隠れマルコフモデルの音声合成への応用

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

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，data-driven, corpus-based, speaker-driven あるいは trainable などと形容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ている．このようなシステムを構築する際に，隠れマルコフモデル (hidden Markov model: HMM) を用いるものが多くなっているが，本文では，これらの中でも特に HMM 自身から音声を合成しようとする方式について，他の手法とも関連づけながらまとめることを目的とする．HMM に基づいた音声合成において用いられる，動的特徴をもった HMM からの音声パラメータ生成法と，メルケプストラム分析合成法について述べた上で，合成システムの構成と，韻律情報（ピッチ，継続長）をスペクトル情報と同時に HMM の枠組でモデル化するアプローチについて述べる．また，HMM からの音声合成手法の応用例を紹介する．

音声合成，隠れマルコフモデル，動的特徴，パラメータ生成

SPEECH SYNTEHSIS BASED ON HIDDEN MARKOV MODELS

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 data-driven, corpus-based, speaker-driven, or trainable approach, by applying statistical learning algorithms. This paper describes one of these approaches: HMM-based speech synthesis in which synthetic speech is generated from HMMs themselves. Algorithms for speech parameter generation from HMMs, and a mel-cepstrum based vocoding technique are reviewed, and an approach to simultaneous modeling of spectrum, pitch and state duration is also described. The relation between the HMM-based approach and other concatenative speech synthesis approaches is also discussed.

speech synthesis, hidden Markov model, dynamic feature, parameter generation

1. まえがき

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，data-driven, corpus-based, speaker-driven あるいは trainable などと形容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ている．これらの方式は，従来の規則に基づいた (rule-based) 合成方式と異なり，大量のデータを用いた自動学習や音声素片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，何らかの形で隠れマルコフモデル (hidden Markov model: HMM) を利用することが多くなっている．HMM は，音声認識の分野において，音声スペクトル系列の統計的モデル化手法として，広く成功を収めている．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が，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．

音声合成における HMM の利用は，(1) 音声データベースの transcription やセグメンテーションに用いるもの (例えば [7])，から，(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，音声素片の inventory を選ぶもの (例えば [8], [9])，(3) HMM 自身から音声を合成しようとするもの (例えば [10]–[12])，など，その形態と HMM への依存の度合は様々である．

(1), (2) はいずれも，音声素片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 [13] 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が，このことは，同時にデータベースに存在しない音は出力できないという限界にも継っている．テキスト音声合成がヒューマンインタフェースのひとつとして広く普及するためには，合成音声の高品質化のみならず，多様な話者性あるいは発話スタイルをもった音声を自在に合成できることが必須と思われるが，どんな大量の音声データを用いたとしても，すべての音声現象を網羅することはできないため，いずれ音声素片を適切に変形する何らかのメカニズムが必要になると予想される．一方，(3) 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をもっている．例えば，音声認識の分野では，近年 HMM の枠組の中で話者適応の問題を取り扱う手法が数多く提案されていることから，音声認識における話者適応と同様の手法を用いることにより，様々な話者の声質を模倣したり，更には，異なる「話者」を異なる「発話スタイル」に対応づけることにより，多様な音声の合成が容易になることも期待される．

このような背景から，本文では，特に上記 (3) の HMM 自身から音声を合成しようとする方式について，他の手法とも関連づけながら述べることを目的とする．以下，**2.** において，HMM に基づいた音声合成において用いられる，動的特徴をもった HMM からの音声パラメータ生成法についてまとめ，**3.**で，メルケプストラム分析合成法について述べた上で，**4.** において，ピッチ情報をスペクトル情報と同時に HMM の枠組でモデル化するアプローチについて述べる．また，**5.**では，合成システムの構成と，他の多くの音声合成手法においても用いられており，本手法でも本質的な役割を果たしているコンテキストクラスタリングに関連した考察を加える．**6.**で，HMM からの音声合成手法の応用例を紹介し，**7.**で結論を述べる．

2. 尤度最大化基準による HMM からの音声パラメータ生成

HMM から音声パラメータを生成することにより，HMM を音声合成や音声スペクトルパラメータの量子化に用いる手法として [14], [10], [11], [15], [16] などをあげることができる．これらのうち，文献 [14] は，乱数発生器を利用する手法 (2 次元単一分布 HMM の各分布の平均と分散に従うガウス雑音系列を状態継続長分だけ発生させ、この系列を、デルタパラメータの標本平均が正か負かによって、昇順あるいは降順に並べる) であるが，それ以外は，与えられた状態系列に対して，出力確率が最大となる音声パラメータ系列を生成すること，あるいはそれと同等なことを基本としている．このとき，音声パラメータは，ひとつの状態が継続している間は一定の値をとり，状態の遷移のたびに不連続に変化してしまう [11] ため，いくつかの対策がとられている．文献 [15] では，離散 HMM を用いて，与えられた状態系列から出力シンボル系列 (ここでは音声パラメータ系列) を生成することを考えているが，出力シンボル i に続き，j が出現する確率を，板倉-斉藤距離により近似し，モデルに追加することにより，出力シンボル系列が不連続に変化することを抑えている．文献 [10] では，自己回帰 HMM を用いており，モデルパラメータ (この場合は音声波形の自己相関) をヒューリスティックな状態滞在確率で重み付け和することにより，出力ベクトル系列を平滑化している．文献 [16] では，2 次元単一分布 HMM を用いており，平均ベクトルを，隣接する状態継続長区間の中心間で線形に補間している．

しかし，上記いずれも，補間や平滑化に関して便宜的な仮定や処理を導入したものとなっている．本節では，動的特徴 [2] をパラメータとして含む連続分布 HMM から尤度最大の意味で最適な音声パラメータ系列を生成する手法 [17] について述べ，動的特徴が重要な役割を果たしていることを示す．

2.1 動的特徴の導入

連続出力分布型 HMM λ が与えられたとき，λ から長さ T の出力ベクトル系列 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ を生成することを考える．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

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau)\boldsymbol{c}_{t+\tau} \tag{1}$$

$$\Delta^2\boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau)\Delta\boldsymbol{c}_{t+\tau} \tag{2}$$

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．

このとき，ある与えられた状態系列 $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$ に沿って，パラメータ系列 $\boldsymbol{O}$ が λ から観測される確率（$\boldsymbol{Q}$, λ の $\boldsymbol{O}$ に対する尤度）$P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda, T)$ を最大にする音声パラメータベクトル系列 $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T\}$ およびそのときの状態系列 $\boldsymbol{Q}$ を求めることを考える [17]．ここで，λ の各状態が，単一ガウス分布をもつとすると，$P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda, T)$ の対数は，

$$\begin{aligned}\log P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda) &= -\frac{1}{2}(\boldsymbol{O}-\boldsymbol{M})'\boldsymbol{U}^{-1}(\boldsymbol{O}-\boldsymbol{M}) - \frac{1}{2}\log|\boldsymbol{U}| \\ &\quad + \text{Const}\end{aligned} \tag{3}$$

と書くことができる．ここで，

$$\boldsymbol{M} = [\boldsymbol{\mu}'_{q_1}, \boldsymbol{\mu}'_{q_2}, ..., \boldsymbol{\mu}'_{q_T}]' \tag{4}$$

$$\boldsymbol{U} = \text{diag}[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, ..., \boldsymbol{U}_{q_T}] \tag{5}$$

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 (1)，(2) 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．一方，式 (1)，(2) の制約下では，$\log P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)$ を最大にする $\boldsymbol{C}$ は線形方程式

$$\frac{\partial \log P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)}{\partial \boldsymbol{C}} = \boldsymbol{0} \tag{6}$$

によって定められる．この方程式は，文献 [17] の高速アルゴリズムにより容易に解くことができ，特に時間方向に再帰的な形式のアルゴリズム [18] として記述することができる．得られる出力ベクトル系列は，静的および動的特徴ベクトルの平均ベクトルだけでなく，それらの共分散行列によって定められることになる．

次に，状態系列 $\boldsymbol{Q}$ が未知の場合，つまり $P(\boldsymbol{O}\,|\,\lambda, T) = \sum_{\boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}\,|\,\lambda, T)$ を式 (1)，(2) の制約下で $\boldsymbol{O}$ に関して最大化することを考えるが，ここでは，これを $\max_{\boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q}\,|\,\lambda, T)$ と近似し，最大化することにする．この問題は，可能な全ての $\boldsymbol{Q}$ について，$P(\boldsymbol{O}, \boldsymbol{Q}\,|\,\lambda, T)$ を最大化する $\boldsymbol{C}$ を求めれば解くことができる．一般には可能な $\boldsymbol{Q}$ の組合せは非常に多く，全ての $\boldsymbol{Q}$ について $P(\boldsymbol{O}, \boldsymbol{Q}\,|\,\lambda, T)$ を求めることは現実的ではないが，異なる状態系列に対する解を再帰的に計算することのできるアルゴリズムにより，準最適なパラメータ系列を効率的に求めることができる [17]．

このように，状態系列 $\boldsymbol{Q}$ を未知とする場合に，適切な状態系列を得るためには，スキップのない left-to-right モデルを考え，状態継続長分布を与える必要がある．状態継続長に関する確率は，$P(\boldsymbol{O}, \boldsymbol{Q}\,|\,\lambda, T) = P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda, T) \cdot P(\boldsymbol{Q}\,|\,\lambda, T)$ と書いたときの $P(\boldsymbol{Q}\,|\,\lambda, T)$ であり，

$$\log P(\boldsymbol{Q}\,|\,\lambda, T) = \sum_{k=1}^{K} \log p_{q_k}(d_{q_k}) \tag{7}$$

で与えられることになる．但し，T の間に通過する状態の数を K，状態 q_k が d_{q_k} 回継続する確率を $p_{q_k}(d_{q_k})$ とする．更に，$P(\boldsymbol{Q}\,|\,\lambda, T)$ のみを最大化するように $\boldsymbol{Q}$ を定めることにすると，$P(\boldsymbol{O}, \boldsymbol{Q}\,|\,\lambda, T) = P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda, T) \cdot P(\boldsymbol{Q}\,|\,\lambda, T)$ の $\boldsymbol{C}$ に関する最大化は，状態系列が既知の場合と同じ問題となり，式 (6) を一度解くだけで，解を得ることができる．状態出力分布が混合ガウス分布の場合には，各混合要素を単一ガウス分布をもった状態と考えて展開し，混合要素のみが未知であるとして状態系列の探索を行えばよい．

式 (1)，(2) の制約がない場合に，$P(\boldsymbol{O}\,|\,\lambda, T)$ を局所最大にする $\boldsymbol{O}$ は，EM アルゴリズムにより求めることができる．更に，式 (1)，(2) の制約下で，$P(\boldsymbol{O}\,|\,\lambda, T)$ を $\boldsymbol{C}$ に関して局所最大化するアルゴリズムを導出することも可能である [19]．この場合，状態出力分布は，混合分布であってもよいし，状態系列だけが既知で混合要素が未知であってもよい．また，状態系列が既知で混合要素が未知の場合に，勾配法により $P(\boldsymbol{O}\,|\,\lambda, T)$ を局所最大化するアルゴリズムも提案されている [20]．

2.2 尤度最大化基準による状態継続長の決定

先で述べたように，$P(\boldsymbol{Q}\,|\,\lambda, T)$ を最大化するように，状態継続長を定めることを考える．$p_{q_k}(d_{q_k})$ が離散的な分布のときには，Viterbi アルゴリズムによって，このよう

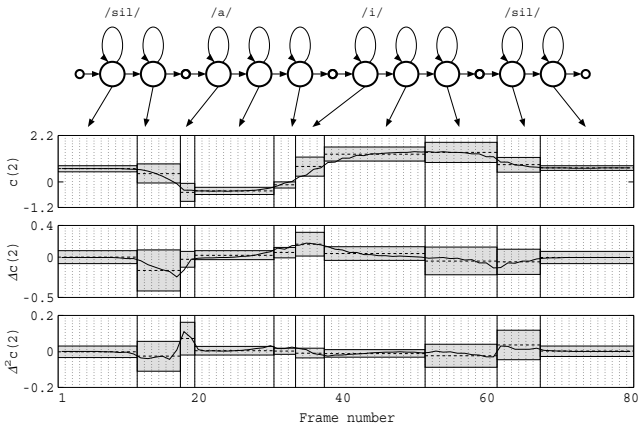

図1　音声パラメータの生成例

な$\boldsymbol{Q}$を定めることができる．また，$p_{q_k}(d_{q_k})$がガウス分布でモデル化されているときには，$\boldsymbol{Q}$を与える$\{d_{q_k}\}_{k=1}^{K}$は

$$d_{q_k} = m_{q_k} + \rho \cdot \sigma_{q_k}^2 \tag{8}$$

$$\rho = \left(T - \sum_{k=1}^{K} m_{q_k}\right) \Big/ \sum_{k=1}^{K} \sigma_{q_k}^2 \tag{9}$$

と簡単に定められる[17]．但し，m_{q_k}と$\sigma_{q_k}^2$は，それぞれ，状態q_kに関するガウス分布の平均と分散である．Tとρは式(9)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(9)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．

なお，状態継続長分布をHMMに含めて学習を行った場合[21]，計算時間がかなり大きくなるが，状態継続長分布を含まないHMMの連結学習の最後の繰り返しで得られる状態滞在確率を用いて状態継続長分布を計算することにより[22]，計算時間を短縮することも可能である．

2.3　音声パラメータの生成例

図1に音韻バランス文(503文)を用いてトレーニングした混合数1の音素モデル`sil`, `a`, `i`, `sil`を結合することによって得たHMMから生成された音声パラメータ系列（メルケプストラムの第2次係数$c_t(2)$のみ）を示す[17]．合わせて，$\Delta c_t(2)$, $\Delta^2 c_t(2)$についても示す．但し，パラメータ生成の際には$P(\boldsymbol{Q} \mid \lambda, T)$だけで状態系列$\boldsymbol{Q}$を定め，状態系列の探索は行なっていない．図中の破線は各状態の平均を，網かけされた部分が標準偏差（対角分散行列を用いており，その分散の平方根）を表している．

式(1)，(2)の制約を用いない場合には，$P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda, T)$を最大にするパラメータ系列は平均ベクトルの系列（図中の破線）となる．それに対して，式(1)，(2)の制約を課した場合には，$\Delta c_t(2)$, $\Delta^2 c_t(2)$は，それぞれ，$c_t(2)$, $\Delta c_t(2)$の軌跡の傾きを表すものとなり，自由度は$c_t(2)$にしかなくなる．しかし，最大化される尤度は，$c_t(2)$, $\Delta c_t(2)$, $\Delta^2 c_t(2)$すべてに関するものであり，これら3種の尤度を妥協させる形で$c_t(2)$の軌跡が定まっている様子が図よりわかる．例えば，それぞれの音素モデルの始めと終りの状態では，動的および静的特徴の分散が比較的大きいため，パラメータ系列はひとつの状態が継続する間にも適切な軌跡を描いて変化している．また，それぞれの音素モデルの中心の状態では，動的および静的特徴の分散は小さく，動的特徴の平均ベクトルはほとんど0であるため，生成されたパラメータ系列は静的特徴の平均ベクトルに近い一定値をとる傾向となっている．以上の例により，HMMから音声パラメータを生成する際に，動的特徴が重要な働きをすることが示されている．

3.　メルケプストラム分析合成系

HMMの音声合成や音声符号化への応用においてはlsp係数や自己回帰係数などを音声パラメータとするものが多いが，音声認識の分野では，MFCC [23], LPC-derived mel-cepstrum (LPC-MCEP) [24], PLP-derived cepstrum (PLP-CEP) [25]など、いずれも非直線周波数軸上で定義されたケプストラムに対応するパラメータが有効であることが知られている．従って，音声合成のためのHMMにおいても，このような音声パラメータを用いることにより，精度良く音声スペクトル系列をモデル化できることが期待される．しかし，MFCC, PLP-CEP, LPC-MCEPなどでは，分析法と整合性をもった音声合成法が知られていないことが問題となる．それに対して，文献[26]のメルケプストラム分析法によれば，MLSAフィルタ[27], [28]を用いることにより，得られたメルケプストラム係数から，直接音声を合成することができる[1]．合成音声は，いわゆるvocoded speechとなるが，近年，励振源の改良などにより，高い音声品質をもつボコーダが数多く提案されていることから[29]–[31]，これらの励振源生成手法を導入することにより，この問題は克服できるものと期待される．また，次節で述べるMSD-HMM[32]によれば，離散値，連続値が混在する音源パラメータの系列をHMMの枠組の中でモデル化することも可能である．

4.　ピッチパタンのモデル化

音声の話者性，発話スタイル，感情表現などは，韻律，つまり，声の高さ(ピッチ)，大きさ(パワー)，音素継続長，ポーズなどに大きく影響される．しかし，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

[1] メルケプストラム分析合成のソースコードは，学術研究用として`http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/`にて公開されている．

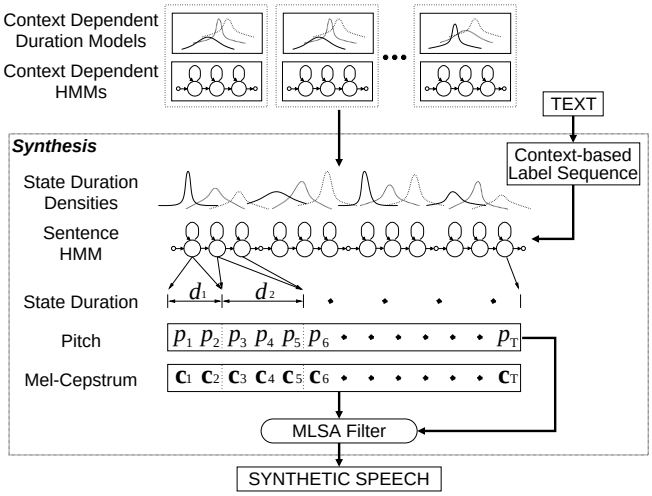

図2　HMM に基づいた音声合成システムの構成

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，そこでは，(1) 無声区間のピッチとして分散の大きな乱数を与える [33]，(2) 無声区間のピッチの値を 0 として，混合分布によりモデル化する [34]，(3) 無声区間のピッチの値は存在するが観測できなかったとし (latent variable と考える)，EM アルゴリズムを適用する [35]，などの便宜的な方法が用いられている．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)[32] 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直接モデル化することが可能となる [36].

5.　HMM に基づいた音声合成システム

5.1　音声合成システムの構成

前節までで述べた要素技術を統合することにより構築される合成システム [37] を図 2に示す.

スペクトル，ピッチモデルは連結学習により学習される．連結学習はラベル境界の情報を必要としないため，システムの自動学習に適している．但し，スペクトルモデル，ピッチモデルを別々に連結学習した場合，両者のモデル間で境界のずれが生じるため，スペクトル，ピッチ，およびそれぞれの動的特徴量を結合したものを特徴ベクトルとして HMM を学習する．

5.2　コンテキストクラスタリング

スペクトル，ピッチ，継続長に影響を与えると考えられるコンテキスト（変動要因）には，アクセント型，品詞，当該・先行・後続音素などがあるが，これらをすべて考慮したコンテキスト依存モデルを構築する．但し，考慮するコンテキストの種類の増加とともにその組合せは指数的に増大するため，モデル当たりの学習データが著しく減少し，モデルのパラメータの推定精度が低下する．また，可能な全てのコンテキストの組合せを網羅する学習データを用意することは実際上不可能であり，学習できなかったコンテキストの組合せが，音声合成時に要求されることになる．この問題を解決するために，決定木を用いたコンテキストクラスタリング [4] を適用する．それぞれのコンテキストは，スペクトル，ピッチ，状態継続長に対して異なる影響を与えると考えられるため，スペクトル，ピッチ，状態継続長，それぞれ別々のクラスタリングを行う．文献 [4] のコンテキストクラスタリング手法は，文献 [36] において MSD-HMM に対し拡張されているため，ピッチのコンテキストクラスタリングもスペクトルと同様に行うことができる．

なお，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 [38]，回帰木 [39]，それらを拡張した MSR[40] などが用いられるが，本方式は，その意味では，回帰木を用いるものに分類される．

5.3　音声素片選択方式との関係

音声素片選択型の音声合成方式では，(a) コンテキストの適合性（target cost, context matching score などと呼ばれる），(b) 素片の接続コスト (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素片を選び，接続することが目標とされる ([41], [42] 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素片を選択するもの ([42], [43] 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素片から選ぶもの ([8], [9], [44] など)，に大別される．方法 (i) の利点は，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，素片の選択を行うことができる点にあるが，コンテキスト間距離の設定がヒューリスティックで難しい，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，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，各クラスタに適切な数の音声素片 (instance) をおくことにより合成時に必要なデータサイズを制御しやすい，などの利点がある．

前節までで述べた HMM からの音声合成手法は，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で，文献 [8], [9] 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素片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・動的特徴量の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．

ところで，音声素片選択法における，基準 (b) のための音声パラメータの不連続の度合のはかり方として，接

続部分の引き続くフレームの音声パラメータの差分を考慮しただけでは，元の音声データ自体が接続部分に対応する位置で不連続な変化をしている可能性があり，必ずしも適切ではない．このため，元の音声データベース中で隣に位置する音声素片の音声パラメータと，合成時に隣に位置する音声素片の音声パラメータとの差分により，これを定義している [41], [42], [9]．このような接続歪みの評価は，データベース中のデルタパラメータと合成時のデルタパラメータとの差ををを評価していることと等価であり，これは，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．また，このことから，接続部分での音声パラメータの平滑化を行なう場合には，ただ不連続をなくせばよいというものではなく，元の音声データが滑らかに変化していたか，不連続な変化をしていたかによって，異なる平滑化を行う必要があることがわかる．文献 [45] では，音声素片の統計量に基づいてこのような平滑化を行っているが，数学的には **2.**で述べたパラメータ生成法と等価なものとなっている．

更に，方法 (ii) の音声素片選択法においては，クラスタ中心に近い音声素片を優先して選択するが，これは，HMMからの音声合成法において，静的特徴量の尤度を考慮していることに対応していると考えられる．

以上の議論から，接続コストを考慮した音声素片の選択，および接続部分の平滑化に関して動的特徴が重要な役割を果たしていること，HMMからの音声合成手法と方法 (ii) の素片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素片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．

6. 応用

本節では，いくつかの応用について簡単に述べる．

6.1 話者補間に基づく音声合成

各話者の音声は，HMMによってモデル化されるため，統計的，情報理論的な尺度に基づいて，複数の話者HMMを補間することにより，多様な声質を実現することができる [46]．ある男性話者によって学習したHMMと，別の女性話者によって学習したHMMを，任意の比率で補間することにより，2人の話者を声質を任意の比率で補間した音声を合成できることが確かめられている．話者を発話スタイルや感情表現に置き換え，代表的ないくつかの発話スタイル/感情表現を補間することにより，任意の発話スタイル/感情表現の音声を合成できる可能性がある．

6.2 話者適応に基づく声質変換

HMMに基づく音声認識の分野では，学習データに含まれない未知話者に対する認識精度を向上させるため，話者適応の技術が種々検討されている．これは，標準的なHMMを用意しておき，未知話者の入力音声が得られた段階で，HMMのパラメータをその話者に適合するように修正するものである．話者適応の手法は，合成音声の声質変換にも応用が可能であり，コードブックマッピングに基づくスペクトル写像に応用した例などが報告されている [47]．前節で述べた音声合成システムでは，合成単位として音素HMMを利用していることから，このような話者適応による音素HMMのパラメータ変換がより適切に機能すると考えられ，多様な声質での音声合成が可能になると期待される．話者適応手法の一例としてMAP/VFS [48], MLLR [6] を用いた場合について，数文章の適応データで目標話者に近い声質をもった合成音声が得られることが確かめられている [49], [50].

6.3 極低ビットレート音声符号化

100ないし数百bit/s程度のビットレートで音声を符号化するためには，音素ボコーダ，あるいはセグメントボコーダを用いるのが最も一般的な方法である [51]–[55]．これらの符号化法では，音声を音素単位，または音響的なセグメント単位などの音声単位に分割し，得られた音声単位のインデックス系列と継続長を復号化器側に伝送する．復号化器側では，伝送されたインデックスと継続長に従い，音声単位を連結することにより音声を合成する．音素ボコーダ，セグメントボコーダにおける符号化器は一種の音声認識器とみなすことができるため，近年のHMMに基づいた音声認識器の性能向上を考慮すると，HMMに基づいて音素ボコーダ，あるいはセグメントボコーダを構築することは一つの有効な方法と考えられる．文献 [56] では，HMMに基づいた音声合成方式を利用することにより，符号化器・復号化器を通してHMMに基づいた音素ボコーダを構築し，その性能を評価しており，ピッチ情報を除いて 146bit/s （26%の無音区間を含む）で，同じくピッチ情報を除いて400bit/s（8 bit/frame×50 frame/s）のベクトル量子化に基づくボコーダと同等の性能を得ている．また，68bit/sの場合にも符号化音声の了解性を保つことができている．

6.4 話者照合システムの安全性の検証

話者照合は入力音声と申告話者の声が同一の話者によるものかを判定する技術であり，音声による本人確認手段として，今後利用の拡大が期待されている．従来，発声すべきテキストがその都度システムから指定されるテキスト指定型話者照合システム [57] を用いれば，テープレコーダ等による録音音声を用いた詐称を防ぐことができるとされてきた．しかし，テキスト音声合成方式の進歩により，合成音声による詐称に関する検討も必要になってきたと考えられる．HMMに基づいた音声合成方式を用いた実験では，1文章の適応データを用いて不特定話

者モデルを適応した場合にも，十分詐称が可能であることが示されている [58]．今後は，このような攻撃に対する話者照合システムの安全性を高めるために，どのような対策をとればよいかを検討する必要がある．

6.5 バイモーダル (audio-visual) 音声合成

実世界における人間同士の音声対話では，音声情報を担うのは音だけではなく，唇形状等の視覚的な情報が影響を与えることが知られており，このような音声のバイモダリティを考慮したヒューマンインタフェースに関する研究が盛んに行われている [59], [60]．トーキングヘッドなどへの応用を考えたバイモーダル音声合成（音声と唇動画像の同時生成）もそのひとつである．前節までで述べた音声合成手法を，このような唇動画像の生成に応用することができる．唇画像のモデル化手法は，唇の形状モデルを用いるモデルベース法，画素値をそのまま扱う画像ベース法に大別することができるが，いずれにおいても，HMM に基づいたパラメータ生成手法により，良好な唇動画像が生成できることが確かめられている [61], [62]．

7. むすび

動的特徴を用いた HMM に基づいて音声を合成する手法について，他の手法とも関連づけながら述べた．その結果，HMM に基づいた音声合成手法は，最近の他の音声合成手法の特徴のいくつかを，HMM という枠組の上で実現していることが示された．更に，音声合成に必要な情報が HMM という統計モデルによって表現されているために，HMM のモデルパラメータを適切に変換することにより，元の音声データベースに存在しない多様な音声を合成できる可能性がある．HMM のモデルパラメータの変換は，音声認識の分野で，話者適応，環境適応などを目的として，広く研究されており，理論的にもよく整備されているため，これらを利用することができる．このような利点の代償として，音声データベース中の音声波形そのものを用いるのではなく，音声パラメータから音声を合成するため，合成音声が，いわゆる vocoded speech となるが，最近提案されている高品質ボコーダの励振源生成手法を導入することにより，この問題は克服できるものと考えられる．

音声認識と音声合成を，HMM に基づいてひとつの枠組でとらえることができるため，テキスト音声合成以外にもいくつかの応用を考えやすくなる．このような応用についてもいくつかを簡単にまとめた．

今後，多様な発話スタイルをもった音声合成の研究を進めるためには，多様な発話スタイルで発声された音声のデータベースの整備が不可欠であるが，どのような音声をどのような収録条件で，またどのくらいの量を収集すればよいのかということがまだわかっておらず，このような目的のための音声データベース構築法を確立する必要がある．

謝辞 本文は，東京工業大学大学院総合理工学研究科小林隆夫教授，益子貴史助手との共同研究の成果に基づいている．研究の初期段階において，HMM に基づいた音声認識に関して様々な御教示を頂いた音声認識研究者の方々に感謝致します。

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 concatenarive 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] 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.

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

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

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

[12] 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.

[13] 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.

[14] A. Ljolje and F. Fallside, “Synthesis of natural sounding pitch contous in isolated utterances using hidden Markov models,” IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-34, no.5, pp.1074–1080, Oct. 1986.

[15] E. P. Farges and M. A. Clements, “Hidden Markov models applied to very low rate speech coding,” Proc. ICASSP, pp.433–436, 1986.

[16] T. Fukada, Y. Komori, T. Aso and Y. Ohora, “A study of pitch pattern generation using HMM-based statistical information,” Proc. ICSLP, pp.723–726, 1994.

[17] 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.

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

[19] 徳田恵一, 益子貴史, 小林隆夫, “尤度最大化基準によるHMMからの音声パラメータ生成法の検討,” 日本音響学会講論集, Sep. 1999(発表予定).

[20] 立和航, 古井貞熙, “HMMによる規則音声合成の検討,” 日本音響学会講論集, 2-3-7, Mar. 1999.

[21] Y. Ariki and M. A. Jack, “Enhanced time duration constraints in hidden Markov modelling for phoneme recognition,” Electronics Letters, vol.25, no.13, June 1989.

[22] 吉村貴克, 徳田恵一, 益子貴史, 小林隆夫, 北村 正, “HMMに基づく音声合成のための状態継続長モデルの構築,” 信学技報, SP98-64, DSP98-85, 1998.

[23] 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.

[24] K. Shikano, “Evaluation of LPC spectral matching measures for phonetic unit recognition,” CMU Technical Report CMU-CS-86-108, Computer Science Department, 1986.

[25] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” J. Acoust. Soc. America, vol. 87, pp.1738–1752, Apr. 1990.

[26] 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井 聖, “メルケプストラムをパラメータとする音声のスペクトル推定,” 信学論(A), vol.J74-A, no.8, pp.1240–1248, Aug. 1991.

[27] 今井 聖, 住田一男, 古市千枝子, “音声合成のためのメル対数スペクトル近似(MLSA) フィルタ,” 信学論(A), vol.J66-A, no.2, pp.122-129, Feb. 1983.

[28] 徳田恵一, 小林隆夫, 深田俊明, 今井 聖, “音声の適応メルケプストラム分析,” 信学論(A), vol.J74-A, no.8, pp.1249–1256, Aug. 1991.

[29] A. V. McCree, T. P. Barnwell III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE Trans. Speech and Audio Processing, vol.3, no.4, pp.242–250, July 1995.

[30] M. Nishiguchi, K. Iijima and J. Matsumoto, “Harmonic vector excitation coding of speech at 2.0 kbps,” Proc. IEEE Workshop on Speech Coding for Telecommunications, pp.39–40, Pocono Manor, Pennsylvania, Sep. 1997.

[31] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoer revisited,” Proc. ICASSP, pp.1303–1306, 1997.

[32] 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.

[33] G. J. Freij, and F. Fallside, “Lexical stress recognition using hidden Markov models,” Proc. ICASSP, pp.135–138, 1988.

[34] U. Jensen, R. K. Moore, P. Dalsgaard, and B Lindberg, “Modeling intonation contours at the phrase level using continuous density hidden Markov models,” Computer Speech and Language, vol.8, no.3, pp.247–260, 1994.

[35] K. Ross, and M. Ostendorf, “A dynamical system model for generating F_0 for synthesis,” Proc. ESCA/IEEE Workshop on Speech Synthesis, pp.131–134, 1994.

[36] 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.

[37] 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・状態継続長の同時モデル化,” 信学技報, SP99-, Aug. 1999.

[38] 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.

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

[40] N. Iwahashi and Y. Sagisaka, “Duration Modelling with Multiple Split Regression,” Proc. EUROSPEECH, pp.329–332, 1993.

[41] 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.

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

[43] 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.

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

[45] M. Plumpe, A. Acero, H. Hon and X. Huang, “HMM-based Smoothing for Concatenative Speech Synthesis,” Proc. ICSLP, vol.6, pp.2751–2754, 1998.

[46] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi and K. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” Proc. EUROSPEECH, vol.5, pp.2523-2526, Sep 1997.

[47] 橋本 誠, 樋口宜男, “話者選択と移動ベクトル場平滑化による声質変換のためのスペクトル写像”, 信学論(D), vol.J80-D-II, no.1, pp.1–9, Jan. 1997.

[48] 高橋淳一, 嵯峨山茂樹, “逐次型話者適応方式 MAP/VFS における分散適応,” 日本音響学会講論集, 2-5-5, Mar. 1995.

[49] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system,” Proc. ICASSP, vol.3, pp.1611–1614, Apr. 1997.

[50] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, ”Speaker adaptation for HMM-based speech synthesis system using MLLR,” Proc. ESCA/COCOSDA Workshop on Speech Synthesis, pp.273–276, Nov. 1998.

[51] S. Roucos and R. M. Schwartz and J. Makhoul, “A segment vocoder at 150 b/s,” Proc. ICASSP-83, pp.61–64, 1983.

[52] Y. Shiraki and M. Honda, “LPC speech coding based on variable-length segment quantization,” IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-36, no.9, pp.1437–1444, Sep 1989.

[53] J. Picone and G. R. Doddington, ”A phonetic vocoder,” Proc. ICASSP-89, pp.580–583, May 1989.

[54] F. K. Soong, “A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis,” Proc. ICASSP-89, pp.584–587, May 1989.

[55] Y. Hirata and S. Nakagawa, “A 100bit/s speech coding using a speech recognition technique,” Proc. EUROSPEECH-89, pp.290–293, Sep 1989.

[56] 広井 順, 徳田恵一, 益子貴史, 小林隆夫, 北村 正, “HMMに基づいた極低ビットレート音声符号化,” 信学技報, SP98-63, DSP98-84, pp.39–44, Sep. 1998.

[57] 松井知子, 古井貞熙, “テキスト指定型話者認識,” 信学論(D-II), vol.J79-D-II, no.5, pp.647–656.

[58] T. Masuko T. Hitotsumatsu, K. Tokuda and T. Kobayashi, “On the security of HMM-based speaker verification systems against imposture using synthetic speech,” Proc. EUROSPEECH, Budapest, Hungary, Sep. 1999 (accepted).

[59] D. G. Stork, and M. E. Hennecke, eds., “Speechreading by Humans and Machines,” NATO ASI Series, Springer Verlag, Berlin, 1996.

[60] 中村 哲, 山本英理, 永井 論, 鹿野清宏, “HMMを用いた音声と唇画像の統合による音声認識と唇画像生成,” 情報処理学会研究会, 音声言語情報処理15-17, Feb. 1997.

[61] 益子貴史, 小林隆夫, 徳田恵一, “HMMを用いた唇動画像の生成,” 信学技報, SP97-6, pp.33–38, May 1997.

[62] 広井 順, 徳田恵一, 益子貴史, 小林隆夫, 北村 正: “HMMに基づいた唇動画像の生成—画像ベースアプローチ—”, 日本音響学会講論集, 2-P-22, pp.311–312, Mar. 1999.