特集 音声情報処理技術の最先端

統計的手法を用いた音声モデリングの高度化とその音声認識への応用

篠田 浩一
東京工業大学情報理工学研究科
shinoda@cs.titech.ac.jp

篠崎 隆宏
ワシントン州立ワシントン大学
staka@u.washington.edu

従来，音声認識においては隠れマルコフモデル(HMM)による音声のモデル化が一般的である．HMMを用いた音声認識は丁寧な読み上げ発声に対しては90%以上の高い認識性能を持つ．しかしながら，日常会話などの通常の話し言葉音声に対してはその性能はまだ十分でない．また，認識性能の著しく低い話者が存在する，周囲雑音の影響を受け性能が劣化する，など多くの課題が残されている．これらの課題の解決のためには，HMMを内包した，より柔軟な統計的モデリング手法が必要とされている．この目的のために多様な手法が活発に研究されているが，ここでは，その中で特に3つのトピック，情報量基準を用いたモデル選択，構造的事後確率最大化による話者適応化，ダイナミックベイジアンネットを用いた音声モデリング，について解説し，今後の展望を述べる．

■音声認識における音声のモデリングとその課題

音声のモデリング手法としての隠れマルコフモデル（Hidden Markov Model; HMM）は1970年代に導入されて以来広く用いられるようになり，現在では最も一般的な手法となっている．HMMの利点としては，音声のゆらぎを確率分布として表現でき頑健であること，モデルパラメータの学習やモデルに基づいた推論を効率的に行うアルゴリズムが存在することが挙げられる．HMMを用いた音声認識は，丁寧な読み上げ発声に対しては90%以上の高い認識性能を持つ．しかしながら，日常会話などの通常の話し言葉音声に対してはその性能はまだ十分でない．また，認識性能の著しく低い話者が存在する，周囲雑音の影響を受け性能が劣化する，などの問題点がある．

筆者らは，これらの問題の大きな原因は，HMMによる認識がまだ皮相的なレベルにとどまっていることにある，と考えている．例えて言えば，海面のさざ波の様子から海底の地形を推測するようなものである．音声認識の性能向上のためには，より一歩踏み込んだ，音声特徴の「内在構造」の解明とその活用が不可欠であろう．

以下に続く3章では，そのような問題意識のもとで進んでいる3つの研究トピックを紹介する．最初の「情報量基準を用いたモデル選択」は，与えられた学習データに対し最適な音声モデリングを行うためのツールとして情報量基準の1つである，記述長最小基準を用いる方法である．モデル構築にかかる，データ量および計算量を効果的に削減する．次の「構造的事後確率最大化による話者適応化」では音声の内在構造として音響空間における階層的な確率分布構造を仮定し，その構造を利用してモデルを話者の音声特徴に適応させる．従来に比べきわめて少量のデータでの認識性能向上が可能である．そして，最後は「ダイナミックベイジアンネットを用いた音響モデリング」である．従来，新しいモデルを実現し評

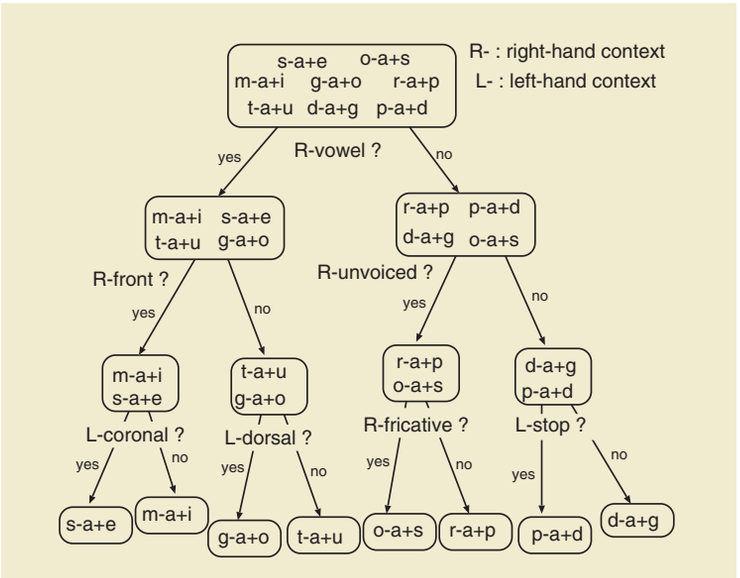

図 -1　音素文脈決定木の例．中心音素が /a/ の場合

価するためには学習や推論アルゴリズムの開発はもとよりプログラムの実装といった多くの労力が必要とされてきた．ベイジアンネットは，HMM を含めさまざまな確率モデルを表現できる柔軟なフレームワークであり，また表現可能なモデル一般に対する学習や推論アルゴリズムが開発されているため，さまざまな確率モデルを容易に実現し評価することができる．

■情報量基準を用いたモデル選択

音声認識におけるモデル選択問題

与えられたデータに対し最適なモデルを選択するモデル選択はパターン認識一般において重要な課題である．HMM を用いた音声認識においては，音声モデルのサイズの最適化がそれに相当する．音声認識では音素を基本単位として音声モデルを作成するが，発声変形の影響を受け，そのコンテキストにより同じ音素でも対応する音響的特徴量が大きく異なる．そこで，前後のコンテキストを考慮したトライフォン（3つ組音素）が認識単位として用いられる．音素の種類は日本語でも英語でも 40 ～ 50 程度であるのに対し，トライフォンの数は日本語で 4,000 種類以上，英語では 1 万種類以上と，著しく増加する．また，トライフォンの種類により出現頻度が著しく異なる．これらのトライフォンのパラメータを学習データから推定する場合，そのままではデータ不足問題が起き，性能が劣化する．そこで，クラスタリングを行い実効パラメータ数を減らす．クラスタリングには，しばしば，次章で説明する音素文脈決定木を用いた状態クラスタリングが使用される．

そこでの課題の 1 つとして，クラスタ数の最適化がある．クラスタ数が小さすぎると単純なモデルとなり，データの音響的特徴を十分表現できない．逆に，クラスタ数が大きすぎると複雑なモデルとなるが，認識単位あたりのデータが少なくなる．どちらの場合も性能が最適なモデルに比べ低くなってしまう．従来は，クラスタリングの停止の基準として，クラスタの分割・マージ前後の尤度比が閾値として用いられてきた．この閾値はもっぱら他のデータを用いた認識実験，あるいはクロスバリデーションにより最適化されていたが，多くのデータ量，計算量が必要であるという問題があった．最適な閾値を自動的に決定する方法が望ましい．

以下に続く節で，クラスタリング手法を説明した上で，情報量基準の 1 つである MDL 基準についてその概略を説明し，最後に MDL 基準を用いたクラスタリングにより閾値を自動決定する方法について述べる．

音素文脈決定木による状態クラスタリング

ここでは，前準備として，音素文脈決定木（Phonetic Decision Tree; PDT）を用いた状態クラスタリングについて説明する．PDT は，そのルートノードが，同じ中心音素を持つすべてのトライフォンの集合に対応し，左右の音素の種類や素性に関する質問でノードを再帰的に 2 分割することで作成される 2 分木である（図 -1）．リーフノードは個々のトライフォンに 1 対 1 に対応している．PDT は中心音素の各状態ごとに作成される．ここでは同じ中心音素を持つトライフォンは，みな状態数が同じ

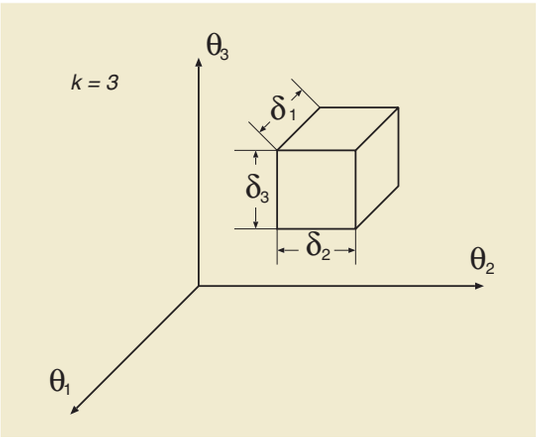

図 -2　MDL 基準におけるパラメータ空間の分割（次元数 k=3 の場合）

で，かつ，同じ状態間遷移を持つと仮定されている．

PDTは，以下のように尤度最大（Maximum Likelihood; ML）基準を用いて作成される．

(1) ルートノードを分割の対象ノードとする．
(2) 対象ノードに対応するすべてのデータサンプルからパラメータが最尤推定量のときの尤度を計算する．
(3) ある質問でノードを分割する．
(4) 分割された2つのノードそれぞれで (2) の手続きを行い，分割前後の尤度差を計算する．
(5) (3) ～ (4) をすべての質問で繰り返し，最も分割前後の尤度差が大きくなる質問を選択し，その質問で対象ノードを2分割する．
(6) 分割されたそれぞれの子ノードを対象ノードとし，(2)～ (5) の手続きを繰り返す．

ML基準では，分割により尤度差が減少することがないため，リーフとトライフォンが1対1に対応するまで，分割が行われる．これではクラスタリングを行う意味がないので，一般には尤度差に対し閾値を設定し，上の手続きの (5) で尤度差が閾値を超えた場合にのみ分割し，それ以外では分割を停止する．分割停止後は，その時点で末端のノードがリーフノードとなり，それに対応するすべての状態のパラメータは共有される．

MDL基準

今，データを $x^N=x_1,...,x_N$，確率モデルを $i=1,...,M$ としたとき，データに対するモデル i の記述長 $DL_{(i)}$ は，以下の式で表される．

$$DL^{(i)}=-\log P_{\hat{\theta}(i)}(x^N)+\frac{k^{(i)}}{2}\log N+\log M \qquad (1)$$

ここで，$k^{(i)}$ はモデル i の次数，$\hat{\theta}^{(i)}$ は データ x^N に対するモデル i のパラメータ $\theta^{(i)}=(\theta_1^{(i)},...,\theta_{k^{(i)}}^{(i)})$ の最尤推定量，$P_{\hat{\theta}^{(i)}}(x^N)$ はデータ x^N に対するモデル i の尤度，M はモデルの個数である．記述長最小（Minimum Description Length; MDL）基準は，データの記述長を最小とする確率モデルが最適な確率モデルであると主張する．

以下，MDL基準の式 (1) の導出をおおざっぱに説明する．詳細については，韓・小林の教科書[1]を参照されたい．MDL基準は，次式に示すように，まず，モデルパラメータを符号化し，そのモデルで指定される確率分布を符号化の関数としてデータを符号化して伝送を行う場合（2段階符号化）の合計の符号長 l を最小にするモデルを選択するものである．

$$l=-\log_K P_{\hat{\theta}}(x^N)+l_0(x^N) \qquad (2)$$

ここで，$l_0(x^N)$ は $\hat{\theta}$ の記述長，K は符号語の個数である．さて，ここでモデルパラメータ $\hat{\theta}$ は一般には実数であり，その符号化には無限長の符号長を要する．それでは符号化ができないので，パラメータ空間を量子化していくつかのセルに分割し，$\hat{\theta}$ の値をその属するセルの代表値に近似する，という手続きをとる．今，V_k をパラメータ θ の張る k 次元空間とし（図-2），$\delta=\{\delta_1,...,\delta_k\}$ を各セルの辺の長さとすると，上の式 (2) は以下のようになる．

$$l\sim-\log_K P_{\hat{\theta}+\delta}(x^N)+\log_K\frac{V_k}{\delta_1\cdots\delta_k} \qquad (3)$$

この式の第2項は，パラメータ空間の中で θ の属するセルを指定するのに要する符号長となる．ここで，δ_j は，パラメータ θ_j の精度を示す値となるが，問題は，最適な δ_j をどのように求めるか，ということである．δ_j が小さすぎると第2項の符号長が大きくなり，逆に δ_j が大きすぎると近似が粗くなり第1項の符号長が大きくなる．どこかに δ_j の最適値が存在することが予想される．モデルが正則であり，かつ，データサンプル N が十分大きいと仮定した場合，符号長を最小にする δ_j は以下のオーダーとなる．

$$\delta_j\sim O(1/\sqrt{N}) \qquad (4)$$

これは，パラメータの精度は $O(1/\sqrt{N})$ 以下にはできないという統計学の常識とも合致する．この値を式 (3) に代入すると，

$$l\sim-\log_K P_{\hat{\theta}}(x^N)+\frac{k}{2}\log_k N+O(1) \qquad (5)$$

という式を得る．さらに，第3項として，モデルが複数ある場合にその中からモデルを選択するのに要する記述長を加えると式 (1) を得る．第2項がモデルの大きさに対するペナルティとなっており，データに対し最適なモデルサイズを持つモデルを選択することができる

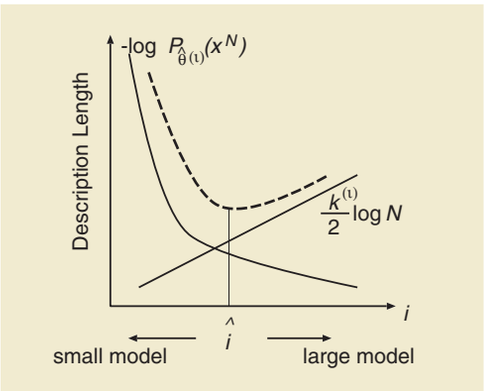

図 -3　MDL 基準とモデルサイズ

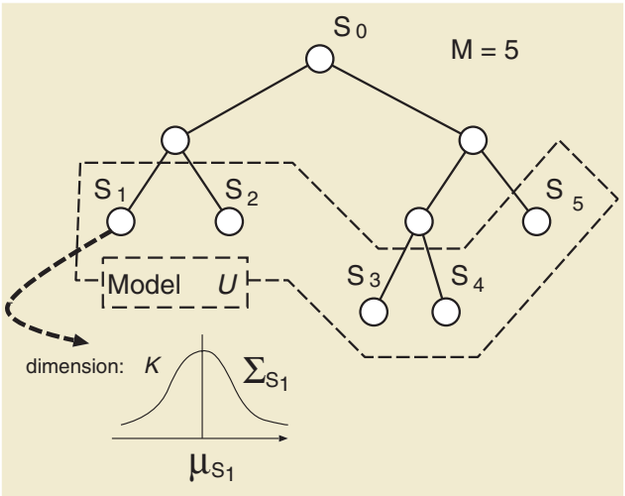

図 -4　MDL 基準の音韻決定木クラスタリングへの適用

（図 -3）．ML 基準に比べ閾値の調節が不要であるという大きな利点がある．

MDL 基準を用いた状態クラスタリング[2)]

HMM の状態クラスタリングに MDL 基準を適用することを考える．今，状態分割の途中で，状態 S が $S_1,...,S_M$ に分割されているケースを考える（図 -4）．$S_1,...,S_M$ から構成されるモデルを U とし，モデル U のデータ $\boldsymbol{O}$ に対する対数尤度を $L(U)$，記述長を $DL(U)$ とする．

まず，式 (1) の第 2 項を計算する．HMM の出力分布としては対角共分散行列を持つ多次元正規分布を用いるのが普通であり，この場合，特徴ベクトルの次元数を K とすると，各分布のパラメータは，K 次元の平均ベクトルと K 次元の対角分散であるから，分布ごとの自由パラメータ数 k は $2\times K\times M$ となる．さらに，式 (1) の第 3 項目は一定と仮定する．そうすると，記述長は以下のようになる．

$$DL(U)=L(U)+KM\log\sum_{m=1}^{M}\Gamma(S_m) \qquad (6)$$

$DL(U)$ を最小にするモデルが最適な状態クラスタである．実際のクラスタリングでは，ML 基準によるクラスタリングと同様の方法を用い，記述長の差分が非減少になる方向に分割を繰り返し，分割すべきノードがなくなった時点で停止する．

MDL 基準によるクラスタリングは ML 基準によるクラスタリングにおける閾値を自動的に決めることに対応する．この観点から，閾値を最適に決めさえすればいいのだから，ML 基準で十分である，という意見がしばしば聞かれる．しかしながら，MDL 基準を用いることの隠れた長所は，実は，ノードごとに独立に，対応するデータ量に応じた別々の閾値が与えられる，という点にある．すなわち，対応するデータ量の多いノードはより大きい閾値を持ち，より分割されにくくなる．それに対し，ML 基準でノードごとに閾値を最適化することは，膨大な手間がかかり事実上不可能であろう．実際，MDL 基準によるクラスタリング結果と，それと全体の状態数が同一になるよう閾値を調整した上での ML 基準によるクラスタリング結果とを比較すると，データ量の多い音素については，ML 基準の方が状態数が多く，データ量の少ない音素については，MDL 基準の方が状態数が多い，という傾向が見られる．このことが，MDL 基準を用いたクラスタリングが，十分に最適化された ML 基準よりも，しばしば性能が良くなる傾向があることの理由と推測される．

■構造的事後確率最大化による話者適応化

話者適応化

話者適応化は，音声認識において使用者の少量の発声を用いて認識システムをその使用者の音響的特徴に適応させる技術である[4)]．近年，誰の声でも事前登録なしで認識する不特定話者認識の実用化が進展しているが，その性能は使用者の発声を登録した特定話者認識の認識性能にはいまだ及んでいない．できるだけ少量の発声で特定話者並みの性能を上げる話者適応化技術の確立が期待されている．HMM を用いた認識の場合，話者適応化は出力分布である多次元混合正規分布の各正規分布のパラメータ，特に，平均ベクトルを適応化の対象とすることが多い．これは，他のパラメータに比べ適応の効果が大きいためである．そこで，以下では，パラメータとして

平均ベクトルをとった場合を想定して説明を進める．

話者適応化手法では，話者間写像の種類，パラメータ推定手法，パラメータ共有構造，のそれぞれについて選択の余地があり，写像については最尤回帰（MLLR）法，パラメータ推定手法については事後確率最大化（MAP）法などが提案されている．ここでは特に3番目のパラメータ共有構造に着目する．話者適応化手法では，ほぼ例外なく，モデルパラメータ共有を行うことで推定すべき自由パラメータ数を減少させ，少ないデータ量での頑健なパラメータ推定を実現している．実用上重要である，きわめてデータ量が少ない場合（数発声程度），むしろ，前者2つよりも性能に与える影響は大きい．逆にいうと共有構造を調節することで，写像の種類，パラメータ推定の手段に寄らず，ほぼ同程度の性能を実現できる．

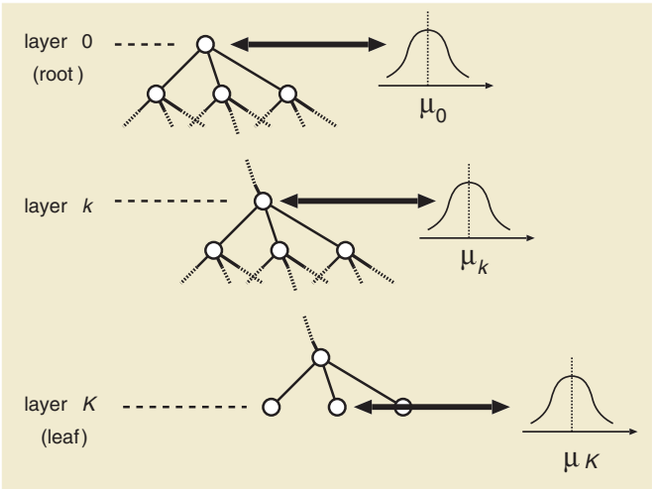

図-5　HMM適応のためのパラメータ共有構造

構造的アプローチ

前節で述べたように，パラメータ共有構造の選択は話者適応化において本質的に重要である．もし，パラメータ共有の構造が固定されている，すなわち，推定すべき（実効的な）自由パラメータ数がデータ量の多少にかかわらず変化しない場合には，想定された範囲と異なる量の適応データ量を与えたときに，かえって認識性能が劣化する可能性がある．これは前章で述べた音声モデリングにおける問題と同様の問題である．実用においては，事前にデータ量を知ることができない場面がほとんどなので，対策として考えられることは，データ量の多少により適応化手法を切り替えることである．しかしながら，切り替えのタイミングを適応語彙や話者の違いに対して頑健に設定することは甚だ困難である．データ量に依存しない，シームレスな適応手法が望まれる．

この問題を解決するために，パラメータの階層的な共有構造（木構造）を作成し，利用できるデータ量に応じてパラメータ共有の程度を変化させる手法がいくつか提案されている．以下，例として，自律的制御を用いた話者適応化（Automatic Model Complexity Control; AMCC）を説明する．この手法では，事前に図-5に示すような，HMMにおける分布の木構造を作成する．木構造を作成する際に用いる分布間距離としては，出力正規分布間の（対称化した）Kullback-Leibler擬距離を使う．ここでKは階層の数（木の深さ）であり，リーフ層（第K層）のノードはHMMの各分布（より正確には各状態に付随する混合正規分布の各混合成分）に1対1に対応する．ルートノード（第0層）は，すべての分布の集合に対応している．中間層のノードは分布の部分集合に対応しており，その要素は，その下層のリーフノードに対応する分布すべてである．木構造の各々のノードN_kに1つの適応パラメータが付随し，そのパラメータが対応する部分集合G_kに属する分布間で共有される．データ量が少ないときには，上位ノードに付随する大局的な適応パラメータを用い，データ量が多くなるに従い，下部のノードに付随するより局所的な適応パラメータを用いることで，データ量による共有構造の切り替えを実現している．

音声認識でしばしば用いられるN-gram言語モデルでは，あるN-gramの出現頻度がきわめて小さく統計的に信頼できない場合には，代わりに(N-1)-gramの出現頻度を修正して用いるバックオフ手法が一般に用いられている．この構造的アプローチも，これと類似の考え方に基づいており，音響的にバックオフを行う手法であると見なすことができる．

構造的事後確率最大化法

最近，篠田とLeeが提案した構造的事後確率最大化（Structural Maximum A Posteriori; SMAP）法[3]は，前述の構造的アプローチをさらに一歩進めたものである．このアプローチでは，データ量が大きい場合のMAP推定の漸近性を保ちつつ，データ量が少ないときには木構造による柔軟なパラメータ共有を行う．この手法では，前節で述べたAMCCと同様の分布木構造を用い，各ノードには多次元正規分布が割り当てられる．あるノードの分布の事前分布としてその親ノードのパラメータを用い，ノードパラメータのMAP推定をルートノードから順にリーフノードまでカスケード的に行う．音響的バックオフにスムージング手法を取り込んだ枠組みと捉えることができる．また，データ量が大きくなるに従い，MAP推定あるいはML推定時の認識性能に漸近的に（上から）

近づくという特長もある．

ここでは，HMMのある状態の1つ混合成分に対応する正規分布$\mathcal{N}(\boldsymbol{x}|\mu,\Sigma)$に注目し，その平均ベクトル$\mu$の適応前後の差分ベクトル$\Delta$を推定する方法を説明する．HMMの他の分布についても同じ方法が適用できる．今，注目している分布に対応する，ルートからリーフまでのノード列を$\{N_0,\ldots,N_k,\ldots,N_k\}$とする．ここで$N_0$はルートノード，$N_k$は混合成分に対応するリーフノードである．ノード$N_k$に付随する差分ベクトルを$\Delta_k$とする．ここで，ノード$N_{k-1}$の差分ベクトル$\Delta_{k-1}$が求まっているときに，それを事前分布のパラメータとして用いてノードkの差分ベクトルΔ_kを求めることを考える．ここで，主に計算の簡便さのために，差分ベクトルの事前分布としては，自然共役事前分布である正規分布をとる．このとき，差分ベクトルΔ_kのMAP推定量は以下のように計算される．

$$\Delta_0=\tilde{\Delta}_0 \tag{7}$$

$$\Delta_k=\frac{\Gamma_k\tilde{\Delta}_k+\tau_k\Delta_k-1}{\Gamma_k+\tau_k},\ k=1,\ldots,K, \tag{8}$$

ここで，$\tilde{\Delta}_k$は，Δ_kの最尤推定量であり，データから求められる．$\tau_k>0$は制御パラメータである．Γ_kはノードkに対応するデータサンプル数に相当する量である．式(8)をルートノードからリーフノードへとカスケード的に適応していくことにより，リーフノードN_k，すなわち，HMMの混合成分に対応する，差分ベクトルΔ_Kが求まり，それより，適応後の平均ベクトルμ'_Kが求まる．実用上は，認識性能は制御パラメータτ_kにさほど敏感ではなく，木の深さkによらず一律に決めても問題は少ない．

さて，これらの式がどのような意味を持つのかをさらに詳しく見ることにする．式(8)を用いた簡単な計算の後，リーフノードの差分ベクトルΔ_Kを求める式を以下のように書くことができる．

$$\hat{\Delta}_K=\sum_{k=0}^{K}w_k\hat{\Delta}_k, \tag{9}$$

すなわち，求めるべき差分ベクトルは木構造における先祖ノードにおける差分ベクトルの最尤推定値の重み付け和で表される．ここで，各階層kの差分ベクトルに対する重み係数w_kは以下の式で表される．

$$w_k=\frac{\Gamma_k}{\Gamma_k+\tau_k}\prod_{i=k+1}^{K}\frac{\tau_i}{\Gamma_i+\tau_i}. \tag{10}$$

ノードN_kに対する重みw_kは，そのノードに対応するデータ量Γ_kが大きくなるに従い大きくなり，また，kが小さくなるに従い小さくなる．すなわち，データ量が少ないときには，大局的な構造を表す上位層のパラメータの寄与が大きく，データ量が多くなるに従い，局所的な構造を示す下位層のパラメータが支配的になる．

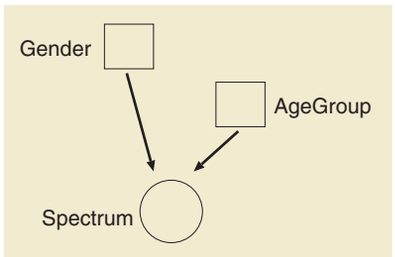

図-6　ベイジアンネットの例

■ダイナミックベイジアンネットを用いた音声モデリング

HMMは音声の時間的および周波数的な変動をモデルの内部状態間の遷移と各状態に応じた出力分布として表現することができ，音声認識における中心的な技術として研究されてきた．しかしHMMは基本的に単純なモデルであり，音声の特性に影響するさまざまな要因を取り込むことが難しい欠点がある．これに対しベイジアンネットは種々の確率変数間の多様な依存関係をグラフとして表現する，柔軟な確率的モデル化のフレームワークである．音声認識へのベイジアンネットの本格的な応用は比較的最近のことであり，1998年のZweigの研究が最初である[5]．本章ではベイジアンネットを用いた音声のモデル化方法および応用例を示す．

ベイジアンネット

ベイジアンネットは確率変数を表すノードと，確率変数間の直接的な依存関係を表す枝により定義される有向無サイクルグラフである．グラフの構造により確率モデルの構造が指定され，各ノードに割り当てられた条件付確率分布（Conditional Probability Distribution: CPD）によりモデルのパラメータが表現される．具体的なCPDの実現方法としては仮定する確率分布や，対象とするノードおよびその親ノードの確率変数が離散変数か連続変数かに応じて複数の方法が考えられる．たとえば，親および子ノードの確率変数がどちらも離散変数の場合はCPDとして条件付確率テーブル（Conditional Probability Table: CPT）を用いることができる．CPTは親ノードの値の組合せごとに子ノードが取る値の確率を表にしたものである．また，親ノードが離散変数で子ノードが連続変数の場合には，親ノードの値の組合せごとのガウス分布や混合ガウス分布などを用いることができる．

図-6に性別，年齢層，スペクトルの関係を表すベ

Female	Male
0.49	0.51

表-1　CPT : Gender

Child	Adult
0.3	0.7

表-2　CPT : AgeGroup

Gender	AgeGroup	Spectrum
Female	Child	N_1
Female	Adult	N_2
Male	Child	N_3
Male	Adult	N_4

表-3　CPD : Spectrum

イジアンネットの例を示す．図において*Gender*は性別，*AgeGroup*は年齢層を表す離散確率変数であり，*Spectrum*はスペクトルの特徴量ベクトルを表す連続確率変数である．離散確率変数に対応するノードは四角，連続確率変数に対応するすノードは丸で表してある．**表-1**に*Gender*および**表-2**に*AgeGroup*のCPTをそれぞれ示す．各行が単一の確率分布を表しており，和は1となる．この例ではどちらの変数も親ノードを持たないため，表は1行となっている．**表-3**にSpectrumのCPDを示す．ノード*Spectrum*は*Gender*および*AgeGroup*を親ノードとして持ち，それらの値の組合せは4通りであるから，CPDは4個の確率分布により定義される．表ではN_1からN_4までの4個の多次元ガウス分布を用いている．

ベイジアンネットにおいて任意のノードはその親ノードが与えられたとき，すべての非子孫ノードと条件付独立である．図-6に示した例では，簡単のため*Gender*, *AgeGroup*, *Spectrum*をそれぞれG, A, Sと表すことにすると，GはAの非子孫でありAの親集合は空であるから，条件付独立関係より式(11)が成り立つ．

$$P(A|G)=P(A) \tag{11}$$

他方，G, A, Sの同時確率分布は条件付確率の規則をくり返し適用することにより式(12)に示すように表すことができる．式(12)は条件付独立関係より導かれた式(11)を用いることにより，式(13)に示すように単純化できる．式(13)の各項は表-1, 2, 3に示したCPDに対応している．このようにベイジアンネットは同時確率分布を要素の積で表すコンパクトな表現となっている．

$$P(G, A, S)=P(G)P(A|G)P(S|G, A) \tag{12}$$
$$=P(G)P(A)P(S|G, A) \tag{13}$$

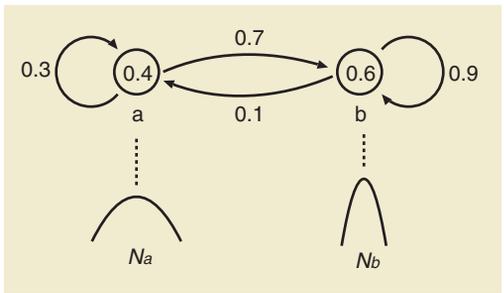

図-7　2状態のエルゴディックHMM

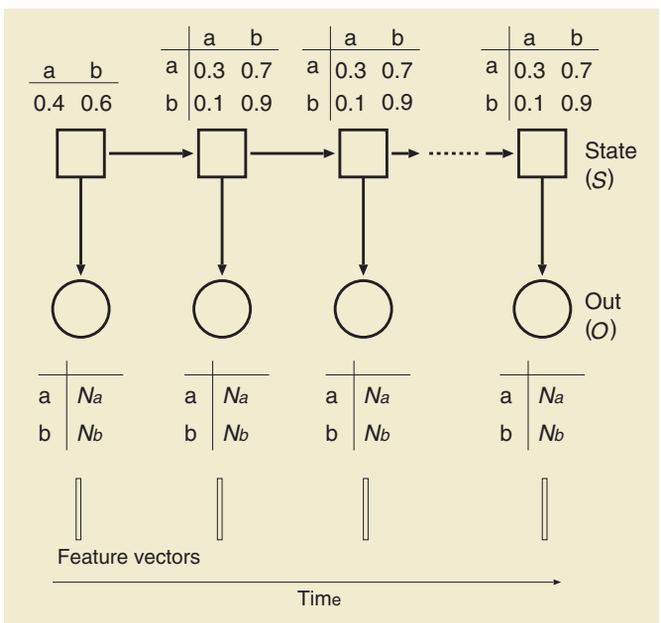

図-8　エルゴディックHMMをエミュレートしたDBN

認識システムの構成

ベイジアンネットは柔軟なフレームワークであり，HMMやN-gramを含め多様な確率モデルを表現することができるが，ここでは簡単のため**図-7**に示すエルゴディックHMMをエミュレートする方法を示す．よりさまざまなモデルの構成法は文献5)や文献6)などが詳しい．図のHMMは2状態aとbからなり，ノード内に示した初期状態確率および枝に示した状態遷移確率を持つ．また状態a,bの出力確率はそれぞれガウス分布N_a, N_bである．

図-8に，対応するベイジアンネットおよび各ノードのCPDを示す．HMMが対象とするデータは特徴量の時系列であるため，ベイジアンネットは対象とする時系列データの長さの分だけ時間方向に一定の構造が繰り返したネットワークとなる．このようなベイジアンネットを

ダイナミックベイジアンネット（DBN）という．ネットワークは繰り返し構造をとるため，DBNを定義するには始めの2スライス分の構造およびパラメータを指定すれば十分である．

図のDBNのタイムスライスはSおよびOの2個のノードを持つ．ノードSはHMMの状態を表しaとbの2通りの値をとる離散確率変数である．ノードOは特徴量ベクトルを表す連続確率変数である．ノードSのCPTは，始めのスライスにおいてはHMMの初期状態確率，2番目以降のスライスにおいては遷移確率に対応している．ノードOのCPDはSの値ごとのガウス分布であり，HMMの出力確率に対応している．HMMにおいてパラメータを推定する問題は，ベイジアンネットにおいてSおよびOのCPDを推定する問題となる．またHMMにおいて最尤状態系列を求める問題は，Oの値が与えられた条件でSにネットワーク全体の尤度が最大となるように値を割り当てる問題となる．

ベイジアンネットを用いることでさまざまなモデルを実現することができるが，モデルの構造が複雑になると非現実的な計算量となってしまう問題がある．このため厳密な確率推論を行う代わりに，近似解法を取り入れたアルゴリズムも研究されている．

音声認識への応用

ベイジアンネットを利用した新しい確率モデルによる音声のモデル化としては，Zweigによる発話クラスを表す2値変数を取り入れたモデルが挙げられる．またStephensonらは，通常の音響特徴量に加えてX線撮影により得られた調音変量を組み合わせた音響モデルを提案している[7]．日本における研究としてはMarkovらによる雑音のモデル化[8]や，篠崎らによる隠れ変数を用いた発話速度変動のモデル化の研究[9]などが挙げられる．

ツールの紹介

ベイジアンネットにおける学習データからのパラメータ学習や確率推論を行うためのアルゴリズムを実装した種々のツールキットが作られ，公開されている．これらのなかで音声認識への応用を特に意識して作成されたフリーなツールキットとしてGMTK[10]がある．大量のデータを扱う音声認識において必要とされる機能が充実している．

■まとめと今後の展望

以上，音声認識に対する新しい統計モデリング手法について説明した．これらの手法はいずれも現在発展途上のものであり，課題も多い．

モデル選択に関しては，最近，ベイズ法を用いたアプローチが新たに提案されており，そこでは，事前分布を適切に設定することで，MDL基準と異なり，データ量が少ない場合でも適用できるという利点がある．また，HMMは隠れ変数を持つ非正則なモデルであり，MDL基準は近似なしで直接適用することはできない．統計的学習理論において，非正則なモデルの複雑度を測る方法の出現が望まれる．

話者適応に関しては，今後より大量な音声データが蓄積されることが予想されており，それを有効に用いる手法，具体的には，EigenVoice法に代表される，話者ごとのパラメータの相関を利用する方法の進展が期待される．そこでは音響的特徴における音韻性と話者性との分離が大きな課題であり，音声の内在構造のより一層の解明が望まれる．

また，現在のところベイジアンネットを用いた音声認識は計算量が多く，多くの研究は小語彙のタスクを対象とするか，既存の認識システムを用いて生成した認識結果候補のリスト（N-best）に対して尤度を計算し直すかたちで行われている．しかしながら，HMMの長い歴史と比較してベイジアンネットが音声認識に応用されるようになったのはごく最近のことであり，今後の発展が期待される．

参考文献

1）韓　太舜，小林欣吾：情報と符号化の数理，培風館（1999）．
2）Shinoda, K. and Watanabe, T.: MDL-based Context-Dependent Subword Modeling for Speech Recognition, J. Acoust. Soc. Jpn.(E), Vol.21, No.2, pp.79-86 (2000).
3）Shinoda, K. and Lee, C.-H.: A Structural Bayes Approach to Speaker Adaptation, IEEE Trans. Speech Audio Processing, Vol.9, No.3, pp.276-287 (2001).
4）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（サーベイ論文），電子情報通信学会論文誌，Vol.J87-D-II, No.2, pp.371-386 (2004). IEEE Trans. Speech Audio Processing, Vol.9, No.3, pp.276-287, (2001).
5）Zweig, G.: Speech Recognition with Dynamic Bayesian Networks, Ph.D. Thesis, University of California, Berkeley (1998).
6）Bilmes, J.: Graphical Models and Automatic Speech Recognition, Technical Report UWEETR-2001-005, University of Washington, Dept. of EE, Seattle, WA (Nov. 2001).
7）Stephenson, T., Bourlard, H., Bengio, S. and Morris, A.: Automatic Speech Recognition using Dynamic Bayesian Networks with Both Acoustic and Articulatory Variables, Proc. ICSLP, pp.951-954 (2000).
8）Markov, K. and Nakamura, S.: Modeling HMM State Distributions with Bayesian Networks, Proc. ICSLP, pp.1013-1016 (2002).
9）Shinozaki, T. and Furui, S.: Time Adjustable Mixture Weights for Speaking Rate Fluctuation, Proc. EUROSPEECH, pp.973-976 (2003).
10）http://ssli.ee.washington.edu/~bilmes/gmtk/

（平成16年7月13日受付）