音声の構造的表象を用いた音声認識における特徴量空間分割とその効果*

◎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み，音声認識においてはノイズとなる．従来の音声認識では，不特定話者モデルのように大量のデータを集めることによって解決を試みてきた．音声の構造的表象 [1] は，音響的実体そのものを直接用いず，実体間の関係（距離）のみをモデル化することにより，非言語的特徴を表現する線形変換性歪み・乗算性歪みに対して不変である音声の表象手法である．既に我々は，この話者不変の表象を用いた音声認識を検討しており，その有効性を示してきた [2, 3]．ただ，本手法は話者性を効果的に消失させる一方で，その強すぎる不変性のために全く異なる単語が同一と見なされてしまう可能性があり，構造間照合時に不適切なマッチングを抑制するような何らかの制約を設ける必要があった．

本稿では，構造的表象を用いた音響的照合において適切な制約を導入するために，特徴量空間分割による構造間照合手法を提案する．連続的に発声された日本語 5 母音系列を認識タスクとして，8 名の学習話者による提案手法のモデルが，数千名による従来手法による HMM とほぼ同等の性能であることを実験的に示す．

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，いわゆる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可能であり，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．

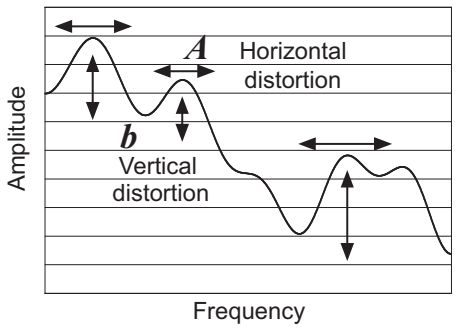

Fig. 1 Spectral distortions caused by $\boldsymbol{A}$ and $\boldsymbol{b}$

線形変換性歪みは，$\boldsymbol{c}$ に対する行列 $\boldsymbol{A}$ の乗算 $\boldsymbol{c'} = \boldsymbol{Ac}$ 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，$\boldsymbol{c}$ に対する $\boldsymbol{A}$ の乗算で表されることが示されている [4]．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 $\boldsymbol{c'} = \boldsymbol{Ac} + \boldsymbol{b}$ で表現される．Fig. 1 は，アフィン変換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．

2.2 音声の構造的表象

空間内に存在する n 点に対して，${}_nC_2$ 個の全ての二点間距離を求めると，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，アフィン変換は，特殊な形式の場合を除けば，構造を歪ませる変換であるからである．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．

分布間距離尺度の一つである Bhattacharyya 距離

*Use of feature space division and its effects in structural speech recognition.
by ASAKAWA Satoshi, MINEMATSU Nobuaki, and HIROSE Keikichi (The University of Tokyo)

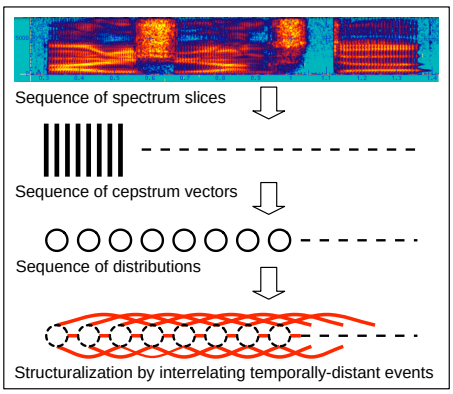

Fig. 2　Structuralization of a single utterance

(BD) を考える．あ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，BD は以下の式で表される．

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

二つの分布に対して線形・非線形を問わず共通の変換が施された場合，変換の前後で BD は不変となる [5]．二つの分布がガウス分布 $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$，$\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ で表現されているとき，BD は，

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

となる．$\boldsymbol{\mu}_{12}$ 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，BD はその前後で不変である．これは BD が空間を歪める距離尺度であることに起因する．音響事象を分布として捉え，音響事象群を分布間距離のみによって定義することで，アフィン変換に対して不変，つまり非言語的特徴に対しておよそ不変な構造を求めることが可能となる．$\boldsymbol{c}$ 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（Fig. 2）．次に，n 個の頂点 $(P_1, P_2, \cdots, P_n, Q_1, Q_2, \cdots, Q_n)$ 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（Fig. 3）．このときの構造間距離を，対応する頂点間距離の和（$\sum_{i=1}^{n} \overline{P_i Q_i}^2$）の最小値として定義する．分布間距離として Bhattacharyya 距離の平方根を用いた場合，

$$\sqrt{\frac{1}{n} \sum_{i<j} (\overline{P_i P_j} - \overline{Q_i Q_j})^2} \quad (3)$$

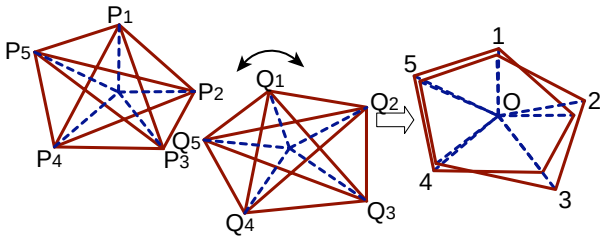

Fig. 3　Acoustic matching after shift and rotation

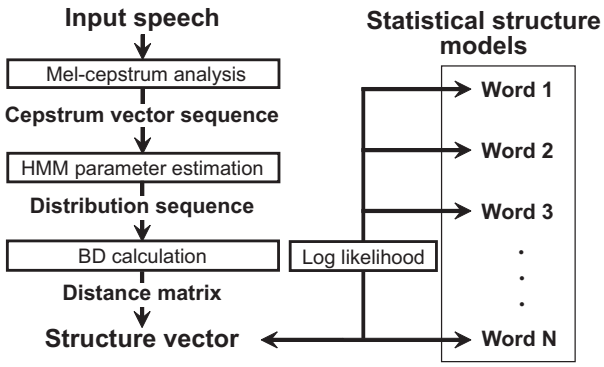

Fig. 4　Framework of the structural recognition

は上記の構造間距離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，$\boldsymbol{b}$，$\boldsymbol{A}$ を求めることなく行なうことが可能であることを意味する．

2.4　構造的表象に基づく音声認識と特徴量空間分割

2.5　構造的表象に基づく音声認識の枠組み

構造的表象に基づく音声認識の枠組みを Fig. 4 に示す．図の左側は入力音声から構造を抽出する手順を表す．まず，入力音声はメルケプストラム分析によりケプストラムベクトル系列に変換され，さらに多次元ガウス分布系列へと変換される．ケプストラム系列から分布系列への変換には HMM の学習アルゴリズムを利用し，各状態における出力確率分布を分布系列として用いる．本枠組みでは，一発声という非常に限られたデータ量からパラメータ推定を行う必要があるため，MAP 推定を用いている．得られた分布系列より，すべての二分布間距離を BD により算出し距離行列を得る．得られた距離行列は対角が 0 の対称行列であるので，上三角成分を並べたベクトル（構造ベクトル）を一発声を表す特徴量として用いる．図の右側はシステムが保持する各単語のテンプレートを表す．学習データより得られる複数の構造ベクトルから，単語毎に平均ベクトル・分散共分散行列を算出し，それをモデルとして保持する（構造統計モデル）．認識時には，入力構造ベクトルと各単語の構造統計モデルとの対数尤度に基づいて認識結果を出力する．

2.6 特徴量空間分割

構造の不変性は線形・非線形問わず，あらゆる変換に対して成り立ち，分布がガウス分布である場合においてはあらゆるアフィン変換に対して不変性が成立する．この非常に強い不変性により効果的に話者性を取り除くことが可能となるが，それと同時に，言語的に全く異なる単語同士が同一と見なされてしまうことが起こりうる．この強すぎる不変性は認識性能の低下を引き起こすと考えられ，許容される変換に対して何らかの制約を導入する必要があると考えられる．

ここで，回転（$\boldsymbol{A}$）に着目する．2.1 節で述べたように，話者性の違い，即ち声道長の違いはケプストラムベクトルに対する変換行列 $\boldsymbol{A}$ によってモデル化することができる．声道長正規化においては，例えば $\boldsymbol{A}$ は下記のように定式化される [7].

$$\boldsymbol{A}=\begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad (4)$$

α はワーピングパラメータで，$|\alpha|<1$ である．α が十分に小さい場合であれば α の高次項を無視することができ，この場合の $\boldsymbol{A}$ は対角およびその付近にのみ値を持つような行列（帯行列）として表現される．変換行列 $\boldsymbol{A}$ に対して上記のような行列のみを許容するような制約を $\boldsymbol{A}$ に対してかけることで，声道長の違いを表現する回転のみを許容し，全く異なる単語同士がマッチングすることを防ぐことができる．以下では，特徴量空間を分割することにより上記の制約を実現する．

特徴量ベクトルとして，$\left(\boldsymbol{c}^T, \boldsymbol{\Delta c}^T\right)^T$ を考える．ここで，$\boldsymbol{c}=(c_1\ ,\ c_2\ ,\ \cdots\ ,\ c_D)^T$，$\boldsymbol{\Delta c}=(\Delta c_1\ ,\ \Delta c_2\ ,\ \cdots\ ,\ \Delta c_D)^T$ は，それぞれケプストラムベクトルとそのデルタを表し，D は次元数である．上記特徴量を用いた場合，アフィン変換

$$\begin{pmatrix} \boldsymbol{c'} \\ \boldsymbol{\Delta c'} \end{pmatrix}=\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}\begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{\Delta c} \end{pmatrix}+\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \quad (5)$$

に対して，あらゆる $\boldsymbol{A}=\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}$ を許容することになる．ここで，特徴量をケプストラムとデルタケプストラムとに分割してそれぞれで Bhattacharyya 距離を算出する場合を考える．

$$\boldsymbol{c'} = \boldsymbol{A}_{11}\boldsymbol{c}+\boldsymbol{b}_1 \quad (6)$$

$$\boldsymbol{\Delta c'} = \boldsymbol{A}_{22}\boldsymbol{\Delta c}+\boldsymbol{b}_2 \quad (7)$$

この場合，$\boldsymbol{A}_{12}=\boldsymbol{A}_{21}=\boldsymbol{0}$ となり，BD はそれぞれの部分特徴量空間における変換 $\boldsymbol{A}_{11}$，$\boldsymbol{A}_{22}$ に対して不変となる．これは，それぞれの部分特徴量空間においても構造の回転として観測されるような $\boldsymbol{A}$ のみを許容することに相当し，回転に対する制約と見なすことができる．同様にして，特徴量ベクトル全体から一部の特徴量のみを分割して構造化することは，その部分特徴量空間内に射影した構造においても変換不変性が成立するような行列 $\boldsymbol{A}$ のみを許容することに相当する．式 (4) のような対角およびその付近のみに値を持つ行列のみを許容する制約をかけるために，以下のように，隣接する w 個の特徴量ブロックをまとめて一つのストリームとし，各ストリームで個別に構造ベクトル算出および照合を行う．

stream 1 : $(c_1, c_2, \cdots, c_w)$
stream 2 : $(c_2, c_3, \cdots, c_{w+1})$
stream 3 : $(c_3, c_4, \cdots, c_{w+2})$
$\vdots$

ブロックサイズ w の値は，変換行列 $\boldsymbol{A}$ において 0 以外の値を持つ部分の幅に対応し，この値が小さいほど構造間照合時の回転に対してより強い制約をかけることを意味する．二つの発声の音響的照合を行う際は，各ストリームでの構造間距離の総和を全体の距離として用いる．

Table 1　Acoustic conditions for the analysis

sampling	16bit / 16kHz
window	25 ms length / 4 ms shift
parameters	Mel cepstrum (0 to 12) + Δ (0 to 12)
distribution	1-mixture Gaussian with a diagonal matrix

3 認識実験

3.1 実験条件

特徴量空間分割による効果を確かめるため，日本語 5 母音系列の連続発声をタスクとして認識実験を行った．日本人成人 16 名（男女各 8 名）より，日本語 5 母音に対して各母音が 1 回ずつ出現する連続発声母音系列（単語数 ${}_5\mathrm{P}_5$=120）を，それぞれ 5 回の発声を収録した．計 9,600 発声のうち，男女各 4 名を学習データとして，残りを評価データとして用いた．Table 1 に示す音響分析条件により，ケプストラム分析および HMM パラメータ推定を行った．

3.2 特徴量空間分割

2.6 節にて述べた特徴量空間分割は，分布系列推定の後に導入した (Fig. 5)．分割前の時点で，分布系列はメルケプストラム 1〜12 次元とパワー項（c_0）とそれらのデルタによる多次元ガウス分布系列として表現されている．事前の予備的実験により，構造的音響照合の際にはデルタパラメータは用いない方が良い性能を示すことが分かっており，以下の実験では構造的音響照合にはケプストラムおよびパワーのみを用

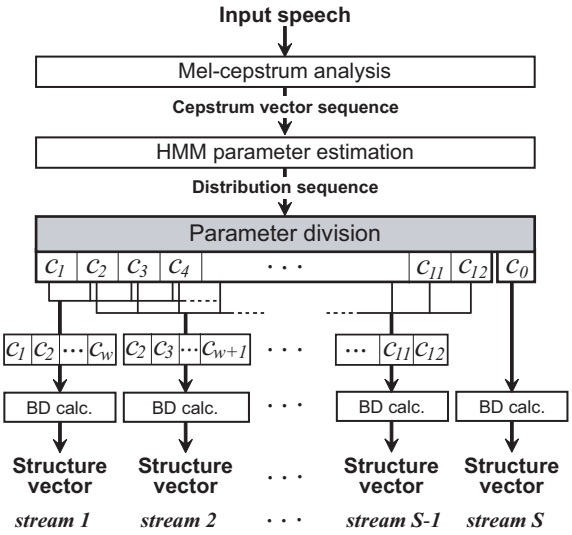

Fig. 5　Structuralization with parameter division

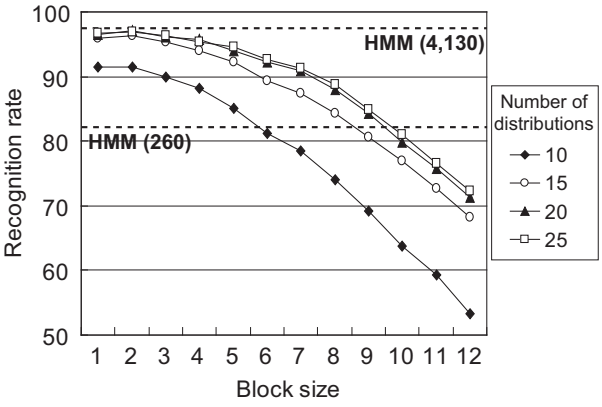

Fig. 6　Recognition rate with parameter division

いる．メルケプストラム 1〜12 次元に対して，ブロックサイズ w にて特徴量空間分割を行う．分割することにより各ストリームは w 次元ガウス分布系列として表現され，それぞれのストリームに対して構造ベクトルが求められる．パワーに関しては，それのみで単独のストリームとして扱う．

認識の枠組みは，Fig. 4 の左側を Fig. 5 により置き換えた形となり，モデル側もストリーム毎に構造統計モデルを保持する．認識時には，各ストリームでの対数尤度の合計に基づいて認識結果を出力する．ブロックサイズは 1〜12 とし，分布系列の分布数は 10，15，20，25 とした．比較実験として，不特定話者 HMM 音響モデルでの認識も行った．学習話者 260 名による状態共有 HMM と 4,130 名による分布共有 HMM[8] を音響モデルとして用い，言語的制約として 120 単語のみを許容する文脈自由文法を用いた．

結果を Fig. 6 に示す．横軸がブロックサイズ，縦軸が認識率を表す．図より，ブロックサイズが小さい場合において認識率が高く，ブロックサイズ 2 の場合が最も良い認識率を示した．分布数に関しては 20 以上の場合ではそれほど変化はなく，結果として分布数 20，ブロックサイズ 2 の場合で 97.1%と最も良い性能を示した．従来の不特定話者 HMM による結果も併せて Fig. 6 に破線で示す．学習話者 260 名および 4,130 名の不特定話者 HMM の認識率はそれぞれ 82.1%，97.4%である．8 名の学習話者による提案手法が，学習話者 260 名の HMM の性能を大きく上回り，4,130 名の HMM とほぼ同等の性能を示すことが確認された．

提案手法において最も良い性能を示したブロックサイズ 2 の場合は，式 (4) において対角要素およびその隣接要素のみが値を持つ場合に相当し，α の 2 次以上の項を無視した場合に相当する．[7] では行列 $\boldsymbol{A}$ の推定において α の 2 次以上の項を無視できると仮定しており，本実験の結果と整合する．以上より，構造間照合における回転に対する制約は特徴量空間を分割することにより実現可能であり，適切な制約を施すことで不適切なマッチングが抑制され，結果として認識率が向上することが実験的に確認された．

4　おわりに

本稿では，構造的表象を用いた音響的照合における不適切なマッチングを抑制することを目的として，特徴量空間分割による構造間照合手法を提案した．連続的に発声された日本語 5 母音系列をタスクとして認識実験を行い，8 名の学習話者による提案手法のモデルが，学習話者 4,130 名による従来の不特定話者 HMM とほぼ同等の性能を示した．構造間照合の際の回転に対する制約として特徴量空間分割による実装を検討し，声道長差異のモデルとして適切な制約を施すことで，不適切なマッチングを抑制し，認識実験において大幅な認識率向上が得られることを確認した．

今後の課題として，より多様な話者性に対応する制約を実現するための最適な特徴量空間分割の方法を検討することが挙げられる．また，子音を含む場合での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] T. Murakami *et al.*, *Proc. ASRU*, 203–208, 2005.
[3] 朝川他，信学技報，SP2006-105，119–124，2006.
[4] M. Pitz and H. Ney, *IEEE Trans. Speech and Audio Processing*, 13(5), 930–944, 2005.
[5] 峯松他，音講論（春）, 147–148, 2007.
[6] N. Minematsu, *Proc. ICASSP*, 585–588, 2004.
[7] 江森，篠田，信学論，J83-D-II(11)，2108–2117，2000.
[8] T. Kawahara *et al.*, *Proc. ICSLP*, 3069–3072, 2004.