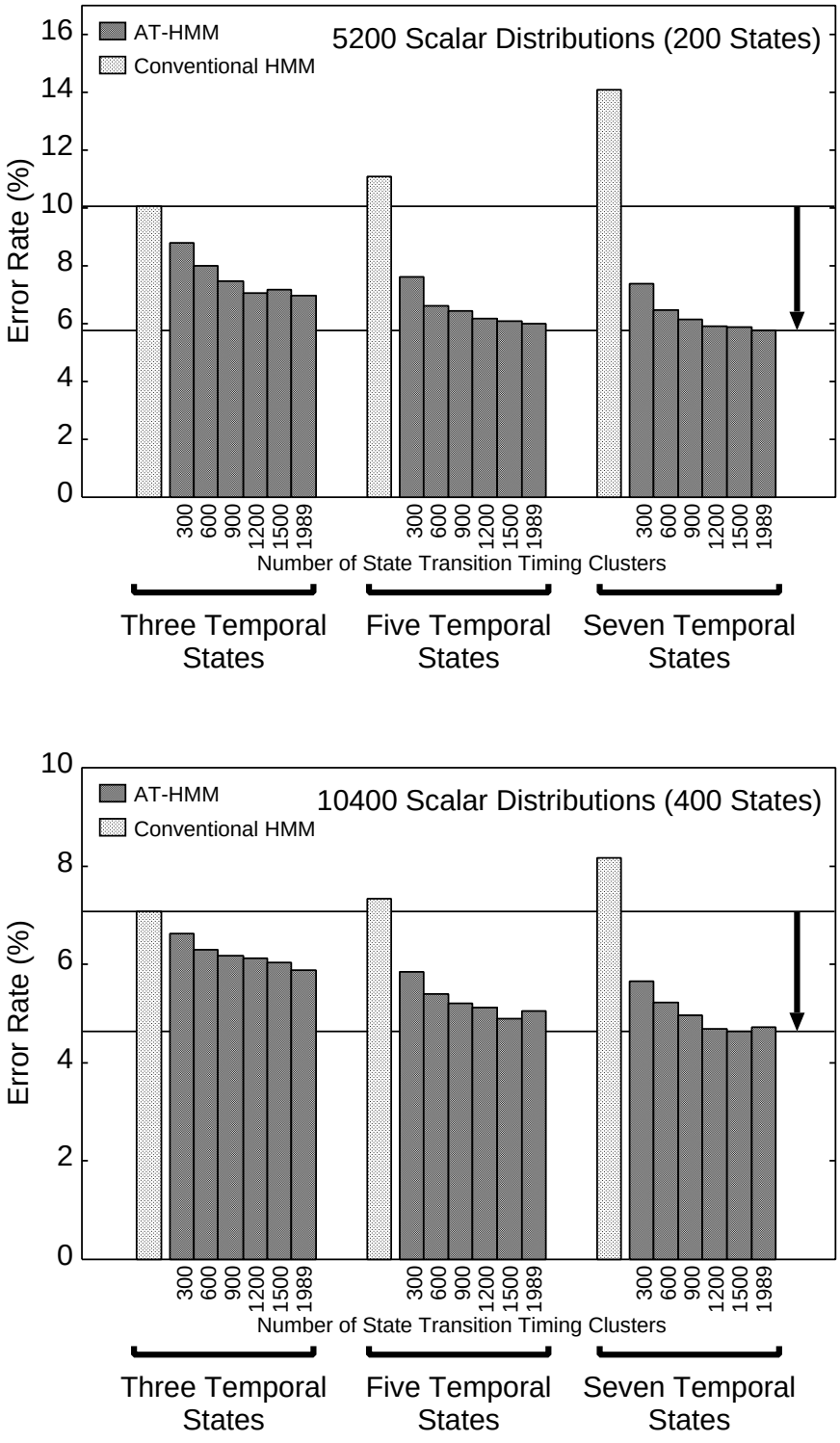


図 3.12: AT-HMM と従来型 HMM における連続音素認識実験の音素誤り率 (1)

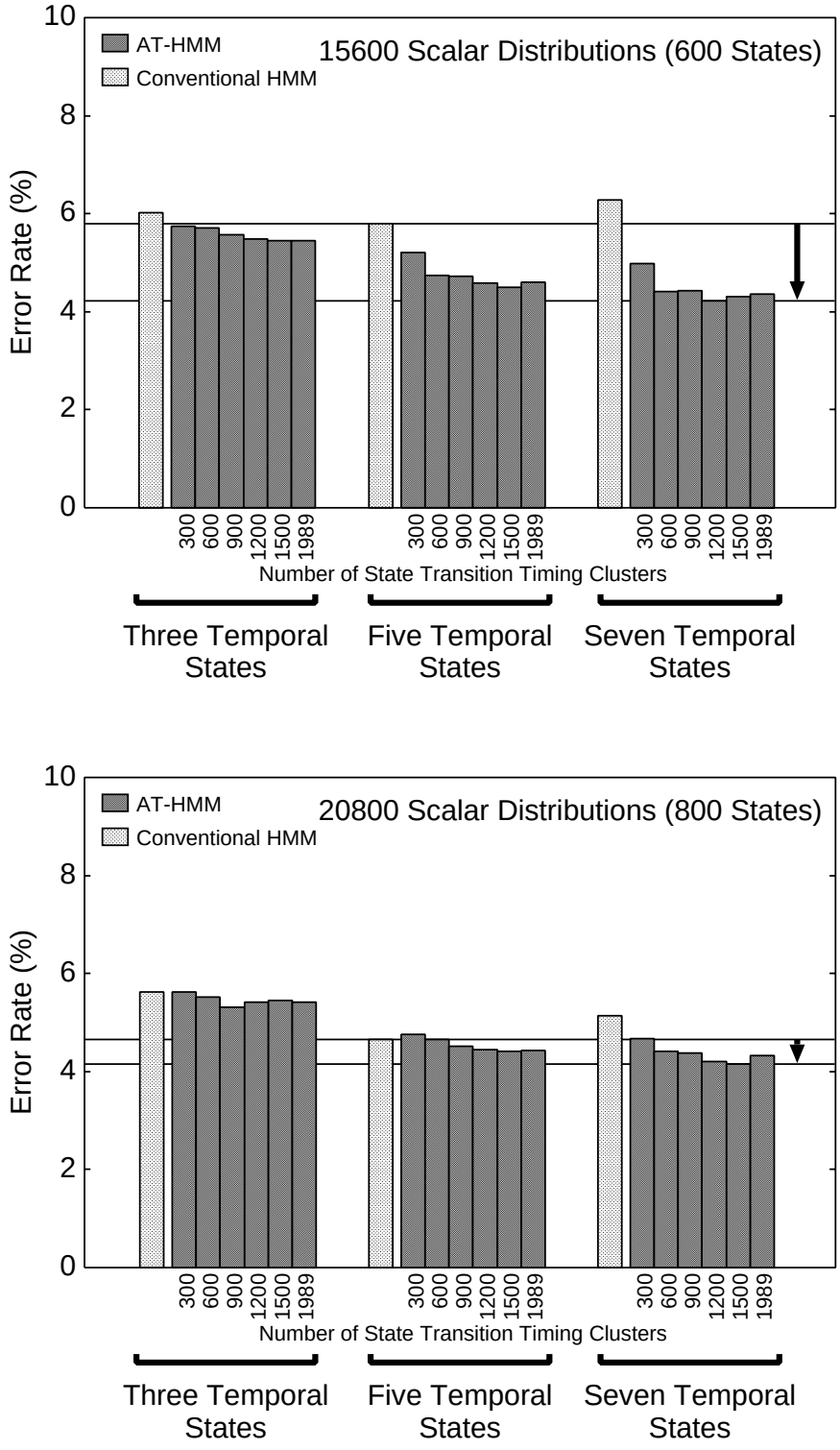


図 3.13: AT-HMM と従来型 HMM における連続音素認識実験の音素誤り率 (2)

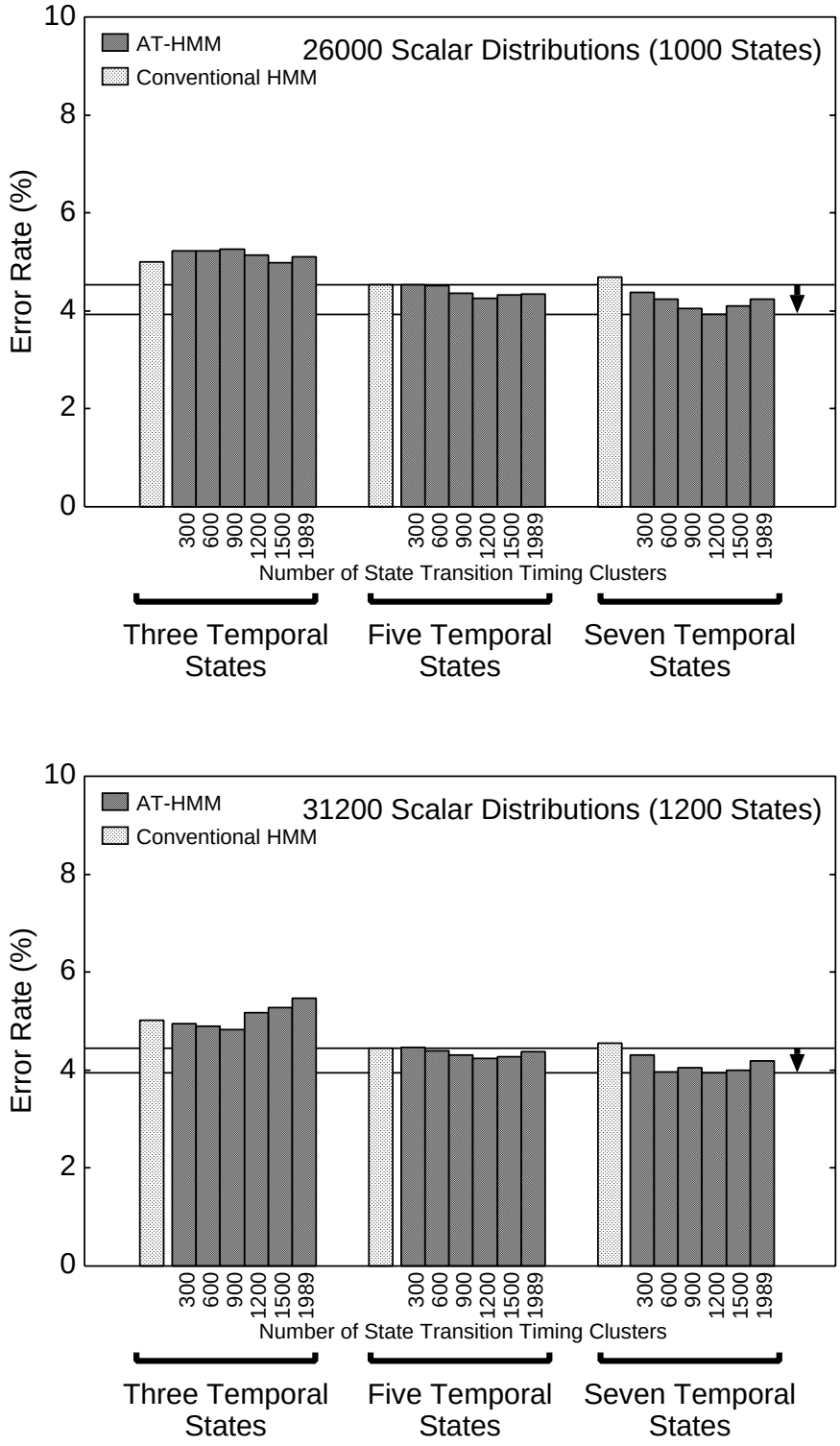


図 3.14: AT-HMM と従来型 HMM における連続音素認識実験の音素誤り率 (3)

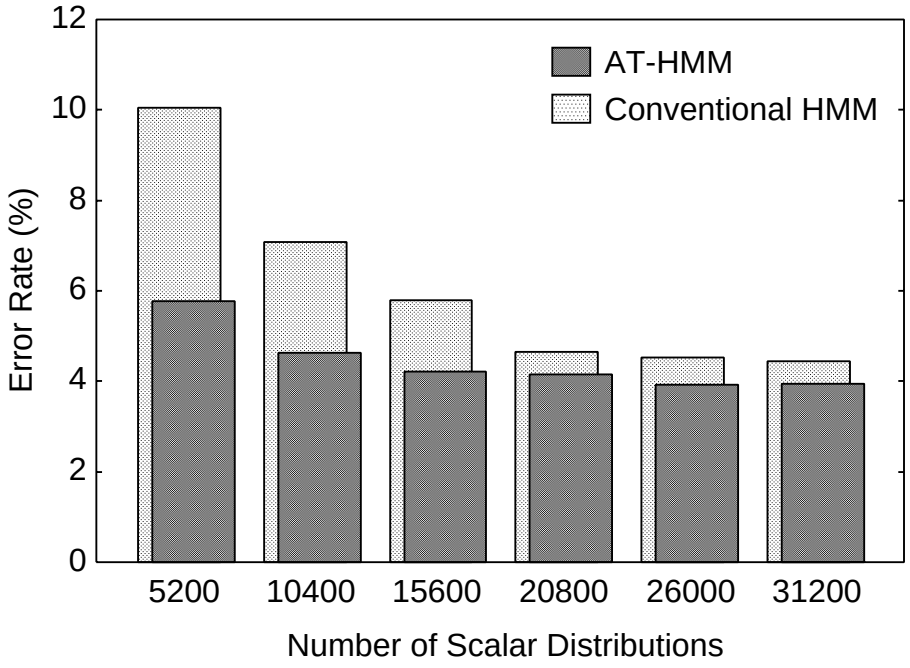


図 3.15: 各々のスカラー分布数のモデルにおける最小音素誤り率

話者 mht，800 状態の状態共有構造から生成した時間方向状態数 7，時間方向共有構造数 1500 の AT-HMM を用い，単語「赤」/aka/に対して個々の特徴量の状態タイミングを Viterbi アルゴリズムにより計算した結果を，第 1 次から第 3 次までの MFCC 特徴量及び，ΔMFCC 特徴量について図 3.16 に示す．本論文で提案した AT-HMM 生成法は，環境依存音素を単位として時間非同期化を行っているため，音素境界では同期した状態遷移として推定されている．またモデル内では，個々の特徴量はお互いに非同期なタイミングの状態遷移として推定されている．この AT-HMM は，時間方向状態数 7 の時間方向共有構造により実現されているため，順序制約を持った状態遷移となっている．従って，第 2.1 節で述べた時間非同期遷移構造の分類において，(B-2) のモデル同期と (B-3) 遷移部分共有，及び (III) 順序制約付きの複合した構造が実現されていることが分かる．

本実験より，時間方向状態数を増加させた従来型 HMM よりも，個々の特徴量の非同期性に着目した AT-HMM の方が，音響特徴ベクトル時系列を頑健かつ詳細にモデル化できることが確認された．更に，時間非同期遷移構造が音素環境に依存していることが確認された．

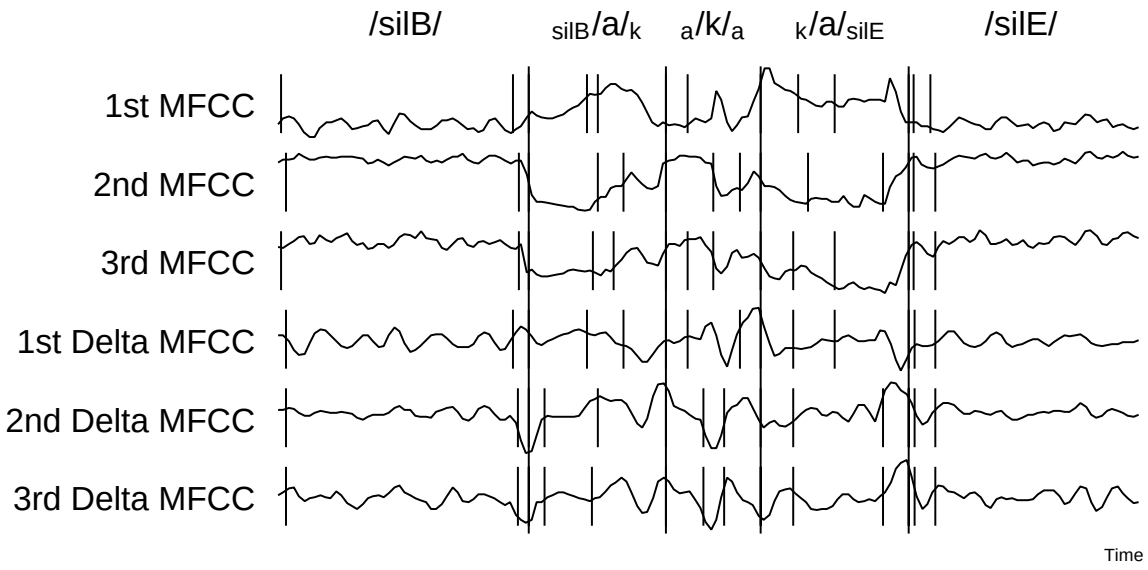


図 3.16: 順序制約付き AT-HMM による単語「赤」/aka/に対する Viterbi セグメンテーション結果

3.6.3 スカラー分布の消滅量

スカラー分布数 31200，時間方向状態数 3 の AT-HMM においては，時間方向共有構造クラスタ数の増加に従って認識率が従来型 HMM よりも低下している．その原因として，複雑な時間非同期遷移構造を持つ音響特徴ベクトル時系列は，少ない時間方向状態数を持つ AT-HMM では十分に表現することができない可能性がある．本論文で提案した AT-HMM 生成法は，状態遷移タイミング $\{l_{i,d}\}$ のクラスタリングにより消滅するスカラー分布が存在する．図 3.9 は，第 2 特徴量の第 2 状態のスカラー分布がクラスタリングにより消滅する場合の模式的な例である．極端に少ない時間方向状態数を持つ AT-HMM を生成することで，このようなスカラー分布の消滅が頻繁に発生している可能性がある．

スカラー分布の消滅量を調べるため，AT-HMM の生成に使用したスカラー HMM に含まれているスカラー分布数に対する，クラスタリングにより消滅したスカラー分布の割合を図 3.17 に示す．この割合は，各話者平均の割合である．また，AT-HMM の時間方向共有構造のクラスタ数は 1989 である．

図のように，時間方向状態数が 3 のモデルは，時間方向状態数 5 と 7 と比較してより多くの分布の消滅が発生しており，少ない時間方向状態数の AT-HMM では認識に必要な情報の欠損が発生している可能性がある．しかし，時間方向状態数を多くするに従って，分

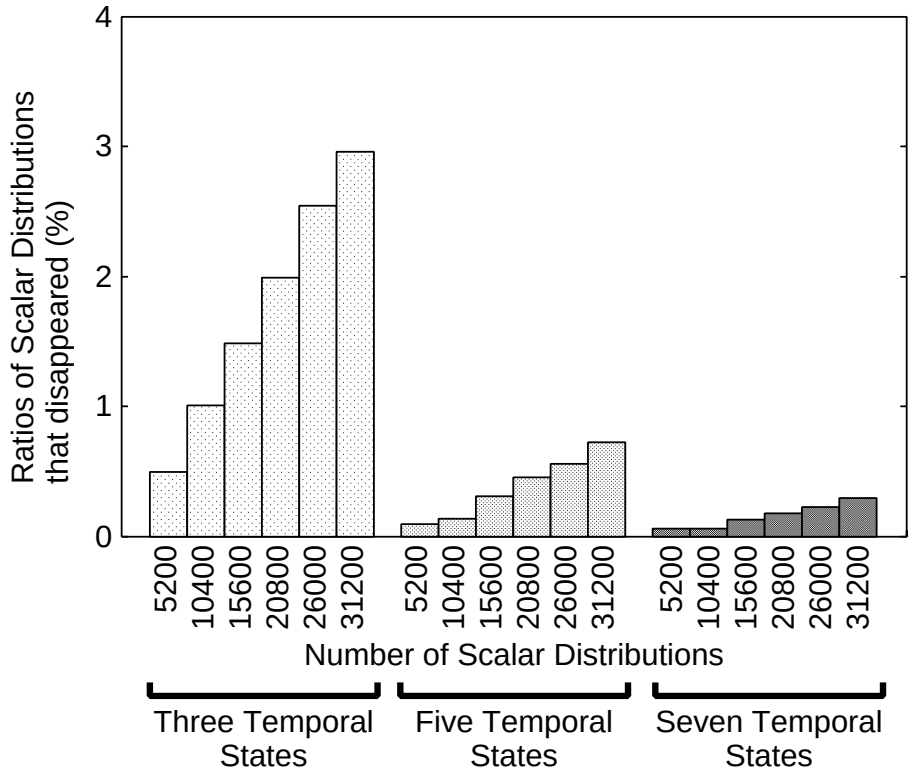


図 3.17: 消滅したスカラー分布の割合

布消滅の発生が減少しており，適当な時間方向共有構造を持つモデルを用いることにより，この問題を回避することができると考えられる．

3.6.4 音素セグメンテーション能力の評価

本論文で提案した AT-HMM は，個々の特徴量間の精密な時間非同期遷移構造を持つモデルである．そのため，各々の音素の境界などを従来型 HMM より，精密に計算することができると考えられる．従来型 HMM と AT-HMM の音素セグメンテーション能力の違いを調べため，評価データに対して Viterbi アルゴリズムにより得られる音素境界と，視察ラベル情報の間の平均誤差を計算した．使用したモデルは，各々のスカラー分布のモデル中で最小の音素誤り率が得られたモデルである．話者は mht である．図 3.18 に，Viterbi セグメンテーションにより得られた音素境界と視察ラベル間の平均誤差を示す．図のように，従来型 HMM の約 14ms の平均誤差が，AT-HMM を用いることにより約 11ms まで低減されている．

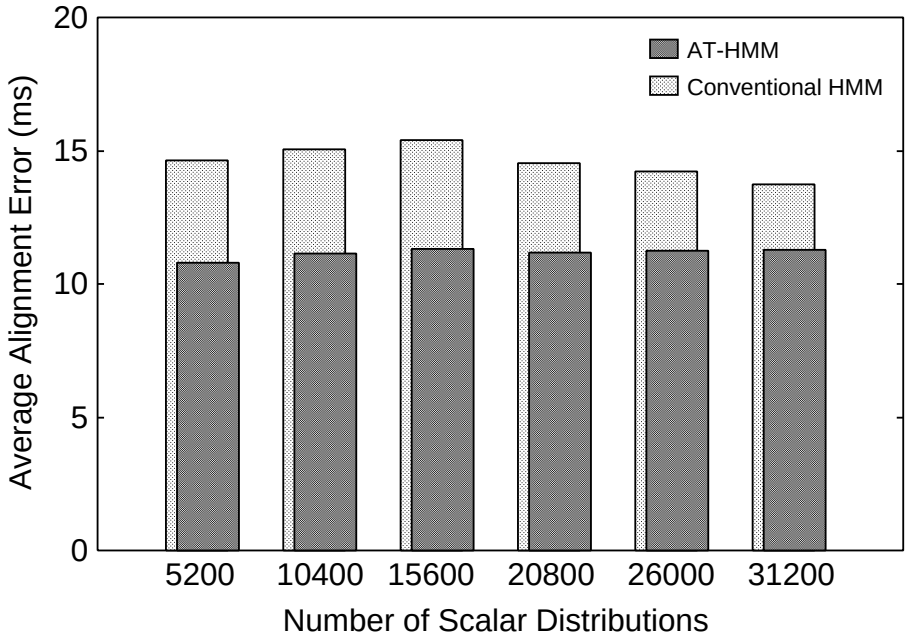


図 3.18: Viterbi セグメンテーションにより計算された音素境界と視察ラベル情報の間の平均誤差

この実験から，AT-HMM が従来型 HMM よりも高い音素セグメンテーション能力を持つことを確認した．

3.7 不特定話者連続音素認識実験

特定話者環境の実験では，順序制約付き AT-HMM は従来型 HMM よりも高い音声認識性能が得られた．本節では，不特定話者環境における順序制約付き AT-HMM の認識実験を行なうことにより，その音声認識性能の評価を行なう．もし，時間非同期遷移構造が話者に独立ならば，本議論で提案しているような単一の時間非同期遷移構造を表す AT-HMM でも音声認識性能が改善すると考えられる．逆に，時間非同期遷移構造が話者に依存している場合，その音声認識性能は従来型 HMM へ近付くと考えられる．

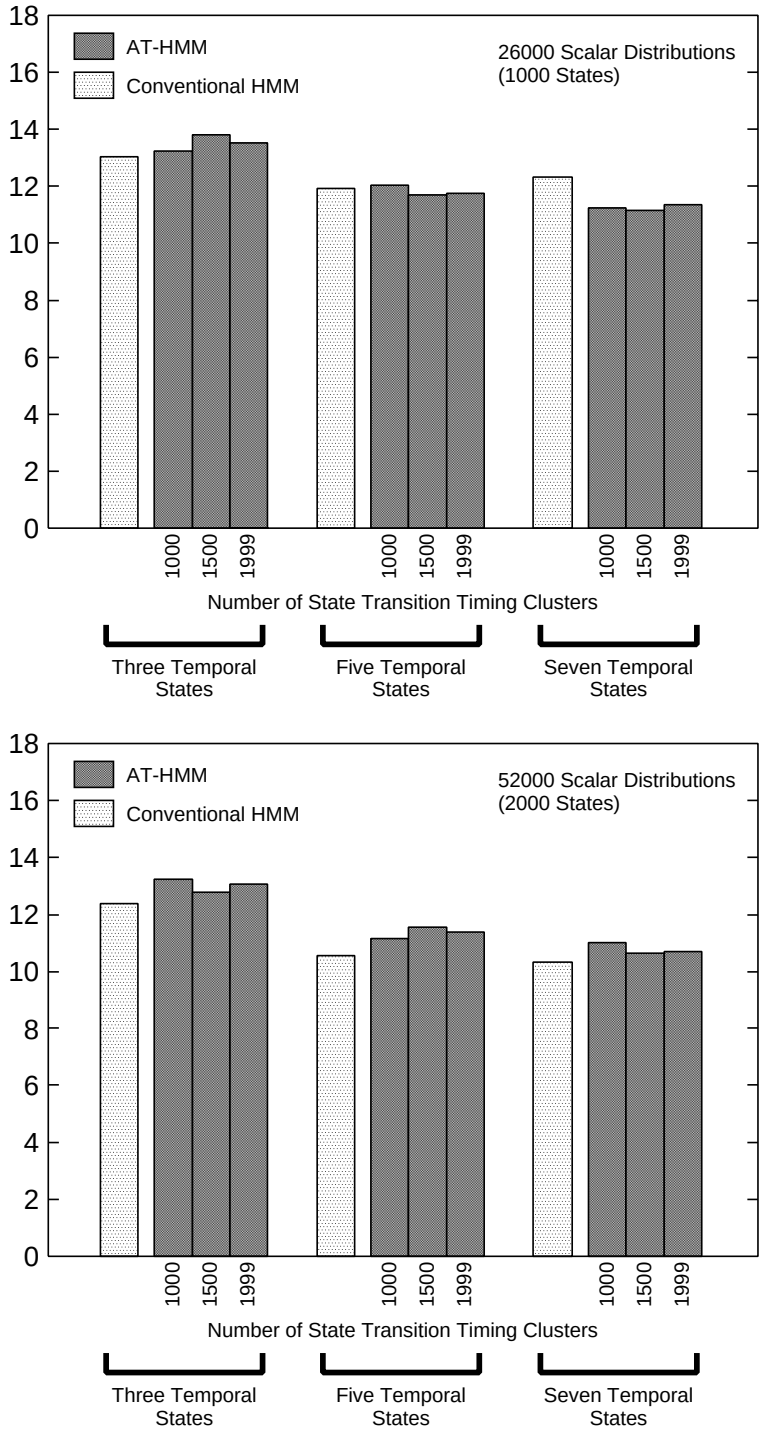


図 3.19: AT-HMM と従来型 HMM の不特定話者連続音素認識実験結果 1

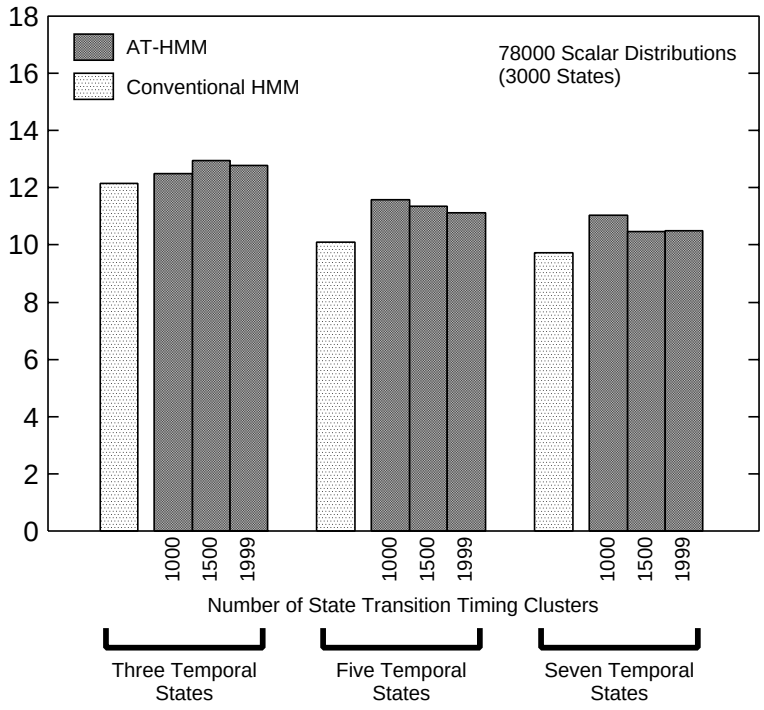


図 3.20: AT-HMM と従来型 HMM の不特定話者連続音素認識実験結果 2

3.7.1 実験条件

不特定話者の日本語音素接続制約付きの連続音素認識実験により評価を行なう．不特定話者用 AT-HMM を生成する際に用いた個々の特徴量のスカラー HMM の状態共有構造は，(C) 完全同期の従来型 HMM の状態共有構造を HTK[7] により音素分類木を基礎とした状態のトップダウンクラスタリングを用いて生成し，個々の特徴量の状態共有構造として用いた．そのため，全ての特徴の状態共有構造は同一である．スカラー HMM の時間方向状態数は 3，状態数は，1000 状態（AT-HMM のスカラー分布数は 26000），2000 状態（AT-HMM のスカラー分布数は 52000），3000 状態（AT-HMM のスカラー分布数は 78000），時間方向共有構造のクラスタ数 M は，500 から 1500 まで 500 毎に 4 種類である．比較実験として，スカラー分布数の等しい従来型 HMM（1000 状態から 3000 状態のモデル）について認識を行った．

学習データには，ATR 研究用日本語音声データベース set-A 中，男性 6 話者（mau, mht, mms, mmy, mnm, msh）の重要単語 5240 単語中の奇数番目と音素バランス単語 216 単語（計 2836 単語 $\times$6 話者）を使用し，評価データには男性 4 話者（mtk, mtm, mtt, mxm）の重

要語 5240 単語中偶数番目の 8 分の 1（計 327 単語 ×4 話者）を使用した．音声波形は，特定話者条件における実験と同様の方法で分析を行い，最後に単語音声単位の CMN（Cepstrum Mean Normalization）により正規化を行った．

3.7.2 AT-HMM の不特定話者連続音素認識性能

AT-HMM の時間方向状態数を 3, 5, 7 とした時の，種々のスカラー分布数を持つモデルの音素誤り率を図 3.19 と図 3.20 に示す．これらの音素誤り率は，4 種類の時間方向共有構造のクラスタ数のモデル中で最も小さな音素誤り率の値である．評価に用いたモデル全体で最も小さな音素誤り率が得られた AT-HMM は，スカラー分布数 52000，時間方向状態 7 の 10.0%であるのに対して，スカラー分布数 78000，時間方向状態数 7 の従来型 HMM が音素誤り率 9.7%が得られており，AT-HMM の認識率は従来型 HMM と同程度，あるいは若干低くなっている．

この原因を調べるため，図 3.21 に特定話者と不特定話者条件における個々の特徴量の振舞いをモデル化したスカラー HMM の状態遷移タイミングを示す．図のように，特定話者条件において個々の特徴量はお互いに異なったタイミングで状態が遷移しているのに対し，不特定話者条件では比較的状態遷移タイミングが揃っている．このことから，特徴量間の非同期性の構造は話者に依存しており，特定話者の場合には安定してモデル化することができ音声認識性能の改善に有効であったのに対し，不特定話者の場合は，話者間で共通な非同期構造は必ずしも有効でないことが一因であり，不特定話者の場合には，共通な構造の AT-HMM は従来の HMM に比べて高い性能が得られなかったと考えられる．

本実験より，時間非同期遷移構造が話者に依存していることを確認した．今後の展開として，複数の時間非同期遷移構造を持つ音響特徴ベクトル時系列に対して，複数の時間非同期遷移構造を持つモデルの研究を行うことにより，不特定話者の音声認識へ発展させることができると考えられる．

3.8 本議論のまとめ

本章では，個別特徴量の非同期性のモデル化について議論した．その中で，個々の特徴量やストリーム間の時間非同期遷移構造には種々のタイプが存在することを述べた．また，MFCC とその時間微分特徴量から構成された音響特徴ベクトルの時系列のモデル化におい

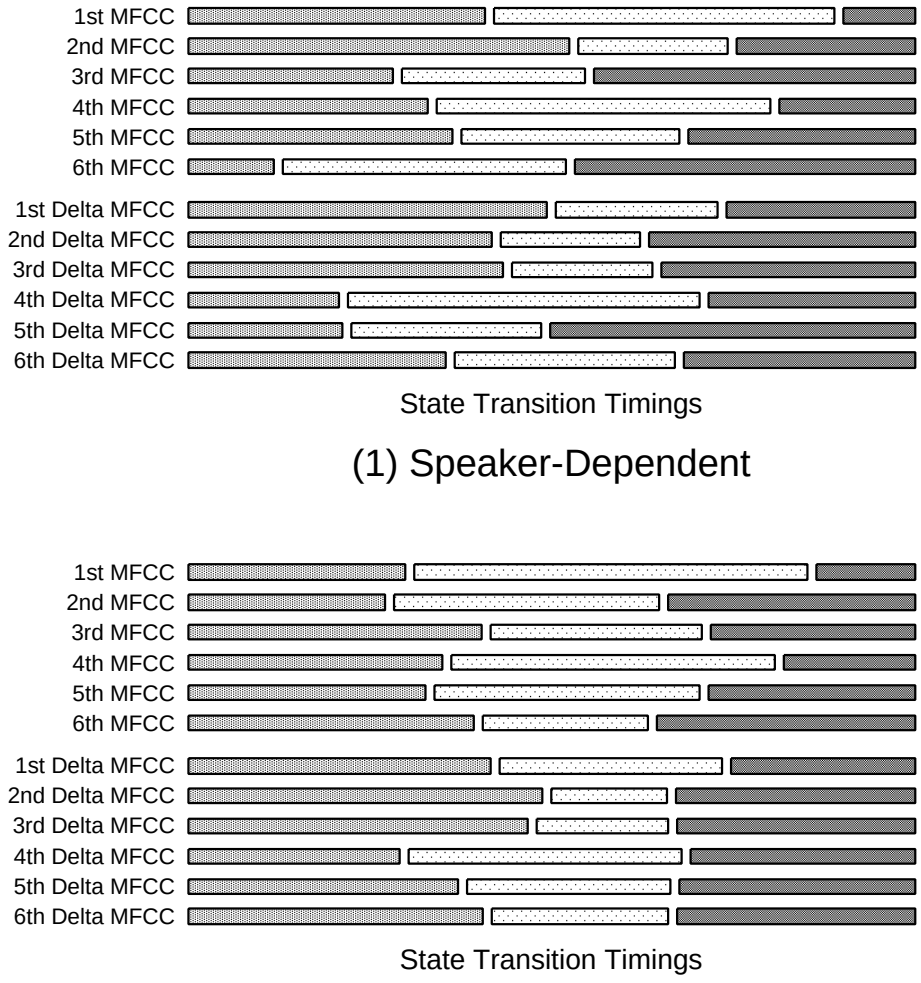


図 3.21: 特定話者と不特定話者条件における，環境依存音素 ₐ/k/ₐ の AT-HMM の生成に使用した個々のスカラー HMM の状態遷移タイミング

て，個々の特徴量の状態遷移に順序の制約を持つ時間非同期遷移構造が有効であることを確認した．更に，順序制約を持つ時間非同期遷移構造は，特定話者及び音素環境に依存していることを確認した．個々の特徴量の時間非同期性を積極的に考慮したモデル化を行なうことにより，特定話者の音素環境依存モデルにおいて，特徴ベクトル時系列の時間的な振舞いを少ないパラメータで詳細にモデル化できることを確認した．

今後の展開としては，不特定話者など，複数の順序制約付き時間非同期遷移構造を持つと考えられる特徴ベクトル時系列のモデル化には，複数の時間非同期遷移構造を持つモデルに関する研究を考えることができる．

第 4 章

個別特徴量の音素環境依存性のモデル化

本章では，個々の特徴量の音素環境依存性に関する議論を行ない，個々の音響特徴量の音素環境依存性や複雑性がお互い異なる音響特徴ベクトル時系列の効果的なモデル化法について議論する．個々の特徴量やストリームに依存したパラメータ共有構造を生成することにより，少ないパラメータ数でより高い統計的信頼性を持つモデルを生成し，音声認識性能を改善させることである．

4.1 パラメータ共有構造の特徴量依存性

音声認識に用いられる音響特徴ベクトルは，MFCC や対数パワー，またそれらの時間微分特徴量などから構成されている．これらの個々の特徴量は，お互いに異なる振る舞いを示している．ケプストラムなどの音響特徴量は，高次よりも低次に多くの音声情報を含んでおり，個々の特徴量をモデル化するために必要なパラメータ数を最適化することにより，少ないパラメータ数で効果的に音声信号をモデル化することができると考えられる．個々の特徴量をモデル化するために必要と考えられる分布数を調べるため，図 4.1 に，1000 状態 8 混合の HMM における，個々の特徴量の平均分布間距離を示す．分布間距離は，Kullback divergence により計算した．この分布間平均距離が大きい程，多様な分布が必要であることを表すと考えられる．図のように，低次の MFCC パラメータは，高時と較べて平均分布間距離が大きく，低次を表現するためには高次元よりも多くの分布が必要と考えられる．また，ΔMFCC よりも MFCC の方が平均分布間距離が大きく，MFCC の方がより多くの分布を必要としていると考えられる．また，音声認識における重要な特徴量であるフォルマントでは，音素環境依存性が個々の特徴量で異なっており，このような複雑性の違いだ

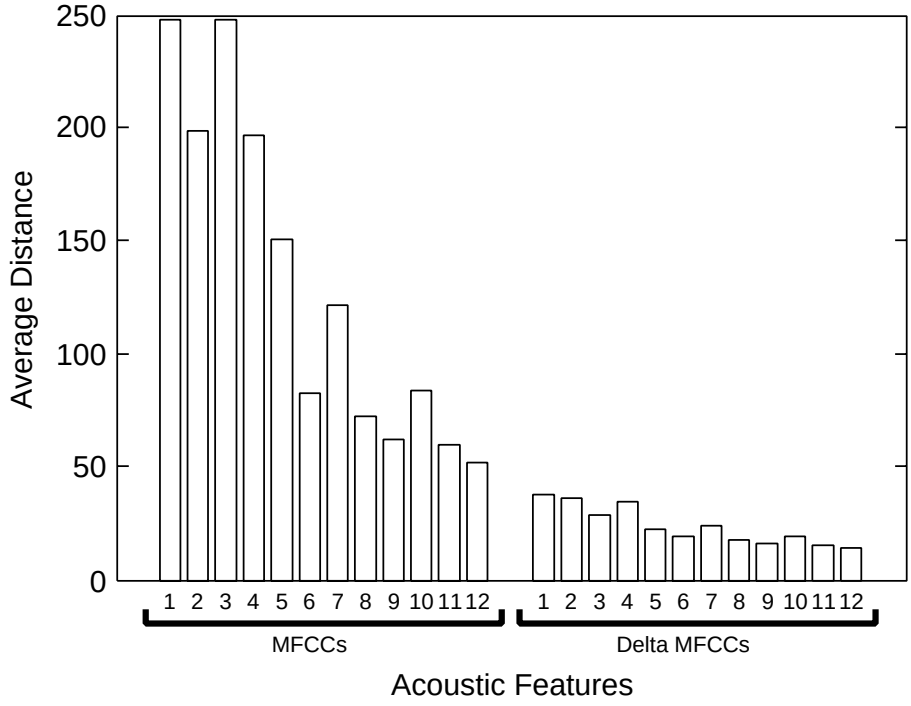


図 4.1: 1000 状態 8 混合の HMM における個々の特徴量の分布間平均距離

けでなく，個々の音響特徴量の音素環境依存性も異なっていると考えられる．個々の特徴量毎に，最適な音素環境依存性を考慮したパラメータ共有構造とパラメータ数を割り当てることにより，より頑健に音響特徴ベクトル時系列をモデル化することができると考えられる．

4.2 特徴量依存音素環境クラスタリング

本節では，個々の特徴量毎に音素環境クラスタリング（Phoneme Environment Clustering: PEC）と呼ぶ効率的な異音共有化手法を基礎とした手法により，個々の特徴に依存したパラメータ構造を自動的に生成する手法として，特徴量依存音素環境クラスタリング法（Feature-Dependent Phoneme Environment Clustering: FD-PEC）を提案する．

まず，FD-PEC の基礎となる手法である PEC について説明した後，個々の特徴量に依存した異音共有構造として特徴量依存音素環境クラスタ構造と，FD-PEC 法について各々述べる．

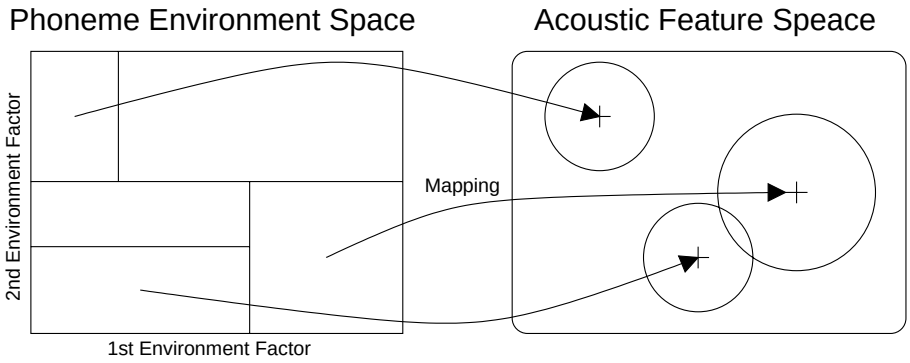


図 4.2: 音素環境空間と音響特徴量空間の間の写像関係の概念図

4.2.1 音素環境クラスタリング

音声認識では，音素や音節を認識単位とした音響モデルが用いられている．個々の音素は，前後の音素環境や，話者，雑音環境などの影響により，その振舞いは大きく変形させられている．高精度な音声認識システムを構成するためには，前後の環境要因を考慮した環境依存音素モデルを用いることが必須である．音素環境クラスタリング（PEC）[32] は，環境依存音素モデルを効率的に生成するための手法の一つである．

PEC は，種々の環境要因の直積で表された「音素環境空間」と音響特徴量空間などのパターン空間の間の写像関係を，音素環境空間と音響特徴量空間の分割により決定する手法である．図 4.2 に，音素環境空間と音響特徴量空間の写像関係の概念図を示す．この図のように，音素環境空間を構成している個々の部分音素環境空間は，音響特徴量空間上の分布への写像関係が定義されている．

音素環境空間として，先行音素環境，当該音素環境，後続音素環境の 3 つ組の例について述べる．先行音素環境の音素集合が {a,u,o}，当該音素環境が {k}，後続音素環境が {i,u} の場合，音素環境空間は {a,u,o}×{k}×{i,u} もしくは，$_{\mathrm{a,u,o}}/\mathrm{k}/_{\mathrm{i,u}}$ として記述され，$3\times1\times2=6$ 個の環境依存音素が含まれている．

次に，尤度を基準とした音素環境クラスタリングのアルゴリズムの流れを下記に示す．

ステップ 1

環境空間内の全ての音素環境クラスタに対して，要因 J（先行音素や当該音素，後続音素など）に関して要因要素配分 A, B で部分空間へ分割した時に，分割されたモデルの尤度の積が最大となる分割 $(J,(A,B))$ を探索する．式 (4.1) は，最大分割の探索を表

している．式中の $L(O|\lambda)$ は，モデル λ に対する学習データ O の対数尤度，(A,B) は，環境要因 J により分割された音素環境クラスタである．分割元音素環境クラスタが ${}_{a,u,o}/\mathrm{k}/_{i,u}$ を先行音素環境で分割した場合，(A,B) の組合せは，$1)A=\{a\},B=\{u,o\}$，$2)A=\{a,u\},B=\{o\}$，$3)A=\{a,o\},B=\{u\}$ の 3 種類である．O_A と O_B は，音素環境クラスタ A と B の学習データの集合である．$(\hat{A,B})$ は最大尤度の得られる分割である．

$$(\hat{A,B}) = \underset{all(A,B)}{\operatorname{argmax}} L(O_A|\lambda_A) + L(O_B|\lambda_B) \tag{4.1}$$

ステップ 2

分割前のモデルの尤度と分割後のモデルの尤度の差を分割ゲインと考え，最大の分割ゲインの得られる音素環境クラスタを分割し，分割後のモデルを再推定する．式 (4.2) に分割ゲイン G の計算式を示す．

$$G = (L(O_A|\lambda_A) + L(O_B|\lambda_B)) - L(O_{A\cup B}|\lambda) \tag{4.2}$$

ステップ 3

音素環境クラスタ数が所望の数になるまで，ステップ 1 とステップ 2 の処理を繰り返す．

このような処理により，部分音素環境空間を，先行音素，当該音素，後続音素のいずれかの環境を逐次的に分割することにより，音響特徴量空間上の分布が分割され，モデルが除々に精密化されると考えられる．

4.2.2 特徴量依存音素環境クラスタ構造

個々の特徴量に依存したパラメータ共有構造を持つ新しい共有構造として，特徴量依存音素環境クラスタ構造を提案する．図 4.3 に，特徴量依存音素環境クラスタ構造の概念図を示す．図のように，第 1 特徴量は 6 つ，第 2 特徴量は 5 つ，第 8 特徴量は 4 つなどのように，特徴量毎にお互いに異なるクラスタ数（パラメータ数）を持つ．また，個々の特徴量に依存した音素環境クラスタを持つ．図の例では，環境依存音素 ${}_{a}/\mathrm{k}/_{a}$ と ${}_{a}/\mathrm{k}/_{i}$ は，第 1 と第 2 特徴量はお互いに異なるクラスタであるが，第 8 特徴量では共有化されている．最適

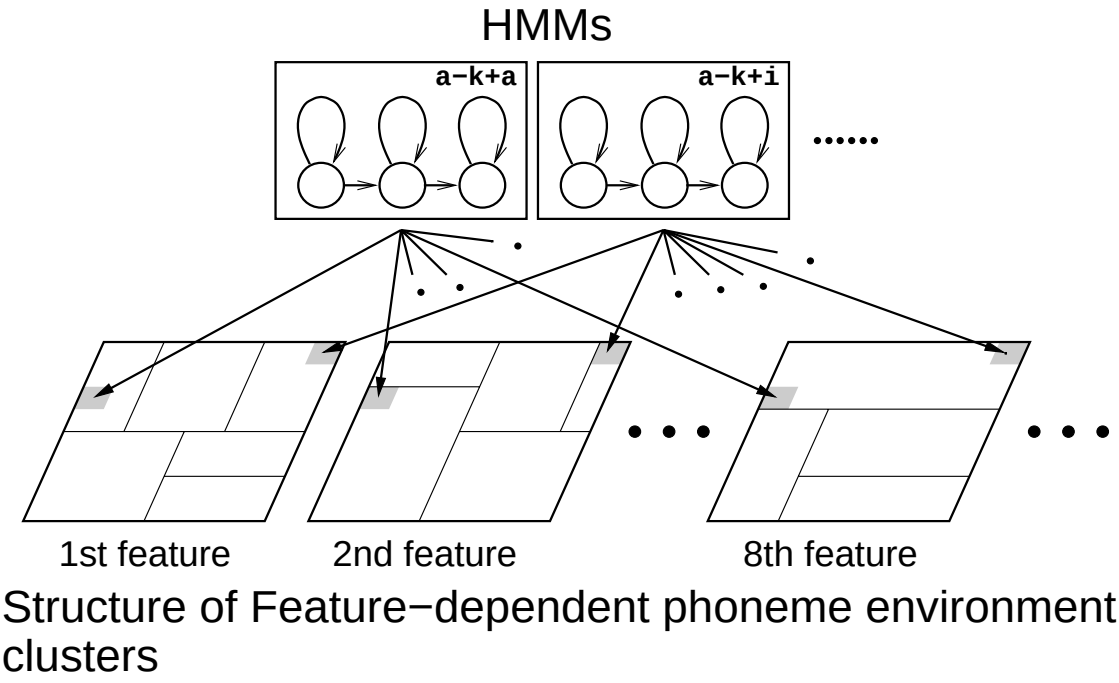


図 4.3: 特徴量依存音素環境クラスタ構造の概念図

な特徴量依存音素環境クラスタ構造を決定することにより，音響特徴ベクトル時系列がより少ないパラメータ数で頑健にモデル化されると考えられる．

4.2.3 特徴量依存音素環境クラスタリング

前章で述べた特徴量依存音素環境クラスタ構造を，自動的に生成するための手法として，特徴依存音素環境クラスタリング法（Feature-Dependent Phoneme Environment Clustering: FD-PEC）を提案する．この手法は，個々の特徴量毎に，尤度を基準とした音素環境クラスタリングを行い，個々の特徴量に依存した音素環境クラスタ構造を生成する手法である．

FD-PEC 法のアルゴリズムの流れを示す．

ステップ 1

個々の特徴量毎に，初期音素環境クラスタを決定する．

ステップ 2

個々の特徴量の音素環境クラスタ集合中で，最も分割ゲインの大きなクラスタを探索する．

ステップ 3

最大の分割ゲインを持つ音素環境クラスタを分割する．

ステップ 4

所望のクラスタ数になるまで，ステップ 2 とステップ 3 の処理を繰り返す．

以上の処理により，特徴量に依存した音素環境クラスタ構造が自動的に生成されると考えられる．このアルゴリズムは尤度を基準としており，分割によるゲインの増加量が小さい音素環境クラスタよりも大きな音素環境クラスタが分割される．従って，個々の特徴量をモデル化するために必要なパラメータ数を自動的に割り当てることができると考えられる．

4.3 特徴量依存逐次状態分割法

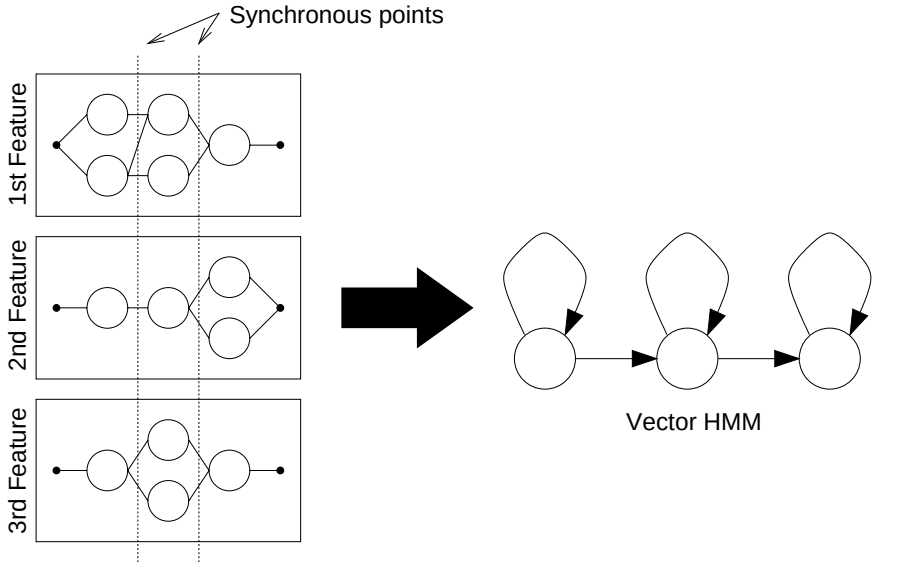
前節で提案した FD-PEC 法は，環境依存音素の共有構造を自動生成するためのアルゴリズムである．しかし，環境依存音素 HMM の個々の状態は，たとえ同一の環境依存音素のモデルであったとしても，お互いに異なる音素環境依存性を持つ．そこで本節では，FD-PEC 法を基礎とした，特徴量依存逐次状態分割法 Feature-Dependent Successive State Splitting: FD-SSS）を提案する．

4.3.1 特徴量依存逐次状態分割法

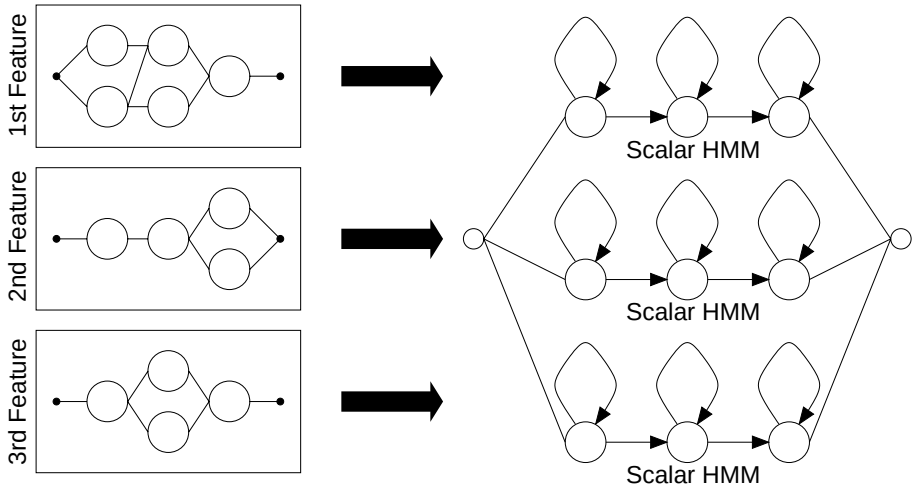
特徴量に依存したスカラー分布共有構造を自動的に生成するためのアルゴリズムの 1 つとして，特徴量依存逐次状態分割法（Feature-Dependent Successive State Splitting: FD-SSS）を提案する．

音響特徴ベクトルの個々の特徴量に対して，最適な状態数やスカラー分布共有構造を解析的に求めることは困難である．FD-SSS は，個々の特徴量に依存した 1 次元の特徴量依存隠れマルコフネットワーク（Feature-Dependent Hidden Markov Network: FD-HMnet）を用意し，全ての FD-HMnet 中の状態に対して最尤逐次状態分割法 (ML-SSS)[35] を基礎とした手法により，時間方向分割と環境方向分割を繰り返しながら，除々にモデルを精密化して行く手法である．ML-SSS 法の詳細は，付録 C に示す．

この FD-SSS 法の実装方法には，多次元ベクトル出力型 HMM を用いる手法と，スカラー出力型 HMM（1 次元 HMM）を用いる手法の，2 つの方法を考えることができる．



(A) Synchronous FD-SSS



(B) Asynchronous FD-SSS

図 4.4: 同期型 FD-SSS と非同期型 FD-SSS

同期型 FD-SSS

個々の特徴量の振舞いをモデル化した FD-HMnet と同じ状態共有構造を持つ，多次元ベクトル出力型 HMM を仮定した手法である．例として図 4.4-(A) に，同期型 FD-SSS によって得られる多次元ベクトル出力型 HMM の構造を示す.

非同期型 FD-SSS

個々の特徴量の振舞いをモデル化した 1 次元の FD-HMnet から得られる，順序制約無し AT-HMM を仮定した手法である．図 4.4-(B) に，非同期型 FD-SSS によって得られる順序制約無し AT-HMM の構造を示す.

本章では，実装の容易かつ AT-HMM への適用を考慮し，非同期型 FD-SSS 法について検証を行った.

4.3.2 非同期型 FD-SSS 法の処理の流れ

非同期型 FD-SSS 法の処理の具体的な流れを図 4.5，手順を次に示す.

ステップ 1

初期FD-HMNet を作成し全ての学習データを用いて学習する．例として，音素当たり 1 状態の FD-HMnet を用いる．音響特徴ベクトルが 26 次元の場合，26 個の FD-HMnet を用意する必要がある.

ステップ 2

全ての特徴量の全ての状態に対して，音素環境と時間方向の両方の分割ゲインを計算し，分割によって最大の分割ゲインが得られる状態を見付ける.

ステップ 3

状態の分割を行い，分割によって影響を受けた状態を再学習する.

ステップ 4

必要な状態数まで分割が進むまでステップ 2 と 3 を繰り返し行う.

以上の処理により，個々の特徴量に依存した FD-HMnet が自動的に生成される．FD-SSS 法により生成された FD-HMnet は，全ての特徴量に対して共通の HMnet を割り当てる従

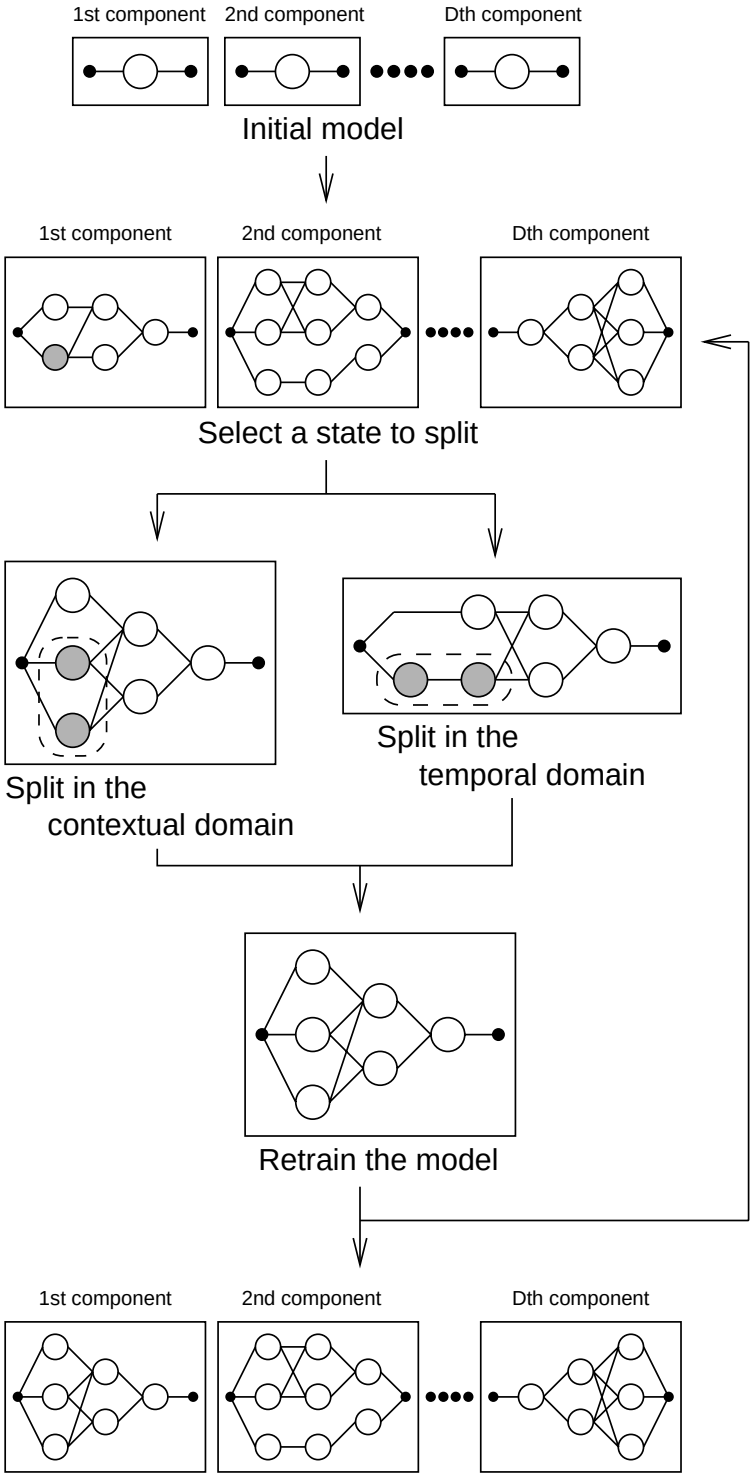


図 4.5: 特徴量依存逐次状態分割法の処理の流れ

来法よりも，音響特徴ベクトル時系列の振舞いを効果的にモデル化することができると考えられる．次節では，FD-SSS 法により生成された FD-HMnet の状態共有構造を持つ順序制約付き AT-HMM に対して，特定話者連続音素認識実験を行う．

4.4 特定話者連続音素認識実験

本節では，個々の特徴量に対して音素環境依存性を考慮したスカラー分布共有構造を持つ順序制約付き AT-HMM の音声認識性能の評価を行なう．全ての特徴量に対して共通のパラメータ共有構造を割り当てる従来の共有構造よりも，パラメータの使用効率の改善によって，より高い認識性能が得られると考えられる．

4.4.1 実験条件

FD-SSS により得られた FD-HMnet 構造を持つ HMM の音声認識性能を評価するため，日本語音素接続制約付きの特定話者連続音素認識実験を行なった．個々の特徴量の FD-HMnet の状態共有構造を持つ，順序制約付き AT-HMM の評価を行なった．また比較実験として，全ての特徴量に対して共通の状態共有構造を ML-SSS 法により生成し，その状態共有構造を持つ順序制約付き AT-HMM の評価を行なった．順序制約付き AT-HMM の時間方向状態数は 5, 7 である．時間非同期遷移構造数 M は 1200 のモデルを用いた．スカラー分布数は 5200（従来型 HMM で 200 状態）から 31200（従来型 HMM で 1200 状態）まで 5200 分布毎に 6 種類変化させた．

学習データには，ATR 研究用日本語音声データベース set-A 中，男性 2 話者（mht, mau）女性 2 話者（fms, ffs）の重要語 5240 単語中の奇数番目と音素バランス単語 216 単語（計 2836 単語 $\times$4 話者）を使用し，評価データには重要語 5240 単語中偶数番目の 4 分の 1（計 665 単語 $\times$4 話者）を使用した．音素ラベルは，付録 A の計 26 音素を使用した．サンプリング周波数 12kHz の波形データをフレーム長 25ms，フレーム周期 5ms，ハミング窓を掛けて分析した．特徴パラメータは，対数パワー，12 次 MFCC，Δ 対数パワー，12 次 ΔMFCC の計 26 次元を使用した．

4.4.2 実験結果

図 4.6 に，FD-SSS により生成した FD-HMnet 構造を持つ AT-HMM と，ML-SSS により生成した全ての特徴量に対して共通の HMnet 構造を持つ AT-HMM の音素誤り率を示す．図 4.6 の上段は，時間方向状態数 5 の順序制約付き AT-HMM，下段は時間方向状態数 7 の順序制約付き AT-HMM の音素誤り率である．

図に示すように，スカラー分布数 5200 から 15600 までは，FD-SSS により生成した AT-HMM の方が，ML-SSS により生成した AT-HMM よりも高い音素認識率が得られた．しかし，スカラー分布数 20800 以上では，ML-SSS により生成した AT-HMM の方が高い音素認識率が得られた．

FD-SSS は個々の特徴量に対して別々の HMnet が生成されるため，従来の全ての特徴量に対して共通の HMnet を割り当てる手法よりも詳細なモデルと考えることができる．従って，FD-SSS により生成された FD-HMnet 構造が学習データに過剰に最適化されてしまうことにより，いわゆる過学習が発生している可能性がある．過学習の問題が発生しているか調べため，学習データに対するモデルの尤度と評価データに対するモデル尤度を図 4.7 と図 4.8 に示す．

図のように，FD-SSS により生成されたパラメータ共有構造を持つ AT-HMM の学習データに対する尤度は，ML-SSS により生成されたそれよりも高い．従って，従来法よりも学習データをより詳細に表現していることがわかる．一方，評価データに対する尤度は，スカラー分布数 10400 を境界として，従来法の方が高い尤度が得られ，提案法のモデルの尤度は低下している．これは，学習データに過剰に適応してしまった結果と考えられる．

本実験により，比較的少ないパラメータ数を持つモデルで，個々の特徴量の音素環境依存性を考慮したクラスタリングの有効性を確認した．しかし，パラメータ数が増加するに従い，いわゆる過学習の問題により音声認識性能が低下することを確認した．

4.4.3 生成された FD-HMnet

参考のため，FD-SSS 法によって個々の特徴量に割り当てられた FD-HMnet の構造の例を付録 D に示す．付録に示すように，個々の特徴量毎に全く異なった状態共有構造が生成されている．付録 C の図 C.5 に示すような，個々の特徴量に対して共通の HMnet を割り当てた従来手法は，後続音素環境の分割が主なのに対して，個々の特徴量の FD-HMnet は，先行及び後続音素環境に対しても分割されていることが分かる．

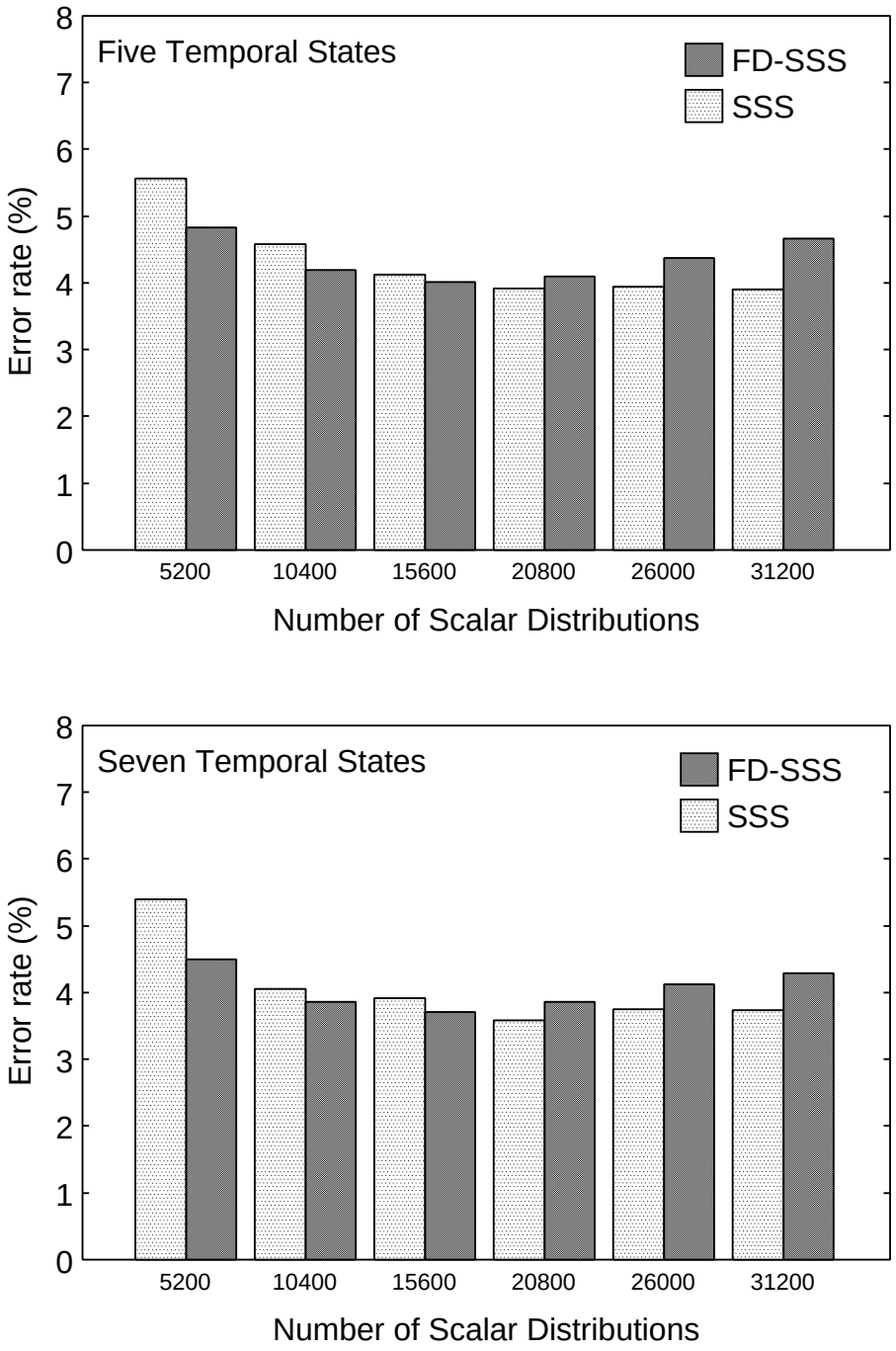


図 4.6: FD-SSS により生成した FD-HMnet 構造を持つ AT-HMM と，ML-SSS により生成した全ての特徴量に対して共通の HMnet 構造を持つ AT-HMM の音素誤り率

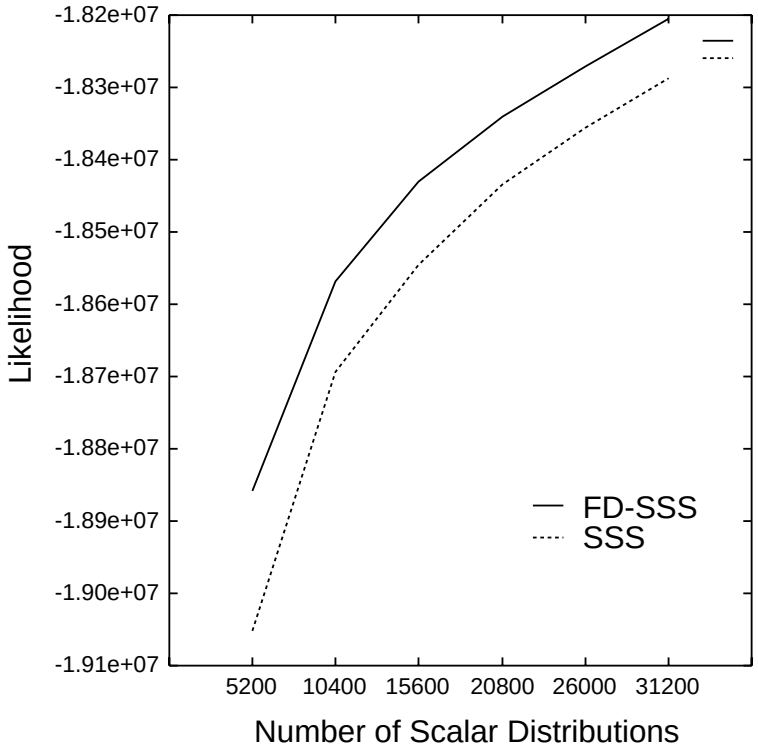


図 4.7: FD-SSS と ML-SSS により生成されたパラメータ共有構造を持つ AT-HMM の学習データに対する尤度

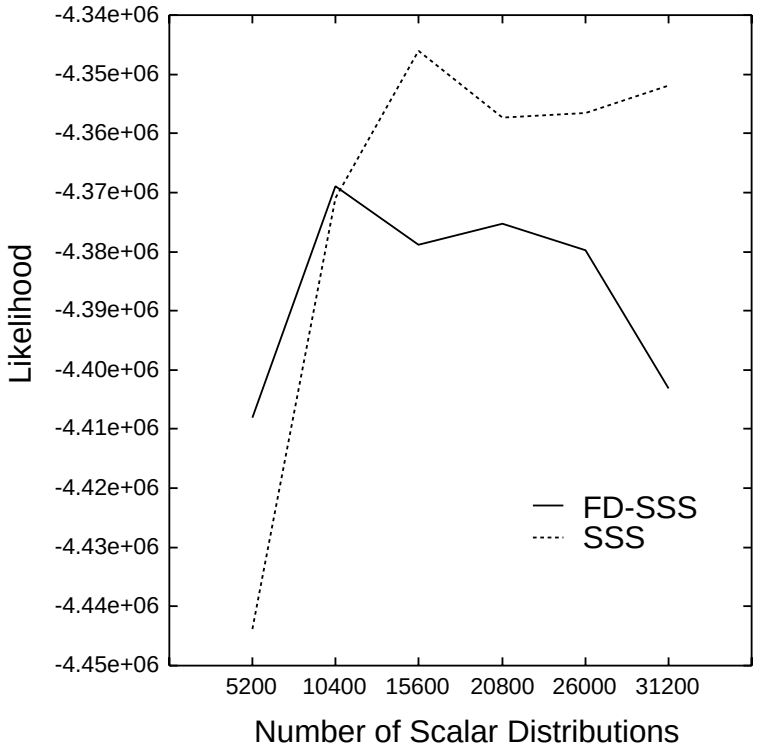


図 4.8: FD-SSS と ML-SSS により生成されたパラメータ共有構造を持つ AT-HMM の評価データに対する尤度

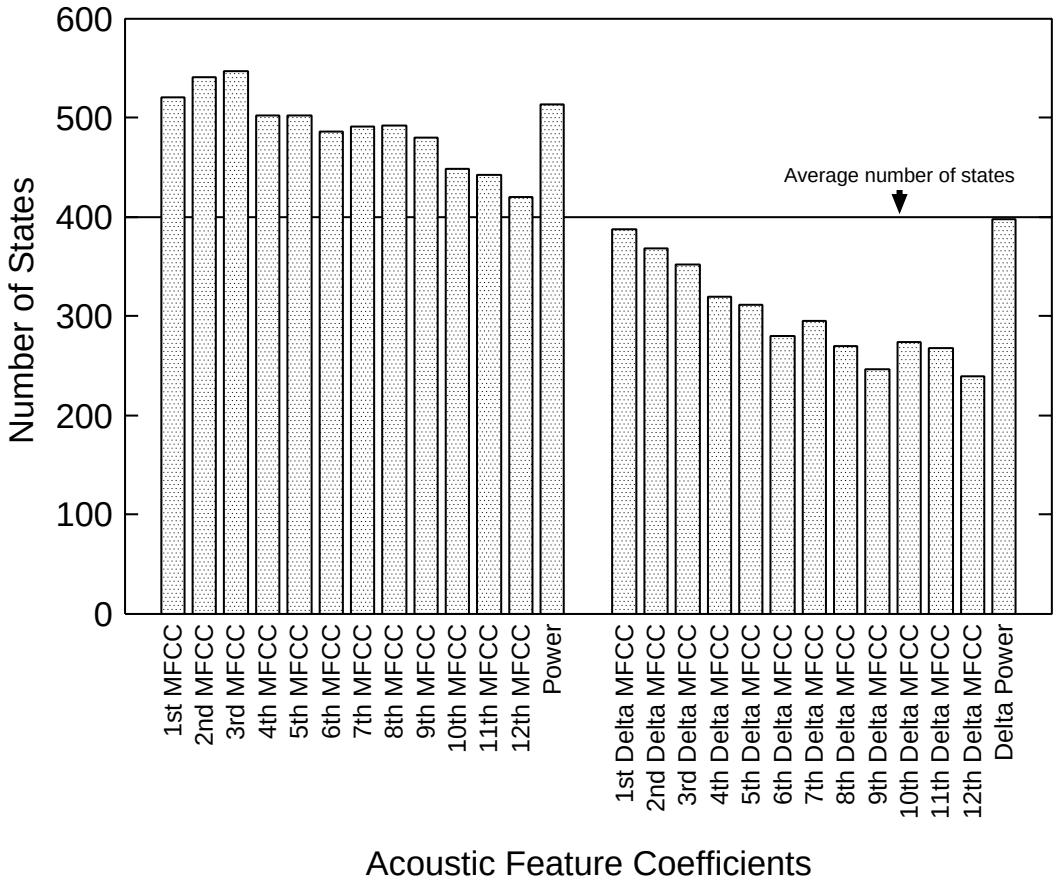


図 4.9: 個々の特徴量に割り当てられた状態数（スカラー分布数 10400）

図 4.9 に，スカラー分布数 10400 の FD-HMnet における個々の特徴量に割り当てられた状態数を示す．図のように，MFCC の高次よりも低次により多くの状態が割り当てられている．また，ΔMFCC の方が MFCC よりも少ない状態数が割り当てられており，図 4.1 の平均分布間距離と類似した状態数の割り当てが自動的に行われている．

4.5 まとめ

本章では，個別特徴量の音素環境依存性のモデル化に関する議論を行なった．音素環境依存性を考慮したトップダウンクラスタリングにより，個々の特徴量に依存したパラメータ共有構造を自動生成する手法を提案した．

音声認識実験を行った結果，比較的少ない総スカラー分布数のモデルにおいて，高い認識性能が得られることを確認した．しかし，総スカラー分布が増加するに従って，いわゆ

る過学習の問題により，音声認識性能の低下が見られた．

本議論では，音素環境依存性を考慮したパラメータ共有構造の生成手法において，全ての特徴量に対して共通の共有構造を割り当てる従来手法を改め，個々の特徴量に依存した共有構造を割り当てる手法を新たに提案した．本研究の意義は，このように音素環境依存性の概念を拡張した点である．しかし，原理的には，従来のベクトルとしての扱いと，本研究で提案した個別特徴量の間には，類似した特性を持つ個別特徴量のクラスタとして扱う手法が考えられる．本研究を踏まえ，このような個別特徴量のクラスタの概念へ拡張し，従来法と提案法を含むより一般化した手法の検討を行うことにより，更に音声認識性能を改善することができる可能性がある．

第 5 章

結論

5.1 本研究の要約

従来より，最も広く用いられている音響モデルである HMM に対する音声認識性能を改善させるため，様々な精密化や頑健化手法が提案されてきた．これらの改善手法の大半は，音声の観測量がベクトルであることを仮定している．しかし，音声認識に用いられる音響特徴ベクトルは，複数の音響特徴量から構成（第 1MFCC，第 2MFCC，また時間微分成分など）されており，個々の音響特徴量の振舞いは互いに異なる．個々の特徴量の特性を考慮したモデル化や，各特徴量の間の相関を考慮した統合により，更に効果的に学習データの振舞いがモデル化できると考えられる．本研究では，このような個々の特徴量の振舞いの効果的なモデル化において，次の 2 つの点について議論した．

5.1.1 個別特徴量の非同期性のモデル化

第 3 章では，個々の特徴量間で値の変化が非同期なタイミングで発生する特徴ベクトル時系列の効果的なモデル化法について議論した．

提案手法のまとめ

個々の特徴量間の時間非同期構造に着目した音響特徴ベクトル時系列のモデル化について検討を行なった．個々の特徴量の値の変化がお互いに非同期に変化する特徴ベクトル時系列を効果的にモデル化するための枠組として，非同期遷移型 HMM（Asynchronous Transition HMM: AT-HMM）と呼ぶ新しい HMM のクラスを提案した．また，時間非同期遷移構造

の分類を行ない，各特徴量間の状態遷移に順序関係を持つ AT-HMM の検討を主に行った．この順序制約を持つ AT-HMM は，本検討で新たに提案した時間方向共有法を用いることにより，従来型 HMM と同じ構造で実現することができるため，Viterbi アルゴリズムなどの高速な尤度計算法を使用することができる．過去の研究において，個々の（サブバンドなどの）ストリーム間の非同期性扱った検討では，非同期性のモデル化に直積型 HMM を用いていた．しかし，直積型 HMM による実現はストリーム数の増加は計算量の多大な増加を招くため，本検討で扱うような個々の音響特徴量間の非同期性を表現することは不可能である．また，時間方向共有構造により実現された順序制約付き AT-HMM の生成法を提案した．

評価実験のまとめ

AT-HMM の音声認識性能の評価を行なった．

大量の時間方向状態数を持つ順序制約付き AT-HMM の音声認識性能を評価するため，特定話者連続音素認識実験を行なった．その結果，時間方向状態数の増加に従い，音声認識性能の改善が得られた．時間方向状態数の増やすことにより，時間非同期遷移構造が精密化され，音声認識性能の改善に繋がったと考えられる．また，順序制約無し/付きの AT-HMM 各々の音声認識性能を評価を行なうため，特定話者切り出し音素認識実験を行なった．その結果，順序制約無し AT-HMM は従来型 HMM より低い音素認識率だったのに対して，順序制約付き AT-HMM は従来型 HMM より高い音素認識率が得られた．MFCC などの音響特徴ベクトル時系列のモデル化には，個々の特徴量の値の変化に順序制約を持つ時間非同期遷移構造が有効であると考えられる．

時間方向状態数を増加させた完全同期な従来型 HMM との比較及び，順序制約付き AT-HMM の生成法における，時間方向共有構造数 M に対する評価を行なうため，特定話者連続音素認識実験を行った．その結果，種々のスカラー分布数を持つモデル全てにおいて，従来型 HMM よりも高い認識率が得られ，誤り削減率は 10%から 40%となった．このことから，AT-HMM は完全同期な従来型 HMM よりも効果的に音声信号をモデル化することができたと考えられる．また，時間方向共有構造をクラスタリングすることにより，環境依存音素毎に別々に時間方向共有構造を決定したモデル（時間方向共有構造のクラスタ数 1989）よりも高い認識率が得られた．時間方向共有構造数 M を適当に設定することにより，サンプル数の少ない環境依存音素に対する時間非同期遷移構造の推定精度の改善効果

があったと考えられる．逆に，時間方向共有構造数 M を少なくするに従い，従来型 HMM の音素認識率へ近付いていった．これは，時間非同期遷移構造が環境依存音素に依存していることを示していると考えられる．

また，音素セグメンテーション能力の評価実験を行った結果，視察ラベルからの平均誤差が，従来型 HMM では約 14ms に対して AT-HMM では約 11ms まで改善された．一方，不特定話者条件では，AT-HMM の認識率は，従来型 HMM と同程度，あるいは若干低くなった．

本議論のまとめ

個別特徴量の非同期性のモデル化について議論を行った．その中で，個々の特徴量やストリーム間の時間非同期遷移構造には種々のタイプが存在することを述べた．また，MFCC とその時間微分特徴量から構成された音響特徴ベクトルの時系列のモデル化において，個々の特徴量の状態遷移に順序の制約を持つ時間非同期遷移構造が有効であることを確認した．更に，順序制約を持つ時間非同期遷移構造は，特定話者及び音素環境に依存していることを確認した．個々の特徴量の時間非同期性を積極的に考慮したモデル化を行なうことにより，特定話者の音素環境依存モデルにおいて，特徴ベクトル時系列の時間的な振舞いを少ないパラメータで詳細にモデル化できることを確認した．

5.1.2 個別特徴量の音素環境依存性のモデル化

本議論では，個々の特徴量の音素環境依存性に関する検討を行ない，個々の音響特徴量の音素環境依存性や複雑性がお互い異なる音響特徴ベクトル時系列の効果的なモデル化法について議論した．

提案手法のまとめ

個々の音響特徴量の音素環境依存性や複雑性がお互い異なる音響特徴ベクトル時系列の効果的なモデル化法について議論した．個々の特徴量毎にお互いに異なる音素環境依存性や複雑性を持つ特徴ベクトル時系列を効果的にモデル化するための手法として，依存音素環境クラスタリング法（FD-PEC）を提案した．また，FD-PEC 法の実現法として，特徴量依存逐次状態分割法（FD-SSS）を提案した．FD-SSS 法により，個々の特徴量に対して

状態単位のスカラー分布共有構造が自動的に生成される．

評価実験のまとめ

FD-SSS 法により得られた特徴量依存隠れマルコフネットワーク（FD-HMnet）と同一の状態共有構造を持つ順序制約付き AT-HMM に対して，特定話者連続音素認識実験を行った．その結果，スカラー分布数 5200 から 15600 までは，FD-SSS により生成した AT-HMM の方が，ML-SSS により生成した AT-HMM よりも高い音素認識率が得られた．しかし，スカラー分布数 20800 以上では，ML-SSS により生成した AT-HMM の方が高い音素認識率が得られた．この認識率低下の原因として，学習データへの過剰な適応がある．このような過学習が発生しているか調べるため，学習データに対する尤度と評価データに対する尤度を調査した．その結果，学習データに対する尤度は総スカラー分布数の増加に従って順調に増加しているのに対して，評価データに対する尤度は 10400 を境に提案法のモデルで尤度の低下が見られた．

本議論のまとめ

個別特徴量の音素環境依存性のモデル化に関する議論を行なった．本議論の意義は，全ての特徴量に対して共通の音素環境依存性を仮定していた従来の共有化手法を，個々の特徴量の音素環境依存性へ展開することにより，環境依存性の概念を拡張した点である．

このような，個々の特徴量の音素環境依存性を考慮した共有化手法を提案し，音声認識実験を行った．その結果，総スカラー分布数の増加に従って音声認識性能の低下が見られたが，比較的少ない総スカラー分布数のモデルにおいて，高い認識性能が得られることを確認した．大量のパラメータ数を持つモデルでの音声認識性能の低下は，いわゆる過学習によるのではないかと考えられる．

5.2 今後の展望

5.2.1 個々の特徴量の時間非同期性に対する今後の展望

今後の展開としては，不特定話者など，複数の順序制約付き時間非同期遷移構造を持つと考えられる特徴ベクトル時系列のモデル化には，複数の時間非同期遷移構造を持つモデ

ルに関する研究を考えることができる．そのための方法として，次に 3 つ述べる．

- AT-HMM の複数混合化

  AT-HMM を複数混合化することにより，複数の時間的な振舞いを表現することができると考えられる．同一の時間方向共有構造を並べた構造を持つ複数混合 AT-HMM だけでなく，お互いに異なる時間方向共有構造を並べた構造を持つ複数混合 AT-HMM を考えることができる．

- AT-HMM のマルチパス化

  複数の AT-HMM を平行に並べた構造を持つマルチパスモデル [54, 55, 56] による時間非同期遷移構造の複数化を考えることができる．このマルチパスモデルにより，種々の時間非同期遷移構造を表現するモデルを考えることができる．

- 条件付き AT-HMM

  話者クラスタリングを基礎とする手法により，予め話者を選択し，対応する AT-HMM で認識する手法を考えることができる．更に，このようなモデルを効果的に用いための手法として，時間非同期遷移構造についての話者クラスタリング法を考えることができる．

その他，AT-HMM の音声認識以外への応用として，個々の特徴量の状態遷移が必ずしも同期しない，オンライン手書き文字認識 [58] などへの応用を考えることができる．

5.2.2 個々の特徴量の音素環境依存性に対する今後の展望

本議論では，音素環境依存性を考慮したパラメータ共有構造の生成手法において，全ての特徴量に対して共通の共有構造を割り当てる従来手法を改め，個々の特徴量に依存した共有構造を割り当てる手法を新たに提案した．本研究の意義は，このように音素環境依存性の概念を拡張した点である．

しかし，原理的には，従来のベクトルとしての扱いと，本研究で提案した個別特徴量の間には，類似した特性を持つ個別特徴量のクラスタとして扱う手法が考えられる．今後の展望として，このような個別特徴量のクラスタの概念へ拡張し，従来法と提案法を含むより一般化した手法へ拡張することにより，更に音声認識性能を改善する手法が得られる可能性がある．

謝辞

本研究を行うにあたり，指導教官として大変有益な御助言と御指導を頂いた北陸先端科学技術大学院大学の下平 博助教授に心より感謝致します．東京大学の嵯峨山 茂樹教授には，北陸先端科学技術大学院大学在職中はもとより、東京大学に移られてからも，本研究全般について昼夜を問わず大変丁寧な御助言と御指導を頂きました．また更に，本研究の基礎となる音声認識技術一般についての知識や研究生活全般にわたって御指導下さいました．心より感謝致します．北陸先端科学技術大学院大学の中井 満助手には，研究室ゼミの時間などに限らず御助言を頂きました．ここに感謝いたします．下平研究室の皆様には，本研究に関する数多くの御討論や御意見を頂きました．ここに深く感謝致します．

聴覚分野からの貴重な御意見を頂きました，北陸先端科学技術大学院大学の赤木 正人教授に感謝致します．同じく北陸先端科学技術大学院大学の島津 明教授，小谷 一孔助教授には審査会を通して貴重なコメントを頂きました．ここに深く感謝致します．

ATR 音声言語コミュニケーション研究所第 1 研究室の中村哲室長には，研究内容について異なった立場からの御意見や御指導を頂きました．ここに深く感謝いたします．また，本研究に対して御意見を頂きました，第 1 研究室の研究員や技術員の方々には，深く感謝いたします．

本研究をまとめることができましたのも，皆様の御指導の賜であり，感謝の意を表しつつ本論文の結びと致します．

参考文献

[1] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell Systems Technical Journal*, Vol. 62, No. 4, pp. 1035–1074, 1983.

[2] L.R. Rabiner, S.E. Levinson, and M.M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition," *Bell Systems Technical Journal*, Vol. 62, No. 4, pp. 1075–1105, 1983.

[3] L.R. Rabiner and B.-H. Juang, "An Introduction to Hidden Markov Models," *IEEE Transactions on Acoustics Speech, Signal Processing*, Vol. 3, No. 1, pp. 4–16, 1986.

[4] 大河内 正明, "Hidden Markov Model に基づいた音声認識," **日本音響学会誌**, Vol. 42, No. 12, pp. 936–941, 1986.

[5] 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.

[6] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book version 3.2*. Cambridge University Engineering Department, 2002.

[8] H. Sakoe, "Two-Level DP-matching – A Dynamic Programming-Based Pattern Matching Algoritm for Connected Word Recognition," *IEEE Transactions on Acoustics Speech, Signal Processing*, Vol. 27, No. 6, pp. 588–595, 1979.

[9] C.S. Myers and L.R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Transactions on Acoustics Speech, Signal*

Processing, Vol. 29, No. 2, pp. 284–297, 1981.

[10] J.S. Bridle, M.D. Brown, and R.M. Chamberlain, "An Algorithm for Connected Word Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 899–902, 1982.

[11] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Transactions on Acoustics Speech, Signal Processing*, Vol. 32, No. 2, pp. 263–271, 1984.

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 39, No. 1, pp. 1–38, 1977.

[13] L.E. Baum and J.A. Eagon, "An Equality with Applications to Statistical Prediction for Functions of Markov Process and to a Model for Ecology," *Bulletin of the American Mathematical Society*, Vol. 73, pp. 360–363, 1967.

[14] 板倉 文忠, 斉藤 収三, "統計的手法による音声スペクトル密度とホルマント周波数の推定," **電子情報通信学会論文誌**, Vol. J53-A, No. 1, pp. 35–42, 1970.

[15] 板倉 文忠, "低ビットレート音声符号化," **電子情報通信学会誌**, Vol. 70, No. 4, pp. 386–391, 1987.

[16] A.M. Noll, "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection," *Journal of the Acoustical Society of America*, Vol. 36, No. 2, pp. 296–302, 1964.

[17] A.V. Oppenheim and R.W. Schafer, "Homomorphic Analysis of Speech," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-16, No. 2, pp. 221–226, 1968.

[18] F. Jelinek, "Self-Organized Language Modelling for Speech Recognition," *IBM Technical Journal*, Watson Research Center, Unpublished, 1985.

[19] E. Bocchieri and B. Mak, "Subspace Distribution Clustering for Continuous Observation Density Hidden Markov Models," *Proceedings of European Conference on Speech Communication and Technology*, Vol. 1, pp. 107–110, 1997.

[20] M. Ostendorf, V.V. Digalakis, and O.A. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 5, pp. 360–378, 1996.

[21] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 127–130, 1988.

[22] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Transactions on Acoustics Speech, Signal Processing*, Vol. 37, No. 12, pp. 1857–1869, 1989.

[23] H. Gish and K. Ng, "A Segmental Speech Model with Applications to Word Spotting," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 447–450, 1993.

[24] L. Deng, M. Aksmanovic, X. Sun, and C.F.J. Wu, "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 507–520, 1994.

[25] B.-H. Juang, "Maximum Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT&T Technical Journal*, Vol. 64, No. 6, pp. 1235–1249, 1985.

[26] 高橋 敏, 嵯峨山 茂樹, "離散混合出力分布型HMM," **日本音響学会平成 *8* 年度秋季研究発表会講演論文集**, Vol. 1, 2-3-2, pp. 51–52, 1996.

[27] M.J. Russell and A.E. Cook, "Experimental Evaluation of Duration Modelling Techniques for Automatic Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5–8, 1985.

[28] L.R. Rabiner, B.-H. Juang, S.E. Levinson, and M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Technical Journal*, Vol. 64, No. 6, pp. 1211–1234, 1985.

[29] S.E. Levinson, "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Computer Speech and Language*, Vol. 1, No. 1, pp. 29–45, 1986.

[30] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1205–1208, 1985.

[31] K.-F. Lee and H.-W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 123–126, 1988.

[32] 嵯峨山 茂樹, "音素環境クラスタリングの原理とアルゴリズム," **電子情報通信学会技術報告**, SP87-86, pp. 1–6, 1987.

[33] 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**, Vol. J76-DII, No. 10, pp. 2155–2164, 1993.

[34] S. Hayamizu, K.-F. Lee, and H.-W. Hon, "Description of Acoustic Variations by Tree-Based Phone Modeling," *Proceedings of International Conference on Spoken Language Processing*, pp. 705–708, 1990.

[35] M. Ostendorf and H. Singer, "HMM Topology Design Using Maximum Likelihood Successive State Splitting," *Computer Speech and Language*, Vol. 11, No. 1, pp. 17–42, 1997.

[36] 堀 貴明, 加藤 正治, 伊藤 彰則, 好田 正紀, "音素決定木に基づく逐次状態分割法による HM-Net の検討," **電子情報通信学会論文誌**, Vol. J80-DII, No. 10, pp. 2645–2654, 1997.

[37] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modeling," *Proceedings of the ARPA Human Language Technology Workshop*, pp. 307–312, 1994.

[38] S.J. Young and P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proceedings of European Conference on Speech Communication and Technology*, pp. 2203–2206, 1993.

[39] X.D. Huang, K.-F. Lee, H.-W. Hon, and M.Y. Hwang, "Improved Acoustic Modeling with the SPHINX Speech Recognition System," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 345–348, 1991.

[40] J. Bellegarda and D. Nahamoo, "Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 13–16, 1989.

[41] D.B. Paul, "The Lincoln Robust Continuous Speech Recognizer," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 449–452, 1989.

[42] X.D. Huang and M.A. Jack, "Unified Techniques for Vector Quantization and Hidden Markov Modeling Using Semi-Continuous Models," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 639–642, 1989.

[43] M.Y. Hwang and X.D. Huang, "Subphonetic Modeling with Markov States – SENONE," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 33–36, 1992.

[44] V. Digalakis and H. Murveit, "Genones: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 537–540, 1994.

[45] 高橋 敏, 嵯峨山 茂樹, "4 階層共有構造の音響モデルによる音声認識," **電子情報通信学会論文誌**, Vol. J82-DII, No. 3, pp. 315–323, 1999.

[46] C. Cerisara, D. Fohr, and J.P. Haton, "Asynchrony in Multi-Band Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1121–1124, 2000.

[47] N. Mirghafori and N. Morgan, "Sooner or Later: Exploring Asynchrony in Multi-Band Speech Recognition," *Proceedings of European Conference on Speech Communication and Technology*, Vol. 2, pp. 595–598, 1999.

[48] H.J. Nock and S.J. Young, "Loosely Coupled HMMs for ASR," *Proceedings of International Conference on Spoken Language Processing*, Vol. 3, pp. 143–146, 2000.

[49] H. Bourlard and S. Dupont, "A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," *Proceedings of International Conference on Spoken Language Processing*, pp. 422–425, 1996.

[50] B.T. Logan and P.J. Moreno, "Factorial HMMs for Acoustic Modeling," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 813–816, 1998.

[51] M.J. Tomlinson, M.J. Russell, and N.M. Brooke, "Integrating Audio and Visual Information to Provide Highly Robust Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 821–824, 1996.

[52] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous Stream Modeling for Large Vocabulary Audio-Visual Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 169–172, 2001.

[53] 中村 哲, 熊谷 健一, 田村 哲嗣, "バイモーダル音声認識における音素境界を越えた同期性のモデル," **日本音響学会平成** *13* **年度秋季研究発表会講演論文集**, Vol. 1, 1-1-13, pp. 25–26, 2001.

[54] K. Okuda, T. Matsui, and S. Nakamura, "Acoustic Model for Robust Speech Recognition of Stressed Japanese Speech," *Proceedings of Hands-free Speech Communication Workshop*, pp. 127–130, 2001.

[55] 松田 繁樹, 中井 満, 下平 博, 嵯峨山 茂樹, "複数の特徴ベクトル軌道を持つ環境依存音素クラスタの生成," **日本音響学会平成** *13* **年度秋季研究発表会講演論文集**, Vol. 1, 1-1-10, pp. 19–20, 2001.

[56] 伊田 政樹, 中村 哲, "雑音 DB を用いたモデル適応化 HMM の SN 比別マルチパスモデルによる雑音下音声認識," **電子情報通信学会技術報告**, Vol. 101, No. 522, pp. 51–55, 2001.

[57] K. Fukunaga, *Introduction to Statistical Pattern Recognition (Second Edition)*. Academic Press, Inc., San Diego, 1990.

[58] 中井 満, 嵯峨山 茂樹, 秋良 直人, 小場 久雄, 下平博, “ストローク HMM によるオンライン手書き文字認識の性能評価,” **電子情報通信学会技術報告**, PRMU2000-36, pp. 9–16, 2000.

本研究に関する発表論文

[1] 松田繁樹, 中井満, 下平博, 嵯峨山茂樹: “非同期遷移型 HMM による音声認識” 電子情報通信学会論文誌. （採録決定済 2003 年 6 月掲載予定）

[2] S. Matsuda, M. Nakai, H. Shimodaira and S. Sagayama: “Feature-Dependent Allophone Clustering,” *Proceedings of International Conference on Spoken Language Processing*, pp. 413–416, 2000.

[3] S. Matsuda, M. Nakai, H. Shimodaira and S. Sagayama: “Asynchronous-Transition HMM,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp.1001-1004, Jun 2000.

[4] S. Sagayama, S. Matsuda, M. Nakai and H. Shimodaira, “Asynchronous-Transition HMM for Acoustic Modeling,” *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, in CD-ROM, 1999.

[5] 松田繁樹, 中井満, 下平博, 嵯峨山茂樹: “状態遷移に順序関係を持つ非同期遷移型 HMM,” **電子情報通信学会技術報告**, SP99-98 pp. 31-36, 1999.

[6] 松田繁樹, 中井満, 下平博, 嵯峨山茂樹: “特徴量間で状態遷移が非同期な HMM,” **情報処理学会研究報告**, 99-SLP-27-4, pp. 25–32, 1999.

[7] 松田繁樹, 中井満, 下平博, 嵯峨山茂樹: “複数混合分布を持つ順序制約付き非同期遷移型 HMM,” **日本音響学会平成** *12* **年度秋季研究発表会講演論文集**, 1-5-11, pp. 21-22, 2000.

[8] 松田繁樹, 中井満, 下平博, 嵯峨山茂樹: “順序制約付き非同期遷移型 HMM による連続音素認識,” **日本音響学会平成** *12* **年度春季研究発表会講演論文集**, 1-8-12, pp. 23-24, 2000.

[9] 松田繁樹, 中井満, 下平博, 嵯峨山茂樹: “非同期遷移型 HMM,” **日本音響学会平成** *11* **年度秋季研究発表会講演論文集**, 1-1-2, pp. 23–24, 1999.

第 A 章

使用した音素ラベル

音素ラベル	単語例	音素ラベル	単語例
a	aisatsu (挨拶)	o	oNgaku (音楽)
b	bakugeki (爆撃)	p	posuto (ポスト)
ch	chikyuu (地球)	q	saqkaku (錯覚)
d	daNsei (男性)	r	reigi (礼儀)
e	eNgeki (演劇)	s	saqka (作家)
f	fukusou (服装)	sh	shihei (紙幣)
g	geNki (元気)	t	taiyou (太陽)
h	haqtatsu (発達)	ts	tsuyu (梅雨)
i	ikioi (勢い)	u	udoN (うどん)
j	jiNkou (人口)	w	wariai (割合)
k	kita (北)	y	yume (夢)
m	masatsu (摩擦)	z	zaimoku (材木)
n	niNshiki (認識)	N	saiN (サイン)

第 B 章

未知モデルの補間法

音声認識を行うためには，音素ネットワークや単語ネットワーク中に存在する環境依存音素すべてのモデルの尤度を計算する必要がある．音素分類木によるトップダウンクラスタリングを用いて生成された従来型モデルは，その分類木の情報を元に不足モデルの状態を学習された状態により割り当てることができる．しかし，本研究で説明した AT-HMM 生成法は，学習データ中に存在した環境依存音素以外のモデルを生成することはできない．時間非同期遷移構造の有無による音声認識性能の違いを評価するには，従来型 HMM と AT-HMM 両方に対して同じモデル補間法を用いる方が望ましい．そこで本研究では，特徴ベクトル時系列の値の変化が連続であることを仮定した手法を共通のモデル補間法として用いる．

図 B.1 に，このモデル補間法の概念図を示す．このモデル補間法は，左側に接続可能なモデルの最終状態と割り当て候補モデルの開始状態間の距離，また右側に接続可能なモデルの開始状態と割り当て候補モデルの最終状態間の距離の和が最小となる候補モデルを割り当てる手法である．具体例として，不足しているモデルが ${}_a/k/_i$ の場合，${}_a/k/_*$ や ${}_*/k/_i$ のモデルが割り当て候補モデルである．また，この不足モデルの左側には，${}_*/a/_k$ が接続し，右側には ${}_k/i/_*$ が接続しなければならない．

式 (B.4) は，不足モデルを ${}_A/B/_C$ とした場合の，補間モデル $\hat{m}_c$ の計算を表している．式 (B.1) は，補間候補モデルの集合 M_C，式 (B.2) と式 (B.3) は，左側と右側接続可能モデルの集合 M_L と M_R である．また，$m(b)$ と $m(e)$ はモデル m の開始状態の終了状態である．d は状態間距離の関数である．

$$M_C = \begin{cases} {}_*/B/_C \cup_A /B/_* & (|{}_*/B/_C \cup_A /B/_*| \neq 0) \\ {}_*/B/_* & (|{}_*/B/_C \cup_A /B/_*| = 0) \end{cases} \tag{B.1}$$

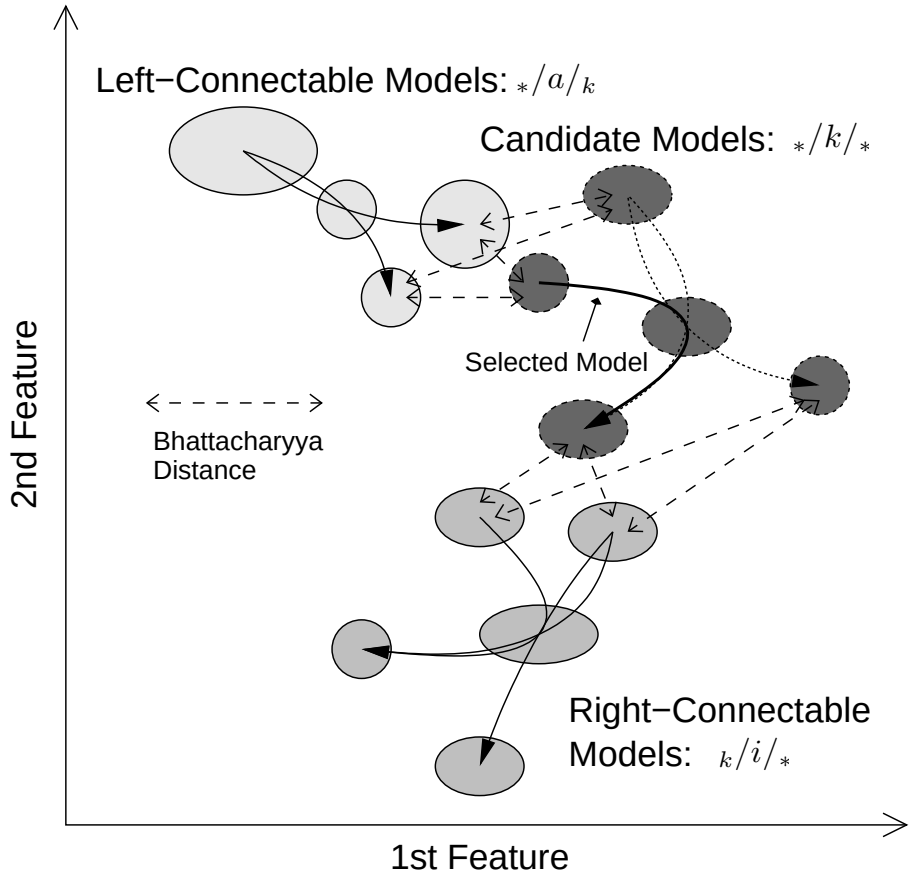


図 B.1: モデル補間法の概念図

$$M_L = \begin{cases} *\!/A/\!_B & (|*\!/A/\!_B| \neq 0) \\ *\!/A/\!_* & (|*\!/A/\!_B| = 0) \end{cases} \tag{B.2}$$

$$M_R = \begin{cases} _B\!/C/\!_* & (|_B\!/C/\!_*| \neq 0) \\ */C/\!_* & (|_B\!/C/\!_*| = 0) \end{cases} \tag{B.3}$$

$$\hat{m_c} = \underset{m_c \in M_C}{\operatorname{argmin}} \sum_{m_l \in M_L} d(m_l(e), m_c(b)) + \sum_{m_r \in M_R} d(m_c(e), m_r(b)) \tag{B.4}$$

本研究では，状態間距離として Bhattacharyya 距離 [57] を用いた．式 (B.5) に Bhattacharyya 距離の定義を示す．式中の $b^{(i)}$ はガウス分布，$\mu^{(i)}$ と $\Sigma^{(i)}$ は平均ベクトルと共分散行列である．

$$d(b^{(1)}, b^{(2)}) \quad = \quad \frac{1}{8}(\mu^{(1)} - \mu^{(2)})^t (\frac{\Sigma^{(1)} + \Sigma^{(2)}}{2})^{-1} (\mu^{(1)} - \mu^{(2)}) +$$

$$\frac{1}{2}\ln\frac{\left|(\Sigma^{(1)}+\Sigma^{(2)})/2\right|}{\left|\Sigma^{(1)}\right|^{1/2}\left|\Sigma^{(2)}\right|^{1/2}} \tag{B.5}$$

このモデル補間法により，あらゆる環境依存音素の尤度を計算することが可能となる．

第 C章

最尤逐次状態分割法

最尤逐次状態分割法（Maximum Likelihood Successive State Splitting: ML-SSS）は，隠れマルコフネットワーク構造（Hidden Markov Network: HMnet）を音素環境クラスタリングを基礎とした状態分割を行いながら，除々にモデルを精密化してゆくアルゴリズムである．分割における環境要因として，先行音素環境，当該音素環境，後続音素環境の 3 つ組の環境要因が用られる．図 C.1 に，ML-SSS 法のアルゴリズムの流れを示す．

C.1　初期モデルの学習

オリジナルの SSS 法 [33] では，初期モデルとして全音素環境を含んだ 1 状態の HMnet から分割を開始している．しかし，この単一状態の初期モデルの場合，計算コストや，ある程度分割が進むまでは，個々の音素モデルが生成されず認識に用いることができないなどの問題があり，一般的には，各音素 1 状態の HMnet が用いられる．

C.2　分割状態の決定と分割処理

分割処理では，HMnet 内の全状態中で，以下の式により求められた分割ゲインが最も大きい状態が分割される．分割を検討する状態を s，分割後の仮状態をそれぞれ q_0, q_1 とした場合の分割ゲインである．また，$A(s)$ を状態 s における状態遷移確率，$B(s)$ は状態 s における出力確率分布のパラメータである．

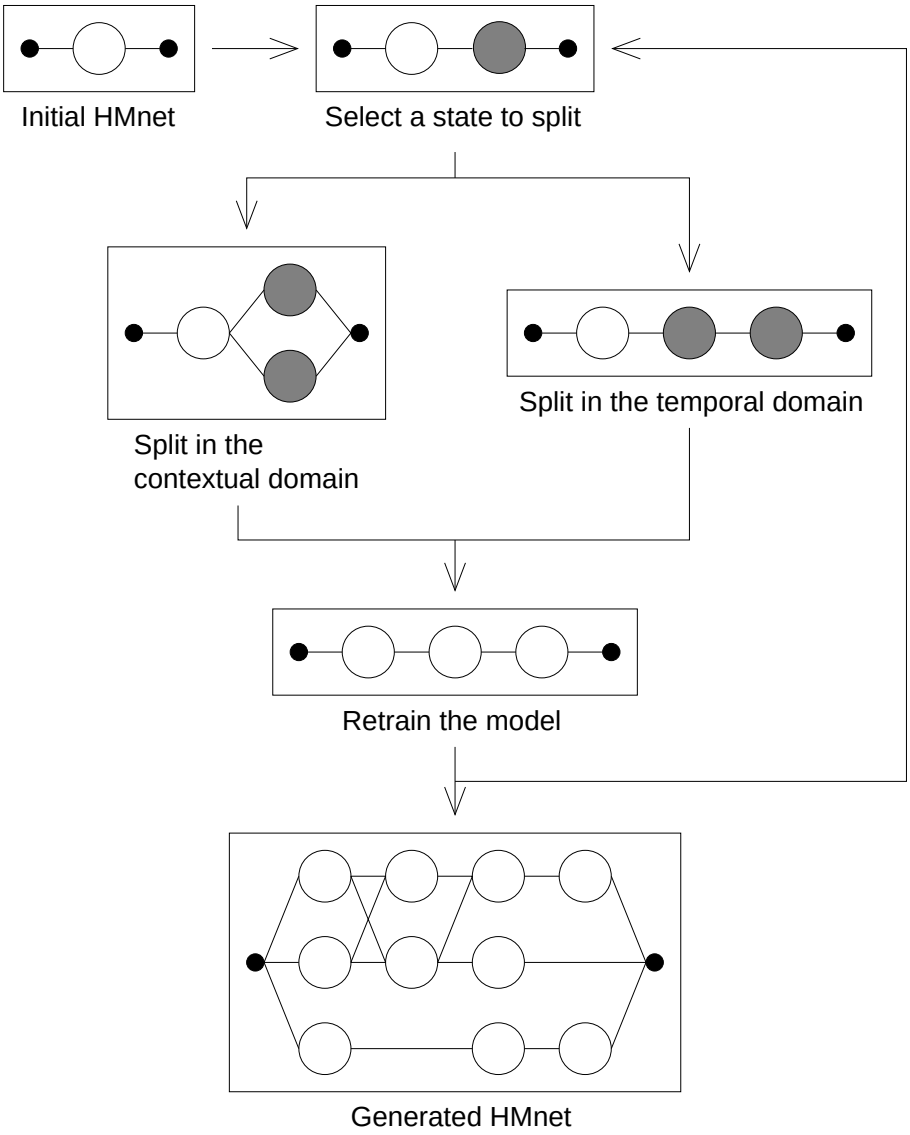


図 C.1: ML-SSS 法の処理の流れ

$$G(s^*, q_0, q_1) = -\sum_t [\xi_t(s^*, s^*) \log a_{s^* s^*} + \gamma_t(s^*) \log p(o_t|s^*, B(s))] \tag{C.1}$$

$$+ \sum_t [\sum_{s=q_0,q_1} \sum_{s=q_0,q_1} \xi_t(s, s') \log a_{ss'} + \sum_{s=q_0,q_1} \gamma_t(s) \log p(o_t|s, B(s))]$$

出力確率分布をガウス分布とした場合,

$$p(o_t|s, B(s)) \sim N(\mu(s), \Sigma(s))$$

また,

$$\begin{aligned}
a_{ss'} &= p(s_t = s|s_{t-1} = s', A(s')) \\
N_1(s) &= \sum_t \gamma_t(s) \\
N_2(s, s') &= \sum_t \xi_t(s, s') \\
\gamma_t(s) &= p(s_t = s|y_1^T, \lambda) \\
\xi_t(s, s') &= p(s_t = s, s_{t-1} = s'|y_1^T, \lambda)
\end{aligned}$$

とすると,

$$G(s^*, q_0, q_1) = -N_2(s^*, s^*) \log a_{s^* s^*} + 0.5 N_1(s^*) \log |\Sigma(s^*)| \tag{C.2}$$

$$+ \sum_{s=q_0,q_1} \sum_{s'=q_0,q_1} N_2(s, s') \log a_{ss'} - \sum_{s=q_0,q_1} 0.5 N_1(s) \log |\Sigma(s)|$$

となる．また，出力確率分布が共分散行列の場合，$|\Sigma(s)| = \sum_{d=1}^{D} \sigma_d^2$ である．

式 (C.2) をそのまま計算した場合，対象とする状態の分割によって影響を受ける状態全てに対して再学習する必要がある．しかし，この計算は大量の計算コストが必要である．そのため，実際の分割ゲインの計算では，分割前の状態が受け持っている音響特徴ベクトル時系列区間内での分割といった近似を用いている．次に，その手法について説明する．

C.2.1 音素環境方向分割のゲイン計算

音素環境方向分割とは，HMnet 内に表現されている各状態の環境部分空間を分割することである．分割例を図 C.2 に示す.

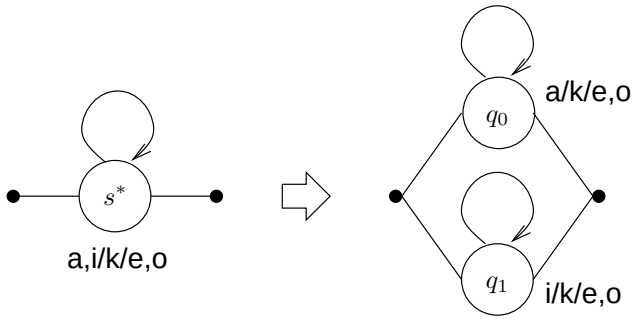


図 C.2: 状態 s^* から状態 q_0 と q_1 への先行音素環境要因による音素環境方向分割

状態を音素環境方向の分割において，分割後の環境部分空間を求めることが重要である．ML-SSS 法では，3 つの環境要因に対して以下に示す k-mean クラスタリング法を基礎としたアルゴリズムを用いて分割後の空間と分割ゲインの計算を行う．この場合の分割ゲインの計算は，式 (C.2) から，以下の様になる．

$$G_C(s^*, q_0, q_1) = 0.5[N_1(s^*)\log|\Sigma(s^*)| - N_1(q_0)\log|\Sigma(q_0)| - N_1(q_1)\log|\Sigma(q_1)|] \qquad \text{(C.3)}$$

アルゴリズムの説明の準備のために，各変数の説明を行う．分割を検討する状態を s，分割後の仮状態を q_0 と q_1 とする．環境要因内の要素の集合を $J = \{a, i, u, e, o, k, s, \cdots, N\}$，各環境要因内の音素集合を $A \subseteq J$，また環境部分空間は $A_l(s) \times A_c(s) \times A_r(s)$ である．

- 初期化 $p = 0$

  仮の分割状態 $\lambda^{(0)}(q_0)$ と $\lambda^{(0)}(q_1)$ に対して状態パラメータの初期値を割り当てる．

$$\lambda^{(0)}(q_0) = \lambda(s) = (\mu(s), \Sigma(s)) \qquad \text{(C.4)}$$

$$\lambda^{(0)}(q_1) = (\mu(s)(1+\epsilon), \Sigma(s))$$

- 繰り返し $p = 1, 2, \cdots$

  1. 新しい 2 分割の発見 $\{A^{(p)}(q_0), A^{(p)}(q_1)\}$

     各 $x_j \in A(s)$ に対して，

$$\sum_{t:x_t=x_j} \log p(o_t|\lambda^{(p-1)}(q_0)) \geq \sum_{t:x_t=x_j} \log p(o_t|\lambda^{(p-1)}(q_1)) \tag{C.5}$$

であるならば，x_j は $A^{(p)}(q_0)$ に属する．それ以外は $A^{(p)}(q_1)$ に属する．x_t は音声サンプルの時刻 t における音素環境である．

2. セントロイドの推定 $\{\lambda^{(p)}(q_k) = (\mu^{(p)}(q_k), \Sigma^{(p)}(q_k)) : k = 0, 1\}$

   標準の最尤パラメータ推定を用いる．

$$\mu^{(p)}(q_k) = \frac{1}{N_k} \sum_{t:x_t \in A^{(p)}(q_k)} o_t \tag{C.6}$$

$$\Sigma^{(p)}(q_k) = \frac{1}{N_k} \sum_{x_t \in A^{(p)}(q_k)} \sum_{t:x_t=x_j} (o_t - \mu^{(p)}(q_k))(o_t - \mu^{(p)}(q_k))^t \tag{C.7}$$

   ここで，$N_k = \sum_{x_j \in A^{(p)}(q_k)} N_j$ である．N_j は，$\{t : x_t = x_j, s_t = s\}$ の要素数である．また，$N_0 + N_1 = N_s$ である．

3. 分散に対する評価

   分割が変更されない場合や，相対的な尤度ゲインの変化が小さい時は停止する．

$$\frac{L^{(p)} - L^{(p-1)}}{|L^{(p-1)}|} < \eta \tag{C.8}$$

   ここで，式 (C.3) より，$L^{(p)} = -N_0 \log |\Sigma^{(p)}(q_0)| - N_1 \log |\Sigma^{(p)}(q_1)|$ である．また，η はヒューリスティックな値である．

C.2.2 時間方向分割のゲイン計算

時間方向分割とは，図 C.3 に示す様に，HMnet 内の状態を時間方向に分割することである．

分割を検討する状態を s，分割後の仮状態を q_0 と q_1，分割後の新しい状態のパラメータを $\lambda = \{\mu(q_0), \sigma^2(q_0), \nu(q_0), \mu(q_1), \sigma^2(q_1), \nu(q_1)\}$ とする．この時，分割前後の γ と ξ に対する関係は，式 (C.9) の様になる．

$$\gamma_t(s^*) = \gamma_t(q_0) + \gamma_t(q_1) \tag{C.9}$$

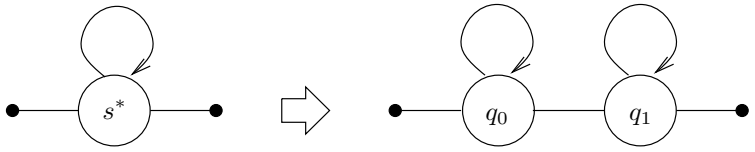


図 C.3: 状態 s^* から状態 q_0 と q_1 への時間方向分割

$$\xi_t(s^*, s^*) = \xi_t(q_0, q_0) + \xi_t(q_1, q_0) + \xi_t(q_1, q_1)$$

また,

$$\tilde{\gamma}_t(q) = p(q_t = q | s_t = s^*, o_1^T) \tag{C.10}$$

$$\tilde{\xi}_t(q, q') = p(q_t = q, q_{t-1} = q' | s_t = s^*, s_{t-1} = s^*, o_1^T)$$

の時,

$$\gamma_t(q) = p(q_t = q | o_1^T) = p(q_t = q, s_t = s^* | o_1^T) = \tilde{\gamma}_t(q)\gamma_t(s^*) \tag{C.11}$$

$$\xi_t(q, q') = p(q_t = q, q_{t-1} = q' | y_1^T) = p(q_t = q, q_{t-1} = q', s_t = s^*, s_{t-1} = s^* | y_1^T)$$

$$= \tilde{\xi}_t(q, q')\xi_t(s^*, s^*)$$

となる．$\tilde{\gamma}_t(q)$ と $\tilde{\xi}_t(q, q')$ は標準のフォワードバックワードアルゴリズムにより，分割前の状態 s^* の受け持つ区間 $(2 \leq t \leq 7)$ の $\gamma_t(s^*) > 0$ な範囲を，分割後の 2 状態 q_0 と q_1 によって計算することにより求めることができる．その様子を図 C.4 に示す.

上述の計算によって求められた $\tilde{\gamma}_t(q)$ と $\tilde{\xi}_t(q, q')$ を用い，新しい λ を推定する.

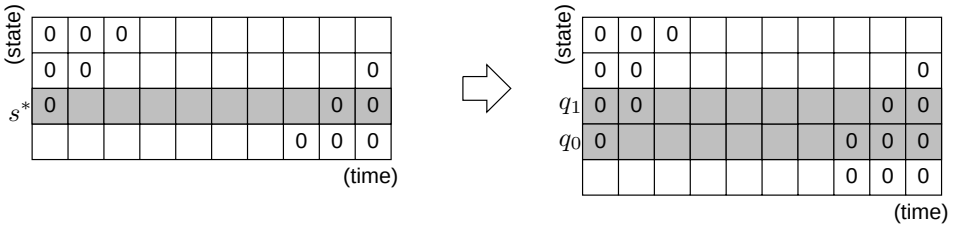


図 C.4: 状態 s から仮状態 q_0 と q_1 への時間方向分割における計算範囲

$$\mu_m(q) = \frac{\sum_t \tilde{\gamma}_t(q)\gamma_t(s^*)o_{t,m}}{\sum_t \tilde{\gamma}_t(q)\gamma_t(s^*)} \tag{C.12}$$

$$\sigma_m^2(q) = \frac{\sum_t \tilde{\gamma}_t(q)\gamma_t(s^*)o_{t,m}^2}{\sum_t \tilde{\gamma}_t(q)\gamma_t(s^*)} - \mu_m(q)^2 \tag{C.13}$$

$$\nu(q) = \frac{\sum_t \tilde{\xi}_t(q,q)\xi_t(s^*,s^*)}{\sum_{q'}\sum_t \tilde{\xi}_t(q',q)\xi_t(s^*,s^*)} \tag{C.14}$$

この様にして求められた分割後のモデルパラメータ λ を用いて，式 (C.2) によって時間方向分割のゲインを計算を行う．

これらの計算によって，各状態に対する分割ゲインを求めることができ，最も大きなゲインを持つ状態の分割処理を行い，その後，状態分割によって影響を受けた状態 (始端，終端を通らずに辿ることのできる全状態) を，次の状態分割に備えて再学習を行わなければならない．この様な分割処理を必要な状態数まで繰り返す．

C.3 全状態の再学習

ここまでの処理で HMnet の構造決定が終了する．そこで最後に，各状態の割り当てられている出力確率分布を実際の HMnet でしようする最終的な形状に変更し，その条件下での再学習を HMnet 全体に対して行う．こうして HMnet の生成が完了する．

C.4 ML-SSS により生成された HMnet の例

ML-SSS 法によって生成された HMnet の例を示すため，音素/k/のみを対象とし，状態数が 10 になるまで分割が進行した時点での構造を図 C.5 に示す．HMnet を生成した時の条件を，次に示す．

- 学習データ

  ATR 研究用日本語音声データベースの男性話者 1 名分 (mht) 中の，重要語 5240 単語中奇数番目及び，音韻バランス単語を使用

- 音響特徴ベクトル

  対数パワー，12 次元 MFCC，Δ 対数パワー，12 次元 ΔMFCC からなる計 26 次元ベクトル

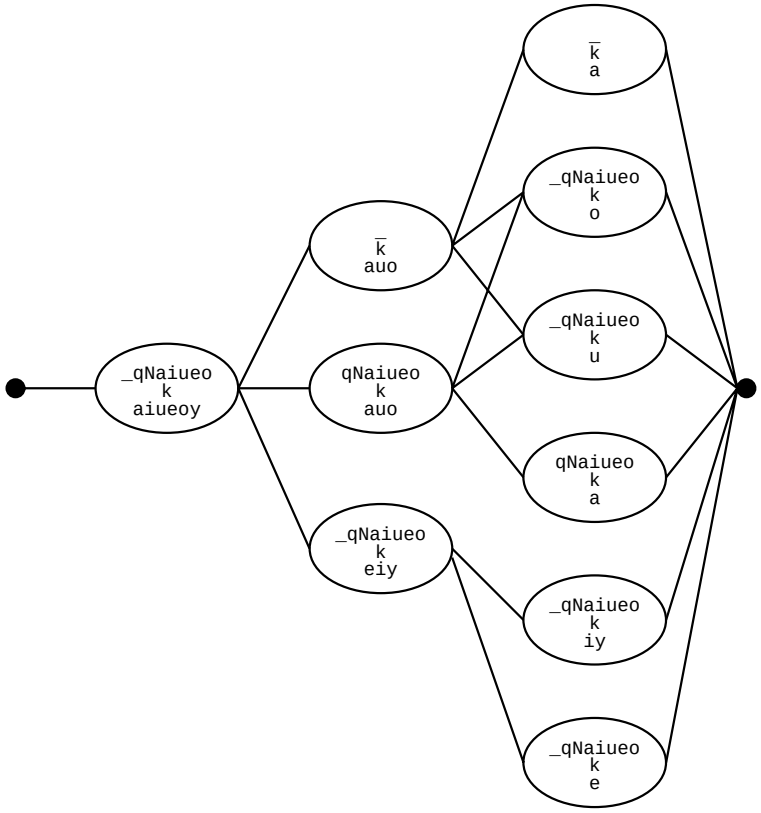


図 C.5: ML-SSS により生成された音素/k/の HMnet

- 音響分析

  サンプリング周波数 20kHz，16bit 量子化，25ms ハミング窓，フレーム周期 8ms

- 使用音素ラベル

  付録第 A 章に示される音素ラベルを使用

  “-” は，無音を表す

- 初期モデル

  時間方向状態数 3 の HMnet を使用

第 D 章

特徴量別逐次状態分割法により生成された FD-HMnet の例

FD-SSS 法により生成された音素/k/の FD-HMnet の例を示す．この FD-HMnet を生成した時の条件を以下に示す．音素/k/のみを対象とし，全状態数が 260 になるまで状態を分割した時点での構造を，図 D.1 に示す．

- 学習データ

  ATR より提供される研究用日本語音声データベースの男性話者 1 名分 (mht) 中の，重要語 5240 単語中奇数番目及び，音韻バランス単語を使用

- 音響特徴ベクトル

  対数パワー，12 次元 MFCC，Δ 対数パワー，12 次元 ΔMFCC からなる計 26 次元ベクトル

- 音響分析

  サンプリング周波数 20kHz，16bit 量子化，25ms ハミング窓，フレーム周期 8ms

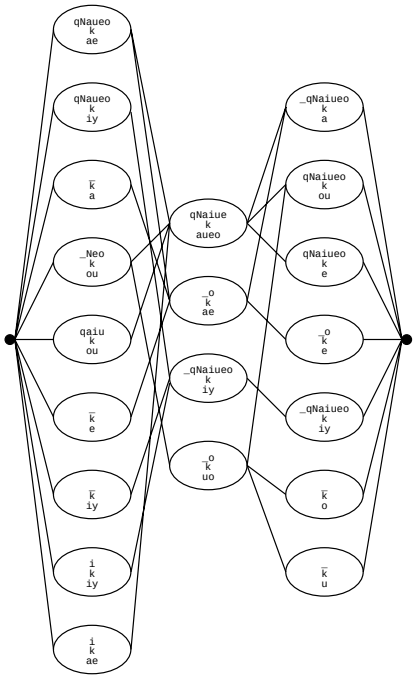
- 使用音素ラベル

  付録第 A 章に示される音素ラベルを使用

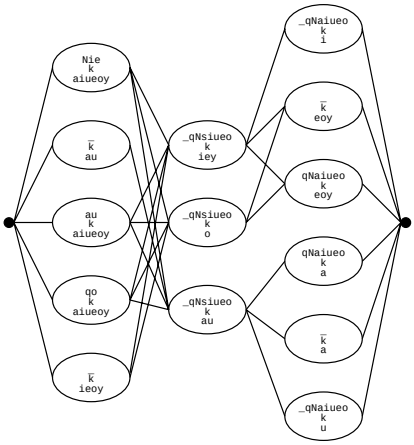
  “-” は，無音を表す

- 初期モデル

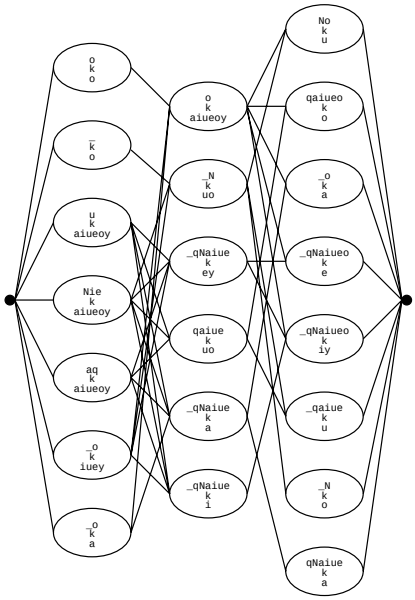
  各特徴量は，時間方向状態数 3 の FD-HMnet を使用



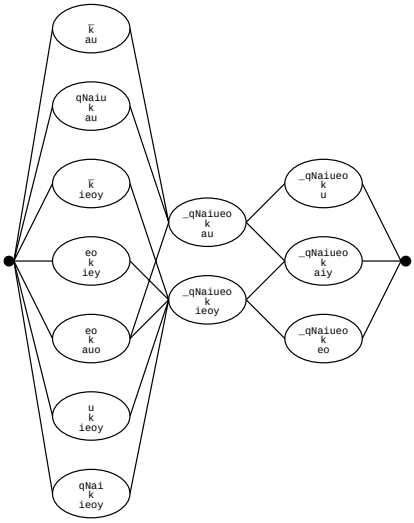
1st MFCC



2nd MFCC

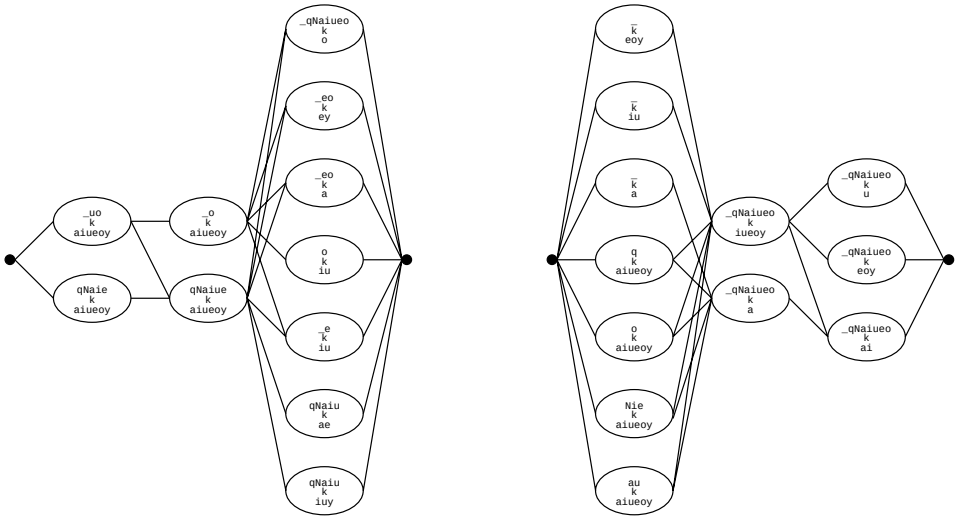


3rd MFCC



4th MFCC

図 D.1: FD-SSS により生成された音素/k/の FD-HMnet



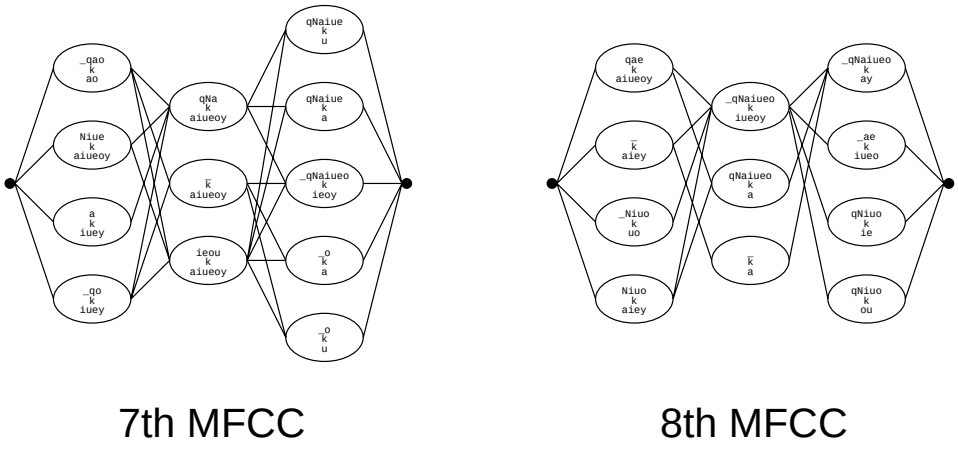


図 D.2: FD-SSS により生成された音素/k/の FD-HMnet（続き）

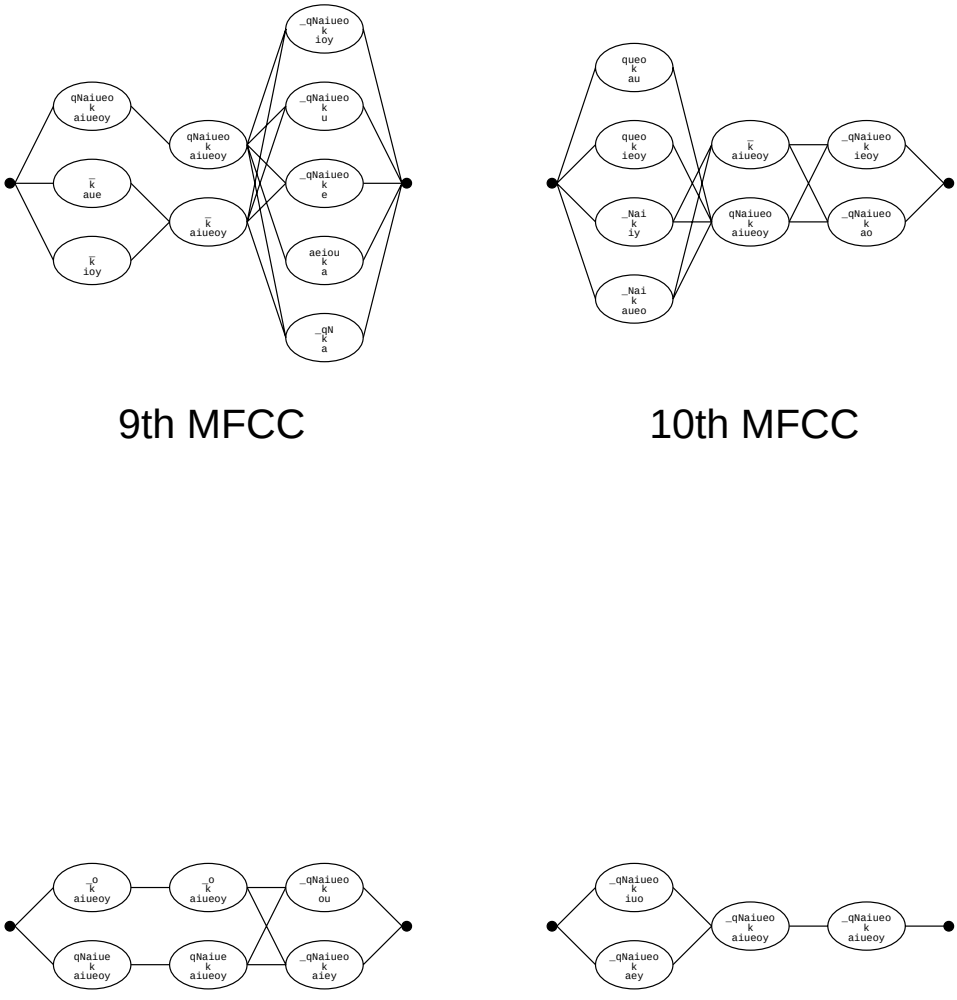
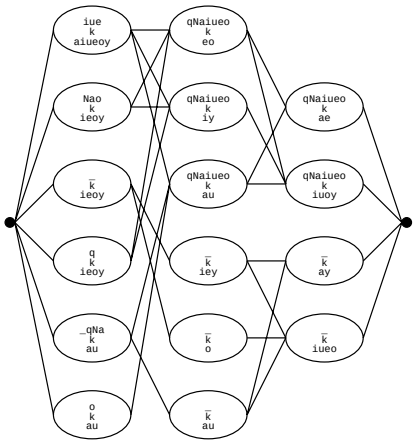
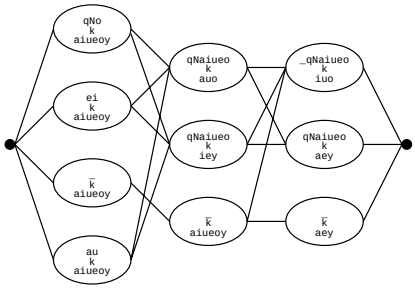


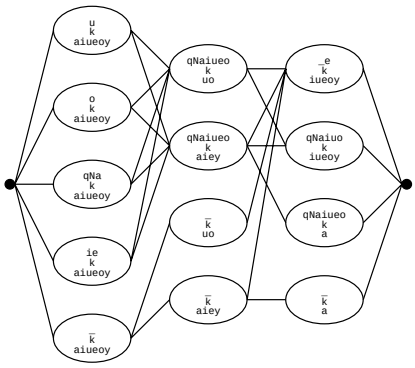
図 D.3: FD-SSS により生成された音素/k/の FD-HMnet（続き）



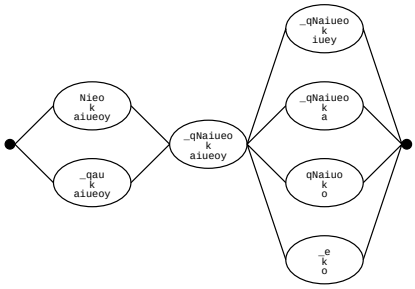
1st Delta MFCC



2nd Delta MFCC

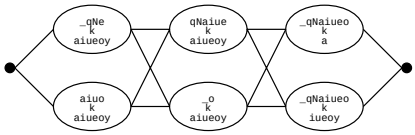


3rd Delta MFCC

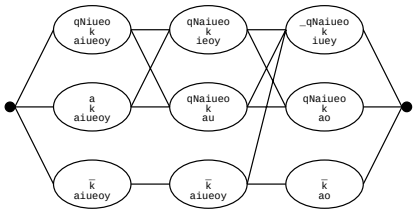


4th Delta MFCC

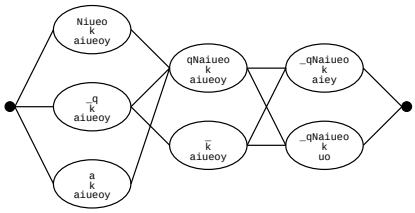
図 D.4: FD-SSS により生成された音素/k/の FD-HMnet（続き）



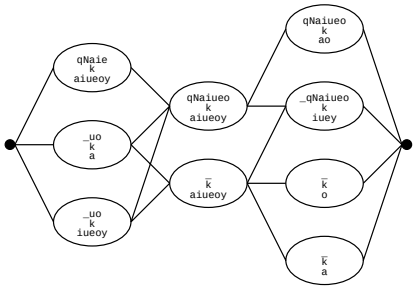
5th Delta MFCC



6th Delta MFCC

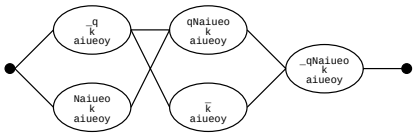


7th Delta MFCC

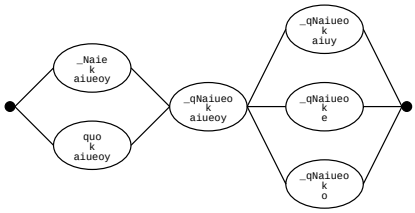


8th Delta MFCC

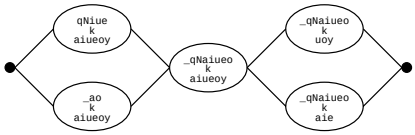
図 D.5: FD-SSS により生成された音素/k/の FD-HMnet（続き）



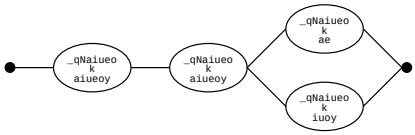
9th Delta MFCC



10th Delta MFCC



11th Delta MFCC



12th Delta MFCC

図 D.6: FD-SSS により生成された音素/k/の FD-HMnet（続き）

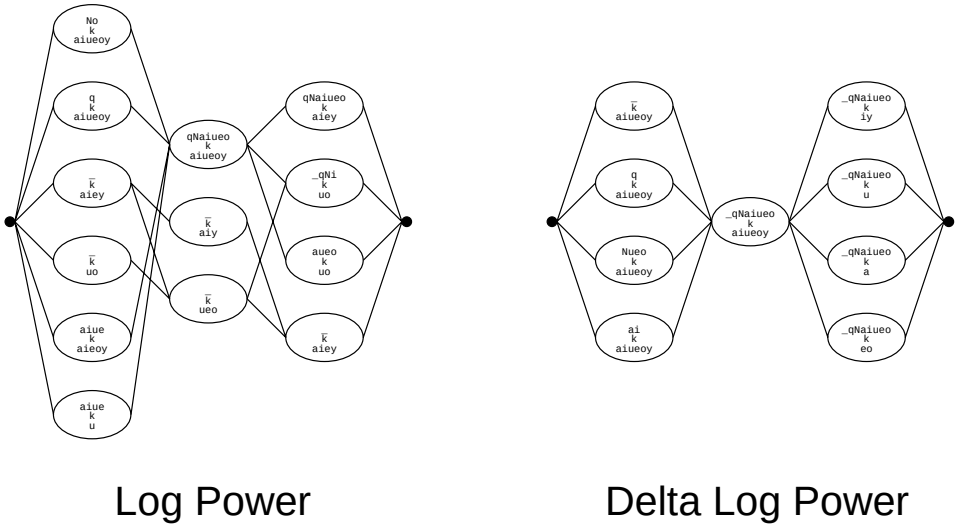


図 D.7: FD-SSS により生成された音素/k/の FD-HMnet（続き）