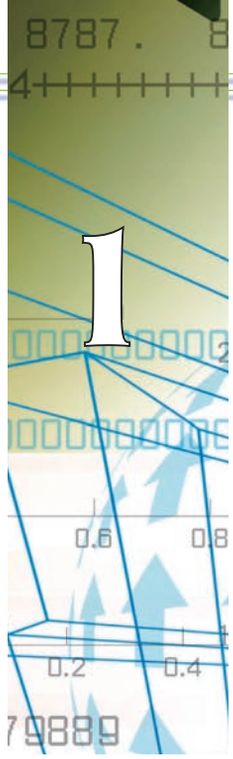

特集 音声情報処理技術の最先端

隠れマルコフモデルによる音声認識と音声合成

徳田 恵一
名古屋工業大学大学院工学研究科
tokuda@ics.nitech.ac.jp

音声認識の分野では，時系列の統計モデルである隠れマルコフモデル（Hidden Markov Model: 以下HMM）が音声パラメータ系列のモデル化手法として有効なことが知られ，実用的なシステムにおいても広く用いられている．本稿では，HMMの定義および関連するアルゴリズムについて，概説した上で，音声認識および音声合成におけるHMMの利用について述べる．また，HMMの限界を指摘した上で，次世代音声モデルとして期待される手法についても触れる．

■背景

音声認識における音声パラメータ系列のモデル化手法として，隠れマルコフモデル（Hidden Markov Model：以下HMM）が，広く用いられている．これは，大規模な音声データベースの整備と，計算機によるデータ処理能力の向上によるところが大きい．HMMの枠組みは，統計モデルという点では単純な考え方であり，数学的に取り扱いやすいという利点を持つ．加えて非常に柔軟であり，たとえば，コンテキスト依存モデル，動的特徴，混合ガウス分布，パラメータ共有手法／コンテキストクラスタリング手法，話者／環境適応化手法などの導入により，HMMに基づいた音声認識システムの性能は大きく改善されてきた．

音声合成においても，音声認識と同様の背景により，コーパスベースと称される音声合成方式の研究が盛んに行われるようになってきた．これらの方式は，従来の規則に基づいた合成方式の多くが発見的な手法に基づいているのに対し，大量のデータを用いた自動学習や音声単位選択に基づいているため，高品質で自然性の高い音声を合成しやすい，というだけでなく，システムの自動学習が可能，音声データ提供話者の個人性，さらには発話様式が合成音によく反映される，などの特徴を持つ．このような音声合成システムを構築する際に，音声認識で用いられてきたHMMが何らかのかたちで利用されることが多くなっている．

このような背景から，本稿では，HMMの定義および関連するアルゴリズムについて簡単にまとめた上で，音声認識および音声合成におけるHMMの利用法について解説することを目的とする．また，HMMの限界を指摘した上で，次世代音声モデルとして期待される手法についても触れる．

■隠れマルコフモデル（HMM）

HMMの定義

HMMは，図-1に示すように，出力ベクトル$\boldsymbol{o}_t$を出力する確率分布が$b_i(\boldsymbol{o}_t)$であるような信号源（状態）が，状態遷移確率$a_{ij}=P(q_t=j \mid q_{t-1}=i)$をもって接続されたものとして定義される．ただし，$i, j$は状態番号とする．音声関連の応用では，出力ベクトル$\boldsymbol{o}_t$は，MFCC, LPC

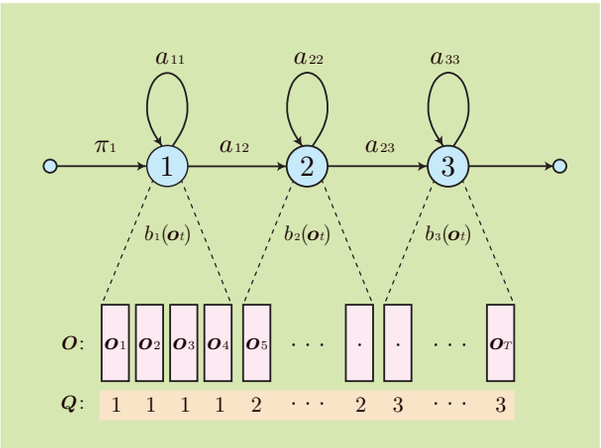

図 -1　隠れマルコフモデル（HMM）の例

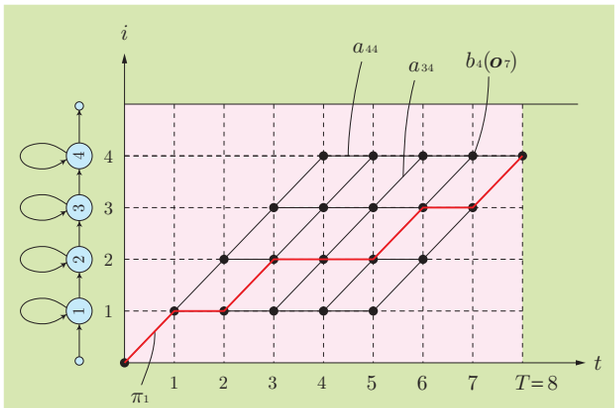

図 -2　HMM の状態遷移

ケプストラムなど，音声の短時間的なスペクトルを表現する音声パラメータである．HMMは時間方向とスペクトル方向の変動を統計的にモデル化しており，さまざまな要因で変動する音声のパラメータ系列の表現として適しているといえる．一般のHMMでは，任意の状態間での遷移が許されるが，音声のモデル化においては，因果性を表現するため，状態を横1列に並べたときに左方向への遷移がない（時間が逆戻りしない）モデル（left-to-right型モデルと呼ばれる）が用いられる．図-1もleft-to-right型モデルの一種である．

出力確率分布としては，多次元ガウス分布の重み付き和で表される多次元ガウス混合分布が用いられることが多いが，ここでは，簡単のため，単一の多次元ガウス分布を仮定する．このとき，ガウス分布の平均ベクトル μ_i と共分散行列 U_i が，出力確率分布 $b_i(o_t)$ を特徴付けるパラメータとなる．

HMMの状態数を N としたとき，HMMのパラメータ λ は，初期状態確率 $\pi=\{\pi_i\}_{i=1}^N$，状態遷移確率 $A=\{a_{ij}\}_{i,j=1}^N$，各状態 i での出力確率 $B=\{b_i(\cdot)\}_{i=1}^N$ により $\lambda=(\pi, A, B)$ で与えられる．このとき，状態が，$Q=\{q_1, q_2,..., q_T\}$ と遷移して，出力ベクトル系列 $O=[o'_1, o'_2,..., o'_T]'$ が出力される確率は，遷移確率と各状態での出力確率を掛け合わせることにより，

$$P(O,Q|\lambda)=\prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \qquad (1)$$

で与えられる．ただし，$'$ は，ベクトルの転置を表す．また，$a_{q_0 i}=\pi_i$ と置いた．したがって，出力ベクトル系列 $O=[o'_1, o'_2,..., o'_T]'$ が λ から出力される確率は，すべての可能な状態遷移の組合せについて和をとることにより，

$$P(O|\lambda)=\sum_{Q} P(O,Q|\lambda)$$
$$=\sum_{Q}\prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \qquad (2)$$

と書くことができる．式（1）は，図-2において，左下端のノードから，右上端のノードへ至る1本のパス上の確率をすべて掛け合わせたものである．各パス（たとえば，図中，赤色のパス）は，それぞれ1つの状態系列 Q に対応する．一方，式（2）は，可能なすべてのパスに対応する確率を加え合わせたものとなる．式（2）で定義される $P(O|\lambda)$ を効率的に計算するアルゴリズムがあり，時間順方向に計算するものを前向きアルゴリズム，時間逆方向に計算するものを後ろ向きアルゴリズムと呼ぶ．

HMMの学習

HMMのモデルパラメータ λ の学習は，与えられた学習用のベクトル系列 O に対して，式（2）で与えられる $P(O|\lambda)$ を最大にする λ を求めることである[☆1]．つまり，

$$\lambda_{\max}=\arg\max_{\lambda} P(O|\lambda) \qquad (3)$$

このようなパラメータ推定法は，最尤推定法あるいはML推定法と呼ばれる．

式（3）の最大化問題は，EMアルゴリズムに基づいて解くことができる．ここでのEMアルゴリズムは，何らかの初期モデルからスタートし，次式で定義される補助関数（Q 関数とも呼ばれる）

☆1 実際には，複数の学習用データ $\{O^{(1)}, O^{(2)},..., O^{(m)}\}$ により，1つのHMMの学習が行われることに注意する．

$$\mathcal{Q}(\lambda,\bar{\lambda}) = \sum_{Q} P(\boldsymbol{Q}|\boldsymbol{O},\lambda)\log P(\boldsymbol{O},\boldsymbol{Q}|\bar{\lambda}) \qquad (4)$$

を最大化する$\bar{\lambda}$を求め，$\lambda \leftarrow \bar{\lambda}$と置き換える操作を繰り返すかたちのアルゴリズムとなっている．関係式

$$\mathcal{Q}(\lambda,\bar{\lambda}) \geq \mathcal{Q}(\lambda,\lambda) \Rightarrow \boldsymbol{P}(\boldsymbol{O}|\bar{\lambda}) \geq \boldsymbol{P}(\boldsymbol{O}|\lambda) \qquad (5)$$

を示すことができ，このことから，上記アルゴリズムの繰り返しにより，$P(\boldsymbol{O}|\lambda)$の値が単調に増加することが保証され，$P(\boldsymbol{O}|\lambda)$の局所的最大点を求めることができる．なお，式(4)の最大化は，前向き・後ろ向きアルゴリズムを利用することにより，効率的に行うことができる．

小規模な孤立単語認識においては，1つのHMMが1つの単語に対応する単語モデルが用いられることもあるが，一般には，1つのHMMは，音素など，比較的短い音声単位をモデル化する．これは，大語彙の単語モデルを学習するための学習用音声データを収集することが容易ではないこと，また，多くの単語モデルを保持することは，記憶容量および計算量の観点から効率的ではないためである．

EMアルゴリズムにおける音素の初期モデルは，学習用の音声データに音素境界が与えられている場合，何らかの初期化手法により比較的簡単に得ることができる．音素境界が付与されていない場合には，音素境界の与えられた少量の音声データを用いて，初期モデルを作り，その後，音素境界の付与されていない大量の音声データにより，連結学習を行うことができる．連結学習は，発声内容に対応した音素の系列に従って，音素HMMを連結し，すべての学習データを使って，すべての音素HMMを同時に学習する方法である．学習用音声データすべてに音素境界が付与されている場合にも，その境界がモデル学習の観点から最適なものとは限らないため，境界付き学習の後，連結学習を行うのが普通である．

■HMMによる音声認識

最適状態系列の探索

音声認識は，与えられた$\boldsymbol{O}$に対して，任意の単語列W（ここでは「文」と呼ぶ）の中から，$P(W|\boldsymbol{O})$を最大にする単語列$W_{\max}$を求める操作である．つまり，

$$\begin{aligned} W_{\max} &= \arg\max_{W} P(W|\boldsymbol{O}) \\ &= \arg\max_{W} \frac{P(\boldsymbol{O}|W)P(W)}{P(\boldsymbol{O})} \\ &= \arg\max_{W} P(\boldsymbol{O}|\lambda_W)P(W) \end{aligned} \qquad (6)$$

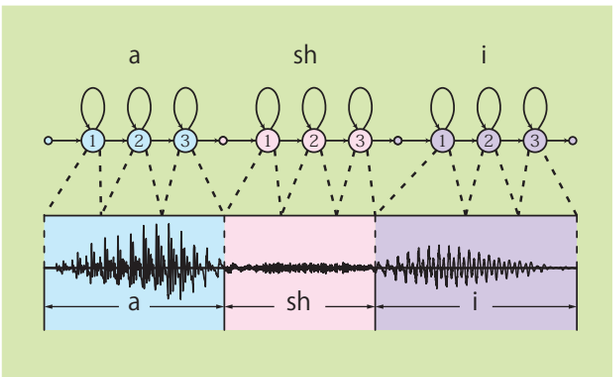

図-3　ビタビアルゴリズムによるセグメンテーション

Wに関する最大化に対して，$P(\boldsymbol{O})$は定数であること，また，$P(\boldsymbol{O}|W)$は，Wに対応するHMMをλ_Wとして，$P(\boldsymbol{O}|\lambda_W)$で与えられることに注意する．なお，$\lambda_W$は，通常，音素モデルを連結して作られる．式(6)は，ベイズ識別則に基づいており，$P(\boldsymbol{O}|\lambda_W)P(W)$が真の確率分布を表しているときには，文認識誤り率を最小にするという意味で最適である．式(6)における$P(W)$は言語モデルと呼ばれ，単語の部分列が出現する確率のモデルの積として与えられることが多い（N-gramモデル）．対応して，$P(\boldsymbol{O}|\lambda_W)$は音響モデルと呼ばれる．

式(6)において，$P(\boldsymbol{O}|\lambda_W)$の部分は，

$$\begin{aligned} P(\boldsymbol{O}|\lambda_W) &= \sum_{\boldsymbol{Q}} P(\boldsymbol{O},\boldsymbol{Q}|\lambda_W) \\ &\simeq \max_{\boldsymbol{Q}} P(\boldsymbol{O},\boldsymbol{Q}|\lambda_W) \end{aligned} \qquad (7)$$

で計算される．この近似はビタビ近似と呼ばれる．与えられたベクトル系列$\boldsymbol{O}$とλに対して，$P(\boldsymbol{O},\boldsymbol{Q}|\lambda)$を最大にする状態系列$\boldsymbol{Q}$と，そのときの$P(\boldsymbol{O},\boldsymbol{Q}|\lambda)$の値とを動的計画法に基づいて効率的に求めるのが，ビタビアルゴリズムである．これは，図-2において，$P(\boldsymbol{O},\boldsymbol{Q}|\lambda)$を最大にするという意味で最適なパスを1本見つけることに対応し，結果として図-3に示すような，状態と音声との時間的な対応関係を得ることができる．

ビタビ近似の下で式(8)の最大化問題は，

$$W_{\max} = \arg\max_{W} \max_{\boldsymbol{Q}} P(\boldsymbol{O},\boldsymbol{Q}|\lambda_W)P(W) \qquad (8)$$

と書くことができる．式(8)の最適化問題を，ビタビアルゴリズムを直接的に適用して解くことは，探索空間が膨大になり，容易ではないことから，ビームサーチ他の手法に基づいた実用的な探索アルゴリズムが数多く提案されている．

コンテキスト依存モデル

各音素のスペクトルパターンは，その前後の音素が何であるかにより，大きく変形を受けることが知られている．そのため，1つの音素に対して，その先行・後続音素に依存した複数のモデルを用意する．このようなモデルをコンテキスト依存モデルと呼ぶ．たとえば，「現実」という発声が次のような音素の系列になるとする．

```
g e N j i ts u
```

このとき，先行・後続音素を考慮したモデルの系列は，たとえば 以下のように表現することができる．

```
sil-g+e g-e+N e-N+j N-j+i j-i+ts i-ts+u
ts-u+sil
```

発声前後の無音区間も一種の音素と見なし，sil という名前を与えていることに注意する．このようなコンテキスト依存モデルは，音素の3組（トライフォン）に依存するため，トライフォンモデルと呼ばれる．各トライフォンモデルは，中心の1音素分の時間長だけをモデル化することに注意する．対応して，先行・後続音素に依存しない通常の音素モデルはモノフォンモデルと呼ばれる．

通常，音素は数十種類あるため，組合せによりトライフォンモデルの総数は膨大なものになる．それに伴い，各モデル当たりの学習データは極端に少なくなり，適切なモデルパラメータを推定することが難しくなる．さらに，大量の学習用音声データを用意しても，すべてのトライフォンがデータ中に出現することは期待できず，学習データに存在しないトライフォンに対応するモデルをつくることができないという問題が起こってくる．

このためコンテキストのクラスタリングが行われる．コンテキストクラスタリングは，類似したコンテキスト（つまり，先行あるいは後続音素）を持つトライフォンを同一視することであり，総モデル数は，モノフォンモデルのそれとトライフォンモデルのそれとの中間となる．クラスタリングは，通常，自動的に行われ，トップダウンに行う方法とボトムアップに行う方法があるが，いずれにせよ，学習データに出現しなかったトライフォンをどのクラスタに割り当てるかが一意に定まる必要がある．決定木に基づいたクラスタリングは，音韻学的な知識に基づいて，このような要請を自然なかたちで満たすことができるため，広く用いられている．決定木に基づいたクラスタリングでは，音韻に関する質問によりクラスタを2分していき（図-4），一種の回帰木を構築する．木をルートノードから辿ることにより，すべてのコンテキストは，必ずいずれかのリーフノードに属することになる．クラスタリングは，図-4に示したように，モデルごとではなく，モデルの状態位置ごとに別々のクラスタリングが行われることが多い．これは，各モデルの左端の状態が先行音素の影響を受けやすいのに対し，右端の状態は後続音素の影響を受けやすいためである．

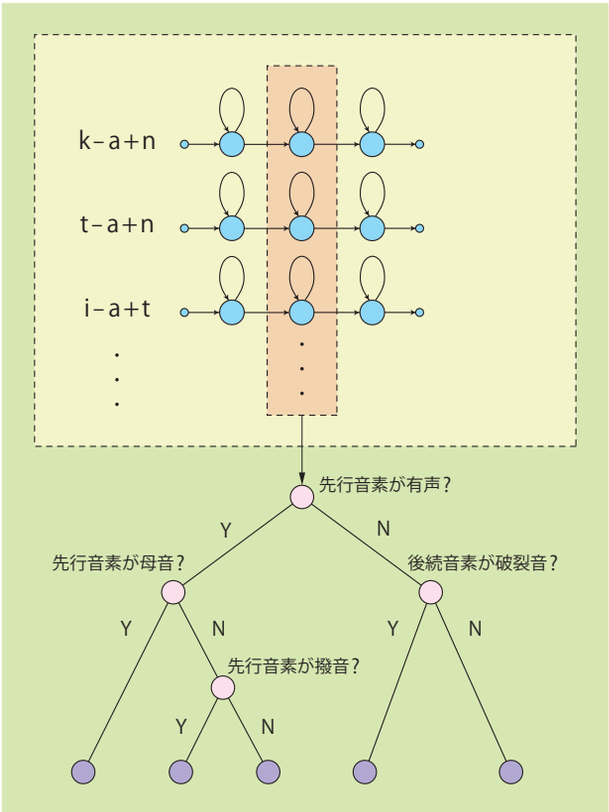

図-4　HMM のコンテキストクラスタリング

なお，HMMの基礎に関しては文献1）～7）などを参照されたい．また，HMMに関連したアルゴリズムの多くは，ツールキットとして利用することができ[8]，そのマニュアルもよい解説書となっている．

■音声合成におけるHMMの利用

単位選択型音声合成におけるHMMの利用

単位選択型の音声合成方式は，大量の音声データベースを音声合成システム内に保持しておき，合成したいテキストに対応する音声単位をそこから選択し，接続することにより，任意のテキストを合成するものである（図-5）．音声単位の選択は，適切に定義されたターゲットコストおよび接続コストと呼ばれるコストの総和を最小にするように行われ，そこではビタビアルゴリズムに類似した動的計画法に基づいたアルゴリズムが用いられる．

単位選択型音声合成方式においては，大量の音声データをあらかじめ音素などの音声単位にセグメンテーショ

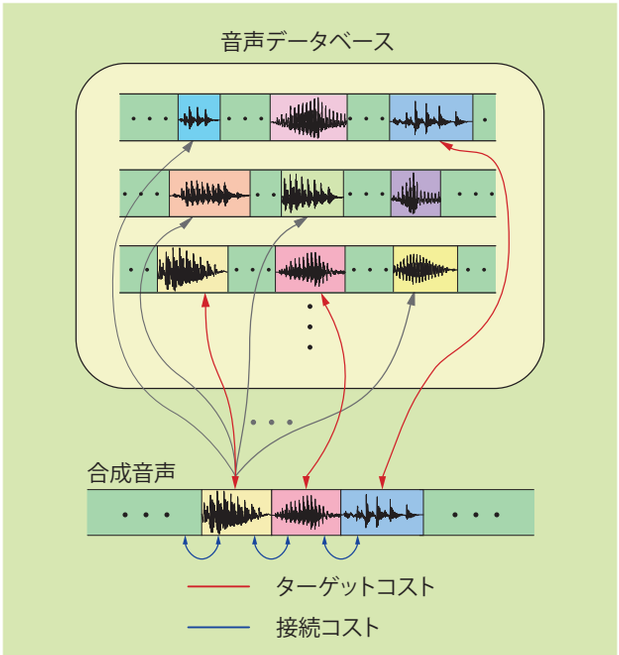

図 -5　単位選択型音声合成システム

ンしておく必要がある．しかし，これをすべて人手で行うことは容易ではないため，HMMを用いた自動セグメンテーションが広く利用されている．つまり，発声に対応する音素の列に従ってHMMを連結し，ビタビアルゴリズムにより，音声パラメータ系列$\boldsymbol{O}$と状態系列$\boldsymbol{Q}$との対応付けを行い，その結果として音素境界を得る（図-3）．

音声単位の長さは，音節，ダイフォン（音素の中心から次の音素の中心まで），音素，可変長単位などが考えられる．連続した音声データから音声単位を選択する場合には，音声単位は短いほど可能な接続点の候補が増えるため，計算量の増加を厭わなければ，接続歪の小さい接続ができる可能性が高くなる．このような観点から，ハーフフォン（音素の半分）を単位として用いているシステムもある．HMMによるセグメンテーションによれば，音素内をさらに細かく（状態に対応する長さで）分割することができるため，これを接続の単位とすることも可能である．このようなシステムでは，HMMの持つ統計量やコンテキストクラスタリング結果を利用することにより，単位選択のためのコストを定義する，音声データベースの量をあらかじめ削減する，動的計画法のための予備選択を行う，などが可能である．

HMMからのパラメータ生成に基づく音声合成

単位選択型の音声合成方式では，音声波形を接続することにより，高品質な合成音声を得ることができるが，さまざまな話者性や発話スタイル，さらには，感情表現を伴った音声を合成することは容易ではない．これに対し，HMMからのパラメータ生成に基づく音声合成方式は，多様な合成音声が比較的容易に実現可能な方式である．

図-6にシステムの基本構成を示す．学習部は，音声認識における学習部とほぼ同様である．大きな違いは，HMMの出力ベクトルがスペクトルパラメータだけではなくF0（基本周波数）に関するパラメータを含むことである．それに伴い，F0に影響を与えると考えられる，アクセント型，品詞，文の長さ，文内での位置，などをコンテキストに含めている．合成部では，与えられた入力テキストをコンテキスト依存ラベル（コンテキスト依存モデル名）の列に変換し，各ラベルに対応するコンテキスト依存音素HMMを順に連結することにより，入力テキストに対応する1つの文HMMを作る．そして，文HMMから，以下で述べる音声パラメータ生成手法を用いて，F0パラメータ系列とスペクトルパラメータ系列を生成し，ソースフィルタ理論に基づく分析合成（ボコーダ）方式により波形を生成する．

音声パラメータ系列$\boldsymbol{O}$の生成は，音素HMMを連結することにより作られたHMMをλ_Wとしたとき，$P(\boldsymbol{O}|\lambda_W)$を最大とするような$\boldsymbol{O}$を定めることによって行われる．この最大化問題は，式（7）と同様のビタビ近似を適用することにより，

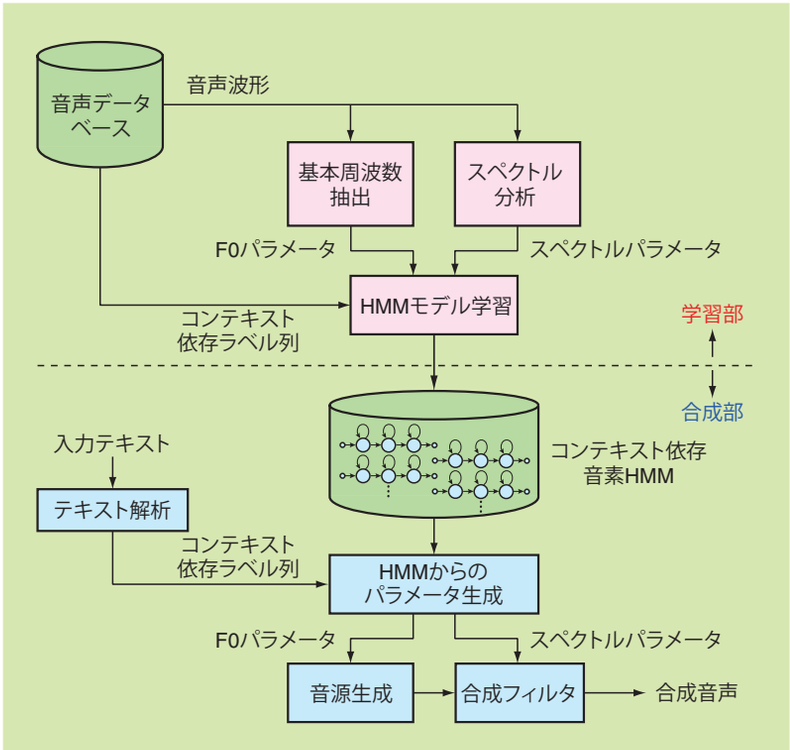

図-6　HMM からのパラメータ生成に基づく音声合成システム

$$\boldsymbol{O}_{\max} = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\lambda_W) \tag{9}$$

$$\simeq \arg\max_{\boldsymbol{O}} \max_{\boldsymbol{Q}} P(\boldsymbol{O},\boldsymbol{Q}|\lambda_W) \tag{10}$$

と書くことができる．音声認識における問題，式 (6) あるいは式 (8) と，ここでの問題，式 (9) あるいは式 (10) は，それぞれ対称的な関係となっている．さらに，

$$P(\boldsymbol{O},\boldsymbol{Q}|\lambda_W) = P(\boldsymbol{O}|\boldsymbol{Q},\lambda_W)P(\boldsymbol{Q}|\lambda_W) \tag{11}$$

と書けることから，$\boldsymbol{Q}$ を $P(\boldsymbol{Q}|\lambda_W)$ だけに基づいて定めた後，$\boldsymbol{O}$ を定めることにすれば，式 (10) の最適化問題は次のように書くことができる．

$$\boldsymbol{Q}_{\max} = \arg\max_{\boldsymbol{Q}} P(\boldsymbol{Q}|\lambda_W) \tag{12}$$

$$\boldsymbol{O}_{\max} = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{Q}_{\max},\lambda_W) \tag{13}$$

式 (12) に関しては，適当な継続長分布を導入することにより解くことができる☆2．一方，式 (13) の最大化は，状態系列が与えられたときの最適パラメータ系列決定問題であり，このような $\boldsymbol{O}$ は各状態の出力確率分布の平均ベクトルの系列で与えられる．すなわち，生成されるパラメータ系列は，1つの状態が継続している間は一定の値（平均ベクトル）をとり，状態の遷移のたびに不連続に変化することになり（図-7 (a)），合成音声にも耳障りな不連続が生ずる．

この問題は，音声認識で広く用いられている動的特徴を考慮することにより解決される．つまり，出力ベクトル $\boldsymbol{o}_t$ は，音声スペクトルパラメータベクトル $\boldsymbol{c}_t$ と，動的な特徴ベクトル $\Delta\boldsymbol{c}_t$ および $\Delta^2\boldsymbol{c}_t$ で構成され，$\boldsymbol{o}_t=[\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$ で表されるとする．$\Delta\boldsymbol{c}_t, \Delta^2\boldsymbol{c}_t$ は，それぞれ，$\boldsymbol{c}_t$ の時間方向に関する1次微分，2次微分に対応するよう，$\{\boldsymbol{c}_{t+l}\}_{l=-L}^{L}$ を線形結合することにより計算されることから，$\boldsymbol{C}=[\boldsymbol{c}_1',\boldsymbol{c}_2',\ldots,\boldsymbol{c}_T']'$ とおけば，$\boldsymbol{O}$ は，$\boldsymbol{O}=\boldsymbol{K}\boldsymbol{C}$ のかたちで与えられることになる．ただし，$\boldsymbol{c}_t$ を M 次元とすれば，$\boldsymbol{K}$ は，$3TM\times TM$ の疎行列であり，1部の要素に動的特徴を計算するための係数を持つ．このとき，$P(\boldsymbol{O}|\boldsymbol{Q}_{\max},\lambda_W)$ を最大にする $\boldsymbol{C}$ は，

$$\frac{\partial \log P(\boldsymbol{K}\boldsymbol{C}|\boldsymbol{Q}_{\max},\lambda_W)}{\partial \boldsymbol{C}} = \boldsymbol{0} \tag{14}$$

と置くことによって得られる線形方程式を解くことで定められる．このように生成された出力ベクトル（ここで

☆2 状態継続長分布は，状態 i が d_i 回継続する確率 $p_i(d_i)$ で与えられ，図-1の構造を持つHMMの場合には，各状態の継続長が状態系列 $\boldsymbol{Q}$ を一意に定める．なお，継続長分布もコンテキスト依存とする．

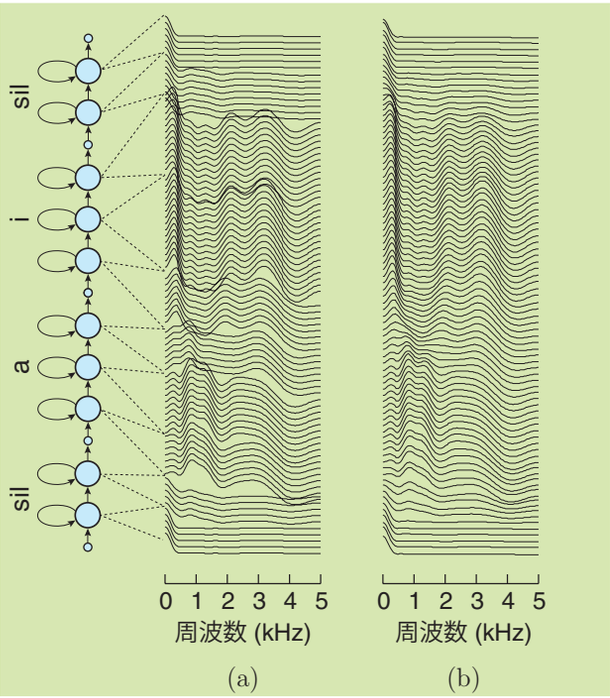

図-7　音素モデル sil，a，i，sil を接続することにより作られた HMM からのパラメータ生成例.
（a）動的特徴なし，（b）動的特徴あり

はメルケプストラム）から計算されたスペクトルの例を図-7（b）に示す．動的特徴を用いない場合には，状態が継続する間，一定のスペクトル形状をとり，状態が遷移するときに不連続な変化を起こしているのに対して，動的特徴を考慮した方法では，滑らかに変化するスペクトル系列が得られている様子が分かる．

HMMからのパラメータ生成に基づく音声合成については文献9），10）を，また，関連したツールキットについては文献11）を参照されたい．なお，コーパス音声合成の歴史と展望については文献12），13）に詳しい．

■HMMの限界と今後

HMMは，音声パラメータ系列のモデル化手法としての有効性が知られ，実用的な音声認識システムにおいて広く用いられるようになったが，その一方でHMMにより達成可能な性能の限界についても明らかとなり，次世代音声モデルの必要性が認識されるに至っている．HMMの主たる制約は，

1. 状態継続区間内での定常性を仮定している．
2. 出力ベクトル間の相関を考慮できない．
3. 時間構造のモデル化能力が弱い．

などである．これらの問題を解決するため，これまでにも，セグメントモデル，トラジェクトリモデル，セグメント入力モデル，隠れトラジェクトリモデル，部分隠れマルコフモデル，隠れセミマルコフモデル，動的システムモデル，スイッチング状態空間モデル，条件付分布モデル，HMMトラジェクトリ法，トラジェクトリHMMなど，さまざまな モデルの提案がなされており，成果が出始めている段階といえる．これらは，総称であったり，個々のモデル名であったりし，また，関連の深いモデルが異なった名前で呼ばれることもあるが，これらの間の関係についても整理されつつある．

HMMに関連した効率的な各種アルゴリズムは，上に列挙した制約に基づいている．したがって，これらを緩和すれば，その代償として計算量の増加が避けられないが，今後の計算機能力の向上がこれを補っていくものと予想される．

そのほかにも，ベイズ的アプローチの導入，ベイジアンネットワークによる確率変数間の依存関係の表現，変分近似あるいはパーティクルフィルタによる学習アルゴリズムなど，統計的機械学習の手法の導入も盛んとなっている．また，モデルの構造や学習アルゴリズムだけでなく，学習基準そのものについても，最小識別誤り学習，最大相互情報量基準推定，最大エントロピーモデル，ベイズ学習など，さまざまな アプローチが研究されている．

しばらくは，「HMMの次」が何に落ち着くのか，目が離せない状況といえよう．今後の展開を楽しみとしたい．

参考文献

1）中川 聖一：確率モデルによる音声認識，電子情報通信学会 (1988).
2）今井　聖：音声認識，共立出版 (1995).
3）Rabinar, L. and Juang, B.-J. 著，古井貞熙 監訳：音声認識の基礎（上）・（下），NTTアドバンストテクノロジ (1995).
4）北　研二，中村　哲，永田昌明：音声言語処理，森北出版 (1996).
5）古井貞熙：音声情報処理，森北出版 (1998).
6）鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄 編著：音声認識システム，オーム社 (2001).
7）安藤彰男：リアルタイム音声認識，電子情報通信学会 (2003).
8）http://htk.eng.cam.ac.uk/
9）徳田恵一：隠れマルコフモデルの音声合成への応用，信学技報，SP99-61, pp.48-54 (Aug. 1999).
10）小林隆夫，徳田恵一：講座 コーパスベース音声合成技術の動向［IV］選－HMM音声合成方式－，電子情報通信学会誌，Vol.87, No.4 (Apr. 2004).
11）http://hts.ics.nitech.ac.jp/
12）匂坂芳典：コーパスベース音声合成，Journal of Signal Processing, Vol.2, No.6 (Nov. 1998).
13）広瀬啓吉：21世紀に向けての音声合成の技術展望，IPSJ Magazine, Vol.41, No.3 (Mar. 2000).

（平成16年7月13日受付）