HMMによる音声合成の基礎

徳田 恵一

名古屋工業大学

知能情報システム学科

〒 466-8555 名古屋市昭和区御器所町

近年，隠れマルコフモデル (hidden Markov model: HMM) は音声認識の一般的な手法となってきた。音声合成においても，大量の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を背景に，コーパスベースと呼ばれる音声合成方式，あるいは音声合成システム構築法が数多く提案されている．このようなシステムを構築する際に，HMM が果たす役割が大きくなっていることから，本文では、HMM を音声合成に利用する方法について，HMM 自身から音声を合成する手法とも関連付けながら，概説することを目的とする。

音声合成，テキスト音声合成，隠れマルコフモデル，HMM，コーパス，

FUNDAMENTALS OF SPEECH SYNTHESIS BASED ON HMM

Keiichi Tokuda

Department of Computer Science

Nagoya Institute of Technology

Gokiso-cho, Shouwa-ku, Nagoya, 466-8555 Japan

The increasing availability of large speech databases makes it possible to construct speech synthesis systems, which are referred to as corpus-based approach, by applying unit selection and statistical learning algorithms. In constructing such a system, the use of hidden Markov models (HMMs) has arisen largely. This paper aims to describe such approaches in relation to an approach in which synthetic speech is generated from HMMs themselves.

speech synthesis, Text-to-Speech Translation, hidden Markov model, HMM, corpus

1. まえがき

近年，大量の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を背景に，隠れマルコフモデル (hidden Markov model: HMM) に代表される統計的手法が，音声認識の一般的なアプローチとなってきた。HMM の枠組は，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もつ．加えて非常に柔軟であり，例えば，コンテキスト依存モデル [1]，動的特徴 [2]，混合ガウス分布 [3]，tying 手法/コンテキストクラスタリング手法 (例えば [4])，話者/環境適応化手法 (例えば [5], [6]) 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を大きく改善してきた．

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（あるいは speaker-driven, trainable など) と称される音声合成方式，あるいは音声合成システム構築法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，更には発話様式が合成音によく反映される，などの特徴をもつ．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 が何らかの形で利用されることが多くなってきており，その利用形態は，

(1) 音声データベースのトランスクリプションやセグメンテーションに用いるもの (例えば [7]).
(2) HMM の尤度や，HMM におけるコンテキストクラスタリングの結果を利用して，音声データベースの中から，音声単位の inventory を選ぶもの (例えば [8], [9]).
(3) ランタイムに，HMM の尤度や，HMM におけるコンテキストクラスタリングの結果を利用して，複数の instance から選ぶもの (例えば [10], [11]).
(4) HMM 自身から音声を合成しようとするもの (例えば [12]–[14]).

などに分類することができる．(1), (2), (3) はいずれも，音声単位の接続に基づいた手法における HMM の利用であり，これらの手法では，PSOLA 法 [15] などの利用により（波形レベルで）自然性の高い合成音声が得られる利点がある．一方，(4) では，合成音声が，いわゆる vocoded speech となる欠点があるものの，HMM のパラメータを適切に変換することにより，データベース中に存在しない様々な音声を出力できる可能性がある．

このような背景から，本文では，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，上記 (4) の HMM 自身から音声を合成しようとする方式についても，他の手法と関連づけながら述べる．

以下，**2.** において，HMM に関する基本的な事項についてまとめる．**3.** では，HMM の利用について，特に音声単位選択方式における利用法を中心に述べる．**4.** で，HMM からの音声合成手法について述べ，**5.** で結論を述べる．

2. 隠れマルコフモデル (HMM)

2.1 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 $a_{ij} = P(q_t = j|q_{t-1} = i)$ をもって接続されたものとして定義される．但し，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 [16], LPC 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，様々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．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とにする．つまり，

$$\begin{aligned} b_i(\boldsymbol{o}) &= \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_i, \boldsymbol{U}_i) \\ &= \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{U}_i|}} \exp\left\{\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu}_i)'\boldsymbol{U}_i^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_i)\right\} \end{aligned} \tag{1}$$

ただし，$'$ は，行列の転置を表す．この場合，ガウス分布の平均ベクトル $\boldsymbol{\mu}_i$ と共分散行列 $\boldsymbol{U}_i$ が，出力確率分布 $b_i(\boldsymbol{o})$ 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 $A =$

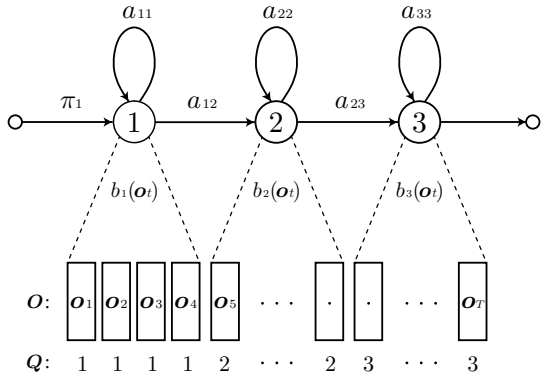

図 1　隠れマルコフモデル (HMM) の例

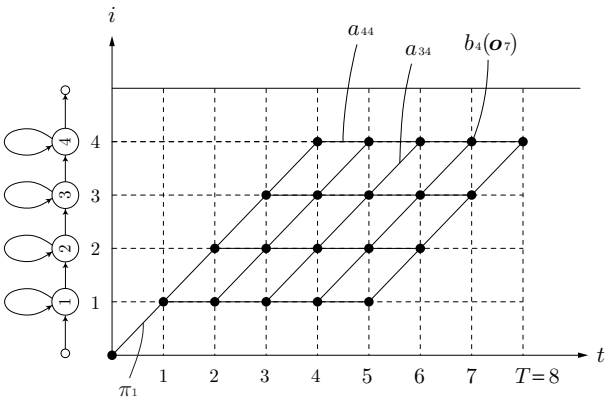

図 2　HMM の状態遷移

$\{a_{ij}\}_{i,j=1}^{N}$，各状態 i での出力確率 $B=\{b_i(\cdot)\}_{i=1}^{N}$ により $\lambda=(A,B,\pi)$ と与えられる．このとき，状態が，$\boldsymbol{Q}=\{q_1,q_2,\ldots,q_T\}$ と遷移して，出力ベクトル系列 $\boldsymbol{O}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\ldots,\boldsymbol{o}_T)$ 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，

$$P(\boldsymbol{O},\boldsymbol{Q}|\lambda)=\prod_{t=1}^{T}a_{q_{t-1}q_t}b_{q_t}(\boldsymbol{o}_t) \tag{2}$$

で与えられる．但し，$a_{q_0 i}=\pi_i$ とおいた．従って，出力ベクトル系列 $\boldsymbol{O}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\ldots,\boldsymbol{o}_T)$ 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，

$$\begin{aligned}P(\boldsymbol{O}|\lambda) &= \sum_{\text{all }\boldsymbol{Q}}P(\boldsymbol{O},\boldsymbol{Q}|\lambda)\\ &= \sum_{\text{all }\boldsymbol{Q}}\prod_{t=1}^{T}a_{q_{t-1}q_t}b_{q_t}(\boldsymbol{o}_t)\end{aligned} \tag{3}$$

と書くことができる．式 (2) は，図 2 において，左下端のノードから，右上端のノードへ到る 1 本のパス上の確率をすべて掛け合わせたものである．一方，式 (3) 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．

2.2　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 $\boldsymbol{O}$ に対して，式 (3) で与えられる観測尤度 $P(\boldsymbol{O}|\lambda)$ を最大にする λ を求めることである[1]．つまり，

$$\lambda_{\max}=\arg\max_{\lambda}P(\boldsymbol{O}|\lambda) \tag{4}$$

このためのアルゴリズムは，EM アルゴリズムに基づいて導出することができ，Baum-Welch 再推定式と呼ばれる．

何らかの初期モデルから始めて，再推定式により与えられる $\overline{\lambda}$ を新たな λ とする操作を繰り返すことにより，$P(\boldsymbol{O}|\lambda)$ の値（$\boldsymbol{O}$ に関する λ のゆう度）が単調に増加することが保証されており，$P(\boldsymbol{O}|\lambda)$ の局所的最大点を求めることができる．一般に，ひとつの HMM は，音素などの比較的短い音声単位をモデル化する．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には，セグメンタル k-means 法によって得ることができる．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，初期モデルをつくり，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 (embedded training) を行う．連結学習は，トランスクリプション（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，連結学習を行うのが普通である．

2.3　最適状態系列の探索

音声認識は，I 個の単語あるいは文章などに対応する HMM を，$\lambda_1,\lambda_2,\ldots,\lambda_I$ として[2]，与えられた $\boldsymbol{O}$ に対して，

$$\begin{aligned}i_{\max} &= \arg\max_{i}P(\lambda_i|\boldsymbol{O})\\ &= \arg\max_{i}\frac{P(\boldsymbol{O}|\lambda_i)P(\lambda_i)}{P(\boldsymbol{O})}\end{aligned} \tag{5}$$

を求める操作である．その際，$P(\boldsymbol{O}|\lambda_i)$ の部分は，

$$\begin{aligned}P(\boldsymbol{O}|\lambda_i) &= \sum_{\boldsymbol{Q}}P(\boldsymbol{O},\boldsymbol{Q}|\lambda_i)\\ &\simeq \max_{\boldsymbol{Q}}P(\boldsymbol{O},\boldsymbol{Q}|\lambda_i)\end{aligned} \tag{6}$$

で計算される．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 $P(\boldsymbol{O},\boldsymbol{Q}|\lambda)$ の値を効率的に求めるのが，Viterbi アルゴリズムである．

2.4　コンテキスト依存モデルとクラスタリング

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，ひとつの音素に対して，その先行・後続音素（音素環境）に依存して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．例えば，「あらゆる現実を，すべて自分の方へねじ曲げたのだ。」の「現実」の部分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以下のように表現することができる．

[1] 実際には，複数の学習用データ $\{\boldsymbol{O}^{(1)},\boldsymbol{O}^{(2)},\ldots,\boldsymbol{O}^{(m)}\}$ により，一つの HMM の学習が行われることに注意する．

[2] これらは音素モデルを連結して作られる．

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

このようなモデルは，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．

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．それにともない，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．更に，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．クラスタリングは，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 [4] は，音韻学的な知識に基づいて，このような要請を自然な形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 3），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に属することになる．クラスタリングは，図 3 に示したように，モデル毎ではなく，モデルの状態位置毎に別々のクラスタリングが行われることが多い．

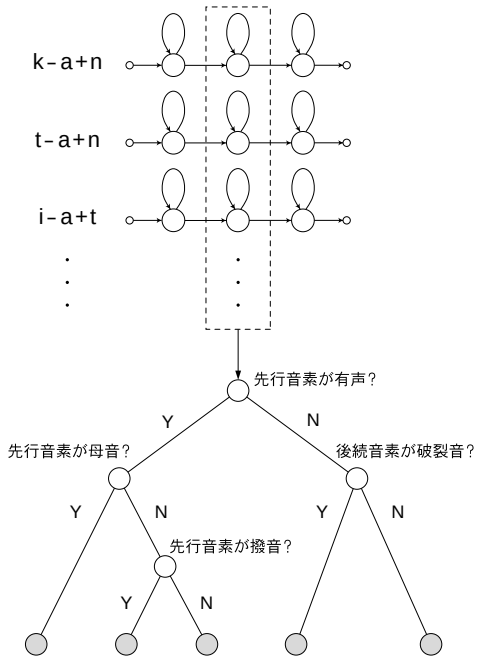

図 3　HMM のコンテキストクラスタリング

なお，HMM の基礎に関しては [17]–[22] などを参照されたい．英語になるが [23] も基本的である．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 [24]，そのマニュアルもよい解説書となっている．

3.　音声合成における HMM の利用

3.1　トランスクリプションとセグメンテーション

単位選択型の音声合成方式は，大量の音声データベースから，合成したい文章に対応する音声単位を選択し，接続することにより，任意の文章を合成するものである．従って，大量の音声データを音素などの音声単位にラベル付け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 を用いた自動的ラベリングが広く利用されている [7]．つまり，トランスクリプションに従って，HMM を連結し，**2.3** で述べた Viterbi アルゴリズムにより，音声の特徴ベクトル系列 $\boldsymbol{O}$ 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．HMM は，不特定多数の話者のデータにより学習しておいた不特定話者 HMM を用いるが，これを初期モデルとして，連結学習を行うことにより，セグメンテーションの精度を高めることができる．

音声データの発声内容からトランスクリプションへの変換は，発音辞書や変換規則に基づいて行うことができるが，無声化やポーズの挿入など，規則的に行うことができない部分もある．このような発音の変動を考慮した音素ネットワーク（複数音素の並列接続やポーズの挿入を許す）に対して Viterbi アルゴリズムを適用することにより，発音の変動を自動的に検出することができる．発声内容が未知の場合にも，HMM に基づいた音素認識を行うことにより，トランスクリプションを得ることができる．通常，合成音声の品質向上をはかるため，自動ラベリングの後，手修正が施されるが，システムの自動構築を目指し，手修正を行わない方式も多い．

音声単位の長さは，diphone, phone，可変長単位 [25] などが考えられるが，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，接続歪の小さい接続ができる可能性が高まる．このような観点から，文献 [26] では，half-phone を単位として用いている．HMM によるセグメンテーションによれば，音素内を更に細かく（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 [11]．

3.2　音声単位の選択とクラスタリング

音声単位のスペクトルに影響を与える変動要因は，主として音韻情報（音素の場合には，先行・当該・後続音

素）であり，これは，音声認識においてトライフォンを考えたことに対応する．音声単位の音響的な性質として，スペクトル[3]だけでなく，ピッチ，継続長なども考慮する場合には，音韻情報だけでなく，アクセントに関連した要因，当該音素の位置，品詞など，様々な変動要因を考慮する必要がある．これらの要因をすべて合わせたものを、ここでは「コンテキスト」と呼ぶことにする．例えば，[27] の日本語音声合成システムにおいては，音素のコンテキストとして，表 1 に示すものを考慮している．なお，音声認識においては，当該音素をコンテキストと考えることはしないが（認識結果の音素名がわからなくなるため），音声合成においては，当該音素名をコンテキストと考えても差障りはない．また，別モジュールで予測された韻律（ピッチや継続長など）を，ここで言うコンテキストと同様に扱うこともある ([28] など)．

コーパスベースの単位選択型音声合成における問題は，このようなコンテキストが与えられたときに，対応する音声単位を音声データベースの中から選び出すことである．選択は，システムの inventory 構築時に行う場合と，ランタイムに行う場合がある．いずれの場合も，音声認識のと同様，unseen コンテキストの問題に対処できなければならない．このため，単位選択型音声合成においては，

(a) コンテキストの適合性（target cost, context matching score などと呼ばれる）
(b) 音声単位の接続コスト（discontinuity, continuity cost, concatenation cost などと呼ばれる）

の二つのコスト関数を最小化するように音声単位を選び，接続することが目標とされる ([29], [28] など)．これを実現するための方法は，

(i) コンテキスト間の距離を定義し，それに基づいて合成時に使用する音声単位を選択するもの ([30], [28], [26] など）
(ii) 予めコンテキストクラスタリングを行い，合成時には，対応するクラスタ中の音声単位から選ぶもの ([9], [11], [31], [32], など）

に大別される．方法 (i) は，基準 (a) のコンテキスト適合性と基準 (b) の接続コストを同時に評価しながら，DP 法により文章に対応する音声単位の列を選択するものである．方法 (i) には，コンテキスト間距離の設定にヒューリスティクスが含まれる，合成時に必要となる音声データのサイズが大きくなる，という短所がある．一方，方法 (ii) では，要求されたコンテキストに対応するクラスタ内の instance すべてを候補として，基準 (b) の接続コストに基づいた DP 法により，文章に対応する音声単位の列を選択する．方法 (ii) には，スペクトル距離などの客観基準によりクラスタリングを行うことができる，決定木に基づいたクラスタリングを行うことにより unseen コンテキストに対する汎化作用が期待できる，各クラスタに適切な数の音声単位 (instance) をおくことによりランタイムに必要なデータサイズを制御しやすい，クラスタ中心からのずれを DP の際のコストに組み込むことができる，などの利点がある．

[3]パワーはスペクトルに含まれるとする．

表 1 コンテキストの例

- { 先行，当該，後続 } 音素
- 当該音素のアクセント句内でのモーラ位置
- { 先行，当該，後続 } の品詞，活用形，活用型
- { 先行，当該，後続 } アクセント句のモーラ長，アクセント型
- 当該アクセント句の位置，前後のポーズの有無
- { 先行，当該，後続 } 呼気段落のモーラ長
- 当該呼気段落の位置
- 文のモーラ長

方法 (ii) において，音声データベースのラベリングが HMM によって行われるのであれば，単位選択のためのクラスタリングも HMM に基づいて行うのは，システム全体に一貫性をもたせる意味で自然なことと考えられる．実際に，そのような観点からのシステムが提案されている ([9], [11] など)．

なお，コーパスベース音声合成の歴史と展望については，[33], [34] を参照されたい．

4. パラメータ生成に基づく HMM 音声合成

4.1 HMM からのパラメータ生成

音声パラメータ生成に基づいた HMM からの音声合成は，与えられた λ に対して，出力確率が最大となる長さ T の出力ベクトル系列を求めること，つまり，

$$\boldsymbol{O}_{\max} = \arg\min_{\boldsymbol{O}} P(\boldsymbol{O}|\lambda, T) \tag{7}$$

を基本とする．λ は，音素 HMM を連結することにより，つくられた文に対応する HMM である．ここでは，問題を簡単化するため，式 (6) と同様の近似を適用する．

$$\begin{aligned} \boldsymbol{O}_{\max} &= \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\lambda, T) \\ &\simeq \arg\max_{\boldsymbol{O}} \max_{\boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}|\lambda, T) \end{aligned} \tag{8}$$

更に，

$$P(\boldsymbol{O}, \boldsymbol{Q}|\lambda, T) = P(\boldsymbol{O}|\boldsymbol{Q}, \lambda, T) P(\boldsymbol{Q}|\lambda, T) \tag{9}$$

と書けることから，$\boldsymbol{Q}$ を $P(\boldsymbol{Q}|\lambda)$ だけに基づいて定めた後，$\boldsymbol{O}$ を定めることにすれば，式 (8) の最適化問題は次のように書くことができる．

$$\boldsymbol{Q}_{\max} = \arg\max_{\boldsymbol{Q}} P(\boldsymbol{Q}|\lambda, T) \tag{10}$$

$$\boldsymbol{O}_{\max} = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{Q}_{\max},\lambda) \quad (11)$$

式 (10) に関しては，**4.2** で検討することとし，ここでは，式 (11) の問題を解くことを考える．式 (2), (1) より，$P(\boldsymbol{O}\,|\,\boldsymbol{Q},\lambda,T)$ の対数は，

$$\begin{aligned}\log P(\boldsymbol{O}\,|\,\boldsymbol{Q},\lambda) &= \log\prod_{t=1}^{T} b_{q_t}(\boldsymbol{o}_t)\\ &= -\frac{1}{2}(\boldsymbol{O}-\boldsymbol{M})'\boldsymbol{U}^{-1}(\boldsymbol{O}-\boldsymbol{M}) - \frac{1}{2}\log|\boldsymbol{U}| \\ &\qquad + \text{Const} \end{aligned} \quad (12)$$

と書くことができる．ここで，

$$\boldsymbol{O} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']' \quad (13)$$

$$\boldsymbol{M} = [\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', ..., \boldsymbol{\mu}_{q_T}']' \quad (14)$$

$$\boldsymbol{U} = \mathrm{diag}[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, ..., \boldsymbol{U}_{q_T}] \quad (15)$$

であり，$\boldsymbol{\mu}_{q_t}$ と $\boldsymbol{U}_{q_t}$ はそれぞれ，状態 q_t の平均ベクトルと共分散行列である．式 (16)，(17) の制約を考えないとき，$P(\boldsymbol{O}\,|\,\boldsymbol{Q},\lambda)$ は $\boldsymbol{O}=\boldsymbol{M}$ のときに最大化されることは明らかである．これは，出力ベクトル系列が平均ベクトルの系列によって与えられることを意味する．

この問題は，音声認識で広く用いられている動的特徴 [2] 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，静的な特徴ベクトル $\boldsymbol{c}_t$ (例えばメルケプストラム) と，動的な特徴ベクトル $\Delta\boldsymbol{c}_t$ (例えばデルタメルケプストラム) および $\Delta^2\boldsymbol{c}_t$ (例えばデルタデルタメルケプストラム) 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．但し，$\Delta\boldsymbol{c}_t$ および $\Delta^2\boldsymbol{c}_t$ の値は，静的特徴ベクトル $\boldsymbol{c}_t$ から

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w_1(\tau)\boldsymbol{c}_{t+\tau} \quad (16)$$

$$\Delta^2\boldsymbol{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w_2(\tau)\Delta\boldsymbol{c}_{t+\tau} \quad (17)$$

により計算されるものとする．ここで，$w_1(\tau)$，$w_2(\tau)$ は動的特徴量を計算するための重み係数である．

式 (16), (17) の条件は，行列形式により，

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C} \quad (18)$$

と線形変換の形で書くことができる．ただし，

$$\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T]^\top \quad (19)$$

とする．$\boldsymbol{c}_t$ が M 次元とすれば，$\boldsymbol{C}, \boldsymbol{O}$ は，それぞれ，TM 次元，$3TM$ 次元である．$\boldsymbol{W}$ は，$3TM \times TM$ の行列であり，1 部の要素に係数 $1, w_1(\tau), w_2(\tau)$ をもち，他の多くの要素は 0 となる．式 (18) の条件の下で，$P(\boldsymbol{O}|\boldsymbol{Q},\lambda)$ を最大にする $\boldsymbol{C}$ は，

$$\frac{\partial \log P(\boldsymbol{W}\boldsymbol{C}|\boldsymbol{Q},\lambda)}{\partial \boldsymbol{C}} = \boldsymbol{0}, \quad (20)$$

とおくことによって得られる線形方程式

$$\boldsymbol{W}^\top\boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{C} = \boldsymbol{W}^\top\boldsymbol{U}^{-1}\boldsymbol{M}^\top. \quad (21)$$

により定められる．$\boldsymbol{W}'\boldsymbol{U}^{-1}\boldsymbol{W}$ が $TM \times TM$ の行列であることから，式 (21) を解くためには $O(T^3M^3)$ の演算量を必要とするが[4]，$\boldsymbol{W}'\boldsymbol{U}^{-1}\boldsymbol{W}$ の特別な性質を利用すれば，コレスキー分解あるいは QR 分解を用いて $O(TM^3L^2)$ の演算量で解くことができる[5]．ただし，$L = \max\{L_-^{(1)}, L_+^{(1)}, L_-^{(2)}, L_+^{(2)}\}$ とする．式 (21) は，[36], [37] のアルゴリズムによって解くこともでき，それによれば，時間方向に再帰的な形で計算を行うことができる [38]. なお，式 (7) および式 (8) を解くアルゴリズムも提案されており，それらは [39] にまとめられている．

生成された出力ベクトル（ここで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 4 に示す．実験条件については [36] を参照されたい．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．それに対して，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わかる．

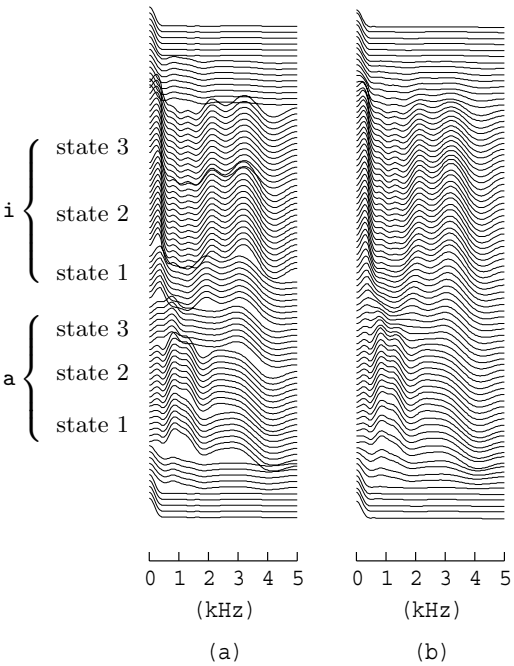

図 4　音素モデル sil, a, i, sil を接続することによりつくられた HMM からのパラメータ生成例．(a) 動的特徴なし，(b) 動的特徴あり．

[4] $\boldsymbol{U}_{q,i}$ が対角行列のときには，$O(T^3M)$ となる．

[5] $\boldsymbol{U}_{q,i}$ 対角行列のときには，$O(TML^2)$ となる．更に，$L_-^{(1)} = -1$, $L_+^{(1)} = 0$, $w^{(2)}(i) \equiv 0$ のときには，$O(TM)$ となる [35].

4.2 尤度最大化基準による状態継続長の決定

図 1 のような構造の HMM を考える．時刻 $t = 1 \sim T$ の間に状態 $i = 1 \sim K$ を通過するとし，状態 i が d_i 回継続する確率を $p_i(d_i)$ とすれば，式 (10) 中の $P(\boldsymbol{Q}|\lambda, T)$ は，

$$\log P(\boldsymbol{Q}|\lambda, T) = \sum_{i=1}^{K} \log p_i(d_i) \tag{22}$$

と書くことができる．ただし，$\sum_{i=1}^{K} d_i = T$ である．図 1 の構造をもつ HMM の場合，式 (2) より，$p_i(d)$ は，

$$p_i(d) = \frac{1 - a_{ii}}{a_{ii}} a_{ii}^{d} \tag{23}$$

と指数分布でモデル化されることになるが，これは継続長を適切に制御するためには単純過ぎるモデルである．そこで，$p_i(d)$ をガウス分布でモデル化することにする．このとき，式 (10) の $\boldsymbol{Q}_{\max}$ を与える $\{d_i\}_{i=1}^{K}$ は

$$d_i = m_i + \rho \cdot \sigma_i^2 \tag{24}$$

$$\rho = \left(T - \sum_{k=1}^{K} m_i \right) \Big/ \sum_{k=1}^{K} \sigma_i^2 \tag{25}$$

と簡単に定められる [40]．但し，m_i と σ_i^2 は，それぞれ，状態 i に関するガウス分布の平均と分散である．T と ρ は式 (25) で関係づけられているため，ρ を与えることにより，T を定めることができる．式 (25) よりわかるとおり，発話速度は，ρ の値が小さいほど速く，大きいほど遅くなる．平均的な発話速度で音声を合成したい場合には，$\rho = 0$ とすればよい．

4.3 ピッチパタンのモデル化

音声のピッチパタンは，有声区間では 1 次元の連続値，無声区間では無声であること表す離散シンボルとして観測されるため，通常の音声認識などで用いられる離散分布 HMM や連続分布 HMM の枠組みを直接適用することはできない．これまでにも，ピッチパタンを HMM，あるいは統計モデルによりモデル化しようとする試みは行なわれているが，無声区間の処理に関して，何らかの便宜的な仮定や手法が用いられていた．一方，可変次元の多空間上における確率分布に基づいた HMM(MSD-HMM)[41] は，離散分布 HMM，混合連続分布 HMM を特別な場合として含むものであり，更に，離散シンボルと連続値が時間的に混在した観測系列をモデル化することができるため，これにより，無声区間を含んだピッチパタンを確率理論的な整合性をもって HMM によりモデル化することが可能となる [42].

4.4 単位接続型音声合成方式との関係

4.1–**4.3** で述べた HMM からの音声合成手法により，スペクトル，ピッチ，継続長を同時に HMM の枠組でモデル化する音声合成システムを構築することができる [27]. 音素 HMM は，表 1 に示すコンテキストを考慮したものであり，その出力ベクトルは，スペクトルとピッチを連結し，更にそれらの動的特徴を連結したもので，出力ベクトルのスペクトルに関する部分とピッチに関する部分に対し，別々に，図 3 と同様のクラスタリングを適用している．継続長分布に関しても同様である．従来，音声合成において，韻律情報を統計モデルで制御する場合には，数量化 I 類 [43]，回帰木 [44]，それらを拡張した CTR [45] などが用いられるが，その意味では，本方式の韻律制御は回帰木を用いるものに分類される．

HMM からの音声合成手法は，コンテキストクラスタリングに基づいているという点で方法 (ii)，特に，HMM を用いてクラスタリングを行っているという点から，文献 [9], [11] との関連が深い．主な相違点のひとつは，各クラスタにある複数の音声単位から一つを選ぶのではなく，これらから計算された統計量によりそのクラスタが表現され，合成時にはこの統計量から静的・動的特徴ベクトルに関する尤度最大化基準より音声パラメータが生成される点である．方法 (ii) の音声単位選択法においては，クラスタ中心に近い音声単位を優先して選択することができるが，これは，HMM からの音声合成法において，静的特徴量に関する尤度を考慮していることに対応している．また，音声単位選択法において，接続コストを考慮することは，HMM からの音声合成法が，動的特徴量に関する尤度を考慮していることと対応している．このような議論から，HMM からの音声合成手法と，方法 (ii) の単位接続型の音声合成法は，一方は各クラスタの統計量から，他方はクラスタ内のマルチテンプレートから音声単位が生成される点が異なっているものの，いずれも類似した原理によって音声を生成していることが理解される [46].

5. むすび

最近のコーパスベースの音声合成における HMM の役割について概説することを目的とし，HMM について簡単にまとめた後，音声合成における HMM の利用法について述べた．また，ピッチ，継続長，スペクトルを同時に HMM でモデル化し，HMM 自身から音声を合成する手法についても，他の手法と関連付けながらまとめた．

自動学習に基づいた音声合成システムの構築に関する研究は，今後，益々盛んになっていくものと思われる．特に，共通に利用可能な音声合成のための音声データベースの整備が進むことにより，音声認識同様，競争的な環境が生まれ，研究が飛躍的に推進することが期待される．

謝辞 ご討論頂く東京工業大学大学院総合理工学研究科小林隆夫教授，益子貴史助手に感謝致します．実験に

ご協力頂く東京工業大学博士課程田村正統氏，名古屋工業大学博士課程吉村貴克氏に感謝します．

文　献

[1] S. Schwartz, Y-L. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, “Context-dependent modeling for acoustic-phonetic of continuous speech,” Proc. ICASSP, pp.1205–1208, 1985.

[2] S. Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. Acoust., Speech, Signal Processing, vol.34, pp.52–59, 1986.

[3] B.-H. Juang, “Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains,” AT&T Technical Journal, vol.64, no.6, pp.1235–1249, 1985.

[4] J. J. Odell, “The use of context in large vocabulary speech recognition,” PhD dissertation, Cambridge University, 1995.

[5] 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. Acoust., Speech, Signal Processing, vol.39, no.4, pp.806-814, Apr. 1992.

[6] M. J .F. Gales, and P. C. Woodland, “Mean and variance adaptation within the MLLR framework,” Computer Speech and Language, vol.10, No.4, pp.249–264, Apr. 1996.

[7] A. Ljolje, J. Hirschberg and J. P. H. van Santen, “Automatic speech segmentation for concatenative inventory selection,” in Progress in Speech Synthesis, ed. J. P. H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, Springer-Verlag, New York, 1997.

[8] R. E. Donovan and P. C. Woodland, “Automatic speech synthesiser parameter estimation using HMMs,” Proc. ICASSP, pp.640–643, 1995.

[9] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith and M. Plumpe, “Recent improvements on Microsoft’s trainable text-to-speech system -Whistler,” Proc. ICASSP, 1997.

[10] H. Hon, A. Acero, X. Huang, J. Liu and M. Plumpe, “Automatic generation of synthesis units for trainable text-to-speech synthesis,” Proc. ICASSP, 1998.

[11] R. E. Donovan and E. M. Eide, “The IBM Trainable Speech Synthesis System,” Proc. ICSLP, vol.5, pp.1703–1706, 1998.

[12] A. Falaschi, M. Giustiniani and M. Verola, “A hidden Markov model approach to speech synthesis,” Proc. EUROSPEECH, pp.187–190, 1989.

[13] M. Giustiniani and P. Pierucci, “Phonetic ergodic HMM for speech synthesis,” Proc. EUROSPEECH, pp.349–352, 1991.

[14] 益子貴史, 徳田恵一, 小林隆夫, 今井 聖, “動的特徴を用いた HMM に基づく音声合成,” 信学論 (D), vol.J79-D-II, no.12, pp.2184–2190, Dec. 1996.

[15] E. Moulines, and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” Speech Communication, no.9, pp.453–467, 1985.

[16] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE trans. Acoust., Speech, Signal Processing, vol.ASSP-33, pp.357–366, Aug. 1986.

[17] 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.

[18] 今井 聖, 音声認識, 共立出版, 1995.

[19] L. Rabinar and B.-J. Juang (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.

[20] 北 研二, 中村 哲, 永田昌明, 音声言語処理, 森北出版, 1996.

[21] 鹿野清宏, 中村 哲, 伊勢史郎, 音声・音情報のディジタル信号処理, 昭晃堂, 1997.

[22] 古井貞熙, 音声情報処理, 森北出版, 1998.

[23] X. D. Huang, Y. Ariki and M. A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, 1990.

[24] http://htk.eng.cam.ac.uk/.

[25] Y. Sagisaka, N. Kaiki, N. Iwahashi and K. Mimura, “ATR ν-talk speech synthesis system,” Proc. ICSLP, pp.483–486, 1992.

[26] B. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal, “The AT&T Next-Gen TTS system,” Proc. Joint ASA, EAA and DAEA Meeting, pp.15–19, Mar. 1999.

[27] 吉村貴克, 徳田恵一, 益子貴史, 小林隆夫, 北村正, “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,” 信学論 (D-II), vol.J83-D-II, no.11, Nov. 2000.

[28] A. W. Black and N. Campbell, “Optimising selection of units from speech databases for concatenative synthesis,” Proc. EUROSPEECH, pp.581-584, Sep 1995.

[29] 岩橋直人, 海木延佳, 匂坂芳典, “音響的距離尺度に基づく複合音声単位選択,” 信学技報, SP91-5, 1991.

[30] A. G. Hauptmann, “SPEAKEZ: A first experiment in concatenation synthesis from a large corpus,” Proc. EUROSPEECH, pp.1701–1704, 1993.

[31] A. W. Black P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” Proc. EUROSPEECH, pp.601–604, Sep 1997.

[32] M. W. Macon, A. E. Cronk and J. Wouters, ”Generalization and discrimination in tree-structured unit selection,” Proc. ESCA/COCOSDA Workshop on Speech Synthesis, Nov. 1998.

[33] 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.

[34] 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.

[35] A. Acero, “Formant analysis and synthesis using hidden Markov models,” Proc. EUROSPEECH, Budapest, Hungary, pp.1047–1050, 1999.

[36] K. Tokuda, T. Kobayashi and S. Imai, “Speech parameter generation from HMM using dynamic features”, Proc. ICASSP-95, pp.660–663, 1995.

[37] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and Satoshi Imai : “An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features”, Proc. EUROSPEECH-95, pp.757–760, 1995.

[38] K. Koishida, K. Tokuda, T. Masuko and T. Kobayashi, “Vector quantization of speech spectral parameters using statistics of dynamic features”, Proc. ICSP, vol.1, pp.247–252, Aug. 1997.

[39] K. Tokuda, Takayoshi Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. ICASSP, Turkey, June 2000.

[40] 徳田恵一, 益子貴史, 小林隆夫, 今井 聖, “動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム,” 日本音響学会誌, vol.53, no.3, pp.192–200, Mar. 1997.

[41] 宮崎 昇, 徳田恵一, 益子貴史, 小林隆夫, “多空間上の確率分布に基づいた HMM とピッチパタンモデリングへの応用,” 信学技報, SP98-11, pp.19–26, Apr. 1998.

[42] 宮崎昇, 徳田恵一, 益子貴史, 小林隆夫, “多空間上の確率分布を用いた HMM とピッチパタン生成の検討,” 信学技報, SP98-12, pp.27–34, Apr. 1998.

[43] 阿部匡伸, 佐藤大和, “区分化音節モデルに基づく基本周波数の 2 階層構造,” 日本音響学会誌, vol.49, no.10, pp.682–690, Oct. 1993.

[44] M. Riley, “Tree-Based Modelling of Segmental Duration,” Talking Machines: Theories, Models, and Designs, Elsevier Science Publishers, pp.265–273, 1992.

[45] N. Iwahashi and Y. Sagisaka, “Statistical Modelling of Speech Segment Duration by Constrained Tree Regression,” IEICE trans, vol.E83-D, no.7, pp.1550–1559, July 2000.

[46] 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.