社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声の構造的表象を用いた日本語母音系列の自動認識

村上　隆夫† 　丸山　和孝† 　峯松　信明†† 　広瀬　啓吉†

† 東京大学大学院情報理工学系研究科
†† 東京大学大学院新領域創成科学研究科
〒 113-0031 東京都文京区本郷 7-3-1
E-mail: †{murakami,maruyama,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　音声コミュニケーションは，音声の生成，収録，伝送，再生，聴取の何れの過程においても非言語的特徴が不可避的に混入するが，これを表現する次元そのものを保有しない音響的普遍構造が提案されている．これは，音声の物理的実体を捨象し，関係のみを捉えることで得られる音声の構造的表象であり，構造音韻論の物理実装として，あるいは音声ゲシュタルトとして解釈される．この構造的表象を用いて，孤立発声された日本語母音系列の音声認識を行なう実験を行なった．音声の物理的実体を知識として一切持たない音声認識についての試みであるが，認識実験の結果，本タスクにおいては，一人の話者で学習された適応・正規化無しの提案モデルが，4,130 人の話者で学習され，CMN（Cepstral Mean Normalization）による正規化が施された従来の音響モデルより良い性能を示した．
キーワード　音響的普遍構造，音声認識，日本語母音系列，最大事後確率推定，スペクトル高域成分除去

Automatic Recognition of Japanese Vowel Sequences Using Structural Representation of Speech

T. MURAKAMI†, K. MARUYAMA†, N. MINEMATSU††, and K. HIROSE†

† Graduate School of Information Science and Technology, University of Tokyo
†† Graduate School of Frontier Sciences, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0031 Japan
E-mail: †{murakami,maruyama,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　When we humans communicate with each other by means of speech, non-linguistic features are inevitably involved in every step of speech production, encoding, transmission, decoding, and hearing. Recently, a new acoustic representation of speech without any dimensions indicating the non-linguistic features was proposed. It captures only the interrelations among speech events and can be interpreted as physical implementation of structural phonology or as speech Gestalt. Recognition experiments of Japanese vowel sequences were carried out to investigate the performance of the new models. The results showed that the new models trained from a single speaker with no normalization can outperform the traditional models trained from 4,130 speakers with CMN.
Key words　acoustic universal structure, speech recognition, Japanese vowel sequences, maximum a posteriori estimation, upper-band spectrum removal

1. はじめに

我々が音声によるコミュニケーションを行なう際，その音声を生成する際には話者の声道形状の特徴，収録・伝送・再生の際にはその音響機器の特性，聴取の際には聴覚特性，といった非言語的特徴が音声に対して不可避的に混入する．従来の音声認識技術は，音響音声学に基づいて音声の物理的実体を捉えてきたが，この実体はこれら非言語的特徴によって不可避的に歪んだものである．それ故に，この歪みに対処するために沢山の話者，環境の音声によって学習された音響モデルが構築され，さらには話者及び環境に対する適応・正規化処理が施されてきたが，尚も認識性能を劣化させてしまう話者が存在する．音声の物理的実体に対する歪みの影響を受けることなく，音声を音響的に記述する方法は他に無いのだろうか．

言語学は音素に対して二通りの定義をしている [1]．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. 前者に基づくものとして，例えば不特定話者音響モデルが挙げられるが，後者に基づく音声認識は（少なくとも我々の知る限り）試みられていない．

近年，上記の非言語的特徴を表現する次元そのものを保有しない「音響的普遍構造」が提案された [2], [3]．これは，複数の音声事象の関係（距離）のみを捉えることで得られる構造的表象であり，音素の後者の定義に基づいている．さらには，この構造的表象は構造音韻論の物理実装として，あるいは音声ゲシュタルトとして解釈される [2], [3]．人間が音声コミュニケーションを行なう際にも，この構造的表象を利用していることが示唆されている [4]．本研究は，音響的普遍構造を利用した音声認識，即ち音素の後者の定義に基づく音声認識の実現を目的としている．ここでは，孤立発声された日本語母音系列を認識タスクとして，音声認識実験を行なった結果を報告する．

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

音声の物理的実体には非言語的特徴が混入し，これが従来の音声認識システムの性能低下を招いてきた．この非言語的特徴は主に加算性雑音，乗算性歪み，線形変換性歪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歪み，線形変換性歪みの二つである．加算性雑音とは，時間軸上の加算，即ち近似的にはスペクトルに対する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言える．これらは場所を移動するなどの対策を練ることで，物理的に抹消することができるので，不可避的な雑音ではない．本研究ではこの加算性雑音は扱う対象とはしない．

乗算性歪みは，スペクトルに対する乗算で表現される歪みであり，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．マイクロフォンなどの伝送特性がその典型例である．また，乗算性歪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ケプストラム平均正規化法（Cepstral Mean Normalization, CMN）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．即ち，話者の声道形状の違いの一部も近似的に乗算性歪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歪みである．

線形変換性歪みは，c に対する行列 A の乗算 $c' = Ac$ で表現される歪みである．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，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 の乗算で表されることが示されている [5]．即ち，声道長の差異，聴覚特性の差異は近似的に線形変換性歪みとして扱うことができる．これらも不可避的な歪みである．

以上をまとめると，音声の物理的実体には非言語的特徴が不可避的に混入し，これらはケプストラムベクトル c に対するアフィン変換 $c' = Ac + b$ で表現される．

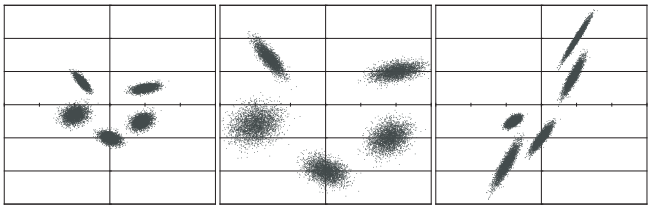

図 1　構造不変の定理
Fig. 1　Theorem of the invariant structure

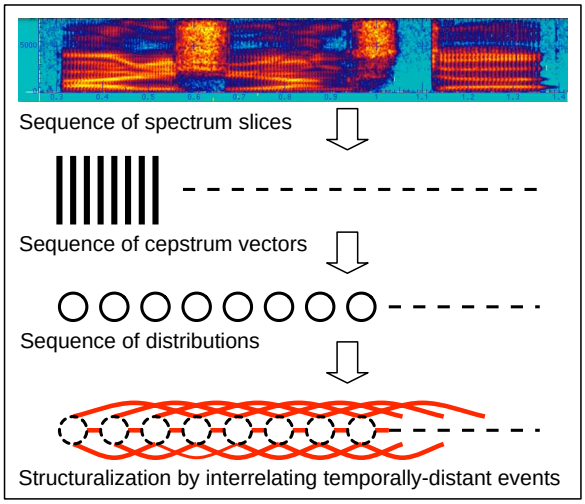

図 2　一発声の構造化
Fig. 2　Structuralization of a single utterance

2.2 音声に内在する音響的普遍構造

音声には非言語的特徴が不可避的に混入し，これはアフィン変換 $Ac + b$ で表現される．これに対して，様々な適応・正規化技術が提案されてきたが，何れの技術も，モデル（0）を入力音声（1）に近づける，もしくは入力音声（1）をモデル（0）に近づけるものに過ぎず，完全に 1 もしくは 0 にすることはできない．これは音声の物理的実体をそのままモデル化しているため，非言語的特徴を表現する次元が残留しているからである．近年提案されている音響的普遍構造は，非言語情報を表現する次元そのものを消失させる技術である．

各音声事象を分布化し，N 個の分布によって構成される構造を考える．N 個の分布に対して ${}_NC_2$ 個の全ての二分布間距離を求めれば，一つの構造を規定したことになるが，アフィン変換は構造を歪ませる変換であるため，不変な構造は「空間」を歪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，全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．即ち距離行列は不変であり，その結果，構造も不変となる（図 1 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．

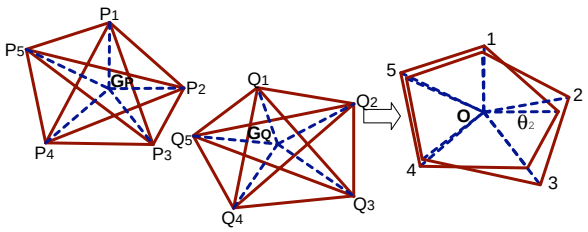

図 3　構造に基づく音響的照合

Fig. 3　Acoustic matching based on the strucuture

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

$0 \leqq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leqq 1$ 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，

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

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歪める距離尺度であることに起因する．MLLR [6] や SAT [7] では，話者性はアフィン変換で記述されるが，この構造はアフィン変換に対して不変であり，音響的普遍構造と呼ばれている．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　一発声の構造化とその音響的照合

音声認識は一発声された音声を対象として扱うが，音声から音声事象の分布系列を得た後，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 2）．

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（$P_1, ..., P_M$, $Q_1, ..., Q_M$）で構成される二つの構造において，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．即ち，図 3 に示されるような枠組みの音響的照合である．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．

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

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 $(\overrightarrow{OP_1}, ..., \overrightarrow{OP_M})$ であり，T は行列 $(\overrightarrow{OP_1}, ..., \overrightarrow{OP_M})$ である．

3.　音声の構造的表象を用いた音声認識

3.1　距離行列に基づく構造の音響的照合

音声の構造的表象に基づいた音声認識を考える．**第 2.3 節**で，一発声された入力音声を構造化した後，それを音響的に照合する方法について述べた．しかしながら，音響的普遍構造は空間を歪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．従って，三角不等式が満たされない可能性があり，直接式 (3) を用いることは出来ない．音素分布間の距離としてバタチャリヤ距離の平方根を用いた場合，

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

が式 (3) を近似することが示されている [8]．これは，距離行列の情報のみを基に，シフト（b）と回転（A）に基づく音響的照合を近似的に行なうことができることを意味する．

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．最尤（Maximum Likelihood; 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．従って，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（Maximum a Posteriori; MAP）推定を検討する．MAP 推定の具体的な枠組みに関しては [9] を参照した．以下，分散共分散行列は全て対角である．また，本報告では孤立発声された日本語母音系列を認識対象として扱う（詳細は**第 4.1 節**で後述する）．従って，ここでは各母音（/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．これらは一発声毎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通りである．

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 ：$\{\mu_i\}$ 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．

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

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

ここで，

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

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

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

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

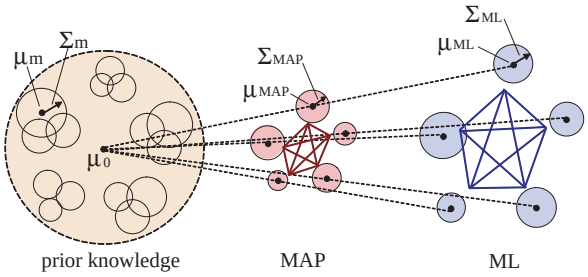

図 4　音声事象分布の最大事後確率推定

Fig. 4　MAP-based estimation of distributions of speech events

$$A = \Sigma_0^{-1} \tag{11}$$

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．本研究では各母音毎に中心前後 14 フレームが用いられたので，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 4 に示す．

3.3　スペクトル高域成分除去

本研究では，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．[10] は，母音のスペクトル包絡の 2.2kHz 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，音声にローパスフィルタを通すことでスペクトル広域成分除去を行なうことを試みた．

4.　日本語母音系列の音声認識実験

4.1　音声の構造的表象を用いた日本語母音系列音声認識

孤立的に発声された日本語母音系列を認識タスクとする認識実験を行なった．/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．音声の構造的表象を用いた日本語母音系列音声認識の枠組みを図 5 に示す．

入力音声の各母音の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．[11] は，構造サイズが調音努力（発話スタイル）を表すことを実験的に示している．従って，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，このうち意味を持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．

認識器に持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（全角分散行列を使用）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なう．これは，**第 3.1 節**において述べたシフト（b）と回転（A）に基づく音響的照合の近似を，構造と構造モデルの間で行なうことに相当する．この距離が最も小さい単語を認識結果とする．

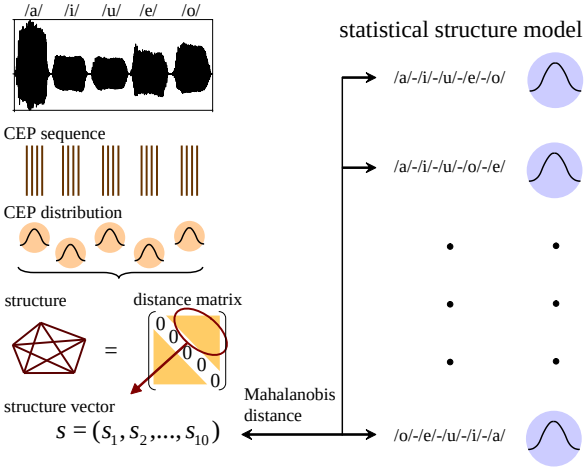

図 5　構造を用いた日本語母音系列の自動認識

Fig. 5　Automatic recognition of Japanese vowel sequences using the structure

4.2　全帯域を用いた構造的表象に基づく認識実験

前節で述べた枠組みを用いて，まずは構造化の際に全帯域を用いる認識実験を行なった．用いた音声資料は，男性 4 名，女性 4 名の計 8 名の話者が，それぞれ 5 母音を 5 回孤立的に発声した音声である．ケプストラムとしては MFCC（1〜12 次元）を使用し，各母音のケプストラム分布は中心前後 14 フレーム（140msec）を用いて推定した．分布の推定方法は，ML 推定，MAP 推定の両方を試みた．各話者毎に，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（$= 8 \times 3,125$）個の構造ベクトルを入力に用いた．また，各構造統計モデルの学習には，評価話者を除く 7 名によって作成される計 21,875 個（$= 7 \times 5^5$）の構造ベクトルを用い，混合数 1，2，4 のガウス分布を推定した．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学習話者の全母音音声（$7 \times 5 \times 5$ 個）を使用した．学習話者の音声事象に対しては，評価話者と当該話者を除く 6 名の全母音音声（$6 \times 5 \times 5$ 個）を使用した．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 として種々の値を使用して検討した（$n = \infty$ で ML 推定）．音響的条件を表 1 に示す．

実験結果を表 2 に示す．第 1 位以外に，2 位，4 位，5 位，10 位の正解率も示している．ML 推定の場合，認識率は 20%となっている．本実験では chance level は 0.8%（$= 1/120$）であることを考えると，ML 推定による音声事象分布群が成す構造にも語彙同定のための情報が含まれていることが分かる．一方，MAP 推定を施すことで認識率は飛躍的に向上する．特に，入力音声の事前知識に対する重み n を小さくするほど認識率が向上する．これは，各音声事象の少量サンプルデータを用いて，事前分布を僅かに修正することで得られる分布群が張る構造には，語彙同定のための情報が多く含まれていることを意味する

表 1　全帯域を用いた場合の音響的条件
Table 1　Acoustic conditions using the full-band spectrum

サンプリング	16bit/16kHz
窓長 / シフト長	25msec / 10msec
パラメータ	MFCC（1〜12 次元）
分布推定方法	ML or MAP

表 2　全帯域を用いた構造による認識結果
Table 2　Recognition results using the structure (full-band)

構造統計モデル＝単一ガウス分布

推定法＼順位	1	2	5	10
ML	19.7%	39.5%	69.3%	89.2%
MAP(n=10)	29.7%	55.7%	84.0%	96.8%
MAP(n=1)	35.9%	82.1%	96.6%	100.0%
MAP(n=0.1)	43.2%	88.5%	98.2%	100.0%
MAP(n=0.01)	53.0%	98.9%	100.0%	100.0%

構造統計モデル＝混合ガウス分布 (混合数 2)

推定法＼順位	1	2	5	10
ML	21.1%	39.3%	69.4%	87.3%
MAP(n=10)	29.1%	52.2%	80.8%	96.0%
MAP(n=1)	53.8%	84.0%	95.6%	100.0%
MAP(n=0.1)	59.1%	87.9%	99.3%	100.0%
MAP(n=0.01)	58.5%	97.2%	100.0%	100.0%

構造統計モデル＝混合ガウス分布 (混合数 4)

推定法＼順位	1	2	5	10
ML	19.1%	33.8%	62.4%	85.1%
MAP(n=10)	23.4%	39.9%	74.8%	93.5%
MAP(n=1)	43.7%	72.9%	93.5%	99.3%
MAP(n=0.1)	57.8%	92.1%	100.0%	100.0%
MAP(n=0.01)	60.9%	93.9%	100.0%	100.0%

(但し，$n=0$ にすると，事象分布が事前分布に等しくなり，距離行列の成分は全て 0 となり，認識不能になる)．構造統計モデルの混合数の増加は，認識結果にあまり影響を与えていない．これは，構造統計モデルには全角分散行列を用いたため，混合数 1 でも十分なモデル化能力が得られていたと考えられる．

4.3　1 母音を既知とした場合の認識実験

構造的表象のみを用いた認識実験では，音声の物理的実体を全く用いていない．そこで，5 母音のうち 1 母音を既知とした場合の認識実験を行なった．1 母音を既知とすることで，構造のシフト・回転の自由度を削減し，認識率は向上するものと予想される．結果を表 3 に示す．/u/または/e/を既知とした場合の認識率が高いが，これは誤認識結果として/a/-/i/-/e/-/u/-/o/が多かったことが原因である．表 4 に，/a/-/i/-/e/-/u/-/o/も正解に含めた場合の認識結果を示す．音声の実体を直接用いず，音声事象分布を推定する際に全帯域を用いた場合においても，ほぼ 100%の性能で語彙を 2/120 まで削減できている．

4.4　高域成分を除去した構造的表象に基づく認識実験

音声にローパスフィルタを通してスペクトル高域成分を除去した場合の実験も行なった．カットオフ周波数は 2kHz，4kHz，8kHz（全帯域）の 3 種類を試みた．ここでは，ケプストラムとして MCEP（1〜12 次元，$\alpha=0.55$）を用いた．また，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，構造統計モデルの学習に用いる話者は 1 名で済むはずである．従って，ここでは構造統計モデルの学習に，男性 1 名による 5 母音 35 回ずつの発声を用いた．これらを 7 つのグループ（1 グループにつき，5 母音 5 回発声）に分割した．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，これを基に構造統計モデルを作成した．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．このときの音響的条件を表 5 に示す．

構造統計モデルとして単一ガウス分布を用いた場合の認識結果を表 6 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 2kHz

表 3　1 母音を既知とした時の認識結果
Table 3　Recognition results with a single vowel given

構造統計モデル＝単一ガウス分布

推定法＼既知母音	a	i	u	e	o
ML	33.8%	27.5%	58.6%	55.6%	35.9%
MAP(n=10)	39.6%	35.8%	72.0%	72.0%	41.6%
MAP(n=1)	40.1%	36.7%	94.0%	94.9%	41.4%
MAP(n=0.1)	44.3%	43.3%	98.2%	98.1%	45.3%
MAP(n=0.01)	53.9%	53.2%	99.4%	99.6%	54.0%

構造統計モデル＝混合ガウス分布 (混合数 2)

推定法＼既知母音	a	i	u	e	o
ML	35.6%	30.8%	56.6%	55.5%	37.2%
MAP(n=10)	39.0%	35.7%	67.6%	70.9%	41.4%
MAP(n=1)	59.0%	55.7%	94.1%	95.0%	60.4%
MAP(n=0.1)	60.7%	59.1%	98.2%	97.7%	61.7%
MAP(n=0.01)	58.9%	58.6%	99.7%	99.8%	58.9%

構造統計モデル＝混合ガウス分布 (混合数 4)

推定法＼既知母音	a	i	u	e	o
ML	38.7%	30.7%	49.3%	46.2%	40.3%
MAP(n=10)	41.1%	34.0%	58.6%	60.1%	45.1%
MAP(n=1)	53.5%	48.8%	86.8%	90.5%	56.5%
MAP(n=0.1)	60.0%	58.5%	98.1%	97.2%	62.0%
MAP(n=0.01)	61.4%	61.2%	99.0%	99.9%	61.4%

表 4　/a/-/i/-/e/-/u/-/o/を含む認識結果
Table 4　Recognition results allowing for /a/-/i/-/e/-/u/-/o/

推定法＼混合数	1	2	4
ML	47.9%	46.5%	38.7%
MAP(n=10)	66.0%	59.8%	43.8%
MAP(n=1)	93.3%	91.3%	79.2%
MAP(n=0.1)	97.9%	97.3%	94.5%
MAP(n=0.01)	99.1%	99.6%	99.2%

表 5　ローパスフィルタを用いた場合の音響的条件
Table 5　Acoustic conditions using LPF

サンプリング	16bit/16kHz
窓長 / シフト長	25msec / 10msec
パラメータ	MCEP (1〜12 次元, α=0.55)
分布推定方法	ML or MAP
カットオフ周波数	2kHz, 4kHz, or full-band

表 6　ローパスフィルタを用いた構造による認識結果
Table 6　Recognition results using the structure (LPF)

	full-band	4kHz	2kHz
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%

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，今回の認識タスクにおいては，

- 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）を用いた音声認識
- 適応・正規化技術を一切用いない音声認識

がいずれも実現可能であることを意味する．

4.5　従来手法との比較実験

比較実験のための従来の音響モデルとして，2 種類の不特定話者音響モデルを用いた．一つは学習話者 4,130 名の混合共有 HMM，もう一つは学習話者 260 名の状態共有 HMM であり，両者とも CMN による話者・環境の正規化を行なった．特徴パラメータとしては全帯域の 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 7 に，3 種類の手法（提案手法，2 種類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，2kHz までの低域成分を持つ入力音声であれば，ローパスフィルタを通すことで，2kHz 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．従って，ローパスフィルタを特徴抽出の一部と考えれば，表 7 の全ての場合において認識率は 100%である．従来手法では，入力音声が全帯域の場合，100%の認識率を実現しているが，ローパスフィルタが施された入力音声に対しては，CMN を施しているにも拘らず認識性能が劣化している．従って，今回のタスクにおいては，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た，と結論づけられる．但し，より厳密な比較実験を行なうには，従来の音響モデルを 2kHz 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．

5. ま と め

音声コミュニケーションは，生成，収録，伝送，再生，聴取の各過程において非言語的特徴が不可避的に混入するが，これを表現するアフィン変換による歪みを次元として保有しない音響の構造的表象が，近年提案されていることを述べた．ここでは，この構造的表象を用いた日本語母音系列音声認識システムを構築した．認識実験を行ない，結果として，

表 7　3 つの手法の性能比較
Table 7　Comparison of the performance among the three methods

methods	full-band	4kHz	2kHz
HMM(260)	100.0%	93.8%	72.3%
HMM(4,130)	100.0%	95.2%	87.5%
Proposed(1)	100.0%	100.0%	100.0%

- 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）を用いた音声認識
- 適応・正規化技術を一切用いない音声認識

が，今回の認識タスクにおいては，100%の認識性能を以って実現可能であることを示した．今後は，子音を含めた連続音声の構造化，加算性雑音に対して頑健な構造の抽出，音声の物理的実体をモデリングする従来手法との融合などについて検討を行なっていく予定である．

文　献

[1] H. A. Gleason, “An introduction to descriptive linguistics,” New York: Holt, Rinehart & Winston (1961)

[2] 峯松信明, 西村多寿子, 西成活裕, 櫻庭京子, “構造不変の定理とそれに基づく音声ゲシュタルトの導出”, 電子情報通信学会音声研究会, SP2005 (2005-5, 発表予定)

[3] 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”, 電子情報通信学会音声研究会, SP2005 (2005-5, 発表予定)

[4] 峯松信明, 松井健, 広瀬啓吉, “構造音韻論の物理実装に基づく新しい音声の音響的表象”, 電子情報通信学会音声研究会, SP2004-27, pp.47-52 (2004-6)

[5] M. Pitz and H. Ney, “Vocal tract normalization as linear transformation of MFCC,” Proc. EUROSPEECH, pp.1445-1448 (2003)

[6] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171-185 (1995)

[7] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, “A compact model for speaker-adaptive training,” Proc. ICSLP, vol.2, pp.1137-1140 (1996)

[8] N. Minematsu, “Yet another acoustic representation of speech sounds,” Proc. ICASSP, pp.585–588 (2004)

[9] 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 Processing, vol.39, no.4, pp.806-814 (1991)

[10] T. Kitamura and M. Akagi, “Speaker individualities in speech spectral envelopes,” JASJ(E), Vol.16, No.5 (1995)

[11] N. Minematsu, S. Asakawa and K. Hirose, “The acoustic universal structure in speech and its correlation to para-linguistic information in speech,” Proc. IWMMS’2004, pp.69-79 (2004)