論　文

音声の構造的表象に基づく日本語孤立母音系列を対象とした音声認識

村上　隆夫[†*a)]　　峯松　信明[††]　　広瀬　啓吉[†]

Recognition of Spoken Sequences of Japanese Isolated Vowels Based on Structural Representation of Speech

Takao MURAKAMI[†*a)], Nobuaki MINEMATSU[††], and Keikichi HIROSE[†]

あらまし　音声には話者の声道形状の特性，音響機器の特性などの非言語的特徴が不可避的に混入する．近年，これらを表現する次元を原理的に保有しない音響的普遍構造が提案されている．これは，音声事象の物理的実体を捨象し，関係のみをとらえることによって得られる音声の構造的表象である．この構造を音声認識において利用することを目的として，日本語孤立母音系列をタスクとする不特定話者音声認識に関する基礎検討を行った．クリーン音声を対象とした認識実験を行った結果，提案手法は音声の物理的実体を明示的に用いずに，学習話者 1 名で不特定話者の母音系列を 100%の性能をもって認識することに成功した．また雑音重畳音声に関しては，雑音下で学習した学習話者 1 名の提案手法が，SS（Spectral Subtraction）及び CMN（Cepstral Mean Normalization）を用いた学習話者 4,130 名の従来手法を上回る結果が得られた．

キーワード　音声の構造的表象，日本語孤立母音系列，音声認識，最大事後確率推定，スペクトル高域成分の均一化

1. まえがき

我々が音声コミュニケーションを行う際，音声の生成時には話者の声道形状の特性，収録・伝送・再生時には音響機器の特性，聴取時には聴覚特性，といった非言語的要因によって音声の音響的特性は不可避的に変形する．従来の音声認識技術は，フォルマントやスペクトル包絡などの音声の物理的実体をとらえてモデル化してきたが，この物理的実体は上記の非言語的要因によって不可避的に多様なひずみを生む．この多様性問題に対して，従来では何百，何千の学習話者を集めて話者バランスのとれた不特定話者音響モデルを構築し，更には話者適応，話者正規化技術の研究が様々な形で行われてきた [1] が，根本的な解決には至っていない．

† 東京大学大学院情報理工学系研究科，東京都
Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
†† 東京大学大学院新領域創成科学研究科，柏市
Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa-shi, 277–8561 Japan
* 現在，(株）日立製作所システム開発研究所
a) E-mail: takao.murakami.nr@hitachi.com

その一方で，人間は非常に偏った音声提示環境のもと，様々な音響ひずみに対する対処法を獲得する．例えば，幼児の聞く声の大部分は両親の声である．更には，対話が自分と相手との音声コミュニケーションで成立することを考えると，偏った音声提示環境が一生続くことが分かる．人の聞く声の約半分は自分の声だからである．音声生成は自らの聴取によるフィードバックを必要とする（スピーチチェイン）．これは，不特定話者音響モデルが構築される何百，何千という学習話者環境とは相反するものである．それにもかかわらず音声は人間にとって一番楽なコミュニケーションメディアである．このような本質的な違いは一体どこから来るのだろうか．

ここで，言語学の観点から音声を眺めてみると，音素には以下の二つが定義されていることが分かる [2]．(1) a phoneme is a class of phonetically-similar sounds and (2) a phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system. 音声の絶対的な特性に基づく従来手法は，すべて第一の定義に基づくものと考えることができる．しかし，これは二つある定義のうちの一方でしかない．

また，近代言語学の祖であるソシュールは言語に対して，“Language is a system of only conceptual differences and phonic differences.” と主張している [3]．このソシュールの言語哲学に啓蒙され，ヤコブソンらは「構造音韻論」と呼ばれる言語学の一分野を確立した．言語音そのものではなく，音と音はどう違うのかに着目し，図 1 に示すような子音三角形，母音三角形を考えた．ここで音と音の違いは弁別素性によって表現されている．図 2 はこの考え方をフランス語の母音，準母音群に適用したものである [4]．本来，弁別素性は二つの言語音の違いを記述するために考案されたが，やがて素性の束で音，すなわち音素を絶対的に表現するようになる．結局，ほかとの違いで各音素を定義するのではなく，素性を用いて音素を絶対的に定義するようになる．音声認識においても音素を弁別素性の束とみなし，弁別素性に着眼した研究例 [5]〜[7] もあるが，これらはすべて音声事象の絶対的特性を用いて素性をとらえたものである．しかし，本来の定義に戻れば，弁別素性は音的差異を表現するために作られたものであり，構造音韻論は話者を問わず同一の幾何学的構造が普遍的に存在すると主張する．

近年，冒頭の非言語的特徴を表現する線形変換性・乗算性ひずみを一切保有しない音響的普遍構造が提案

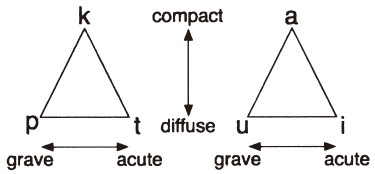

図 1　ヤコブソンによる子音と母音の三角形
Fig. 1　Consonant and vowel triangles proposed by Jakobson.

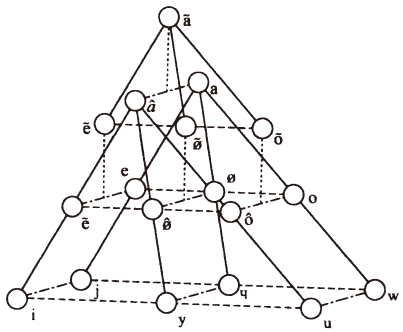

図 2　ヤコブソンによるフランス語の幾何学的音韻構造
Fig. 2　Jakobson’s geometrical structure of French.

された [8], [9]．これは，音声の物理的実体を捨象し，関係のみをとらえることによって得られる音声の構造的表象である．音素の定義で見れば，第二の定義に基づくものであり，構造音韻論の物理実装に相当する．この構造を英語発音の音響的分析に利用する試みが既に行われており，非言語的特徴の違いに対して頑健な分析が行えることが実験的に示されている [10]．

本論文ではこの構造を音声認識に利用することに関する基礎検討を行う．まず **2.** において音響的普遍構造の概要を述べ，一発声の構造化及び構造に基づく音響的照合が可能であることを示す．**3.** では，構造を用いて日本語孤立母音系列を対象とした音声認識を行う枠組みについて説明する．**4.** 及び **5.** では，それぞれクリーン環境下，雑音環境下での日本語孤立母音系列の認識実験について，従来手法との比較実験も含めて述べる．

2.　音声の構造的表象

2.1　音声に不可避的に混入する非言語的特徴

音声に混入する非言語的特徴は主に加算性雑音，乗算性ひずみ，線形変換性ひずみの三種類に分類される．このうち，音声に「不可避的に」混入するものは乗算性ひずみ，線形変換性ひずみの二つである．加算性雑音とは，時間軸上の加算で表現される雑音であり，テレビ・ラジオなどの背景雑音がその典型例といえる．これらは場所を移動するなどして物理的に抹消することができるという意味において，不可避的な雑音ではない．

乗算性ひずみは，スペクトルに対する乗算で表現されるひずみであり，これはケプストラムベクトル c に対するベクトル b の加算 $c' = c + b$ に相当する．マイクロホンなどの音響機器の特性がその典型例である．また，乗算性ひずみを消失させるために，入力音声のケプストラムからその平均値を減算する CMN（Cepstral Mean Normalization）があるが，これによって話者性の違いによる影響も軽減できる．すなわち，話者の声道形状の違いの一部も近似的に乗算性ひずみとして扱うことができる．音声は必ずある話者によって発声され，ある音響機器によって収録されるので，これらは不可避的なひずみである．

線形変換性ひずみは，c に対する行列 A の乗算 $c' = Ac$ で表現されるひずみである．話者の声道長の差異，聴取者の聴覚特性の差異を表すために，工学的には対数スペクトルに対して周波数ウォーピングが施

されるが，単調増加かつ連続である周波数ウォーピングは，c に対する A の乗算で表される [11]．すなわち，声道長の差異，聴覚特性の差異は近似的に線形変換性ひずみとして扱うことができる．これらも不可避的なひずみである．

以上をまとめると，音声に不可避的に混入する非言語的要因による音変形は，ケプストラムベクトル c に対する $c' = Ac + b$ という変換で簡単に表現される．これはアフィン変換と呼ばれる．MLLR [12] や SAT [13] でも，話者性がアフィン変換で記述される．図 3 は，アフィン変換 $Ac + b$ が対数スペクトルに与える影響を示したものである．A は対数スペクトルの水平変化，b は垂直変化を引き起こす．なお，単一のアフィン変換は最も簡素な声質変換に相当するが，性別変換や年齢変換に必要なスペクトル変形はこの変換で容易に実装可能である．

2.2 音声に内在する音響的普遍構造

観測された音声から，各音声事象（例えば各音素）を分布化し，N 個の分布によって構成される構造を抽出することを考える．N 個の分布に対して ${}_NC_2$ 個のすべての二分布間距離を求めれば，一つの幾何学的構造を規定したことになる．音声に不可避的に混入する非言語的特徴は，アフィン変換 $Ac + b$ で表現されるが，これに対して不変な構造を抽出することが構造音韻論の物理実装のための必要条件と考えられる．しかしながら，アフィン変換は構造をひずませる変換である．そのため，不変な構造は「空間」をひずませることで抽出される．

構造不変の定理：意味のある記述が分布としてのみ可能な物理現象を考える．分布群に対して，すべての二分布間距離を求める（距離行列）．二分布間距離として，バタチャリヤ距離，カルバック・ライブラ距離，ヘリンガー距離などを用いた場合，各分布に対して単一の任意一次変換を施しても，二分布間距離は不変である．すなわち距離行列は不変であり，その結果，構造も不変となる（図 4 参照）．

以下，バタチャリヤ距離を用いて話を進める．二つの分布の確率密度関数をそれぞれ $p_1(x)$，$p_2(x)$ とすると，バタチャリヤ距離は以下の式で表される．

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \tag{1}$$

$0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leq 1$ を確率として解釈すれば，これは自己情報量となり，単位は [bit] となる．二つの分布がガウス分布で表現されているとき，バタチャリヤ距離は，

$$\begin{aligned} &BD(p_1(x), p_2(x)) \\ &= \frac{1}{8}\mu_{12}^T \left(\frac{\sum_1 + \sum_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2}\ln \frac{|(\sum_1 + \sum_2)/2|}{|\sum_1|^{\frac{1}{2}}|\sum_2|^{\frac{1}{2}}} \end{aligned} \tag{2}$$

となる．μ_{12} は $\mu_1 - \mu_2$ である．このとき，二つの分布に対して共通のアフィン変換 $Ac + b$ をかけた場合，バタチャリヤ距離はその前後で不変である．これは，バタチャリヤ距離が空間をゆがめる距離尺度であることに起因する．したがって，バタチャリヤ距離行列によって規定されるこの構造もアフィン変換に対して不変となる（音響的普遍構造）．c に A を掛ける演算は構造の回転として観測され，b を加える演算は構造のシフトとして観測される．

2.3 一発声の構造化と構造に基づく音響的照合

音声認識は一発声された音声を対象として扱うが，音声からケプストラム系列を求め，そこから音声事象分布（ケプストラム分布）の系列を得た後，任意の二分布間距離を求めれば一発声の構造化も可能である（図 5）．次に，二つの構造の構造間差異を求めることで，求めた構造を音響的に照合することを考える．M 個の頂点で構成される二つの構造（$P_1, \ldots, P_M$,

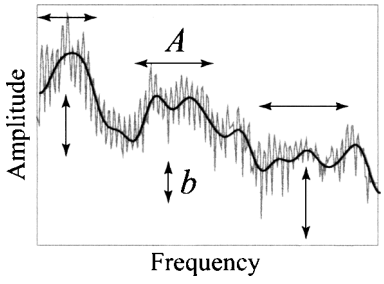

図 3　A, b によって引き起こされるスペクトルひずみ
Fig. 3　Spectrum distortions caused by A and b.

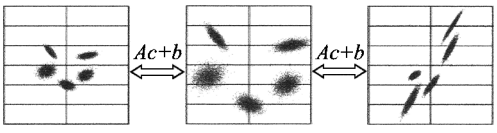

図 4　構造不変の定理（これらがすべて同一の構造となる）
Fig. 4　Theorem of the invariant structure.

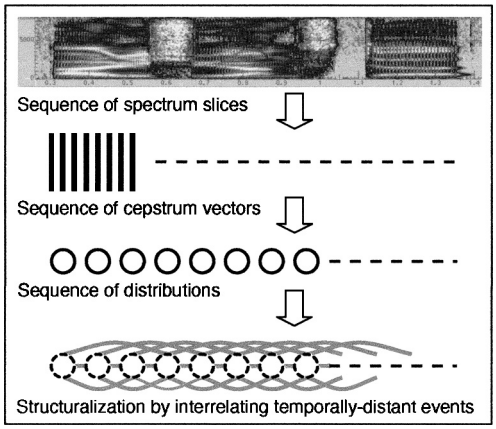

図 5　一発声の構造化
Fig. 5　Structuralization of a single utterance.

$Q_1, \ldots, Q_M$）において，構造 Q をシフト（b）と回転（A）のみで構造 P に近づけ，対応する頂点間距離の和（$\sum_{i=1}^{M} \overline{P_iQ_i}^2$）の最小値を求めることで構造間差異を定義する．すなわち，図 6 に示されるような枠組みの音響的照合である．（シフト（b）は乗算性ひずみ，回転（A）は線形変換性ひずみを表現するため，図 6 の操作によって得られるスコアは話者（環境）適応・正規化後の音響照合スコアとなる．これが A や b を求めることなく算出される．）二つの構造が N 次元ユークリッド空間内にある場合，その構造間差異は以下の式によって導出される [14].

$$\sum_{i=1}^{M} \overline{OP_i}^2 + \overline{OQ_i}^2 - 2\sum_{i=1}^{N} \sqrt{\alpha_i}, \qquad (3)$$

O は両構造の重心である（構造をシフトさせて重心を重ねる）．α_i は N 次正方行列 S^tTT^tS の固有値である．S は行列（$\overrightarrow{OP_1}, \ldots, \overrightarrow{OP_M}$）であり，$T$ は行列（$\overrightarrow{OQ_1}, \ldots, \overrightarrow{OQ_M}$）である．しかしながら，音響的普遍構造は空間をひずませることで得られるため，ユークリッド空間内には存在しない．したがって，三角不等式が満たされない場合があり，直接式 (3) を用いることはできない．ここで，分布間距離としてバタチャリヤ距離の平方根を用いた場合，シフト（b）及び回転（A）後の $\sum |\theta_i|$（$|\theta_i| = \angle P_iOQ_i$）が十分に小さければ，

$$\sqrt{\sum_{i<j} (\overline{P_iP_j} - \overline{Q_iQ_j})^2} \qquad (4)$$

が上記構造間差異を近似することが示されている [15]. 式 (4) は距離行列のうち意味をもつ上三角成分をベクトル（これを「構造ベクトル」と定義する）として並べたときのユークリッド距離に対応する．したがって，構造に基づく音響的照合は，距離行列のみを用いて近似的に行うことができる．以上より，音響的普遍構造を用いた音声認識が可能であることが示唆される．

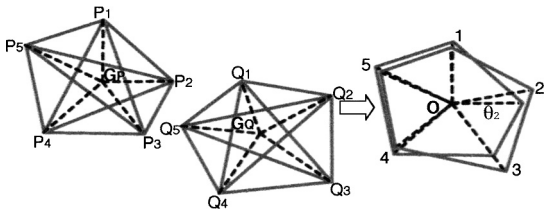

図 6　構造に基づく音響的照合
Fig. 6　Structure-based acoustic matching.

提案手法と従来の適応・正規化手法との違いを明確にしておきたい．本手法はスペクトルスムージングによってピッチ情報が取り除かれるように，非言語的特徴を音声の物理的特性から除去する方法論である．換言すれば，非言語的特徴を表現する次元そのものを消す方法である．従来のように音そのものを扱う方法論では，この次元が常に残るため，（例えば A や b を具体的に求め）常に音を定位させる必要があった．

3. 音声の構造的表象を用いた日本語孤立母音系列の認識

3.1 音声の構造的表象を用いた日本語孤立母音系列の認識の枠組み

本論文では，構造を用いて日本語孤立母音系列を認識対象とした音声認識を行うことを考える．これは，/a/，/i/，/u/，/e/，/o/の各母音が一回ずつ孤立的に発声されたものを一つの単語とみなし（語彙サイズ ${}_5P_5 = 120$），それを認識するタスクである．これを，構造を用いて認識する枠組みを図 7 に示す．

まず入力音声から各母音の音声事象分布（ケプストラム分布）を求め，これを構造化する．このとき，構造サイズ（構造の大きさ）が一定値となるよう正規化する．ここで構造サイズは，求めた距離行列（構造を一意に規定する）に対して Ward 法を用いたボトムアップクラスタリングを行い，その結果得られた樹形図の高さとして定義する．Ward 法は，統合により生じる累積ひずみが最小となる二要素を順次統合していく方法であり，最終的に得られる樹形図の高さは全要素を一点（重心）で代表させたときのひずみとなる．これはコードブックサイズが 1 のときの VQ ひずみに相当し，音素群を構造として見た場合の構造の半径

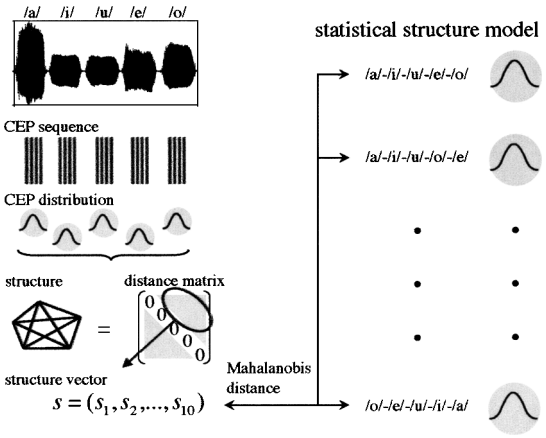

図 7　構造を用いた日本語孤立母音系列の認識
Fig. 7　Recognition of spoken sequences of Japanese isolated vowels using the structure.

に相当する量として解釈することができる．母音群を対象とした場合，その重心位置には「あいまい母音（schwa）」が存在する．英語の強勢母音，弱勢母音を考えれば分かるように，生成された母音と重心母音との距離は「どの程度の調音努力を払って該当母音を生成したか」に相当する．[16] では，これらの考察に基づいて構造サイズを調音努力（発話スタイルの一種）と解釈することで種々の分析を行っている．したがって，構造サイズの正規化は音声認識において有効であると考えられる．

構造として実際に求めるのは距離行列であり，このうち意味をもつ成分は上三角成分であるので，これをベクトルとして並べた「構造ベクトル」（10 次元）を特徴ベクトルとして用いる．この特徴ベクトルには個々の母音を絶対的，個別的に同定するために必要な特徴量は存在しない．提案手法は音の同定は一切行わず，音の系列として存在する語の同定を行う手法である．ある音時系列に存在する事象間距離を（離れているものを含め）すべて集めることでその音時系列を表象する．言語音の絶対的価値を論ずるのが音声学であり，相対的価値を論ずるのが音韻論である．従来の方法はすべて前者に基づくものであり，提案手法は後者を物理的に解釈した方法である．

認識器にもたせる構造モデルは以下のように作成する．複数の/a/-/i/-/u/-/e/-/o/の構造ベクトルから 10 次元ガウス分布（全共分散行列を使用）を求め，これを/a/-/i/-/u/-/e/-/o/の「構造統計モデル」とする．他の 119 個（/i/-/a/-/u/-/e/-/o/など）の構造統計モデルは，/a/-/i/-/u/-/e/-/o/の構造統計モデルの要素を交換することで得られる．最終的に 120 個の構造統計モデルが得られ，これを認識に利用する．

構造の音響的照合は，入力構造ベクトルと各構造統計モデルとのマハラノビス距離を算出することで行う．これは，式 (4) のシフト（b）と回転（A）に基づく音響的照合の近似を，入力構造と構造統計モデルの間で行うことに相当する．この距離が最も小さい単語を認識結果とする．

3.2　音声事象分布の最大事後確率推定

音響的普遍構造を音声認識に利用する場合，一発声された音声から音声事象分布を推定する必要がある．最ゆう（ML）推定は分布の推定手法として広く用いられているが，得られるデータ量 n が少ないときに不適切な分布を推定する可能性がある．したがって，一発声を構造化する本研究においては，この問題が顕著となる．

そこで，音声事象分布の最大事後確率（MAP）推定を検討する．MAP 推定の具体的な枠組みに関しては [17] を参照した．以下，分散共分散行列はすべて対角である．また，ここでは日本語孤立母音系列を認識対象として扱うので，各母音（/a/, /i/, /u/, /e/, /o/）の孤立発声を複数用意し，これを事前知識として用いる．これらは一発声ごとにガウス分布化される（計 M 個）．MAP 推定に用いるパラメータは以下のとおりである．

μ_m ： m 番目の発声の平均ベクトル
Σ_m ： m 番目の発声の対角共分散行列
μ_0 ： $\{\mu_m\}$ の平均 $(= \frac{1}{M}\sum_{m=1}^{M}\mu_m)$
Σ_0 ： $\{\Sigma_m\}$ の平均 $(= \frac{1}{M}\sum_{m=1}^{M}\Sigma_m)$
S_μ ： $\{\mu_m\}$ の対角共分散行列
$(= \frac{1}{M}\sum_{m=1}^{M}(\mathrm{DIAG}(\mu_m - \mu_0))^2)$
Ω ： $= \Sigma_0 S_\mu^{-1}$
μ_{ML} ： 入力発声の平均ベクトル（ML 推定）
Σ_{ML} ： 入力発声の対角共分散行列（ML 推定）

ここで，$\mathrm{DIAG}(x)$ は，ベクトル x の要素を対角成分に並べた対角共分散行列である．これらを用いて，MAP 推定では入力発声の分布を以下のように推定する．

$$\mu_{MAP} = \hat{\mu_0} \quad (5)$$

$$\Sigma_{MAP} = \hat{B}\hat{A}^{-1} \quad (6)$$

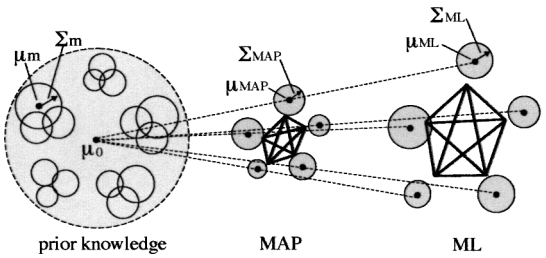

図 8　音声事象分布の最大事後確率推定
Fig. 8　MAP-based estimation of distributions of speech events.

ここで,

$$\hat{\mu_0} = \Omega(\Omega + nE)^{-1}\mu_0 + n(\Omega + nE)^{-1}\mu_{ML} \quad (7)$$

$$\hat{B} = B + \frac{n}{2}\Sigma_{ML} + \frac{n}{2}\Omega(\mathrm{DIAG}(\mu_{ML} - \mu_0))^2(\Omega + nE)^{-1} \quad (8)$$

$$B = E \quad (9)$$

$$\hat{A} = A + \frac{n}{2}E \quad (10)$$

$$A = \Sigma_0^{-1} \quad (11)$$

である．μ_{MAP} は μ_0 と μ_{ML} の内挿値をとり，n の増加につれて μ_{ML} に近づく．本論文では各母音ごとに中心前後 14 フレームが用いられる．したがって，本来 $n = 14$ であるが，この値を変化させて入力発声の事前知識に対する重みを調節することが可能である．音声事象分布の MAP 推定の様子を図 8 に示す．なお, [10] では一発声に基づく英語発音の構造化に上記の MAP 推定を適用し，少量データの場合でも安定した発音の構造化が可能であることを実験的に示している．

3.3　スペクトル高域成分の均一化

本論文では，音声に不可避的に混入する非言語的特徴をアフィン変換 $Ac+b$ で表現しているが，これは簡素なモデルであるため，音響的普遍構造が非言語的特徴を消失させる効果は限られている可能性がある. [18] は，母音のスペクトル包絡の 2.2 kHz 以上の帯域には話者性の情報が多く含まれていることを実験的に示している．これに基づいて，話者性をより効果的に消失させるために，本論文では音声に LPF（ローパスフィルタ）を通す，あるいは雑音を重畳することでスペクトル広域成分を均一化させることを試みた．

図 9 に，5 名の話者が発声した/a/のスペクトル包絡を示す．上・中央・下の図は，それぞれクリーン音声・LPF（カットオフ周波数：2 kHz）を施した音声・白色雑音（SNR=10 [dB]）を重畳した音声に対応する．話者による違いがスペクトル高域成分によく表れているが，LPF を通すことでそれが下に，白色雑音を重畳することでそれが上にそろえられ，話者性の消失が効果的に行われていることが分かる．

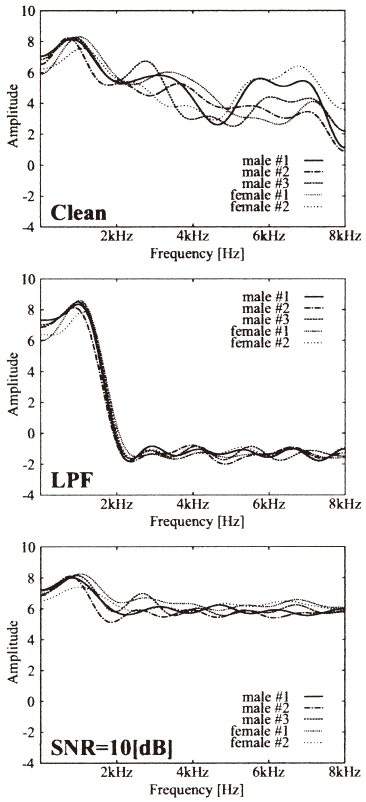

図 9　5 名話者の/a/のスペクトル包絡
Fig. 9　Spectral envelopes of /a/ of 5 speakers.

ただし，実際に図 9 中央及び下の音声を聞いてみると，話者性は完全には消失されていないことが容易に確認できる．また，個々の音韻性についても，母音間差異は十分に残存していることが確認できる．

4.　構造的表象を用いたクリーン環境下での日本語孤立母音系列の認識実験

4.1　実験条件と実験結果

本節から実際に行った認識実験について述べる．ここで，一発声の構造化によって，音声に不可避的に混入する非言語的特徴が十分に消失されるのであれば，以下の三つの音声認識の可能性が見えてくる．

- 音声の物理的実体を明示的に用いない音声認識
- 一人の話者で学習された音響モデル（構造モデル）

を用いた不特定話者音声認識
- 適応・正規化技術が一切不要な音声認識

これらを満たす音声認識の可能性を検証するべく認識実験を行った．

まずはクリーン環境下での認識実験を行った．入力構造ベクトルについては，男性 4 名，女性 4 名の計 8 名の評価話者が，5 母音を 5 回孤立発声したデータを用いて抽出した．まずこの音声に，カットオフ周波数 2 kHz，4 kHz，8 kHz（全帯域）のいずれかの LPF を通した．次にケプストラムとしてメルケプストラム MCEP（α=0.55）（1〜12 次元）[19] を抽出した．ここで α は周波数伸縮パラメータであり，サンプリング周波数が 16 kHz のとき，α=0.55 とすればバーク尺度に対してよい近似を与えることが知られている [20]．各母音のケプストラム分布は中心前後 14 フレーム（140 ms）を用いて，12 次元ガウス分布（対角共分散行列）を推定した．分布の推定方法は，ML 推定，MAP 推定の両方を試みた．そして各話者ごとに，3,125（$=5^5$）個の/a/-/i/-/u/-/e/-/o/の構造ベクトルを抽出し，計 25,000（$=8\times3,125$）個の構造ベクトルを入力に用いた．なお，他の母音列については，/a/-/i/-/u/-/e/-/o/の要素を交換することで得られ，認識結果も同様に入れ換わるだけであるため，入力音声には用いなかった．

構造統計モデルについてだが，高域成分を用いない構造化によって，仮に非言語的特徴が完全に消失できるのであれば，学習話者は 1 名で済むはずである．したがって，ここでは構造統計モデルの学習に，上記の評価話者とは別の男性 1 名が 5 母音を 35 回孤立発声した音声資料を用いた．これらを七つのグループ（1 グループにつき，5 母音 5 回孤立発声）に分割し，各グループに対し，3,125（$=5^5$）個の/a/-/i/-/u/-/e/-/o/の構造ベクトルを抽出した．得られた計 21,875（$=7\times5^5$）個の構造ベクトルを用いて 10 次元ガウス分布（全角共分散行列）を推定し，これを構造統計モデルとして使用した．MAP 推定に用いる事前知識は，評価話者の音声事象に対しては，7 グループ内の全母音データ（$7\times5\times5$ 個）を用いた．学習話者の音声事象に対しては，当該グループを除く 6 グループ内の全母音データ（$6\times5\times5$ 個）を用いた．ここでは，MAP 推定時の入力音声の事前知識に対する重み n として種々の値（n=10, 1, 0.1, 0.01）を使用して検討した（$n=\infty$ で ML 推定）．このときの音響的条件を表 1 に示す．

認識結果を表 2 に示す．認識率は，高域成分除去を施すことで飛躍的に向上する．その際，MAP 推定の重み n を小さくすることによる効果が見られなくなったのは，ML 推定の認識率が向上したためと見られる．注目すべきは，カットオフ周波数が 2 kHz の時に MAP 推定を用いることで，100%の認識率が得られている点である．これは，本節冒頭で述べた三つの音声認識が，今回の認識タスクでは 100%の性能をもっていずれも実現可能であることを意味する．なお，これは重み n が本来の値 n=14 のときでも成立する（**4.2** 参照）．

表 1　クリーン環境下での認識実験における音響的条件
Table 1　Acoustic conditions in the recognition experiments in the clean environment.

サンプリング	16 bit / 16 kHz
窓	窓長 25 ms，シフト長 10 ms
音声特徴量	MCEP（α=0.55）（1〜12 次元）
音声事象分布	12 次元ガウス分布（対角共分散行列）
分布推定方法	ML or MAP（n=10, 1, 0.1, 0.01）

表 2　クリーン環境下での日本語孤立母音系列の認識結果
Table 2　Recognition results of spoken sequences of Japanese isolated vowels in the clean environment.

推定法 \ 帯域	full-band	4 kHz	2 kHz
ML	24.7 %	47.9%	86.8%
MAP(n=10)	42.9 %	62.7%	100.0%
MAP(n=1)	42.6 %	62.1%	100.0%
MAP(n=0.1)	45.7 %	60.8%	99.9%
MAP(n=0.01)	70.3 %	65.4%	96.7%

4.2　構造サイズの正規化による効果

4.1 では構造サイズの正規化を行っていたが，正規化を行う場合と行わない場合の性能比較も行った．ここでは，音響的条件は表 1 のまま，学習話者を評価話者を除く 7 名にした（**4.1** の七つのグループをこの 7 名で置き換える）場合と，男性話者 1 名（**4.1** と同じ）にした場合で，それぞれ実験を行った．また，MAP 推定における重み n は，本来の値である n=14 と設定した．結果を表 3 に示す．

表 3 から，構造サイズの正規化による効果を確認することができる．特に，学習話者 1 名のときに構造サイズの正規化による効果がよく見られる．これは，学習話者が 1 名しかいない場合，評価話者との調音努力のミスマッチが大きくなってしまうため，その正規化による効果が表れたものと考えられる．また，ML 推定のときに効果が大きいことも分かる．これは，ML 推定の場合は抽出される構造が不安定であるため，構造サイズもばらついたのが原因であると考えられる．

表 3 構造サイズの正規化の効果
Table 3 The effect caused by normalization of the size of the structure.

学習話者=7 名

推定法 \ 帯域	full-band	4 kHz	2 kHz
ML（正規化なし）	29.4%	35.6%	83.9%
ML（正規化あり）	35.6%	43.2%	82.2%
MAP（n=14，正規化なし）	43.3%	70.4%	99.8%
MAP（n=14，正規化あり）	41.1%	63.7%	99.8%

学習話者=1 名

推定法 \ 帯域	full-band	4 kHz	2 kHz
ML（正規化なし）	11.6%	26.0%	68.8%
ML（正規化あり）	24.7%	47.9%	86.8%
MAP（n=14，正規化なし）	41.5%	47.9%	99.4%
MAP（n=14，正規化あり）	43.0%	62.8%	100.0%

4.3 従来手法との比較実験

従来手法による認識性能も求めた．従来の音響モデルとして，まずは 2 種類の不特定話者音響モデルを用意した．一つは学習話者 4,130 名の混合共有 HMM（ATR/BLA，triphone，混合数 256），もう一つは学習話者 260 名の状態共有 HMM （ASJ-JNAS，triphone，混合数 64）である（共に全帯域を用いて学習）[21]．更に，**4.1** の男性話者 1 名の音声データにカットオフ周波数 2 kHz の LPF を施し，それを用いて学習した音響モデル（混合数 1）も用意した（これを用意した理由については後述する）．この三つの音響モデルのいずれに対しても，CMN による話者・環境の正規化を行った．特徴パラメータとしては MFCC（1〜12 次元），ΔMFCC（1〜12 次元），及び ΔE を用いた（計 25 次元）．言語的制約としては，120 単語のみを許容する文脈自由文法を用いた．

表 4 に，4 種類の手法（提案手法，三つの音響モデルを用いた従来手法）による認識実験結果を示す．括弧内の数字は音響モデル（構造モデル）の学習時に使用した話者数である．提案手法においては，2 kHz までの低域成分をもつ入力音声であれば，LPF を通すことで，2 kHz 以上の高域成分を除去した構造統計モデルの条件と合わせることができる．したがって，LPF を特徴抽出の一部と考えれば，表 4 のすべての場合において認識率は 100%である．

全帯域を用いて学習された不特定話者音響モデルは，入力音声が全帯域の場合に 100%の認識率を実現しているが，LPF が施された入力音声に対しては，CMN を施しているにもかかわらず認識性能が劣化している．したがって，この認識タスクにおいては，1 人の話者で学習された提案手法が，4,130 人の話者で学習された従来手法より良い性能を示している．ただし，

表 4 クリーン環境下での四つの手法の認識性能
Table 4 Recognition performance of the four methods in the clean environment.

手法 \ 入力音声の帯域	full-band	4 kHz	2 kHz
full band HMM (260)	100.0%	93.8%	72.3%
full band HMM (4,130)	100.0%	95.2%	87.5%
limited band HMM (1)	88.8%	88.8%	88.8%
Proposed (1)	100.0%	100.0%	100.0%

より厳密な比較実験を行うには，従来の音響モデルを 2 kHz 以上の高域成分を除去した音声で学習させる必要がある．

これが，もう一つの音響モデルを用意した理由である．三つ目の音響モデルは 2 kHz までの帯域で学習されており，更には提案手法と同じ学習データを用いて作成されているため（同様に評価音声に対しては常に前処理として 2 kHz の LPF を行う），より厳密な比較実験であるといえる．この場合においても，提案手法は従来手法より上回る性能が得られた．この従来手法について，各評価話者に対する認識性能を調べたところ，すべての誤認識は 2 名の女性話者によるものであった（79.6%と 31.2%）．これは，カットオフ周波数 2 kHz の LPF では，話者性が完全には消失されないことを示唆する．**3.3** で述べたように，カットオフ周波数 2 kHz の LPF を通した音声を聞いてみると，話者性は確かに完全には消失されていないことが容易に分かる．その残りの話者性は構造化によって消失され，その結果として提案手法は 100%の性能を示した，と解釈できる．

5. 構造的表象を用いた雑音環境下での日本語孤立母音系列の認識実験

5.1 雑音による音声の構造的表象のひずみ

本節では，雑音環境下での日本語孤立母音系列を認識することを考える．音響的普遍構造は乗算性ひずみ・線形変換性ひずみに対して原理的に不変であるが，ここでは加算性雑音が構造に与える影響について考える．クリーンな音声，雑音，雑音下の音声のパワースペクトルをそれぞれ $|S(f)|^2$，$|N(f)|^2$，$|X(f)|^2$ とし，S と N の独立性を仮定すれば，

$$|X(f)|^2 \approx |S(f)|^2 + |N(f)|^2 \tag{12}$$

が成立する．式 (12) は対数パワースペクトル（$x(f) = \log|X(f)|^2$）上では，

$$x(f) \approx \log(\ \exp(s(f)) + \exp(n(f))\ ) \tag{13}$$

と表される．したがって，加算性雑音はケプストラムに対して非線形変換を施すため，音響的普遍構造はその形状がひずむものと予想される．図 10 は男性話者の 5 母音を，Ward 法によるボトムアップクラスタリングを用いて樹形図化したものである．このときの音響的条件を表 5 に示す．左はクリーンな音声の樹形図であり，右はこれに SNR=10 [dB] の白色雑音を加えたものである．構造サイズは正規化しているが，構造形状が雑音によってひずんでいる．特に/i/と/u/の距離が短くなっているが，これは/i/と/u/の第一フォルマントが互いに近傍にあり，他のフォルマントが雑音に埋もれたためと考えられる．

5.2 クリーンな構造統計モデルを用いた認識実験

前節において，雑音下では構造はその形状がひずむことを実験的に示したが，クリーン音声で学習された構造統計モデルを用いて雑音環境下の音声を認識する場合，入力音声の構造形状のひずみが原因で認識性能が低下することが予想される．これをより詳細に調べるため，以下の実験を行った．

評価音声としては **4.1** で使用した，8 名話者（男性 4 名，女性 4 名）が 5 母音を 5 回孤立発声したデータを用い，ここから各話者ごとに 3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の音声を得た．その各々に SNR=∞（clean），20, 10, 0 [dB] となるような白色雑音を重畳し，LPF（カットオフ周波数：2 kHz）を施した後，ML 推定または MAP 推定（$n = 10, 1, 0.1, 0.01$ のうち最適なもの）によって音声事象分布を得た．ここから入力構造ベクトル（計 $8 \times 5^5 = 25,000$ 個）を得た．また，入力音声における雑音の影響を軽減するため，LPF 後に SS（Spectral Subtraction）（$\alpha = 2.0$，$\beta = 0.5$）を行う場合も試みた．ここで α，β はそれぞれ over-estimation factor，flooring factor である [22]．また，雑音パワースペクトルの推定には 300 ms の白色雑音区間を用いた．構造統計モデルは，**4.1** で作成した学習話者 1 名のクリーンな構造統計モデルをそのまま使用した．音響的条件は表 6 のとおりである．

実験結果を表 7 に示す．予想どおり，雑音下では認識性能が劣化している．SS による性能改善も見られるが，クリーン環境の性能に及ぶまでには至っていない．低 SNR のときに MAP 推定による効果が見られなくなったのは，事前知識の推定をクリーンな環境で行っているため，雑音下の入力音声との間でミスマッチが生じたことが原因と考えられる．

5.3 雑音下の構造統計モデルを用いた認識実験

4.1 では，学習話者 1 名の構造統計モデルを用いて，クリーン環境における日本語孤立母音系列を 100%認識することに成功した．ここで，構造統計モデルの学習に必要な話者が 1 名で十分ならば，極めて高品質な音声合成器を用いて，入力音声の雑音環境と合致する音声を合成し，それをもとに構造統計モデル（及び事前知識）をオンラインで学習させることも可能である．これは構造に基づく音声知覚の運動理論 [23] と解釈することができるが，少なくとも人間は完璧な合成器を持っている．

ここでは，雑音下の構造統計モデルの性能を調べるため，入力音声の SNR が既知との仮定のもと，学習

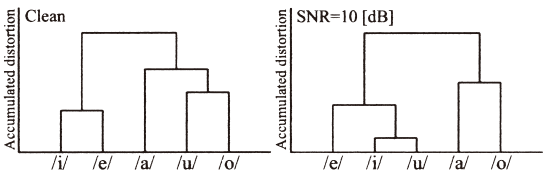

図 10 男性話者の 5 母音の樹形図
Fig. 10 Tree diagrams of 5 vowels of a male speaker.

表 5 分析実験における音響的条件
Table 5 Acoustic conditions in the analysis experiments.

サンプリング	16 bit / 16 kHz
窓	窓長 25 ms，シフト長 10 ms
パラメータ	FFTcep.（1〜12 次元）
音声事象分布	12 次元ガウス分布（対角共分散行列）
分布推定方法	MAP（n=14）
SNR	∞（clean），10 [dB]

表 6 雑音環境下での認識実験における音響的条件
Table 6 Acoustic conditions in the recognition experiments in the noisy environment.

サンプリング	16 bit / 16 kHz
窓	窓長 25 ms，シフト長 10 ms
音声特徴量	MCEP（α=0.55）（1〜12 次元）
音声事象分布	12 次元ガウス分布（対角共分散行列）
分布推定方法	ML or MAP（n=10, 1, 0.1, 0.01 のうち最適なもの）
SNR	∞（clean），20, 10, 0 [dB]

表 7 クリーンな構造統計モデルを用いた認識結果
Table 7 Recognition results using the statistical structure models trained in the clean environment.

	w/o SS		with SS	
SNR	ML	MAP	ML	MAP
∞	86.8%	100.0%	-	-
20 [dB]	49.2%	95.6%	64.8%	99.8%
10 [dB]	25.9%	33.0%	48.9%	43.0%
0 [dB]	5.4%	9.7%	8.0%	9.3%

表 8 雑音下の構造統計モデルを用いた認識結果
Table 8 Recognition results using the statistical structure models trained in the noisy environment.

	full-band		2 kHz	
SNR	ML	MAP	ML	MAP
∞	24.7%	70.3%	86.8%	100.0%
20 [dB]	73.9%	92.9%	67.9%	99.8%
10 [dB]	77.4%	99.1%	68.1%	86.7%
0 [dB]	73.9%	87.0%	71.1%	85.1%

話者 1 名の雑音下の構造統計モデル（及び事前知識）を用いた認識実験を行った．**4.1** で使用した，男性話者 1 名による 7 グループのデータ（1 グループにつき，5 母音 5 回孤立発声）に対して，各グループごとに 3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の音声を得た．その各々に，入力音声と同じ SNR となるよう白色雑音を重畳した．ここから，計 21,875（$= 7 \times 5^5$）個の/a/-/i/-/u/-/e/-/o/構造ベクトルを求め，構造統計モデルの学習に用いた．入力音声は **5.2** と同じものを使用した．ただし，SS は行っていない．また，本実験では白色雑音を重畳することで，図 9 に示すようにスペクトル高域成分を上にそろえることができるので，LPF を用いない場合（全帯域）についても試みた．音響的条件は表 6 のとおりである．

結果を表 8 に示す．表 7 よりはるかに良い認識性能が得られている．これは，入力構造と構造統計モデルとの間で雑音環境のミスマッチがなくなったためと考えられる．また，全帯域を用いた場合においては，クリーン環境より雑音環境下の方が高い認識率が得られ，低 SNR では LPF を施した場合より良い性能が得られている．これは，白色雑音を重畳することで，フォルマントの情報を保ちつつ，スペクトル高域成分をそろえることができたためと思われる．ただし，雑音レベルが非常に大きいとき（SNR=0 [dB]）においては，認識性能が劣化している．これは，音声が雑音に埋もれて音韻差異が不明りょうになったためと見られる．

5.4 従来手法との比較実験

SS（$\alpha = 2.0$，$\beta = 0.5$）を用いた従来手法との比較実験も行った．雑音パワースペクトルの推定には 300 ms の白色雑音区間を用いた．音響モデルは **4.3** で使用した，学習話者 4,130 名の混合共有 HMM，学習話者 260 名の状態共有 HMM の二つを用いた．CMN による話者・環境の正規化も行っている．

実験結果を表 9 に示す．提案手法の認識性能も併せて載せている．この際，提案手法では全帯域を用いた場合と 2 kHz の LPF を通した場合のうち，良い性能が得られた方を載せている．用いる帯域を雑音レベルに応じて使い分けることで，学習話者 1 名の提案手法が学習話者 4,130 名の従来手法（SS 及び CMN を適用）を上回る結果を得ていることが分かる．

表 9 雑音環境下での三つの手法の認識性能
Table 9 Recognition performance of the three methods in the noisy environment.

SNR	HMM (260)	HMM (4,130)	Proposed (1)
∞	100.0%	100.0%	100.0%
20 [dB]	100.0%	98.8%	99.8%
10 [dB]	94.3%	97.2%	99.1%
0 [dB]	83.0%	86.8%	87.0%

6. むすび

音声に不可避的に混入する話者の声道形状の特性，音響機器の特性などの非言語的特徴を表現する次元を保有しない音声の構造的表象が提案されている．本論文では，この構造を音声認識に利用することに関する基礎検討を行った．日本語孤立母音系列をタスクとする様々な認識実験を行い，その結果，クリーン環境下においては，音声事象分布の MAP 推定，及びスペクトル高域成分の均一化を施すことで，

- 音声の物理的実体を明示的に用いない音声認識
- 一人の話者で学習された音響モデル（構造モデル）を用いた不特定話者音声認識
- 適応・正規化技術が一切不要な音声認識

が 100%の認識性能を以って，いずれも実現可能であることを認識実験によって示した．従来手法との比較実験においても，学習話者 1 名の提案手法が学習話者 4,130 名の従来手法（CMN を適用）を上回る結果が得られた．雑音環境下においても，雑音下で学習した学習話者 1 名の提案手法が学習話者 4,130 名の従来手法（SS 及び CMN を適用）を上回る結果が得られた．[10] では孤立母音群から構成される構造を英語発音学習に利用しているが，本論文では孤立母音群を認識タスクとする提案手法の有効性を示した．

今後の検討課題としては，提案手法の連続音声認識への拡張，子音を含めた音声認識への拡張，提案手法と従来手法との融合などが挙げられる．現在，HMM 学習を用いた連続音声の構造化，及び，日本語連続母音系列を対象とした音声認識が検討されているので参照されたい [24].

文　献

[1] 篠田浩一，“確率モデルによる音声認識のための話者適応化技術，” 信学論（D-II），vol.J87-D-II, no.2, pp.371–386,

Feb. 2004.
[2] H.A. Gleason, An introduction to descriptive linguistics, Holt, Rinehart & Winston, New York, 1961.
[3] F. Saussure, Cours de linguistique general, publie par Charles Bally et Albert Schehaye avec la collaboration de Albert Riedlinge, Payot, Lausanne et Paris, 1916.
[4] R. Jakobson and J. Lotz , Notes on the French phonemic pattern, Hunter, N.Y., 1949.
[5] A. Gutkin and S. King, “Structural representation of speech for phonetic classification,” Proc. ICPR, vol.3, pp.438–441, 2004.
[6] T. Fukuda and T. Nitta, “Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition,” IEICE Trans. Inf. & Syst., vol.E87-D, no.5, pp.1110–1118, May 2004.
[7] L. Deng, G. Ramsay, and D. Sun, “Production models as a structural basis for automatic speech recognition,” Speech Commun., vol.33, no.2-3, pp.93–111, 1997.
[8] N. Minematsu, “Mathematical evidence of the acoustic universal structure in speech,” Proc. ICASSP, pp.889–892, 2005.
[9] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, “Theorem of the invariant structure and its derivation of speech Gestalt,” Proc. SRIV, pp.47–52, 2006.
[10] 朝川　智，峯松信明，広瀬啓吉，“音声の構造的表象に基づく英語学習者発音の音響的分析,” 信学論（D），vol.J90-D, no.5, pp.1249–1262, May 2007.
[11] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech Audio Process., vol.13, no.5, pp.930–944, 2005.
[12] C.J. Leggetter and P.C. Woodland, “Maximum likelihood speaker adaptation of continuous density hidden Markov models,” Comput. Speech Lang., vol.9, pp.171–185, 1995.
[13] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive training,” Proc. ICSLP, vol.2, pp.1137–1140, 1996.
[14] 峯松信明，志甫　淳，村上隆夫，丸山和孝，広瀬啓吉，“音声の構造的表象とその距離尺度,” 信学技報，SP2005-13, 2005.
[15] 峯松信明，“音声の音響的普遍構造のひずみに着眼した外国語発音の自動評定,” 信学技報，SP2003-180, 2004.
[16] N. Minematsu, S. Asakawa, and K. Hirose, “Paralinguistic information represented as distortion of the acoustic universal structure in speech,” Proc. ICASSP, vol.1, pp.261–264, 2006.
[17] 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. Signal Process., vol.39, no.4, pp.806–814, 1991.
[18] T. Kitamura and M. Akagi, “Speaker individualities in speech spectral envelopes,” J. Acoust. Soc. Jpn. (E), vol.16, no.5, pp.283–289, 1995.
[19] 音声信号処理ツールキット：http://www.sp.nitech.ac.jp/~tokuda/SPTK/index-j.html
[20] 今井　聖，音声認識，共立出版，1995.
[21] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, “Recent progress of open-source LVCSR engine Julius and Japanese model repository,” Proc. ICSLP, pp.3069–3072, 2004.
[22] 鹿野清宏，中村　哲，伊勢史郎，音声・音情報のディジタル信号処理，昭晃堂，1997.
[23] 柏野牧夫，“音声知覚の運動理論をめぐって,” 日本音響学会秋季講演論文集，1-2-10, pp.243–246, 2004.
[24] S. Asakawa, N. Minematsu, and K. Hirose, “Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics,” Proc. INTERSPEECH, pp.890–893, 2007.

（平成 19 年 4 月 24 日受付，8 月 19 日再受付）

村上　隆夫

2004 東大・工・電子情報卒．2006 同大大学院情報理工学系研究科修士課程了．同年（株）日立製作所入社．現在システム開発研究所に所属．

峯松　信明　（正員）

1990 東大・工・電気卒．1995 同大大学院工学系研究科博士課程了．博士（工学）．同年豊橋技術科学大学情報工学系助手．2000 東大大学院工学系研究科助教授．2004 同新領域創成科学研究科助教授．2002 瑞国 KTH 客員研究員．音声分析・認識・合成・応用，音声知覚，音声学・音韻論と，幅広い観点から音声コミュニケーションを研究．日本音響学会，情報処理学会，日本音声学会，人工知能学会等各会員．

広瀬　啓吉　（正員）

1972 東大・工・電気卒．1977 同大大学院工学系研究科博士課程了．工博．同年東大工学部電気工学科講師．1994 同電子工学科教授．1999 同大大学院新領域創成科学研究科教授．2004 年 10 月より同情報理工学系研究科教授．1987 米国 MIT 客員研究員．音声言語情報処理分野一般，特に韻律に着目した研究．IEEE，米国音響学会，ISCA，日本音響学会，情報処理学会，人工知能学会，言語処理学会等各会員．