# 判別分析と構造表象を用いた話者の多様性に超頑健な音声認識*

◎朝川 智，喬 宇，峯松 信明，広瀬 啓吉（東京大学）

## 1 はじめに

年齢，性別，個人性，音響機器など，音声には不可避的に非言語的情報が混入し，これが音声の音響的特徴を変形させ，音声記号としては同一音であったとしても，物理的には異なる特性を持った音響現象として観測される。従来の音声工学では，数千・万の話者の音声から同一記号音を収集し，スペクトル包絡に代表される音の実体を統計的にモデル化することで多様性問題の解決を図って来た。非言語的要因は一般に時不変であり，静的なバイアス項として音響量を変形する。上記のような統計モデルでの解決は，非言語的要因による音声の変動を母集団からのサンプリングに伴うランダム雑音としてモデル化することとなり，例えば不特定話者音響モデルは，音声に含まれる話者性は時間軸上でランダムに変化することを前提としている。これは，明らかに，事実にそぐわない。

幼児の言語獲得は，両親の発声を真似ることで行なわれる。音声模倣と呼ばれるこの活動は，霊長類ではヒトのみに見られる。しかし，親の太い声を真似る幼児はいない。彼らは音は模倣しない。霊長類以外では，音声模倣は鳥やクジラに見られるが，この場合は音を模倣する [1]。即ち，ヒトの幼児は個体サイズを超えた音声模倣をする。言い換えれば，静的なバイアスを超えて，父親の声と自らの声に同一性を感覚する。この場合注意すべきは，幼児は，与えられた発声をモーラに分割することは困難であるため [2]，両発声をストリングとして比較し同一性を感覚することも困難であれば，親の声から抽出した各モーラを自らの声で再生することも困難となることである。発達心理学では，幼児が模倣する対象を語全体の音形／語ゲシュタルトなどの言葉で表現している [2]。

音声合成における話者変換や音声認識における話者適応において，話者の違いは音響空間の写像としてモデル化される。話者の違いが静的なバイアスであれば，この写像は静的となる。写像で対応づける二話者が変われば当然写像関数は変わるが，任意の写像を施しても不変なる音響量が定義できれば，それは話者不変量となり，それが語の全体的な様態を表象していれば，発達心理学の言う語ゲシュタルトの物理的定義として議論することができる [3]。

本稿では筆者等が提唱する，微分可能かつ可逆な任意の写像に対して不変な音声の全体的・構造的表象 [4] が，孤立単語認識というタスクにおいて，話者の多様性に対して如何に頑健に機能するのか，を実験的に示す。この時,「強すぎる不変性問題」「高すぎる次元数問題」を解決する必要が生じるが，これらを特徴量空間分割及び判別分析を通して解決する。

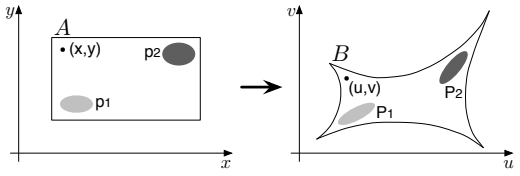


Fig. 1 微分可能かつ可逆な二つの空間 A と B

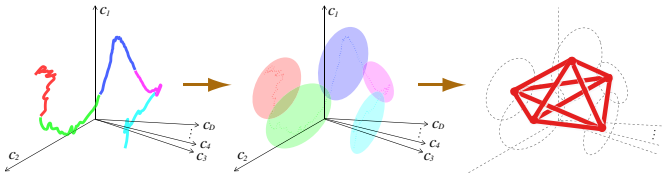

Fig. 2 発声の構造的・全体的・話者不変的表象

## 2 話者不変な音声の構造的表象とその照合

### 2.1 発声の構造化とそれに基づく話者不変表象

Fig. 1 に示す写像を考える。微分可能かつ可逆な変換関数で対応付けられた二つの空間である。空間 A の点 $(x, y)$ は空間 B の点 $(u, v)$ へ写像される。空間 A における事象 $p_1$ と $p_2$ を考える。全ての事象は点ではなく確率密度関数として存在する。$p_i$ に対応する空間 B の事象を $P_i$ とする。$P_i$ も確率密度関数となる。この時 f-divergence は両空間で常に等しい [5]。

$$f_{div}(p_1(x,y), p_2(x,y)) = f_{div}(P_1(u,v), P_2(u,v))$$

但し，$f_{div}$ は下記で定義される。

$$f_{div}(p_1(x,y), p_2(x,y)) = \int p_2(x,y) g\left(\frac{p_1(x,y)}{p_2(x,y)}\right) dxdy$$

バタチャリヤ距離も f-divergence の一種であり，本稿では写像不変な分布間距離として，これを用いる。

写像不変量のみを用いて音声ストリームを表象すれば，それは話者不変量となる。Fig. 2 にその手順を示す。ケプストラム時系列を分布系列へと変換し，任意の二分布間距離をバタチャリヤ距離として計測し，距離行列として発声を表象する。当然この距離行列は話者不変となる。また，距離行列は幾何学的構造を規定するため，この表象は構造的表象となる。

### 2.2 構造表象における音響的照合

話者 A のある発声を構造化し，話者 B のある発声を構造化する。この時，両発声を同一の分布系列としてモデル化し，構造化する。距離行列をベクトルとして考え（構造ベクトル），両構造ベクトルのユークリッド距離を計算すると，それは Fig. 6 に示すように，両構造を近づけ（シフト），両者が重なるように回転させて得られる,「対応する二事象間距離の総和

*Speech recognition with super robustness to speaker variability based on discriminant analysis and speech structures. by S. Asakawa, Y. Qiao, N. Minematsu, and K. Hirose (The University of Tokyo)

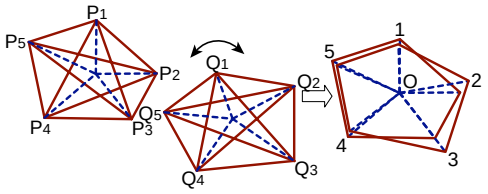


Fig. 3　二発声間の構造的照合

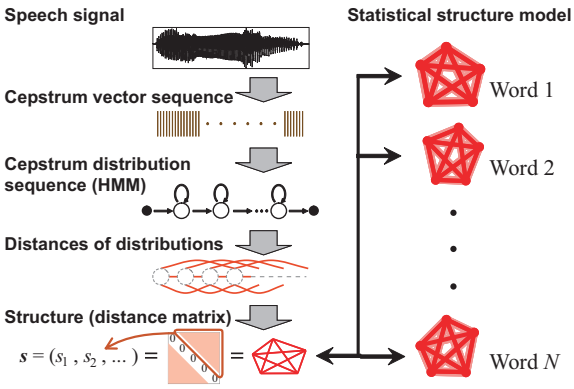


Fig. 4　構造表象に基づく孤立単語音声認識

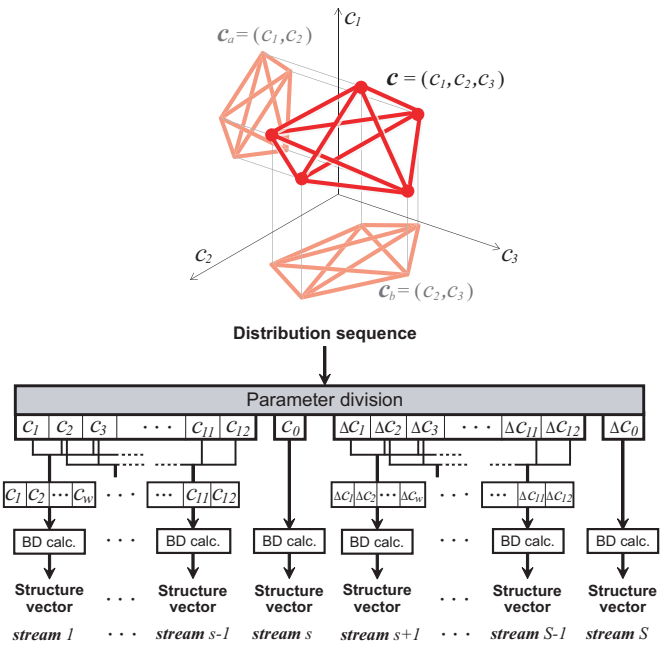


Fig. 5　部分空間への射影に基づく特徴量空間の分割

の最小値」を近似できることが示されている [4]。例えば音響空間としてケプストラム空間を考えれば，シフトは音響機器特性の違いを表現し，回転は声道長の違いを表現する。即ち発声を構造化して得られるユークリッド距離は，両発声を音響機器や声道長の違いに関して適応／正規化して計算される音響スコアを近似することになる。音の実体をモデル化する従来の音響モデルでは，このシフトや回転の演算子を具体的に推定し，実際に音響モデル（あるいは入力音声）を変形した上で照合を行なう必要があるが，構造表象ではこれらの操作を明示的に行なう必要は一切ない。しかし，これらの操作を行なった後に計算される音響スコアが得られる。これが構造表象に基づく音響照合の枠組みであり，話者の違いを効果的に消失させた上での音響スコア計算が容易に可能となる [4]。

以上の議論をまとめたものが Fig. 4 である。一発声の構造化は MAP-based な HMM 学習を通して行ない，推定された分布群を用いて構造ベクトルを得る。各単語のモデルは学習データより構成される該当単語の構造ベクトル群の多次元ガウス分布である。

## 3　二つの問題とその解決策，及び改善策

### 3.1　強すぎる不変性問題とその解決

音声の構造表象では，音声の絶対的な物理特性は一切捨象され，音声のコントラストのみが抽出される。そして，任意の写像前後の二構造は，同一の発声として扱われる。これは極めて強い不変性を有することを意味しており，言語的に異なる二発声を同一と見なすことが容易に起きる。任意写像に不変という構造表象の利点は，言語的要因による変形までを不変にする欠点となり得る。ここで，例えば声道長の違いによる変形にのみ，不変性を有する音響照合方式が必要となる。声道長の変化はケプストラムベクトル $c$ に対して，行列 $A$ を掛ける演算 $Ac$ で近似できるが，この時 $A$ は，対角成分付近以外は零となる帯行列となる [6]。帯行列による写像に関してのみ不変性を有するような音響照合は，特徴量空間を適切に次元分割し，個々の部分空間で構造的照合を行なうことで実装できることを先行研究で報告している [7]。具体的には最低次から連続する $w$ 個の次元で部分空間を張り，その空間で構造表象を構成し構造照合を行なう。同様の操作を次元を一つずらした $w$ 次元の部分空間でも行ない，最高次までこれを繰り返す。最終的な音響照合スコアは各部分空間でのスコアの和として定義する。Fig. 5 上図に，3 次元で構成される全空間を二つの 2 次元部分空間に射影した図を示す。下図にケプストラム及び Δ ケプストラムで構成される特徴量に対して空間分割する様子を示す。Δ 項を使う場合は，ケプストラムとは別々に部分空間を構成した。

### 3.2　高すぎる次元数問題とその解決

音声の実体をモデル化する場合，音声ストリームに $N$ 個の事象があれば，モデル化対象数は $N$ である。しかし，構造表象の場合 ${}_NC_2$ 個となり，モデル化対象数は $O(N^2)$ で増加し，パラメータ次元数の増加を容易に招く。その一方で，個々のパラメータの独立性が高いとは言えず，適切なパラメータ次元数の削減は，認識精度の向上に寄与することが期待される。既に先行研究にて PCA や LDA による次元圧縮を行なうことで，構造ベクトルをより低次元かつ識別的なパラメータへと変換する方法を提案しており [8]，本稿では Fig. 5 において特徴量分割を施し，各ストリームで構造ベクトルを算出した時点でまず LDA を導入して次元削減し，次元削減された特徴ベクトルを結合してできる拡大構造ベクトルに対して再度 LDA を施してより識別的なパラメータとする方法を採択した（2 段階 LDA）。Fig. 6 に 2 段階 LDA の様子を示す。

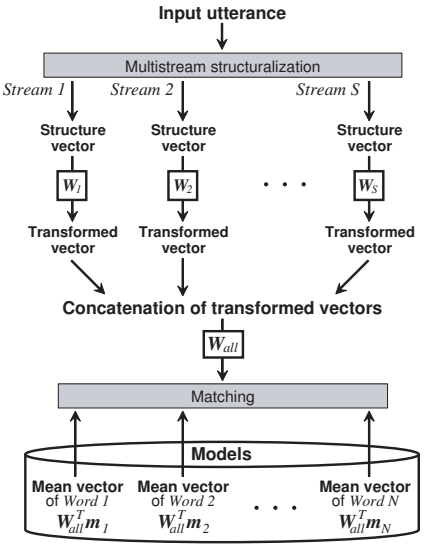


Fig. 6　2 段階 LDA を通して行なう構造的音響照合

### 3.3　ストリーム間構造距離の導入

Fig. 6 では個々のストリームは全く独立に扱われている。ここでは，各ストリームによって張られる構造間の距離の情報も，語彙同定に有効に寄与すると考えた。そして，$S$ 本のストリームに対して ${}_SC_2$ 個だけ得られるストリーム間構造距離をベクトル化し，拡大構造ベクトルに連結する形で導入し，これを 2 段目の LDA の入力として学習される識別器についても，実験的に検討することとした。

なお，下記の実験では，単語モデルは分布ではなく点として定義している。即ち，各単語の学習データに対して 1 段階目の LDA 後に得られる拡大構造ベクトルの平均ベクトルとして単語モデルを定義している。認識時には，入力される拡大構造ベクトルと，各単語モデル（ベクトル）とを 2 段階目の LDA（Fig. 6 では $W_{all}$）を通して比較し，認識結果を得る。

## 4　多様な話者性を伴う孤立単語認識実験

### 4.1　二つの孤立単語音声認識タスク

日本語 5 母音を並び替えて構成される母音単語セット（語彙数 120）と，東北大・松下単語音声データベースの音韻バランス単語（語彙数 212）の 2 種類のタスクを用意した。母音は話者の違いによってその音響的特徴が大きく変化するが，無声の摩擦音や破裂音は話者の影響が比較的小さい。これらは音の実体を捉えても話者の影響は少ないことを意味する。音のコントラストを捉えることで不変量を定義する提案手法は，前者のタスクには最適な方法であるが，後者のタスクには必ずしも最適な枠組みとは言えない。本稿では両タスクにおける構造的音声認識の性能を実験的に算出することで，その妥当性について検討する。

### 4.2　多様な話者性を伴う音声入力

本稿では自然発声の音声のみならず，これらに対して周波数ウォーピングを施し，声道長を伸縮することに等しい変換をかけた音声も入力音声として使用した。具体的な変換方法は [6] で示されている行列演算を用いてケプストラムを線形変換し，幅広い多様な話者性を人工的に生成した。

Table 1　音響分析条件（自然音声入力/構造モデル）

| | |
|---|---|
| サンプリング | 16 bit / 16kHz（5 母音単語） |
| | 12 bit / 16kHz（212 単語） |
| 窓長 | 25 ms length / 10 ms shift |
| 音響特徴量 | MCEP (0〜12) + Δ (0〜12) |
| 分布 | 対角分散行列による単一ガウス分布 |
| 分布数 | 20（5 母音単語）/ 25（212 単語） |
| 分布推定 | MAP 推定 |

Table 2　音響分析条件（自然音声入力/単語 HMM）

| | |
|---|---|
| サンプリング | 16 bit / 16kHz（5 母音単語） |
| | 12 bit / 16kHz（212 単語） |
| 窓長 | 25 ms length / 10 ms shift |
| 音響特徴量 | MFCC (1〜12) + Δ (1〜12) + ΔP |
| 分布 | 対角分散行列による単一ガウス分布 |
| 分布数 | 25（5 母音単語）/ 25（212 単語） |
| 分布推定 | ML 推定 |

### 4.3　音響分析条件と実験条件

自然音声を用いた場合の構造抽出及びモデリングに関する音響分析条件を Table 1 に示す。また，比較対象として構築した単語 HMM 構築時の音響分析条件を Table 2 に示す。なお，変換音声を対象とした実験では，特徴量として FFT ケプストラム（0〜16 次元）を用いた。これはメル化は周波数ウォーピングに相当する操作であり，ある特定のウォーピングパラメータによる声道長変換とほぼ等しい操作となるためである。比較実験で用いる単語 HMM も同様，FFT ケプストラム（0〜16 次元）を用いて学習した。

なお，学習・評価話者数であるが，5 母音単語に関しては学習・評価ともに男女 4 名ずつであり，各話者が各単語を 5 回ずつ発声している。そのため，学習データ数は 40 発声／語であり，評価データ数も 40 発声／語となる。音韻バランス 212 単語は，学習・評価ともに男女 15 名ずつであり，各話者が各単語を 1 回ずつ発声している。そのため，学習データ数は 30 発声／語であり，評価データ数も 30 発声／語である。

### 4.4　実験結果

Fig. 7 に，自然音声を入力した場合の認識結果を示す。上図が 5 母音単語，下図が音韻バランス単語である。共に，横軸はブロックサイズ $w$，縦軸が単語認識率である。MSS は特徴量ストリームの次元分割を意味し（Multiple Stream Structuralization），LDA は判別分析の導入を，Δ は Δ 項の導入を，ISD はストリーム間距離（Inter-Stream Distance）の導入を表す。Δ を伴わない場合，ストリーム数は $S$=14 − $w$ であるが，伴う場合は $S$=2(14 − $w$) となる。各ストリームに対して構成される構造のエッジ数は 5 母音単語の場合 ${}_{20}C_2$=190，音韻バランス単語の場合 ${}_{25}C_2$=300 である。2 段階 LDA は，前段で次元数を $D$ まで削減し，後段では拡大構造ベクトル次元数がクラス数以上である時のみ，次元数は「クラス数 −1」へと削減

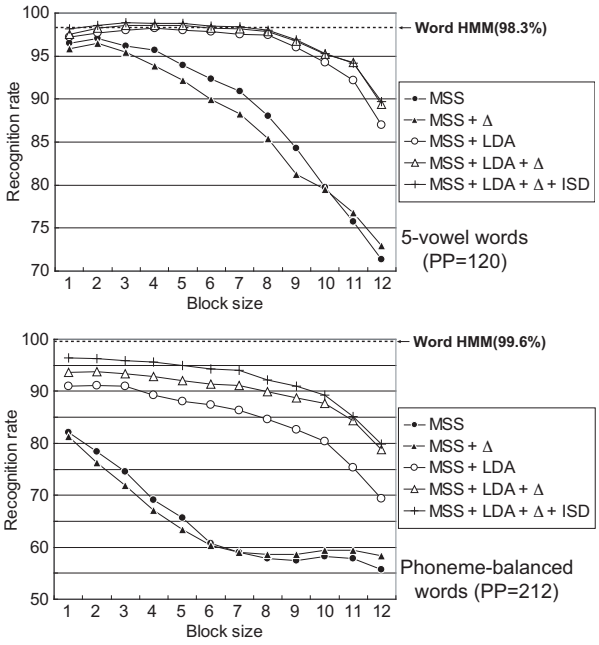


Fig. 7　自然音声を入力した場合の認識率

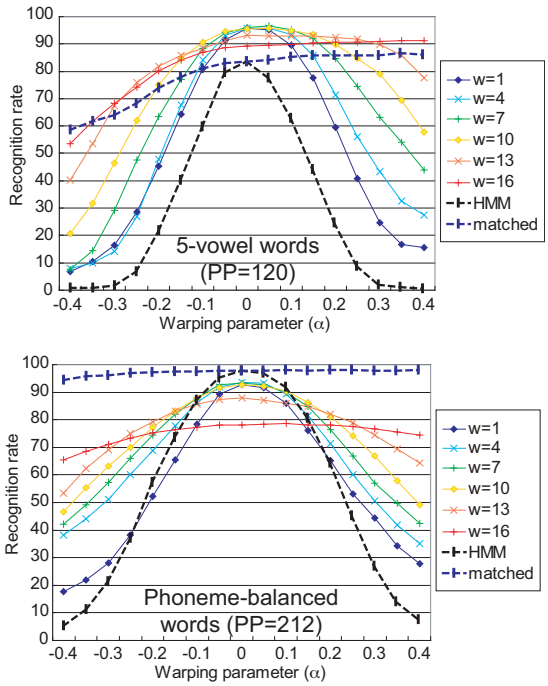


Fig. 8　変換音声を入力した場合の認識率

される。種々の場合を検討したが，Fig. 7 は両者において $D = 20$ とした結果である。なお，単語 HMM の結果は 5 母音単語の場合が 98.3%，音韻バランス単語の場合が 99.6%であった。構造モデルの最高性能は 5 母音単語の場合，MSS+LDA+Δ+ISD の条件下で，$w$=3 の時に 98.9%となり，音韻バランス単語の場合は上記条件下で，$w$=1 の時に 96.4%となった。

評価データとして変換音声を入力した場合の結果を Fig. 8 に示す。MSS+LDA+Δ+ISD を用いた結果である。横軸はウォーピングパラメータ $\alpha$ であり，値の負/正によって声道長が伸/縮する。$\alpha$=−0.4 で声道長が約倍に，+0.4 で約半分になる。−0.4 から 0.05 ステップで増加させ，+0.4 までの 17 通りについて実験した。5 母音単語，音韻バランス単語両者とも，$w$=1,4,7,10,13,16 の場合を示している。なお HMM とは，構造モデルと同一学習データで構築した単語 HMM の性能であり，matched とは，各 $\alpha$ で学習データも変換させ，17 通りの HMM を構築した場合の性能である。学習／評価間にミスマッチが無い場合の性能であり，話者適応による理論的な最高性能である。

### 4.5　検討と考察

Fig. 7 より LDA 導入の効果が分かる。特に，音韻バランス単語の場合に極めて大きい。これは逆に言えば，無声子音など，構造表象と相性の悪い音声が音韻バランス単語には存在していることが一因であると考えられる。但し，無声子音までを含めて話者変換を一つの関数で表現できる可能性もある。今回の場合，MAP 推定時の事前分布としては，両タスクに対して，一つずつ事前分布を学習データより与えた。音韻バランス単語では無声子音も母音も同一の事前分布を用いており，若干不適切な分布推定が行なわれた可能性もある。最低限の音実体カテゴリーを利用して，適切な分布推定を行なうことによる，MSS 及び LDA 導入後の性能向上について今後検討したい。Δ 項は，LDA 導入前は性能を下げる方向に働いたが，導入後は上げる方向に働いている。また ISD は，音韻バランス単語の方が寄与度は高い。

Fig. 8 より，構造表象の話者性に対する極めて強い頑健性が分かる。話者の違いも音韻の違いも音色（スペクトル包絡）の違いであるため，前者に対する不変性を高めれば，後者に対する不変性も自ずと高まる。$w$ がその制御パラメータとなっており，$w$ が大きければ話者不変性が高くなり，単語同定率は落ちる。このトレードオフは，図からも窺える。5 母音単語の場合，FFT-ケプストラムを使用したことにより，HMM の性能が劣化してしまった。これとの比較となってしまうが，構造表象の場合は，例えば $w$=16 では，ほぼ全ての条件で matched を超える性能が得られた。話者適応せずとも，話者適応したかのように機能する構造表象の特性が明確に現れた結果となった。一方，音韻バランス単語であるが，常に matched と同等の性能が得られた訳では無い。しかし，同一条件下で学習した HMM と比較すると（例えば $w$=10），無ミスマッチ時（$\alpha$=0）の性能劣化が小さく，その分ミスマッチが存在する時の性能向上が非常に大きい。これも構造表象の特性が明確に現れた結果と言える。

## 参考文献


[1] 岡ノ谷，音講論，1-7-15，pp.1555–1556 (2008-3)
[2] 早川，月刊言語，vol.35，no.9，pp.62–67，大修館書店 (2006)
[3] 峯松他，人工知能学会全国大会論文集，1F2-3，pp.1–4 (2007)
[4] N. Minematsu, Proc. ICASSP, pp.pp.889–892 (2005)
[5] 喬他，信学技報，SP2008-51，pp.49-54 (2008)
[6] 江森他，電子情報通信学会論文誌 D-II，vol.J83-DII，no.11，pp.2108–2117 (2000)
[7] S. Asakawa *et al.*, Proc. ICASSP, pp.4097–4100 (2008)
[8] 喬他，音講論，2-P-2 (2008-9)