音声の構造的表象を用いた音声認識における構造間比較手法の検討*

◎朝川智, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

従来の音声認識技術は，音声の音響的実体，即ちスペクトルそのものをモデル化してきたが，この実体は話者の声道形状の特性，マイクロフォンの特性などの非言語的特徴によって不可避的に歪む．近年，上記の非言語的特徴を表現する線形変換性歪み・乗算性歪みを原理的に持たない音響的普遍構造が提案された [1]．この構造的表象は，音声の音響的実体を捨象し，実体間の関係（距離）のみをモデル化する物理表象であり，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 [2]．既に我々は，本表象の音声認識への応用について検討を始めている [3, 4]．本稿では，[4] での検討課題である「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，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 音声に内在する音響的普遍構造

ある音響事象をケプストラムベクトルにより構成される音響空間内の分布として定義する．N 個の分布に対する ${}_N\mathrm{C}_2$ 個の分布間距離を求めることは，N 点によって張られる構造を音響空間内において定義することになる．分布間距離としてバタチャリヤ距離の平方根を用いる．バタチャリヤ距離は，任意の二分布間に対して如何なる共通のアフィン変換を施しても距離は不変である．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 [2]．ケプストラム $\boldsymbol{c}$ に対するアフィン変換 $\boldsymbol{Ac}+\boldsymbol{b}$ は，$\boldsymbol{A}$ の乗算は声道長特性，聴覚特性など話者・聴取者の生理学的な差異を表現し，ベクトル $\boldsymbol{b}$ の加算は音響機器などの差異を表現することを考えると，音響事象分布間の距離行列，即ち構造は，これら非言語的特徴に凡そ依存しない普遍的な構造となる（音響的普遍構造）[1]．

2.2 一発声の構造化と構造に基づく音響的照合

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，一発声の構造化が可能である．次に，二つの構造の構造間距離を求めることで，構造を音響的に照合することを考える．それぞれ N 個の頂点（$P_1, \cdots, P_N$ 及び $Q_1, \cdots, Q_N$）で構成される二つの構造において，構造間距離 D は下記の式で近似的に求まることが示されている [5]．

$$D = \sqrt{\frac{1}{N}\sum_{i<j}\left(p_{ij}-q_{ij}\right)^2} \qquad (1)$$

ただし，$p_{ij} = \overline{P_iP_j}, q_{ij} = \overline{Q_iQ_j}$ であり，p_{ij}, q_{ij} はそれぞれ構造 P，Q を表す距離行列の ij 要素である．従って，構造に基づく音響的照合は距離行列のみから近似的に行うことができる．

2.3 構造に基づく連続発声認識

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 (1) 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連続発声での認識率低下の原因となっている [4]．また，式 (1) 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．

本研究では，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．

3 DP に基づく構造間比較手法

3.1 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 $d(i,j)(1 \le i \le M, 1 \le j \le N)$ とすると，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出来る．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 (1) を用いて構造間距離を算出する．

4 評価実験

4.1 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．Table 1 の音響分析条件の下で，状態数は 25 として構造化を行った．

---

*A method of structural matching between two word utterances for speech recognition using structural representation of speech.
by Satoshi Asakawa, Nobuaki Minematsu and Keikichi Hirose (University of Tokyo)

Table 1　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）
帯域	Fullband

Table 2　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}	$\mathbf{2.16 \times 10^{-4}}$

Table 2 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　認識実験

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 ${}_5\mathrm{P}_5 = 120$）に対して，それぞれ 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．各発声に対して，Table 1 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 [4] 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 Fig 1 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，入力の状態数をモデルから $-3 \sim +3$ と変化させて認識を行った．結果を Fig 2 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．

4.3　考察

4.1 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，

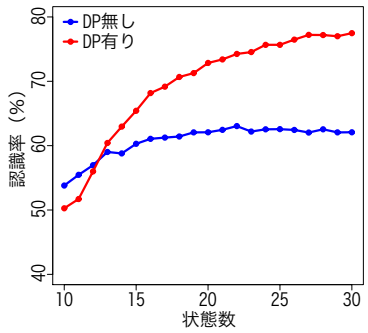

Fig. 1　入力とモデルが同一状態数の場合の認識率

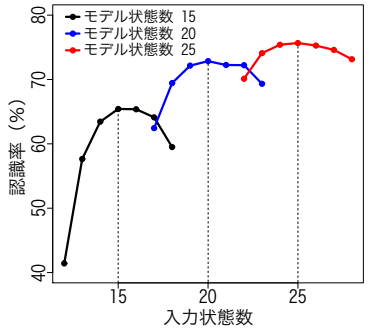

Fig. 2　入力とモデルで状態数が異なる場合の認識率

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ことは，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 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 [3] 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

音声の構造的表象に基づく音声認識において問題となっていた，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，その有効性を実験的に示した．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．

参考文献

[1] N. Minematsu, Proc. ICASSP, 889–892, 2005.
[2] N. Minematsu et al, Proc. ISCA Turorial and Research Workshop on SRIV, 47–52, 2006.
[3] T. Murakami et al, Proc. ASRU, 203–208, 2005.
[4] 朝川他, 音講論（秋）, 1-2-7, 2006.
[5] 峯松他, 信学技報, SP2005-13, 9–12, 2005.