論　文

決定木に基づく音素コンテクスト・次元・状態位置の同時クラスタリングによる音響モデリング

全　炳河[†a)]　　徳田　恵一[†b)]　　北村　正[†c)]

Decision Tree Based Simultaneous Clustering of Phonetic Contexts, Dimensions, and State Positions for Acoustic Modeling

Heiga ZEN[†a)], Keiichi TOKUDA[†b)], and Tadashi KITAMURA[†c)]

あらまし　近年，連続音声認識システムにおける音響モデルとして，前後の音素環境を考慮した音素コンテクスト依存隠れマルコフモデルが広く利用されている．音素コンテクスト依存隠れマルコフモデルを利用する場合，総モデル数が増加し，システムが非常に多くの自由パラメータを含むことになるため，統計的に信頼できるパラメータを推定することが困難になる．このため，様々なパラメータ共有手法が提案されており，中でも音素決定木に基づく状態共有法は，優れた解決法の一つである．しかし，状態単位の共有構造では特徴ベクトルの全次元に同一の共有構造を構築するため，各特徴量に対し，異なる共有構造を構築できない，適切なパラメータ数を割り当てることができない，といった問題点がある．本論文では，記述長最小化基準に基づく次元分割法を導入することにより音素決定木を拡張した，音素・次元決定木を提案する．更に，状態位置に関する分割条件を加え，音素コンテクスト・次元・状態位置を決定木に基づき同時にクラスタリングする手法を提案する．不特定話者連続音声認識実験の結果，提案法は従来の音素決定木に基づく状態共有法と比較して 13〜15%誤り率を削減することが示された．

キーワード　隠れマルコフモデル，音素決定木，コンテクストクラスタリング，記述長最小化基準，次元分割

1. まえがき

近年，計算機の高速化，大規模音声コーパスの整備などに伴い，大語い連続音声認識に関する研究が盛んに行われている．音声認識における音響モデルとしては，統計モデルの一種である隠れマルコフモデル (Hidden Markov Model; HMM) が広く用いられている．音声波形は様々な分析手法を用いて特徴ベクトル系列に変換され，これを学習データとして Baum-Welch アルゴリズムなどの学習アルゴリズムを用いて HMM のパラメータが学習される．HMM を用いた音声認識システムは，十分な学習データ量が与えられれば，高い認識性能を示すことが知られている．

数字単語認識のように語いが限られた認識タスクでは，各単語に対してモデルを用意することが可能である．しかし，大語い連続音声認識のように語い数が数千から数十万に及ぶタスクでは，各単語のモデルを用意することは学習データ量の問題から困難である．このため，大語い連続音声認識においては，音響モデルとして音素や音節などのサブワード単位をモデル化したサブワード HMM を用意し，認識時にはそれらを連結することにより単語を表現する手法が一般に用いられる．この方法では，各サブワード HMM 当りの学習データ量を十分に確保できるため，統計的に信頼性の高い HMM を学習することができる．

一般に音素の音響的特徴は前後の音素環境により大きく変化することが知られている．音素環境によりサブワード単位を区別してモデル化を行う音素コンテクスト依存 HMM は，音響的特徴をより精密にモデル化できると考えられ，多くの音声認識システムにおいて利用されている．

しかし，前後一つの音素環境を考慮してモデル化を

† 名古屋工業大学大学院工学研究科情報工学専攻，名古屋市
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan
a) E-mail: zen@ics.nitech.ac.jp
b) E-mail: tokuda@ics.nitech.ac.jp
c) E-mail: kitamura@nitech.ac.jp

行う triphone においても，その総モデル数は数千から数万に及び，学習データにすべての triphone が出現しない，各モデル当りの学習データ量が不足しパラメータ推定精度が低下する，といった問題が生じる．このため，様々なモデルパラメータ共有手法が提案されており [1]〜[4]，中でも音素決定木 (Phonetic Decision Trees; P-DT) による HMM 状態共有法 [4] は，この問題の優れた解決法の一つである（図 1）．本手法は，前後の音素環境を分割条件としてコンテクスト依存モデルの HMM 状態（出力分布は対角共分散単一ガウス分布）の集合に対してトップダウンにクラスタリングを行い，クラスタリング終了時の決定木の葉ノードに含まれる状態を共有することにより，HMM 状態共有構造を構築する手法である．また，学習データ中に存在しないモデルを決定木をたどり対応する葉ノードに割り当てることで生成することができる．

本手法に関して，分割するノード及びコンテクスト分割条件を選択するための基準 [5]〜[7]，音素環境以外の変動要因の考慮 [8]〜[12]，単一ガウス分布から混合ガウス分布への拡張 [13], [14] など，様々な検討がこれまでになされている．

しかしながら，これらの手法はすべて，HMM 状態単位の共有構造を構築している．音声認識に広く用いられている特徴量である MFCC などは，メル周波数軸上での対数スペクトルのフーリエ係数と考えることができる．低次の MFCC は対数スペクトル包絡の概形を表現しており，高次と比較してより多くの音韻情報を含んでいると考えられる．また，音声認識においては，MFCC など各時刻における静的な特徴だけでなく，その時間変化に対応する動的特徴の有効性が確認されている [15]．ところが，対数スペクトル包絡の概形を表現している静的特徴は，その時間変化である動的特徴と比較して，より多くの音韻情報を含んでいると考えられるにもかかわらず，従来の状態共有法では，すべての次元に同じ共有構造を構築するため，各特徴量に対してパラメータ数を変化させることができない．また，各特徴量はそれぞれ異なるコンテクスト依存性をもつと考えられるが，状態単位の共有構造ではこれを表現することができない．

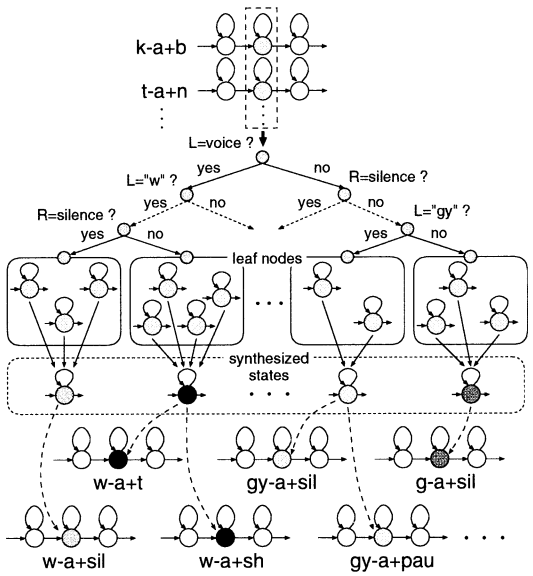

図 1　音素決定木 (P-DT) による HMM 状態共有法の概要
Fig. 1　An overview of the P-DT based state tying technique.

本論文では，この問題を解決する一手法として，記述長最小化 (Minimum Description Length; MDL) 基準に基づく次元分割法を導入することにより音素決定木を拡張し，音素コンテクスト及び次元を分割条件とする決定木によるクラスタリング法を提案する．更に，HMM の状態位置に関する分割条件を導入し，音素コンテクスト・次元・状態位置を同時にクラスタリングする手法を提案する．本論文で提案する手法は，決定木によるコンテクスト依存モデル構築のための，MDL 基準に基づく統一的な枠組みである．

次章で音素コンテクストと次元を分割条件とするクラスタリング法について述べ，**3.** で更に状態位置に関する分割条件を導入することにより，音素コンテクスト・次元・状態位置を決定木に基づき同時にクラスタリングする手法について述べる．**4.** でそれらの評価実験を示す．

2. 音素・次元決定木に基づくクラスタリング法

2.1 特徴量依存音素決定木

特徴ベクトルの各次元に対して異なるパラメータ共有構造を構築する手法はいくつか提案されているが [16], [17]，特にコンテクストクラスタリングと組み合わせた手法として，特徴量依存逐次状態分割法 (Feature Dependent Successive State Splitting; FD-SSS) [17] が提案されている．FD-SSS は，ゆう度最大化基準に基づく逐次状態分割法 (Maximum Likelihood Successive State Splitting; ML-SSS) [18] を拡張した手法で，特徴ベクトルの各次元に対して個別に ML-SSS を適用し，パラメータ共有構造を構築す

る．各次元に対して個別に共有構造を構築する手法は，決定木に基づく手法に対しても適用可能である．本論文では，各次元に個別に音素決定木を構築し，パラメータ共有構造を構築する手法を，特徴量依存音素決定木 (Feature Dependent Phonetic Decision Trees; FDP-DT) と呼ぶことにする（図 2）．FDP-DT では各当該音素の各状態位置の各次元に音素決定木を構築し，1 次元の分布を共有する．

音声認識における状態出力確率分布には，対角共分散混合ガウス分布が利用されることが多い．これは，コンテクスト以外の要因（性別，話者性，話速など）による音響的特徴の変動をモデル化する場合，単一ガウス分布では自由度が不足するためである．通常，音素決定木に基づく状態共有法では，各葉ノードに対応するクラスタの出力確率分布を混合ガウス分布化する．一般に，MFCC などの音声特徴量の各係数間には，相関が存在する．葉ノードに対応するクラスタの出力確率分布が対角共分散単一ガウス分布の場合，図 3 に示すように P-DT，FDP-DT とも次元間の相関を表現できない．しかし，P-DT では出力確率分布を混合ガウス分布化することにより，非対角要素を有する共分散行列を近似できる．これに対して FDP-DT の各状態の出力確率分布は，各次元における音素決定木の葉ノードに対応するクラスタの 1 次元分布からなり，単純にクラスタの出力確率分布を混合ガウス分布化するだけでは特徴量間の相関を近似することができない(注1)．

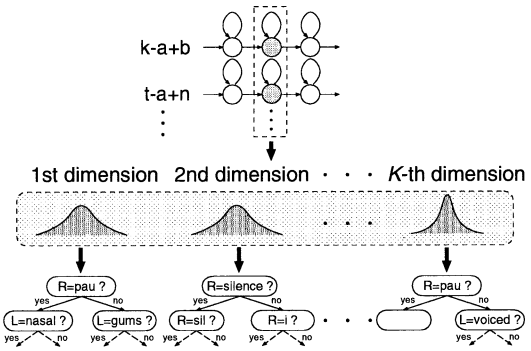

図 2　特徴量依存音素決定木 (FDP-DT) によるクラスタリング法の概要

Fig. 2　An overview of the FDP-DT based clustering technique.

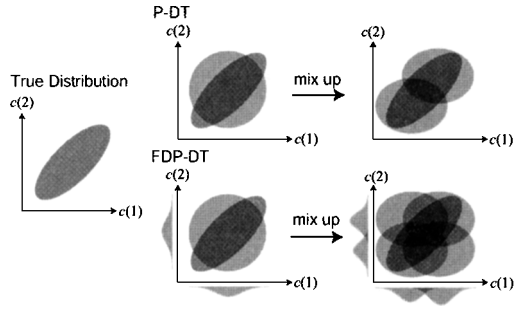

図 3　FDP-DT によるクラスタリング法の問題点

Fig. 3　Problem in FDP-DT based clustering technique.

2.2　音素・次元決定木

前章で述べた問題点を解決するため，MDL 基準に基づく次元分割法を導入し P-DT を拡張し，各次元に対して適切なコンテクスト依存共有構造を構築しつつ，次元のまとまりを保ち混合ガウス分布化に適したパラメータ共有構造を構築する手法を提案する．

MDL 基準に基づく音素決定木によるクラスタリング法 [5] において，決定木中のノード S をコンテクストに関する分割条件（以後質問とする）q を用いて子ノード S_{q+} 及び S_{q-} に分割するとき，分割前後の記述長 (Description Length; DL) の変化量 $\Delta^{(q)}_{\mathrm{DL}}(S)$ は，以下のように計算される(注2)．

$$\begin{aligned}\Delta^{(q)}_{\mathrm{DL}}(S) = \frac{1}{2}\Big\{ &\Gamma(S_{q+})\log|\boldsymbol{\Sigma}(S_{q+})| \\ &+ \Gamma(S_{q-})\log|\boldsymbol{\Sigma}(S_{q-})| \\ &- \Gamma(S)\log|\boldsymbol{\Sigma}(S)|\Big\} + K\log\Gamma(S_0)\,. \end{aligned} \tag{1}$$

ただし，K は特徴ベクトルの次元数，$\boldsymbol{\Sigma}(\cdot)$ は各ノードに対応するクラスタの出力確率分布（単一ガウス分布）の共分散行列，S_0 は決定木の根ノードを表す．また，$\Gamma(\cdot)$ は学習データ中に各ノードに対応するクラスタが出現する頻度（学習データ数）であり，

$$\gamma_t(S) = \frac{\alpha_t(S)\beta_t(S)}{\sum_S \alpha_t(S)\beta_t(S)} \tag{2}$$

$$\Gamma(S) = \sum_{t=1}^{T} \gamma_t(S) \tag{3}$$

(注1)：文献 [19] において，各次元に対して個別に共有構造を決定した後，モデル全体を並列化することにより特徴量間の相関を近似する手法が提案されているが，この手法では学習データ中に出現しないモデルの混合重みを推定することが困難という問題がある．このことから，特徴ベクトルの各次元に対して異なる共有構造を構築可能であり，かつ次元のまとまりを保つことが可能な，パラメータ共有構造構築法が必要であることが理解される．

(注2)：文献 [5] において，モデルの複雑さの変化量 $K\log\Gamma(S_0)$ を c 倍し，モデルの大きさを様々に変化させた場合の認識性能の評価が行われており，$c = 2.0$ のとき最も良い認識性能が得られたことが報告されている．本論文では予備実験の結果，$c = 1.0$ のとき最も良い認識性能が得られたため，モデルの複雑さの変化量を調節する項は導入しなかった．

のように計算される．上式において，$\alpha_t(S)$ は時刻 t，ノード S に対応するクラスタの出力確率分布における前向き確率，$\beta_t(S)$ は時刻 t，ノード S に対応するクラスタの出力確率分布における後ろ向き確率であり，$\gamma_t(S)$ は，時刻 t においてノード S に対応するクラスタの出力確率分布に滞在する確率を表す．

ここで，式 (1) は，各ノードに対応するクラスタの出力確率分布が対角共分散単一ガウス分布であることを仮定することにより導出されている [4], [5]．よって式 (1) は，以下のように変形できる．

$$\Delta^{(q)}_{\mathrm{DL}}(S) = \sum_{k=1}^{K} \Delta^{(q)}_{\mathrm{DL}}(S,k)\ , \tag{4}$$

$$\begin{aligned}\Delta^{(q)}_{\mathrm{DL}}(S,k) = \frac{1}{2}\Big\{ &\Gamma(S_{q+})\log\sigma^2(S_{q+},k) \\ &+\Gamma(S_{q-})\log\sigma^2(S_{q-},k) \\ &-\Gamma(S)\log\sigma^2(S,k)\Big\} + \log\Gamma(S_0)\ .\end{aligned} \tag{5}$$

ただし，$\sigma^2(S_{q+},k)$，$\sigma^2(S_{q-},k)$，$\sigma^2(S,k)$ はそれぞれ，共分散行列 $\mathbf{\Sigma}(S_{q+})$，$\mathbf{\Sigma}(S_{q-})$，$\mathbf{\Sigma}(S)$ の $[k,k]$ 要素である．式 (5) より，分割前後の DL の変化量 $\Delta^{(q)}_{\mathrm{DL}}(S)$ は，各次元の DL の変化量 $\Delta^{(q)}_{\mathrm{DL}}(S,k)$ の総和となることが分かる．そこで本論文では，ノード S を $\Delta^{(q)}_{\mathrm{DL}}(S,k) < 0$ である次元からなる分布をもつノード S_1 と $\Delta^{(q)}_{\mathrm{DL}}(S,k) \geq 0$ である次元からなる分布をもつノード S_2 に次元分割する．その後，ノード S_1 にのみ質問 q を適用してコンテクストの分割を行う．本手法を「MDL 基準に基づく次元分割法」と呼ぶ．MDL 基準に基づく次元分割法の概要を図 4 に示す．提案法における次元の分割及び分割の決定法は以下のように表現することができる．

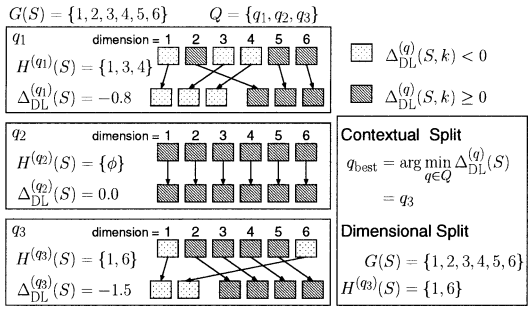

図 4　MDL 基準に基づく次元分割法の概要
Fig. 4　An overview of the MDL-based dimensional-split technique.

$$H^{(q)}(S) = \left\{ k \;\middle|\; k \in G(S),\ \Delta^{(q)}_{\mathrm{DL}}(S,k) < 0 \right\}\ , \tag{6}$$

$$\Delta^{(q)}_{\mathrm{DL}}(S) = \sum_{k\in H^{(q)}(S)} \Delta^{(q)}_{\mathrm{DL}}(S,k)\ , \tag{7}$$

$$q_{\mathrm{best}} = \arg\min_{q\in Q} \Delta^{(q)}_{\mathrm{DL}}(S)\ . \tag{8}$$

ただし，$G(S)$ はノード S に存在する次元の集合，$H^{(q)}(S)$ は質問 q によりコンテクスト分割が行われる次元の集合であり，$G(S) \in H^{(q)}(S)$ である．本手法により得られる次元分割を伴う木構造を音素・次元決定木 (Phonetic and Dimensional Decision Trees; PD-DT) と呼ぶ．PD-DT によるクラスタリング法の概要を図 5 に示す．また，提案手法の手順を以下にまとめる．

（1）　共有構造をもたないコンテクスト依存 HMM を学習する．

（2）　中心音素が同じ triphone の状態を集め根ノードとする．

（3）　決定木のすべてのノードに対して式 (6)〜(8) を用いて $\Delta^{(q_{\mathrm{best}})}_{\mathrm{DL}}(S)$ 及び $H^{(q_{\mathrm{best}})}(S)$ を計算する．

（4）　全ノード中最も $\Delta^{(q_{\mathrm{best}})}_{\mathrm{DL}}(S)$ が小さいノード S' を選択する．

（5）　ノード S' をノード S_1 及びノード S_2 に $H^{(q_{\mathrm{best}})}(S')$ に従い次元分割する．ただし，ノード S_1 の分布は集合 $H^{(q_{\mathrm{best}})}(S')$ に含まれる次元から構成され，ノード S_2 の分布は $G(S) - H^{(q_{best})}(S')$ に含まれる次元から構成される．

（6）　ノード S_1 を質問 q_{best} を用いてノード $S_{q_{\mathrm{best}}+}$ 及びノード $S_{q_{\mathrm{best}}-}$ にコンテクスト分割する．

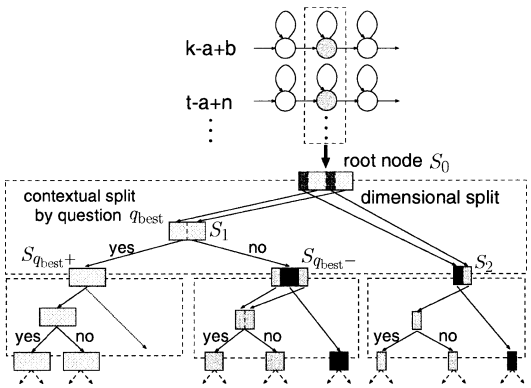

図 5　音素・次元決定木によるクラスタリング法の概要
Fig. 5　An overview of the PD-DT based clustering technique.

（7） すべてのノードに対して $\Delta^{(q)}_{\mathrm{DL}}(S) \geq 0$ であれば，クラスタリングを終了する．さもなくば，(3) に戻る．

2.3 PD-DT と他の手法との関係

PD-DT は，P-DT 及び FDP-DT をその特殊形として含む，より一般化された手法である．制約条件 $H^{(q)}(S) = G(S)$ 下では，次元分割が生じないため，P-DT と等しい．また，制約条件

$$H^{(q)}(S) = \left\{ k \,\middle|\, \arg\min_{k \in G(S)} \Delta^{(q)}_{\mathrm{DL}}(S,k), \Delta^{(q)}_{\mathrm{DL}}(S,k) < 0 \right\}$$

下では，各次元に対し個別に音素決定木を構築をする場合と等価であり，つまり，FDP-DT と等しい．

なお，4 階層共有構造 HMM [16] では，まずコンテクスト依存の状態共有構造を構築し，各状態の出力確率分布を混合ガウス分布化した後，各特徴量ごとの共有構造をボトムアップに構築する．このため，特徴量間の相関を考慮した上で共有構造を決定できる．しかし，コンテクスト依存の共有構造は全特徴量で共通であるため，特徴量ごとに異なるコンテクスト依存の共有構造を構築することはできない．これに対して PD-DT では，決定木構築時において状態出力確率分布が対角共分散単一ガウス分布である必要があり，最終的に混合ガウス分布化した際に際的な分割が得られるわけではないものの，決定木を構築する過程で分布を次元分割することで，より適切なコンテクスト依存のパラメータ共有構造を構築できると考えられる．

3. 音素・次元・状態位置決定木

2.2 において，MDL 基準に基づく次元分割法を導入することにより P-DT を拡張し，音素コンテクストと次元を分割条件とする決定木に基づくクラスタリング法を提案した．分割条件としては，このほかに HMM の状態位置が考えられる．これまでも，状態位置に関する分割条件を P-DT に追加した研究はいくつかあるが [14], [20]，決定木の根ノード付近で状態位置に関する分割が生じ，状態位置ごとに決定木を構築する場合と，ほとんど同じ結果となることが報告されている．しかし，状態位置への依存性は音素コンテクストと同様に，各次元ごとに異なると考えられる．そこで本章では，音素コンテクストに関する質問に加え，状態位置に関する質問を導入し，決定木に基づき音素コンテクスト・次元・状態位置を同時にクラスタリングする手法を提案する．

3.1 状態位置に関する分割条件

本論文では，隣り合う状態位置の分布が共有され得ると考え，これを状態位置に関する質問として導入する．各 HMM が N 個の状態からなる場合，以下に示すような質問を追加する．

- 第 1 状態である．
- 第 1, 2 状態である．
  ：
- 第 $1, 2, \ldots, N$ 状態である．

本論文では，P-DT に状態位置に関する質問を導入した手法を，音素・状態位置決定木 (Phonetic and State positional Decision Trees; PS-DT) と呼ぶことにする．同様に，PD-DT に状態位置に関する質問を導入した手法を，音素・次元・状態位置決定木 (Phonetic, Dimensional and State positional Decision Trees; PDS-DT) と呼ぶことにする．

3.2 MDL 基準に基づくクラスタリングへの導入

P-DT 及び PS-DT において，DL の変化量（式 (1)）は以下のように変形できる．

$$\Delta^{(q)}_{\mathrm{DL}}(S) = \Delta^{(q)}_{\mathrm{ML}}(S) + K \log \Gamma(S_0) \,, \tag{9}$$

$$\begin{aligned} \Delta^{(q)}_{\mathrm{ML}}(S) = \frac{1}{2} \Big\{ & \Gamma(S_{q+}) \log |\Sigma(S_{q+})| \\ & + \Gamma(S_{q-}) \log |\Sigma(S_{q-})| \\ & - \Gamma(S) \log |\Sigma(S)| \Big\} \,. \end{aligned} \tag{10}$$

ただし，$\Delta^{(q)}_{\mathrm{ML}}(S)$ は分割前後のゆう度の変化量，$K \log \Gamma(S_0)$ は分割前後のモデルの複雑さの変化量を表す項である．同様に，PD-DT 及び PDS-DT において，DL の変化量（式 (5)）は以下のように変形できる．

$$\Delta^{(q)}_{\mathrm{DL}}(S,k) = \Delta^{(q)}_{\mathrm{ML}}(S,k) + \log \Gamma(S_0) \,, \tag{11}$$

$$\begin{aligned} \Delta^{(q)}_{\mathrm{ML}}(S,k) = \frac{1}{2} \Big\{ & \Gamma(S_{q+}) \log \sigma^2(S_{q+},k) \\ & + \Gamma(S_{q-}) \log \sigma^2(S_{q-},k) \\ & - \Gamma(S) \log \sigma^2(S,k) \Big\} \,. \end{aligned} \tag{12}$$

ただし，$\Delta^{(q)}_{\mathrm{ML}}(S,k)$ は k 次元における分割前後のゆう度の変化量，$\log \Gamma(S_0)$ は分割前後のモデルの複雑さの変化量を表す項である．式 (9)〜(12) より，MDL 基準に基づくクラスタリング法は，ゆう度最大化基準に基づくクラスタリング法において，ゆう度変化量のしき

い値が分割前後のモデルの複雑さの変化量（P-DT 及び PS-DT では $K \log \Gamma(S_0)$，PD-DT 及び PDS-DT では $\log \Gamma(S_0)$）に設定されている場合とみなすことができる．

状態位置に関する質問が導入されている場合，根ノードのデータ数 $\Gamma(S_0)$ は，すべての状態位置のデータ数の総和になり，状態位置に関する質問が導入されていない場合と比較して，根ノードのデータ数は増加する．その結果，分割前後のモデルの複雑さの変化量の項は増加するため（$\log \Gamma(S_0)$ が増加するため），各状態位置ごとに決定木を構築した場合と比較して，葉ノード数は減少すると考えられる．そこで本論文では，2 種類の $\Gamma(S_0)$ の計算法について検討を行った．

case 1　$\Gamma(S_0) = \sum_{n=1}^{N} \Gamma(S_0^n)$

case 2　$\Gamma(S_0) = \sum_{n \in S} \Gamma(S_0^n)$

ただし，$\Gamma(S_0^n)$ は n 番目の状態位置のデータ数を示す．case 1 では，根ノード近傍で状態位置に関する質問が適用され，各状態位置ごとに決定木を構築することと等価になった場合，分割前後のモデルの複雑さの変化量が増加してしまうため，状態位置ごとに決定木を構築した場合と比較して葉ノード数は減少する．しかし case 2 では，各状態位置ごとに決定木を構築することと等価になった場合，状態位置ごとに決定木を構築した場合と等価な決定木が得られる．本論文では，case 1 及び case 2 を用いた PS-DT をそれぞれ PS-DT (1)，PS-DT (2) と呼ぶことにする．同様に，case 1 及び case 2 を用いた PDS-DT をそれぞれ PDS-DT (1)，PDS-DT (2) と呼ぶことにする．

3.3 PDS-DT により構築される共有構造

提案手法により構築される共有構造の例を図 6 に示す．図の下側の番号は，分布のインデックスを表し，横方向に現れる同じ値は，同じインデックス番号の分布を共有していることを表し，縦方向に現れる同じ値は，一つの多次元分布内の異なる次元に対応していることを表す．図 6 において，インデックスが 3 番，7 番，8 番の分布は状態位置をまたいで共有されている．特に，7 番と 8 番の分布は，すべての状態位置で共有されている．また，1，2，3，4，5，6 番の分布はすべて 2 次元分布であるが，それぞれ特徴ベクトルの異なる次元から構成されている．例えば，1 番と 2 番は $c(1)$ と $c(4)$ から構成されているが，3 番は $c(2)$ と $c(5)$ から構成されている．このように PDS-DT では，特徴ベクトルの各次元をグループ化しながら，各次元

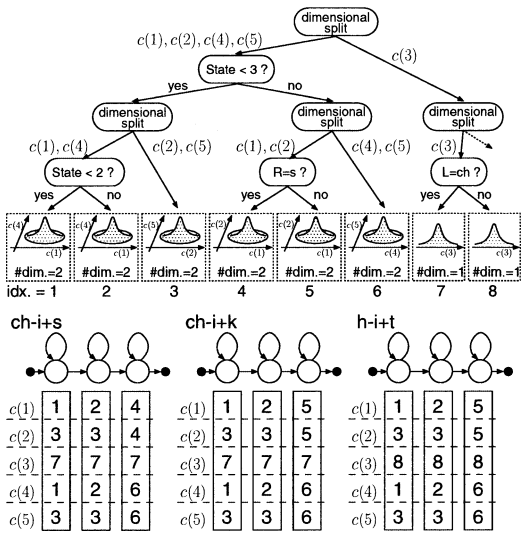

図 6　PDS-DT により構築される共有構造の概要
Fig. 6　An overview of sharing structure constructed by PDS-DT.

に対してコンテクスト及び時間方向に異なった共有構造を構築することが可能である．

このような共有構造は，[21], [22] で提案された順序制約付き非同期遷移型 HMM の共有構造と，次元ごとに異なる共有構造をもち得る，状態位置方向の共有が可能，という意味で似ている．ただし，順序制約付き非同期遷移型 HMM は，FD-SSS を用いて各次元のコンテクスト依存共有構造を構築した後，それらを統合する際に時間方向共有構造を構築するため，複数の次元で一つの多次元分布を構成することはできない．これに対して提案手法では，決定木に基づきコンテクスト・次元・状態位置がクラスタリングされるため，同一クラスタに分類された複数の次元で，一つの多次元分布を構成することができる．

4. 評価実験

提案法の有効性を評価するため，二つの不特定話者連続音声認識実験を行った．一つめの実験では，次元分割法の導入の有効性を評価する．二つめの実験では，次元分割法と状態位置に関する質問との組合せの有効性を評価する．

4.1 PD-DT の評価実験

4.1.1 実験条件

データベースとして ATR 日本語連続音声データベース b-set を用いた．男性話者 6 名 503 文章（a〜j セット）のうち各話者の 450 文章（a〜i セット）を

表 1　音響分析条件
Table 1　Acoustic analysis conditions.

サンプリング周波数	16 kHz
フレーム長/フレーム周期	25 ms / 5 ms
分析窓	ブラックマン窓
特徴パラメータ	メルケプストラム 1〜18 次
	Δ メルケプストラム 0〜18 次
	Δ^2 メルケプストラム 0〜18 次

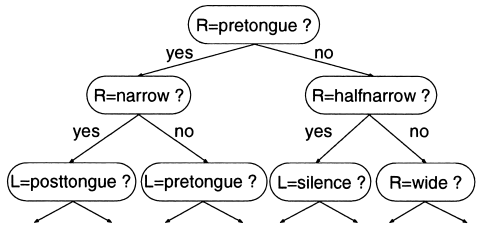

図 7　構築された P-DT の例（当該音素/k/の第 2 状態）
Fig. 7　Example of P-DT for the 2nd state of phoneme /k/.

学習に用い，残り 53 文章（j セット）を評価に用いた．音声分析条件を表 1 に示す．3 状態の left-to-right HMM を用いて 37 音素をモデル化した．また，決定木構築時の音素コンテクストに関する質問は 118 個用意した．日本語の音素連鎖を考慮したネットワーク上で，HTK のビタビアルゴリズムを用いて認識を行った（音素タイプライタ）．認識性能を音素誤り率（%エラー）で評価した．ラベル挿入ペナルティを変化させ，挿入誤り数と削除誤り数がほぼ等しくなったときの音素誤り率を各認識実験の結果とした．また，本研究では leave-one-out 法を用い，話者 6 名中の 1 名を除く 5 名を用いて音響モデルを学習し，残りの 1 名を評価に用いた．以後示すすべての実験結果は，六つの実験結果の平均である．パラメータ共有構造を構築する手法として，以下の三つについて実験を行った．

- 全次元共通で音素決定木を構築 (P-DT)
- 特徴ベクトルの各次元に音素決定木を構築 (FDP-DT)
- 音素・次元決定木を構築 (PD-DT)：提案手法 1

上記のすべての条件について，MDL 基準に基づき決定木を構築した．

4.1.2　実 験 結 果

はじめに，各手法 (P-DT, FDP-DT, PD-DT) を用いて構築された決定木の例を図 7〜図 9 に示す．これらはそれぞれ，当該音素/k/の第 2 状態における決定木の例である．図より，FDP-DT 及び PD-DT では，各次元に対し異なる共有構造を構築していることが分かる．

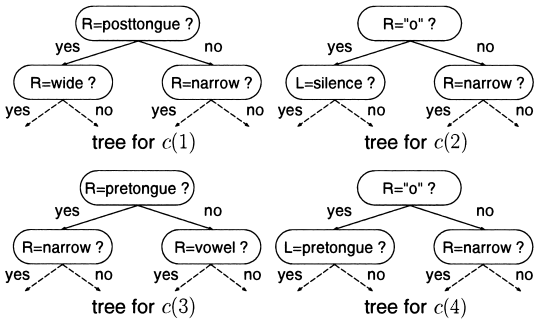

図 8　構築された FDP-DT の例（当該音素/k/の第 2 状態）
Fig. 8　Examples of FDP-DT for the 2nd state of phoneme /k/.

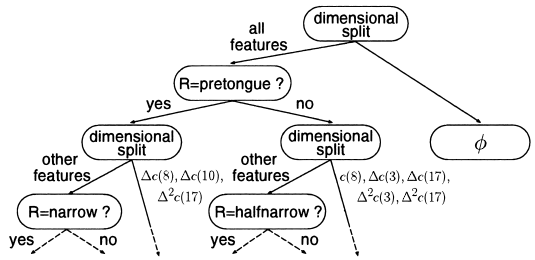

図 9　構築された PD-DT の例（当該音素/k/の第 2 状態）
Fig. 9　Example of PD-DT for the 2nd state of phoneme /k/.

表 2　各手法における分布数及びパラメータ数
Table 2　Average numbers of distributions (leaf-nodes) and free-parameters for each technique.

techniques	#distributions	#parameters
P-DT	3,044	344,624
FDP-DT	136,592	273,182
PD-DT	93,838	311,916

各手法を用いて構築した音響モデルの分布数（決定木の葉ノード数）及び単混合分布時のパラメータ数を表 2 に示す．なお，表 2 に示すパラメータ数には，分布の平均値，分散値を表現するためのモデルパラメータの総数を示した．FDP-DT 及び PD-DT により構築された音響モデルは，P-DT により構築されたモデルと比較して，総分布数が大きく増加しており，より高い表現能力をもつと考えられる．しかし，音響モデル中のパラメータ数は減少しており，過学習は生じにくいと思われる．

次に，特徴ベクトルの各次元に割り当てられた分布数を図 10 に示す．図より，FDP-DT 及び PD-DT は，メルケプストラムの高次より低次に多くの分布を割り当てている．同様に，動的特徴量より静的特徴量に多くの分布を割り当てている．このように，FDP-DT 及び PD-DT は，コンテクストによる変動が大きい次元には多数の分布を，変動が小さい次元には少数の分布を割り当てることにより，表現能力を自動的に調整することができる．それに対して，状態単位の共有構造では，すべての次元で同数の分布が割り当てられるため，各次元の表現能力を調整できない．

図 11 に認識実験結果を示す．PD-DT は P-DT と比較して，ほぼ同じパラメータ数において約 8%の誤り削減率が得られた．各葉ノードに対応するクラスタの出力確率分布が対角共分散単一ガウス分布である場合，FDP-DT と PD-DT の認識性能に大きな違いはなかった．しかし，混合ガウス分布化した場合，FDP-DT では改善が得られないのに対し，PD-DT では誤り率が減少した．これは，PD-DT では各次元に対して異なるコンテクスト依存共有構造を構築可能であり，かつ次元のまとまりを保つことができるため，状態出力確率分布を混合ガウス分布化することにより，クラスタリングの結果まとめられた次元間に相関を有するようなガウス分布を近似できるためだと考えられる．また文献 [16] おいて，状態共有 HMM から 4 階層共有構造 HMM を構築しても，状態共有 HMM の認識性能を上回ることがなかったことが報告されている．これに対して，PD-DT によりコンテクスト依存パラメータ共有構造を構築した HMM は，P-DT によりコンテクスト依存状態共有構造を構築した混合数 8 の状態共有 HMM より高い認識性能が得られた．このことから，決定木構築過程において次元分割を行い，各特徴量に対して異なるコンテクスト依存共有構造を構築することが有効であることが推察される．

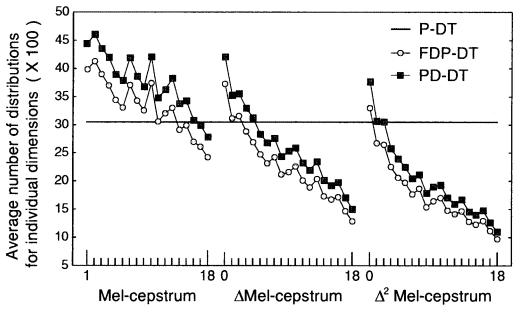

図 10　各特徴量に割り当てられた分布数
Fig. 10　Average numbers of distributions for each dimension.

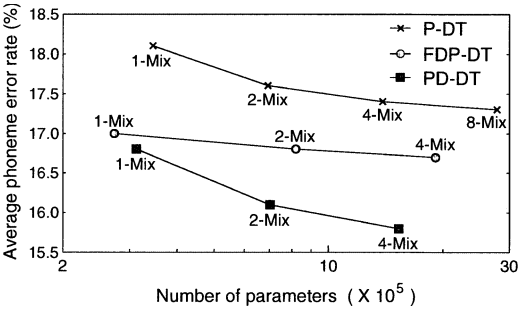

図 11　各手法の音素誤り率
Fig. 11　Average phoneme error rates for each technique.

4.2　PDS-DT の評価実験

4.2.1　実験条件

学習データには，日本音響学会の音素バランス文からなる研究用連続音声データベース (ASJ-PB) の男性話者すべてと，新聞記事読上げ音声コーパス (ASJ-JNAS) のうち男性話者 100 名分を利用した．評価データには，ASJ-JNAS のうち，学習データに含まれない話者 23 名の 100 文章 (IPA-98-TestSet) を用いた．発話単位でケプストラム平均正規化を行った．対角共分散単一ガウス分布の状態出力確率分布をもつ left-to-right HMM を用いて 43 種類の音素をモデル化した．共有構造の違いによる認識性能の変化を評価するため，混合分布化は行わなかった．各 HMM 当りの状態数を 3，4，5 と変化させ評価実験を行った．また，決定木構築時の音素コンテクストに関する質問は 150 個用意した．音響分析条件や認識実験の条件は，**4.1.1** と同様である．

4.2.2　実験結果

構築された PDS-DT の例を図 12 に示す．この例では，決定木の根ノードにおいて，$\Delta^2 c(14)$〜$\Delta^2 c(17)$ が次元分割されており，状態位置に関する質問が適用されていない．$\Delta^2 c(15)$，$\Delta^2 c(17)$ では，状態位置間の変動よりも先行音素による変動が大きかったため，質問 “L=o ?” が適用されている．最終的に構築された音素・次元・状態位置決定木では，$\Delta^2 c(14)$〜$\Delta^2 c(17)$ に対して状態位置に関する質問が全く適用されなかった．このため，$\Delta^2 c(14)$〜$\Delta^2 c(17)$ において，状態位置をまたいだ分布の共有が生じている．

次に，構築された音響モデルの分布数及びパラメータ数を表 3 に示す．表中の#distributions の括弧内

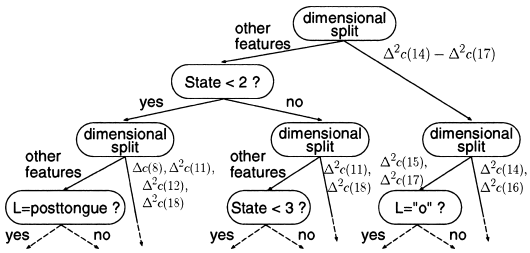

図 12　構築された PDS-DT の例（当該音素/N/，状態数=3）

Fig. 12　Example of PDS-DT for the phoneme /N/ (#states=3).

表 3　各手法の分布数（葉ノード数）及びパラメータ数

Table 3　Numbers of states, distributions (leaf nodes), and free-parameters for each technique.

#states	techniques	#distributions	#param.
3	P-DT	6,955	778,960
	PS-DT (1)	6,480 (0%)	725,760
	PS-DT (2)	6,955 (0%)	778,960
	PD-DT	186,596	645,560
	PDS-DT (1)	176,582 (0.25%)	603,842
	PDS-DT (2)	185,743 (0.23%)	644,004
4	P-DT	8,255	924,560
	PS-DT (1)	7,584 (0%)	850,080
	PS-DT (2)	8,255 (0%)	924,560
	PD-DT	233,109	783,912
	PDS-DT (1)	217,633 (0.60%)	722,234
	PDS-DT (2)	234,350 (0.54%)	780,568
5	P-DT	9,322	1,044,064
	PS-DT (1)	8,404 (0.93%)	941,248
	PS-DT (2)	9,236 (0.90%)	1,034,432
	PD-DT	272,817	899,424
	PDS-DT (1)	251,260 (2.28%)	814,408
	PDS-DT (2)	275,220 (2.11%)	893,574

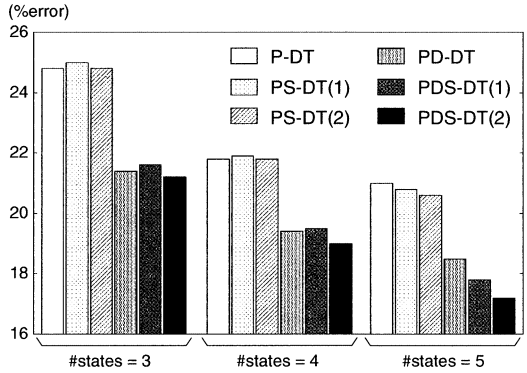

図 13　認識実験結果

Fig. 13　Recognition experimental results.

は，全分布中において状態位置をまたいで共有されている分布の割合を示す．各 HMM の状態数が 3 または 4 の場合，PS-DT では状態位置をまたいだ状態の共有は生じなかったが，PDS-DT では状態位置をまたいだ分布の共有が生じている．状態数が 5 のとき，PDS-DT では全体の約 2%の分布が状態位置をまたいで共有されていた．

認識実験結果を図 13 に示す．状態数が 3 または 4 の場合，PDS-DT の認識性能は PD-DT とほぼ同等，P-DT や PS-DT と比較して 13〜15%の誤り削減率が得られた．状態数が 5 のとき，PDS-DT は PD-DT と比較して 7%，P-DT や PS-DT と比較して 16%の誤り削減率が得られた．

実験結果より，状態位置に関する分割条件は，状態単位の共有構造では余り効果が得られないが，次元分割法と組み合わせることで，特に状態数が多いとき有効であることが明らかとなった．

5. むすび

本論文では,「MDL 基準に基づく次元分割法」を導入することにより，音素決定木による状態クラスタリング法を拡張し，音素コンテクスト及び次元を分割条件とする，音素・次元決定木 (PD-DT) を提案した．更に，状態位置に関する分割条件を導入し，音素コンテクスト・次元・状態位置を同時にクラスタリングする，音素・次元・状態位置決定木 (PDS-DT) を提案した．提案法では，モデルの各次元に対して異なるコンテクスト及び状態位置依存共有構造を構築することが可能であり，より精度の高い音響モデルを構築することができる．不特定話者連続音素認識タスクにおいて，提案手法を従来の音素決定木に基づく状態共有法と比較したところ，13〜16%の音素誤り削減率が得られた．

今後は，他手法との詳細な比較検討，大語い連続音声認識への適用を行う予定である．

謝辞　本論文に関して熱心な御討論と有益な御助言を頂いた，名古屋工業大学南角吉彦氏，宮島千代美助手，豊田中央研究所吉村貴克博士，東京工業大学益子貴史助手，ATR 音声言語コミュニケーション研究所松田繁樹博士に感謝する．また，有益なコメントを頂いた査読者に感謝する．

文　献

[1] K.-F. Lee, “Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition,” IEEE Trans. Acoust. Speech Signal Process., vol.38, no.4, pp.599–609, April 1990.

[2] 鷹見淳一，嵯峨山茂樹，“逐次状態分割法による隠れマルコフ網の自動生成，” 信学論（D-II），vol.J76-D-II, no.10, pp.2155–2164, Oct. 1993.

[3] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," Proc. ICASSP 93, vol.II, pp.311–314, Minneapolis, U.S.A., April 1993.
[4] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. ARPA Workshop on Human Language Technology, pp.307–312, Berlin, Germany, March 1994.
[5] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, March 2000.
[6] W. Chou and W. Reichl, "Decision tree state tying based on penalized Bayesian information criterion," Proc. ICASSP 99, vol.1, pp.345–348, Phoenix, U.S.A., March 1999.
[7] 渡部晋治，南　泰浩，中村　篤，上田修功，"ベイズ的アプローチに基づく状態共有型 HMM 構造の学習，" 信学技報，SP2002-14, April 2002.
[8] W. Reichl and W. Chou, "A unified approach of incorporating general features in decision tree based acoustic modeling," Proc. ICASSP 99, vol.2, pp.573–576, Phoenix, U.S.A., March 1999.
[9] I. Shafran and M. Ostendorf, "Use of higher level linguistic structure in acoustic modeling for speech recognition," Proc. ICASSP 2000, vol.3, pp.1643–1646, Istanbul, Turkey, June 2000.
[10] 樋口晋也，武田一哉，板倉文忠，"言語情報を用いた音響モデルの作成，" 2001 音響学秋季講論集，分冊 I，no.1-Q-21, pp.181–182, Oct. 2001.
[11] 五十川賢造，篠田浩一，嵯峨山茂樹，"形態素情報と単語内位置情報を用いた話し言葉音声認識のための音響モデル，" 情処学研報，2002-SLP-44, pp.111–116, Dec. 2002.
[12] H. Suzuki, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Speech recognition using voice-characteristic-dependent acoustic models," Proc. ICASSP 2003, vol.I, pp.740–743, Hong Kong, China, April 2003.
[13] 加藤恒夫，黒岩眞吾，清水　徹，樋口宜男，"混合分布 HMM における Tree-based クラスタリング，" 信学論（D-II），vol.J83-D-II, no.11, pp.2128–2136, Nov. 2000.
[14] H.J. Nock, Context clustering for triphone-based speech recognition, Master Thesis, Cambridge University, Aug. 1996.
[15] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process., vol.34, no.1, pp.52–59, Feb. 1986.
[16] 高橋　敏，嵯峨山茂樹，"4 階層共有構造の音響モデルによる音声認識，" 信学論（D-II），vol.J82-D-II, no.3, pp.315–323, March 1999.
[17] S. Matsuda, M. Nakai, H. Shimodaira, and S. Sagayama, "Feature-dependent allophone clustering," Proc. ICSLP 2000, vol.I, pp.413–416, Beijing, China, Oct. 2000.
[18] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language, vol.11, no.1, pp.17–41, Jan. 1997.
[19] 松田繁樹，中井　満，下平　博，嵯峨山茂樹，"複数混合分布を持つ順序制約付き非同期遷移型 HMM，" 2000 音響学秋季講論集，分冊 I，no.1-5-11, pp.21–22, Oct. 2000.
[20] A. Lazaridès, Y. Normandin, and R. Kuhn, "Improving decision trees for acoustic modeling," Proc. ICSLP 96, vol.2, pp.1053–1056, Philadelphia, U.S.A., Oct. 1996.
[21] 松田繁樹，中井　満，下平　博，嵯峨山茂樹，"状態遷移に順序関係を持つ非同期遷移型 HMM，" 信学技報，SP99-98, Dec. 1999.
[22] S. Sagayama, S. Matsuda, M. Nakai, and H. Shimodaira, "Asynchronous-transition HMM for acoustic modeling," Proc. ICASSP 2000, vol.II, pp.1001–1004, Istanbul, Turkey, June 2000.

（平成 15 年 5 月 28 日受付，16 年 1 月 6 日再受付）

全　炳河

平 11 鈴鹿高専・電子情報卒．平 13 名工大・工・知能情報システム卒．平 15 同大大学院博士前期課程了．現在，同博士後期課程在学中．音声認識・合成の研究に従事．日本音響学会会員．

徳田　恵一　（正員）

昭 59 名工大・工・電子卒．平元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．平 16 名工大大学院（情報工学専攻）教授．工博．音声言語情報処理，マルチモーダル情報処理，統計的学習理論の研究に従事．平 13 電気通信普及財団賞，平 13 本会論文賞，猪瀬賞各受賞．日本音響学会，人工知能学会，情報処理学会，IEEE，ISCA 各会員．

北村　正　（正員）

昭 48 名工大・工・電子卒．昭 53 東工大大学院博士課程了．同年東工大精密工学研究所助手．昭 58 名工大・工・電子工学科講師．昭 59 同助教授．平 7 名工大知能情報システム学科教授．平 15 名工大大学院（情報工学専攻）教授．工博．音声情報処理，マルチメディア情報処理，動画像処理の研究に従事．日本音響学会，情報処理学会，IEEE，ISCA 各会員．