論　文

動的特徴を用いたHMMに基づく音声合成

益子　貴史[†]　　徳田　恵一[††]　　小林　隆夫[†]　　今井　聖[†]

HMM-Based Speech Synthesis Using Dynamic Features

Takashi MASUKO[†], Keiichi TOKUDA[††], Takao KOBAYASHI[†], and Satoshi IMAI[†]

あらまし　隠れマルコフモデル（HMM）からの動的特徴を用いた音声スペクトルパラメータ生成アルゴリズムに基づく規則音声合成システムの新たな枠組みを提案している．本システムで用いるパラメータ生成アルゴリズムでは，HMMで学習した静的，動的特徴の統計情報に従って連続的に遷移するスペクトル系列を生成することができる．規則音声合成にこのアルゴリズムを適用することにより，滑らかで自然性の高い音声を合成できると考えられる．本論文ではこのHMMに基づく規則音声合成システムの枠組みを示し，韻律生成部を除く合成システムを構築した．生成されたスペクトルパラメータを用いて合成した音声の主観評価実験により動的特徴の有効性を示すと共に，合成単位である音素HMMの構成について，音素環境依存性など，いくつかの検討を行っている．

キーワード　規則音声合成，隠れマルコフモデル，動的特徴，メルケプストラム分析，MLSAフィルタ

1. まえがき

隠れマルコフモデル（HMM）は，音声スペクトル系列の統計的モデル化手法であり，音声認識等の分野でその有効性が示されている．HMMによりモデル化された音声の統計的な特徴量から音声パラメータを生成することができれば，より柔軟な音声合成ができることが期待できる．

このような観点から，混合連続出力分布型HMMから音声スペクトルパラメータを生成する手法を提案した[1]～[3]．ほかにもHMMに基づく音声合成に関する報告がいくつかなされているが[5]～[9]，提案アルゴリズムはこれらの手法とは異なり，ゆう度最大化の基準により，HMMで学習した静的，動的特徴の統計情報に従って連続的な音声パラメータ系列を生成することができるという特徴がある．更に，メルケプストラム分析法[10]およびMLSAフィルタ[11]と組み合わせることにより，HMMから生成したスペクトルパラメータから直接音声を合成することができるという利点がある．

本論文では，音声合成システムにおける新たな枠組みとして，HMMからのパラメータ生成アルゴリズムを利用した規則音声合成システム[4]について述べる．提案するシステムでは音声単位として環境依存の音素HMMを用いており，データベースに基づく自動学習によりパラメータを自動的に決定することができる．近年実用化されている規則音声合成システムの多くは，スペクトルパラメータや波形などの領域において，音素，音節などの音声単位を接続することによって音声を合成する方式がとられている．しかしこれらの方式では，環境依存の音声単位を用いたり，複数音素の音声単位を用いた場合でも，音声単位の接続部分に必ず不連続が生じるため，何らかのヒューリスティックな接続規則を導入する必要がある．これに対し，HMMに基づく提案手法では，HMMで学習した統計情報に基づいてゆう度最大となるようにパラメータの変形がなされるため，音素境界部においても適切な変形による接続が行われると考えられる．また，話者性や発話スタイルなどの多様性を要求された場合に，これらを規則によって制御することは今のところ難しく，それぞれに対応した音声データを用意した場合には，記憶すべきデータが膨大になるという問題がある．それに対して，本方式では，HMMによってモデル化された表現を用いているため，音声単位当りのデータ量が少

[†] 東京工業大学精密工学研究所，横浜市
Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama-shi, 226 Japan
[††] 名古屋工業大学工学部知能情報システム学科，名古屋市
Faculty of Engineering, Nagoya Institute of Technology, Nagoya-shi, 466 Japan

ない上に，音声認識の分野で研究されているHMMに基づく話者適応の手法を応用することにより，多様な合成音声を得ることが容易であると期待される．

2. HMMに基づくパラメータ生成

連続出力分布型HMMのパラメータセットをλで表し，λからある状態遷移系列に沿って長さTの出力ベクトル系列$\{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$を生成することを考える[1]〜[3]．ここでは，HMMのそれぞれの状態は，ガウス分布でモデル化した状態継続長分布$p_q(d_q)$をもつとする．但し$p_q(d_q)$は，状態qがd_qフレーム継続する確率を表す．また，簡単のため，HMMは単一出力分布型left-to-rightモデルであると仮定する．

HMMの状態遷移系列を$\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$とし，$\boldsymbol{q}$に沿って出力されるパラメータ系列からなるベクトルを$\boldsymbol{o} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']'$とする．与えられたHMM λに対し，出力ベクトル$\boldsymbol{o}$は，$P(\boldsymbol{q}, \boldsymbol{o} \mid \lambda, T)$を$\boldsymbol{o}$と$\boldsymbol{q}$に関して最大化することにより得られる．状態遷移確率$P(\boldsymbol{q} \mid \lambda, T)$にかける重みを$a_d$とすると，出力ベクトル$\boldsymbol{o}$と状態遷移系列$\boldsymbol{q}$の同時生起確率$P(\boldsymbol{q}, \boldsymbol{o} \mid \lambda, T)$の対数は，

$$\log P(\boldsymbol{q}, \boldsymbol{o} \mid \lambda, T) = a_d \log P(\boldsymbol{q} \mid \lambda, T) + \log P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda, T) \quad (1)$$

と表すことができる．重みa_dは状態遷移確率$P(\boldsymbol{q} \mid \lambda, T)$と出力確率$P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda, T)$が全体の確率に及ぼす影響の比率を制御するパラメータである．$P(\boldsymbol{q}, \boldsymbol{o} \mid \lambda, T)$を最大化する際，状態遷移系列$\boldsymbol{q}$は，相対的に，$a_d$が小さければ出力確率による影響を，逆に$a_d$が大きければ状態遷移確率による影響を強く受けて決まる．

HMMはleft-to-rightモデルであると仮定しているため，状態遷移確率$P(\boldsymbol{q} \mid \lambda, T)$は状態継続長分布$p_q(d_q)$のみにより決定されるので，$T$フレーム中に$K$個の状態を遷移するとすると，

$$\log P(\boldsymbol{q}, \boldsymbol{o} \mid \lambda, T) = a_d \sum_{k=1}^{K} \log p_{q_k}(d_{q_k}) - \frac{1}{2}\log|\boldsymbol{U}| - \frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})'\boldsymbol{U}^{-1}(\boldsymbol{o}-\boldsymbol{\mu}) - \text{Const.} \quad (2)$$

となる．ここで，

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \ldots, \boldsymbol{\mu}_{q_T}']' \quad (3)$$

$$\boldsymbol{U} = \mathrm{diag}[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \ldots, \boldsymbol{U}_{q_T}] \quad (4)$$

であり，$\boldsymbol{\mu}_{q_t}$および$\boldsymbol{U}_{q_t}$はそれぞれ状態q_tの平均および共分散である．d_{q_k}はk番目の状態の継続長を表し，$\sum_{k=1}^{K} d_{q_k} = T$となる．また，Const.はそれぞれの出力ガウス分布の正規化係数の対数の和である．

$\boldsymbol{q}$が与えられたとき，$P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda, T)$を最大化する$\boldsymbol{o}$は，静的特徴$\boldsymbol{c}_t = [c_t(1), c_t(2), \ldots, c_t(M)]'$（例えばメルケプストラム係数，但し$M$は次数）のみを考慮する場合（$\boldsymbol{o}_t = \boldsymbol{c}_t$），$\boldsymbol{o} = \boldsymbol{\mu}$となる．このとき，個々のフレームでの出力は，前後のフレームでの出力とは独立に，そのフレームに対応する分布の平均となるため，ある状態から次の状態へ遷移する部分でスペクトルに不連続が生じてしまう．これを避けるため，出力パラメータに動的特徴を導入して$\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$とする．$\Delta^{(0)}\boldsymbol{c}_t = \boldsymbol{c}_t$，$\Delta^{(1)}\boldsymbol{c}_t = \Delta\boldsymbol{c}_t$，$\Delta^{(2)}\boldsymbol{c}_t = \Delta^2\boldsymbol{c}_t$とおけば

$$\Delta^{(n)}\boldsymbol{c}_t = \sum_{i=-L^{-(n)}}^{L^{+(n)}} w^{(n)}(i)\boldsymbol{c}_{t+i}, \quad n = 0, 1, 2 \quad (5)$$

と定義することができる．但し，$L^{-(0)} = L^{+(0)} = 0$，$w^{(0)} = 1$とする．$w^{(1)}(i)$，$w^{(2)}(i)$は，例えば

$$w^{(1)}(i) = \begin{cases} i/2, i = -1, 0, 1 \\ 0, \quad \text{otherwise} \end{cases} \quad (6)$$

$$w^{(2)}(i) = w^{(1)}(i) * w^{(1)}(i) \quad (7)$$

と定義される．但し，$*$は畳込みを表す．このとき，$\boldsymbol{o}$は静的なパラメータ系列からなるベクトル$\boldsymbol{c} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \ldots, \boldsymbol{c}_T']'$から

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \quad (8)$$

により計算される．但し，

$$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_T]' \quad (9)$$

$$\boldsymbol{w}_t = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}] \quad (10)$$

$$\boldsymbol{w}_t^{(n)} = [\underset{1\text{st}}{\boldsymbol{0}_{M\times M}}, \ldots, \boldsymbol{0}_{M\times M}, \underset{(t-L^{-(n)})\text{-th}}{w^{(n)}(-L^{-(n)})\boldsymbol{I}_{M\times M}}, \ldots, \underset{t\text{-th}}{w^{(n)}(0)\boldsymbol{I}_{M\times M}}, \ldots, \underset{(t+L^{+(n)})\text{-th}}{w^{(n)}(L^{+(n)})\boldsymbol{I}_{M\times M}}, \boldsymbol{0}_{M\times M}, \ldots, \underset{T\text{-th}}{\boldsymbol{0}_{M\times M}}]', \quad n = 0, 1, 2 \quad (11)$$

ここで，$\boldsymbol{0}_{M\times M}$，$\boldsymbol{I}_{M\times M}$はそれぞれすべての要素が0である$M \times M$の行列，および$M \times M$の単位行列とする．また，

$$\boldsymbol{c}_t = \boldsymbol{0}_M, \quad t < 1, T < t \tag{12}$$

と仮定している．ここで，$\boldsymbol{0}_M$ はすべての要素が 0 である $M \times 1$ のベクトルである．従って，$\boldsymbol{o}$ に関する $P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)$ の最大化は，静的なパラメータ系列からなるベクトル $\boldsymbol{c} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \ldots, \boldsymbol{c}_T']'$ に関する最大化となる．式 (2)，(8) より，$P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)$ を最大化する $\boldsymbol{c}$ は，次式の連立線形方程式

$$\boldsymbol{W}'\boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}'\boldsymbol{U}^{-1}\boldsymbol{\mu} \tag{13}$$

の解として与えられる [1]〜[3]．

3. 音声合成システムの構成

図 1 に，提案する音声合成システムのブロック図を示す．システムは，大きく学習部と合成部の二つの部分に分けることができる．学習部では，まず音声データベースからメルケプストラム分析 [10] によりメルケプストラム $\boldsymbol{c}_t$ を求める．更に $\boldsymbol{c}_t$ の値からデルタメルケプストラム $\Delta\boldsymbol{c}_t$ およびデルタデルタメルケプストラム $\Delta^2\boldsymbol{c}_t$ を式 (5)〜(7) により計算し，これらを特徴ベクトルとして音素 HMM を学習する．合成部では，まず合成したい任意のテキストを音素列に変換する．この音素列に従って前述の音素 HMM を接続し，与えられたテキストに対応する一つの文 HMM をつくる．

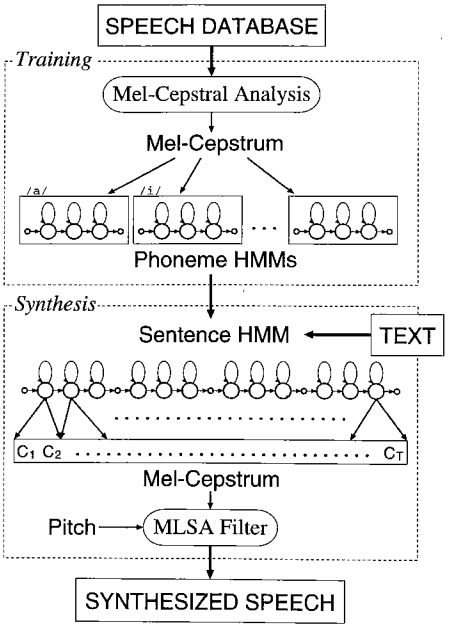


図 1 HMM を用いた規則音声合成システム
Fig. 1 Speech synthesis-by-rule system using HMM.

この文 HMM から，2. で述べたパラメータ生成アルゴリズムによりメルケプストラム系列を生成する．これに適当なピッチを与え，MLSA フィルタ [11] を用いて合成音声を得る．以下，この音声合成システムの具体的な構築例およびこれにより生成されたスペクトルの例を示す．

3.1 音声データベース

HMM の学習データとして，ATR 日本語音声データベースの話者 MHT による音韻バランス文 503 文を用いた．サンプリング周波数は 10 kHz である．本システムでは，/子音 + 拗音/や/子音 + 無声化母音/も一つの音素と考え，データベースに付属するラベルデータに基づいて，学習データを 60 種類の音素ラベルおよび無音ラベル（表 1）にラベル付けした．また，合成単位として triphone HMM を用いているが，学習データ中に存在する triphone は 3,518 種類である．

3.2 HMM 学習部

フレーム長 25.6 ms，フレーム周期 5 ms のブラックマン窓を用いて，メルケプストラム分析により 0 次から 15 次までのメルケプストラム $\boldsymbol{c}_t$ を求め，式 (5)〜(7) によりデルタメルケプストラム $\Delta\boldsymbol{c}_t$ およびデルタデルタメルケプストラム $\Delta^2\boldsymbol{c}_t$ を計算した．

使用した HMM は 3 状態または 5 状態の left-to-right モデルであり，パラメータ生成時の計算量を削減するため，それぞれの状態の出力分布は単一の対角共分散ガウス分布とした．本論文では，まず明示的な状態継続長分布をもたない通常の HMM の学習を行い，学習データに対する Viterbi セグメンテーションにより状態継続長分布を求め，モデルに付加した．

まず音素ラベルに従って，Baum-Welch アルゴリズムにより monophone HMM を学習した．次に，トレーニングデータ中に存在するすべての triphone について，それぞれの中心音素に対応する monophone HMM をコピーすることにより triphone HMM を生成した．そして，monophone HMM および triphone HMM のそれぞれのセットに対して連結学習によるパ

表 1 音素ラベルの種類
Table 1 Phoneme labels used in the system.

母音	a, i, u, e, o
子音	N, m, n, y, w, r, p, pp, t, tt, k, kk, b, d, dd, g, ch, cch, ts, tts, s, ss, sh, ssh, h, f, ff, z, j, my, ny, ry, by, gy, py, ppy, ky, kky, hy
子音 + 無声化母音	pi, pu, ppi, ki, ku, kku, chi, cchi, tsu, su, shi, shu, sshi, sshu, hi, fu
無音	sil

ラメータの再推定を行った．triphoneの種類は非常に多く，出現頻度の少ないものに対しては十分なトレーニングデータが得られない．このようなトレーニングデータの不足を補い，またシステムで保持するデータ量を削減するために，中心音素が同じtriphone HMMの，それぞれのモデル内での位置が等しい状態の集合に対して，文献[12]の手法により状態のクラスタリングを行い，同じクラスタに属する状態の出力分布を共有化した．この出力分布を共有化したtriphone HMM (tied triphone HMM) に対して再び連結学習によるパラメータの再推定を行った．

最後に，トレーニングデータに対してViterbiセグメンテーションを行い，monophone HMMおよびtriphone HMMそれぞれについて状態継続長分布を求め，単一ガウス分布でモデル化した．但し状態継続長分布は，出力分布の共有とは独立に個々の状態でもつこととした．

3.3 音声合成部

合成時には，まず合成したい任意のテキストを音素列に変換し，この音素列に従ってHMM学習部で学習したtriphone HMMを接続して与えられたテキストに対応する一つの文HMMをつくる．このとき，学習データ中に出現しないtriphoneに対しては，monophone HMMを利用する．この文HMMから**2.**で述べたアルゴリズムによりメルケプストラム系列を生成する．但しここでは，式(2)の状態遷移確率の重みを $a_d \to \infty$，すなわち状態遷移系列 $\boldsymbol{q}$ は，出力確率 $P(\boldsymbol{o}|\boldsymbol{q},\lambda,T)$ とは独立に状態遷移確率 $P(\boldsymbol{q}|\lambda,T)$ のみにより決定されることとする．このとき，与えられた全体のフレーム長 T に対し，個々の状態の継続長 $\{d_{q_k}\}_{k=1}^{K}$ は次式により求められる．

$$d_{q_k} = m_{q_k} + \rho \cdot \sigma_{q_k}^2 \tag{14}$$

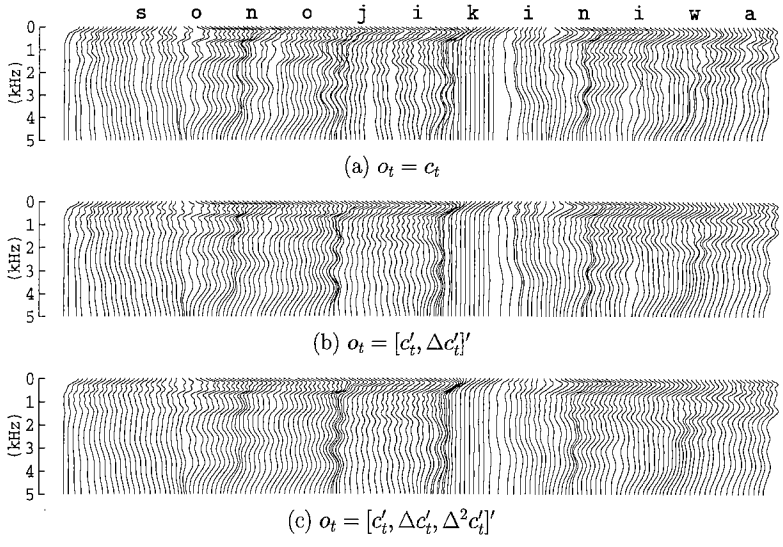
$$\rho = \left(T - \sum_{k=1}^{K} m_{q_k}\right) \Big/ \sum_{k=1}^{K} \sigma_{q_k}^2 \tag{15}$$

ここで，m_{q_k}，$\sigma_{q_k}^2$ はそれぞれ状態 q_k の状態継続長分布の平均および分散である．HMMは単一混合出力分布型left-to-rightモデルと仮定しているため，状態遷移系列 $\boldsymbol{q}$ は式(14)で得られた状態継続長 $\{d_{q_k}\}_{k=1}^{K}$ により一意に決定される．

このようにして生成されたメルケプストラム系列に適当なピッチを与え，MLSAフィルタを用いて合成音声を生成する．

3.4 生成スペクトルの例

図2に5状態HMMから生成された発声「その時期には」(/s-o-n-o-j-i-k-i-n-i-w-a/) のスペクトルの例を示す．図2(a)は静的特徴であるメルケプストラムのみを用いて学習したHMMから生成されたスペクトル，(b)はメルケプストラムとデルタメルケプストラムを用いて生成されたスペクトル，(c)はメルケプストラムおよびデルタ，デルタデルタメルケプスト



(a) $o_t = c_t$

(b) $o_t = [c_t', \Delta c_t']'$

(c) $o_t = [c_t', \Delta c_t', \Delta^2 c_t']'$

図2 HMMから生成されたスペクトルの例
Fig. 2 Examples of spectra generated from HMMs.

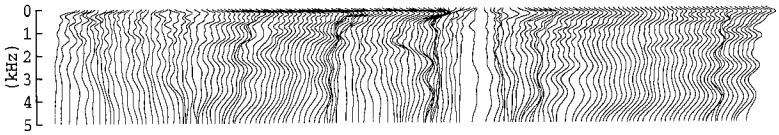


図3 実際の発声「その時期には」のスペクトル
Fig. 3 Spectra of real voice pronouncing /s-o-n-o-j-i-k-i-n-i-w-a/.

ラムを用いて生成されたスペクトルである．

図2(a)より，静的特徴のみの場合には，同じスペクトルが数フレーム続き，その後スペクトルが急激に変化していることがわかる．これらは，それぞれある状態の継続区間と次の状態へ遷移する部分に対応している．これは，静的特徴のみの場合には前後のスペクトルを考慮していないため，個々のフレームにおいて独立に対応する状態の出力分布のゆう度を最大化する値，すなわち平均値が出力されるためである．これに対し，動的特徴を考慮する場合には，あるフレームでの出力パラメータがその前後のフレームのパラメータに影響するため，静的特徴の平均値の列では必ずしもゆう度最大にはならない．ゆう度最大化基準により，静的特徴の平均のみではなく動的特徴の平均，更に静的，動的特徴の共分散を考慮して，連続的に変化するスペクトルを生成することができる（図2(b)，(c)）．

図2(b)，(c)を実際の発声（図3）と比べると，HMMの学習時にスペクトルパラメータが平均化されるために平たんなスペクトルとなっているものの，実際の発声に近いスペクトル系列が得られている．実際の発声から分析して得られたピッチを用いて合成した音声の非公式な受聴の結果，(a)はスペクトルの急激な変化による不連続感があるが，(b)，(c)は滑らかに接続されていることがわかった．また，分析合成音と比較すると多少音質が低下しているものの，個々の音韻をはっきりと知覚することができ，明りょう度の劣化は見られなかった．

4. 実　　験

動的特徴の効果およびパラメータ共有を行った際の総分布数と音質との関係を調べるため，対比較試験による主観評価実験を行った．被験者は，主観評価実験の経験はあるが本方式による合成音には慣れていない男性9名である．評価用データとしてトレーニングデータとは異なる12文章を合成した．但し合成音声のピッチは，実際の発声から分析して得られたものを用いた．12文章に含まれるtriphoneは延べ619個，このうち学習データ中に存在せずmonophoneで置き換えられたものは36個（5.8%）である．この12文章を4文章ずつ3セットに分け，個々の被験者はこれらのうちの1セット4文章について評価を行った．従って，1文章当り3人に評価されたことになる．被験者には特に評価基準に関する指示は与えておらず，2種類の音声を1組として聴取し，どちらがよいかを相対的かつ主観的に比較判断してもらった．この際，判定の信頼度については特に検討しておらず，得られた評価値をそのまま集計し，実験結果として示してある．

4.1 動的特徴の効果

まず，動的特徴の効果を調べるため，パラメータ共有を行わないtriphoneモデルを用いて，静的特徴であるメルケプストラム $\boldsymbol{c}$ のみを用いて学習・生成した音声，$\boldsymbol{c}$ のみから生成されたパラメータを隣接する状態の継続区間の中心間で線形に補間して生成した音声，$\boldsymbol{c}$ と $\Delta\boldsymbol{c}$ を用いて学習・生成した音声，$\boldsymbol{c}$ と $\Delta\boldsymbol{c}$，$\Delta^2\boldsymbol{c}$ を用いて学習・生成した音声の4通りについて比較を行った．使用したHMMは5状態left-to-rightモデルで，それぞれの状態の出力分布は単一の対角共分散ガウス分布とした．

実験結果を図4に示す．図より，動的特徴を用いた場合には，静的特徴のみの場合やその線形補間によるものよりも良い結果が得られている．これは，動的特徴を利用することによって，パラメータの変化の仕方も考慮し，与えられたHMMに対してゆう度が最大となるパラメータ系列を生成しているためであり，HMMで学習した統計情報を十分に活用しているためであると考えられる．

4.2 パラメータ共有

次に，HMMの総分布数と音質との関係を調べるため，特徴ベクトルとして $\boldsymbol{c}$，$\Delta\boldsymbol{c}$，$\Delta^2\boldsymbol{c}$ を用い，パラメータ共有を行った3状態および5状態triphone HMMそれぞれについて，状態クラスタリングの条件を変えることにより総分布数の異なるいくつかのHMMのセットについて主観評価実験を行った．3状態モデルでは，triphoneモデル（総分布数10554），総分

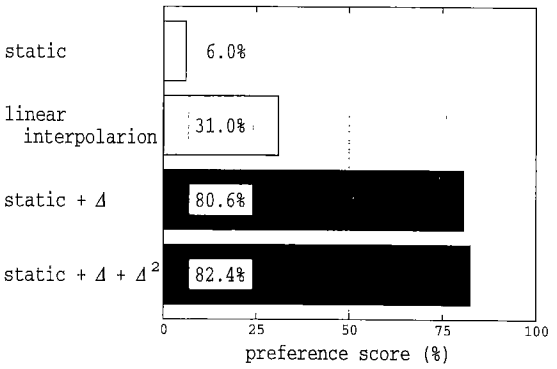
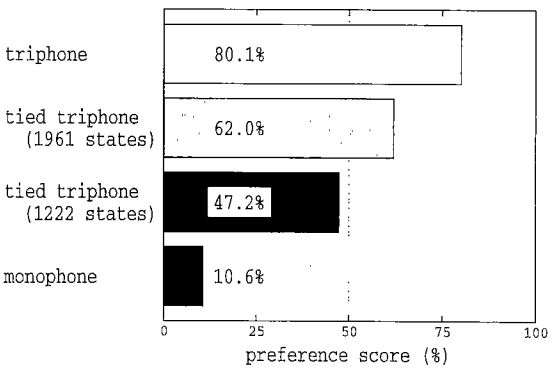


図4　動的特徴の効果
Fig. 4　An effect of dynamic features.

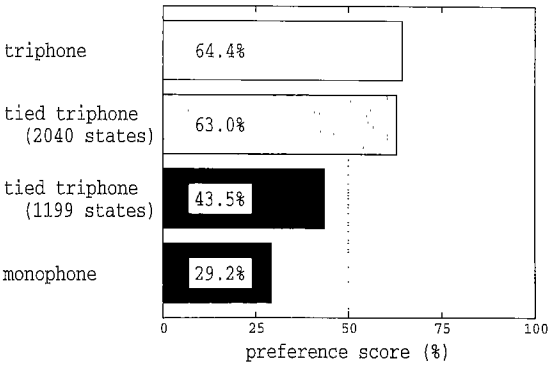
布数1961および1222のtied triphoneモデル，monophoneモデル（総分布数183）の4通り，5状態モデルでは，triphoneモデル（総分布数17590），総分布数2040および1199のtied triphoneモデル，monophoneモデル（総分布数305）の4通りについて実験を行った．但し，monophoneモデルの状態継続長分布は音素環境依存とした．

図5より，全体的に総分布数が多い程音質がよく，総分布数が減少するに従って音質が劣化することがわかる．非公式な受聴および生成されたスペクトルの観察により，分布数が多い場合には細かなスペクトルの変化を表現することができるが，分布数が少ない場合にはスペクトルが平均化されるために明りょう性が低下することがわかっている．またその半面，分布数が多い場合には分布当りのトレーニングデータが十分得られないために適当なモデルパラメータを推定することができず聴感上の不連続感が増すが，分布数が少ない場合にはより滑らかに変化するスペクトルが生成される．5状態triphoneモデルの場合には，分布数が多く不連続感が強くなるためにスコアが低下していると考えられる．なお，3状態および5状態いずれのmonophoneモデルを用いた場合にも了解度の低下は見られなかった．

次に，3状態および5状態tied triphoneモデルから生成された音声4通りについて主観評価実験を行った結果を図6に示す．図より，総分布数が約2,000の場合には3状態モデルおよび5状態モデルはほぼ同等であるが，総分布数が約1,200の場合には3状態モデルよりも5状態モデルの方がよい結果を示している．総分布数が同じならば，3状態モデルは5状態モデルよりもパラメータ空間方向の分解能が高く，逆に5状態



(a) 3-state HMMs.



(b) 5-state HMMs.

図5　総分布数と合成音の品質との関係
Fig. 5　Relation between total number of output distributions and synthesized voice quality.

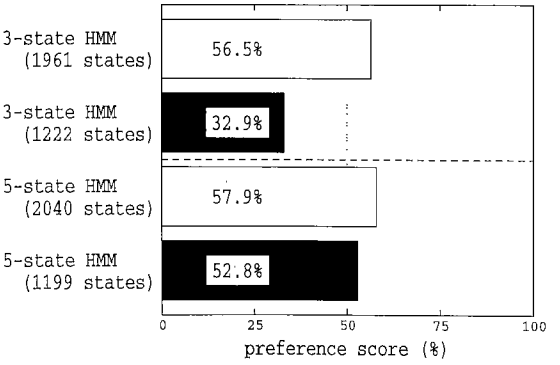


図6　3状態モデルと5状態モデルの比較
Fig. 6　Comparison between 3-state models and 5-state models.

モデルは3状態モデルよりも時間方向の分解能が高いと考えることができ，実験結果から分布数が少ない場合にはパラメータ空間方向よりも時間方向の分解能を高くした方がよいと考えられる．

但し，音質は学習データの量に強く依存すると考えられるため，学習データ量との関係など，更に詳細な検討をする必要がある．

5. むすび

本論文では，規則音声合成システムにおける新たな枠組みとして，動的特徴を利用したHMMに基づく音声パラメータ生成アルゴリズムを用いた規則音声合成システムについて述べ，合成単位である音素HMMの構成についていくつかの検討を行った．動的特徴を利用することにより，それぞれの状態がもつ静的，動的特徴の分布（平均および共分散）を考慮してパラメータを生成することができ，主観評価実験により，動的特徴のパラメータ生成に対する効果を確認した．

今後の課題としては，HMMに基づくパラメータ生成の枠組みで，音素環境や文脈などのさまざまな環境要因を考慮してピッチ，パワー，継続長の制御を行い，音質明りょう度や単語了解度，自然性などの合成システム全体の性能を評価することが挙げられる．また，音声認識の分野で研究されているHMMに基づく話者適応の技術を応用した多様な声質での音声合成などが挙げられる．

文　献

[1] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," ICASSP-95, pp.660–663, May 1995.

[2] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," EUROSPEECH-95, pp.757–760, Sept. 1995.

[3] 徳田恵一，益子貴史，小林隆夫，今井　聖，"HMMからの音声パラメータ生成アルゴリズム，" 信学技報，SP95-122, Jan. 1996.

[4] 益子貴史，徳田恵一，小林隆夫，今井　聖，"メルケプストラムをパラメータとするHMMに基づく音声合成，" 信学技報，SP95-123, Jan. 1996.

[5] 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, pp.1074–1080, 1986.

[6] E.P. Farges and M.A. clements, "An analysis-synthesis hidden Markov model of speech," ICASSP-88, pp.323–326, 1988.

[7] M. Giustiniani and P. Pierucci, "Phonetic ergodic HMM for speech synthesis," EUROSPEECH-91, pp.349–352, 1991.

[8] R.E. Donovan and P.C. Woodland, "Improvements in an HMM-based synthesizer," EUROSPEECH-95, pp.573–576, 1995.

[9] T. Fukada, Y. Komori, T. Aso, and Y. Ohara, "Fundamental frequency contour modeling using HMM and categorical multiple regression technique," J. Acoust. Soc. Jpn. (E), vol.16, no.5, pp.261–272, 1995.

[10] 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.

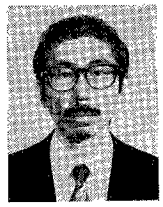
[11] 今井　聖，住田一男，古市千枝子，"音声合成のためのメル対数スペクトル近似（MLSA）フィルタ，" 信学論（A），vol.J66-A, no.2, pp.122–129, Feb. 1983.

[12] S.J. Young and P.C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," Computer Speech and Language, vol.5, no.3, pp.369–383, 1994.

（平成8年4月30日受付，8月5日再受付）

益子　貴史　（正員）

平5東工大・工・情工卒．平7同大大学院博士前期課程了（知能科学専攻）．同年同大学精密工学研究所助手．音声の分析・合成，音声認識の研究に従事．日本音響学会会員．



徳田　恵一　（正員）

昭59名工大・工・電子卒．平1東工大大学院博士課程了．同年東工大電気電子工学科助手．平8名工大知能情報システム学科助教授．工博．音声分析，音声合成・符号化，音声認識，ディジタル信号処理の研究に従事．日本音響学会，IEEE各会員．

小林　隆夫　（正員）

昭52東工大・工・電気卒．昭57同大大学院博士課程了．同年東工大精密工学研究所助手．工博．現在同助教授．ディジタルフィルタ，音声の分析・合成，音声認識の研究に従事．日本音響学会，IEEE各会員．

今井　聖　（正員）

昭34東工大・工・電気卒．昭39同大大学院博士課程了．同年東工大精密工学研究所助手．昭43同大助教授．昭54同大教授．工博．ディジタル信号処理，音声の合成および認識の研究に従事．昭45年度精機学会論文賞受賞．著書「ディジタル信号処理」など．計測自動制御学会，日本音響学会，IEEE，ASA各会員．