隠れマルコフモデルに基づいた韻律の統計モデル化手法と感情音声の生成
Prosody Modeling Based on Hidden Markov Models and Emotional Speech Generation

名古屋工業大学 大学院工学研究科
Graduat School of Engineering, Nagoya Institute of Technology

徳田恵一
Keiichi Tokuda

The increasing availability of large speech databases makes it possible to construct speech synthesis systems by applying statistical learning algorithms. This work aims at one of these corpus-based approaches: HMM-based speech synthesis in which synthetic speech is generated from HMMs themselves. The HMMs are widely used statistical models to characterize the sequence of speech spectra and have successfully been applied to speech recognition systems. From these facts, it is considered that the HMM is useful for modeling pitch patterns of speech. However, we cannot apply the conventional discrete or continuous HMMs to pitch pattern modeling since the observation sequence of pitch pattern is composed of one-dimensional continuous values and a discrete symbol which represents "unvoiced." The MSD-HMM which we have proposed includes discrete HMM and continuous mixture HMM as special cases, and further can model the sequence of observation vectors with variable dimension including zero-dimensional observations, i.e., discrete symbols. As a result, MSD-HMMs can model pitch patterns without heuristic assumption. Using the MSD-HMMs, we constructed an HMM-based speech synthesis system in which spectrum, pitch and state duration are modeled simultaneously in a unified framework of HMM. Through the four year period of research, we also improved the related techniques and algorithms used in the system and applied it to various types of speech synthesis: emotional speech synthesis, eigenvoice-based speech synthesis, etc.

Key words: HMM-based speech synthesis, fundamental frequency, prosody

1 研究の目的

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，動的特徴，混合ガウス分布，tying 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，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，(3) HMM 自身から音声を合成しようとするものなど，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，本研究では，コーパスに基づいた統計的なアプローチによる韻律のモデル化・生成手法を確立するため、隠れマルコフモデル (HMM) に基づいた一連の手法を提案し、感情音声他への適用を行い、多様な音声合成が可能となることを示した．以下に、各研究項目と成果をまとめる。

2 韻律の統計的モデル化手法の確立

HMM は音声スペクトル列の統計的モデル化手法として音声認識の分野で広く用いられており，ピッチパタンについても，HMM によりモデル化することができれば，音韻情報および韻律情報の統一的なモデル化を可能とし，よりきめ細かな音声信号モデルの構築に役立つものと思われる．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 HMM や連続 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，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，(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，(3) 無声区間のピッチの値は観測できなかったとする，などの方法が用いられている．本研究では，これらとは異なり，無声区間を含んだピッチパタンを直接 HMM によりモデル化することを可能とするため，可変次元の多空間上における確率分布に基づいた HMM を新たに定義し，拡張された HMM のモデルパラメータの再推定手法を与える．

2.1 多空間上の確率分布

空間インデックス $g = 1, 2, \ldots, G$ により参照される G 個の R^n 空間 $(R_1^{n_1}, R_2^{n_2}, \ldots, R_G^{n_G})$ を考える (図 1)．これらの空間はそれぞれ異なった次元 n_g をもつが，それらのうちのいくつかが同じ次元であってもよい．また，各空間上には，確率密度関数 $(\mathcal{N}_1^{n_1}, \mathcal{N}_2^{n_2}, \ldots, \mathcal{N}_G^{n_G})$ と，重み $(w_1, w_2, \ldots, w_G)$ が定義されている．但し，$\sum_{g=1}^{G} w_g = 1$ とする．

観測事象 $\boldsymbol{o}$ は，n 次元のベクトル $\boldsymbol{x}$ と，$\boldsymbol{x}$ がどの空間から出力されたかを表す空間インデックスの集合 X からなるものとする．すなわち

$$\boldsymbol{o} = (X, \boldsymbol{x}) \tag{1}$$

但し，X に含まれる空間インデックスが表す空間は，すべて $\boldsymbol{x}$ と同じ n 次元でなければならない．このとき，$\boldsymbol{o}$ の観測確率は，次式で定義することができる．

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g^{n_g}(V(\boldsymbol{o})) \tag{2}$$

但し，

$$S(\boldsymbol{o}) = X, \quad V(\boldsymbol{o}) = \boldsymbol{x} \tag{3}$$

零次元空間からの観測事象 $\boldsymbol{o}$ は，空間インデックスの集合 X だけからなり，$\boldsymbol{x}$ の値は存在しないが，記述の便宜上，$\mathcal{N}_g^0(V(\boldsymbol{o})) = 1$ と定義する．

図 1 の例では，観測事象 $\boldsymbol{o}_1$ は，空間インデックスの集合 $X = \{1, 2, G\}$ と，3 次元のベクトル $\boldsymbol{x}_1$ からなっている．従って，観測値 $\boldsymbol{x}_1$ は，三つの 3 次元空間 R_1^3, R_2^3, R_G^3 のいずれかから出力されたものであり，その出力確率は $w_1\mathcal{N}_1^3(\boldsymbol{x}) + w_2\mathcal{N}_2^3(\boldsymbol{x}) + w_G\mathcal{N}_G^3(\boldsymbol{x})$ で与えられる．以上で定義される確率分布は，$n_g = 0$ $(g = 1, 2, \ldots, G)$ のとき，離散分布と等価となり，また，$n_g = m$ $(g = 1, 2, \ldots, G)$, $X \equiv \{1, 2, \ldots, G\}$ のときには，m 次元 G 混合の連続分布と等価となることから，これらを一般化したものであることがわかる．

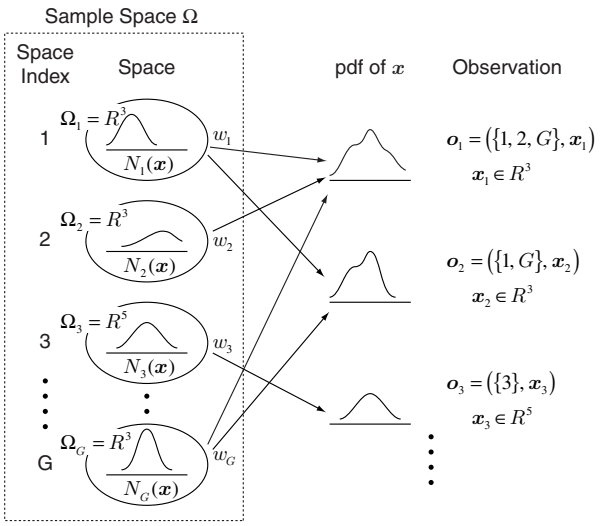

図 1: 多空間上の確率分布と観測事象

2.2 多空間上の確率分布に基づく HMM

前節で定義した多空間上の確率分布を用いて HMM λ を定義する．λ は状態遷移行列 $A = [a_{ij}]_{N\times N}$ と各状態 i での出力確率

$$b_i(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_{ig}\mathcal{N}_{ig}^{n_g}(V(\boldsymbol{o})) \tag{4}$$

からなる．但し，$a_{q_0 j}$ を初期状態確率とする．従って，各状態 i は，それぞれ G 個の分布 $(\mathcal{N}_{i1}^{n_1}, \mathcal{N}_{i2}^{n_2}, \ldots, \mathcal{N}_{iG}^{n_G})$ と空間の重み $(w_{i1}, w_{i2}, \ldots, w_{iG})$ をもつ．

このとき，観測系列 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ の出力確率は

$$P(\boldsymbol{O}|\lambda) = \sum_{\text{all } \boldsymbol{q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{o}_t) \tag{5}$$

$$= \sum_{\text{all } \boldsymbol{q},\boldsymbol{l}} \prod_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t l_t} \mathcal{N}_{q_t l_t}^{n_{l_t}}(V(\boldsymbol{o}_t)) \tag{6}$$

と書くことができる．但し，$\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ を状態系列，$\boldsymbol{l} = \{l_1, l_2, \ldots, l_T\}$ を観測系列 $\boldsymbol{O}$ に対して許される空間インデックスの系列 $\boldsymbol{l} \in \{S(\boldsymbol{o}_1) \times S(\boldsymbol{o}_2) \times \ldots \times S(\boldsymbol{o}_T)\}$ としている．

ここで定義した多空間上の確率分布に基づく HMM (以下，MSD-HMM) に関する学習アルゴリズムについては，[1] 他に詳しく記述した．

2.3 ピッチパタンのモデル化への適用

多空間上の確率分布は，離散分布と混合連続分布を特別な場合として含むため，新たに拡張された HMM は，離散 HMM と混合連続分布 HMM を特別な場合として含むものであり，更に，観測時刻毎に異なる次元の $\boldsymbol{x}$ が観測される場合に対応することができる．

音声のピッチの値は，有声区間においては連続値をとり，無声区間においては値が存在しない．このような信号は，有声区間の値を 1 次元空間における 1 次元確率分布からの出力値とし，無声区間を，2 節において定義した零次元の空間からの出力として，これらが時間的に混成された信号ととらえることができる．すなわち 2.1 節において $G = 2$，$n_1 = 1$，$n_2 = 0$ としたとき，この二つの空間上の確率分布に基づく HMM はピッチパタンを観測系列として直接扱うことができる．

このような考えに基づいた音声合成システムを構築し，有効性を確認した [2].

3 HMM による韻律生成手法の確立

これまでに開発した HMM から基本周波数パターンを生成する手法を更に高性能化するため、「ガンマ分布による継続時間長モデル」を導入した。これにより、より少ない HMM のモデルパラメータ数で同等の基本周波数パターンを生成できることを示した。また、「有声・無声境界でのダイナミックスを考慮した基本周波数パターンのモデル化手法」を導入し、生成される基本周波数パターンの自然性の向上をはかった。

3.1 継続長の高精度モデル化

ガウス分布の定義域は一般に実数全体に及ぶことから，負数域でもその確率が 0 以外の値をとることがある．そのため，継続長をガウス分布を用いてモデル化した場合，正数域のみの生起確率を積分しても 1 とならない．さらに，その分布形状は平均 μ を中心に左右対称であるため，継続長のモデル化には必ずしも最適であるとは言い難い．

一方，ガンマ分布はガウス分布と同様，パラメトリックな分布であり，その密度関数は以下の式で定義される．

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} \exp(-x\beta), \quad x > 0 \tag{7}$$

分布形状を図 2 に示す．ガンマ分布は，正数域で定義される分布である．さらに，ガウス分布に比べ，

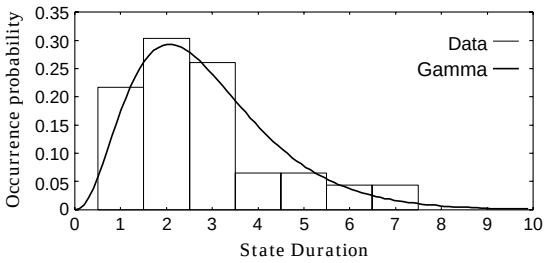

図 2: ガンマ分布の分布形状

その分布形状が左右非対称であることなどから，より精度の高い継続長のモデル化が可能であると期待できる．

本研究では，音声パラメータをスキップなしの left-to-right HMM でモデル化し，状態継続長分布を音声パラメータのモデルにおける連結学習の際のトレリス上で求める．状態継続長のガンマ分布の次元は HMM の状態に対応している．

クラスタリングにおけるデータセットの分割は，ガウス分布を用いた場合 [2] と同様，分割前後でのデータセットの記述長の変化に基づいて行われる．

評価実験では，従来のガウス分布の場合に比べ，モデル数が約 24% 減少したにも関らず，合成音による対比較試験では，ほぼ同等かそれ以上のスコアを得ることを確認し，より精度の高い継続長モデルを構築できることを示した [3].

3.2 F0 の高精度モデル化

MSD-HMM による F0 のモデル化においては，文献 [4] のパラメータ生成アルゴリズムを用いることにより，動的特徴量を考慮した F0 生成を行っていた．ところが，有声/無声境界の動的特徴量を計算しておらず，有声/無声境界の動的特徴量を無声を表すシンボルとしてピッチパターンをモデル化していた．このため，ピッチパターンを生成したとき，有声/無声境界で動的特徴量が考慮されないため，短い無声区間を挟んだ 2 つの有声区間の間で不連続なピッチパターンが生じ，原音声と比較してアクセントが異なって聞こえる場合があった．また，静的特徴量が有声であるのに，動的特徴量は無声であるといった矛盾が生じていた．そこで，当該フレーム及び前後最近傍の有声フレームの静的特徴量より動的特徴量を求め，有声/ 無声境界の動的特徴量を考慮してより精度の高いピッチパターンのモデルを構築する．また，ピッチパターン生成においても有声/無声境界の動的特徴量を考慮してパラメータ生成を

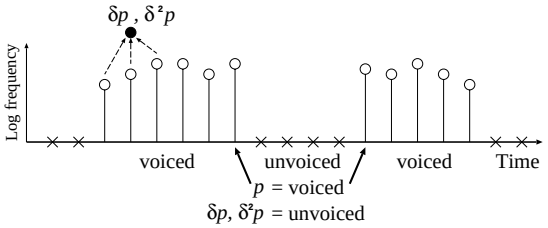

図 3: 従来の動的特徴量の計算

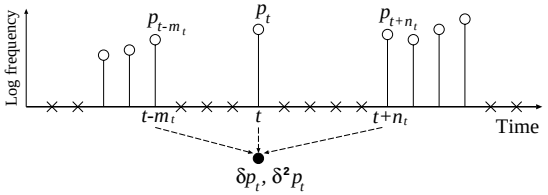

図 4: 境界を考慮した動的特徴量の計算

行い，不連続な区間がない自然なピッチパターンを生成する．

文献 [2] では図 1 のように，当該フレームとその前後それぞれ 1 フレームずつで計算した 1 次及び 2 次の回帰係数から微分係数に対応する値を求め，動的特徴量としてモデル化した．時刻 t における 1 次と 2 次の動的特徴量 δp_t, $\delta^2 p_t$ は以下のようになる．

$$\delta p_t = \frac{1}{2}p_{t+1} - \frac{1}{2}p_{t-1} \quad (8)$$

$$\delta^2 p_t = p_{t-1} - 2p_t + p_{t+1} \quad (9)$$

δp_t, $\delta^2 p_t$ はそれぞれ，計算に利用される静的特徴量が全て有声の場合のみ計算され，無声区間及び有声/無声境界の有声フレームについては，動的特徴量を計算していない．そこで図 2 に示すように，当該フレーム及び前後最近傍の有声フレームから回帰係数を計算する．得られた回帰係数から微分係数に対応する値を求め，これを動的特徴量とする．時刻 t の有声フレームにおいて，前後最近傍フレームまでのフレーム数をそれぞれ m_t, n_t としたとき，当該フレームの 1 次及び 2 次の微分係数を求める式は以下のようになる．

$$\delta p_t = \frac{m_t p_{t+n_t}}{m_t n_t + n_t^2} + \frac{(n_t - m_t)p_t}{m_t n_t} - \frac{n_t p_{t-m_t}}{m_t^2 + m_t n_t} \quad (10)$$

$$\delta^2 p_t = 2\left\{\frac{p_{t-m_t}}{m_t^2 + m_t n_t} - \frac{p_t}{m_t n_t} + \frac{p_{t+n_t}}{m_t n_t + n_t^2}\right\} \quad (11)$$

式 (10), (11) は $m_t = n_t = 1$ のとき，それぞれ式 (8), (9) と等しい．音声の最初の有声/無声境界では，当該フレームより前の有声フレーム，最後の有声/無声境界では当該フレームより後の有声フレームが存在しない．このため，式 (10), (11) において音声

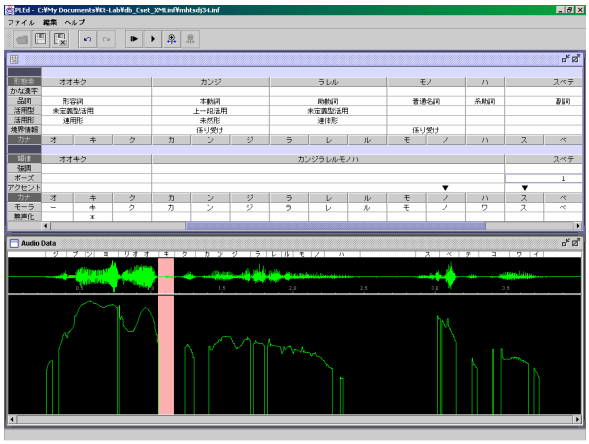

図 5: 韻律・言語情報編集ツール PLEd の概観

の最初の有声/無声境界では m_t, 最後の有声/無声境界では n_t を無限大として計算する.

以上の手法によれば，有声/無声境界での動的特徴量を考慮することができ，短い無声区間を挟んだ有声区間の境界において不連続な区間が生ぜず，自然で滑らかな F0 パターンが生成されることを確認した [5].

4 感情を含む韻律の統計的モデル化手法の開発

これまでの成果をもとに、多様な声質、感情、発話スタイルの韻律生成が可能なシステムを開発し、合成音声の評価を行った。特に、音声データベースの韻律ラベルの自動付与、感情音声データベースのラベリング手法などについても検討を行った。また、発話スタイル、感情のモデル化を考慮したモデル学習アルゴリズムについて更なる改良を行った。

4.1 韻律生成 HMM のための学習データ作成ツール

統計モデルを学習するための学習用音声データには，音素ラベルおよび音素境界を付与するのみでなく，品詞，アクセントなどの言語・韻律情報を付与する必要がある．計算機によるデータ処理能力の向上を背景に，大量の音声データを学習に利用することが可能となったが，これらのラベル付作業は人手を介するため，大量の学習データ投入の障害となっていた．ピッチパターンについては，ToBI, J-ToBI などの記述法が提案されているが，ピッチパターン以外にも，品詞，活用形，活用形，強調，文構造，発音など付与すべき情報は複数あり，それらはお互いに何らかの相互関係をもっているため，独立に取り扱った場合には，データ間の不整合性を引き起こす恐れがある．本研究では，韻律合成用 HMM のための学習データを一括管理し，効率良く入力・編集する手法について検討を行い，実際に編集ツールを試作した [6]．図 5 に概観を示す.

言語・韻律情報の記述法を XML に基づいて定義した．本記述法を用いることにより，アクセント，品詞，活用形，活用形，強調，文構造，発音など必要となる複数の情報を一元管理することが可能となり，これらを矛盾なく，入力・編集することができる．本記述ファイルは，何らかの GUI により編集されることを想定しており，GUI を備えた編集プログラムを実装することにより，効率的な言語・韻律情報の入力作業が可能となることが期待される．また，GUI を備えた編集プログラムが具備する機能についても検討を行った．GUI 開発の際には，記述ファイル自身が高い可読性をもつため，容易にデバッグが行える利点がある.

なお、本研究では，韻律合成用 HMM のための学習データを入力・編集することを想定したが，本方式に依存した部分は，ほとんどないため，作成された言語・韻律情報データを他の音声合成方式で利用することも可能である.

4.2 感情音声合成

ユーザ補助やロボット等とのコミュニケーションを目的として，合成音声に多様な表現をさせようという試みが，数多く行われている．その中で，近年，合成音声の感情表現に関する研究例も数多く報告されている．我々は，HMM 音声合成システムの利点を活かし，感情音声をモデルの学習に用いることで音声の合成を試みた [7].

今後，個性豊かな音声合成システムを構築するためには，一般ユーザの発声した音声により，システムの学習を行う必要があると考えられる．しかし，学習用の感情音声を提供する発話者がナレータ，声優などの専門家でない場合には，安定して感情表現することは容易ではないという問題がある．また，文章の意味内容や感嘆詞によって，一文章中でも感情の表出度合が変化することが考えられる．このような学習データに基づいて構築されたシステムでは，所望の感情表現をともなった合成音声が得られない可能性がある.

そこで本研究では，感情の表出度合に関する情報 (以下，感情強度とする) を学習データに与え，モデ

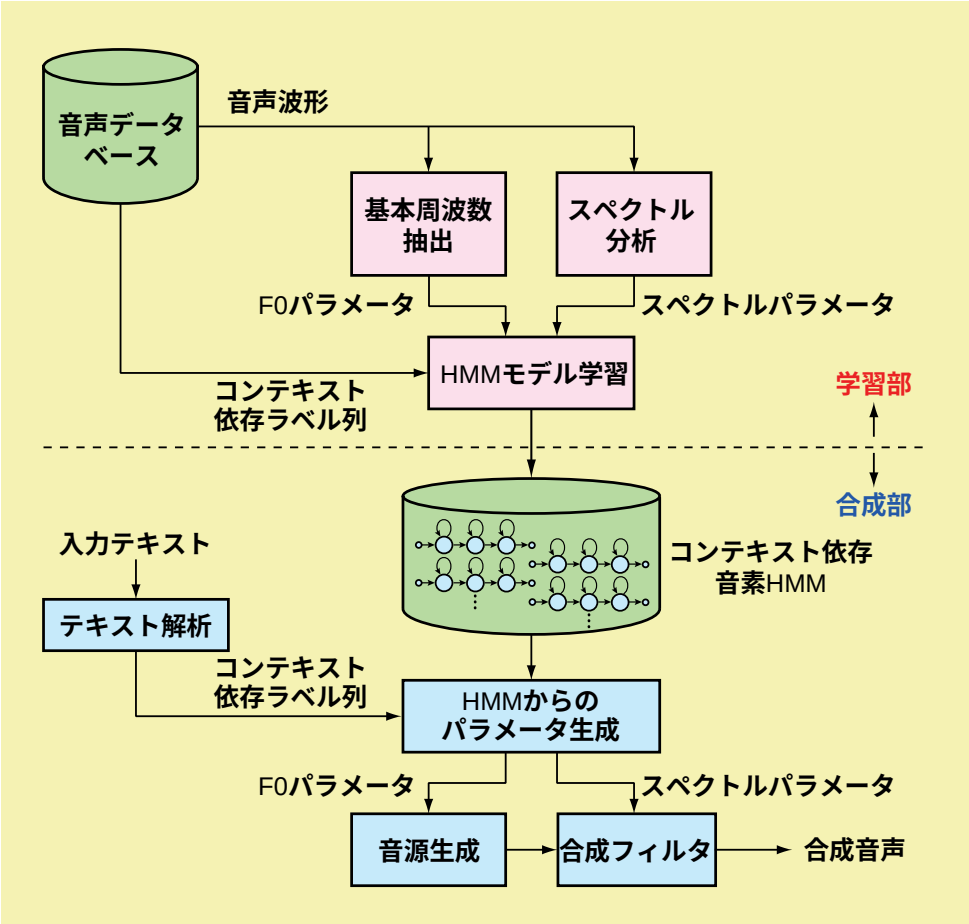

図 6: HMM 音声合成システムのブロック図

ル学習におけるコンテキストとして使用する手法を提案した．

実際に，は感情音声を用いた HMM 感情音声合成において，感情強度を考慮してモデル化を行い，合成音声の評価，合成用テキストに与える感情強度の検討を行った．主観評価実験により，モデル化の際にコンテキストとして感情強度を考慮することで，合成音声に異なる感情表現を与えられること，強い感情強度を与えることで，より自然な怒りの感情表現を含む音声が合成できることが確認された．

5 統計的モデルを用いた韻律生成システムの開発

これまでの成果をもとに、多様な声質、感情・発話スタイルの韻律の生成が可能なシステムを開発し、音声合成によって評価した。また、「固有声手法を」導入することににより、ユーザーが好みの声質を自在に設定可能なシステムを構築した。

5.1 HMM 音声合成ツールキットの開発

HMM 音声合成システムの概要は図 6 に示す通りである．システムは学習部，合成部から構成される．

学習部では，音声データからスペクトルパラメータとしてメルケプストラム，基本周波数パラメータとして対数基本周波数を求め，これらの 1 次及び 2 次の動的特徴量をフレーム毎に結合して特徴ベクトルとする．スペクトルは通常の連続分布 HMM，基本周波数は多空間上の確率分布に基づいた HMM(MSD-HMM)，継続長は HMM の各モデルの状態継続長を多次元のガウス分布でそれぞれモデル化する．スペクトル，基本周波数，継続長は音素，アクセント型，形態素など様々な要因によって変動することから，モデル化の際，これらの変動要因の組み合わせ (コンテキスト) 毎に別々にモデル化し，MDL 基準に基づく決定木によるコンテキストクラスタリングを適用して分布の共有化を行っている．この際，スペクトル，基本周波数，継続長はそれぞれ異なる変動要因に依存すると考えられるため，別個にクラスタリングを行う．また，MDL 基準による Tree-based ク

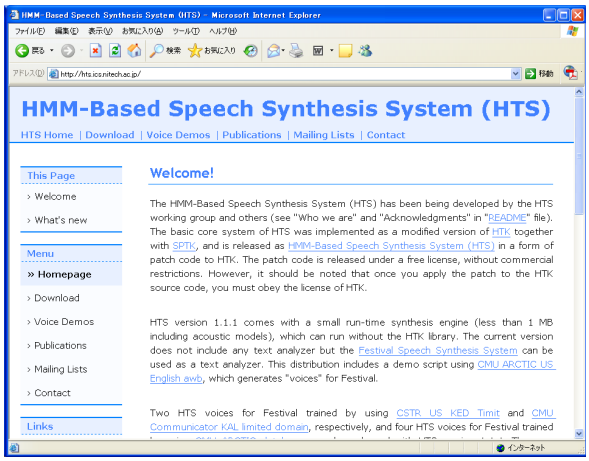

図 7: HMM 音声合成ツールキット HTS (http://hts.ics.nitech.ac.jp/)

ラスタリングは，において，MSD-HMM に対して拡張されている．

合成部では，まず，合成する文章を変動要因を考慮したラベル列に変換し，得られたラベル列に従ってコンテキスト依存 HMM を結合し，文章に対応する HMM を構成する．次に継続長分布に従って各状態の継続長を決定し，メルケプストラム列及び基本周波数パターンをパラメータ生成アルゴリズム ([4] の case 1) により生成し，MLSA フィルタを用いて波形を合成する．

これら一連のアルゴリズムを，共通の研究基盤として提供するため，オープンソースのフリーソフトウェアとして公開した．図 7 にウェブサイトのホームページを示す．平成 15 年度の延べダウンロード件数は，1000 件近くとなっており，実際に世界的に有数の研究機関で広く利用されている．

5.2 固有声手法による多様な声質の実現

HMM 音声合成システムは，HMM 自身から音声パラメータを出力し，音声を合成することを特徴としており，HMM のパラメータを変換することによって様々な声質の音声を合成できるという利点がある．実際に，話者適応，話者補間の手法を適用することにより，様々な声質を容易に生成できることを確かめている．話者適応の手法では，少量の適応データにより目標話者の声質を模擬できるが，適応データが入手できない場合には適用できないため，実際に存在しない話者の声質を実現したいという用途には向かない．一方，話者補間の方法によれば，適応データなしで様々な声質を実現できるが，声質の自由度

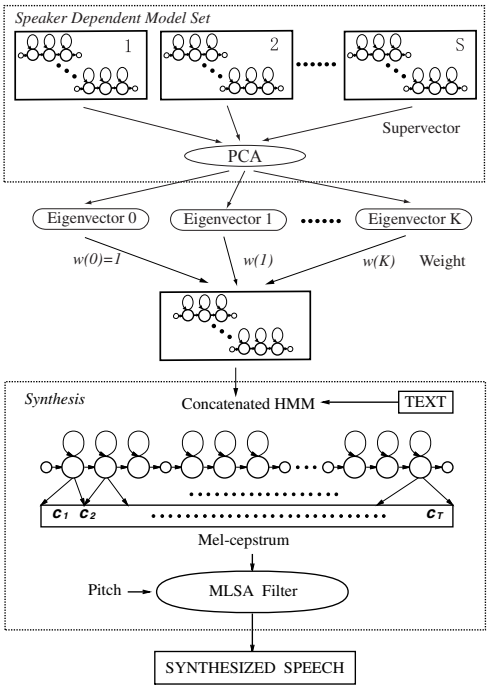

図 8: 固有声システムのブロック図

を高くするため，代表話者の数を増やした場合に，それぞれの話者にどのような重みを設定すれば所望の声質が得られるかをユーザが把握するのが難しくなる．

適応データが入手できない場合に，ユーザが所望の声質を容易に得られるようなシステムの構築を目指して，固有声 (eigenvoice) を用いた音声合成システムを提案した（図 8）．固有声の手法は，顔認識のために用いられる固有顔 (eigenface) の手法を基にしており，多人数の特定話者 HMM の集合を主成分分析（PCA）などの手法により，少数の自由パラメータで表現する手法であり，少量の学習データで話者適応をすることが可能となる．各固有ベクトルは（固有声ベクトルと呼ぶ）性別などの類型的な声質を表している可能性があり，本文では，固有声ベクトルの各成分の重みの値により合成音声の声質がどのように変化するのかについて検討を行った．

図 8 に固有声手法に基づく音声合成システムを示す．音声合成システムは，固有声ベクトル生成部と音声合成部から構成される．固有声ベクトル生成部では，音声データベースからメルケプストラム分析によりメルケプストラムを求め， 静的特徴量と動的特徴量 (Δ および Δ^2) を特徴ベクトルとして特定話者の HMM を作成する．各特定話者 HMM の平均ベクトルからスーパーベクトルを作り，それらに PCA を施すことによって得られた固有声ベクトル $\boldsymbol{e}(j)$ を音声合成システムに蓄積する．

ユーザは，それぞれの好みの声質に合わせて $w(j), j = 1, 2, \ldots, K$ の値を決定する．固有ベクトル $\boldsymbol{e}(j)$ を $w(j)$ により線形結合することにより，所望の声質を表すスーパーベクトルが得られれば，合成音声の生成過程は，通常の HMM 音声合成システムと同様である．まず，合成したい任意のテキストをコンテキストに基づいたラベル列に変換する．このラベル列に従って HMM を結合し，1 つの文 HMM を構成する．状態継続長分布に従って各状態の継続長を決定し，尤度最大化基準に基づくパラメータ生成アルゴリズムにより，文 HMM からメルケプストラム列を生成する．最後に，生成したメルケプストラムから，MLSA フィルタにより音声を合成する．基本周波数についても，HMM によってモデル化することにより，固有声の手法を適用することができる．実際に，多様な韻律的特徴をもった音声の合成が可能であることを示した．

6 まとめ

以上，本研究による得られた成果をまとめた．コーパスに基づいた統計的なアプローチによる韻律のモデル化・生成手法を理論的な面から確立し，更に，実際的なシステムを構築することにより，その妥当性，有用性を示すことができた．

紙面の関係で，記述しきれなかった成果を以下に列挙しておく．詳しくは発表論文リストにあげた論文を参照されたい．

- 音声合成システムの自動構築法
- 唇動画像と音声の同時生成
- HMM の状態共有法
- 音声品質改善のためのポストフィルタ
- 混合励振源モデル
- DAEM アルゴリズムによる音声合成用 HMM の学習
- 新しい統計モデルである trajectory-HMM の提案と音声合成への応用
- アクセント情報の自動ラベリング
- 継続長分布付 HMM 学習による音声合成
- 英語、中国語、ポルトガル語など，他言語への拡張

今後は，本研究の成果を生かし，人間のように自然で多様な音声を自在に生成することの可能な「仮想声優システム」を構築したいと考えている．

参考文献

[1] 徳田 恵一, 益子 貴史, 宮崎 昇, 小林 隆夫, “多空間上の確率 分布に基づいた HMM,” 信学論 (D-II), vol.J83-D-II, no.7, pp.1579–1589, July 2000.

[2] 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.

[3] 石松喜伸, 徳田恵一, 益子貴史, 小林隆夫, 北村正, “HMM 音声 合成のおけるガンマ分布状態継続長モデルの検討,” 電子情報通信学会技術研究報告, vol.101, no.352, SP2001-81, pp.57-62, Oct. 2001.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. of ICASSP 2000*, vol.3, pp.1315–1318, June 2000.

[5] 全 炳河，徳田 恵一，益子 貴史，小林 隆夫，北村 正, “有声/ 無声境界の動的特徴量を考慮したピッチのモデル化,” 電子情報通信学会技術研究報告, vol.101, no.325, SP2001-70, pp.53-58, Sep. 2001.

[6] 徳田恵一，水谷伸晃，酒向慎司，石松喜伸，吉村貴克，江本喜久 男，河井 恒, “韻律生成 HMM のための学習データ作成ツール,” 日 本音響学会 2003 年春季研究発表会講演論文集, vol.I, 1-6-19, pp.259-260, Mar. 2003.

[7] 都築亮介, 全 炳河, 徳田恵一, 北村 正, ”HMM 音声合成における 感情表現のモデル化,” 電子情報通信学会技術研究報告, vol.103, no.264, pp.25-30, Aug. 2003.