社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59 (41)
2005／12／22

音声の構造的表象を用いた雑音環境下における日本語母音系列の自動認識

村上　隆夫† 　丸山　和孝† 　朝川　智†† 　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,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

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が，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，関係のみを捉えることによって得られる音声の構造的表象である．本稿では，まず雑音環境下での音響的普遍構造に関する分析実験を行なった．その結果，加算性雑音によって音響的普遍構造の形状が歪むものの，スペクトル高域成分を均一化させることで，話者性がさらに消失されることが示された．次に，加算性雑音下での日本語母音系列の認識実験を行ない，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，音声認識，日本語母音系列，加算性雑音，スペクトル高域成分の均一化

Automatic recognition of Japanese vowel sequences in noise using structural representation of speech

T. MURAKAMI†, K. MARUYAMA†, 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,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Non-linguistic features such as vocal tract shapes and acoustic devices are inevitably involved in speech. Recently, a new representation of speech without any dimensions indicating the non-linguistic features was proposed. It discards the absolute properties of speech events and captures only the interrelations among them. In this paper, first, analysis experiments of the representation in noise were conducted. The results showed that though additive noise distorts the representation, it can remove much of speaker individuality by modifying the upper-band spectrum to be uniform. In the next, recognition experiments of Japanese vowel sequences in noise were done. The results showed that the proposed method trained from a single speaker in the matched condition can outperform the conventional method trained from 4,130 speakers with SS and CMN.

Key words　structural representation of speech, speech recognition, Japanese vowel sequences, additive noise, uniform upper-band spectrum

1. はじめに

音声には，その生成の際に話者の声道形状の特性，伝送・収録の際には音響機器の特性，さらには聴取の際には聴取者の聴覚特性，といった非言語的特徴が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，この実体は上記の非言語的特徴によって不可避的に歪んだものである．このため，不特定話者モデルに代表されるデータ集めによる解決策は，限界があるものと考えられる．

言語学は音素に対して以下の二つを定義している [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. 不特定話者モデルは 1) に基づく技術である．音素を弁別素性の束，即ち構造とみなし，弁別素性に着眼した研究例 [2]〜[4] もあるが，音声事象の絶対的特性を捉えているという点において，これもやはり 1) に基づくものと言える．これに対して近年，冒頭で述べた非言語的特徴を原理的に保有しない音声表象として「音響的普遍構造」[5], [6] が提案されたが，これは音声事象同士の関係を捉えることで得られる幾何学的構造であり，2) に基づくものである．本研究は，音響的普遍構造の音声認識への利用を目的としている．

筆者らは既に孤立発声された日本語母音系列という簡単な認識タスクにおける実験を行ない，結果，男性 1 名で学習された提案手法が，(2kHz の LPF を適用することで) 100%の認識率を実現することに成功した [7]．そのときの音声はクリーン音声であったが，本稿では雑音環境下での音声認識を考える．まず，雑音下では話者性がさらに消失されることに着眼し，その消失度合いについての定性的・定量的分析を行なう．次に，雑音下における日本語母音系列の認識実験を行ない，SS (Spectral Subtraction) を用いた従来手法との比較実験を行なう．

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，加算性雑音・乗算性歪み・線形変換性歪みの 3 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．加算性雑音は，時間軸上の加算で表現される雑音であり，背景雑音がその典型例であるが，これは場所を移動するなどの対策によって，物理的に抹消することが可能なので，原理的には不可避的ではないと考えられる．

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 $c' = c + b$ で表現される．音響機器の特性がその典型例である．また，CMN (Cepstral Mean Normalization) によって，話者性の違いによる影響も軽減できることを考慮すると，話者の声道形状の違いの一部も乗算性歪みであると考えられる．音声は必ずある話者・ある音響機器によって発声・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，さらには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 の乗算で表すことができる [8]．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．

以上より，音声に不可避的に混入する非言語的特徴は，c に対するアフィン変換 $c' = Ac + b$ で近似的に表現される．

2.2 音声に内在する音響的普遍構造

近年提案されている音響的普遍構造は，上述したアフィン変換 $Ac + b$ に対して原理的に不変な音声表象である．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求めることを考える．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．不変な構造は「空間」を歪ませることで抽出される．

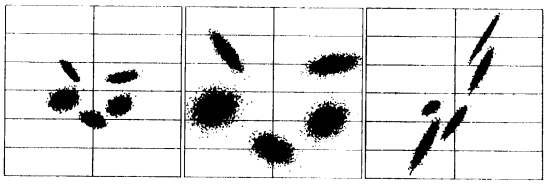

図 1 構造不変の定理
Fig. 1 Theorem of the invariant structure

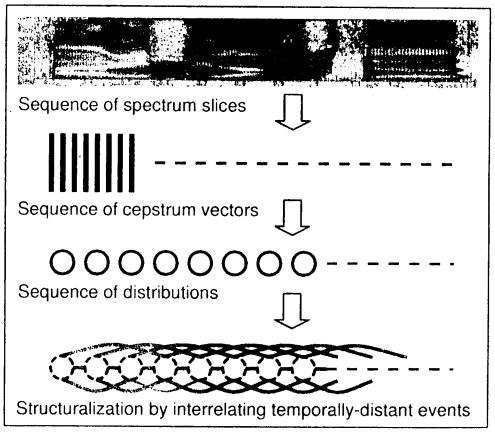

図 2 一発声の構造化
Fig. 2 Structuralization of a single utterance

> 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）．

バタチャリヤ距離で話を進める．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，

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

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

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

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

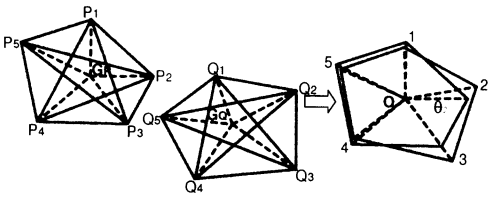

図 3　構造に基づく音響的照合

Fig. 3　Acoustic matching based on the strucuture

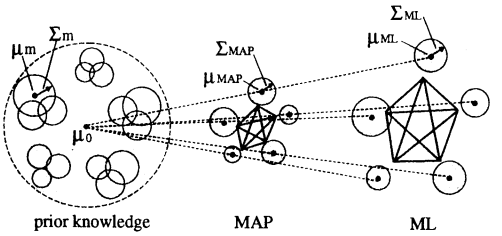

図 4　音声事象分布の最大事後確率推定

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

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．ここから，バタチャリヤ距離は空間を歪める距離尺度であることが分かる．MLLR [9] や SAT [10] 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となる．これが音響的普遍構造と呼ばれている構造である．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．

3.　音声の構造的表象を用いた音声認識

3.1　一発声の構造化とその音響的照合

音声の構造的表象を用いた音声認識を考える．まず，一発声された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（即ち距離行列）を求めることで音声を構造化する（図 2）．

次に，M 個の頂点（$P_1, ..., P_M, Q_1, ..., Q_M$）で構成される二つの構造のうち，一方をシフト（$b$）と回転（$A$）のみでもう一方に近づけることで音響的照合を行なうことを考える（図 3）．このときの構造間差異を，対応する頂点間距離の和（$\sum_{i=1}^{M} \overline{P_iQ_i}^2$）の最小値として定義することにする．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，

$$\sqrt{\sum_{i<j}(\overline{P_iP_j} - \overline{Q_iQ_j})^2} \tag{3}$$

は上記の構造間差異を近似することが示されている [6]．式 (3) は，距離行列のうち意味を持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相当する．

3.2　音声事象分布の最大事後確率推定

一発声された音声から音声事象分布を推定するにあたって，データ量 n が少ないために最尤（Maximum Likelihood; ML）推定では不適切な分布を推定する可能性がある．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 [11] を参照した．以下，分散共分散行列は全て対角である．また，本研究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 **3.4** 節）ので，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 ： $\{\mu_m\}$ 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．

$$\mu_{MAP} = \hat{\mu_0} \tag{4}$$

$$\Sigma_{MAP} = \hat{B}\hat{A}^{-1} \tag{5}$$

ここで，

$$\hat{\mu_0} = \Omega(\Omega + nE)^{-1}\mu_0 + n(\Omega + nE)^{-1}\mu_{ML} \tag{6}$$

$$\hat{B} = B + \frac{n}{2}\Sigma_{ML} + \frac{n}{2}\Omega(\mathrm{DIAG}(\mu_{ML} - \mu_0))^2(\Omega + nE)^{-1} \tag{7}$$

$$B = E \tag{8}$$

$$\hat{A} = A + \frac{n}{2}E \tag{9}$$

$$A = \Sigma_0^{-1} \tag{10}$$

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り，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恐れがある．[7] では，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 [12] ことを踏まえて，音声に LPF を通すことによるスペクトル高域成分の均一化を行なっている．本稿ではさらに，白色雑音の付与によるスペクトル高域成分の均一化も試みる（詳細は第 **4**，**5** 章）．

3.4　構造を用いた日本語母音系列音声認識の枠組み

本研究では孤立発声された日本語母音系列を認識タスクとす

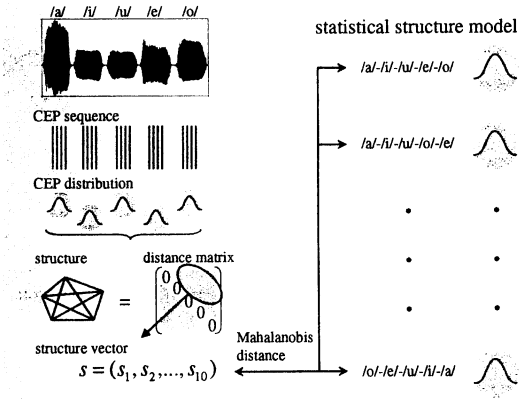

図 5 構造を用いた日本語母音系列の自動認識

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

る．各母音は一回ずつ出現する（語彙サイズは ${}_5P_5 = 120$）．このような認識タスクを構造のみ用いて認識する枠組みを図 5 に示す．まず，入力音声から距離行列を求めることで音声を構造化する．この際，構造サイズは調音努力を表し [13]，また，雑音下では構造サイズが小さくなる（詳細は第 **4.3** 節）ので，これを正規化する．特徴量として用いるのは，第 **3.1** 節で述べた構造ベクトル（10 次元）である．次に，認識器に持たせる構造モデルを以下のようにして得る．複数の/a/-/i/-/u/-/e/-/o/構造ベクトルから 10 次元ガウス分布を求め，これを「構造統計モデル」として使用する．他の 119 個のモデルは/a/-/i/-/u/-/e/-/o/モデルの要素を交換することで得られる．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のマハラノビス距離を求めることで行なう．尚，本タスクにおいては，マハラノビス距離の最小値は尤度の最大値に対応する．

4. 雑音環境下における音声の構造的表象

4.1 雑音による音声の構造的表象の歪み

音響的普遍構造は乗算性歪み・線形変換性歪みに対して原理的に不変であるが，本稿では雑音環境下の認識実験を行なうため，加算性雑音が音響的普遍構造に与える影響について考えてみる．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 $|X(f)|^2$，$|S(f)|^2$，$|Y(f)|^2$ とし，

$$|Y(f)|^2 \approx |X(f)|^2 + |S(f)|^2 \tag{11}$$

が成立すると仮定する．式 (11) は対数パワースペクトル $(y(f) = \log|Y(f)|^2)$ 上においては，

$$y(f) \approx \log(\exp(x(f)) + \exp(n(f))) \tag{12}$$

と表される．従って，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歪むものと予想される．図 6 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，分析条件は表 1 のとおりで，MAP 推定 $(n = 14)$ を用いて音声事象分布を求めている．左はクリーンな音声の樹形図

表 1 分析実験における音響的条件

Table 1 Acoustic conditions in the analysis experiments

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	FFTcep.（1〜12 次元）

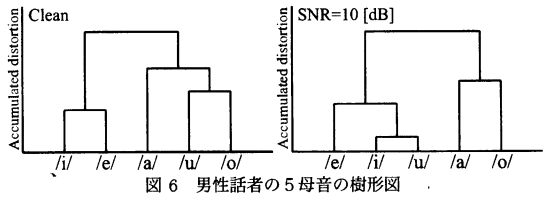

図 6 男性話者の 5 母音の樹形図

Fig. 6 Tree diagrams of 5 vowels of a male speaker

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歪んでいる．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．

4.2 雑音による話者性の消失に関する定性的分析

[7] では音声に LPF を通し，話者性が多く含まれるスペクトル高域成分を下に揃えることで，日本語母音系列の認識性能を向上させた．スペクトル高域成分を音素間で均一化させる別の方法は，上に揃える方法である．これは加算性雑音の重畳によって実現される．図 7 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：2kHz）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．話者による違いが高域によく表れているが，白色雑音の重畳により，それが上に揃えられ，話者性の消失が効果的に行なわれていることが分かる．但し，実際に図 7 中央及び下の音声を聞いてみたところ，話者性は完全には消失されていなかった．

4.3 雑音による話者性の消失に関する定量的分析

雑音による話者性の消失について定量的に調べるため，雑音を付与した場合の話者間構造差異・話者内構造差異の分散分析を行なった．8 名話者（男性 4 名，女性 4 名）が 5 母音を 5 回発声したデータを音声資料として用いた．ここから各話者毎に 5 個の/a/-/i/-/u/-/e/-/o/の音声を得た．これに SNR $= \infty$ (clean)，20, 10, 0[dB] の白色雑音を重畳し，表 1 に示す分析条件でケプストラムを求め，ML 推定 or MAP 推定 $(n = 14)$ によって分布化した．求めた音声事象分布から，各話者毎に構造を抽出した．この際，全構造のサイズが等しくなるように正規化を施す場合と，施さない場合の 2 通りを試みた．抽出した構造から，話者間構造差異 $({}_8C_2 \times 5^5 = 700$ 個)，及び話者内構造差異 $(8 \times {}_5C_2 = 80$ 個) を求め，分析対象とした．

分散分析の結果，どの場合でも危険率は $p < 0.001$ で話者間差異の方が話者内差異より大きくなった．これは，話者特有の発声の癖，方言差などの「構造の話者性」が消失し得ないことが原因と見られる．話者間差異の平均，話者内差異の平均を表 2 に示す．括弧内の数値は，構造サイズを正規化した場合の結果である．ML 推定の場合も MAP 推定の場合も，以下のよ

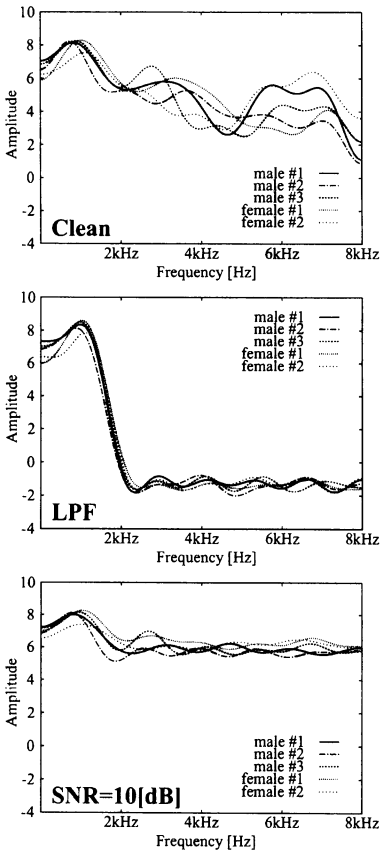

図 7　5名話者の/a/のスペクトル包絡

Fig. 7　Spectral envelopes of /a/ of 5 speakers

うな傾向が見られた．まず，構造サイズを正規化しない場合，SNR の低下とともに構造サイズが小さくなり，その結果，話者間差異・話者内差異も減少している．そこで，構造サイズを正規化した場合を見ると，この場合でも雑音の重畳によって話者間差異・話者内差異が減少しているのが分かる．これは，雑音を重畳することで「音の話者性」の消失，及び話者内の発声の揺れ（即ち「構造揺らぎ」）の抑制が行なわれている効果と見られる．但し，SNR の低下に伴い，話者間差異・話者内差異が増加している．これは低 SNR においては，音韻差異が不明瞭に（構造サイズが非常に小さく）なり，構造形状が不安定になったことが原因と見られる．

5.　雑音環境下での日本語母音系列の認識実験

5.1　クリーンな構造統計モデルを用いた認識実験

第 **4.1** 節において，雑音下では構造はその形状が歪むことを示したが，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歪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，以下の実験を行なった．

音声資料は第 **4.3** 節と同じものを用いた．ここから各話者毎に 3,125 ($=5^5$) 個の/a/-/i/-/u/-/e/-/o/の音声を得た．そ

表 2　分析結果（括弧内は構造サイズの正規化有り）

Table 2　Analysis results (Parenthesized data are obtained with the size normalization)

音響事象分布の推定法= ML 推定

SNR[dB]	話者間差異の平均	話者内差異の平均	構造サイズ
∞	1.47 (1.17)	0.99 (0.86)	12.5 (12.5)
20	0.60 (0.81)	0.36 (0.58)	6.7 (12.5)
10	0.43 (0.85)	0.25 (0.63)	4.3 (12.5)
0	0.25 (0.95)	0.17 (0.73)	2.4 (12.5)

音響事象分布の推定法= MAP 推定

SNR[dB]	話者間差異の平均	話者内差異の平均	構造サイズ
∞	1.13 (0.93)	0.61 (0.55)	11.5 (11.5)
20	0.56 (0.66)	0.25 (0.40)	6.3 (11.5)
10	0.40 (0.74)	0.19 (0.49)	4.0 (11.5)
0	0.22 (0.90)	0.14 (0.64)	2.0 (11.5)

の各々に SNR=∞ (clean)，20, 10, 0[dB] の白色雑音を重畳し，LPF（カットオフ周波数：2kHz）を施した後，ML 推定 or MAP 推定（n = 10, 1, 0.1 or 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 8×5^5 = 25,000 個）を得た．また，雑音による影響を軽減するため，LPF 後に SS (α = 2.0, β = 0.5) を行なう場合も試みた．その際，雑音パワースペクトルの推定には 300ms の白色雑音区間を用いた．構造統計モデルは，評価話者を除く 7 名によるクリーン音声から得た計 21,875 ($=7\times5^5$) 個の/a/-/i/-/u/-/e/-/o/構造ベクトルを用いて学習させた．このときの音響的条件を表 3 に示す．

実験結果は表 4 のとおりである．予想通り，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，事前分布の推定をクリーンな環境で行なっているため，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．

5.2　雑音下の構造統計モデルを用いた認識実験

[7] では，男性話者 1 名で学習した構造統計モデルを用いて，クリーン環境における日本語母音系列を 100%認識したが，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，評価音声の雑音環境と合致する音声を合成し，それを基に構造統計モデル（及び事前知識）をオンラインで学習させることも可能と考えられる．これは構造に基づく音声知覚の運動理論 [14] 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，評価音声の SNR が既知との仮定のもと（即ち，matched な条件），学習話者男性 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なった．男性話者が 5 母音を 35 回発声したデータを 7 つのグループに分け，各グループ毎に 3,125 ($=5^5$) 個の/a/-/i/-/u/-/e/-/o/の音声を得た．その各々に，評価音声と同じ SNR となるよう白色雑音を重畳した．ここから，計 21,875 ($=7\times5^5$) 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．音響的条件

表 3　認識実験における音響的条件

Table 3　Acoustic conditions in the recognition experiment

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	MCEP（α=0.55）（1〜12 次元）
分布推定方法	ML or MAP

表 4　クリーンな構造統計モデルを用いた認識結果

Table 4　Recognition results using the clean structure models

	w/o SS		SS	
SNR	ML	MAP	ML	MAP
∞	82.9%	99.9%	-	-
20[dB]	55.0%	98.5%	68.7%	99.9%
10[dB]	38.5%	39.5%	57.0%	52.8%
0[dB]	12.7%	12.4%	18.2%	13.1%

は表 3 と同じであるが，白色雑音を重畳することで，スペクトル高域成分を揃えることができるので，LPF を用いない場合（full band）についても試みた．また，SS は行なっていない．

結果を表 5 に示す．表 4 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無くなったためと考えられる．また，full band の場合においては，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揃えることができたためと思われる．但し，雑音レベルが非常に大きいとき（SNR=0[dB]），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瞭になったためと考えられる．

5.3　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なった．雑音パワースペクトルの推定には 300ms の白色雑音区間を用いた．音響モデルは，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 2 通りの不特定話者モデルを用いた．特徴量は，全帯域の MFCC（1〜12 次元），Δ MFCC（1〜12 次元），及びΔ E であり（計 25 次元），CMN による話者・環境の正規化も行なった．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．

実験結果を表 6 に示す．提案手法では，full band の場合と 2kHz の場合のうち良い性能が得られた方を載せている．括弧内の数値は，学習話者数である．低 SNR において，提案手法はいずれの従来手法よりも良い性能を得ていることが分かる．

表 5　雑音下の構造統計モデルを用いた実験結果

Table 5　Recognition results using the noisy structure models

	full band		2kHz	
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%

表 6　3 つの手法の認識性能

Table 6　Recognition performance of the three methods

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. ま と め

本稿では，まず加算性雑音における音響的普遍構造の特性を調べるために分析実験を行ない，雑音下では構造はその形状が歪むものの，スペクトル高域成分を均一化させることで，話者性をさらに消失させることができることを示した．次に，日本語母音系列を認識タスクとして雑音下の認識実験を行なった．その結果，入力音声と同じ SNR の音声で学習した構造統計モデル（及び事前知識）を用いることで，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）を上回る結果が得られた．これは，十分高品質な音声合成器があれば，オンラインで構造統計モデルを作成し，高い認識性能を得ることができることを示唆する．今後は子音を含めた連続音声認識，及び従来手法との融合を検討する予定である．

文　献

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

[2] A. Gutkin *et al.*, "Structural representation of speech for phonetic classification," Proc. ICPR, vol.3, pp.438-441 (2004)

[3] T. Fukuda *et al.*, "Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition," IEICE transactions, vol.E87-D, no.5, pp.1110-1118 (2004)

[4] L. Deng *et al.*, "Production models as a structural basis for automatic speech recognition," Speech Communication, vol.33, no.2-3, pp.93-111 (1997)

[5] 峯松信明他，"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"，信学技報，SP2005-12，pp.1-8 (2005-5)

[6] 峯松信明他，"音声の構造的表象とその距離尺度"，信学技報，SP2005-13，pp.9-12 (2005-5)

[7] 村上隆夫他，"音声の構造的表象を用いた日本語母音系列の自動認識"，信学技報，SP2005-14，pp.13-18 (2005-5)

[8] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in Cepstral space," IEEE Trans. Spech Audio Processing , vol.13, no.5, pp. 930-944 (2005)

[9] C.J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, pp.171-185 (1995)

[10] T. Anastasakos, *et al.*, "A compact model for speaker-adaptive training," Proc. ICSLP, vol.2, pp.1137-1140 (1996)

[11] C.H. Lee, *et al.*, "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)

[12] T. Kitamura *et al.*, "Speaker individualities in speech spectral envelopes," JASJ(E), Vol.16, No.5 (1995)

[13] N. Minematsu, *et al.*, "The acoustic universal structure in speech and its correlation to para-linguistic information in speech," Proc. IWMMS'2004, pp.69-79 (2004)

[14] 柏野牧夫，"音声知覚の運動理論をめぐって"，音講論，1-2-10，pp.243-246 (2004)