修士論文

音声の構造的表象を用いた音声認識に関する基礎的研究

2006 年 2 月 3 日

指導教員　峯松 信明 助教授

東京大学大学院情報理工学系研究科
電子情報学専攻　46439

村上 隆夫

内容梗概

我々が音声によるコミュニケーションを行なう際，その音声を生成する際には話者の声道形状の特性，収録・伝送・再生の際には音響機器の特性，聴取の際には聴覚特性，といった非言語的特徴が不可避的に混入する．従来の音声認識技術は音声の物理的実体を捉え，モデル化してきた．しかしながら，この物理的実体は上記の非言語的特徴の影響を不可避的に受け，ある意味歪んだ実体となる．この問題に対処するため，従来では沢山の学習話者を集めることで不特定話者音響モデルを構築し，さらには話者適応，話者正規化技術の研究が様々な形で行なわれてきたが，根本的な解決には至っていない．

近年，上記の非言語的特徴を表現する線形変換性・乗算性歪みを一切保有しない音響的普遍構造が提案されている．これは，音声の物理的実体を捨象し，関係のみを捉えることによって得られる音声の構造的表象であり，構造音韻論の物理実装として解釈される．

本研究では，この構造を音声認識に利用することに関する基礎的な実験を行なった．まず，音声の構造化による非言語的特徴の消失に関する定量的分析を行ない，その効果を示した．次に，認識タスクとして孤立５母音系列を考え，これを構造のみを用いて認識する実験を行なった．その結果，音声事象分布の最大事後確率（Maximum A Posteriori; MAP）推定，及びローパスフィルタを用いたスペクトル高域成分の均一化により，認識性能が飛躍的に向上し，音声の物理的実体を明示的に用いない提案手法が，学習話者１名で 100%の性能を以って認識することに成功した．従来手法との比較実験も行ない，学習話者１名の提案手法が，CMN（Cepstral Mean Normalization）による正規化を施した学習話者 4,130 名の従来手法を上回る結果が得られた．

雑音環境下における孤立５母音系列の認識実験も行なった．この際，構造は加算性雑音によって歪むものの，同時にスペクトル高域成分に多く含まれる話者性を消失させる効果があることを定量的分析により示した．ここでは，学習話者が１名で済むのであれば雑音下で構造統計モデルをオンラインで学習することも可能，という仮定に基づき，雑音下で学習された構造統計モデルを用いた認識実験を行なった．その結果，学習話者１名の提案手法が，SS（Spectral Subtraction）による雑音処理を行ない，CMN による正規化を施した学習話者 4,130 名の従来手法を上回る結果が得られた．

最後に認識タスクを孤立５母音系列から連続５母音系列に拡張し，これを構造を用いて認識する実験を行なった．この際，HMM を用いて連続音声を構造化する枠組みを検討し，その学習アルゴリズムとして変分ベイズ（Variational Bayes; VB）法を導入した．認識実験によってその効果を確認し，結果として連続５母音系列音声を構造のみを用いて 60%以上の性能で認識することができることを示した．

目次

図目次

表目次

第1章

序論

1.1　はじめに

音声認識システムは進歩を遂げ，現在では電話サービスやカーナビゲーションシステムなどに搭載されるまでに至った．しかしながら，その認識性能は人間のそれに未だ遠く及ばない．例えば，[1] は仮に人間が生涯耳にするデータを機械の学習に用いたとしても，現在の音声認識技術では人間の認識性能に遠く及ばないであろうことを予測している．

その要因の一つとして挙げられるのが，音声に混入する非言語的特徴の存在である．我々が音声によるコミュニケーションを行なう際，その音声を生成する際には話者の声道形状の特徴，収録・伝送・再生の際にはその音響機器の特性，聴取の際には聴覚特性，といった非言語的特徴が音声に対して混入する．しかも，これらは人間同士がコミュニケーションを行なう際においても，音声認識システムが認識を行なう際においても「不可避的に」混入するものである．この非言語的特徴によって，音声の物理的実体は変化するため，学習に用いた音声データと認識対象となる入力音声データでこれら非言語的特徴が異なると，認識性能の劣化に繋がる．

従来の音声認識技術は，この問題を解決するために沢山の話者の音声データを集め，それを用いて学習されたモデルを構築してきた．このようなモデルは不特定話者音響モデルと呼ばれる．しかしながら，この不特定話者音響モデルを用いても，性能を劣化させる話者が必ず存在する．このため，学習したモデルを入力話者に近付ける（即ち，話者適応），あるいは入力音声の話者性をモデルに近付ける（即ち，話者正規化）研究も様々な形で行なわれてきたが，根本的な解決には至っていない．その一方で，音声は人間にとって一番楽なコミュニケーションメディアである．これは何故なのであろうか．

例えば言語学は，音素に対して以下の二つを定義している [2]．1) a phoneme is a class of phonetically-similar sounds and 2) a phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system. 音声の「絶対的な」特性に基づく従来手法は，全て前者の定義に基づくものと考えることができる．しかし，これは二つある定義のうちの一方でしかない．

近年，上記の非言語的特徴を表現する次元そのものを保有しない「音響的普遍構造」が提案された [3, 4, 5, 6]．これは，複数の音声事象（例えば音素）の関係（距離）のみを捉えることで得られる音声の構造的表象であり，音声の「相対的な」特性に基づく手法である．従って，これは音素の後者の定義に基づいている．このような非言語的特徴による影響を受けることの無い音声の物理的表象に基づいて，人間がコミュニケーションを行なっていることも知覚実験によって示唆されている [7]．本研究はこの音響的普遍構造を音声認識へ利用することを目的とする．

1.2　本論文の構成

本論文は，全 9 章で構成される．**第 2 章**では，従来の音声認識システムの枠組みと，それを実現するための要素技術を説明する．**第 3 章**では，音声に不可避的に混入する非言語的特徴の数学的モデルについて説明し，その非言語的特徴に対する従来手法を述べる．**第**

4 章では，上記の非言語的特徴を表現する次元を保有しない音声表象である音響的普遍構造について説明し，それと同様に音声の相対関係に基づく他の研究例を紹介する．

第 5 章からは，構造を用いた音声認識及びそれに関連する実験を行なう．**第 5 章**では，構造を用いた音声認識に先立って，音声の構造化による非言語的特徴の消失度合いを定量的に分析する．**第 6 章**では，まず単純な認識タスクとして孤立５母音系列の音声認識を考え，これを構造のみを用いて認識する枠組みを述べた後，クリーン環境下における認識実験を行なう．**第 7 章**では，雑音環境下における孤立５母音系列も認識することを考える．その際，まず従来の雑音処理技術を紹介する．次に，雑音環境下における構造の歪み，及び話者性のさらなる消失に関する分析実験を行なう．その結果を基に，雑音環境下の孤立５母音系列を，構造を用いて認識する実験を行なう．**第 8 章**では，認識タスクを孤立５母音系列から連続５母音系列に拡張し，これを構造を用いて認識する枠組み，及び実際に行なった認識実験について述べる．最後に，**第 9 章**で本論文をまとめ，今後の展望について述べる．

尚，本論文では 3 つの付録が掲載されている．**付録 A** では，**第 6 章**で述べる，孤立５母音系列の音声事象分布の最大事後確率推定の式を導出する．**付録 B** では，**第 8 章**で述べる，連続５母音系列の音声事象分布を，変分ベイズ法を用いて最大事後確率推定するアルゴリズムを導出する．**付録 C** では，本研究に関連する分布及び関数について簡単にまとめる．

第2章

従来の音声認識システム

2.1　はじめに

本章では，従来における音声認識がどのように行なわれるのかを説明するため，まず従来の連続音声認識[1]システムの枠組みを説明する．これは，音声から音声特徴量の時系列データを抽出し，その後音響モデルと言語モデルを用いたサーチを行なうことで，認識結果を出力する，というものである．ここでは，本研究に直接は関連しない言語モデル，デコーダについては説明を省略し，音声特徴量と音響モデルについてより詳細な説明を行なう．

2.2　従来の音声認識システムの枠組み

図 2.1 に，従来の連続音声認識システムの枠組みを示す．まず入力音声 S に対して音声分析を行ない，音声の音韻的特徴[2]をよく表す音声特徴量の時系列データ X を取り出す．次に，この X に対して，音響モデル（Acoustic Model）・言語モデル（Language Model）を用いたサーチを行なう．具体的には，音声特徴量の時系列データ X が現れたときに，その発話内容が単語列 W である事後確率 $P(W|X)$ を考え，これを最大化する $\hat{W}$ を求める．即ちサーチは，

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

を満たす $\hat{W}$ を求めるタスクとして帰着されるが，これはベイズの定理により，

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(X|W)P(W) \tag{2.2}$$

と変形できる．音響モデルは，単語列 W が発声されたときに，音声特徴量の時系列データ X が現れる事後確率 $P(X|W)$ を記述するモデルであり，音声の音響的な特徴を表現する．一方，言語モデルは単語列 W の出現確率 $P(W)$ を記述するモデルであり，ある単語の後ろにはどのような単語がよく用いられるか，というような発話内容の言語的性質を表すモデルである．このように従来の音声認識システムは，入力音声から音韻的特徴をよく表す音声特徴量を抽出した後，音響的制約及び言語的制約の両方の観点からサーチを行ない，その認識結果として単語列 $\hat{W}$ を出力する．

2.3　音声特徴量

2.3.1　ケプストラム

音声分析で抽出される音声特徴量として現在最も広く用いられているのがケプストラム（Cepstrum）である．音声分析において，音声波形からケプストラムを抽出するまでの様子を図 2.2 に示す．まず音声波形から，数十ミリ秒程度のフレームを切り出し，その区間に対して離散フーリエ変換（Discrete Fourier Transform; DFT）を施し，スペクトルを抽出

[1]連続音声認識とは，単語毎に区切らず，文章として自然に発声された音声を認識するものである．
[2]音声がどの音素に対応するかを示す特徴

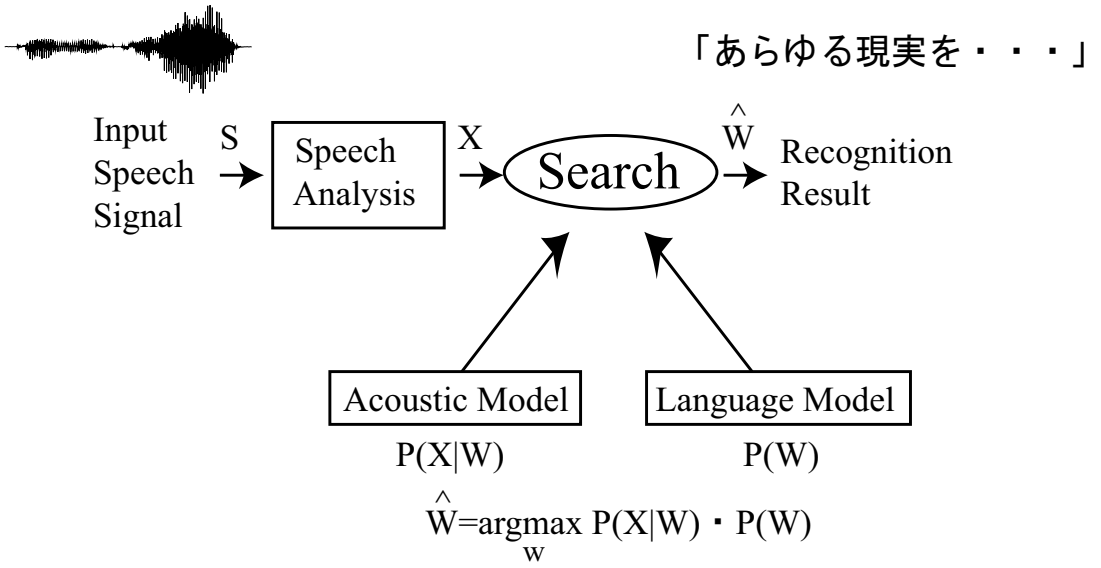

図 2.1: 従来の連続音声認識システムの枠組み

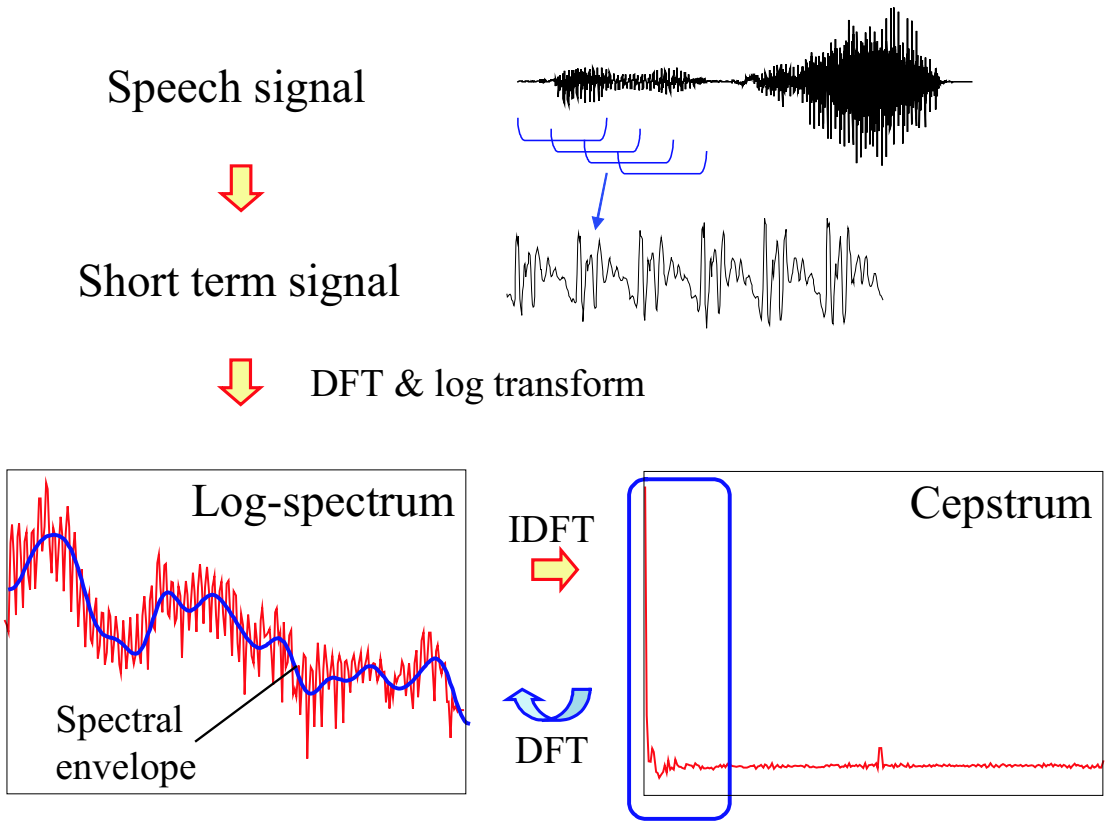

図 2.2: 音声分析

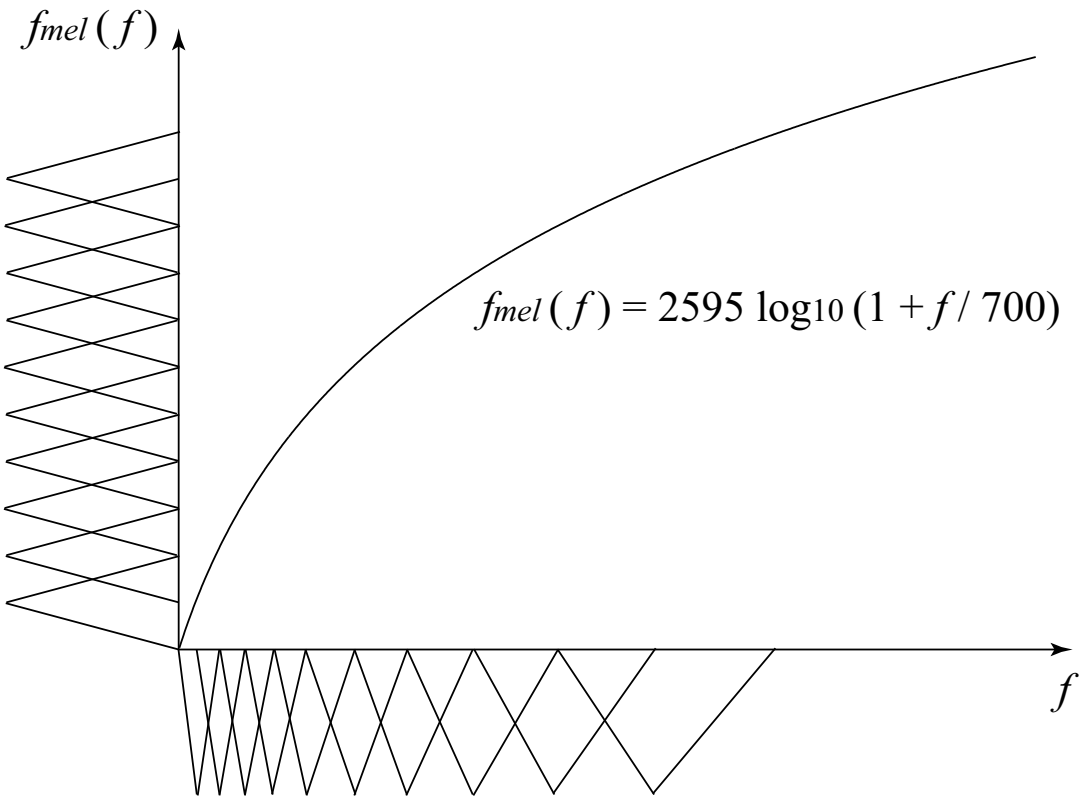

図 2.3: メル周波数とその軸上に等間隔で配置された三角窓

する．その後，対数パワースペクトルに対して逆離散フーリエ変換（Inverse DFT; IDFT）を施したものがケプストラムである．このケプストラムのうちの低次項のみを離散フーリエ変換すると，スペクトル包絡（Spectrum Envelope）が得られる．声道管の共鳴によって強められた周波数をフォルマント周波数と呼び，これらはスペクトル包絡の山の部分におよそ相当するが，音声の音韻的特徴はフォルマント周波数によく表れる．つまりケプストラムは，音声の音韻的特徴を効率良く表すことのできるパラメータである．

2.3.2　メル尺度に基づくケプストラム

人間の音の高さの感覚はメル尺度と呼ばれるが，これは音の周波数に対してほぼ対数に近い特性を示し，人間の周波数分解能は低い周波数ほど細かく，高い周波数ほど粗いことが知られている．これをケプストラムに反映させた音声特徴量が提案されている．MFCC（Mel-Frequency Cepstrum Coefficient）はその一つである．MFCC は，図 2.3 に示すようにメル周波数（メル尺度化された周波数）軸上に等間隔で配置された三角窓を用意し，フィルタバンク分析を行なうことで求められる．尚，メル周波数 f_{mel} は周波数 f[Hz] に対して，

$$f_{mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.3}$$

などの周波数ウォーピングを施すことで得られる．各窓毎に，対応する周波数帯域のパワースペクトルを求め，それに窓の大きさの重みを付けて和をとることでメルスペクトルが得

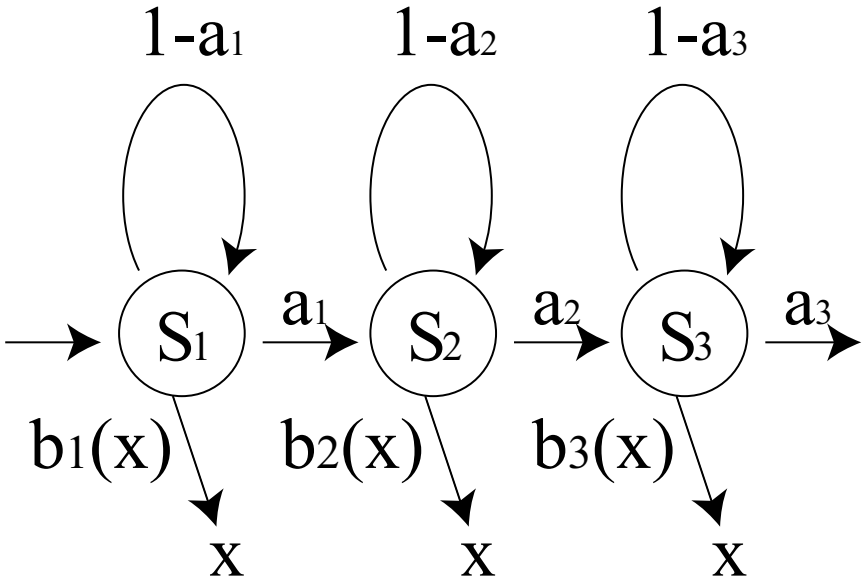

図 2.4: 隠れマルコフモデル（HMM）

られる．これに離散コサイン変換を施すことで，MFCC が求められる．

2.3.3　Δケプストラム

/y/，/w/などの半母音はスペクトルの動きそのものにその音韻的特徴が表れていると考えることができるが，これを考慮し，ケプストラムの時間軸に対する動き（即ちスペクトルの時間変化量）の情報も音声特徴量としてよく用いられる．これは Δ ケプストラムと呼ばれる．従来の音声認識システムでは，上記のメル尺度に基づくケプストラムと，その Δ ケプストラムを合わせたものが，音声特徴量としてよく用いられる．MFCC 及び ΔMFCC がその例である．

2.4　音響モデル

2.4.1　隠れマルコフモデル（HMM）

音響モデルは，ある単語 W が発声されたときに音声特徴量の時系列データ X が出力される確率 $P(X|W)$ を記述するモデルであり，現在では，図 2.4 に示すような隠れマルコフモデル（Hidden Markov Model; HMM）を用いるのが主流となっている．S_i は i 番目の状態，a_i は状態 S_i から状態 S_{i+1} への状態遷移確率，$b_i(x)$ は状態 S_i から音声特徴量 x が出力される出力確率である．出力確率 $b_i(x)$ の分布形としては，ガウス分布を複数用意し，その重み付け和で $b_i(x)$ を表現する「混合ガウス分布」がよく用いられる．

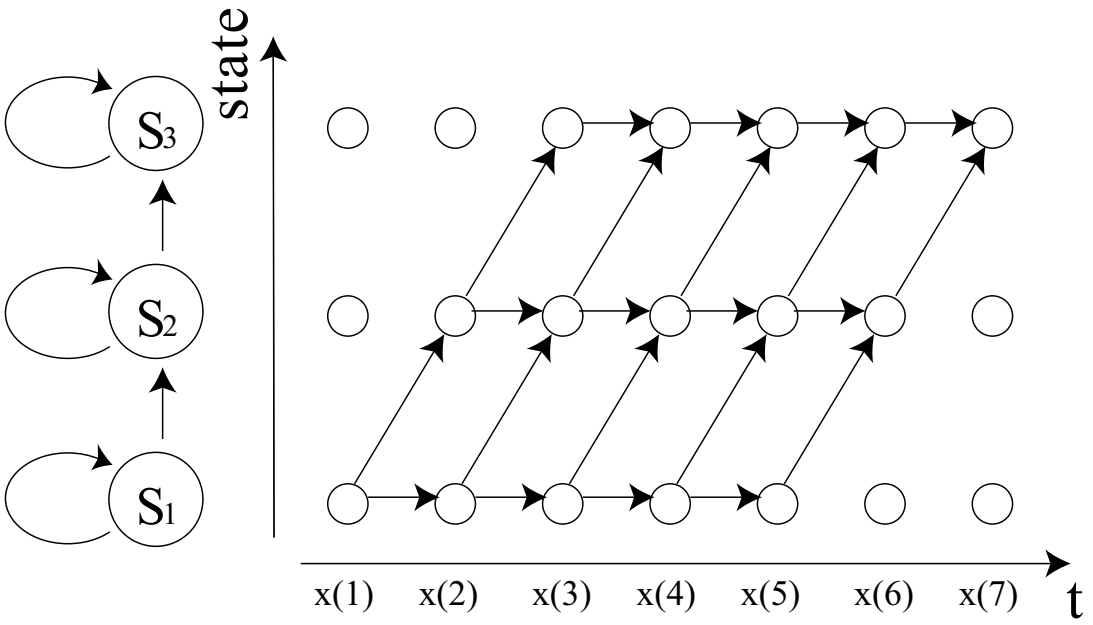

図 2.5: HMM の状態遷移の経路

2.4.2　HMM を用いた音声特徴量の出現確率の計算

HMM を用いて $P(X|W)$ を求める方法について説明する．まずは各単語毎に一つの HMM が割り当てられる場合（単語 HMM）で考える．図 2.5 は，単語 W の HMM から音声特徴量の時系列データ $X = \{x(1), x(2), \cdots, x(7)\}$ が出力される場合の可能な状態遷移の経路を表している．ある 1 つの経路を通って X が出力される確率は，その経路の状態遷移確率 a_i と経路上の各状態での出力確率 $b_i(x)$ の積によって計算できる．図 2.5 に示された経路全てに対してこの確率を求めて和をとることで，音素 W の HMM から音声特徴量 X が出力される確率，即ち $P(X|W)$ を求めることができる．しかし，全ての経路からの出力確率の和をとると計算量が増大してしまうため，実際には最も出力確率の大きな経路のみを計算し，その確率値で $P(X|W)$ を近似する「ビタビアルゴリズム」が用いられる．

また，連続音声認識では通常，音素毎に一つの HMM が割り当てられる（音素 HMM）．この場合は，音素 HMM 同士を次々に結合してから図 2.5 の枠組みを考えることで，単語 W に対する $P(X|W)$ を記述できる．尚，前後の音素環境を考慮しない音素 HMM はモノフォン，前後の音素環境を考慮する音素 HMM はトライフォンと呼ばれる．

2.4.3　HMM の学習

HMM において学習すべきパラメータは $\theta = \{a_i, b_i(x)\}$ であるが，これは最尤（Maximum Likelihood; ML）推定に基づいて行なわれる．即ち，学習データから音声特徴量の時系列データ X が観測されたとき，その尤度を最大化する θ を求める問題に帰着され，

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta) \tag{2.4}$$

が求めるパラメータとなる．しかし，HMM の場合は隠れ変数[3]が存在し，式 (2.4) を解析的に解くのは困難である．このため，実際には式 (2.4) の局所最適解を求める Baum-Welch アルゴリズムが用いられる．

Baum-Welch アルゴリズムでは前向き変数 $\alpha_i(t)$，後向き変数 $\beta_i(t)$ と呼ばれる変数が登場する．これらは，時刻 t における状態が i であれば 1，そうでなければ 0 をとる隠れ変数を z_{ti} とすると，

$$\alpha_i(t) = P(z_{ti} = 1, x_1, \cdots, x_t|\theta) \tag{2.5}$$

$$\beta_i(t) = P(x_{t+1}, \cdots, x_T|z_{ti} = 1, \theta) \tag{2.6}$$

と表すことができるものである．この前向き変数 $\alpha_i(t)$ 及び後向き変数 $\beta_i(t)$ を用いて，時刻 t における状態が i である確率 $\overline{z_{ti}}$ を，

$$\overline{z_{ti}} = P(z_{ti} = 1|X, \theta) \tag{2.7}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{2.8}$$

のようにして求めることができる．上記は，パラメータ θ と学習データからの音声特徴量の時系列データ X を用いれば，そのデータ系列の各々が特定の状態から生じた確率を求めることができることを意味する．式 (2.5) から式 (2.8) を用いれば，新しいパラメータを最尤推定によって求めることができる．例えば，出力確率 $b_i(x)$ の分布形として単一ガウス分布 $\mathcal{N}(x; \mu_i, \sigma_i^2)$ を用いる場合，パラメータ $\theta = \{a_i, \mu_i, \sigma_i^2\}$ に対して，新しいパラメータ $\hat{\theta} = \{\hat{a}_i, \hat{\mu}_i, \hat{\sigma}_i^2\}$ を，

$$\hat{a}_i = \frac{\sum_{t=1}^{T-1} \alpha_i(t) a_i b_i(x_{t+1}) \beta_{i+1}(t+1)}{\sum_{t=1}^{T-1} \alpha_i(t)\beta_i(t)} \tag{2.9}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} \overline{z_{t,i}} x_t}{\sum_{t=1}^{T} \overline{z_{t,i}}} \tag{2.10}$$

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \overline{z_{t,i}} (x_t - \mu_i)^2}{\sum_{t=1}^{T} \overline{z_{t,i}}} \tag{2.11}$$

のようにして求めることができる．式 (2.9) から式 (2.11) の算出には，式 (2.5) から式 (2.8) を求めておく必要があり，一方，式 (2.5) から式 (2.8) の算出には，パラメータを式 (2.9) から式 (2.11) によって求めておく必要があるので，両者は互いに依存関係にある．しかしながら，このようにして得たパラメータ $\hat{\theta}$ は，パラメータ θ に対して常に，

$$P(X|\theta) \leq P(X|\hat{\theta}) \tag{2.12}$$

が成立するので，式 (2.5) から式 (2.8) の算出と，式 (2.9) から式 (2.11) の算出を繰り返す反復アルゴリズムによって，パラメータは局所最適解に収束する．

[3]外部から直接観測することができない変数．HMM の枠組みでは，データ系列 X が観測されたとき，その各々がどの状態から生じたものなのかまでを観測することはできない．

第3章

音声に不可避的に混入する非言語的特徴とそれに対する従来手法

3.1　はじめに

前章では従来の音声認識システムとそれを支える要素技術について説明した．この従来の音声認識技術が抱える大きな問題点として，音声に混入する非言語的特徴の存在が挙げられる．音声認識は，音声から発声された文字列の情報を抽出するタスクであるが，その音声には話者の声道形状の特性，マイクロフォンなどの音響機器の特性，テレビ・ラジオなどの背景雑音といった，音声認識をするにあたって必要の無い非言語的特徴が混入する．この非言語的特徴によって音声の物理的実体（即ち音声自体，スペクトル自体，及びケプストラム自体）は変化する．このため，学習データに用いた音声と，認識対象である入力音声との間で非言語的特徴が異なると（即ち「ミスマッチ」が生じると），音声認識の性能が劣化してしまう．しかも，音声は必ずある話者によって発声されねばならず，何らかのマイクロフォンを用いて収録せねばならない．つまり，話者の声道形状の特性や音響機器の特性は，音声に「不可避的に」混入することになる．

本章では，まず音声に不可避的に混入する非言語的特徴の数学的モデルについて説明する．次に，この非言語的特徴に対する従来手法である不特定話者音響モデル，及び様々な話者適応／話者正規化技術を紹介する．

3.2　音声に不可避的に混入する非言語的特徴の数学的モデル

[3, 4, 5, 6] では，音声に不可避的に混入する非言語的特徴を数学的にモデル化している．以下，それについて説明する．音声に混入する非言語的特徴は主に加算性雑音，乗算性歪み，線形変換性歪みの三種類に分類される．このうち，音声に「不可避的に」混入するものは乗算性歪み，線形変換性歪みの二つである．加算性雑音とは，時間軸上の加算で表現される雑音であり，テレビ・ラジオなどの背景雑音がその典型例と言える．これらは場所を移動するなどして物理的に抹消することができるので，不可避的な雑音ではない．

乗算性歪みは，スペクトルに対する乗算で表現される歪みであり，元のスペクトル特性に対して伝達関数を 1 つかけ合わせたもの（即ちフィルタ）に相当する．これはケプストラムベクトル c に対するベクトル b の加算 $c' = c + b$ に相当する（ケプストラムは対数パワースペクトルの IDFT で定義されるため）．マイクロフォンなどの音響機器の特性がその典型例である．また，乗算性歪みを消失させるために，入力音声のケプストラムからその平均値を減算するケプストラム平均正規化法（Cepstral Mean Normalization; CMN）[8] があるが，これによって話者性の違いによる影響も軽減できる．さらには，入力音声のケプストラムの出現確率を混合ガウス分布で表現した GMM （Gaussian Mixture Models）で話者識別が行なわれることがあるが [9]，これもケプストラムの平均値で話者性を表現できることを意味する．即ち，話者の声道形状の違いの一部も近似的に乗算性歪みとして扱うことができる．前述のとおり，音声は必ずある話者によって発声され，ある音響機器によって収録されるので，これらは不可避的な歪みである．

線形変換性歪みは，c に対する行列 A の乗算 $c' = Ac$ で表現される歪みである．話者の声道長の差異，聴取者の聴覚特性の差異を表すために，工学的には対数スペクトルに対して

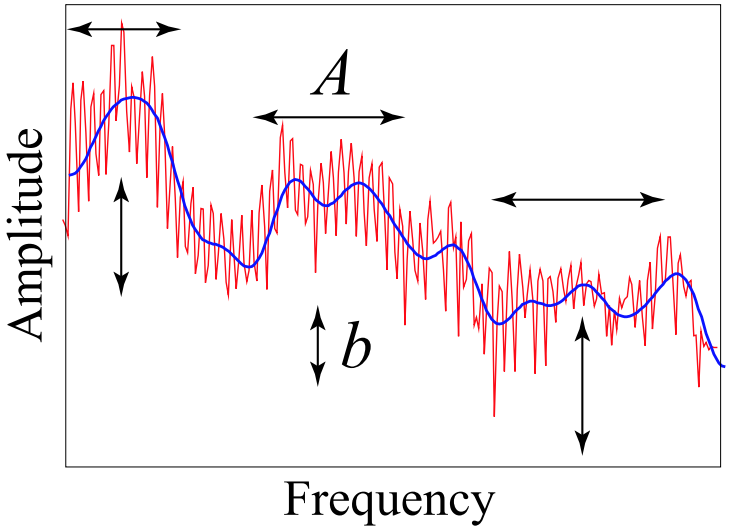

図 3.1: アフィン変換 $Ac + b$ が対数スペクトルに与える影響

周波数ウォーピングが施されるが，単調増加かつ連続である周波数ウォーピングは，c に対する A の乗算で表されることが示されている [10]．即ち，声道長の差異，聴覚特性の差異は近似的に線形変換性歪みとして扱うことができる．これらも不可避的な歪みである．

以上をまとめると，音声に不可避的に混入する非言語的特徴は，ケプストラムベクトル c に対する $c' = Ac + b$ という変換で簡単に表現される．これはアフィン変換と呼ばれる．図 3.1 は，アフィン変換 $Ac + b$ が対数スペクトルに与える影響を示したものである．A は対数スペクトルの水平変化，b は垂直変化を引き起こす．例えば，**第 2.3.1 節**においてフォルマント周波数が音韻的特徴を表すと記したが，共鳴周波数は基本的に音響管の長さに依存するため，フォルマント周波数は性別や年齢（即ち声道長）に依存することになる．

3.3　不特定話者音響モデル

話者が変わればケプストラムも変化する．従来では，この問題に対処するために何百，何千，何万もの話者で学習された音響モデルが構築されてきた．このモデルは不特定話者音響モデル（Speaker Independent HMM; SI-HMM）と呼ばれる．不特定話者音響モデルは，認識対象となる入力音声との話者性の違いをうまく吸収するためのモデルであるが，これを用いても認識性能を劣化させる話者がどうしても存在する．これは，不特定話者音響モデルでは話者性の違いを吸収しきれないことが原因であり，「集める」ことが根本的な解決にはならないことを意味する．尚，不特定話者モデルは入力話者を特定しないことを意図したモデルであるが，逆に特定の入力話者のために構築された音響モデルは特定話者音響モデル（Speaker Dependent HMM; SD-HMM）と呼ばれる．

3.4　話者適応／話者正規化

認識対象となる入力音声との話者性の違いをさらに低減させるために，話者適応や話者正規化と呼ばれる技術が研究されている．話者適応は，得られた入力音声を用いて，不特

定話者音響モデルのパラメータを入力話者に合わせるように変更する，つまり入力話者に適応させる技術である．話者正規化は，これとは逆に，モデルに合わせるように入力話者の特徴量を変更する手法である．ここでは，話者適応技術に必要とされる要件を述べた後，話者適応技術として代表的な MAP 適応，MLLR などを含めた様々な技術を紹介する．

3.4.1　話者適応の要件

話者適応技術の要件は以下の 2 点に集約される [11].

1. 入力音声データ量がごく少数のときでも，優れた性能向上を実現する.（高速適応性）
2. 入力音声データ量が多くなるに従い，入力話者で十分に学習された特定話者モデルの性能に近づく.（最尤推定への漸近性）

まず単純な話者適応法として「追加学習」を考える．これは，入力音声データに対しても学習データの場合と同様に，ML 推定の枠組み（Baum-Welch アルゴリズム）で音響モデルを学習していく，という手法である．この場合，仮に入力音声データが無限に得られるものとすれば，そのモデルは入力話者で学習された特定話者モデルと同等になり，話者性のミスマッチ問題は理論上解決されたことになる（要件 2. が最尤推定への漸近性と呼ばれるのはこれが理由である）.

しかし，この枠組みでは入力データ量が少ないとき，そのデータに含まれない HMM の状態が未適応となる問題（「未観測問題」）が生じる．また，入力データ量が少ないときにパラメータの推定精度が低いという問題（「データ不足問題」）も生じるため，結果として，不特定話者モデルよりかえって認識性能が劣化してしまう．このような問題に対しては，パラメータの少ないモデルを音響モデルとは別個に用意する（これは「適応モデル」と呼ばれる）などの対処が考えられる.

しかし，その一方でパラメータ数の少ない適応モデルを用いると，モデルの単純さが原因で，入力データ量が多くなると認識性能が次第に頭打ちとなり，要件 2. が満たされない問題（「モデル過小問題」）が生じる．一般に，モデルのパラメータ数の大小という観点からすると，要件 1. と要件 2. はトレードオフの関係にあると言える.

3.4.2　MAP 適応

最大事後確率（Maximum A Posteriori; MAP）推定は，入力データ量が少ないとき，ML 推定より頑健なパラメータ推定が行なわれ，かつ入力データ量が多くなるに従い，漸近的に ML 推定に近づくことが数学的に保証される推定方法である.

ML 推定では，入力データ X が得られたとき，パラメータ θ を (2.4) 式により推定する．これに対して，MAP 推定では θ もある確率密度分布に従って分布する確率変数とみなし，X を得た後パラメータが θ である事後確率を最大化する．即ち，

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|X) \tag{3.1}$$

によってパラメータ θ の推定を行なう．ここで，ベイズの定理により式 (3.1) は，次のように変形される．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta)P(\theta) \tag{3.2}$$

ここで，$P(\theta)$ は入力データ X が与えられていない時点での θ の事前分布である．仮に θ に関する知識が何も無ければ，式 (3.2) は式 (2.4) に一致する．即ち，MAP 推定量は ML 推定量と一致する．入力データ X が少ないときは $P(\theta)$ が，多いときは $P(X|\theta)$ が MAP 推定値に大きな影響を与えるようになる．

このような枠組みで話者適応を行なう手法が MAP 適応 [12, 13] である．[12] では出力確率密度分布の分散パラメータとして対角共分散行列を用いた場合の，平均パラメータと分散パラメータの MAP 適応を提案しており，[13] では分散パラメータとして全共分散行列を用いた場合の，HMM の全てのパラメータの MAP 適応を提案している．$P(\theta)$ の設定については，不特定話者モデルのパラメータ値 θ_0 を用いて，入力データ X が得られなかったときの MAP 推定値 $\hat{\theta}$ が不特定話者モデルのパラメータ θ_0 に一致するようにする．こうすることで，MAP 推定値 $\hat{\theta}$ は不特定話者モデルのパラメータ θ_0 から出発し，入力データ量が多くなるにつれて，ML 推定値に近づいていくようになる．

MAP 適応は，話者適応の要件のうち要件 2.（最尤推定への漸近性）は満たしているが，要件 1.（高速適応性）はまだ不十分であると言える．これは，HMM のパラメータ数が多すぎることが原因で，例えば入力データが少ないときに「未観測問題」が生じ，出現しなかった音素環境の HMM には不特定話者モデルのパラメータがそのまま使われることになる．

3.4.3　ケプストラムの平行移動（$c + b$）に基づく話者適応法

MAP 適応は，音響モデルのパラメータを直接推定する適応手法である．これに対して，HMM のガウス分布を変換するためのパラメータを「適応モデル」として用意し，これを推定する話者適応法も提案されている（この手法は「変換写像法」とも呼ばれる）．この手法では，一般に，推定するパラメータ数が MAP 適応法と比較して少なく，その結果として，MAP 適応よりも優れた高速適応性が実現される．

まず，ケプストラムの平行移動（$c + b$）に基づく話者適応法を紹介する．**第 3.2 節**で述べた CMN は，その代表的な話者正規化手法であると言える[1]．この手法は，推定すべきパラメータ b を全ガウス分布で共有したものと捉えることができる．話者性による影響をより低減するには，HMM の各ガウス分布ごとに異なる b を推定するのが望ましいと考えられるが，その場合においては「未観測問題」が生じてしまう．そこで [14] は，事前にガウス分布の分布間距離に基づいてガウス分布群の木構造を求めておき，各ノードにベクトル b を割り当てることで，入力データ量に応じて用いるノードを制御する AMCC（Automatic Model Complexity Control）を提案している．このとき，木構造のルートノードはガウス分布全体を表し，リーフノードは各ガウス分布に 1 対 1 で対応する．従って，入力データが得られるに従って，より詳細な音素環境のベクトル b を用いることができるようになる．

[1] CMN は乗算性歪みを正規化する手法なので，音響機器の特性に対する「環境正規化手法」でもある．

また，[15] ではガウス分布の相関関係ではなく，ベクトル b の相関関係に基づいてガウス分布を木構造化する VFC（Vector Field Correlations）を提案している．

3.4.4　ケプストラムに対するアフィン変換（$Ac+b$）に基づく話者適応法

次に，ケプストラムに対するアフィン変換（$Ac+b$）に基づく話者適応法を紹介する．**第 3.2 節**から分かるように，話者の声道形状の違いはケプストラム c に対するアフィン変換 $Ac+b$ で近似的に表現することができる．アフィン変換による話者適応法として最も代表的なのが MLLR（Maximum Likelihood Linear Regression）[16] である．その扱い易さと性能の良さから，MLLR は広く使われてきた．

MLLR は，HMM の各ガウス分布の平均ベクトル μ にアフィン変換 $A\mu+b$ を施すことで行なわれる．A と b は ML 推定の枠組みで求められる．全ガウス分布で一つの A，b を共有する場合は，入力データ量が少ないときに良い性能を示すが，入力データ量が増加すると「モデル過小問題」が生じる．これは大局的なアフィン変換 $Ac+b$ では話者性を表現する能力が限られていることを意味する．この問題を解決するために，[16] では HMM のガウス分布を複数のグループに分け，同一グループ内で A と b を共有することを試みている．但し，グループ数を増やし過ぎると「適応モデル」のパラメータ数が増大するため，今度は「データ不足問題」が生じてしまう．[16] ではこれに対処するため，行列 A として対角共分散行列を用いることも検討している．

上記は平均ベクトルのみを適応する MLLR であるが，分散共分散行列も合わせて適応する手法も提案されている [17, 18]．[17] は，ガウス分布 $\mathcal{N}(x;\mu,\Sigma)$ を $\mathcal{N}(x;A\mu+b,A\Sigma A^T)$ と変換する手法を提案しており，制約付き MLLR（constrained MLLR）と呼ばれる．一方，[18] はより厳密な適応を行なうため，分散共分散行列 Σ に対しては平均ベクトル μ とは別の変換行列 H を用意し，$\mathcal{N}(x;A\mu+b,H\Sigma H^T)$ と変換する手法を提案している．これは制約無し MLLR（unconstrained MLLR）と呼ばれる．但し，制約無し MLLR の方が制約有り MLLR よりもパラメータ数は多い．

また [19] は，MLLR における A と b を各グループ毎に複数用意し，その重み付け和で適応を行なう MLST（Maximum Likelihood Stochastic Transformation）を提案し，さらなる性能向上を実現させている．MLLR と MLST の関係は，ガウス分布と混合ガウス分布の関係と同値であると解釈できる．近年では，周波数ウォーピングを表現するように A を推定する APT（All-Pass Transform）が提案され，MLLR に対する性能向上を実現させている [20]．**第 3.2 節**で単調増加かつ連続である周波数ウォーピングが Ac で表されることを述べたが，その逆は必ずしも成立しない．つまり，任意の A が周波数ウォーピングを表現するとは限らないために，上記のような結果が得られたのではないかと筆者は解釈している．

3.4.5　Eigenvoice

Eigenvoice[21] は，顔画像認識で用いられている主成分分析手法 EigenFace[22] の考え方を基に提案されたもので，主成分分析（Principal Component Analysis; PCA）を用いて，効率的なパラメータ数の削減を行なうものである．

まず，N 個の特定話者モデル（及び 1 個の不特定話者モデル）を用意する．各特定話者モデル毎に，HMM の全ガウス分布の平均ベクトルを繋げ合わせた話者ベクトル（supervector）を作成する．状態数を s，混合数を m とすると，話者ベクトルの次元数は $D = s \times m$ である．次に，得られた N 個の話者ベクトルに対してそれぞれ主成分分析を行なうことで，N 個の固有ベクトル

$$e(j) = (e_1(j), ..., e_D(j))^T \quad (j = 0, ..., N-1) \tag{3.3}$$

を求める．$e(0)$ は，話者ベクトルの全話者に対する平均に相当し，$e(1), ..., e(N-1)$ は，分散への貢献度順に整列された主成分である．高次元の固有ベクトルを捨て，$K+1$ 次元の固有ベクトル $e(0), ..., e(K)$ を残すことで，パラメータ数の削減を行なう．

残された固有ベクトルを用いて，入力話者の話者ベクトル P を，

$$P = e(0) + \sum_{j=1}^{K} w(j) \times e(j) \tag{3.4}$$

と表すことを考える．このとき，各固有ベクトルの重み $w(j)$ $(j = 1, ..., K)$ は ML 推定によって求められる．パラメータ数は K だが，これは D よりはるかに小さいので，少量のデータでも十分な推定が可能となる．また，分散共分散行列と遷移確率においては不特定話者モデルのものを用いる．Eigenvoice は MLLR よりもさらに優れた高速適応性を持つが，入力データ量が増加すると「モデル過小問題」が生じる．

3.4.6　MAP 適応，MLLR，及び Eigenvoice の組み合わせ手法

ここまでに紹介した MAP 適応，MLLR，及び Eigenvoice の入力データ量に応じた性能の違いを，不特定話者モデル及び（十分に学習された）特定話者モデルの性能と合わせて簡単に図示すると，図 3.2 のようになる．話者適応の要件は高速適応性及び最尤推定への漸近性であるが，この 3 つの手法はいずれもこの 2 つの要件を同時に満たすことができない．

そこで，これらを組み合わせた手法が提案されている．例えば，MLLR で適応を行なった平均ベクトルを，MAP 適応における事前分布に用いる手法が提案されている [25]．これは，入力データ量が多いときに最尤推定への漸近性が保証され，かつ MAP 適応法よりも優れた高速適応性を持つ手法である．認識実験の結果，入力データ量に依らず，MLLR 単独，MAP 適応単独のときよりも良い性能を実現している．また，Eigenvoice 提案時，同様にして Eigenvoice で適応を行なった後に MAP 適応を行なう試みもなされており，入力データ量が多いときの性能向上を実現させている [21]．

さらには，MLLR における変換行列をあらかじめ多数の学習話者に対して求め，それら変換行列に対して Eigenvoice と同様の手法でパラメータ数削減を行なうことで，Eigenvoice と MLLR を組み合わせる EigenMLLR[26] が提案されており，入力データ量が少ないときに MLLR よりも高い認識性能を実現させている．Eigenvoice と比較すると，Eigenvoice の固有ベクトルより，EigenMLLR の固有行列の方がパラメータ数が少ないので，学習話者のデータ量が少なくて済むなどの利点がある．

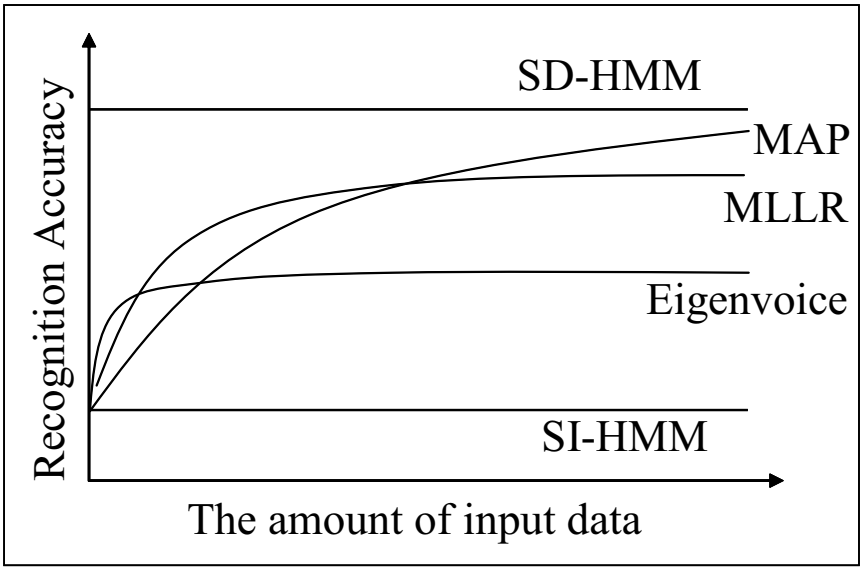

図 3.2: ３つの話者適応法の入力データ量に応じた性能の違い

3.4.7　ヤコビ適応法を用いた加算性雑音，乗算性歪み，及び話者の声道長に対する同時適応

[23] は，背景雑音に対する高速な適応手法であるヤコビ適応法を提案している．[24] はこれを拡張して，背景雑音，音響機器の特性，及び話者の声道長に対して同時に，かつ高速に適応する手法を提案している．

上記の非言語的特徴に対する同時適応は非線形問題となる[2]が，ヤコビ適応法では，Taylor 展開の一次項を用いて音響空間の近傍を線形近似する．C_S，C_N，C_H をそれぞれ入力話者のクリーン音声，加算性雑音，乗算性歪み（スペクトル領域における伝達関数）に対応したケプストラムとし，λ を話者の声道長伸縮係数とする．C_Y を観測された入力話者のケプストラムとし，

$$C_Y = \Psi(C_S, \lambda, C_N, C_H) \tag{3.5}$$

という関数で表されると仮定する．λ，C_N，C_H は時間が経つにつれて僅かずつ変化するものと仮定し，その変動分を $\Delta\lambda$，ΔC_N，ΔC_H とすると，それによる C_Y の変動分 ΔC_Y は，

$$\Delta C_Y = J_\lambda \Delta\lambda + J_N \Delta C_N + \Delta C_H \tag{3.6}$$

と表せる．ここで J_λ，J_N はそれぞれ声道長に関するヤコビベクトル，雑音に関するヤコビ行列であり，$J_\lambda = \partial C_Y / \partial\lambda$，$J_N = \partial C_Y / \partial C_N$ である．J_λ，J_N は入力データが観測される前に，あらかじめ求めておくことができる．観測された入力音声から，それに対応する HMM の状態毎に ΔC_Y を求める．そこから最小自乗誤差法を用いて $\Delta\lambda$，ΔC_N，ΔC_H を推定し，式 (3.6) を用いて HMM の全ガウス分布の平均値を更新することで適応を行なう．

このような枠組みで [24] は認識実験を行ない，その効果を示している．但し，話者の声道長に対する適応の効果は，背景雑音及び音響機器の特性に対する効果と比べると小さい．

[2]例えば，**第 7.3.1 節**では加算性雑音がケプストラムに対して非線形変換を施すことについて論じる．

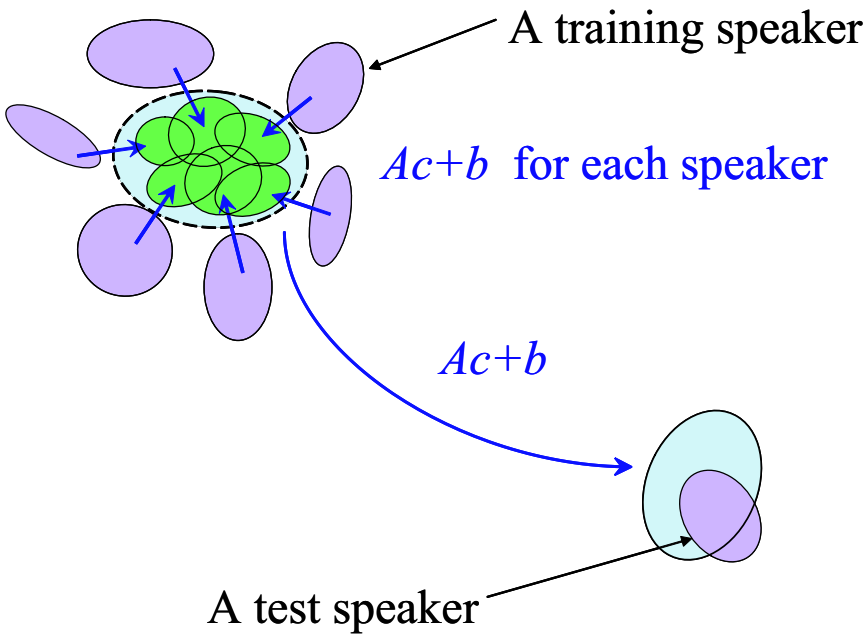

図 3.3: SAT 及び MLLR を用いた話者適応

3.4.8　学習話者の正規化・選択

MAP 適応や MLLR などの話者適応技術は，あらかじめ学習した不特定話者モデルを入力話者に適応するものであるが，そもそも不特定話者モデルは，入力話者と学習話者との話者性を違いを吸収するために構築されたものである．従って，話者適応を行なう場合においては，音響モデルはもはや不特定話者モデルである必要がない，と考えることができる．しかしながら，音響モデルの学習に十分な大量の音声データを一人の話者に発声させるのは，その話者にとって大変な労力を必要とする話である．

SAT（Speaker Adaptive Training）[27] は，複数の話者の音声データを得た後，各話者毎にアフィン変換 $Ac + b$ を施し，その上で音響モデルの学習を行ない，架空の話者による特定話者モデルを作り上げる．このとき，学習するパラメータは架空話者の HMM のパラメータの他に，各話者に対する変換パラメータ A 及び b があるが，これらは ML 推定の枠組みで学習される．このような音響モデルを作成することで，学習話者の話者性の差異に起因する分散を小さくすることができ，適応時の認識性能を向上させることができる．[27] では，SAT を用いて作成した音響モデルと不特定話者モデルの 2 つに対して MLLR による話者適応を施し，その性能を比較することで SAT による効果を確認している．SAT 後，MLLR を用いて話者適応を行なう様子を図 3.3 に示す．近年では，前述の EigenMLLR に SAT を組み合わせることでもその効果が確認されている [28].

一方，複数の学習話者のうち，入力話者に音響的に近い話者セットを求めた後，その話者セット内における各話者のモデルに対して MLLR を行なう手法も提案されており，入力データ量が少ないときに特に効果を上げている [29].

第4章

音声の構造的表象

4.1　はじめに

前章では，音声にはケプストラム c に対するアフィン変換 $Ac + b$ で表現される非言語的特徴が不可避的に混入し，これに対して様々な適応・正規化技術が提案されていることを述べた．しかしながら，これらの何れの技術も，モデル（0）を入力音声（1）に近づける，もしくは入力音声（1）をモデル（0）に近づけるものに過ぎず，完全に 1 もしくは 0 にすることはできない．これは音声の物理的実体をそのままモデル化しているため，非言語的特徴を表現する次元が残留しているからである．近年提案されている音響的普遍構造 [3, 4, 5, 6] は，非言語的特徴を表現する次元そのものを消失させる技術である．

本章では，この音響的普遍構造について，その言語学的背景である構造音韻論から説明を行なう．また本手法との関連として，音声の相対関係に基づく他の研究例も紹介する．

4.2　構造音韻論

近代言語学の祖であるソシュールは言語に対して，“Language is a system of only conceptual differences and phonic differences.” “What defines a linguistic element, conceptual or phonic, is the relation in which it stands to the other elements in the linguistic system.” “The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from the others.” などと主張している [30]．このソシュールの言語哲学に啓蒙され，ヤコブソンらは「構造音韻論」と呼ばれる言語学の一分野を確立した．これは，弁別素性[1]を用いて 2 つの音素の違いや，全ての音素間差異によって構成される幾何学的構造を議論する学問分野である．例えば，ヤコブソンはフランス語の母音と準母音を弁別素性を用いて立体的に構造化している（図 4.1）[31]．音声認識において音素を弁別素性の束とみなし，弁別素性に着眼した研究例 [32, 33, 34] もあるが，これらは全て音声事象の「絶対的特性」を捉えている．しかし，上記から分かるように弁別素性はもともと音的差異を表現するために作られたものであり，構造音韻論は同一の幾何学的構造が話者を問わず普遍的に存在すると主張する．**第 3.2 節**では話者性などの非言語的特徴の数学的モデルについて説明した．音響的普遍構造は，この非言語的特徴に対して不変な幾何学的構造として提案されたもので，これは構造音韻論の物理実装に相当する．

4.3　音声に内在する音響的普遍構造

各音声事象（例えば各音素）を分布化し，N 個の分布によって構成される構造を考える．N 個の分布に対して ${}_NC_2$ 個の全ての二分布間距離を求めれば，一つの構造を規定したことになる．音声に不可避的に混入する非言語的特徴は，アフィン変換 $Ac + b$ で表現される（**第 3.2 節**）．これに対して不変な構造を抽出することが構造音韻論の物理実装のための条件と考えられるが，アフィン変換は構造を歪ませる変換である．このため，不変な構造は「空間」を歪ませることで抽出される．

[1]音素と音素を区別するために用いる音声的特徴．例えば/b/と/p/は有声／無声によって区別される．

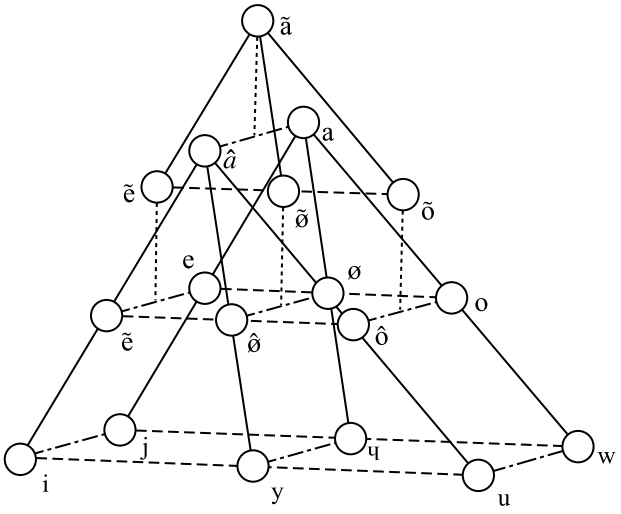

図 4.1: ヤコブソンによるフランス語の幾何学的音韻構造

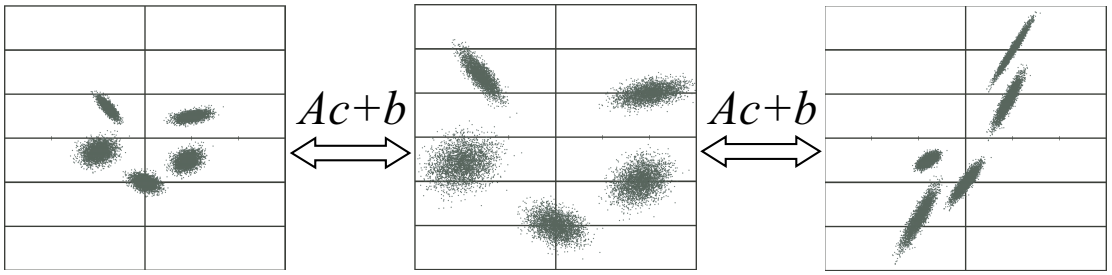

図 4.2: 構造不変の定理（これらが全て同一の構造となる）

構造不変の定理 [5] ：意味のある記述が分布としてのみ可能な物理現象を考える．分布群に対して，全ての二分布間距離を求める（距離行列）．二分布間距離として，バタチャリヤ距離，カルバック・ライブラ距離，ヘリンガー距離などを用いた場合，各分布に対して単一の任意一次変換を施しても，二分布間距離は不変である．即ち距離行列は不変であり，その結果，構造も不変となる（図 4.2 参照）．

以下，バタチャリヤ距離を用いて話を進める．二つの分布の確率密度関数をそれぞれ $p_1(x)$，$p_2(x)$ とすると，バタチャリヤ距離は以下の式で表される．

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \tag{4.1}$$

$0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leq 1$ を確率として解釈すれば，これは自己情報量となり，単位は [bit] となる．二つの分布がガウス分布で表現されているとき，バタチャリヤ距離は，

$$BD(p_1(x), p_2(x)) = \frac{1}{8}\mu_{12}^T \left(\frac{\sum_1 + \sum_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2}\ln \frac{|(\sum_1 + \sum_2)/2|}{|\sum_1|^{\frac{1}{2}}|\sum_2|^{\frac{1}{2}}} \tag{4.2}$$

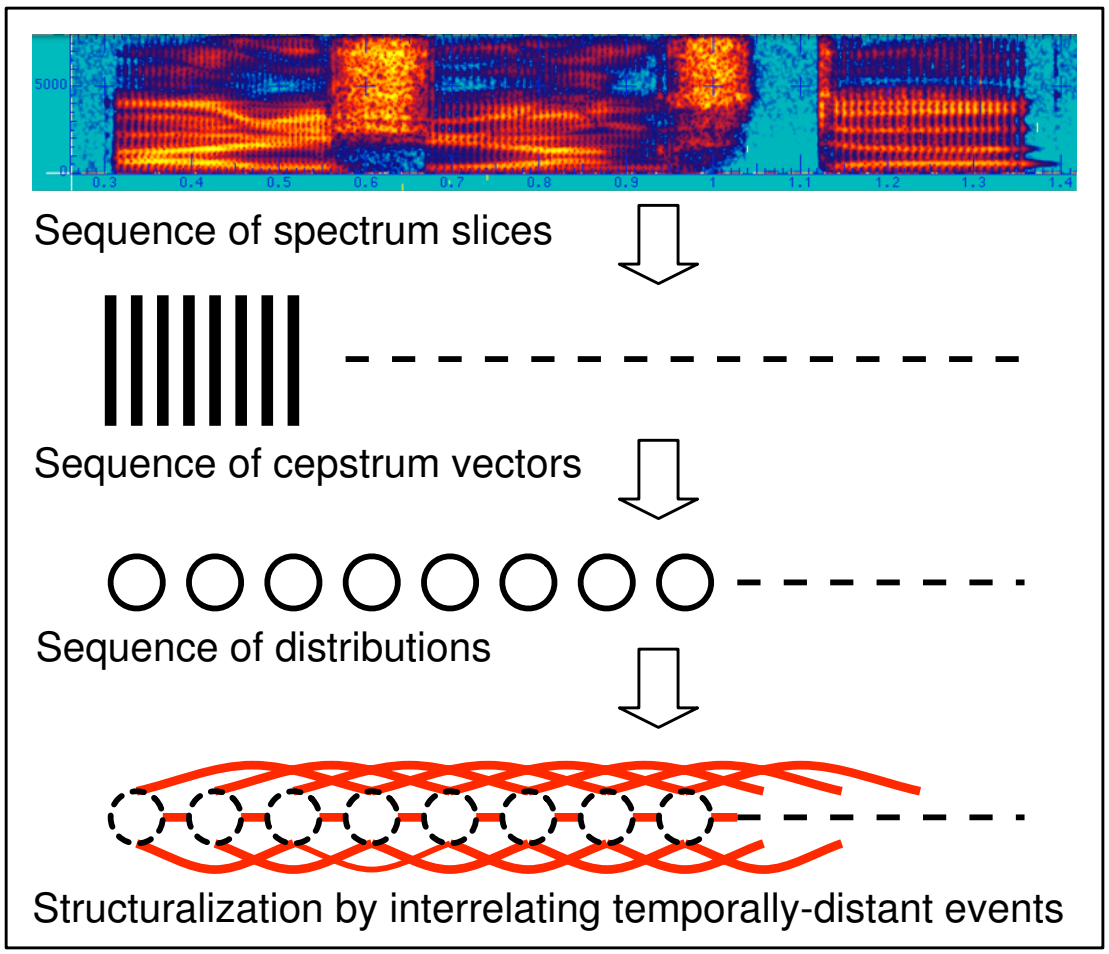

図 4.3: 一発声の構造化

となる．μ_{12} は $\mu_1 - \mu_2$ である．このとき，二つの分布に対して共通のアフィン変換 $Ac + b$ をかけた場合，バタチャリヤ距離はその前後で不変である．これは，バタチャリヤ距離が空間を歪める距離尺度であることに起因する．MLLR[16] や SAT[27] では，話者性はアフィン変換で記述されるが，この構造はアフィン変換に対して不変である．c に A を掛ける演算は構造の回転として観測され，b を加える演算は構造のシフトとして観測される．この構造が音響的普遍構造と呼ばれているものである．尚，[6] は音響的普遍構造の接点として，**第 4.2 節**で述べた言語学の他，心理学，言語障害学，神経生理学，脳科学，及び音楽学の観点からこの構造を考察している．

4.4　一発声の構造化と構造に基づく音響的照合

音声認識は一発声された音声を対象として扱うが，音声からケプストラム系列を求め，そこから音声事象分布（ケプストラム分布）の系列を得た後，任意の二分布間距離を求めれば一発声の構造化も可能である（図 4.3）．次に，二つの構造の構造間差異を求めること

で，求めた構造を音響的に照合することを考える．M 個の頂点（$P_1, ..., P_M, Q_1, ..., Q_M$）で構成される二つの構造において，構造 Q をシフト（b）と回転（A）のみで構造 P に近づけ，対応する頂点間距離の和（$\sum_{i=1}^{M} \overline{P_iQ_i}^2$）の最小値を求めることで構造間差異を定義する．即ち，図 4.4 に示されるような枠組みの音響的照合である．二つの構造が N 次元ユークリッド空間内にある場合，その構造間差異は以下の式によって導出される．

$$\sum_{i=1}^{M} \overline{OP_i}^2 + \overline{OQ_i}^2 - 2\sum_{i=1}^{N} \sqrt{\alpha_i}, \tag{4.3}$$

O は両構造の重心である（構造をシフトさせて重心を重ねる）．α_i は N 次正方行列 S^tTT^tS の固有値である．S は行列（$\overrightarrow{OP_1}, ..., \overrightarrow{OP_M}$）であり，$T$ は行列（$\overrightarrow{OP_1}, ..., \overrightarrow{OP_M}$）である．しかしながら，音響的普遍構造は空間を歪ませることで得られるため，ユークリッド空間内には存在しない．従って，三角不等式が満たされない可能性があり，直接式 (4.3) を用いることは出来ない．ここで，分布間距離としてバタチャリヤ距離の平方根を用いた場合，シフト（b）及び回転（A）後の $\sum |\theta_i|$（$|\theta_i| = \angle P_iOQ_i$）が十分に小さければ，

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j} - \overline{Q_iQ_j})^2} \approx \sqrt{\frac{1}{M}\sum_{i}(\overline{OP_i} - \overline{OQ_i})^2} \tag{4.4}$$

$$\approx \sqrt{\frac{1}{M}\sum_{i}\overline{P_iQ_i}^2} \tag{4.5}$$

という近似式が成立することが示されている [35]．式 (4.4) の左辺は距離行列[2]のうち意味を持つ上三角成分をベクトル（これを「構造ベクトル」と定義する）として並べたときのユークリッド距離に対応する．従って，構造に基づく音響的照合は，距離行列のみを用いて近似的に行なうことができる．以上より，音響的普遍構造を用いた音声認識が可能であることが示唆される．**第 6 章**から**第 8 章**では，それぞれ日本語の孤立 5 母音系列音声，連続 5 母音系列音声を構造を用いて認識する枠組みについて述べる．

4.5　音声の相対関係に基づく他の研究例

4.5.1　母音間の差分ベクトルを用いた母音系列の認識

音声の相対関係に基づく研究例は他にもある．例えば [36] では，入力音声から観測される 2 母音をそれぞれ特徴ベクトル x_p, x_q で代表し，その相対関係であるベクトル $\Delta_{pq} = x_p - x_q$ を認識に利用している．入力として 2 母音が与えられた場合，そこから Δ_{pq} を算出し，その 2 母音が母音組 (v_i, v_j) である確からしさを求めることを考える．まず，母音組 (v_i, v_j) の特徴ベクトルをそれぞれ x_{v_i}, x_{v_j} として $\Delta v_iv_j = x_{v_i} - x_{v_j}$ としたとき，あらかじめ用意していたデータから，Δv_iv_j の平均ベクトル $\bar{\Delta} v_iv_j$，分散共分散行列 Σv_iv_j を求めておく．次に，

[2]分布間距離によって構成される行列．例えば，構造 P の距離行列の (i, j) 成分には $\overline{P_iP_j}$ が格納される．

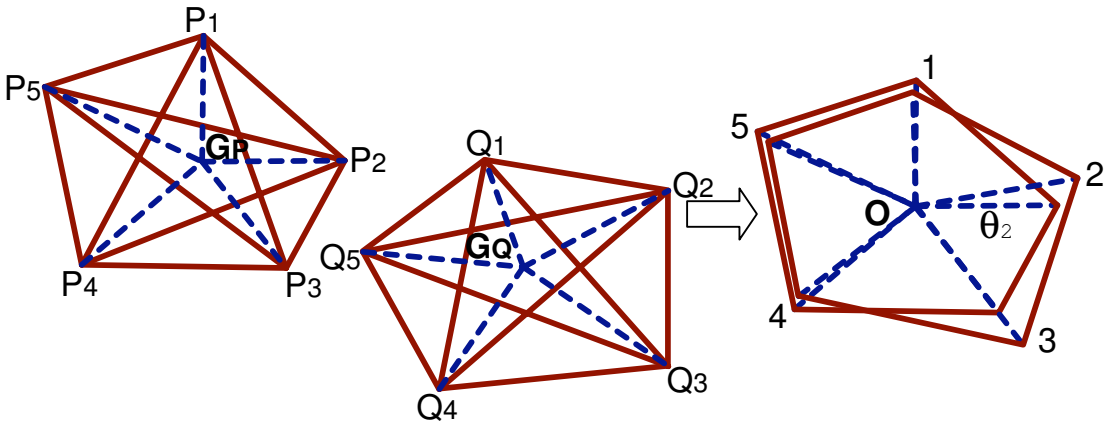

図 4.4: 構造に基づく音響的照合

以下に定義される，母音組 (v_i, v_j) に対する特徴ベクトル (x_p, x_q) の整合度 $Rv_iv_j(x_p, x_q)$ を求める．

$$\begin{aligned} & Rv_iv_j(x_p, x_q) \\ = \; & -(\Delta_{pq} - \bar{\Delta}v_iv_j)^t\Sigma^{-1}v_iv_j(\Delta_{pq} - \bar{\Delta}v_iv_j) \\ & -\log|\Sigma v_iv_j| \end{aligned} \tag{4.6}$$

これは正規分布に対する対数尤度に相当する．長さ n の入力母音系列に対しては，ここから長さ 2 の任意の部分列（${}_nC_2$ 個）を取り出し，式 (4.6) の整合度を求めた後，それらの和としてその系列の整合度 $R(c_s)$ を求めている．

本研究と良く似た研究であると言えるが，音声を構造化した場合はその構造サイズを正規化することで調音努力[3]を正規化することができる [37] という利点を持っている（**第 6.6 節**で，その効果を認識実験結果に基づいて論じる）．また，[36] ではベクトル（絶対値成分及び方向成分）を捉えているため，(構造の回転として観測される）声道長の違いに対しては，構造と比較すると頑健ではないと考えられる．但し本研究においては，図 4.4 のような構造の回転とシフトに基づく音響的照合を行なった場合，本来別の単語であるものが，構造の回転によって偶然同一の単語と見做される可能性がある．これは従来手法との融合が必要不可欠であることを示唆するものと考えられる．

4.5.2　COSMOS 法

[38, 39] は複数の音響モデル間の任意の相互距離を求め，その後多次元尺度法を用いて複数の音響モデルを二次元空間上にマッピングする COSMOS 法を提案している．[38] ではこれを雑音環境下の音声認識に，[39] では複数の音声コーパスの俯瞰的分析に利用している．

このとき，2 つの音響モデルの相互距離としては，対応する音素環境の HMM の，対応する状態における出力確率密度分布の分布間距離を次々に算出し，その重み付け和（HMM

[3]個々の音声事象を他と明確に区別するように発声する努力

の重みとして例えば出現頻度を用いる）を求めることで定義されている．出力確率密度分布の分布間距離については，対応する混合ガウス分布からそれぞれ任意のガウス分布を取り出し，ガウス分布間のバタチャリヤ距離を次々に算出していき，混合重みで重み付け和をとることで求めている．従って，この手法も音響的普遍構造との共通性が多く見られる．音響的普遍構造では，それを構成する単位が音声事象分布であるのに対し，COSMOS 法ではその単位は一つの音響モデルであると言うことができる．

第5章

音声の構造化による非言語的特徴の消失に関する定量的分析

5.1　はじめに

前章では，音声に不可避的に混入する非言語的特徴に対して，それを表現する次元を持たない音声表象として提案されている音響的普遍構造について説明した．本研究は，この構造を用いた音声認識を目的とするが，本章ではそれに先立って，音声の構造化により非言語的特徴がどの程度消失されるのかを調べることとする．具体的には，非言語的特徴として話者性に着眼し，音声の構造化によるその消失度合いを定量的に分析する．

5.2　孤立５母音系列の収録

まずは孤立５母音系列の収録を行ない，成人男性 8 名，女性 7 名に対して孤立５母音を 5 回発声させた．この際，調音努力が構造サイズに影響を与えてしまうので [37]，以下に示す方法で発話スタイルの統一を試みた．まず孤立５母音のサンプル音声をスピーカー提示し，その直後に，発話スタイル，母音の継続長などを真似る形で発声させた．尚，F_0 については話者にとって自然な F_0 で，かつ一定となるよう指示した．最終的に，各話者の 5 つの５母音構造を取り出し，そのサイズが極端に大きい・小さい話者を除いて，男女 4 名ずつを分析対象とした．

5.3　話者性の消失に関する定量的分析

5.3.1　実験条件

8 名の話者の 5 つの構造に対して，話者間構造差異（${}_8C_2 \times 5^2 = 700$ 個），話者内構造差異（$8 \times_5 C_2 = 80$ 個）を求め，両者の比較を通して話者性の消失の度合いを検討した．このときの分析条件を表 5.1 に示す．尚，各母音の音声事象分布（ケプストラム分布）は，母音中心部の前後 14 フレーム（140msec）を用いて推定した．

構造差異の定義としては，(I) 構造のシフト＆回転を行なわずに式 (4.5) を用いる方法と，(II) シフト＆回転後の構造差異近似式である式 (4.4) 左辺を用いる方法を試みた．また，音声を構造化した場合，構造サイズの正規化を行なった上で二つの構造を比較することも可能となる．これは調音努力の正規化を意味するので [37]，構造を音声認識に利用しようとする場合などにおいて，特に有効と考えられる．そこで，(II') 各構造のサイズを平均サイズと等しくなるよう正規化した後の式 (4.4) 左辺による構造差異も検討した．

それぞれ話者間差異と話者内差異の平均を求め，話者間・話者内差異に対する分散分析を行ない，その有意差を検定した．

5.3.2　実験結果

結果を表 5.2 に示す．(I) と (II) を比較すると，(II) の方が話者間差異が小さく，5%水準で有意差がないことが分かる．これは，音声の構造化によって話者性の消失が行なわれて

表 5.1: 音響的条件（第 5.3 節）

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	FFTcep.（1〜12 次元）
音声事象分布	単一ガウス分布（全共分散行列）

表 5.2: 分析実験の結果

構造差異の導出方法	(I)	(II)	(II')
話者間差異の平均	13.8	3.92	3.29
話者内差異の平均	4.99	3.68	3.04
危険率 (p)	0.0%	13.9%	1.97%

いることを示す．また，正規化を行なわない方が，より話者性が消失されることが (II) と (II') を比較することで分かる．これは，正規化を行なわない場合，構造サイズがばらつくので，話者間・話者内差異のばらつきが共に大きくなるのが原因と見られる．

次に話者内差異に注目してみる．同一話者内では $A = E$ と $b = \vec{0}$ であると仮定すると，話者内差異は (I) と (II) で等しくなると考えられる．しかし表 5.2 を見ると，(II) の方が (I) より小さい．これは，式 (4.4) が $\sum |\theta_i|$ が十分に小さいことを前提とした近似式であることが原因の一つだと考えている．$|\theta_i|$ が大きい場合，$|\overline{P_iO} - \overline{Q_iO}| < \overline{P_iQ_i}$ が成立し，その結果として，式 (4.4) 左辺は構造差異を小さく見積もる傾向がある．

話者性の消失度合いを示す異なる指標として,「全話者間差異のうち，話者内差異の平均を下回るものの割合」を (II)，(II') に対して求めてみると 48.1%，39.4%であった．両分布が同一である場合 50%であることを考慮すると，(II) は話者性がよく消失していると考えられる．また，(II') は有意差があると判定されたものの，(I) に比べれば構造差異は十分小さく，話者間差異の約 4 割は話者内差異の平均を下回る．

ここで，話者間差異 < 話者内差異を満たす男性話者と女性話者の構造，及び同男性話者の二つの構造を，多次元尺度法により二次元空間にマッピングした時の図を図 5.1，図 5.2 に示す．話者性の消失に対する構造化の有効性を示す一例であると言える．

5.3.3　分析条件による変動

第 5.3.2 節では，シフト長 10msec，分析対象区間 140msec という条件であったが，同一音声区間に対して，シフト長を短くすることでフレーム数を上げる，即ちより詳細に音響事象を捉えようとすると，話者性の消失度合いはどのように変動するのかを調べてみた．

(II) について，シフト長を 10msec，5msec，2msec とした場合の検定結果を表 5.3 に示す．シフト長を短くしてフレーム数を上げる，即ち分析分解能を上げると，有意差が大きくなることが分かる．アフィン変換 $Ac + b$ は音声に不可避的に混入する非言語的特徴に対する

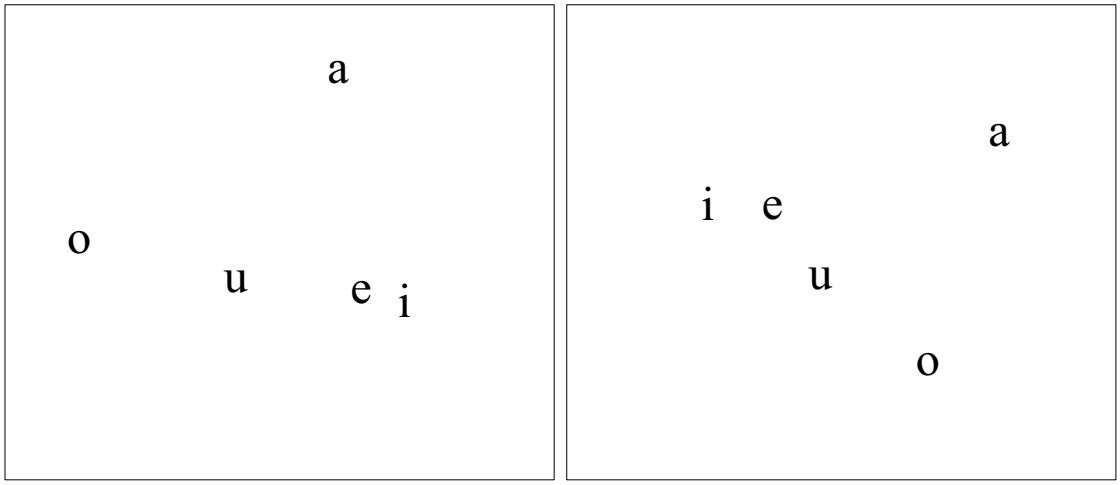

図 5.1: 異なる話者（男性と女性）の二つの構造

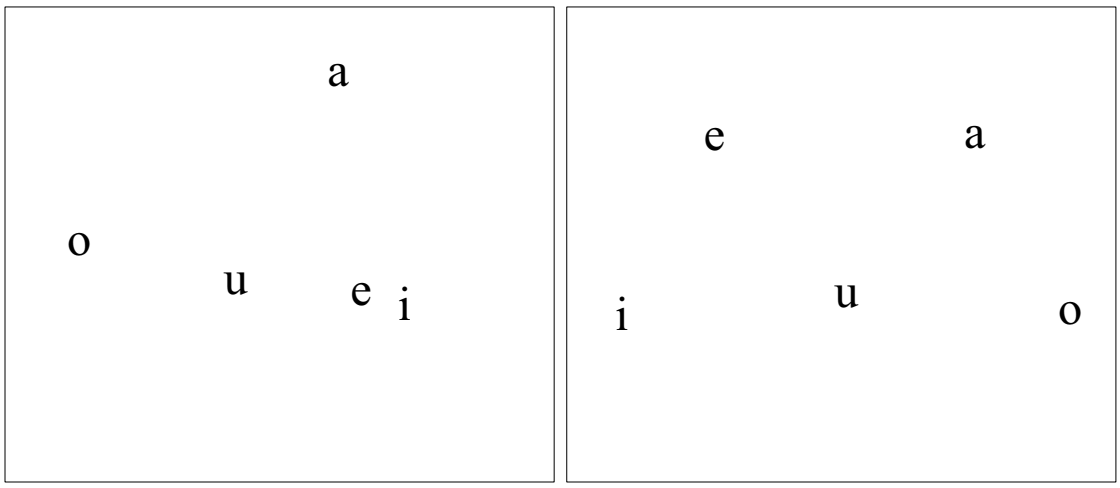

図 5.2: 同一話者（男性）の二つの構造

表 5.3: 分析条件による有意差検定結果の変動

シフト長	フレーム数	危険率 (p)
10msec	14	13.9%
5msec	28	2.78×10^{-6}%
2msec	70	1.06×10^{-9}%

簡素なモデルであり，構造化による非言語的特徴の消失の効果は限られている可能性がある．また，ここで音声の構造化によって消失される話者性は，話者の声道形状の特性による「音の話者性」であり，各話者特有の発声の癖，方言差などの「構造の話者性」，及び不可避な「構造揺らぎ」は依然として構造に反映される．表 5.3 の結果は，これらに起因するものと考えることができる．分析分解能を下げることでこれらを非明瞭化することは可能であるが，この場合は，話者間差異と同時に音韻差異までも不明瞭になる恐れがある．

第6章

音声の構造的表象を用いた孤立５母音系列音声認識

6.1　はじめに

本章から，構造を用いた音声認識に関する基礎的研究として，その認識の枠組みや認識実験について述べていく．本章ではまず，簡単な認識タスクとして孤立５母音系列の音声認識を考え，これを構造のみを用いて認識する枠組みを述べる．この際，i) 一発声から構造化する場合は音声事象分布推定のためのデータ量が少ない，ii) アフィン変換 $Ac + b$ は簡素なモデルであるため，非言語的特徴を表現する能力が限られている，といった問題が生じる．これらの問題に対しては，それぞれ音声事象分布の最大事後確率（MAP）推定，スペクトル高域成分の除去という手法を導入した．

そして，本章ではこれに基づいて行なった，クリーン環境下における認識実験を述べる．ここで，一発声の構造化によって，音声に不可避的に混入する非言語的特徴が十分に消失されるのであれば，以下の 3 つの質問について考えたい．

- 音声の物理的実体を明示的に用いない音声認識は可能だろうか？
- 一人の話者で学習された音響モデル（構造モデル）を用いた不特定話者音声認識は可能だろうか？
- 適応・正規化技術を一切用いない不特定話者音声認識は可能だろうか？

この 3 つの質問に対する答えを求めるべく，認識実験を行なった．構造を用いた認識実験においては，全帯域を用いる場合と，LPF（ローパスフィルタ）を用いて高域成分を除去した場合の両方において実験を行ない，構造サイズを正規化することによる効果も検討した．また，従来手法との比較実験も行なった．

6.2　音声の構造的表象を用いた孤立５母音系列音声認識の枠組み

まず，認識タスクとして孤立的に発声された日本語母音系列の音声認識を考える．これは，/a/, /i/, /u/, /e/, /o/の各母音が一回ずつ孤立的に発声されたものを一つの単語とみなし（語彙サイズ ${}_5P_5 = 120$），それを認識するタスクである．これを，構造を用いて認識する枠組みを図 6.1 に示す．

まず入力音声からから各母音の音声事象分布（ケプストラム分布）を求め，これを構造化する．このとき，構造サイズ（構造の大きさ）が一定値となるよう正規化する．[37] は，構造サイズが調音努力（発話スタイル）を表すことを実験的に示している．従って，構造サイズの正規化は音声認識において有効であると考えられる．構造として実際に求めるのは距離行列であり，このうち意味を持つ成分は上三角成分であるので，これをベクトルとして並べた「構造ベクトル」（10 次元）を特徴ベクトルとして用いる．この特徴ベクトルには個々の母音を絶対的，個別的に同定するために必要な特徴量は存在しない．

音声認識システムに持たせる構造モデルは以下のように作成する．複数の/a/-/i/-/u/-/e/-/o/の構造ベクトルから 10 次元ガウス分布（全共分散行列を使用）を求め，これを/a/-/i/-/u/-/e/-/o/の「構造統計モデル」とする．他の 119 個（/i/-/a/-/u/-/e/-/o/など）の構

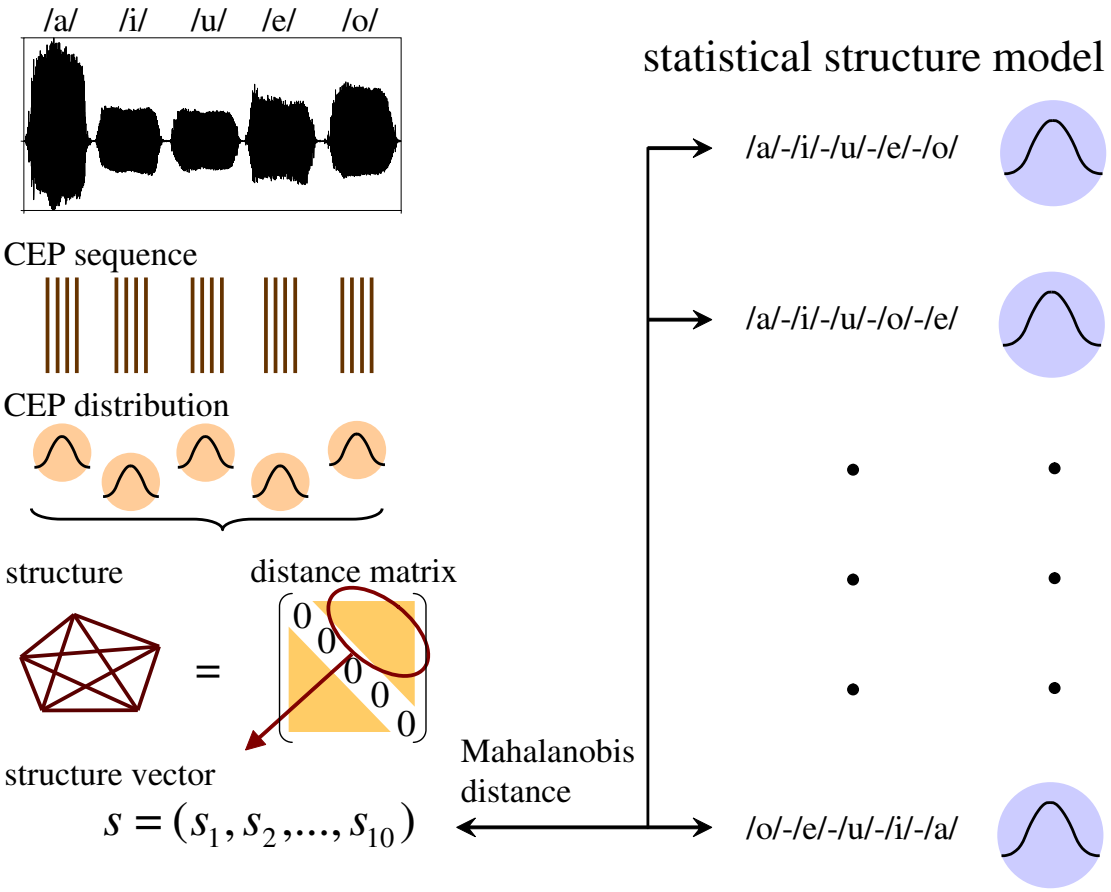

図 6.1: 構造を用いた孤立 5 母音系列の音声認識の枠組み

造統計モデルは，/a/-/i/-/u/-/e/-/o/の構造統計モデルの要素を交換することで得られる．最終的に 120 個の構造統計モデルが得られ，これを認識に利用する．

構造の音響的照合は，入力構造ベクトルと各構造統計モデルとのマハラノビス距離を算出することで行なう．これは，**第 4.4 節**において述べたシフト（b）と回転（A）に基づく音響的照合の近似を，構造と構造統計モデルの間で行なうことに相当する．この距離が最も小さい単語を認識結果とする．

6.3　より頑健な構造化のために

6.3.1　音声事象分布の最大事後確率推定

音響的普遍構造を音声認識に利用する場合，一発声された音声から音声事象分布を推定する必要がある．最尤（ML）推定は分布の推定手法として広く用いられているが，得られるデータ量 n が少ないときに不適切な分布を推定する可能性がある．従って，一発声を構造化する本研究においては，この問題が顕著となる．

そこで，音声事象分布の最大事後確率（MAP）推定を検討する．MAP 推定の具体的な

枠組みに関しては [12] を参照した．以下，分散共分散行列は全て対角である．また，ここでは孤立発声された日本語母音系列を認識対象として扱うので，各母音（/a/, /i/, /u/, /e/, /o/）の孤立発声を複数用意し，これを事前知識として用いる．これらは一発声毎にガウス分布化される（計 M 個）．MAP 推定に用いるパラメータは以下の通りである．

$$
\begin{aligned}
\mu_m &: m \text{ 番目の発声の平均ベクトル} \\
\Sigma_m &: m \text{ 番目の発声の対角共分散行列} \\
\mu_0 &: \{\mu_m\} \text{ の平均 } (= \tfrac{1}{M}\textstyle\sum_{m=1}^{M} \mu_m) \\
\Sigma_0 &: \{\Sigma_m\} \text{ の平均 } (= \tfrac{1}{M}\textstyle\sum_{m=1}^{M} \Sigma_m) \\
S_\mu &: \{\mu_m\} \text{ の対角共分散行列} \\
&\quad (= \tfrac{1}{M}\textstyle\sum_{m=1}^{M} (\mathrm{DIAG}(\mu_m - \mu_0))^2) \\
\Omega &: = \Sigma_0 S_\mu^{-1} \\
\mu_{ML} &: \text{入力発声の平均ベクトル（ML 推定）} \\
\Sigma_{ML} &: \text{入力発声の対角共分散行列（ML 推定）}
\end{aligned}
$$

ここで，DIAG(x) は，ベクトル x の要素を対角成分に並べた対角共分散行列である．これらを用いて，MAP 推定では入力発声の分布を以下のように推定する．

$$\mu_{MAP} = \hat{\mu_0} \tag{6.1}$$

$$\Sigma_{MAP} = \hat{B}\hat{A}^{-1} \tag{6.2}$$

ここで，

$$\hat{\mu_0} = \Omega(\Omega + nE)^{-1}\mu_0 + n(\Omega + nE)^{-1}\mu_{ML} \tag{6.3}$$

$$\hat{B} = B + \frac{n}{2}\Sigma_{ML} + \frac{n}{2}\Omega(\mathrm{DIAG}(\mu_{ML} - \mu_0))^2(\Omega + nE)^{-1} \tag{6.4}$$

$$B = E \tag{6.5}$$

$$\hat{A} = A + \frac{n}{2}E \tag{6.6}$$

$$A = \Sigma_0^{-1} \tag{6.7}$$

である．μ_{MAP} は μ_0 と μ_{ML} の内挿値をとり，n の増加につれて μ_{ML} に近づく．本研究では各母音毎に中心前後 14 フレームが用いられたので，本来 $n = 14$ であるが，この値を変化させて入力発声の事前知識に対する重みを調節することが可能である．音声事象分布の MAP 推定の様子を図 6.2 に示す．

6.3.2　スペクトル高域成分の均一化

本研究では，音声に不可避的に混入する非言語的特徴をアフィン変換 $Ac + b$ で表現しているが，これは簡素なモデルであるため，音響的普遍構造が非言語的特徴を消失させる効

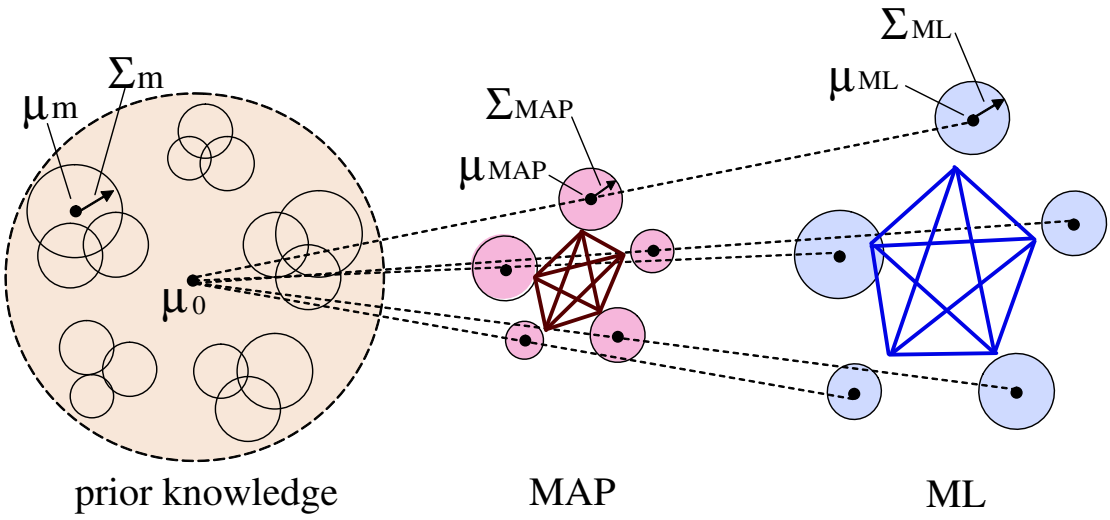

図 6.2: 音声事象分布の最大事後確率推定

果は限られている可能性がある．[40] は，母音のスペクトル包絡の 2.2kHz 以上の帯域には話者性の情報が多く含まれていることを実験的に示している．これに基づいて，話者性をより効果的に消失させるために，本章では音声に LPF （ローパスフィルタ）を通すことでスペクトル広域成分を均一化させることを試みた．

図 6.3 に，5 名の話者が発声した/a/のスペクトル包絡を示す．上・中央・下の図は，それぞれクリーン音声・LPF（カットオフ周波数：2kHz）を施した音声・白色雑音（SNR=10[dB]）を重畳した音声に対応する．話者による違いが高域によく表れているが，LPF を通すことでそれが（下に）揃えられ，話者性の消失が効果的に行なわれていることが分かる．尚，白色雑音の重畳によってもスペクトル高域成分が（上に）揃えられている．**第 7 章**では音声に雑音を重畳することで，スペクトル広域成分を均一化させることも試みている．

但し，実際に図 6.3 中央及び下の音声を聞いてみると，話者性は完全には消失されていないことが分かる．

6.4　全帯域を用いた構造的表象に基づく認識実験

6.4.1　実験条件

本節から，実際に行なった認識実験について説明する．まずは構造化の際に全帯域を用いる認識実験を行なった．用いた音声資料は，**第 5.2 節**で収録して分析対象とした，男性 4 名，女性 4 名の計 8 名の話者による，孤立 5 母音 5 回発声のデータである．ケプストラムとしては MCEP（α=0.55）（1〜12 次元）を使用し，各母音のケプストラム分布は中心前後 14 フレーム（140msec）を用いて推定した．分布の推定方法は，ML 推定，MAP 推定の両方を試みた．各話者毎に，3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の構造ベクトルを抽出し，計 25,000（$= 8 \times 3,125$）個の構造ベクトルを入力に用いた．尚，他の母音列について

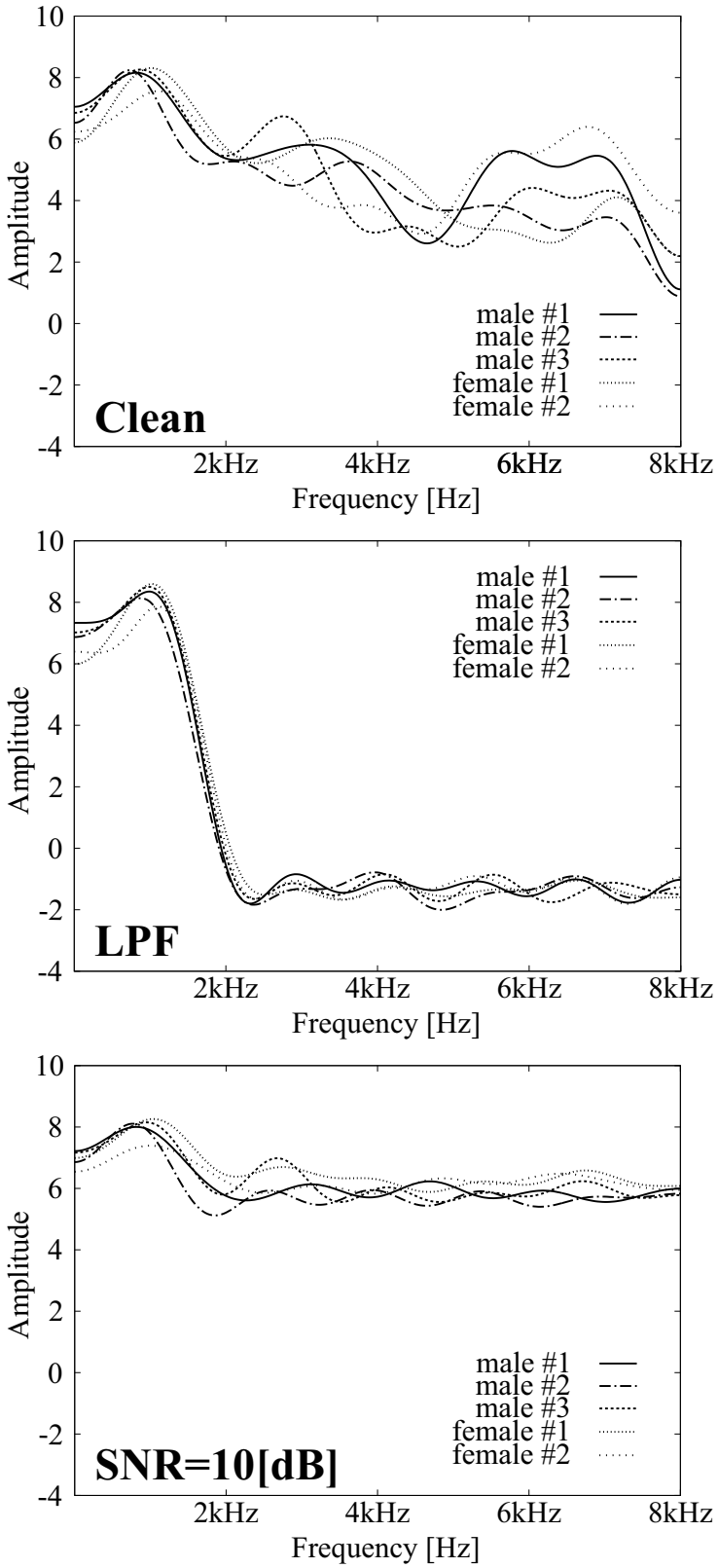

図 6.3: 5 名話者の/a/のスペクトル包絡

表 6.1: 音響的条件（第 6.4 節）

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	MCEP（α=0.55）（1〜12 次元）
音声事象分布	単一ガウス分布（対角共分散行列）
分布推定方法	ML or MAP

は，/a/-/i/-/u/-/e/-/o/の要素を交換することで得られ，認識結果も同様に入れ換わるだけであるため，入力音声には用いなかった．

各構造統計モデルの学習には，評価話者を除く 7 名によって作成される計 21,875 個（$= 7 \times 5^5$）の構造ベクトルを用い，混合数 1，2，4 のガウス分布を推定した．MAP 推定における事前知識のためのデータとしては，評価話者の音声事象に対しては学習話者の全母音音声（$7 \times 5 \times 5$ 個）を使用した．学習話者の音声事象に対しては，評価話者と当該話者を除く 6 名の全母音音声（$6 \times 5 \times 5$ 個）を使用した．ここでは，MAP 推定時の入力音声の事前知識に対する重み n として種々の値を使用して検討した（$n = \infty$ で ML 推定）．このときの音響的条件を表 6.1 に示す．

6.4.2　実験結果

実験結果を表 6.2 に示す．第 1 位以外に，2 位，4 位，5 位，10 位の正解率も示している．まず，構造統計モデルとして単一ガウス分布を用いた場合を見ると，ML 推定の場合，認識率は約 36%となっている．本実験では chance level は 0.8%（$= 1/120$）であることを考えると，ML 推定による音声事象分布群が成す構造にも語彙同定のための情報が含まれていることが分かる．さらに，MAP 推定を施すことで認識率は飛躍的に向上する．特に，入力音声の事前知識に対する重み n を小さくするほど認識率が向上する．これは，各音声事象の少量サンプルデータを用いて，事前分布を僅かに修正することで得られる分布群が張る構造には，語彙同定のための情報が多く含まれていることを意味する（但し，$n = 0$ にすると，事象分布が事前分布に等しくなり，距離行列の成分は全て 0 となり，認識不能になる）．一方，構造統計モデルの混合数の増加は，認識結果にあまり影響を与えていない．これは，構造統計モデルには全共分散行列を用いたため，混合数 1 でも十分なモデル化能力が得られていたと考えられる．また，混合ガウス分布を用いた場合，MAP 推定を施すことで認識率が低下しているところがあるが，第 2 位までを正解とした場合には，高い認識率が得られている．第 2 位までを正解として含めると，どの場合も認識率が飛躍的に向上するが（最大で約 99%），これは誤認識結果として/a/-/i/-/e/-/u/-/o/が多かったことが原因である．**第 6.4.3 節**で，/a/-/i/-/e/-/u/-/o/も正解に含めた場合の認識実験も行なう．

6.4.3　１母音を既知とした場合の認識実験

構造的表象のみを用いた認識実験では，音声の物理的実体を全く用いていない．そこで，5 母音のうち 1 母音を既知とした場合の認識実験を行なった．1 母音を既知とすることで，

表 6.2: 全帯域を用いた構造による認識結果

構造統計モデル＝単一ガウス分布

推定法＼順位	1	2	5	10
ML	35.6%	55.0%	80.3%	92.5%
MAP(n=10)	41.1%	82.1%	94.3%	100.0%
MAP(n=1)	41.5%	83.0%	94.9%	100.0%
MAP(n=0.1)	45.5%	90.7%	99.4%	100.0%
MAP(n=0.01)	68.7%	99.4%	100.0%	100.0%

構造統計モデル＝混合ガウス分布（混合数 2）

推定法＼順位	1	2	5	10
ML	36.4%	55.5%	79.2%	90.9%
MAP(n=10)	26.2%	79.6%	98.0%	100.0%
MAP(n=1)	28.7%	80.9%	98.2%	100.0%
MAP(n=0.1)	34.7%	90.4%	99.5%	100.0%
MAP(n=0.01)	65.2%	98.6%	100.0%	100.0%

構造統計モデル＝混合ガウス分布（混合数 4）

推定法＼順位	1	2	5	10
ML	34.8%	52.7%	80.4%	93.2%
MAP(n=10)	31.8%	72.6%	90.1%	98.5%
MAP(n=1)	33.5%	73.5%	89.9%	98.1%
MAP(n=0.1)	25.0%	85.9%	99.2%	100.0%
MAP(n=0.01)	55.2%	87.3%	90.9%	98.9%

構造のシフト・回転の自由度を削減し，認識率は向上するものと予想される．結果を表 6.3 に示す．/u/または/e/を既知とした場合の認識率が高いが，これは前述したように，誤認識結果として/a/-/i/-/e/-/u/-/o/が多かったためである．表 6.4 に，/a/-/i/-/e/-/u/-/o/も正解に含めた場合の認識結果を示す．音声の実体を直接用いず，音声事象分布を推定する際に全帯域を用いた場合においても，ほぼ 100%の性能で語彙を 2/120 まで削減できている．

6.5　高域成分を除去した場合の認識実験

6.5.1　実験条件

音声に LPF を通してスペクトル高域成分を除去した場合の実験も行なった．カットオフ周波数は 2kHz，4kHz，8kHz（全帯域）の 3 種類を試みた．ケプストラムとしては MCEP（α=0.55）（1〜12 次元）を用いた．また，高域成分を用いない構造化によって，仮に非言

表 6.3: 1 母音を既知とした時の認識結果

構造統計モデル＝単一ガウス分布

推定法＼既知母音	a	i	u	e	o
ML	48.7%	41.7%	66.5%	60.9%	49.2%
MAP(n=10)	55.8%	53.3%	92.9%	80.5%	55.8%
MAP(n=1)	55.9%	53.5%	93.5%	81.4%	55.9%
MAP(n=0.1)	54.3%	52.8%	97.3%	88.8%	54.3%
MAP(n=0.01)	68.7%	68.7%	99.9%	100.0%	68.7%

構造統計モデル＝混合ガウス分布（混合数 2）

推定法＼既知母音	a	i	u	e	o
ML	52.3%	43.2%	62.9%	59.6%	53.5%
MAP(n=10)	39.3%	38.7%	90.4%	77.3%	39.3%
MAP(n=1)	41.0%	40.5%	91.2%	78.5%	41.0%
MAP(n=0.1)	40.5%	39.2%	96.3%	88.5%	40.5%
MAP(n=0.01)	65.4%	65.2%	99.6%	99.8%	65.4%

構造統計モデル＝混合ガウス分布（混合数 4）

推定法＼既知母音	a	i	u	e	o
ML	49.7%	52.5%	63.8%	58.9%	50.5%
MAP(n=10)	41.3%	37.3%	87.1%	73.4%	41.3%
MAP(n=1)	40.8%	37.7%	88.6%	75.2%	40.8%
MAP(n=0.1)	28.2%	27.5%	95.9%	88.9%	28.2%
MAP(n=0.01)	56.7%	56.4%	92.1%	100.0%	55.8%

表 6.4: /a/-/i/-/e/-/u/-/o/を含む認識結果

推定法＼混合数	1	2	4
ML	54.6%	52.9%	53.4%
MAP(n=10)	83.1%	81.0%	79.7%
MAP(n=1)	83.6%	81.6%	81.1%
MAP(n=0.1)	89.4%	90.7%	91.5%
MAP(n=0.01)	99.9%	99.5%	88.9%

表 6.5: 音響的条件（第 6.5 節）

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	MCEP（α=0.55）（1〜12 次元）
音声事象分布	単一ガウス分布（対角共分散行列）
分布推定方法	ML or MAP
カットオフ周波数	2kHz, 4kHz, or full-band

表 6.6: LPF を用いた構造による認識結果

推定法＼帯域	full-band	4kHz	2kHz
ML	24.7 %	47.9%	86.8%
MAP(n=10)	42.9 %	62.7%	100.0%
MAP(n=1)	42.6 %	62.1%	100.0%
MAP(n=0.1)	45.7 %	60.8%	99.9%
MAP(n=0.01)	70.3 %	65.4%	96.7%

語的特徴が完全に消失できるのであれば，構造統計モデルの学習に用いる話者は 1 名で済むはずである．従って，ここでは構造統計モデルの学習に，男性 1 名による孤立５母音 35 回ずつの発声を用いた．これらを 7 つのグループ（1 グループにつき，孤立５母音 5 回発声）に分割した．各グループに対し，3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の構造ベクトルを抽出し，これを基に構造統計モデルを作成した．MAP 推定に用いる事前知識は，評価話者の音声事象に対しては，7 グループ内の全母音データ（$7 \times 5 \times 5$ 個）を用いた．学習話者の音声事象に対しては，当該グループを除く 6 グループ内の全母音データ（$6 \times 5 \times 5$ 個）を用いた．このときの音響的条件を表 6.5 に示す．

6.5.2　実験結果

構造統計モデルとして単一ガウス分布を用いた場合の認識結果を表 6.6 に示す．認識率は，高域成分除去を施すことで飛躍的に向上する．その際，MAP 推定の重み n を小さくすることによる効果が見られなくなったのは，ML 推定の認識率が向上したためと見られる．注目すべきは，カットオフ周波数が 2kHz の時に MAP 推定を用いることで，100%の認識率が得られている点である．これは，本章の冒頭で述べた，

- 音声の物理的実体を明示的に用いない音声認識
- 一人の話者で学習された音響モデル（構造モデル）を用いた不特定話者音声認識
- 適応・正規化技術を一切用いない不特定話者音声認識

が，今回の認識タスクでは，100%の性能を以っていずれも実現可能であることを意味する．尚，**第 6.6 節**でも示すが，これは重み n が本来の値である n=14 のときでも成立する．

表 6.7: 構造サイズの正規化の有無による性能比較

学習話者＝７名

推定法＼帯域	full-band	4kHz	2kHz
ML（正規化無し）	29.4%	35.6%	83.9%
ML（正規化有り）	35.6%	43.2%	82.2%
MAP（n=14，正規化無し）	43.3%	70.4%	99.8%
MAP（n=14，正規化有り）	41.1%	63.7%	99.8%

学習話者＝１名

推定法＼帯域	full-band	4kHz	2kHz
ML（正規化無し）	11.6%	26.0%	68.8%
ML（正規化有り）	24.7%	47.9%	86.8%
MAP（n=14，正規化無し）	41.5%	47.9%	99.4%
MAP（n=14，正規化有り）	43.0%	62.8%	100.0%

6.6　構造サイズの正規化による効果

構造サイズの正規化を行なう場合と行なわない場合の実験結果の比較も行なった．このとき，音響的条件は表 6.5 のまま，学習話者を評価話者を除く 7 名（**第 6.4.1 節**と同じ）にした場合と，男性話者 1 名（**第 6.5.1 節**と同じ）にした場合で，それぞれ実験を行なった．また，MAP 推定における重み n は，本来の値である n=14 と設定し，構造統計モデルとしては単一ガウス分布を用いた．結果を表 6.7 に示す．

表 6.7 から，構造サイズの正規化による効果を確認することができる．特に，学習話者 1 名のときに構造サイズの正規化による効果がよく見られる．これは，学習話者が 1 名しかいない場合，評価話者との調音努力のミスマッチが大きくなってしまうため，その正規化による効果が表れたものと考えられる．また，ML 推定のときに効果が大きいことも分かる．これは，ML 推定の場合は抽出される構造が不安定であるため，結果として構造サイズも不安定になったのが原因であると考えられる．

6.7　従来手法との比較実験

比較実験のための従来の音響モデルとして，まずは 2 種類の不特定話者音響モデルを用意した．一つは学習話者 4,130 名の混合共有 HMM，もう一つは学習話者 260 名の状態共有 HMM である（共に全帯域を用いて学習）．さらに，**第 6.5.1 節**の男性話者 1 名の音声データにカットオフ周波数 2kHz の LPF を施し，それを用いて学習した音響モデルも用意した（これを用意した理由については後述する）．この 3 つの音響モデルのいずれに対しても，CMN による話者・環境の正規化を行なった．特徴パラメータとしては MFCC（1〜

表 6.8: 孤立５母音系列に対する 4 つの手法の性能比較

手法＼評価音声の帯域	full-band	4kHz	2kHz
full band HMM(260)	100.0%	93.8%	72.3%
full band HMM(4,130)	100.0%	95.2%	87.5%
limited band HMM(1)	88.8%	88.8%	88.8%
Proposed(1)	100.0%	100.0%	100.0%

12 次元），ΔMFCC（1〜12 次元），及び ΔE を用いた（計 25 次元）．言語的制約としては，120 単語のみを許容する文脈自由文法を用いた．

表 6.8 に，4 種類の手法（提案手法，3 つの音響モデルを用いた従来手法）による認識実験結果を示す．括弧内の数字は音響モデル（構造モデル）の学習時に使用した話者数である．提案手法においては，2kHz までの低域成分を持つ入力音声であれば，LPF を通すことで，2kHz 以上の高域成分を除去した構造統計モデルの条件と合わせることができる．従って，LPF を特徴抽出の一部と考えれば，表 6.8 の全ての場合において認識率は 100%である．

全帯域を用いて学習された不特定話者音響モデルの場合，入力音声が全帯域の場合，100%の認識率を実現しているが，LPF が施された入力音声に対しては，CMN を施しているにも拘らず認識性能が劣化している．従って，この認識タスクにおいては，1 人の話者で学習された提案手法が，4,130 人の話者で学習された従来手法より良い性能を示している．但し，より厳密な比較実験を行なうには，従来の音響モデルを 2kHz 以上の高域成分を除去した音声で学習させる必要がある．

これが，もう一つの音響モデルを用意した理由である．3 つ目の音響モデルは 2kHz までの帯域で学習されており，さらには提案手法と同じ学習データを用いて作成されているため，より厳密な比較実験であると言える．この場合においても，提案手法は従来手法より上回る性能が得られた．この従来手法について，各評価話者に対する認識性能を調べたところ，全ての誤認識は 2 名の女性話者によるものであった（79.6%と 31.2%）．これは，カットオフ周波数 2kHz の LPF では，話者性が完全には消失されないことを示唆する．**第 6.3.2 節**でも述べたとおり，カットオフ周波数 2kHz の LPF を通した音声を聞いてみると，話者性は確かに完全には消失されていないことが分かる．その残りの話者性は構造化によって消失され，その結果として提案手法は 100%の性能を示した，と解釈することもできる．

ここまでをまとめると，提案手法は，音声事象分布の MAP 推定，LPF を用いた高域成分除去，そして構造サイズの正規化を行なうことで，学習話者 1 名で 100%の認識性能を実現することに成功した．さらには，比較実験の結果，学習話者 4,130 名を含むいずれの従来手法をも上回る性能が得られた．ここでの認識タスクは，まだ非常に単純なものではあるものの，本章で述べた認識実験の結果は，提案手法の非常に高いポテンシャルを示したものと言うことができる．提案手法では各母音を認識することなく，母音同士のコントラストのみを捉え，全体を通してその単語を認識している，ということも注目すべき点である．

第7章

雑音環境下における音声の構造的表象を用いた孤立５母音系列音声認識

7.1　はじめに

前章では，構造を用いたクリーン環境下における孤立 5 母音系列の認識実験について報告した．そこでは，学習話者 1 名の提案手法で 100%の認識性能が実現された．本章では，さらに雑音環境下における孤立 5 母音系列の認識実験について報告する．**第 3.2 節**において，加算性雑音は物理的な抹消が可能という意味において不可避的ではないと述べたが，実際は，例えばカーナビゲーションシステムに音声認識機能を搭載することを考えたとき，背景雑音の音声認識への影響は大きな問題となる．音響的普遍構造は，乗算性歪み及び線形変換性歪みに対して不変なものであったが，加算性雑音に対しては構造の形状が歪むと考えられる．しかし同時に，母音のスペクトル高域成分に多く含まれる話者性の情報 [40] を消失させる効果を持つことが予想される．

本章ではまず，加算性雑音に対する従来手法について整理し，そのうち認識実験で使用するスペクトルサブトラクション（SS）について，より詳細に説明する．次に，加算性雑音による構造の歪み，及びその際の話者性の消失についての定量的分析の結果を報告する．その後，構造を用いた雑音下の日本語母音系列の認識実験について，従来手法との比較実験を含めた報告を行なう．

7.2　従来の雑音処理技術

7.2.1　雑音処理技術の分類

従来の雑音処理技術は，以下の 3 つに大別することができる．

1. 音声入力部での雑音処理
2. 音声分析部での雑音処理
3. 認識部での雑音処理

1. としては，複数のマイクロフォンを用いて雑音信号を推定し，これを観測信号から差し引くもの等がある．2. は音声信号から音声特徴量を抽出するまでの段階で雑音による影響を軽減するもので，代表的なのは**第 7.2.2 節**で説明するスペクトルサブトラクション（SS）[41] である．この手法は実装が容易であり，かつ雑音による影響を軽減する効果が大きいため広く用いられている．これはスペクトル上で雑音による影響を軽減するものであるが，ケプストラム上で雑音による影響を軽減するものとして，例えば [42] は HEQ（Histogram Equalization）を用いた雑音処理手法を提案している．**第 7.3.1 節**でも述べるように，加算性雑音はケプストラムに対して非線形変換を施すことが数学的にも導かれる．そこで [42] では，ケプストラムに非線形変換を施すことで正規化を行なっている．その他，対数スペクトルの時間軌跡に対して周波数特性約 1〜10Hz のバンドパスフィルタを施す RASTA（Relative Spectra）[43] がよく知られている．このとき，低域成分の除去は乗算性歪みの除去に相当する（CMN もケプストラムの直流成分を除去するハイパスフィルタと解釈できる）．一方，高域成分の除去は対数スペクトルの急激な変化を抑制する働きを持つが，例

えば日本語のモーラ持続長がおよそ 140ms 程度（約 7Hz）であることからも，このフィルタの妥当性を窺い知ることができる．

3. は音声特徴量を抽出した後の耐雑音処理で，よく知られているものとして HMM 合成法 [44] がある．これは，無雑音音声の HMM と雑音の HMM から目的の雑音環境の音声 HMM を合成する方法である．また，**第 3.4.7 節**で紹介したヤコビ適応法も 3. に該当する．

7.2.2　スペクトルサブトラクション（SS）

スペクトルサブトラクション（Spectral Subtraction; SS）[41] は，定常的な加算性雑音に対する単純で効果的な手法として広く用いられている．時刻 k の音声信号，雑音信号，入力信号（雑音混じりの音声信号）を $s(k)$，$n(k)$，$x(k)$ とすると，

$$x(k) = s(k) + n(k) \tag{7.1}$$

が成立する．両辺にフーリエ変換を施すと，

$$X(f,m) = S(f,m) + N(f,m) \tag{7.2}$$

となる（f：周波数，m：フレーム番号）．これらは複素スペクトルである．ここから，

$$|X(f,m)|^2 = |S(f,m)|^2 + |N(f,m)|^2 + 2|S(f,m)||N(f,m)|\cos(\theta_{S-N}(f,m)) \tag{7.3}$$

と求めることができる．ここで，$\theta_{S-N}(f,m)$ は，$S(f,m)$ と $N(f,m)$ の位相差を表す．音声と雑音が無相関であれば，$\cos(\theta_{S-N}(f,m))$ の期待値は 0 となる．そこで SS では，

$$|X(f,m)|^2 \approx |S(f,m)|^2 + |N(f,m)|^2 \tag{7.4}$$

と近似し，$X(f,m)$ の位相を用いて $S(f,m)$ を，

$$\hat{S}(f,m) = \begin{cases} H(f,m)\cdot X(f,m) & (H(f,m) \geq \beta \text{（または } H(f,m) \geq 0)) \\ \beta \cdot X(f,m) & (H(f,m) < \beta \text{（または } H(f,m) < 0)) \end{cases} \tag{7.5}$$

と推定する．ここで，

$$H(f,m) = \frac{\left\{|X(f,m)|^2 - \alpha \cdot |\hat{N}(f,m)|^2\right\}^{\frac{1}{2}}}{|X(f,m)|} \tag{7.6}$$

である（$|\hat{N}(f,m)|^2$ は，$|N(f,m)|^2$ の推定値）．α はサブトラクション係数（Overestimation factor），β はフロアリング係数（flooring factor）と呼ばれる．例えば，音声認識エンジン Julius rev.3.3 においては，$\alpha = 2.0$，$\beta = 0.5$ をデフォルト値としている [45]．$|\hat{N}(f,m)|^2$ は，無音声区間における雑音パワースペクトルを用いて推定される．具体的には，その平均スペクトルを用いる方法や，

$$|\hat{N}(f,m)|^2 = \begin{cases} |\hat{N}(f,m-1)|^2 & (\text{音声区間}) \\ \gamma|\hat{N}(f,m-1)|^2 + (1-\gamma)|X(f,m)|^2 & (\text{無音声区間}) \end{cases} \tag{7.7}$$

と推定する方法がある（γ としては $0.7 \leq \gamma \leq 0.95$ がよく使われる）[46]．

SS の問題点としては，

- 雑音が定常であると仮定している．
- $\cos(\theta_{S-N}(f,m))$ は 0 ではない．
- 低 SNR では音声／非音声の判断が困難である．

といった点が挙げられる．例えば [47] は，$\theta_{S-N}(f,m)$ が $[-\pi,\pi]$ で一様分布である場合の $\phi=\cos(\theta_{S-N}(f,m))$ の確率密度関数 $f(\phi)$ が，

$$f(\phi)=\frac{1}{\pi\sqrt{1-\phi^2}} \tag{7.8}$$

となるために，$\cos(\theta_{S-N}(f,m))$ が 0 付近の値をとる確率は小さくなることを指摘している．そこで [47] では $|X(f,m)|^2$ として，

$$\overline{|X(f,m)|^2}=\sum_{\tau=0}^{T-1}\beta_\tau|X(f,m-\tau)|^2 \quad (\textstyle\sum\beta_\tau=1) \tag{7.9}$$

のように移動平均をとったものを用いることで，各フレームにおける $\cos(\theta_{S-N}(f,m))$ の値をスムージングし，$\cos(\theta_{S-N}(f,m))$ が 0 付近をとる確率密度を高くする SS-SMT（SS with Smoothing of Time Direction）を提案している．

また，[48] は，SS は低 SNR では音声／非音声の判断が困難であること，さらには動的に変動する雑音への追随性が不十分であることを指摘し，音声／非音声の判断を一切行なわずに雑音スペクトルの推定値を，

$$|\hat{N}(f,m)|^2=\gamma|\hat{N}(f,m-1)|^2+(1-\gamma)|X(f,m)|^2 \tag{7.10}$$

により求める CSS（Continuous Spectral Subtraction）の高い耐雑音ロバスト性を明らかにしている．但しこの方法では，パワースペクトルの大きな音素が先行する場合，それに続くパワーの弱い音素のスペクトルがマスクされやすい，という問題点がある．

7.3　加算性雑音による音声の構造的表象の歪み及び話者性の消失

7.3.1　雑音による音声の構造的表象の歪み

音響的普遍構造は乗算性歪み・線形変換性歪みに対して原理的に不変であるが，本稿では雑音環境下の認識実験を行なうため，加算性雑音が音響的普遍構造に与える影響について考える．クリーンな音声，雑音，雑音下の音声のパワースペクトルをそれぞれ $|X(f)|^2$，$|S(f)|^2$，$|Y(f)|^2$ とし，

$$|Y(f)|^2\approx|X(f)|^2+|S(f)|^2 \tag{7.11}$$

が成立すると仮定する．式 (7.11) は対数パワースペクトル（$y(f)=\log|Y(f)|^2$）上では，

$$y(f)\approx\log(\ \exp(x(f))+\exp(n(f))\ ) \tag{7.12}$$

表 7.1: 音響的条件（第 7.3 節）

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	FFTcep.（1〜12 次元）
音声事象分布	単一ガウス分布（対角共分散行列）

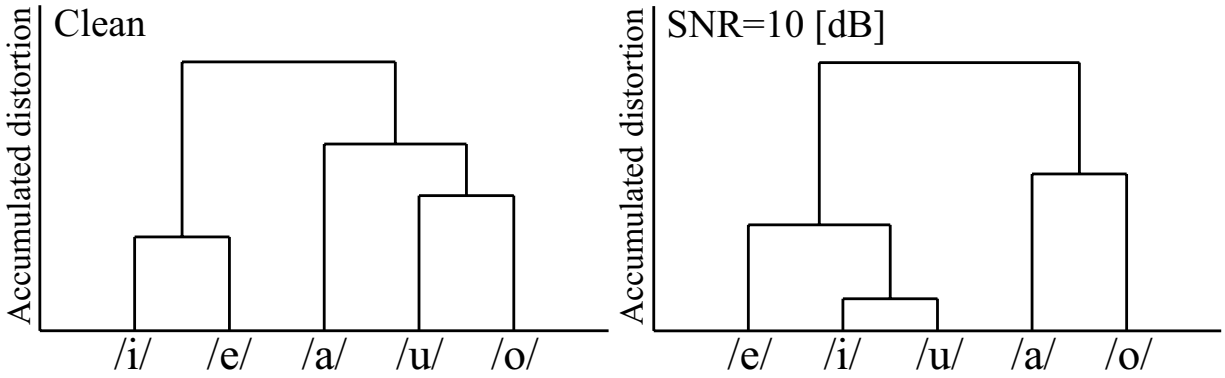

図 7.1: 男性話者の 5 母音の樹形図

と表される．従って，加算性雑音はケプストラムに対して非線形変換を施すため，音響的普遍構造はその形状が歪むものと予想される．図 7.1 は男性話者の 5 母音を，Ward 法によるボトムアップクラスタリングを用いて樹形図化したものである．このとき，分析条件は表 7.1 のとおりで，MAP 推定（$n = 14$）を用いて音声事象分布を求めている．左はクリーンな音声の樹形図であり，右はこれに SNR=10[dB] の白色雑音を加えたものである．構造サイズは正規化しているが，構造形状が雑音によって歪んでいる．特に/i/と/u/の距離が短くなっているが，これは/i/と/u/の第一フォルマントが互いに近傍にあり，他のフォルマントが雑音に埋もれたためと考えられる．

7.3.2　雑音による話者性の消失に関する定量的分析

第 6.5 節では音声に LPF を通し，話者性が多く含まれるスペクトル高域成分を下に揃えることで，日本語母音系列の認識性能を向上させた．スペクトル高域成分を音素間で均一化させる別の方法は，上に揃える方法である．図 6.3 において，白色雑音の重畳によってもスペクトル高域成分が揃えられ，話者性の消失が効果的に行なわれることを示した．

ここでは雑音による話者性の消失について定量的に調べるため，雑音を付与した場合の話者間構造差異・話者内構造差異の分散分析を行なった．音声試料としては，**第 5.2 節**で収録して分析対象とした，8 名話者（男性 4 名，女性 4 名）による孤立 5 母音 5 回発声のデータを用い，ここから各話者毎に 5 個の/a/-/i/-/u/-/e/-/o/の音声を得た．これに SNR $= \infty$（clean）, 20, 10, 0[dB] の白色雑音を重畳し，表 7.1 に示す分析条件でケプストラムを求め，ML 推定 or MAP 推定（$n = 14$）によって分布化した．求めた音声事象分布から，各話者毎に構造を抽出した．この際，全構造のサイズが等しくなるように正規化を施す場合

表 7.2: 雑音下での分析実験の結果（括弧内は構造サイズの正規化有り）

音響事象分布の推定法＝ ML 推定

SNR[dB]	話者間差異の平均	話者内差異の平均	構造サイズ
∞	1.47（1.17）	0.99（0.86）	12.5（12.5）
20	0.60（0.81）	0.36（0.58）	6.7（12.5）
10	0.43（0.85）	0.25（0.63）	4.3（12.5）
0	0.25（0.95）	0.17（0.73）	2.4（12.5）

音響事象分布の推定法＝ MAP 推定

SNR[dB]	話者間差異の平均	話者内差異の平均	構造サイズ
∞	1.13（0.93）	0.61（0.55）	11.5（11.5）
20	0.56（0.66）	0.25（0.40）	6.3（11.5）
10	0.40（0.74）	0.19（0.49）	4.0（11.5）
0	0.22（0.90）	0.14（0.64）	2.0（11.5）

と，施さない場合の 2 通りを試みた．抽出した構造から，話者間構造差異（${}_8C_2 \times 5^5 = 700$ 個），及び話者内構造差異（$8 \times_5 C_2 = 80$ 個）を求め，分析対象とした．

分散分析の結果，危険率はどの場合においても $p < 0.001$ となった．例えば，話者特有の発声の癖，方言差などの「構造の話者性」は，音声の構造化によって消失し得ない．そこで話者間差異・話者内差異の平均を表 7.2 に示す．括弧内の数値は，構造サイズを正規化した場合の結果である．ML 推定の場合も MAP 推定の場合も，以下のような傾向が見られた．まず，構造サイズを正規化しない場合，SNR の低下とともに構造サイズが小さくなり，その結果，話者間差異・話者内差異も減少している．そこで，構造サイズを正規化した場合を見ると，この場合でも雑音の重畳によって話者間差異・話者内差異が減少しているのが分かる．これは，雑音を重畳することで「音の話者性」の消失，及び話者内の発声の揺れ（即ち「構造揺らぎ」）の抑制が行なわれている効果と見られる．但し，構造サイズを正規化した場合，SNR をより下げていくと（例えば SNR=10[dB]），話者間差異・話者内差異は増加する．これは低 SNR においては，音韻差異が不明瞭に（構造サイズが非常に小さく）なり，構造形状が不安定になったためと見られる．

7.4　雑音下の孤立 5 母音系列の認識実験

7.4.1　クリーンな構造統計モデルを用いた認識実験

第 7.3.1 節において，雑音下では構造はその形状が歪むことを実験的に示したが，クリーン音声で学習された構造統計モデルを用いて雑音環境下の音声を認識する場合，入力音声の構造形状の歪みが原因で認識性能が低下することが予想される．これをより詳細に調べるため，以下の実験を行なった．

表 7.3: 音響的条件（第 7.4 節）

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	MCEP（α=0.55）（1〜12 次元）
分布推定方法	ML or MAP
音声事象分布	単一ガウス分布（対角共分散行列）

表 7.4: クリーンな構造統計モデルを用いた認識結果

	w/o SS		with SS	
SNR	ML	MAP	ML	MAP
∞	82.9%	99.9%	-	-
20[dB]	55.0%	98.5%	68.7%	99.9%
10[dB]	38.5%	39.5%	57.0%	52.8%
0[dB]	12.7%	12.4%	18.2%	13.1%

音声資料は，8 名話者（男性 4 名，女性 4 名）による孤立 5 母音 5 回発声のデータ（**第 5.2 節**）を用い，ここから各話者毎に 3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の音声を得た．その各々に SNR=∞（clean）, 20, 10, 0[dB] となるような白色雑音を重畳し，LPF（カットオフ周波数：2kHz）を施した後，ML 推定または MAP 推定（$n = 10, 1, 0.1, 0.01$ のうち最適なもの）によって音声事象分布を得た．ここから入力構造ベクトル（計 $8 \times 5^5 = 25,000$ 個）を得た．また，入力音声における雑音の影響を軽減するため，LPF 後に SS（$\alpha = 2.0$, $\beta = 0.5$）を行なう場合も試みた．その際，雑音パワースペクトルの推定には 300ms の白色雑音区間を用いた．構造統計モデルは，評価話者を除く 7 名によるクリーン音声から得た計 21,875（$= 7 \times 5^5$）個の/a/-/i/-/u/-/e/-/o/構造ベクトルを用いて学習させた．このときの音響的条件は表 7.3 のとおりである．

実験結果を表 7.4 に示す．予想通り，雑音下では認識性能が劣化している．SS による性能改善も見られるが，クリーン環境の性能に及ぶまでには至っていない．低 SNR のときに MAP 推定による効果が見られなくなったのは，事前知識の推定をクリーンな環境で行なっているため，雑音下の入力音声との間でミスマッチが生じたことが原因と考えられる．

7.4.2　雑音下の構造統計モデルを用いた認識実験

第 6.5 節では，男性話者 1 名で学習した構造統計モデルを用いて，クリーン環境における孤立 5 母音系列を 100%認識したが，構造統計モデルの学習に必要な話者が 1 名で十分ならば，極めて高品質な音声合成器を用いて，評価音声の雑音環境と合致する音声を合成し，それを基に構造統計モデル（及び事前知識）をオンラインで学習させることも可能と考えられる．これは構造に基づく音声知覚の運動理論 [49] と解釈することができるが，少なくとも人間は完璧な合成器を持っている．

ここでは，雑音下の構造統計モデルの性能を調べるため，評価音声の SNR が既知との仮

表 7.5: 雑音下の構造統計モデルを用いた認識結果

	full band		2kHz	
SNR	ML	MAP	ML	MAP
∞	24.7%	70.3%	86.8%	100.0%
20[dB]	73.9%	92.9%	67.9%	99.8%
10[dB]	77.4%	99.1%	68.1%	86.7%
0[dB]	73.9%	87.0%	71.1%	85.1%

表 7.6: 雑音下の孤立 5 母音系列に対する 3 つの手法の性能比較

SNR	HMM(260)	HMM(4,130)	Proposed(1)
∞	100.0%	100.0%	100.0%
20[dB]	100.0%	98.8%	99.8%
10[dB]	94.3%	97.2%	99.1%
0[dB]	83.0%	86.8%	87.0%

定のもと，学習話者 1 名の雑音下の構造統計モデル（及び事前知識）を用いた認識実験を行なった．**第 6.5 節**で使用した，男性話者が 5 母音を 35 回発声したデータを 7 つのグループに分け，各グループ毎に 3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の音声を得た．その各々に，評価音声と同じ SNR となるよう白色雑音を重畳した．ここから，計 21,875（$= 7 \times 5^5$）個の/a/-/i/-/u/-/e/-/o/構造ベクトルを求め，構造統計モデルの学習に用いた．音響的条件は**第 7.4.1 節**と同じく表 7.3 のとおりである．但し，白色雑音を重畳することで，スペクトル高域成分を揃えることができるので，LPF を用いない場合（full band）についても試みた．また，SS は行なっていない．

結果を表 7.5 に示す．表 7.4 よりはるかに良い認識性能が得られている．これは，入力構造と構造統計モデルとの間で雑音環境のミスマッチが無くなったためと考えられる．また，full band の場合においては，クリーン環境より雑音環境下の方が高い認識率が得られ，低 SNR では LPF を施した場合より良い性能が得られている．これは，白色雑音を重畳することで，フォルマントの情報を保ちつつ，スペクトル高域成分を揃えることができたためと思われる．但し，雑音レベルが非常に大きいとき（SNR=0[dB]）においては，認識性能が劣化している．これは，音声が雑音に埋もれて音韻差異が不明瞭になったためと見られる．

7.4.3　従来手法との比較実験

SS（$\alpha = 2.0$，$\beta = 0.5$）を用いた従来手法との比較実験も行なった．雑音パワースペクトルの推定には 300ms の白色雑音区間を用いた．音響モデルは，学習話者 4,130 名の混合共有 HMM，学習話者 260 名の状態共有 HMM の 2 通りの不特定話者モデルを用いた．特徴量は，全帯域の MFCC（1〜12 次元），ΔMFCC（1〜12 次元），及び ΔE であり（計 25 次元），CMN による話者・環境の正規化も行なった．言語的制約としては，120 単語のみを許容する文脈自由文法を用いた．

実験結果を表 7.6 に示す．提案手法の認識性能も合わせて載せている．この際，提案手法では full band の場合と 2kHz の場合のうち良い性能が得られた方を載せている．また，括弧内の数値は学習話者数である．用いる帯域を雑音レベルに応じて使い分けることで，学習話者 1 名の提案手法が学習話者 4,130 名の従来手法（SS 及び CMN を適用）を上回る結果を得ていることが分かる．

第8章

音声の構造的表象を用いた連続５母音系列音声認識

8.1　はじめに

前章までにおいて，孤立的に発声された日本語母音系列を認識タスクとする音声認識実験の結果を報告した．本章では，連続的に発声された日本語母音系列を認識タスクとして考える．この際，連続音声から構造を抽出する必要が生じるが，ここでは HMM を用いて各音声事象を各状態に対応付けることを考え，その学習アルゴリズムとして変分ベイズ法を導入する．認識実験を行なうことでその効果を検討し，従来手法との比較実験も行なう．

8.2　HMM を用いた連続音声の構造化

本章では，音声事象数 N が既知（N=5）であるものとして考える．まずは，ケプストラムが一次元の場合について考える．このとき，連続音声から HMM を用いて音声事象分布を推定し，構造化を行なう枠組みは図 8.1 のようになる．ここで，$X = \{x_1, x_2, \cdots, x_T\}$ は連続音声から求めたケプストラム系列，$\theta = \{a_i, \mu_i, S_i | i = 1, \cdots, N\}$ は HMM のパラメータであり，a_i は状態 i から $i+1$ への状態遷移確率，μ_i 及び S_i はそれぞれ状態 i の出力確率密度分布の平均と精度（分散の逆数）である．連続音声からケプストラム系列を求め，この系列のみを用いて HMM を学習させる．このとき，各状態が各音声事象に対応するものと考えられるので，各状態における出力確率密度分布を用いて，各音声事象分布 $p(c_i|X, N)$ を推定する．c_i は状態 i に対応する（即ち i 番目の音声事象における）ケプストラムを表す．

ここで，HMM の学習アルゴリズムとして ML 推定の枠組み（即ち，Baum-Welch アルゴリズム）を用いた場合，パラメータ θ を確定的変数[1]とみなし，

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta, N) \tag{8.1}$$

を満たす $\hat{\theta}$ を点推定[2]する（状態数 N を既知としている）．しかし，このとき学習に用いるデータは一発声された入力音声であるから，データ量の少なさに起因して，うまくパラメータ $\hat{\theta}$ を推定できない可能性がある．そこで，ベイズ推定の枠組みに基づく，HMM の学習アルゴリズムの導入を検討する．ベイズ推定は，データ量が少ないときにおける，ML 推定より（さらには MAP 推定より）優れた推定方法として知られている．その理由は，ベイズ推定ではあらゆるパラメータ（及び隠れ変数）を確率変数と見做し，それらに関する期待値をとり，周辺化する点にある．また，ベイズ推定の枠組みでは，モデル数（この場合，音声事象数 N）の選択も自動的に行なうことができるため，音声事象数が未知である場合にも有利と考えられる．

ベイズ推定ではパラメータ θ を確率変数として扱い，X が得られた後の θ の事後分布 $p(\theta|X, N)$ を推定する．学習アルゴリズムは後述するが，パラメータの事後分布 $p(\theta|X, N)$

[1]その値が一点で定められる変数．一方，その値を取る「確率」が定められる変数を確率変数と呼ぶ．
[2]その値を一点のみ推定すること．

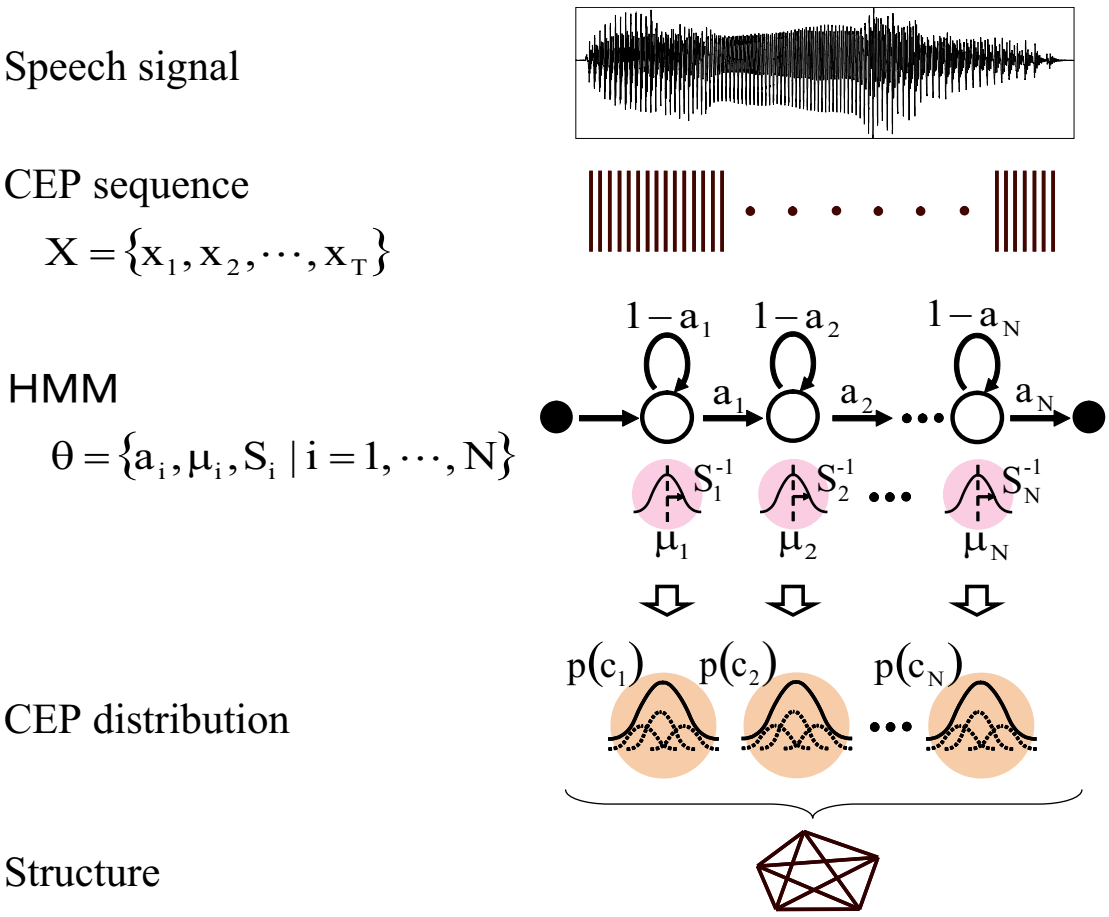

図 8.1: HMM を用いた連続音声の構造化の枠組み

を求めた後，音声事象分布をベイズ推定する場合は，

$$p(c_i|X,N) = \int p(c_i,\theta|X,N)d\theta \tag{8.2}$$

$$= \int p(c_i|\theta,X,N)p(\theta|X,N)d\theta \tag{8.3}$$

$$= \iint \mathcal{N}(c_i;\mu_i,{S_i}^{-1})p(\mu_i,S_i|X,N)d\mu_i dS_i \tag{8.4}$$

のようにして求められる．式 (8.3) から式 (8.4) の変形は $p(\theta) = \prod_{j=1}^{N} p(\mu_j,S_j)p(a_j)$ というパラメータの独立性の仮定，及び確率密度関数の積分値は 1 であることによる（例えば $\int p(a_j)da_j = 1$）．ベイズ推定では，パラメータ θ のあらゆる可能性を考慮し，その確率的重み付け和でもって分布を推定する．式 (8.4) では $p(\mu_i,S_i|X,N)$ がパラメータの事後分布であるが，式 (8.4) はこれを混合重みとしたときの，混合数 ∞ の混合ガウス分布（後述するが，これは一般化 t 分布となる）と見做すことができる．

尚，音声事象分布を MAP 推定する場合は，θ の事後分布 $p(\theta|X,N)$ を最大化するパラメータ θ_{MAP} を求め，この一点で代表させる（即ち，点推定）ことで式 (8.4) を，

$$p(c_i|X,N) = \mathcal{N}(c_i;\mu_{iMAP},{S_{iMAP}}^{-1}) \tag{8.5}$$

とする．この場合，音声事象分布はガウス分布として求まる．

8.3　変分ベイズ法を用いた音声事象分布の最大事後確率推定

8.3.1　変分ベイズ法

ベイズ推定に基づく HMM の学習アルゴリズムを導入するにあたって，HMM は隠れ変数が存在するため，パラメータの事後分布 $p(\theta|X,N)$ を解析的に求めることが困難となる．図 8.1 の枠組みで考える．時刻 t における状態が i であれば 1，そうでなければ 0 をとる隠れ変数 z_{ti} の集合を $Z=\{z_{ti}|t=1,\cdots,T,\ i=1,\cdots,N\}$ とすると，

$$\begin{aligned} p(X,Z|\theta,N) &= \prod_{t=1}^{T}\prod_{i=1}^{N}\left[\frac{\sqrt{S_i}}{\sqrt{2\pi}}\exp\left\{-\frac{S_i}{2}(x_t-\mu_i)^2\right\}\right]^{z_{t,i}} \\ &\quad\times\prod_{t=1}^{T-1}\prod_{i=1}^{N}(a_i)^{z_{t,i}z_{t+1,i+1}}(1-a_i)^{z_{t,i}z_{t+1,i}}\times a_N \end{aligned} \tag{8.6}$$

となる．パラメータの事後分布 $p(\theta|X,N)$ は，ベイズの定理を用いて，

$$p(\theta|X,N) = \frac{p(X|\theta,N)p(\theta|N)}{\int p(X|\theta,N)p(\theta|N)d\theta} \tag{8.7}$$

$$= \frac{\int p(X,Z|\theta,N)dZp(\theta|N)}{\int\int p(X,Z|\theta,N)dZp(\theta|N)d\theta} \tag{8.8}$$

と変形できる．$p(\theta|N)$ はパラメータの事前分布であり，$p(X,Z|\theta,N)$ は式 (8.6) より求まるが，上記積分を計算するのが困難となる．

変分ベイズ（Variational Bayes; VB）法は，このパラメータの事後分布を近似的に求める手法として提案されている．詳細は [50] に記されているが，パラメータ θ が $p(\theta)=\prod_{j=1}^{N}p(\theta_j)$ のように分解されるとき，パラメータの事後分布 $p(\theta|X,N)$，隠れ変数の事後分布 $p(Z|X,N)$ をそれぞれ近似する変分事後分布 $q(\theta|N)$，$q(Z|N)$ [3]を導入し，

$$q(\theta_j|N) = c_1 p(\theta_j|N)\exp\langle\log p(X,Z|\theta,N)\rangle_{q(Z|N),q(\theta_{-j}|N)} \tag{8.9}$$

$$q(Z|N) = c_2\exp\langle\log p(X,Z|\theta,N)\rangle_{q(\theta|N)} \tag{8.10}$$

として求める．ここで，$\langle f(x)\rangle_{g(x)}$ は $f(x)$ の $g(x)$ に関する期待値 $\int f(x)g(x)dx$ であり，θ_{-j} は θ の中の θ_j を除いたパラメータ集合 $\theta-\{\theta_j\}$ である．このように，式 (8.9) 及び式 (8.10) は，$q(Z|N)$ や $q(\theta|N)$ に関する期待値をとり，周辺化する操作を行なう．また，c_1，c_2 はそれぞれ $\int q(\theta_j|N)d\theta_j=1$，$\sum_Z q(Z|N)=1$ となるための規格化定数である．式 (8.9) 及び式 (8.10) は互いに依存関係にあるので，反復アルゴリズムによって逐次推定する．こうすることで q は局所最適解に収束する．このときの q （q^* と表すことにする）は最適変分事後分布と呼ばれる．

[3] q はいずれも事後分布であるが，[50] では表記を簡単にするため X を省略している．

また，変分ベイズ法では N の変分事後分布 $q(N)$ を導入し，式 (8.9) 及び式 (8.10) によって求めた最適変分事後分布 $q(\theta_j|N)^*$，$q(Z|N)^*$ を用いて，

$$\begin{aligned} q(N)^* &= c_3 p(N) \exp\left\{ \left\langle \log \frac{p(X,Z|\theta,N)}{q(Z|N)^*} \right\rangle_{q(Z|N)^*,q(\theta|N)^*} \right. \\ &\quad \left. + \sum_{j=1}^{N} \left\langle \log \frac{p(\theta_j|N)}{q(\theta_j|N)^*} \right\rangle_{q(\theta_j|N)^*} \right\} \end{aligned} \tag{8.11}$$

とすることで，最適変分事後分布 $q(N)^*$ を求めることができる．これが最大となる N を選択することによって，変分ベイズ法の枠組み内でモデル選択を行なうことができる．例えば，[51] ではテキスト文書の話題分割に変分ベイズ法を利用しているが，話題数未知の場合に対しても拡張している．本研究においても，音声事象分布の推定に変分ベイズ法を用いた場合は，音声事象数 N が未知の場合に対しても拡張することができる．

8.3.2　音声データセットを用いたパラメータの事前分布の設定

第 8.3.1 節で説明した変分ベイズ法の枠組みにおいて，パラメータの事前分布 $p(\theta|N)$ を設定することを考える．ここで，$p(\theta|N) = \prod_{i=1}^{N} p(\mu_i, S_i|N)p(a_i|N)$ より，$p(\mu_i, S_i|N)$ 及び $p(a_i|N)$ を求めればよい．それぞれの事前分布の形としては，正規-ガンマ分布，ベータ分布を用いる．これは，正規分布，二項分布の共役事前分布[4]がそれぞれ正規-ガンマ分布，ベータ分布であることによる．このとき，

$$\begin{aligned} p(\mu_i, S_i|N) &= p(\mu_i|S_i, N)p(S_i|N) && (8.12) \\ &= \mathcal{N}(\mu_i; \nu_i^{(0)}, (\omega_i^{(0)} S_i)^{-1}) \mathcal{G}(S_i; \alpha_i^{(0)}, \beta_i^{(0)}) && (8.13) \\ p(a_i|N) &= \mathcal{B}(a_i; \kappa_{1,i}^{(0)}, \kappa_{0,i}^{(0)}) && (8.14) \end{aligned}$$

と表される（$\mathcal{G}$ はガンマ分布，$\mathcal{B}$ はベータ分布）．$\nu_i^{(0)}$，$\omega_i^{(0)}$，$\alpha_i^{(0)}$，$\beta_i^{(0)}$，$\kappa_{1,i}^{(0)}$，$\kappa_{0,i}^{(0)}$ はパラメータ θ が従う分布のパラメータであり，超パラメータと呼ばれる．添字の数字は，式 (8.9) 及び式 (8.10) を反復した回数を意味する．ここでは事前分布なので，0 である．パラメータの事前分布の設定は，この超パラメータを設定することで行なわれる．

ここでは，音声データセットをあらかじめ用意しておき，そこから事前分布の超パラメータを設定することを考える．本章では，連続 5 母音を認識タスクとして扱うので，音声データセットは連続 5 母音の音声データによって構成され，120 単語全てが登場する．各音声データから，ML 推定の枠組み（Baum-Welch アルゴリズム）で HMM を学習する．音声データの数を M 個とし，m 番目の音声データで学習された HMM のパラメータを $\theta'_m = \{a'_{m,i}, \mu'_{m,i}, S'_{m,i}|i = 1, \cdots, N\}$ とする．この各々のパラメータの平均値及び分散値を，それぞれ $E(\cdot)$ 及び $V(\cdot)$ で表すことにする．事前分布の超パラメータを，この $E(\cdot)$ 及び $V(\cdot)$ を用いて設定することを考える．その詳細は**第 8.3.3 節**で記すことにするが，この

[4]事前分布と事後分布が同じ分布族になる分布

とき，HMM の各状態毎に M 個の観測値の $E(\cdot)$ 及び $V(\cdot)$ を求めた場合，事前分布の超パラメータの設定は各状態毎に行なわれることになる．一方，$M \times N$ 個の観測値の $E(\cdot)$ 及び $V(\cdot)$ を求め，全状態でそれらを共有した場合，事前分布の超パラメータは全状態共通の値が設定されることになる．音声事象数 N が未知である場合，後者の方が扱い易いものと考えられるので，本論文では後者を採択した．

8.3.3　パラメータの最適変分事後分布の導出アルゴリズム

変分ベイズ法を用いた，パラメータの最適変分事後分布の導出アルゴリズムを本節にて記す．但し，そこで登場する ξ_i 及び ϕ_i は，変分事後分布 $q(\mu_i|N) = \mathcal{T}(\mu_i; \nu_i, \xi_i, \phi_i)$ の超パラメータである（$\mathcal{T}$ は一般化 t 分布）．これは，隠れ変数の最適変分事後分布の導出の際に用いられる．また，$\alpha_t(i)$ 及び $\beta_t(i)$ は，**第 2.4.3 節**で説明した前向き変数及び後向き変数を，パラメータ θ に関して期待値をとり，周辺化したものである．

さらに，$\overline{z_{t,i}}$ は x_t が i 番目の状態に帰属する確率，$\overline{N}_i$ は i 番目の状態に帰属する平均データ数に相当し，$\overline{x}_i$（$\overline{C}_i$）は i 番目の状態に帰属するデータの平均（分散）に対応する（厳密には，$\overline{C}_i$ を $\overline{N}_i$ で割ったものが分散に相当する）．$\delta_{N,i}$ はクロネッカのデルタである．

i) 事前分布の超パラメータの設定

$$\nu_i^{(0)} = E(\mu'_{m,i}) \tag{8.15}$$
$$\omega_i^{(0)} = \sigma'^2_{i,0}/\tau'^2_{i,0} \tag{8.16}$$
$$\alpha_i^{(0)} = 1/\sigma'^2_{i,0} \tag{8.17}$$
$$\beta_i^{(0)} = 1 \tag{8.18}$$
$$\kappa_{1,i}^{(0)} = \frac{E(a'_{m,i})^2(1 - E(a'_{m,i}))}{V(a'_{m,i})} - E(a'_{m,i}) \tag{8.19}$$
$$\kappa_{0,i}^{(0)} = \frac{E(a'_{m,i})(1 - E(a'_{m,i}))^2}{V(a'_{m,i})} - (1 - E(a'_{m,i})) \tag{8.20}$$
$$\xi_i^{(0)} = \frac{\beta_i^{(0)}}{\omega_i^{(0)}\alpha_i^{(0)}} \tag{8.21}$$
$$\phi_i^{(0)} = 2\alpha_i^{(0)} \tag{8.22}$$

と設定する．但し，

$$\tau'^2_{i,0} = V(\mu'_{m,i}) \tag{8.23}$$
$$\sigma'^2_{i,0} = E(S'_{m,i}) \tag{8.24}$$

である．$l = 0$ とする．

ii) （ベイズ的）前向き変数及び後向き変数の更新

$$\alpha_1(i) = \begin{cases} \exp(B_{1,x_1}) & (\text{i=1 のとき}) \\ 0 & (\text{それ以外のとき}) \end{cases} \tag{8.25}$$

$$\alpha_{t+1}(i) = \alpha_t(i-1)\exp\left(A_{1,i-1} + B_{i,x_{t+1}}\right) + \alpha_t(i)\exp\left(A_{0,i} + B_{i,x_{t+1}}\right) \tag{8.26}$$

$$\beta_T(i) = \begin{cases} \exp(A_{1,N}) & (\text{i=N のとき}) \\ 0 & (\text{それ以外のとき}) \end{cases} \tag{8.27}$$

$$\beta_{t-1}(i) = \beta_t(i)\exp\left(A_{0,i} + B_{i,x_t}\right) + \beta_t(i+1)\exp\left(A_{1,i} + B_{i+1,x_t}\right) \tag{8.28}$$

と漸化的に求める．但し，

$$B_{i,x_t} = \frac{1}{2}\left[\psi(\alpha_i^{(l)}) - \log\beta_i^{(l)} - \frac{\alpha_i^{(l)}}{\beta_i^{(l)}}\left\{\frac{\phi_i^{(l)}}{\phi_i^{(l)}-2}\xi_i^{(l)} + (x_t - \nu_i^{(l)})^2\right\}\right] \tag{8.29}$$

$$A_{1,i} = \psi(\kappa_{1,i}^{(l)}) - \psi(\kappa_{0,i}^{(l)} + \kappa_{1,i}^{(l)}) \tag{8.30}$$

$$A_{0,i} = \psi(\kappa_{0,i}^{(l)}) - \psi(\kappa_{0,i}^{(l)} + \kappa_{1,i}^{(l)}) \tag{8.31}$$

である．ψ はディガンマ分布である．

iii) 隠れ変数の期待値の更新

$$\overline{z_{t,i}} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)} \tag{8.32}$$

$$\overline{z_{t,i}z_{t+1,i+1}} = \frac{\alpha_t(i)\exp(A_{1,i} + B_{i+1,x_{t+1}})\beta_{t+1}(i+1)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)} \tag{8.33}$$

$$\overline{z_{t,i}z_{t+1,i}} = \frac{\alpha_t(i)\exp(A_{0,i} + B_{i,x_{t+1}})\beta_{t+1}(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)} \tag{8.34}$$

iv) 変分事後分布の超パラメータの更新

$$\overline{N}_i = \sum_{t=1}^T \overline{z_{t,i}} \tag{8.35}$$

$$\overline{x}_i = \frac{\sum_{t=1}^T \overline{z_{t,i}}x_t}{\sum_{t=1}^T \overline{z_{t,i}}} \tag{8.36}$$

$$\overline{C}_i = \sum_{t=1}^T \overline{z_{t,i}}(x_t - \overline{x}_i)^2 \tag{8.37}$$

として，

$$\nu_i^{(l+1)} = \frac{\omega_i^{(0)}\nu_i^{(0)} + \overline{N}_i\overline{x}_i}{\omega_i^{(0)} + \overline{N}_i} \tag{8.38}$$

$$\omega_i^{(l+1)} = \omega_i^{(0)} + \overline{N}_i \tag{8.39}$$

$$\alpha_i^{(l+1)} = \alpha_i^{(0)} + \frac{1}{2}\overline{N}_i \tag{8.40}$$

$$\beta_i^{(l+1)} = \beta_i^{(0)} + \frac{1}{2}\overline{C}_i + \frac{\omega_i^{(0)}\overline{N}_i}{2(\omega_i^{(0)} + \overline{N}_i)}(\nu_i^{(0)} - \overline{x}_i)^2 \tag{8.41}$$

$$\kappa_{1,i}^{(l+1)} = \kappa_{1,i}^{(0)} + \sum_{t=1}^{T-1} \overline{z_{t,i}z_{t+1,i+1}} + \delta_{N,i} \tag{8.42}$$

$$\kappa_{0,i}^{(l+1)} = \kappa_{0,i}^{(0)} + \sum_{t=1}^{T-1} \overline{z_{t,i}z_{t+1,i}} \tag{8.43}$$

$$\xi_i^{(l+1)} = \frac{\beta_i^{(l+1)}}{\omega_i^{(l+1)}\alpha_i^{(l+1)}} \tag{8.44}$$

$$\phi_i^{(l+1)} = 2\alpha_i^{(l+1)} \tag{8.45}$$

と更新後，$l = l + 1$ とする．

v) 超パラメータの収束判定

超パラメータの収束判定を行なう．収束するまで **ii)〜iv)** を繰り返す．収束したとき，最適変分事後分布が得られたことになる（収束後，添字の (l) を取り外す）．

8.3.4　音声事象分布の最大事後確率推定

第 8.3.3 節で求めた最適変分事後分布 $q(\mu_i, S_i|N)^* = \mathcal{N}(\mu_i;\nu_i,(\omega_i S_i)^{-1})\mathcal{G}(S_i;\alpha_i,\beta_i)$ を $p(\mu_i, S_i|X, N)$ として式 (8.4) に代入することで，音声事象分布をベイズ推定の枠組みで求めることができる．実際に式 (8.4) に代入してみると，

$$p(c_i|X, N) = \mathcal{T}(c_i;\nu_{c_i},\xi_{c_i},\phi_{c_i}) \tag{8.46}$$

が得られる（混合数 ∞ の混合ガウス分布は一般化 t 分布となる）．但し，

$$\nu_{c_i} = \nu_i \tag{8.47}$$

$$\xi_{c_i} = (\omega_i + 1)\xi_i \tag{8.48}$$

$$\phi_{c_i} = \phi_i \tag{8.49}$$

である．しかし，このようにベイズ推定の枠組みで音声事象分布を求めた場合，音声を構造化するにあたって，一般化 t 分布同士の分布間距離を求めなければならない．このため，ここでは孤立 5 母音の認識タスクのときと同様，MAP 推定の枠組みで音声事象分布を求

めることを考える．その場合，音声事象分布は式 (8.5) のように正規分布の形となり，

$$\mu_{iMAP} = \nu_i \tag{8.50}$$
$$S_{iMAP} = \alpha_i/\beta_i \tag{8.51}$$

となる．

尚，孤立 5 母音の認識タスクのときにおいては，本来入力発声のデータ量である変数 n を，入力発声の事前知識に対する重みと見なして様々な値に変えていたが，今回それに相当する変数は $\overline{N}_i$ である．但し，**第 8.3.3 節の ii)〜iv)** は繰り返されるので，超パラメータが収束される前に $\overline{N}_i$ を変更すると，(ベイズ的）前向き変数，後ろ向き変数まで影響が及び，アライメントがうまく行なわれない可能性がある．従ってここでは，超パラメータの収束後に式 (8.35) の $\overline{N}_i$ に適当な定数 w をかけて，超パラメータを再度計算し直すことにした．この際，式 (8.37) の $\overline{C}_i$ は，i 番目の状態に帰属するデータの分散値に $\overline{N}_i$ をかけたものに相当するので，$\overline{C}_i$ にも定数 w をかけるようにした．

8.3.5　多次元への拡張

これまでの話は，ケプストラムが一次元の場合のものであったが，実際はケプストラムは多次元ベクトルであるので，**第 8.3.3 節**のアルゴリズムを多次元に拡張する必要がある．これは，ケプストラムベクトルの各要素の独立性を仮定すれば容易に実現できる．

但し，式 (8.29)，及び式 (8.38) から式 (8.41) の計算は注意を要する．式 (8.29) では，左辺が一次元だが，右辺が多次元になっている．ここでは，$\exp(B_{i,x_t})$ が出力確率密度としての意味を持つので，右辺の各ベクトル成分の和をとることで B_{i,x_t} を求めることにした．また，式 (8.38) から式 (8.41) においては，$\overline{N}_i$ だけが一次元になっているが，これを各成分の値が $\overline{N}_i$ であるベクトルに拡張することにした．

8.4　連続 5 母音系列の認識実験

8.4.1　連続 5 母音系列の収録

成人男性 8 名，女性 8 名に対して，連続 5 母音系列の収録を行なった．${}_5P_5 = 120$ 個の単語をそれぞれ 1 回ずつ発声させ，それを 5 回繰り返した．この際，調音努力が構造サイズに影響を与えてしまうので [37]，連続 5 母音系列のサンプル音声をスピーカー提示し，その直後に発話スタイル，母音の継続長などを真似る形で発声させた．F_0 については LHHLL となるように発声させるようにした．ここから，男性 4 名，女性 4 名の音声データを評価データとして選び，残りの男性 4 名，女性 4 名のデータを学習データとした．

8.4.2　実験条件

評価データから，音声特徴量として MCEP（1〜12 次元），Δ MCEP（1〜12 次元），及び Δ E を抽出した（計 25 次元）．Δ ケプストラムを用いたのは，今回の認識タスクは連続

表 8.1: 音響的条件（第 8.4 節）

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	MCEP（α=0.55）（1〜12 次元）
	+ ΔMCEP（1〜12 次元） + ΔE（計 25 次元）
音声事象分布	単一ガウス分布（対角共分散行列）
分布推定方法	ML , MAP w/o VB, or MAP with VB
状態数	N=15
カットオフ周波数	full-band
MAP 推定における重み	w=1
評価話者	計 8 名（男性 4 名，女性 4 名）
学習話者	計 8 名（男性 4 名，女性 4 名）

5 母音系列であり，ケプストラムの時間軸に対する変動量も重要と考えられるためである．抽出した音声特徴量を用いて HMM を学習させ，そこから音声事象分布を求めた．この際，

1. Baum-Welch アルゴリズムで HMM を学習後，その出力確率密度分布をそのまま音声事象分布とする方法（ML）
2. Baum-Welch アルゴリズムで HMM を学習後，その出力確率密度分布を用いて，**第 6.3.1 節**の枠組みの MAP 推定を行ない，音声事象分布を推定する方法（MAP w/o VB）
3. 変分ベイズ法を用いて HMM を学習後，**第 8.3.4 節**の枠組みの MAP 推定を行なうことで，音声事象分布を推定する方法（MAP with VB）

の 3 通りを試みた．変分ベイズ法を利用する場合は，式 (8.50) 及び式 (8.51) で用いられる超パラメータ ν_i，α_i，β_i の全てのベクトル成分の変動分が，超パラメータ更新前の値の 1.0×10^{-4} 倍以下になっているかどうかを，**第 8.3.3 節**の **v)** における超パラメータの収束判定の基準とした．また，**第 8.3.3 節**の **ii)〜iv)** の最大反復回数を 200 にした．このようにして求めた音声事象分布を構造化し，計 4,800 個（= 8 話者 × 120 単語 × 5 回発声）の全ての単語を含む構造ベクトルを入力に用いた．尚，構造サイズは正規化されている．

構造統計モデルは，各単語毎に学習データから計 40 個（= 8 話者 × 5 回発声）の構造ベクトルを抽出し，構造サイズ正規化後，これらを用いて単一ガウス分布を推定した．実験的条件を表 8.1 に示す．MAP w/o VB の重みは，MAP with VB と同様，各状態毎に対応するフレーム数を（ML 推定の枠組みで）求めた後，そのフレーム数に定数 w をかけることで定義した．以下の実験では，表 8.1 の実験的条件において，状態数，カットオフ周波数，MAP 推定における重みのうちのいずれかを変更することで行なった．

8.4.3　状態数を変化させたときの実験結果

表 8.1 の実験的条件から，状態数を $N = 5$，15，25 と変化させたときの実験結果を表 8.2 に示す．状態数の増加に伴って，認識性能が向上していることが分かる．これは，各状態へのアライメントがより細かくなっているのが原因とみられる．例として，図 8.2 に，男

表 8.2: 状態数を変化させたときの認識結果

推定法＼状態数	5	15	25
ML	19.9 %	32.4%	31.9%
MAP w/o VB	30.8 %	59.1%	61.5%
MAP with VB	34.1 %	60.4%	62.5%

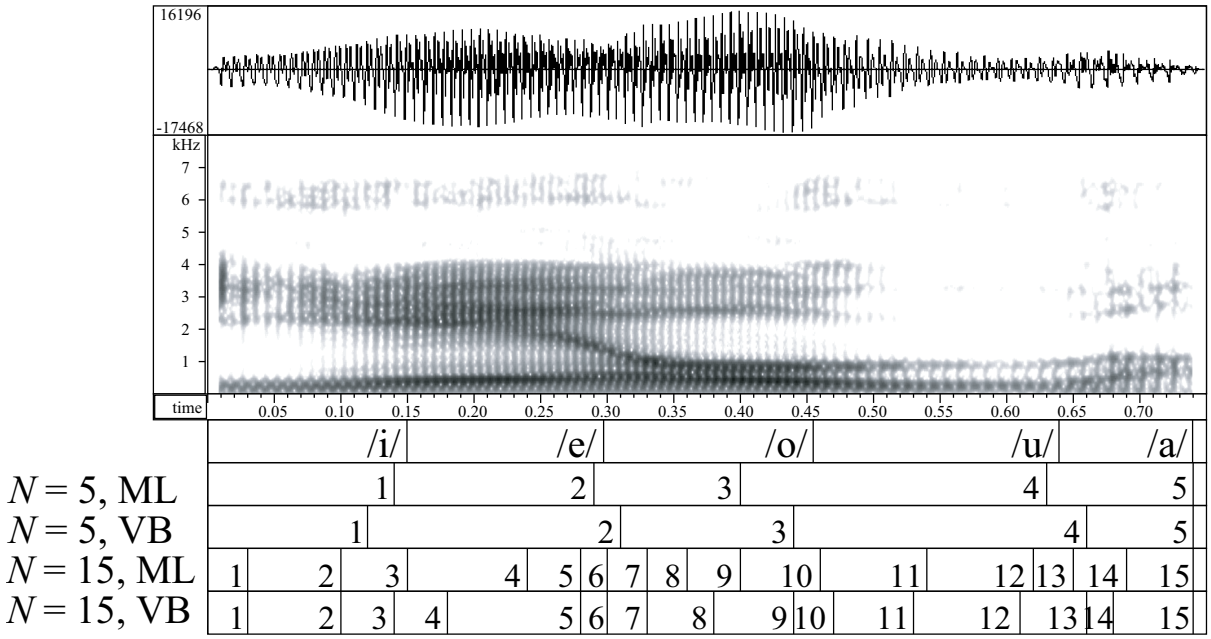

図 8.2: 男性話者の/i/-/e/-/o/-/u/-/a/に対する状態単位のアライメント

性話者の/i/-/e/-/o/-/u/-/a/に対して，HMM の状態単位のアライメントを行なったものを示す．図の下方に 5 種類のアライメントを載せている．一番上の段は人手によるアライメントである．2 番目以降は認識器によるもので，数字が状態番号を指す．このとき，状態数としては 5 状態または 15 状態，HMM の学習アルゴリズムとしては，Baum-Welch アルゴリズム（ML）または変分ベイズ法（VB）のいずれかを選択している．5 状態のときは/o/と/u/などのアライメントがずれてしまっているが，15 状態のときはアライメントが細かい分，アライメントのずれは，より微小なものになっている．

但し，アライメントのずれの問題が解決されたとしても，各母音でスペクトルが安定している区間に対応する HMM の状態は，話者毎に，さらには発声毎に異なる可能性がある．式 (4.5) による構造に基づく音響的照合は，一方をシフト及び回転させた後の，両者の対応する各音声事象分布の差異を近似的に求めるものであるが，上記が故に，スペクトルが安定している音声事象同士がうまく照合されない可能性がある．表 8.2 の結果が，状態数を 25 まで増やしても 60%程度に留まっているのは，これが主な原因の一つと考えられる．

また，MAP with VB と MAP w/o VB を比較すると，特に状態数が少ない場合に変分ベイズ法による効果があることが分かる．MAP w/o VB では，Baum-Welch アルゴリズムによって点推定された HMM パラメータを用いて事後確率を最大化するが，点推定され

表 8.3: 用いる帯域を変化させたときの認識結果（状態数 N=15）

推定法＼帯域	full-band	4kHz	2kHz
ML	32.4 %	37.6%	41.5%
MAP w/o VB	59.1 %	58.6%	61.1%
MAP with VB	60.4 %	59.1%	61.4%

表 8.4: MAP 推定における重みを変化させたときの認識結果（状態数 N=15）

推定法＼重み	0.001	0.01	0.02	0.05	0.1	1	10	100
MAP w/o VB	49.5%	61.8%	62.8%	62.0%	60.5%	59.1%	59.1%	54.1%
MAP with VB	48.1%	61.0%	62.9%	62.2%	61.5%	60.4%	59.5%	53.1%

た HMM パラメータにおいてアライメントがずれていると，それによる影響を受け易いものと考えられる．一方，MAP with VB では，式 (8.9) 及び式 (8.10) のように，$q(Z|N)$ や $q(\theta|N)$ に関する期待値をとり，周辺化しながら求めた最適変分事後分布を用いるため，状態数が少ないときも，より頑健に最大事後確率を推定することができたものと見ている．

8.4.4　用いる帯域を変化させたときの実験結果

表 8.1 の実験的条件から，カットオフ周波数を full-band，4kHz，2kHz と変化させたときの実験結果を表 8.3 に示す．LPF を施すことで，孤立 5 母音の認識タスクの時ほどではないが，認識率が向上していることが分かる．用いる帯域を狭めると，アライメントがよりうまく行なわれなくなる可能性がある．しかし，それと同時に LPF はスペクトル高域成分に多く含まれる話者性を消失させる効果があるので，その観点からは音声事象分布がより頑健に推定されたものと考えられる．

8.4.5　MAP 推定における重みを変化させたときの実験結果

表 8.1 の実験的条件から，MAP 推定における重みを，w=0.001 から w=100 まで変化させたときの実験結果を表 8.4 に示す．重みを w=1 から変化させると，変分ベイズ法による効果があまり見られなくなってくることが分かる．変分ベイズ法では $q(Z|N)$ や $q(\theta|N)$ に関する期待値をとり，周辺化しながら最適変分事後分布を求めるため，MAP with VB は MAP w/o VB よりも優れた最大事後確率を推定することができると考えられる．しかし，重みを w=1 としたときが，事後確率を最大化する場合に相当すると考えられるので，重み w がそこからずれることで，MAP with VB の MAP w/o VB に対する優位性が失われたのが原因である，と見ている．

8.4.6　従来手法との比較実験

従来手法との比較実験も行なった．従来の音響モデルとしては，3 種類のものを用意した．一つは学習話者 4,130 名の混合共有 HMM，一つは学習話者 260 名の状態共有 HMM，

表 8.5: 連続 5 母音系列に対する 4 つの手法の性能比較（状態数 N=15）

手法＼評価音声の帯域	full-band	4kHz	2kHz
full band HMM(260)	82.1%	69.0%	12.2%
full band HMM(4,130)	97.4%	89.6%	16.8%
limited band HMM(8)	96.9%	96.9%	96.9%
Proposed(8)	63.4%	63.4%	63.4%

もう一つは提案手法と同じ学習データ（学習話者 8 名）に対して，カットオフ周波数 2kHz の LPF を通して学習したトライフォンの HMM である（この場合，評価時には前処理として常に LPF を行なう）．いずれも CMN による話者・環境の正規化を行ない，特徴パラメータとしては MFCC（1〜12 次元），ΔMFCC（1〜12 次元），及び ΔE を用いた（計 25 次元）．また，言語的制約としては，120 単語のみを許容する文脈自由文法を用いた．

提案手法を含めた 4 つの手法の認識性能を表 8.5 に示す．括弧内の数字は学習話者数である．ここで，提案手法としては，状態数 N=25，カットオフ周波数 2kHz，MAP 推定における重み w=1 としたときの MAP with VB の認識性能を載せている．全帯域を用いた従来手法では，入力音声のカットオフ周波数の低下に伴い，認識性能が大きく劣化している．一方，提案手法，及びそれと同じ音声データで学習された従来手法では，前処理として常にカットオフ周波数 2kHz の LPF を施しているため，いずれの帯域の評価音声に対しても同じ性能を示している．尚，提案手法と同じ学習データを用いた従来手法の認識性能を見ると，連続 5 母音系列に対する性能の方が，表 6.8 で示した孤立 5 母音系列に対する性能（88.8%）よりも良いことが分かる．これは，認識タスクが孤立 5 母音系列のときは学習話者が 1 名だったのに対し，今回では学習話者が 8 名であったことが原因と見られる．

このとき，提案手法は，それと同じ学習データの従来手法の性能には，どの帯域の入力音声に対しても及ばない．この主な原因として考えられるのが，話者毎，発声毎にばらつく音声事象と状態との対応付けである．両者が本来対応する音声事象分布同士をうまく照合するような，構造に基づく音響的照合などを検討していく必要があるだろう．また，提案手法と従来手法との融合も検討事項として挙げられる．

第9章

結論

9.1　まとめ

音声には話者の声道形状の特性，音響機器の特性などの非言語的特徴が不可避的に混入するが，これを表現する次元を保有しない音声の構造的表象が提案されている．本研究は，この構造を音声認識に利用することに関する基礎的検討を行なった．

様々な認識実験を行ない，その結果，孤立 5 母音系列の音声認識というタスクにおいては，音声事象分布の MAP 推定，及びスペクトル高域成分の均一化を施すことで，

- 音声の物理的実体を明示的に用いない音声認識
- 一人の話者で学習された音響モデル（構造モデル）を用いた不特定話者音声認識
- 適応・正規化技術を一切用いない不特定話者音声認識

が 100%の認識性能を以っていずれも実現可能であることを，認識実験によって示した．従来手法との比較実験においても，学習話者 1 名の提案手法が学習話者 4,130 名の従来手法（CMN を適用）を上回る結果が得られた．

また，雑音環境下における孤立 5 母音系列音声に対しても，学習話者が 1 名で済むのであれば，雑音下の構造統計モデルをオンラインで学習させることも可能，という仮定の下で認識実験を行ない，その結果として，学習話者 1 名の提案手法が学習話者 4,130 名の従来手法（SS 及び CMN を適用）を上回る結果が得られた．

最後に，孤立 5 母音系列から連続 5 母音系列へと認識タスクを拡張し，HMM を用いた連続音声の構造化を検討した．この際，HMM の学習アルゴリズムとして変分ベイズ法を導入し，その効果を認識実験によって確認した．また認識実験の結果，連続 5 母音系列音声を構造のみを用いて 60%以上認識することができることを示した．

9.2　今後の展望

連続 5 母音系列という認識タスクにおいては，比較実験において提案手法と同じ音声データを用いて学習された従来手法を上回ることが出来なかった．この大きな理由として考えられるのが，話者毎，発声毎に異なる，各音声事象と HMM の各状態との対応付けである．構造に基づく音響的照合のアルゴリズム等の改良を検討していく必要があると考えられる．また，本研究では音声事象数が既知との仮定の下，認識実験を行なったが，音声事象数が未知の場合についても式 (8.11) を用いる等して対応していく必要がある．その際，変分ベイズ法の局所最適性も同時に解決する併合分割操作付き変分ベイズ学習 [52] も一つの参考になるだろう．さらには，提案手法の子音を含めた連続音声認識への拡張，提案手法と従来手法との融合が今後の検討課題として挙げられる．

謝辞

まず，二年間に渡り本研究を進めるにあたって，熱心に度重なる御指導を頂いた指導教員の峯松信明助教授並びに広瀬啓吉教授に深く感謝します．また，研究活動を様々な面で支えて下さった高橋登技官，秘書の武田祥子さん，笠島恵美さんに深く感謝します．

また，この研究を進めていく上で，日々多くの議論を交わし合った博士課程の朝川智氏に深く感謝します．朝川智氏には，**第 5.3 節**及び**第 7.3.1 節**において，音響的普遍構造を視覚化する方法も教えて頂きました．**第 6.4 節**及び**第 6.7 節**における実験を行なうにあたって，多大なる御協力を頂いた，修士課程の丸山和孝氏にも深く感謝します．

さらには，東京大学 音声・言語・コミュニケーション研究会の 2005 年度学生交流会において，様々な視点からの御質問及びコメントを頂き，関連研究として**第 4.5.1 節**の [36] を紹介して頂いた，東京大学大学院 情報理工学系研究科 システム情報学専攻の嵯峨山茂樹教授に深く感謝します．大変参考になりました．

財団法人 C&C 振興財団からは，国際会議 EUROSPEECH 2005 及び ASRU 2005 での発表の際に，会議出席のための助成金を頂きました．2 度にも渡って支援して頂いたことは，大変有り難く思っております．ここに深く感謝します．

本研究の実験における音声資料を収集する際には，研究室内，さらには研究室外の様々な方々に御協力頂きました．誠に感謝しております．

日頃の研究室生活においても，卒論時代から大変御世話になった博士課程の八木裕司氏，卒論時代からの同期でお互い切磋琢磨し合った阿部悠氏，修士課程に入ってから共に研究に励み，辛いこともお互い笑い飛ばした浅野泰史氏，高田靖也氏，EUROSPEECH 2005 において，開催地であるポルトガルに関する色々な情報を提供してくれたレポルダオ・アントニオ氏，ASRU 2005 の発表練習に協力してもらい，鋭い質問及びコメントをくれたナリニョオ・ホアン氏，いつも研究室で沢山のお菓子や紅茶を提供してくれたデガー・エルハン氏ら，その他大勢の研究室の皆様のお陰で，実に楽しく過ごすことができました．ここに深く感謝の意を述べたいと思います．どうも有難うございました．

2006 年 2 月 3 日
村上 隆夫

参考文献

[1] R.K. Moore, “A comparison of the data requirements of automatic speech recognition systems and human listeners,” Proc. EUROSPEECH, pp.2581-2584 (2003)

[2] H. A. Gleason, “An introduction to descriptive linguistics,” New York: Holt, Rinehart & Winston (1961)

[3] N. Minematsu, “Yet another acoustic representation of speech sounds,” Proc. ICASSP, pp.585-588 (2004)

[4] N. Minematsu, “Mathematical evidence of the acoustic universal structure in speech,” Proc. ICASSP, pp.889-892 (2005)

[5] 峯松信明，西村多寿子，西成活裕，櫻庭京子，“構造不変の定理とそれに基づく音声ゲシュタルトの導出”，電子情報通信学会技術研究報告，SP2005-12，pp.1-8 (2005)

[6] 峯松信明，西村多寿子，“音声の相対音感　～音声と音楽の同質性に関する一考察～”，電子情報通信学会技術研究報告，SP2005-131，pp.121-126 (2005)

[7] 峯松信明，松井健，広瀬啓吉，“構造音韻論の物理実装に基づく新しい音声の音響的表象”，電子情報通信学会技術研究報告，SP2004-27，pp.47-52 (2004)

[8] B. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” J. Acoust. Soc. Am., vol.55, pp.1304-1312 (1974)

[9] D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech and Audio Processing, vol.3, no.1, pp.72-83 (1995)

[10] M. Pitz, S. Molau, R. Schlüter and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” Proc. EUROSPEECH, pp.1445-1448 (2001)

[11] 篠田浩一，“確率モデルによる音声認識のための話者適応化技術”，電子情報通信学会論文誌，D-II Vol. J87-D-II No.2，pp.371-386 (2004)

[12] C.H. Lee, C.H. Lin and B.H. Juang, “A study on speaker adaptation of the parameters of continuous density hidden Markov models,” IEEE Trans. Signal Processing, vol.39, no.4, pp.806-814 (1991)

[13] J.L. Gauvain and C.H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains,” IEEE Trans. Speech and Audio Processing, vol.2, no.2, pp.291-298 (1994)

[14] K. Shinoda and T. Watanabe, “Speaker adaptation with autonomous control using tree structure,” Proc. EUROSPEECH, pp.1143-1146 (1995)

[15] 高橋敏，嵯峨山茂樹，“学習移動ベクトルの相関関係を用いた音響モデルの話者適応化”，電子情報通信学会論文誌，D-II，vol.J82-D-II，no.3，pp.324-331 (1999)

[16] C.J. Leggetter and P.C. Woodland, “Maximum likelihood speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171-185 (1995)

[17] V.V. Digalakis, D. Rtischev and L.G. Neumeyer, “Speaker adaptation using constrained estimation of Gaussian mixtures,” IEEE Trans. Speech and Audio Processing, vol.3, no.5, pp.357-366 (1995)

[18] M.J.F. Gales and P.C. Woodland, “Mean and variance adaptation within the MLLR framework,” Computer Speech and Language, vol.10, pp.249-264 (1996)

[19] V.D. Diakoloukas and V. V. Digalakis, “Maximum-likelihood stochastic-transformation adaptation of hidden Markov models,” IEEE Trans. Speech and Audio Processing, vol.7, no.2, pp.177-187 (1999)

[20] J. McDonough and A. Waibel, “Performance comparisons of all-pass transform adaptation with maximum likelihood linear regression,” Proc. ICASSP, vol.1, pp.313-316 (2004)

[21] R. Kuhn, P. Nguyen, J.C. Janqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field and M. Contolini, “Eigenvoices for Speaker Adaptation,” Proc. ICSLP, pp.1771-1774 (1998)

[22] M. Kirby and L. Sirovich, “Application of the Karhunen-Loève procedure for the characterization of human faces,” IEEE Trans. Pattern Anal. Machine Intell., vol.12, no.1, pp.103-108 (1990)

[23] S. Sagayama, Y. Yamaguchi, S. Takahashi and J. Takahashi, “Jacobian approach to fast acoustic model adaptation,” Proc. ICASSP, vol.2, pp.835-838 (1997)

[24] H. Shimodaira, N. Sakai, M. Nakai and S. Sagayama, “Jacobian joint adaptation to noise, channel and vocal tract length,” Proc. ICASSP, vol.1, pp.197-200 (2002)

[25] V.V. Digalakis and L.G. Neumeyer, “Speaker adaptation using combined transformation and Bayesian methods,” IEEE Trans. Speech and Audio Processing, vol.4, no.4, pp.294-300 (1996)

[26] K. Chen, W. Liau, H. Wang and L. Lee, “Fast speaker adaptation using Eigenspace-based maximum likelihood linear regression,” Proc. ICSLP, vol.3, pp.742-745 (2000)

[27] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, “A compact model for speaker-adaptive training”, Proc. ICSLP, vol.2, pp.1137-1140 (1996)

[28] V. Doumpiotis and Y. Deng, “Eigenspace-based MLLR with speaker adaptive training in large vocabulary conversational speech recognition,” Proc. ICASSP, vol.1, pp.357-360 (2004)

[29] M. Padmanabhan, L.R. Bahl, D. Nahamoo and M.A. Picheny, “Speaker clustering and transformation for speaker adaptation in speech recognition systems,” IEEE Trans. Speech and Audio Processing, vol.6, pp.71-77 (1998)

[30] フェルディナン・ド・ソシュール，一般言語学講義，岩波書店 (1940)

[31] ローマン・ヤコブソン，構造的音韻論，岩波書店 (1996)

[32] A. Gutkin and S. King, “Structural representation of speech for phonetic classification,” Proc. ICPR, vol.3, pp.438-441 (2004)

[33] T. Fukuda and T. Nitta, “Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition,” IEICE transactions, vol.E87-D, no.5, pp.1110-1118 (2004)

[34] L. Deng, G. Ramsay and D. Sun, “Production models as a structural basis for automatic speech recognition,” Speech Communication, vol.33, no.2-3, pp.93-111 (1997)

[35] 峯松信明，“音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定”，電子情報通信学会技術研究報告，SP2003-180，pp.31-36 (2004)

[36] 下平博，木村正行，“母音間の相対関係に基づく不特定話者母音系列の認識”，電子情報通信学会論文誌，vol. J71-A，no.8，pp.1515-1522 (1988)

[37] N. Minematsu, S. Asakawa and K. Hirose, “The acoustic universal structure in speech and its correlation to para-linguistic information in speech,” Proc. IWMMS, pp.69-79 (2004)

[38] M. Shozakai and G. Nagino, “Improving robustness of speech recognition performance to aggregate of noises by two-dimensional visualization,” Proc. EUROSPEECH, pp.921-924 (2005)

[39] 庄境誠，“複数音声コーパスの俯瞰的分析”，電子通信学会技術研究報告，SP2005-112，pp.43-48 (2005)

[40] T. Kitamura and M. Akagi, “Speaker individualities in speech spectral envelopes,” JASJ(E), Vol.16, No.5 (1995)

[41] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113-120 (1979)

[42] Á de la Torre, J.C. Segura, C. Benítez, A.M. Peinado and A.J. Rubio, “Non-linear transformations of the feature space for robust speech recognition,” Proc. ICASSP, vol.1, pp.401-404 (2002)

[43] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech and Audio processing, vol.2, no.4, pp.578-589 (1994)

[44] F. Martin, K. Shikano and Y. Minami, “Recognition of noisy speech by composition of hidden Markov models,” Proc. EUROSPEECH, vol.2, pp.1031-1034

[45] 大語彙連続音声認識システム Julius：http://julius.sourceforge.jp/

[46] S.V. Vaseghi and B.P. Milner, “Noise compensation methods for hidden Markov model speech recognition in adverse environments,” IEEE Trans. Speech and Audio processing, vol.5, pp.11-21 (1997)

[47] 北岡教英，赤堀一郎，中川聖一，“スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識”，電子情報通信学会論文誌，Vol.J83-D-II，No.2，pp.500-508 (2000)

[48] 庄境誠，中村哲，鹿野清宏，“音声強調手法 E-CMN/CSS の自動車環境内での音声認識における評価 ”，電子情報通信学会論文誌，Vol.J81-D-II，No.1，pp.1-9 (1998)

[49] 柏野牧夫，“音声知覚の運動理論をめぐって”，日本音響学会秋季講演論文集，1-2-10，pp.243-246 (2004)

[50] 上田修功，“ベイズ学習 [I]，[II]，[III]，[IV]”，電子情報通信学会誌，vol. 85，no.4 (pp.265-271)，no.6 (pp.421-426), no.7 (pp.504-509)，no.8 (pp.633-638) (2002)

[51] T. Koshinaka, K. Iso and A. Okumura, “An HMM-based text segmentation method using variational bayes approach and its application to LVCSR for broadcast news,” Proc. ICASSP, vol.1, pp.485-488 (2005)

[52] 上田修功，“最良モデル探索のための変分ベイズ学習”，人工知能学会論文誌，vol.16，no.2，pp.299-308 (2001)

発表文献

[1] 村上隆夫，峯松信明，広瀬啓吉，“音声の構造化による非言語情報の消失に関する定量的分析”，日本音響学会秋季講演論文集，2-P-9，pp.379-380 (2004)

[2] 峯松信明，村上隆夫，丸山和孝，広瀬啓吉，志彫淳，西村多寿子，西成活裕，“構造不変の定理に基づく音声の構造的表象とその距離尺度”，日本音響学会春季講演論文集，1-5-13，pp.25-26 (2005)

[3] 丸山和孝，村上隆夫，峯松信明，広瀬啓吉，“音声の構造的表象に基づく音響的照合に関する実験的検討”，日本音響学会春季講演論文集，1-5-14，pp.27-28 (2005)

[4] 峯松信明，志甫淳，村上隆夫，丸山和孝，広瀬啓吉，“音声の構造的表象とその距離尺度”，電子情報通信学会技術研究報告，SP2005-13，pp.9-12 (2005-5)

[5] 村上隆夫，丸山和孝，峯松信明，広瀬啓吉，“音声の構造的表象を用いた日本語母音系列の自動認識”，電子情報通信学会技術研究報告，SP2005-14，pp.13-18 (2005-5)

[6] 朝川智，峯松信明，伊勢井敏子・ヤッコラ，村上隆夫，広瀬啓吉，“音声の構造的表象に基づく非母語話者の英語発音分析”，電子情報通信学会技術研究報告，SP2005-24，pp.25-30 (2005)

[7] T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose, “Japanese vowel recognition based on structural representation of speech,” Proc. EUROSPEECH, pp.1261-1264 (2005)

[8] 村上隆夫，朝川智，峯松信明，広瀬啓吉，“加算性雑音による音声の構造的表象の歪みに関する実験的検討”，日本音響学会秋季講演論文集，1-P-1，pp.153-154 (2005)

[9] 村上隆夫，丸山和孝，峯松信明，広瀬啓吉，“音声の構造的表象を用いた加算性雑音音声認識”，日本音響学会秋季講演論文集，1-P-2，pp.155-156 (2005)

[10] T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose, “Japanese vowel recognition using external structure of speech,” Proc. ASRU, pp.203-208 (2005)

[11] 村上隆夫，丸山和孝，朝川智，峯松信明，広瀬啓吉，“音声の構造的表象を用いた雑音環境下における日本語母音系列の自動認識”，電子情報通信学会技術研究報告，SP2005-130, pp.115-120 (2005)

[12] N. Minematsu, T. Nishimura, T. Murakami, and K. Hirose, “Speech recognition only with supra-segmental features — hearing speech as music —,” Proc. Speech Prosody (2006-5, submitted)

付録 A

孤立５母音系列の音声事象分布の最大事後確率推定

A.1　パラメータの事前分布の設定

まずは，ケプストラムが一次元であるとして考える．入力発声の音声事象分布をガウス分布で表現し，その平均値，分散値（及び精度）を μ，σ^2（及び S）とする．その ML 推定値 μ_{ML}，σ^2_{ML} は，入力発声として観測されたケプストラム系列 $X = \{x_1, \cdots, x_n\}$ の平均値及び分散値に一致する．ここで求めるものは，その MAP 推定値 μ_{MAP}，σ^2_{MAP} である．

事前分布のための音声データは一発声毎にガウス分布化される（計 M 個）訳であるが，その m 番目の発声の平均値，分散値（及び精度）を μ'_m，σ'^2_m（及び S'_m）とする．

事前分布 $p(\mu, S)$ の形としては，正規-ガンマ分布を用いる．これは，正規分布の共役事前分布が正規-ガンマ分布であることによる．このとき，

$$
\begin{aligned}
p(\mu, S) &= p(\mu|S)p(S) && \text{(A.1)}\\
&= \mathcal{N}(\mu; \nu, (\omega S)^{-1})\mathcal{G}(S; \alpha, \beta) && \text{(A.2)}
\end{aligned}
$$

と表される（$\mathcal{N}$ は正規分布，$\mathcal{G}$ はガンマ分布）．このとき，事前分布のパラメータは，ν，ω，α，β となり，これをあらかじめ用意していた音声データを用いて設定することになる．ここでは，[12] を参考に以下のように設定する．

$$
\begin{aligned}
\nu &= E(\mu'_m) && \text{(A.3)}\\
\omega &= \sigma_0^2/\tau_0^2 && \text{(A.4)}\\
\alpha &= 1/\sigma_0^2 && \text{(A.5)}\\
\beta &= 1 && \text{(A.6)}
\end{aligned}
$$

但し，τ_0^2，σ_0^2 は，

$$
\begin{aligned}
\tau_0^2 &= V(\mu'_m) && \text{(A.7)}\\
\sigma_0^2 &= E(\sigma'^2_m) && \text{(A.8)}
\end{aligned}
$$

である．

A.2　パラメータの事後分布の導出

$X = \{x_1, x_2, ..., x_n\}$ が観測された後，事後分布 $p(\mu, S|X)$ を求める．これは，

$$
\begin{aligned}
p(\mu, S|X) &\propto p(X|\mu, S)p(\mu, S) && \text{(A.9)}\\
&\propto \sqrt{\hat{\omega}S}\exp\left\{-\frac{\hat{\omega}S}{2}(\mu - \hat{\nu})^2\right\} S^{\hat{\alpha}-1}\exp(-\hat{\beta}S) && \text{(A.10)}\\
\therefore\ p(\mu, S|X) &= \mathcal{N}(\mu; \hat{\nu}, (\hat{\omega}S)^{-1})\mathcal{G}(S; \hat{\alpha}, \hat{\beta}) && \text{(A.11)}
\end{aligned}
$$

のようにして求められる．但し，

$$\hat{\nu} = \frac{\omega\nu + n\mu_{ML}}{\omega + n} \tag{A.12}$$

$$\hat{\omega} = \omega + n \tag{A.13}$$

$$\hat{\alpha} = \alpha + \frac{1}{2}n \tag{A.14}$$

$$\hat{\beta} = \beta + \frac{1}{2}n\sigma^2_{MAP} + \frac{\omega n}{2(\omega + n)}(\nu - \mu_{ML})^2 \tag{A.15}$$

である．式 (A.12) から式 (A.15) は，X が観測された後の式 (A.3) から式 (A.6) の更新値に相当する．ここから，

$$\mu_{MAP} = \hat{\nu} \tag{A.16}$$

$$\sigma^2_{MAP} = \frac{\hat{\beta}}{\hat{\alpha}} \tag{A.17}$$

を得る．これをケプストラムが多次元ベクトル（但し，各ベクトル成分の独立性を仮定）の場合に拡張することで，**第 6.3.1 節**のような式が得られる．

付録Ｂ

変分ベイズ法を用いた連続５母音系列の音声事象分布の最大事後確率推定

B.1　パラメータの事前分布の設定

事前分布 $p(\mu_i, S_i|N)$ の超パラメータの値の設定については，認識タスクが孤立 5 母音系列の場合と同様の枠組みを考える．それは超パラメータを，

$$\nu_i^{(0)} = E(\mu'_{m,i}) \tag{B.1}$$
$$\omega_i^{(0)} = \sigma'^2_{i,0}/\tau'^2_{i,0} \tag{B.2}$$
$$\alpha_i^{(0)} = 1/\sigma'^2_{i,0} \tag{B.3}$$
$$\beta_i^{(0)} = 1 \tag{B.4}$$

とするものである．但し，

$$\tau'^2_{i,0} = V(\mu'_{m,i}) \tag{B.5}$$
$$\sigma'^2_{i,0} = E(S'_{m,i}) \tag{B.6}$$

である．式 (B.1) から式 (B.6) は，式 (A.3) から式 (A.8) に対応している．

次に，事前分布 $p(a_i|N)$ の超パラメータの値の設定について考える．ここでは，それらについても，事前に用意した音声データを利用して設定することにする．ここで，ベータ分布 $\mathcal{B}(a_i; \kappa_{1,i}^{(0)}, \kappa_{0,i}^{(0)})$ の平均値 $E(a_i)$，分散値 $V(a_i)$ はそれぞれ，

$$E(a_i) = \frac{\kappa_{1,i}^{(0)}}{\kappa_{1,i}^{(0)} + \kappa_{0,i}^{(0)}} \tag{B.7}$$

$$V(a_i) = \frac{\kappa_{1,i}^{(0)}\kappa_{0,i}^{(0)}}{(\kappa_{1,i}^{(0)} + \kappa_{0,i}^{(0)})^2(\kappa_{1,i}^{(0)} + \kappa_{0,i}^{(0)} + 1)} \tag{B.8}$$

である．そこで，

$$\kappa_{1,i}^{(0)} = \frac{E(a'_{m,i})^2(1 - E(a'_{m,i}))}{V(a'_{m,i})} - E(a'_{m,i}) \tag{B.9}$$

$$\kappa_{0,i}^{(0)} = \frac{E(a'_{m,i})(1 - E(a'_{m,i}))^2}{V(a'_{m,i})} - (1 - E(a'_{m,i})) \tag{B.10}$$

とすることで $\kappa_{1,i}^{(0)}$ と $\kappa_{0,i}^{(0)}$ を設定することにした．これは，ベータ分布 $\mathcal{B}(a_i; \kappa_{1,i}^{(0)}, \kappa_{0,i}^{(0)})$ の平均値が $E(a'_{m,i})$，分散値が $V(a'_{m,i})$ となるものである．

B.2　パラメータの最適変分事後分布の導出

式 (8.6) 及び式 (8.9) から，

$$q(\mu_i, S_i|N) = q(\mu_i|S_i, N)q(S_i|N) \tag{B.11}$$
$$= \mathcal{N}(\mu_i; \nu_i, (\omega_i S_i)^{-1})\mathcal{G}(S_i; \alpha_i, \beta_i) \tag{B.12}$$
$$q(a_i|N) = \mathcal{B}(a_i; \kappa_{1,i}, \kappa_{0,i}) \tag{B.13}$$

と求めることができる．但し，

$$\overline{z_{t,i}} = \langle z_{t,i} \rangle_{q(Z|N)} \tag{B.14}$$

$$\overline{z_{t,i} z_{t+1,i+1}} = \langle z_{t,i} z_{t+1,i+1} \rangle_{q(Z|N)} \tag{B.15}$$

$$\overline{z_{t,i} z_{t+1,i}} = \langle z_{t,i} z_{t+1,i} \rangle_{q(Z|N)} \tag{B.16}$$

$$\overline{N}_i = \sum_{t=1}^{T} \overline{z_{t,i}} \tag{B.17}$$

$$\overline{x}_i = \frac{\sum_{t=1}^{T} \overline{z_{t,i}} x_t}{\sum_{t=1}^{T} \overline{z_{t,i}}} \tag{B.18}$$

$$\overline{C}_i = \sum_{t=1}^{T} \overline{z_{t,i}} (x_t - \overline{x}_i)^2 \tag{B.19}$$

として，

$$\nu_i = \frac{\omega_i^0 \nu_i^0 + \overline{N}_i \overline{x}_i}{\omega_i^0 + \overline{N}_i} \tag{B.20}$$

$$\omega_i = \omega_i^0 + \overline{N}_i \tag{B.21}$$

$$\alpha_i = \alpha_i^0 + \frac{1}{2} \overline{N}_i \tag{B.22}$$

$$\beta_i = \beta_i^0 + \frac{1}{2} \overline{C}_i + \frac{\omega_i^0 \overline{N}_i}{2(\omega_i^0 + \overline{N}_i)} (\nu_i^0 - \overline{x}_i)^2 \tag{B.23}$$

$$\kappa_{1,i} = \kappa_{1,i}^0 + \sum_{t=1}^{T-1} \overline{z_{t,i} z_{t+1,i+1}} + \delta_{N,i} \tag{B.24}$$

$$\kappa_{0,i} = \kappa_{0,i}^0 + \sum_{t=1}^{T-1} \overline{z_{t,i} z_{t+1,i}} \tag{B.25}$$

である（導出過程は省略するが，[50, 51] が参考になる）．$\overline{z_{t,i}}$ は x_t が i 番目の状態に帰属する確率，$\overline{N}_i$ は i 番目の状態に帰属する平均データ数，$\overline{x}_i$（$\overline{C}_i$ を $\overline{N}_i$ で割ったもの）は i 番目の状態に帰属するデータの平均（分散）に相当することを考慮すると，式 (B.20) から式 (B.23) は，式 (A.12) から式 (A.15) と対応していることが分かる．また，式 (B.24) 及び式 (B.25) は [51] と完全に一致する．

B.3　隠れ変数の最適変分事後分布の導出

B.3.1　$q(Z|N)$ の算出

式 (B.6) 及び式 (B.8) から隠れ変数の最適変分事後分布 $q(Z|N)$ を求めることが出来るが，その導出過程で $q(\mu_i|N)$（$= \int q(\mu_i, S_i|N) dS_i$）が登場するので，式 (B.12) を用いて $q(\mu_i|N)$ を先に求めておくと，

$$q(\mu_i|N) = \mathcal{T}(\mu_i; \nu_i, \xi_i, \phi_i) \tag{B.26}$$

となる（$\mathcal{T}$ は一般化 t 分布）．但し，

$$\xi_i = \frac{\beta_i}{\omega_i \alpha_i} \tag{B.27}$$

$$\phi_i = 2\alpha_i \tag{B.28}$$

である．これを用いて，

$$\begin{aligned} q(Z|N) &= c_4 \prod_{t=1}^{T} \prod_{i=1}^{N} \exp\left(z_{t,i} B_{i,x_t}\right) \\ &\quad \times \prod_{t=1}^{T-1} \prod_{i=1}^{N} \exp\left(z_{t,i} z_{t+1,i+1} A_{1,i} + z_{t,i} z_{t+1,i} A_{0,i}\right) \times \exp(A_{1,N}) \end{aligned} \tag{B.29}$$

と求めることができる．但し，c_4 は $\int q(Z|N)dZ = 1$ となるための規格化定数であり，

$$B_{i,x_t} = \frac{1}{2}\left[\psi(\alpha_i) - \log\beta_i - \frac{\alpha_i}{\beta_i}\left\{\frac{\phi_i}{\phi_i - 2}\xi_i + (x_t - \nu_i)^2\right\}\right] \tag{B.30}$$

$$A_{1,i} = \psi(\kappa_{1,i}) - \psi(\kappa_{0,i} + \kappa_{1,i}) \tag{B.31}$$

$$A_{0,i} = \psi(\kappa_{0,i}) - \psi(\kappa_{0,i} + \kappa_{1,i}) \tag{B.32}$$

である．ψ はディガンマ関数である（ここまでの導出過程においても，[50, 51] が参考になる）．

B.3.2　$\overline{z_{t,i}}$，$\overline{z_{t,i}z_{t+1,i+1}}$，$\overline{z_{t,i}z_{t+1,i}}$ の算出

式 (B.14) から式 (B.16) より，パラメータの最適変分事後分布を求めるには $\overline{z_{t,i}}$，$\overline{z_{t,i}z_{t+1,i+1}}$，$\overline{z_{t,i}z_{t+1,i}}$ が必要であるため，これらを $q(Z|N)$ から求める．ここで，$p(X, Z|N)$ について考える．$p(X, Z|N)$ は，

$$p(X, Z|N) = \int p(X, Z, \theta|N)d\theta \tag{B.33}$$

$$= \int p(X, Z|\theta, N)p(\theta|N)d\theta \tag{B.34}$$

$$= \langle p(X, Z|\theta, N)\rangle_{p(\theta|N)} \tag{B.35}$$

より，式 (8.6) の $p(X, Z|\theta, N)$ を θ に関して期待値をとり，周辺化したものであることが分かる．また，

$$p(X, Z|N) = p(X|N)p(Z|X, N) \tag{B.36}$$

$$\approx p(X|N)q(Z|N) \tag{B.37}$$

と変形し，式 (B.29) を式 (B.37) に代入することで，

$$\begin{aligned} p(X, Z|N) &\approx c_5 \prod_{t=1}^{T} \prod_{i=1}^{N} \exp\left(z_{t,i} B_{i,x_t}\right) \\ &\quad \times \prod_{t=1}^{T-1} \prod_{i=1}^{N} \exp\left(z_{t,i} z_{t+1,i+1} A_{1,i} + z_{t,i} z_{t+1,i} A_{0,i}\right) \times \exp(A_{1,N}) \end{aligned} \tag{B.38}$$

が得られる（c_5 は $\iint p(X,Z|N)dXdZ=1$ となるための規格化定数）．式 (B.38) と式 (8.6) を見比べると，$\exp B_{i,x_t}$ が状態 i で x_t を出力する出力確率密度，$\exp A_{1,i}$ が状態 i から $i+1$ への状態遷移確率，$\exp A_{0,i}$ が状態 i から i への状態遷移確率にそれぞれ対応していることが分かる．但し，これらは θ に関して期待値をとり，周辺化したものである．

これらを用いて，θ に関して期待値をとり，周辺化した（ベイズ的）前向き変数 $\alpha_t(i)$ 及び後向き変数 $\beta_t(i)$ を導入することができる．この場合，

$$\alpha_t(i) = p(z_{t,i}=1, x_1, ..., x_t|N) \tag{B.39}$$
$$\beta_t(i) = p(x_{t+1}, ..., x_T|z_{t,i}=1, N) \tag{B.40}$$

であるから，$\alpha_t(i)$，$\beta_t(i)$ は以下のような漸化式で求められる．

$$\alpha_1(i) = \begin{cases} \exp(B_{1,x_1}) & (\text{i=1 のとき}) \\ 0 & (\text{それ以外のとき}) \end{cases} \tag{B.41}$$
$$\alpha_{t+1}(i) = \alpha_t(i-1)\exp\left(A_{1,i-1}+B_{i,x_{t+1}}\right)+\alpha_t(i)\exp\left(A_{0,i}+B_{i,x_{t+1}}\right) \tag{B.42}$$
$$\beta_T(i) = \begin{cases} \exp(A_{1,N}) & (\text{i=N のとき}) \\ 0 & (\text{それ以外のとき}) \end{cases} \tag{B.43}$$
$$\beta_{t-1}(i) = \beta_t(i)\exp\left(A_{0,i}+B_{i,x_t}\right)+\beta_t(i+1)\exp\left(A_{1,i}+B_{i+1,x_t}\right) \tag{B.44}$$

これらを用いて，$\overline{z_{t,i}}$ を以下のように求めることができる．

$$\overline{z_{t,i}} \approx p(z_{t,i}=1|X,N) \tag{B.45}$$
$$= \frac{p(z_{t,i}=1,X|N)}{p(X|N)} \tag{B.46}$$
$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{B.47}$$

式 (B.47) は式 (2.7) と同様の式であるが，これは，$\overline{z_{t,i}}$ は x_t が i 番目の状態に帰属する確率を表すものであることからも頷ける．式 (B.47) によって求めた $\overline{z_{t,i}}$ を用いることで，あらゆる θ の可能性を考慮した上でのアライメントを行なうことができる．また，$\overline{z_{t,i}}$ と同様にして $\overline{z_{t,i}z_{t+1,i+1}}$，$\overline{z_{t,i}z_{t+1,i}}$ についても，以下のように求めることができる．

$$\overline{z_{t,i}z_{t+1,i+1}} = \frac{\alpha_t(i)\exp(A_{1,i}+B_{i+1,x_{t+1}})\beta_{t+1}(i+1)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{B.48}$$

$$\overline{z_{t,i}z_{t+1,i}} = \frac{\alpha_t(i)\exp(A_{0,i}+B_{i,x_{t+1}})\beta_{t+1}(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{B.49}$$

式 (B.41) から式 (B.44)，及び (B.47) から式 (B.49) は [51] と完全に一致する．

以上をまとめることで，**第 8.3.3 節**における，変分ベイズ法を用いた事後分布の近似的導出アルゴリズムが得られる．その後，音声事象分布は式 (8.50) 及び式 (8.51) を用いて MAP 推定される．

付録C

本論文に関連する分布及び関数

C.1　正規分布（ガウス分布）

正規分布（ガウス分布）は,

$$\mathcal{N}(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{C.1}$$

で表される．μ が平均値，σ^2 が分散値である．

C.2　ガンマ分布

ガンマ分布は,

$$\mathcal{G}(x;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)\quad(x\geq 0) \tag{C.2}$$

で表される（$\alpha>0$，$\beta>0$）．$\Gamma(\alpha)$ はガンマ関数であり,

$$\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}\exp(-t)dt \tag{C.3}$$

で表される．ガンマ分布は，右に歪んだ単峰型の分布である．

C.3　ベータ分布

ベータ分布は,

$$\mathcal{B}(x;\kappa_1,\kappa_0)=\frac{1}{B(\kappa_1,\kappa_0)}x^{\kappa_1-1}(1-x)^{\kappa_0-1}\quad(0\leq x\leq 1) \tag{C.4}$$

で表される（$\kappa_1>0$，$\kappa_0>0$）．ここで，$B(\kappa_1,\kappa_0)$ はベータ関数であり,

$$B(\kappa_1,\kappa_0)=\frac{\Gamma(\kappa_1)\Gamma(\kappa_0)}{\Gamma(\kappa_1+\kappa_0)}=\int_0^1 t^{\kappa_1-1}(1-t)^{\kappa_0-1}dt \tag{C.5}$$

で表される．

C.4　t 分布

t 分布は,

$$\mathcal{T}(x;\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2} \tag{C.6}$$

で表される（$\nu>0$）．t 分布は，正規分布と比べて裾が長い分布である．ν は自由度と呼ばれ，ν が ∞ に近づくほど一般化 t 分布は正規分布に近づく．これを一般化して,

$$\mathcal{T}(x;\nu,\xi,\phi)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\phi}\Gamma\left(\frac{\nu}{2}\right)}\left\{1+\frac{(x-\xi)^2}{\nu\phi}\right\}^{-(\nu+1)/2} \tag{C.7}$$

としたものが一般化 t 分布と呼ばれる（$\nu>0$，$\phi>0$）.

C.5　ディガンマ関数

ディガンマ関数は，ガンマ関数の対数微分であり，

$$\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x} \tag{C.8}$$

で表される．