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

2006－SLP－64 (21)
2006／12／21

音声の構造的表象に基づく日本語母音系列連続発声の認識

朝川 智† 村上 隆夫†† 峯松 信明† 広瀬 啓吉†††

† 東京大学大学院新領域創成科学研究科 〒277-8561 千葉県柏市柏の葉 5-1-5
†† 日立製作所 〒100-8280 東京都千代田区丸の内 1-6-6
††† 東京大学大学院情報理工学系研究科 〒113-0033 東京都文京区本郷 7-3-1
E-mail: †{asakawa,murakami,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，連続的に発声された日本語母音系列を認識タスクとして構造的表象に基づく音声認識を検討する．連続発声の分布系列化の際には HMM の学習を利用し，パラメータ推定手法として ML 推定のほかに MAP 推定およびベイズ推定の各手法を比較する．さらに，連続発声構造間の照合時に生じる，音響事象と HMM の状態との対応付けのズレの問題および状態数の異なる構造間照合の問題に対して，DP に基づく構造間比較手法を提案し，その有効性を実験的に検討する．
キーワード 音声の構造的表象，音声認識，日本語母音系列，DP マッチング

Recognition of continuous utterances of Japanese vowel sequences based on structural representation of speech

Satoshi ASAKAWA†, Takao MURAKAMI††, Nobuaki MINEMATSU†, and Keikichi HIROSE†††

† Grad. School of Frontier Sciences, Univ. of Tokyo, 5-1-5, Kashiwanoha Kashiwa, Chiba, 277-8562 Japan
†† Hitachi, Ltd., 1-6-6, Marunouchi, Chiyoda-ku, Tokyo 100-8280 Japan
††† Grad. School of Info. Sci. and Tech., Univ. of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033 Japan
E-mail: †{asakawa,murakami,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, recognition experiments of continuous utterances of Japanese vowel sequences using stractural representation of speech are carried out. Distribution sequences are obtained from continuous utterances by training HMMs, and ML-based, MAP-based and Bayes-based methods are investigated for the estimation of the HMM parameters. Furthermore, to solve problems in continuous utterances recognition, a DP-based structural matching method is proposed and examined in recognition experiments.
Key words structural representation of speech, speech recognition, Japanese vowel sequences, DP matching

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，即ち音素の内部構造に着眼した研究例 [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] が提案された．これは音声事象同士の関係を捉えることで得られる構造的表象であり，2) に基づくものである．本研究は，この音声の構造的表象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 [6]．

本稿では，認識タスクを連続発声へと拡張する．この認識タスクでは連続発声から分布系列を推定する必要があるが，本研究では分布系列推定に HMM を利用する．一発声という非常に限られたデータ量で HMM の分布パラメータを推定する必要があるため，ML 推定だけでなく，MAP 推定および変分ベイズ法に基づくパラメータ推定を行い，各手法の比較検討する．さらに，連続発声構造間の照合時に生じる音響事象と HMM の状態との対応付けのズレの問題，および状態数の異なる構造間の照合の問題に対して，DP に基づく構造間比較手法を提案し，その有効性を実験的に検討する．

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混入する歪みとしては，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としても表現されうる雑音であり，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 $\boldsymbol{c}$ に対するベクトル $\boldsymbol{b}$ の加算 $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{b}$ 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，CMN (Cepstral Mean Normalization) によって，話者性の違いの影響を軽減できることから，話者の声道形状の違いの一部も近似的に乗算性歪みであると考えられる．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といえる．

線形変換性歪みは，$\boldsymbol{c}$ に対する行列 $\boldsymbol{A}$ の乗算 $\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c}$ 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，$\boldsymbol{c}$ に対する $\boldsymbol{A}$ の乗算で表されることが示されている [7]．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 $\boldsymbol{c}$ に対するアフィン変換 $\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}$ で表現される．図 1 は，アフィン変換 $\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}$ が対数スペクトルに与える影響を示したものである．対数スペクトルの水平変化 (フォルマントシフト) は $\boldsymbol{A}$ の乗算として，垂直変化は $\boldsymbol{b}$ の加算として記述され，これら $\boldsymbol{A}$, $\boldsymbol{b}$ が話者や収録環境ごとに変化し，様々な歪みを生む．

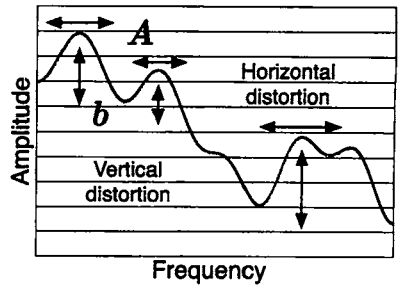

図 1 スペクトルの水平・垂直歪みとアフィン変換

2.2 音声に内在する音響的普遍構造

空間内に存在する N 点に対して，${}_N\mathrm{C}_2$ 個の全ての二分布間距離を求めると，その N 点で張られる構造は一意に規定される．即ち事象群に対して，全ての二事象間距離を求めると，その事象群を構造的に表象することになる．ケプストラム空間内の N 点に対して構造を考えた場合，その構造は非言語的特徴によって不可避的に歪む．何故なら，非言語的特徴はアフィン変換としてモデル化されるからである．この不可避的に歪む構造は，空間を歪ませることによって不変な構造として定義可能となる．

> 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．

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 $p_1(\boldsymbol{x})$, $p_2(\boldsymbol{x})$ とすると，バタチャリヤ距離は以下の式で表される．

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

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

$$BD(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) = \frac{1}{8}\boldsymbol{\mu}_{12}^T \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right)^{-1} \boldsymbol{\mu}_{12} + \frac{1}{2}\ln \frac{|(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}} |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \qquad (2)$$

となる．T は転置を，$|\cdot|$ は行列式を表す．$\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ ($\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$) は $p_1(\boldsymbol{x})$ ($p_2(\boldsymbol{x})$) の平均ベクトルおよび分散共分散行列であり，$\boldsymbol{\mu}_{12}$ は $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ 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 $\boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}$ 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(注1)．これは，バタチャリヤ距離が空間を歪

(注1)：ある条件を満たす変換であればアフィン変換に限らず非線形変換においても変換不変性は成立する [8]

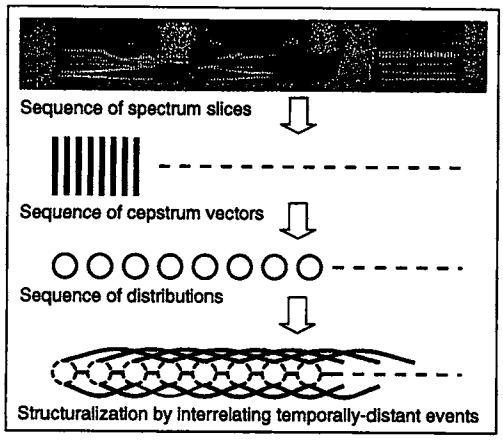

図 2　一発声の構造化

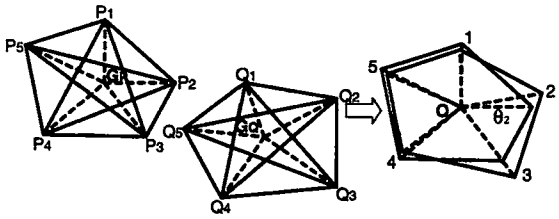

図 3　構造に基づく音響的照合

める距離尺度であることに起因する．この構造はアフィン変換に対して不変であり，凡そ非言語的特徴に依存しない普遍的な構造となる（音響的普遍構造）．$\boldsymbol{c}$ に $\boldsymbol{A}$ を掛ける演算は構造の回転として，$\boldsymbol{b}$ を加える演算は構造のシフトとして観測されることとなる．

2.3　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，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 $(P_1, P_2, \cdots, P_M, Q_1, Q_2, \cdots, Q_M)$ で構成される二つの構造のうち，一方をシフト（$\boldsymbol{b}$）と回転（$\boldsymbol{A}$）のみにより他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）．このときの構造間距離を，対応する頂点間距離の和の最小値として定義する．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，

$$\sqrt{\frac{1}{M}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \tag{3}$$

は上記の構造間距離を近似することが示されている [9]．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し，距離行列の情報のみを基に，シフト（$\boldsymbol{b}$）と回転（$\boldsymbol{A}$）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．即ち，声道長正規化（$\boldsymbol{A}$），環境適応（$\boldsymbol{b}$）後の音響スコアを，これらの処理を明示的に行うことなく計算することになる．

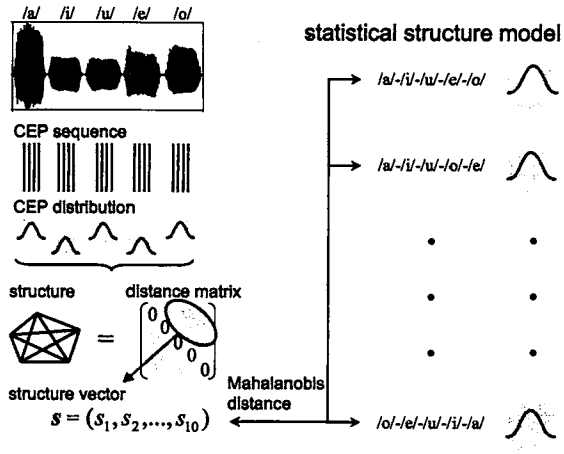

図 4　構造を用いた孤立発声日本語母音系列認識

2.4　構造的表象に基づく日本語母音孤立発声認識

筆者らは既に，孤立的に発声された日本語 5 母音系列を認識タスクとして，構造的表象に基づく音声認識を検討してきた [6]．認識の枠組みを図 4 に示す．入力音声よりケプストラムベクトル系列を求め，各母音に対してケプストラム分布を算出する．各分布間の距離を算出し，得られた距離行列（即ち構造ベクトル）を特徴量として用いる．認識器が持つモデルは，学習話者 1 人より得られる複数の構造ベクトルからガウス分布を求め，これを「構造統計モデル」として使用する．入力とモデルとの音響的照合は，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求め，距離が最小となる単語を認識結果として出力する．

認識実験では，スペクトル高域成分除去と MAP 推定を組み合せることで，1 人の話者で構築した構造統計モデルが 100%の認識率を達成し，4130 人の話者により構築された従来の HMM よりも高い性能および頑健性が得られたことが確認されている [6]．

3.　構造的表象に基づく連続発声認識

3.1　HMM を用いた連続発声の構造化

本研究では，認識タスクを連続発声に拡張する．連続発声から分布系列を求めるためには，孤立発声の場合とは異なり連続発声を適切に分節化し分布系列を求める必要があるが，本研究では分布系列の推定に HMM 学習を用いる．連続発声から音響事象分布系列を推定する枠組みを図 5 に示す．一発声から得られるケプストラム系列より，HMM を学習することによって出力確率分布系列を推定する．ここで，$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$ は連続音声から求めたケプストラム系列，$\theta = \{\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ は HMM のパラメータであり，$\boldsymbol{a}$ は状態遷移確率，$\boldsymbol{\mu}$ 及び $\boldsymbol{\Sigma}$ はそれぞれ出力確率密度分布の平均ベクトルと分散共分散行列である．HMM の学習により得られた出力確率密度分布を用いて，各音響事象分布 $p(\boldsymbol{c}|X)$ を推定する．一発声という非常に限られたデータから HMM を学習する必要があるため，本分析では ML 推定・MAP 推定・ベイズ推定の各推定法を用いて，以下の 3

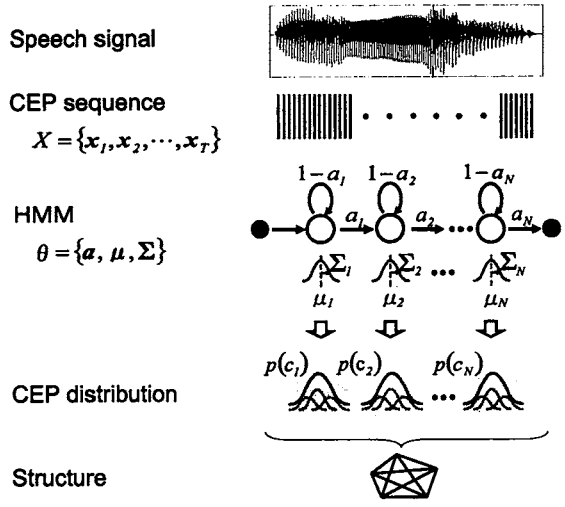

図 5 HMM を用いた連続音声の構造化の枠組み

表 1 分析条件

サンプリング	16bit / 16kHz
窓	窓長 25msec, シフト長 10msec
パラメータ	MCEP + Δ + ΔE (25 次元)
音響事象分布	単一ガウス分布 (対角共分散行列)
分布推定方法	ML, MAP, MAP with VB
状態数	N=5, 15, 25
帯域	Fullband, 4kHz, 2kHz

つの手法により音響事象分布推定を行い，各手法の比較を行う．

ML Baum-Welch アルゴリズムで HMM を学習し，その出力確率密度分布をそのまま音響事象分布とする．

MAP Baum-Welch アルゴリズムで HMM を学習し，その出力確率密度分布と事前分布を用いて音響事象分布を MAP 推定する．

MAP with VB 変分ベイズ法 [10], [11] により HMM を学習する．音響事象分布を算出する際には，変分ベイズ学習により最終的に得られる最適変分事後分布に対して，これを最大化するパラメータを音響事象分布として用いる (MAP 近似)．

3.2 日本語母音系列連続発声認識

3.2.1 認識実験

日本人成人 16 名 (男女各 8 名) より，日本語 5 母音連続発声系列の録音を行った．各母音は一回ずつ出現し，語彙数は ${}_5P_5 = 120$ となる．各単語に対して各話者 5 回の発声を収録した．以下では，このうち男女各 4 名の音声データを学習データとして，残りを評価データとして用いた．音声データより得られるケプストラムベクトルを基に，各発声に対して，ML・MAP・MAP with VB の 3 通りの手法により音響事象分布を求め，各手法の比較を行った．構造統計モデルは，各単語について計 40 個 (= 8 話者× 5 発声) の学習データの構造ベクトルよりガウス分布を推定した．評価データとして 4800 個 (= 8 話者× 5 発声× 120 単語) の構造ベクトルを入力として用い，入力と各構造統計モデルとのマハラノビス距離が最小となるモデルを認識結果として出力する．分析条件を表 1 に示す．以下の実験では，3 種類の分布推定手法に対して，状態数及び帯域を変化させたときの認識率を違いを調べた．

表 2 状態数を変化させたときの認識結果

推定法＼状態数	5	15	25
ML	19.9 %	32.4%	31.9%
MAP	30.8 %	59.1%	61.5%
MAP with VB	34.1 %	60.4%	62.5%

表 3 帯域を変化させたときの認識結果

推定法＼帯域	Full-band	4kHz	2kHz
ML	32.4 %	37.6%	41.5%
MAP	59.1 %	58.6%	61.1%
MAP with VB	60.4 %	59.1%	61.4%

3.2.2 実験結果

状態数を 5, 15, 25 と変化させた場合 (帯域は Fullband) の認識結果を表 2 に示す．3 つの分布推定手法を比較すると，MAP 及び MAP with VB において認識率は飛躍的に向上している．また，VB を導入することにより若干の性能向上が確認できる．状態数の違いに関しては，状態数が増加するに伴い認識性能が向上していることが確認できる．

次に，用いる帯域を Fullband, 4kHz, 2kHz と変化させたとき (状態数は 15) の認識結果を表 3 に示す．孤立 5 母音の認識タスク [6] の時ほどの認識率向上は見られなかったものの，2kHz まで帯域を制限することにより若干の認識性能の向上が確認できた．

3.2.3 考察

分布推定手法に関しては，MAP 及び MAP with VB での認識率向上の結果から，[6] での結果と同様，事前知識を導入することにより安定した分布推定が可能となることがわかる．さらに VB の導入による若干の性能向上に関しては，音響事象の境界推定のズレがより小さくなったためと考えられる．状態数に関しては，連続発声では音素遷移部分が存在するため，状態数を増やすことで適切なモデル化が可能となるといえる．ただ，状態数を増やすことにより，音響事象と HMM の状態との対応付けのズレが生じやすくなる．音響事象と HMM の状態の対応付けは話者や発話ごとに異なるが，本実験では同一の状態番号を持つ状態が同一の音響事象に対応していると見なして構造比較を行っており，この対応関係のミスマッチが認識率が 60% 程度にとどまった理由の一つとして挙げられる．

帯域制限の効果に関しては，若干の性能向上が見られるものの，孤立発声の場合ほどの認識率向上は見られなかった．アフィン変換による非言語的特徴の表現は，最も単純な形でのモデル化であり，構造化により非言語的特徴が完全に消失するわけではない．音声の高域成分には話者の個人性の情報が強く含まれており [12]，帯域制限することでより個人性を削除できるが，一方で，連続発声では調音結合が生じるため，帯域を制限することにより音響事象境界の推定が難しくなると考えられる．

4. DP に基づく構造間比較手法

前節まで述べてきたように，一発声の構造化が可能となり，構造間の距離が定義されれば，構造に基づく音声認識が可能となる．しかし，連続発声を構造化する場合，音響事象と HMM の状態の対応付けが話者や発話ごとに異なることが問題となる．式 (3) による構造間の音響的照合では，同一の状態番号を持つ状態が同一の音響事象に対応していると見なして構造比較を行っており，この対応関係のミスマッチが前節の認識実験での認識率低下の原因となっていると考えられる．また，式 (3) は状態数が同一であることを前提としているため，状態数が異なる場合には構造間比較を行うことが出来ない．

本節では，分布系列に対して DP マッチングを行い，分布系列同士の対応関係を求めた上で構造間の比較を行う手法を提案する．本手法により，音響事象と状態対応のズレの解消と，状態数の異なる構造間の比較が可能となることを実験的に示す．

4.1 DP による対応関係の導出と構造間比較

DP マッチングは，時間軸を非線形に伸縮する時間正規化を行う手法であり，2 つの系列間の対応付けと累積距離が求められる．2 つの分布系列を $P = \{P_1, P_2, \cdots, P_M\}$, $Q = \{Q_1, Q_2, \cdots, Q_N\}$ とし，分布 P_i と Q_j との距離を $d(i,j)(1 \leqq i \leqq M, 1 \leqq j \leqq N)$ とすると，$(1,1)$ から (i,j) までの累積距離 $g(i,j)$ は，例えば以下のように定式化される．

$$g(i,j) = \min \begin{bmatrix} g(i,j-1) & + & w_1 d(i,j) \\ g(i-1,j-1) & + & w_2 d(i,j) \\ g(i-1,j) & + & w_3 d(i,j) \end{bmatrix}$$

w_1, w_2, w_3 はそれぞれのパスに対する荷重係数である．各点において選択されたパスを辿ることで系列間の対応関係を求めることができる．以下では，$d(i,j)$ を分布 P_i と Q_j とのバタチャリヤ距離により定義する．よって，DP を行う際には距離行列だけでなく分布そのものに関する情報も必要となる．

分布系列間の対応関係が得られた後，対応関係に基づいて状態を分割もしくは併合する．例えば，P_i に対して Q_j と Q_{j+1} が対応していた場合，P_i を 2 つに分割，もしくは Q_j と Q_{j+1} を 1 つに併合する，という操作を行う．分割・併合は距離行列の要素を複製・統合することにより実装される．分割・併合後の距離行列から式 (3) を用いて構造間距離を算出する．

従来までの音響的照合手法が，話者適応などにより非言語的特徴の差を吸収した上で，DP や HMM により時間的対応を取る音響的照合であるのに対して，本手法は，DP により時間的対応をとった上で，非言語的特徴を消失させた音響的照合を行っていると捉えることができる．

4.2 評価実験

4.2.1 構造間距離の比較

2 名の話者より日本語母音系列連続発声 (120 単語) を各 5 回収録したデータを用いて，異話者間での構造間距離を DP 無しと DP 有り (提案手法) により算出した．600 発声× 600 発声の全 360,000 通りの組み合せで，同一単語間は 3,000 通り，異単語間は 357,000 通りとなる．表 4 の音響分析条件の下で，

表 4 音響分析条件

サンプリング	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
パラメータ	MCEP + Δ + ΔE (25 次元)
音声事象分布	単一ガウス分布 (対角共分散行列)
分布推定方法	MAP with VB
帯域	Fullband

表 5 DP 無し・有りでの構造間距離の比較

		DP 無し	DP 有り
同一単語間	平均	4.82×10^{-2}	$\mathbf{4.64 \times 10^{-2}}$
	分散	6.10×10^{-5}	$\mathbf{3.81 \times 10^{-5}}$
異単語間	平均	7.50×10^{-2}	$\mathbf{8.06 \times 10^{-2}}$
	分散	1.33×10^{-4}	$\mathbf{2.16 \times 10^{-4}}$

状態数は 25 として構造化を行った．表 5 に，同一単語間・異なる単語間での構造間距離の平均・分散を示す．DP を導入することにより，同一単語間の構造間距離は減少し，異なる単語間では構造間距離が増加している．分散に関しては，同一単語間ではかなり小さくなっているのが確認できる．

4.2.2 認識実験

DP に基づく構造間比較の音声認識における有効性を調べるため，日本語母音系列連続発声を認識タスクとして，DP を用いない場合と DP を用いる場合で認識実験を行った．日本人成人 16 名 (男女各 8 名) より，日本語 5 母音連続発声系列 (単語数 ${}_5P_5 = 120$) に対して，それぞれ 5 回の発声を収録し，男女各 4 名の音声データを学習データ，残りを評価データとして用いた．各発声に対して，表 4 の音響分析条件にて音響事象分布を推定し，各分布間の距離を算出することにより構造を求めた．認識器の持つモデルは，各単語について計 40 個 (= 8 話者× 5 発声) の学習データの構造ベクトルからガウス分布 (構造統計モデル) を求める．評価データとしては，4800 個 (= 8 話者× 5 発声× 120 単語) の構造ベクトルを入力として用い，入力と各構造統計モデルとのマハラノビス距離が最小となるようなモデルを認識結果として出力する．

DP を用いない場合は第 3. 節と同様の枠組みとなるが，DP を用いる場合ではモデル学習時と認識時に DP を導入する．学習時には，学習データ間で DP を用いて対応関係を求めた上でガウス分布を算出する．認識時には，入力とモデルとの対応関係を DP により求めた上でマハラノビス距離を算出する．入力とモデルとの DP の際には分布そのものの情報が必要となるため，モデル学習時に分布モデルも作成し，モデルとして保持する．但し，認識における音響的照合 (マハラノビス距離の算出) の際には，分布そのものの情報は明示的には用いられていない点に注意する．

まず，入力とモデルとで状態数 N を同一として，$N = 10 \sim 30$ に変化させて認識率を算出した．結果を図 6 に示す．DP 無しの場合では，状態数 23 のときに 63.0%が最大であったのに対し，DP 有りでは状態数 30 で 77.48%の認識率となった．

入力とモデルの状態数が異なる場合での認識性能を調べるため，モデル状態数として $N = 15, 20, 25$ とし，入力の状態数を

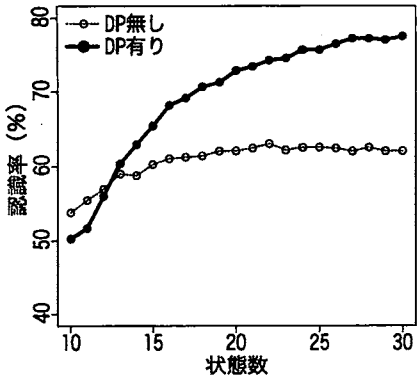

図 6　入力とモデルが同一状態数の場合の認識率

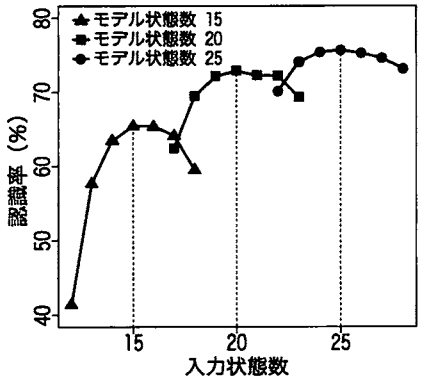

図 7　入力とモデルで状態数が異なる場合の認識率

モデルから −3 ～ +3 と変化させて認識を行った．結果を図 7 に示す．およそ 1,2 状態の違いであれば，それほど大きな性能の劣化は無く，数%程度の認識率の低下で認識が可能である．

4.2.3 考　　察

4.2.1 節の実験結果より，DP を導入することで，構造間距離について同一単語間と異なる単語間との差，つまり，正解と不正解の違いがより明確になったことが実験的に確認された．また，同一単語間での分散がより小さくなった点に関しては，DP により状態のズレが解消されたことにより，同一単語間の構造間距離のばらつきが抑えられたと解釈できる．

4.2.2 節の認識実験では，DP を用いることで認識性能が大幅に向上し，さらに入力とモデルの状態数が異なる場合でも構造間比較が可能となることが実験的に確認された．ただ，依然として孤立発声の認識 [6] での性能には及ばない．原因としては，モデル学習に用いるデータが [6] と比較して少ないことが挙げられる．構造統計モデル構築時の分布推定方法について検討する必要がある．また，認識率を話者別で見てみると，認識率に若干の開きがあることがわかった．構造化による非言語的特徴の消失は完全ではないため，話者によって認識率の差が生じたものと考えられる．状態数が異なる場合の認識に関しては，1,2 状態の違いはおよそ許容範囲であることが確認されたが，その許容範囲の妥当性は今後検討する必要がある．

5. ま と め

日本語母音系列の連続発声を認識タスクとして，構造的表象を用いた認識実験を行った．音声の音響的実体を全く用いず，実体間の関係のみをモデル化する音声認識において，約 63%の認識性能を示した．さらに，音響事象と状態対応のズレと，異なる状態数の構造間比較の問題を解決する方法として，DP に基づく構造間比較手法を提案した．認識実験において 77.48%の認識率となり，DP に基づく構造間比較手法の導入による認識性能向上を確認した．

今後の課題として，現状では既知として与えている状態数を自動推定する手法と，子音を含む音声の構造化手法を検討し，より実用に近い認識タスクでの認識実験を行う予定である．

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. Int. Conf. Production Research (ICPR'2004), 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] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. Int. Conf. Acousitcs, Speech, & Signal Processing (ICASSP'2005), pp.889–892 (2005)

[6] T. Murakami *et al.*, "Japanese vowel recognition using external structure of speech," Proc. IEEE Automatic Speech Recognition and Understanging Workshop (ASRU'2005), pp.203–208, 2005.

[7] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," Proc. European Conf. Speech Communication and Technology (EUROSPEECH'2005), pp.1445–1448, 2003.

[8] 峯松信明他，"音声の構造的表象を通して考察する幼児の音声模倣と言語獲得"，人工知能学会 AI チャレンジ研究会，SIG-Challenge–0624–6，pp.35–42，2006.

[9] 峯松信明他，"音声の構造的表象とその距離尺度"，電子情報通信学会技術研究報告，SP2005-13，pp.9–12，2005.

[10] 上田修功，"ベイズ学習 [I]，[II]，[III]，[IV]"，電子情報通信学会誌，vol.85，no.4 (265–271)，no.6 (421–426)，no.7 (504–509)，no.8 (633–638)，2002.

[11] 越仲孝文他，"HMM の変分ベイズ学習によるテキストセグメンテーション及びその映像インデキシングへの応用"，電子情報通信学会論文誌，Vol.J89–D，No.9，pp.2113–2122，2006.

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