論　文

多空間確率分布HMMによるピッチパターン生成

益子　貴史[†]　　徳田　恵一[††]　　宮崎　昇[†*]　　小林　隆夫[†]

Pitch Pattern Generation Using Multi-Space Probability Distribution HMM

Takashi MASUKO[†], Keiichi TOKUDA[††], Noboru MIYAZAKI[†*], and Takao KOBAYASHI[†]

あらまし　隠れマルコフモデル（HMM）に基づいてピッチパターンとスペクトル系列を同時にモデル化及び生成する手法について述べる．ピッチパターンは，連続値をとる有声区間と値をもたない無声区間の時系列として表現されるため，通常のHMMではモデル化することができない．そこで本論文では，多空間上の確率分布に基づくHMM（multi-space probability distribution HMM: MSD-HMM）を適用し，ピッチパラメータとスペクトルパラメータを結合した特徴パラメータを用いてピッチとスペクトルを統一的にモデル化する手法を提案する．また，MSD-HMMにおける決定木に基づくコンテクストクラスタリング手法を導出し，ピッチやスペクトルの変動要因を考慮したモデルの構築手法について述べる．更に，ゆう度最大化基準に基づくパラメータ生成手法を用いることにより，実音声を近似したピッチパターン及びスペクトル系列を生成できることを示す．

キーワード　多空間確率分布，隠れマルコフモデル，コンテクストクラスタリング，ピッチパターン生成，音声合成

1. まえがき

隠れマルコフモデル（hidden Markov model: HMM）は，音声スペクトル系列の統計モデルとして，音声認識の分野において広く用いられている．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るが非常に柔軟であり，例えば，コンテクスト依存モデル[1]，動的特徴[2]，混合ガウス分布[3]，tying手法（例えば[4]），話者/環境適応化手法（例えば[5]）などの導入により，音声認識における認識率を大きく改善してきた．このことから，ピッチパターンについても，HMMによりモデル化することができれば，音韻情報及び韻律情報の統一的なモデル化を可能とし，より精密な音声信号モデルの構築に役立つことが期待される．また，我々はゆう度最大化基準によりHMMから音声スペクトルパラメータ系列を生成するアルゴリズム[6]を利用した音声合成システム[7]を提案しており，滑らかで自然性の高い音声スペクトル系列が得られること，また話者適応技術を応用することにより容易に多様な声質で音声合成できること[8],[9]を示した．このような観点から，スペクトル及びピッチを同時にHMMによりモデル化することができれば，韻律情報を含めて多様なテキスト音声合成が可能となることが期待される．

HMMをピッチパターンの生成に用いる試みはいくつか行われているが[10],[11]，ピッチパターンは，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を表す離散シンボルとして観測されるため，通常の離散HMMや連続HMMを直接用いることはできず，何らかの工夫が必要となる．例えば，（1）無声区間のピッチとして分散の大きな乱数を与える[12]，（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[13]，（3）無声区間のピッチの値は観測できなかったとしてEMアルゴリズムを適用する[14]，などの方法が用いられている．これに対し，本論文では，ピッチパターンを有声を表す1次元空間からの出力と無声を表す0次元空間からの出力が時間的に混在した系列としてとらえ，多空間上で定義される確率分布に基づくHMM（multi-space probability distribution HMM: MSD-HMM）[15]を用いてモデル化し，ゆう

[†] 東京工業大学大学院総合理工学研究科，横浜市
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan
[††] 名古屋工業大学知能情報システム学科，名古屋市
Faculty of Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan
[*] 現在，NTTコミュニケーション科学基礎研究所

度最大化基準によりピッチパターンを生成する手法を提案する．提案手法では，ピッチパラメータとスペクトルパラメータを結合して一つの特徴パラメータとすることにより，ピッチとスペクトルの統一的なモデル化，生成が可能となる．

ピッチやスペクトルは，音素や構文，韻律情報など，様々な要因によって変化する．これらの変動要因のすべての組合せ（本論文ではすべての要因を含めてコンテクストと呼ぶ）に対してそれぞれ別々にモデルを構築できれば，より精度の高いモデルが得られ，実音声をよく近似したピッチパターン及びスペクトル系列を生成できると考えられる．しかし，学習データが限られているため，考慮する変動要因の数が増大するにつれてモデル当りの学習データが減少し，モデルの信頼性が低下する．また，すべてのコンテクストを網羅する学習データを用意することは現実には不可能であるため，学習データに出現しないコンテクストへ対応するために，何らかの工夫が必要となる．そこで，HMM の状態を共有化してモデルパラメータの信頼性を向上させるとともに，学習データに出現しないコンテクストに対応するモデルも用意できるようにするため，決定木に基づくコンテクストクラスタリング手法 [4] を MSD-HMM に拡張し，状態のクラスタリングを行う．

以下，**2.** で MSD-HMM 及びそれに基づくピッチパターンのモデル化について述べ，**3.** で MSD-HMM における決定木に基づくコンテクストクラスタリングについて述べる．**4.** で提案手法を用いたピッチパターン及びスペクトル系列のモデル化及び生成について述べ，実音声から抽出されたピッチパターンと HMM から生成されたピッチパターンとの比較を行い，その有効性について検討する．

2. 多空間上の確率分布に基づく HMM によるピッチパターンのモデル化

2.1 多空間上の確率分布に基づく HMM

G 個の空間 $\Omega_1, \Omega_2, \ldots, \Omega_G$ からなる標本空間 Ω を考える．

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{1}$$

各空間 Ω_g は n_g 次元の実空間 R^{n_g} とする．G 個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．

各空間 Ω_g の確率を w_g，つまり，$P(\Omega_g) = w_g$，ただし，$\sum_{g=1}^{G} w_g = 1$ とする．更に，$n_g > 0$ のときには，各空間のもつ確率密度関数を $\mathcal{N}_g(\boldsymbol{x})$，$\boldsymbol{x} \in R^{n_g}$ とする．ただし，$\int \mathcal{N}_g(\boldsymbol{x}) d\boldsymbol{x} = 1$ とする．また，$n_g = 0$ のときには，空間 Ω_g は一つの標本点だけからなるとする．条件 $\sum_{g=1}^{G} w_g = 1$ 及び $\int \mathcal{N}_g(\boldsymbol{x}) d\boldsymbol{x} = 1$ より，全標本空間の確率 $P(\Omega)$ は，$P(\Omega) = 1$ となる．

以上の標本空間に対して，本文で考える観測事象は，空間インデックスの集合 X と，n 次元のベクトル $\boldsymbol{x}$ からなる確率変数 $\boldsymbol{o}$ によって表される．すなわち

$$\boldsymbol{o} = (X, \boldsymbol{x}) \tag{2}$$

ただし，X に含まれる空間インデックスが表す空間は，すべて同じ次元をもつとする．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．

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g(V(\boldsymbol{o})) \tag{3}$$

ただし，

$$S(\boldsymbol{o}) = X, \quad V(\boldsymbol{o}) = \boldsymbol{x} \tag{4}$$

0 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，$V(\boldsymbol{o}) = \boldsymbol{x}$ は存在しないことを注意しておく．つまり，式 (3) において，$n_g = 0$ のとき $\mathcal{N}_g(\cdot)$ は存在しないが，ここでは記述の簡便性のため，$\mathcal{N}_g(\cdot) = 1$，$n_g = 0$ と定義している．

この多空間上の確率分布を出力確率分布としてもつ新たな HMM λ を考える．λ の状態数を N としたとき，λ は初期状態確率 $\pi = \{\pi_i\}_{i=1}^{N}$，状態遷移確率 $A = \{a_{ij}\}_{i,j=1}^{N}$，出力確率 $B = \{b_i(\cdot)\}_{i=1}^{N}$ からなる．このような HMM を多空間出力分布 HMM（multi-space probability distribution HMM: MSD-HMM）と呼び，通常の離散 HMM や連続 HMM と同様に EM アルゴリズムに基づくモデルパラメータの最ゆう推定法を導くことができる [15].

2.2 MSD-HMM によるピッチパターンのモデル化

ピッチパターンは連続値をとる有声区間と値をもたない無声区間の時系列として表されるため，通常の連続 HMM や離散 HMM では直接モデル化することができない．そこで，ピッチパターンを有声区間に対応する 1 次元空間 Ω_1 と無声区間に対応する 0 次元空間 Ω_2 の二つの空間から出力される観測事象と考え（図 1），MSD-HMM によりモデル化する．有声の空

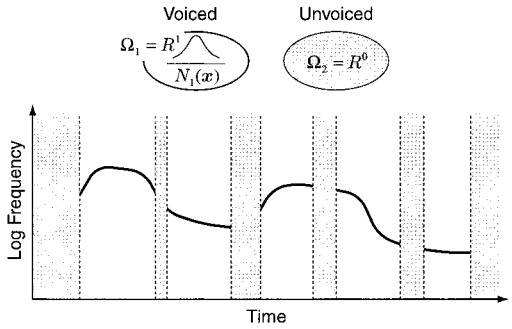


図 1　二つの空間によるピッチのモデル化
Fig. 1　Pitch pattern modeling on two spaces.

間は複数存在してもよいが，ここでは簡単のため有声の空間が一つであるとする．

有声/無声を表す空間のインデックスの集合を X，有声区間におけるピッチの値を $\boldsymbol{x}$，ピッチに関する観測事象を $\boldsymbol{o} = (X, \boldsymbol{x})$ とする．$X = \{1\}$ のときには有声区間を表し，$\boldsymbol{x}$ は 1 次元のピッチの値である．また，$X = \{2\}$ のときには無声区間を表し，$\boldsymbol{x}$ は 0 次元（$\boldsymbol{x}$ は値をもたない）となる．このとき，MSD-HMM の状態 i の出力確率分布 $b_i(\boldsymbol{o})$ は，

$$b_i(\boldsymbol{o}) = \begin{cases} w_{i1}\mathcal{N}_{i1}(V(\boldsymbol{o})), & S(\boldsymbol{o}) = \{1\} \\ w_{i2}, & S(\boldsymbol{o}) = \{2\} \end{cases} \qquad (5)$$

と書くことができる．ここで，w_{i1} は有声となる確率，w_{i2} は無声となる確率を表し，$\mathcal{N}_{i1}(V(\boldsymbol{o}))$ は 1 次元ガウス分布である．各状態の出力確率分布を式 (5) で定義することにより，HMM の枠組みでピッチパターンを直接モデル化することができる．

3. 決定木に基づくコンテクスト クラスタリング

音声のスペクトルやピッチは，アクセント型，構文情報，当該・先行・後続音素など，様々な要因の影響を受けて変動する．そのため，これらの変動要因を考慮し，各要因の組合せ（本論文ではすべての要因を含めてコンテクストと呼ぶ）に対してそれぞれ別々にモデルを構築できれば，精度の高いモデルが得られると期待できる．しかし，考慮する変動要因が増加するとその組合せが指数的に増加するため，モデル当りの学習データが著しく減少し，モデルパラメータの推定精度が低下する．また，可能なすべてのコンテクストを網羅する学習データを用意することは現実には不可能であるため，学習データ中に存在しないコンテクストのモデルが必要となった場合に，何らかの工夫をする必要がある．そこで，連続 HMM に対する決定木に基づくコンテクストクラスタリング手法 [4] を MSD-HMM に拡張し，状態のクラスタリングを行うことを考える．

決定木は 2 分木であり，それぞれの節ごとにコンテクストを二つに分割する質問が用意されている．すべてのコンテクストは根からそれぞれの節の質問に従って木を下って行き，葉のうちのどれかに達するため，いったん決定木を構築すれば，学習データに出現しないコンテクストについても対応するモデル（クラスタ）が一意に決定される．

決定木は以下のようにして構築される．

（1） すべてのコンテクストを一つにまとめ，根とする．

（2） 存在するすべての葉に対して，用意されているすべての質問を適用し，分割の前後でゆう度が最も大きく増加する葉と質問のセットを選ぶ．ゆう度変化の最大値がしきい値以下であれば終了する．

（3） 手順 2 で選ばれた葉を二つに分割し，新たな葉を二つ作る．古い葉は節となって手順 2 で選ばれた質問を保持し，新しく作られた葉に枝を延ばす．

（4） 手順 2 に戻る．

3.1 コンテクスト クラスタリングにおけるゆう度の近似計算

すべての状態の集合をクラスタセット $\mathcal{S} = \{s_1, s_2, \ldots\}$ に分割した場合の，学習データ $\boldsymbol{O}$ に対する対数ゆう度 $\log(P(\boldsymbol{O}|\lambda_{\mathcal{S}}))$ の近似値 $\mathcal{L}$ を求める．まず，$n_g = 0$ である空間のインデックスの集合を $\mathcal{G}_0$，$n_g > 0$ である空間のインデックスの集合を $\mathcal{G}_1$ と定義する．各状態 i または各クラスタ s の各空間 $g \in \mathcal{G}_1$ 上の出力分布は単一の対角共分散ガウス分布であるとし，その平均ベクトル及び共分散行列をそれぞれ $\boldsymbol{\mu}_{ig}$，$\boldsymbol{\Sigma}_{ig}$ または $\boldsymbol{\mu}_{sg}$，$\boldsymbol{\Sigma}_{sg}$ とする．また，クラスタ s における空間 g の重みを w_{sg} とし，時刻 t においてクラスタ s，空間 g をとる確率を $\gamma_t(s, g)$ とする．更に，観測事象 $\boldsymbol{o}_t$ の空間インデックス集合が空間インデックス g を含むような t の集合を $T(\boldsymbol{O}, g) = \{t | g \in S(\boldsymbol{o}_t)\}$ と定義する．

計算に先立ち，文献 [4] と同様，以下の仮定をおく．i）クラスタリングの途中においては，時刻 t において状態 i，空間 g をとる確率 $\gamma_t(i, g)$ の値は一定である．ii）状態遷移確率のゆう度への寄与は無視できる．状態遷移確率はゆう度に対して大きな影響を与える

が，その寄与は $\gamma_t(i,g)$ が変化したときのみ変化する．$\gamma_t(i,g)$ はクラスタリングの途中においては一定であると仮定されているので，状態遷移確率のゆう度への影響も一定であると仮定され，クラスタリングに影響を与えない．iii）学習データ $\boldsymbol{O}$ に対する対数ゆう度は各時刻 t における状態 i の出力確率 $b_i(\boldsymbol{o}_t)$ の対数を $\gamma_t(i,g)$ で重み付けした平均で近似される．

このとき，学習データ $\boldsymbol{O}$ に対する対数ゆう度の近似値 $\mathcal{L}$ は，$(\cdot)^T$ を行列の転置として，

$$\begin{aligned}\mathcal{L} &= \sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\sum_{g=1}^{G}\log(b_s(\boldsymbol{o}_t))\gamma_t(s,g)\\ &= \sum_{s\in\mathcal{S}}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)}\log(b_s(\boldsymbol{o}_t))\gamma_t(s,g)\\ &= \sum_{s\in\mathcal{S}}\Bigg\{\sum_{g\in\mathcal{G}_0}\sum_{t\in T(\boldsymbol{O},g)}\log(w_{sg})\gamma_t(s,g)\\ &\quad+\sum_{g\in\mathcal{G}_1}\sum_{t\in T(\boldsymbol{O},g)}-\frac{1}{2}(n_g\log(2\pi)+\log|\boldsymbol{\Sigma}_{sg}|\\ &\quad-2\log w_{sg}+(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg})^T\boldsymbol{\Sigma}_{sg}^{-1}\\ &\quad\cdot(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}))\gamma_t(s,g)\Bigg\} \qquad (6)\end{aligned}$$

で表される．ここで，共分散行列の再推定規則より

$$\boldsymbol{\Sigma}_{sg}=\frac{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)\cdot\left(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}\right)\left(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}\right)^T}{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)} \qquad (7)$$

が成り立っている．共分散行列は対角であると仮定しているため，

$$\sum_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)\left(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}\right)^T\boldsymbol{\Sigma}_{sg}^{-1}\cdot\left(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}\right)=n_g\sum_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g) \qquad (8)$$

が成り立ち，ゆう度の近似値 $\mathcal{L}$ は

$$\begin{aligned}\mathcal{L} &= \sum_{s\in\mathcal{S}}\Bigg\{\sum_{g\in\mathcal{G}_0}\log w_{sg}\sum_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)\\ &\quad+\sum_{g\in\mathcal{G}_1}-\frac{1}{2}(n_g(\log(2\pi)+1)+\log|\boldsymbol{\Sigma}_{sg}|\\ &\quad-2\log w_{sg})\sum_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)\Bigg\}\\ &= \sum_{s\in\mathcal{S}}\sum_{g=1}^{G}-\frac{1}{2}(n_g(\log(2\pi)+1)+\log|\boldsymbol{\Sigma}_{sg}|\\ &\quad-2\log w_{sg})\sum_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g) \qquad (9)\end{aligned}$$

で求めることができる．ただし，記述の簡便性のため，$g\in\mathcal{G}_0$，つまり $n_g=0$ のとき $\log|\boldsymbol{\Sigma}_{sg}|=0$ としている．

式 (9) における $\gamma_t(s,g)$，w_{sg} 及び $\boldsymbol{\Sigma}_{sg}$ は，

$$\gamma_t(s,g)=\sum_{c\in\mathcal{C}(s)}\gamma_t(c,g) \qquad (10)$$

$$\begin{aligned}w_{sg} &= \frac{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)}{\sum\limits_{h=1}^{G}\sum\limits_{t\in T(\boldsymbol{O},h)}\gamma_t(s,h)}\\ &= \frac{\sum\limits_{c\in\mathcal{C}(s)}\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(c,g)}{\sum\limits_{c\in\mathcal{C}(s)}\sum\limits_{h=1}^{G}\sum\limits_{t\in T(\boldsymbol{O},h)}\gamma_t(c,h)} \qquad (11)\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{sg} &= \frac{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)\cdot\left(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}\right)\left(V(\boldsymbol{o}_t)-\boldsymbol{\mu}_{sg}\right)^T}{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)}\\ &= \frac{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)V(\boldsymbol{o}_t)V(\boldsymbol{o}_t)^T}{\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)}\\ &\quad-\frac{\left(\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)V(\boldsymbol{o}_t)\right)\cdot\left(\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)V(\boldsymbol{o}_t)\right)^T}{\left(\sum\limits_{t\in T(\boldsymbol{O},g)}\gamma_t(s,g)\right)^2}\end{aligned}$$

$$= \frac{\sum_{c\in\mathcal{C}(s)} \left(\boldsymbol{\Sigma}_{cg} + \boldsymbol{\mu}_{cg}\boldsymbol{\mu}_{cg}^T\right) \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(c,g)}{\sum_{c\in\mathcal{C}(s)} \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(c,g)} - \frac{\left(\sum_{c\in\mathcal{C}(s)} \boldsymbol{\mu}_{cg} \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(c,g)\right) \cdot \left(\sum_{c\in\mathcal{C}(s)} \boldsymbol{\mu}_{cg} \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(c,g)\right)^T}{\left(\sum_{c\in\mathcal{C}(s)} \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(c,g)\right)^2} \tag{12}$$

で求められる．ここで $\mathcal{C}(s)$ は，クラスタ s に含まれる状態の集合である．これにより，学習データに対するクラスタセットのゆう度（式 (9)）はそれぞれの状態 c における平均 $\boldsymbol{\mu}_{cg}$，共分散 $\boldsymbol{\Sigma}_{cg}$，そして $\gamma_t(c,g)$ のみで計算される．

3.2 クラスタの分割によるゆう度変化

3.1 の仮定から一つの親クラスタを複数の子クラスタに分割する作業は他のクラスタには影響を与えないため，クラスタの分割によるゆう度の変化量を求めるには，着目している一つのクラスタを分割した際の局所的なゆう度変化のみを考慮すればよい．クラスタの分割はクラスタセット $\mathcal{S}$ の一つの親クラスタ p を子クラスタのセット $\mathcal{D}(p)$ に置き換えるので，分割後のゆう度は

$$\begin{aligned}\mathcal{L} = &\sum_{s\in\mathcal{S},s\neq p} \sum_{g=1}^{G} -\frac{1}{2}(n_g(\log(2\pi)+1) \\ &+ \log|\boldsymbol{\Sigma}_{sg}| - 2\log w_{sg}) \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(s,g) \\ &+ \sum_{d\in\mathcal{D}(p)} \sum_{g=1}^{G} -\frac{1}{2}(n_g(\log(2\pi)+1) \\ &+ \log|\boldsymbol{\Sigma}_{dg}| - 2\log w_{dg}) \sum_{t\in T(\boldsymbol{O},g)} \gamma_t(d,g)\end{aligned} \tag{13}$$

で求められる．対数ゆう度の変化 $\delta\mathcal{L}$ は，子クラスタのセット $\mathcal{D}(p)$ に関するゆう度の和と，親クラスタ p に関するゆう度との差分となり，

$$\begin{aligned}\delta\mathcal{L} = &\sum_{d\in\mathcal{D}(p)} \sum_{g=1}^{G} -\frac{1}{2}(\log|\boldsymbol{\Sigma}_{dg}| - 2\log w_{dg}) \\ &\sum_{t\in T(\boldsymbol{O},g)} \gamma_t(d,g) \\ &- \sum_{g=1}^{G} -\frac{1}{2}(\log|\boldsymbol{\Sigma}_{pg}| - 2\log w_{pg}) \\ &\sum_{t\in T(\boldsymbol{O},g)} \gamma_t(p,g)\end{aligned} \tag{14}$$

となる．

分割によってある子クラスタ d の $\gamma_t(d,g)$ がしきい値以下になった場合には，その分割における対数ゆう度の変化 $\delta\mathcal{L}$ を 0 とすることで学習データ量が過度に減少した子クラスタの生成を防ぐことができる．

4. HMM に基づく音声合成システムへの適用

MSD-HMM によるピッチパターンのモデル化性能を検討するため，HMM に基づく音声合成システム [7] に適用し，ピッチパターン及びスペクトル系列のモデル化，生成を行った．

4.1 システムの構築

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 文のうち 450 文を用いた．サンプリング周波数は 10 kHz，分析周期はスペクトル，ピッチともに 5 ms とした．25.6 ms 長ブラックマン窓を用いてメルケプストラム分析 [16] を行い，0〜15 次メルケプストラムを求めた．また，データベースに付属するピッチデータから対数基本周波数を求めた．更に，メルケプストラム及び対数基本周波数それぞれについて動的特徴量を求め，これらを特徴パラメータとした．

動的特徴量としては，時刻 t でのメルケプストラムを $\boldsymbol{c}_t$，対数基本周波数を p_t として，メルケプストラムについては式 (15)，(16) で表されるデルタ及びデルタデルタメルケプストラム $\Delta\boldsymbol{c}_t$，$\Delta^2\boldsymbol{c}_t$ を，また対数基本周波数については式 (17)，(18) で表される当該フレームとその前後それぞれ 3 フレームずつで計算した 1 次回帰係数 $\delta^l p_t$，$\delta^r p_t$ を用いた．

$$\Delta\boldsymbol{c}_t = \frac{1}{2}(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}) \tag{15}$$

$$\Delta^2\boldsymbol{c}_t = \frac{1}{4}(\boldsymbol{c}_{t+2} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-2}) \tag{16}$$

$$\delta^l p_t = \frac{1}{14}(-3p_{t-3} - 2p_{t-2} - p_{t-1} + 6p_t) \quad (17)$$

$$\delta^r p_t = \frac{1}{14}(3p_{t+3} + 2p_{t+2} + p_{t+1} - 6p_t) \quad (18)$$

ただし，対数基本周波数の回帰係数は，計算に必要なフレームがすべて有声の場合のみ計算し，無声区間及び有声/無声の境界付近で $\delta^l p_t$ または $\delta^r p_t$ を計算できないフレームについては，$\delta^l p_t$，$\delta^r p_t$ の一方または両方を無声として扱う．したがって，同じフレームであっても静的特徴量と動的特徴量がすべて有声または無声となるとは限らず，静的特徴量は有声であるが動的特徴量は無声であるフレームも存在する．

使用した HMM は 3 状態 left-to-right モデルであり，音素単位のコンテクスト依存モデルである．スペクトルとピッチをそれぞれ別々の HMM でモデル化した場合，スペクトルとピッチで音素境界を合わせるためには，何らかの工夫が必要となる．また，ピッチのみを特徴ベクトルとした場合では，有声区間，無声区間ともに音素に関する情報が不足するため，音素境界を適切に学習することができない．そこで，スペクトルパラメータとピッチパラメータを合わせて一つの観測事象とする．したがって，時刻 t での観測事象 $\boldsymbol{o}_t$ は，

$$\boldsymbol{o}_t = (\boldsymbol{o}_t^c, \boldsymbol{o}_t^p, \boldsymbol{o}_t^l, \boldsymbol{o}_t^r) \quad (19)$$

ただし，

$$\boldsymbol{o}_t^c = [\boldsymbol{c}_t^T, \Delta\boldsymbol{c}_t^T, \Delta^2\boldsymbol{c}_t^T]^T \quad (20)$$

$$\boldsymbol{o}_t^p = (X_t^p, \boldsymbol{x}_t^p) \quad (21)$$

$$\boldsymbol{o}_t^l = (X_t^l, \boldsymbol{x}_t^l) \quad (22)$$

$$\boldsymbol{o}_t^r = (X_t^r, \boldsymbol{x}_t^r) \quad (23)$$

となる（図 2）．ここで，X_t^p，X_t^l，X_t^r はそれぞれ対数基本周波数及びその回帰係数についての有声または無声を表す空間のインデックスを要素とする集合である．また，$\boldsymbol{x}_t^p$，$\boldsymbol{x}_t^l$，$\boldsymbol{x}_t^r$ はそれぞれ有声の場合は p_t，$\delta^l p_t$，$\delta^r p_t$ の値をとり，無声の場合は値をもたない．

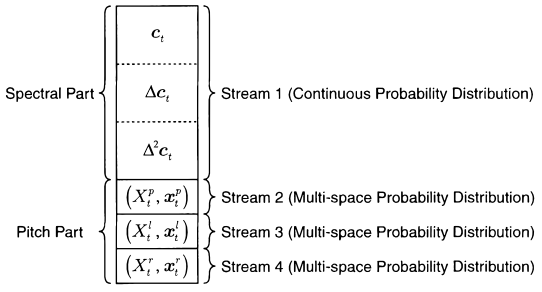


図 2　観 測 事 象
Fig. 2　Observation.

HMM の学習時には，観測事象 $\boldsymbol{o}_t$ を図 2 で表されるように四つのストリームに分け，メルケプストラムに関する部分は通常の単一対角共分散ガウス分布で，また対数基本周波数及びその回帰係数は，有声の空間の数を 1 として **2.** で述べた多空間確率分布でモデル化する．このように，スペクトルとピッチを合わせて一つの観測事象とすることにより，スペクトル系列とピッチパターンを統一的にモデル化することができ，更に無声が続く区間においてもスペクトルパラメータに基づいて適切に状態遷移が定まることが期待できる．

音韻及び韻律に関する変動要因は，文献 [11], [17] などを参考にして表 1 のように定めた．モーラ数，アクセント核位置，モーラ位置については，値をそのままカテゴリーとしており，可能なすべてのカテゴリーの数を数え上げることは困難であるため，表 1 には学習データに出現したカテゴリーの数を示した．学習データ内には，表 1 の要因の組合せにより，22,980 種類のラベルが現れた．スペクトルに影響を与えるコンテクストとピッチに影響を与えるコンテクストは大きく異なると考えられるため，特徴ベクトルをスペクトル部とピッチ部とに分け（図 2），更に HMM の状態位置別に，それぞれ別々にコンテクストクラスタリングを行った．

コンテクストクラスタリングで用いた質問は，以下のとおりである．

（1） {当該，先行，後続} 音素の種類．

（2） {当該，先行，後続} 音素が {母音，半母音，有声破裂音，無声破裂音，有声摩擦音，無声摩擦音 1（`/s/`，`/sh/`など），無声摩擦音 2（`/h/`，`/f/` など），破擦音，流音，鼻音，無音} に含まれる．

（3） アクセント核位置 n のアクセント句の 1 番目のモーラである $(0 \leq n)$．

表 1　変動要因とカテゴリー数
Table 1　Factors and number of categories.

当該音素	50 カテゴリー
前後音素	50 × 50 カテゴリー
前後アクセント句境界	4 × 4 カテゴリー
品詞	10 カテゴリー
当該アクセント句の	
モーラ数	13 カテゴリー
アクセント核位置	11 カテゴリー
当該モーラ位置	13 カテゴリー

（4） アクセント核位置 n のアクセント句の 2 番目から i 番目のモーラである $(2 \leqq n, 2 \leqq i \leqq n)$.

（5） アクセント核位置 n のアクセント句の $n+1$ 番目から j 番目のモーラである $(0 \leqq n, n+1 \leqq j)$.

（6） 当該アクセント句の直前が {0 型のアクセント句，0 型以外のアクセント句，文頭，ポーズ} かつ直後が {1 型のアクセント句，1 型以外のアクセント句，文末，ポーズ} である.

（7） 当該音素の所属するアクセント句に最初に出現する自立語の品詞の種類.

ここで，{...} は括弧内から一つを選択することを表す．なお，文のはじめと終わりの無音区間及び文中のポーズについては，通常の音素と同様 HMM でモデル化しているが，コンテクストは考慮せず，クラスタリングも行わなかった．

クラスタリングの終了条件であるクラスタの分割による対数ゆう度の変化量のしきい値はスペクトル部とピッチ部で同じ値とし，しきい値を 10，30，50 と変えることによって 3 種類のモデルセットを作成した．それぞれのモデルセットの総状態数を表 2 に示す．

ピッチ部の第 1 状態に対して実際に構築された決定木の例を図 3 に示す．図 3 中，“sil” は文のはじめと終わりの無音区間，“silence” は “sil”，文中のポーズ，及び無声破裂音の直前の促音のカテゴリーを表し，“L-*” 及び “R-*” は当該音素の左環境及び右環境を表している．また，“1to13_a0” は当該音素がアクセント型 0 型のアクセント句の 1 番目から 13 番目までのモーラに含まれることを表し，“low-tail” は当該音素の含まれるアクセント句が 0 型以外のアクセント型かつ文末であることを表している．

表 2 各モデルセットの総状態数
Table 2 Number of states of HMM sets.

モデルセット	総状態数	
	スペクトル	ピッチ
クラスタリング前	68,940	68,940
クラスタリング後 (1)	3,285	11,552
クラスタリング後 (2)	688	3,133
クラスタリング後 (3)	433	1,579

合成時には，まず合成したい任意のテキストをコンテク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学習した HMM を結合し，一つの文 HMM を構成する．この文 HMM から，文献 [6] に示されるゆう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メルケプストラム系列及びピッチパターンを生成する．ただし，ピッチパターンについては有声区間でのみパラメータ生成を行う．MLSA フィルタ [18], [19] を用いることにより，生成したパラメータ系列から直接音声を合成することができる．

4.2 ピッチパターン及びスペクトル系列の生成例

学習データに存在する文章に対するピッチパターンの生成例を図 4 に，また学習データに存在しない文章に対する生成例を図 5 に示す．図中，点線が実音声のピッチパターンであり，実線が生成されたピッチパターンである．パラメータ生成時の状態系列は，実音声との比較のため，実音声に HMM をビタビアラインメントすることにより得られたものを用いた．また，各フレームの有声/無声は，ここでは簡単にピッチの静的特徴の有声/無声の空間重みに従って決定した．つまり，ピッチの静的特徴に対する有声の空間重みが無声の空間重みより大きい状態を有声状態とし，有声状態の継続区間を有声区間，それ以外を無声区間とした．ここで，有声/無声の決定を音素単位ではなく状態単

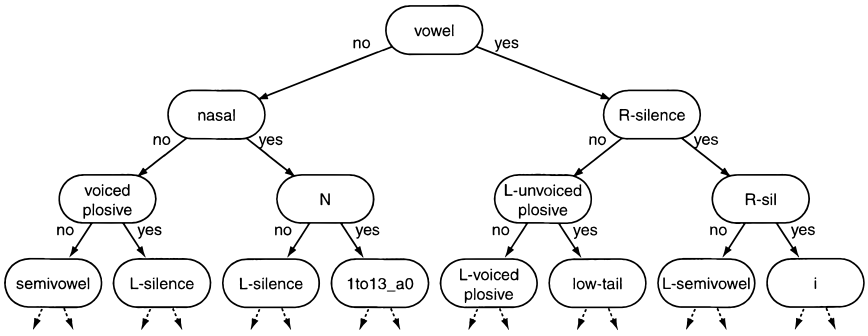
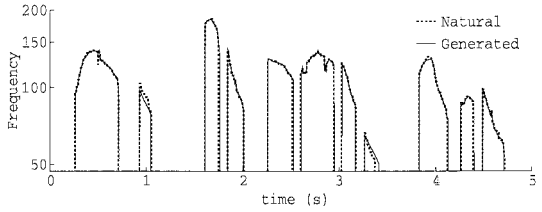
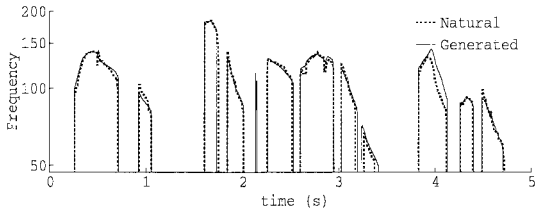


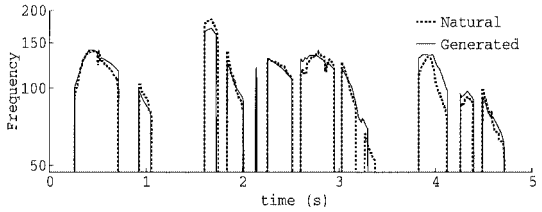
図 3 決定木の例（ピッチ部第 1 状態）
Fig. 3 An example of a decision tree.



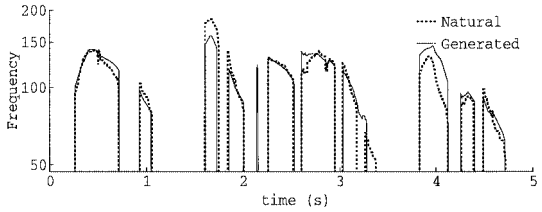
(a) クラスタリング前のモデル（68,940 分布）



(b) ピッチ部の総状態数 11,552 分布のモデル

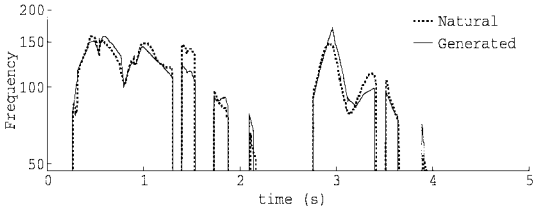


(c) ピッチ部の総状態数 3,133 分布のモデル

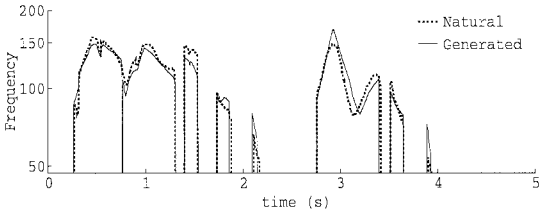


(d) ピッチ部の総状態数 1,579 分布のモデル
「おごりを捨て謙虚な姿勢を取り戻さねば冬は過ごせない」

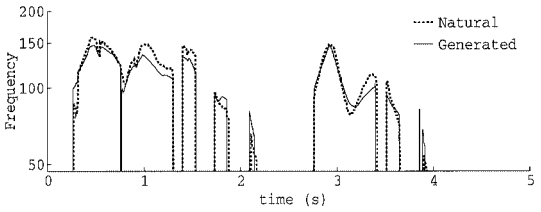
図 4　ピッチパターンの生成例（学習データに含まれる文章）
Fig. 4　Examples of generated pitch patterns for a training sentence.



(a) ピッチ部の総状態数 11,552 分布のモデル



(b) ピッチ部の総状態数 3,133 分布のモデル



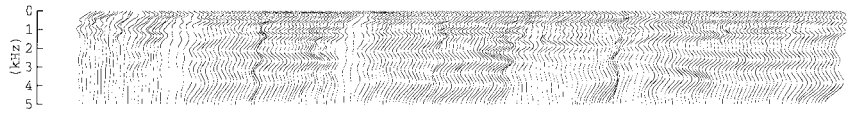
(c) ピッチ部の総状態数 1,579 分布のモデル
「だんだん自分が恐ろしくなって家に逃げ帰った」

図 5　ピッチパターンの生成例（学習データに含まれない文章）
Fig. 5　Examples of generated pitch patterns for a test sentence.

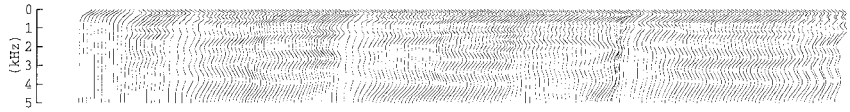
位で行うのは，/ky/，/py/などの「子音＋拗音」も一つの音素として扱っており音素内で無声音から有声音に遷移する場合があるため，また，連結学習によって有声音（無声音）のモデルが有声区間（無声区間）のみではなく隣接している無声区間（有声区間）も学習している可能性があるためである．

図 4 (a) はクラスタリング前のモデルから生成したピッチパターンである．クラスタリング前では学習データが一つしかないモデルが多数存在し，そのようなモデルはただ一つのパターンをモデル化しているために，実音声のピッチパターンと生成されたピッチパターンがほぼ一致することがわかる．また，図 4 (b), (c), (d) より，コンテクストクラスタリングを行った場合にも，学習データのピッチ包絡をほぼ再現していることがわかる．更に，学習データに含まれない文章のピッチパターンを生成した場合（図 5）にも，文章中に現れる 40 種類のラベルのうち 34 種類が学習データに現れないラベルであったにもかかわらず，実音声のピッチパターンをよく近似したピッチパターンが得られている．

ピッチ部の総状態数 11,552 のモデルについて，学習内データ 450 文章及び学習外データ 53 文章に対して求めた実音声との対数基本周波数の rms 誤差は，それぞれ 0.049（0.071 [oct]）及び 0.112（0.162 [oct]）となった．提案手法では，学習データに対するゆう度を



(a) 実音声



(b) 生成スペクトル（スペクトル部の総状態数 3,285）

図 6　学習データに含まれない文章のスペクトルの生成例（「だんだん自分が ···」）
Fig. 6　Examples of generated spectra for a test sentence.

基準としてモデル化しているため，誤差最小という意味では必ずしも最適なモデルになるとは限らない．しかし，図 4，図 5 からもわかるとおり基本周波数の全体的なずれによる部分も多く，聴感上では rms 誤差ほどの不自然さは感じられなかった．

また，図 6 に学習データに含まれない文章について生成したスペクトルの例を示す．図 6 より，スペクトルに関しても，実音声をよく近似したスペクトル系列が得られていることがわかる．スペクトルのモデル化性能のみについて考えると，スペクトルとピッチを同時にモデル化する場合よりもスペクトルのみをモデル化する場合の方が性能が良いと考えられる．しかし，合成音声に対する非公式な受聴により，両者に品質上の差がほとんどないことを確認しており，ピッチを加えたことによるモデル化性能の劣化はわずかであると考えられる．一方，スペクトルとピッチを同時にモデル化することにより，無声が続く区間でも適切に状態遷移が定まること，また合成時にスペクトルとピッチの同期が自動的にとれることから，スペクトルとピッチを同時にモデル化する本手法は利点が多いと考えられる．

5. むすび

本論文では，MSD-HMM におけるコンテクストクラスタリング手法を導出し，ピッチパターン及びスペクトル系列をゆう度最大化基準に基づいて統一的にモデル化及び生成する手法を提案した．音素や構文，韻律情報などのピッチ及びスペクトルの変動要因の組合せに対してコンテクストクラスタリングを行うことにより，学習データに出現しないコンテクストに対しても，自然発声を良く近似したパラメータ系列を生成できることを示した．

今回の実験では，有声/無声を簡単にピッチの静的特徴から定めたため，有声/無声の誤りがいくつかのフレームで起きていた．そこで，ゆう度最大化基準に基づく有声/無声の決定方法，動的特徴を考慮した有声/無声の決定方法，各状態がもつ二つの空間をサブ状態と考えて展開することにより有声/無声を状態系列として制御する手法 [20] などについて検討を行う必要がある．また，今回の実験ではクラスタ分割によるゆう度の変化量をしきい値として用いたが，最適なしきい値を決定することは難しく，またスペクトル部とピッチ部で同じしきい値を用いることが適当であるとは限らないため，MDL 基準に基づく状態数の決定法 [21] などの検討も必要となる．更に，用いるコンテクストの種類や質問の検討，精度の高い状態継続長のモデル化手法の検討も今後の課題となる．

謝辞　研究を進めるあたり，実験に御協力を頂いた東京工業大学大学院修士課程加藤寿彦氏（現ヤマハ（株））に感謝します．本研究の一部は文部省科学研究費補助金（基盤研究 B（2）課題番号 10555125），（財）中部電力基礎技術研究所研究助成金によった．

文　献

[1] S. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, “Context-dependent modeling for acoustic-phonetic recognition of continuous speech,” Proc. ICASSP85, pp.1205–1208, March 1985.

[2] S. Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-34, no.1, pp.52–59, Feb. 1986.

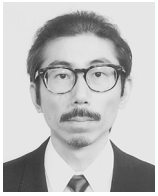
[3] B.-H. Juang, “Maximum-likelihood estimation for mixture multivariate stachastic observations of Markov chains,” AT&T Tech. J., vol.64, no.6,

pp.1235–1249, July 1985.
[4] J.J. Odell, The use of context in large vocabulary speech recognition, Ph.D. Dissertation, Cambridge University, March 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 Process., vol.39, no.4, pp.806–814, April 1991.
[6] 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,” 音響誌，vol.53, no.3, pp.192–200, March 1997.
[7] 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-II），vol.J79-D-II, no.12, pp.2184–2190, Dec. 1996.
[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system,” Proc. ICASSP97, pp.1611–1614, April 1997.
[9] 田村正統，益子貴史，徳田恵一，小林隆夫，“HMM 音声合成に基づく声質変換における話者適応手法の検討,” 音講論集，vol.1, no.2-P-13, pp.319–320, March 1998.
[10] A. Ljolje and F. Fallside, “Synthesis of natural sounding pitch contous in isolated utterances using hidden Markov models,” IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-34, no.5, pp.1074–1080, Oct. 1986.
[11] 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, Sept. 1995.
[12] G.J. Freij and F. Fallside, “Lexical stress recognition using hidden Markov models,” Proc. ICASSP88, pp.135–138, April 1988.
[13] U. Jensen, R.K. Moore, P. Dalsgaard, and B. Lindberg, “Modelling intonation contours at the phrase level using continuous density hidden Markov models,” Computer Speech and Language, vol.8, no.3, pp.247–260, July 1994.
[14] 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, Sept. 1994.
[15] 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, April 1998.
[16] 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.
[17] 阿部匡伸，佐藤大和，“音節を制御単位とする基本周波数の 2 階層構造,” 信学技報，SP92-5, May 1992.
[18] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” Proc. ICASSP92, pp.137–140, March 1992.
[19] 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.
[20] 加藤寿彦，益子貴史，小林隆夫，徳田恵一，“多空間出力分布並列 HMM によるピッチパタン生成,” 音講論集，vol.1, no.1-7-19, pp.217–218, March 1998.
[21] 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，SP99-59, Aug. 1999.

（平成 11 年 8 月 19 日受付，12 月 27 日再受付）

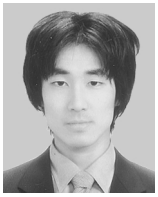
益子　貴史　（正員）

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）．同年東工大精密工学研究所助手．現在東工大大学院総合理工学研究科物理情報システム創造専攻助手．音声分析・合成・認識，マルチモーダルインタフェースの研究に従事．日本音響学会，IEEE，ISCA 各会員．



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・合成・符号化・認識，ディジタル信号処理，マルチモーダルインタフェースの研究に従事．日本音響学会，人工知能学会，情報処理学会，IEEE 各会員．



宮崎　　昇

平 7 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）．現在，NTT コミュニケーション科学基礎研究所勤務．在学中，音声合成の研究に従事．日本音響学会会員．

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．同助教授を経て現在東工大大学院総合理工学研究科物理情報システム創造専攻教授．工博．ディジタルフィルタ，音声分析・合成・符号化・認識，マルチモーダルインタフェースの研究に従事．日本音響学会，IEEE，ISCA 各会員．