# 音素列表象を併用した相対関係特徴を音響単位とする音響モデリング*

☆齋藤大輔, 松浦良, 峯松信明, 広瀬啓吉 (東大)

## 1 はじめに

音声は音韻のみならず，年齢や性別，声道長や音響機器などの諸要因により不可避的に変化し，これらの情報が伝達されることになる．言語的情報を抽出するという音声認識の目的からは，その他の情報は歪みと考えられる．これらの歪みに対処するため，従来の音声認識では大量データを用いた音響モデルの学習や，適応・正規化技術が広く用いられてきた．

近年，これらの歪みに影響を受けない音響的不変構造を用いた音声認識の枠組みが提案されている [1]．この枠組みでは音声の音響的実体そのものは直接用いず，その相対関係（距離情報）のみをモデル化することにより，声道特性や伝送系の違いなどの歪みに原理的に不変な音声認識を実現している [2, 3]．

上記の不変構造に基づく音声認識の枠組みは，ごく少数の学習話者によって不特定話者音声認識を実現できる．しかし単語単位のモデル化であるため，大語彙音声認識や任意語彙への対応などが困難という問題があった．従来の音声認識の枠組みでは，音素などのより細かい言語単位で音響モデルを用意し，未知語に対してもその音素列さえわかれば，当該単語の音響モデルを構成できる．このような観点から，筆者らは音響的不変構造に基づく音声認識においても，より細かい音響単位として音声事象間の相対関係に着目し，これらの単位を用いた単語音響モデルの再構成を検討してきた [4]．本稿では，学習データの音素列情報を相対関係のラベルとして併用する事で，先行研究と比べて，より柔軟な音響モデリングを検討する．提案手法は先行研究に対して教師ありモデリングの枠組みとして位置づけられ，本稿では先行研究との違いおよび融合の可能性についても考察する．

## 2 音声の構造的表象

### 2.1 非言語的特徴による音響的実体の歪み

音声の音響的実体は非言語的特徴によって不可避的に歪むが，これらは大きく乗算性歪みと線形変換性歪みに分けられる．

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る．ケプストラム空間では，この種の歪みは加算演算 $c' = c + b$ として表現される．マイクロフォンの音響特性差異がその典型例である．また話者の声道形状差異も一部近似的に乗算性歪みであると考えられる．音声は必ず発話者を伴い，音響機器によって収録されるため，これらの歪みは不可避である．

線形変換性歪みはケプストラム空間において行列 $A$ による線形変換 $c' = A\,c$ で表現される歪みである．スペクトル表現においては，話者の声道長差異や聴取者の聴覚特性差異は周波数ウォーピングとして考えられる．周波数ウォーピングはケプストラム空間において線形変換で記述されることが示されている [5]．すなわち声道長差異や聴覚特性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．

以上をまとめると，音声の音響的実体に不可避的に混入する非言語的特徴は，ケプストラム空間においてアフィン変換 $c' = A\,c + b$ で表現される．これらの $A, b$ が話者や収録環境によって多様に変化し，音声の音響的実体に様々な歪みが混入する事になる．

### 2.2 音声の構造的表象

ユークリッド空間において $N$ 角形の形状は ${}_NC_2$ 個の全ての頂点間距離を規定する事で一意に定めることができる．すなわち事象群に対して，全ての事象間距離を求めることでその事象群を構造的に表象することになる．しかしケプストラム空間において $N$ 点の「**点間距離**」によって構造を規定した場合，その構造は非言語的特徴によって不可避に歪む．なぜなら，非言語的特徴はケプストラム空間における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場合を除けば，構造を歪ませる変換である為である．しかしこの不可避に歪む構造は空間自体を歪ませる事で不変構造として定義することができる．

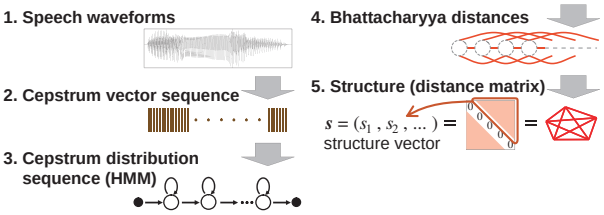
「**分布間距離**」の一つである Bhattacharyya 距離 (以下 BD と記述) を考えた場合，任意の二つの分布の確率密度関数を $p_1(x), p_2(x)$ として以下で表される．

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

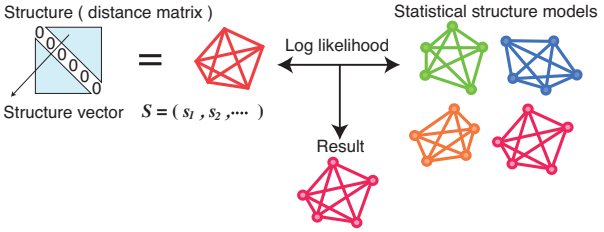
二つの分布に対して共通のアフィン変換 $Ac + b$ を施した場合，BD は変換前後で不変となる．なおこの不変性は非線形変換においても成立する [6]．

すなわちケプストラム空間において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

* “Acoustic modeling using speech contrast as its unit and phoneme sequence as labels” by SAITO Daisuke, MATSUURA Ryo, MINEMATSU Nobuaki, and HIROSE Keikichi (The Univ. of Tokyo)




(a): 一発声からの構造抽出




(b): 構造的音声認識の枠組み

Fig. 1　発声の構造化と構造的音声認識の枠組み

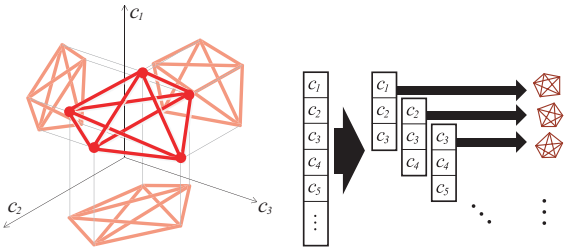



Fig. 2　特徴量空間分割によるマルチストリーム化

よって定義することで，変換不変，すなわち非言語性歪みにおよそ不変な構造を求める事ができる．

### 2.3　一発声の構造化と構造的音声認識の枠組み

一発声を一つの構造的表象で記述する場合を考える．Fig. 1(a) に一発声の音声からの構造的表象の抽出の流れを示す．音声の時系列信号は，まず短時間スペクトル系列からケプストラム系列へと変換される．得られたケプストラム系列もまた時系列信号であるが，これを適当な時間区間において音響事象の分布としてとらえ，その分布の時系列へと変換する（このとき各分布に対応する時間長は分布によって異なる）．これら系列中の各分布に対して全ての組み合わせの分布間距離を求めることで一発声が構造化される．

構造化された発声は距離行列の上三角成分をベクトルとみなすことで記述でき（構造ベクトル），これらを統計的にモデル化する事で単語毎の構造モデルを得る．認識時には，入力発声の構造ベクトルを最尤に出力する単語モデルを認識結果とする（Fig. 1(b)）．

### 2.4　特徴量空間分割

構造の不変性は変換の線形・非線形を問わず成立し，分布がガウス分布である場合はあらゆるアフィン変換に対して不変性が成り立つ．構造的表象に基づく音声認識はこの不変性に依って話者性を効果的に取り除く枠組みであるが，一方で不変性によって単語間の差異を表すような変換に対しても不変となり，単語識別力の低下を招く可能性がある．そのため不変性を抑制する必要がある．ここでは，線形変換性歪みを表す行列 $A$ について，その帯行列性に着目する [3]．ケプストラムベクトルに対して，連続する $w$ 個の要素から構成される $w$ 次元ベクトルを構成し，これを低次から要素をずらす事で複数の特徴量ストリームを構成する．そして，構造不変性が各部分空間においても成立すると仮定し，各ストリームから構造を抽出する（Fig. 2）．認識時には全てのストリームで構造比較を行い，尤度の合計により評価を行う．これにより過剰な不変性を抑制する事で単語識別力を向上させることができる [3]．以降 $w$ をブロックサイズ（BS）と呼ぶ．

## 3　相対関係を音響単位とするモデリング

### 3.1　クラスタリングによるパラメータ共有

構造に基づく音声認識では，単語単位でのモデリングであるため，大語彙音声認識への拡張や辞書にない単語に対する認識が困難という問題点がある．従来の音声認識では音素や音素片などより細かな単位で音響モデルを構築する事で，これらの問題を解決している [7]．例えば新しく辞書に登録する単語に関して，その語の音素表記が未知でも，登録用発声を音素単位の音響モデルで連続音素認識し，結果の音素列を辞書に登録する手法が検討されている [8]．

構造に基づく音声認識においても最小単位の導出を考えることでより柔軟なモデリングが可能となる．この観点から筆者らは，構造を構成する最小構成要素である音声事象間の相対関係を音響単位としてモデル化し，全語彙を用いてこの単位を適切に学習する枠組みを検討してきた [4, 9]．今，$N$ 個の音響事象列で単語を表し，ストリーム数 $s$ で特徴量空間分割した上で構造化する場合を考える．ある 2 つの音響事象に着目すると，これらの間の相対関係は $s$ 個の事象間距離で表す事ができる．この $s$ 次元のベクトルを以下では幾何学的な解釈から「エッジベクトル」と呼ぶ．このとき一つの単語は ${}_NC_2$ 個の $s$ 次元エッジベクトルによって構成されることになる．エッジベクトルが類似している事は，音響事象間の相対関係が似ている事に相当し，全単語に含まれるエッジベクトルのうち，類似しているものを同一のモデルで共有する事でより効率的な学習が可能になる（共有モデル）．この共有モデルは音響単位として用いる事ができる．共有の実装はエッジベクトルのベクトル空間における，K-means クラスタリングによって行った．この枠組

みによる学習および認識の流れは以下のようになる．

1. 認識したい単語毎に $N$ 個の分布系列に変換後，特徴量空間分割に基づく構造化により構造抽出を行い，構造統計モデルを作成する．
2. 単語数 $W$ としたとき，構造統計モデルの各エッジの平均ベクトル（計 $W \times_N C_2$ 個）に対して，クラスタ数 $K$ で K-means クラスタリングを行う（エッジ群に対するパラメータ共有）．
3. 1. で用いた全学習データを，2. で得られた $K$ 個のクラスタに分類し，共有エッジモデルを学習する．各単語モデルを共有関係と $K$ 個の共有エッジモデルから再構成する．
4. 従来の構造音声認識と同様に単語認識を行う．

上記の枠組みによって，先行研究 [4] では，認識率を向上させるとともに，登録用発声を用いて単語モデルを追加する事が可能となっている．すなわち相対関係に対して音響単位を構成する事で，話者不変性を有する構造に基づく音声認識の枠組みに，モデル拡張性を加えていることになる．

### 3.2 音素列表象を併用した相対関係のモデリング

前述のクラスタリングに基づく共有エッジモデリングにより，構造に基づく音声認識をより柔軟に拡張する事が可能になる．しかし未知語の登録の際には登録用発声を必要とするなど，既存の音素 HMM に基づく音響モデリングと比べると，まだ十分ではないといえる．従来の音声認識は音素列のラベル情報[1]をもとに，そのラベルに対応する音響モデルを組み合わせる事で，より大きな単位のモデルを構築することが可能である．一方，構造に基づく音声認識においても相対関係に何らかのラベルが存在すれば，そのラベルに対応するエッジモデルを組み合わせて，モデルを拡張する事ができると考えられる．

先行研究のパラメータ共有において，K-means クラスタリングはクラスの情報をエッジベクトルに割り当てていると解釈する事ができる．すなわち教師なしの枠組みによって相対関係のラベル（$K$ 個のカテゴリ）を導出していると考えられる．一方，個々の相対関係に明示的にラベルを付与する事で，教師ありの枠組みで相対関係をモデル化する事ができる．そこで本研究では音素列表象を利用して相対関係に対してラベル情報を付与することを考える．音素列表記は絶対量に対するラベル付けと考えられ，そのままでは相対関係に対するラベルではない．今回，特徴量として分布間の距離をとってエッジベクトルを導出

[1]本稿におけるラベル付けとは対象となる現象を有限個のシンボル・カテゴリの系列・配置で表す事を意味する．

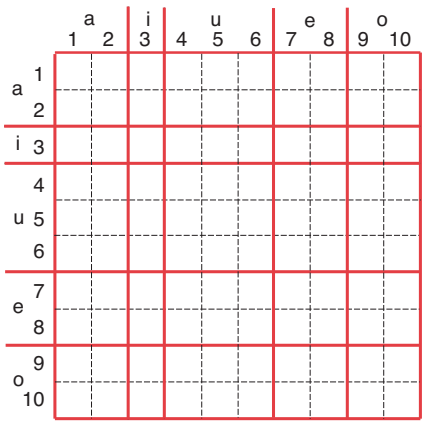



Fig. 3　音素とエッジの対応関係

するのと同様，ラベルとして，音素のペアを用いることを検討する．

例として，Fig. 3 のように発声/aiueo/ を考える．これは一発声が構造化された距離行列を表しており，小さな格子が一つの分布間相対関係を表しているとする．一方，赤い大きな格子は構造を表す距離行列上に対応する音素情報をマッピングしたものである．音素ラベルのペアをラベルと考えれば，赤い格子で囲われた事象間相対関係に当該ラベルを付与していることになる．例えば (1, 3)，(2, 3)，(3, 1), (3, 2) の格子には “a-i” というラベルが付与される．このラベル情報に基づいて，各音素ペア毎にエッジベクトルを確率モデルでモデル化すればよい．単語モデルの構築時には音素表記から Fig 3 のような距離行列を構築し，それぞれの格子に学習した確率モデルを割り当てる．このとき，各音素の状態数は学習データから決定し，今回の認識タスクにおいて/sil/と/a/ を 3 状態，/i//u//e//o/ を，各々4 状態とした．/sil/ は単語前後の無音区間に対応する．

## 4 実験

### 4.1 実験条件

提案手法の有効性を確かめるため，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 母音によって構成される連続発声母音系列（単語数 120 単語）を各 5 回収録した．構造化のための音響分析条件を Table 1 に示す．音素列ラベリングとして，構造化のための HMM とは別に 7 状態 HMM を 1 発声毎に学習し，そのビタビアライメントと構造化時の分布状態との関係から導出した[2]．学習データとして男女各 4 名の 120 単語，各 5 発声ずつを用いて，音素ペアモデルを学習した．各音素ペアモデルに対応するエッジベクトルのモデル化として混合数 1〜512 までの対角共分散 GMM を用いた．認識時には，学習に

[2]各状態が前後の無音及び母音に対応するとした．

Table 1　構造化のための音響分析条件

| | |
|---|---|
| サンプリング | 16 bit / 16 kHz |
| 窓 | 25 ms ハミング窓 / 10 ms シフト |
| ケプストラム | MCEP (0–12) |
| 状態数 | MAP 推定に基づく 25 状態 HMM |
| BS | 2（=12 次元エッジベクトル） |

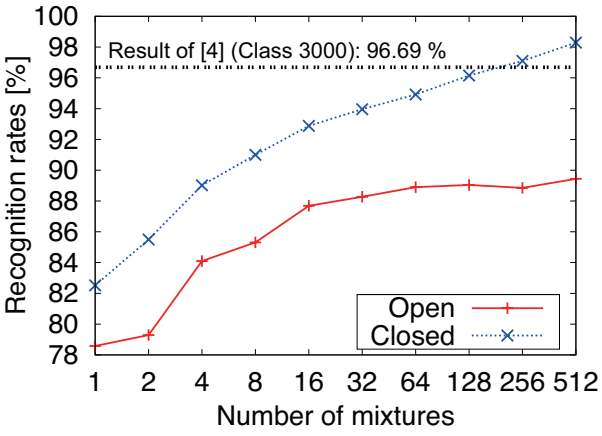



Fig. 4　認識結果

用いた男女各 4 名の 120 単語各 5 発声（Closed），学習時とは異なる成人男女各 4 名の 120 単語，各 5 発声（Open）を評価した．

### 4.2　実験結果

認識実験の結果を Fig. 4 に示す．Fig. 4 において，点線は先行研究 [4] の K-means クラスタリングによって共有を行った結果を表す．Open における提案手法の最高性能は混合数 512 の時，89.44 % であった．Closed の場合に比べて 16 混合から認識率がおよそ収束している．先行研究の結果には及ばないものの，先行研究が単語単位のモデル化（分布数 ${}_{25}C_2=300$）であるのに比べて，提案手法は音素ペア単位でモデル化を行っているため（音素ペア数 ${}_6C_2=21$），タスクオープンかつコンパクトなモデルとなっており，より汎用であるといえる．

## 5　考察

先行研究と提案手法のモデル化の違いについて考察する．先行研究におけるクラスタリングはエッジベクトルの類似度に基づいており，相対関係の物理的側面をモデル化しているといえる．一方提案手法のラベリングは音素ペアに基づいており，文字表記から得られる言語的な相対関係をモデル化している．Fig. 5 は両手法のラベルを単語/aiueo/ に付与したものである．両者を比較するとおよそ音素ペアの格子に沿ってクラスタの関係（距離の大小）が変化しているが，格子内に異なるエッジモデルのクラスタも混じっている．また対角項は異なる音素ペアラベルであるが，それぞれ同一音素のペアのため，距離の小さい値で共通している．このように両手法は相対関係を異なる視点からモデル化しており，これらを融合する事でより効果的なモデル化が可能になると考えられる．

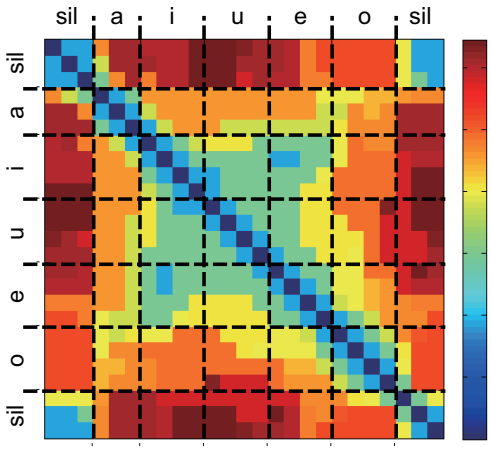



Fig. 5　クラスタリングによるモデリング ($K=20$) と音素ペア (${}_6C_2=21$) との関係（単語/aiueo/, 色は各クラスタの平均ベクトルの絶対値で付与されている.）

## 6　おわりに

本稿では，構造的表象に基づく音声認識において，音素列表象を併用した相対関係のモデル化を提案した．提案したモデリングでは，音素表記のペアを相対関係の明示的なラベルとして用い，教師ありの枠組みによって相対関係の音響単位を導出する．そのため音素表記から構造モデルを構築でき，より柔軟な認識の枠組みが実現できる．日本語 5 母音系列を用いた認識実験の結果，提案手法の有効性を示した．

今後の課題として，子音を含めた場合の検討が挙げられる．音素ペアのラベルは音素数の 2 乗で増大する為，クラスタリングとの融合を適切に行う事が重要となってくる．また提案手法はテキストから構造モデルを生成できることから，構造からの音声合成と組み合わせる事で，話者不変な表象を介したテキスト音声合成を今後検討していく予定である [10]．

## 参考文献


[1] N. Minematsu et al., Proc. ICASSP, pp.889–892, 2005.
[2] Y. Qiao et al., Proc. ASRU2007, pp.576–581, 2007.
[3] S. Asakawa et al., ICASSP2008, pp.4097–4100, 2008.
[4] 齋藤他, 信学技報, SP2009-77, pp.7–12, 2009.
[5] M. Pitz, H. Ney, IEEE Trans. Speech and Audio Processing, Vol.13, pp.930–944, 2005.
[6] 峯松他, 音講論（春）, 1-P-12, 2007.
[7] 田中他, 日本音響学会誌, vol. 42, No. 11, pp.860–868, 1986.
[8] 中川他, 電気学会論文誌 C, vol.118-C, No.6, pp.865–872, 1998.
[9] 松浦他, 音講論（秋）, 2-P-4, 2008.
[10] D. Saito et al., INTERSPEECH, pp.1837–1840, 2008.