音声事象の相対関係を音響単位とした未知語用音響モデルに関する実験的検討*

◎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, 2]．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 [3, 4]．

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．そこで本研究では，音響的不変構造に基づく音声認識において，より細かい音響事象を基本単位として用意しておき，辞書にない単語音声をこれらの音響単位によってモデル化する枠組みについて検討する．本稿では，パラメータ共有によって得られた音声事象間の相対関係（事象間距離）を用いて，未知単語の単語モデルを構築する手法を提案し，その有効性を示す．

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．

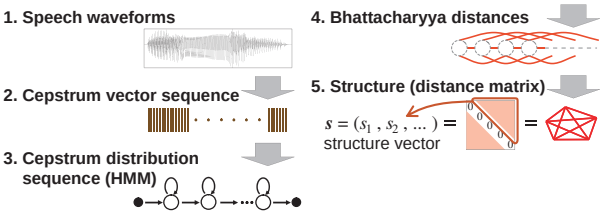
$$BD(p_1, p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x) p_2(x)}\, dx \qquad (1)$$

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]．

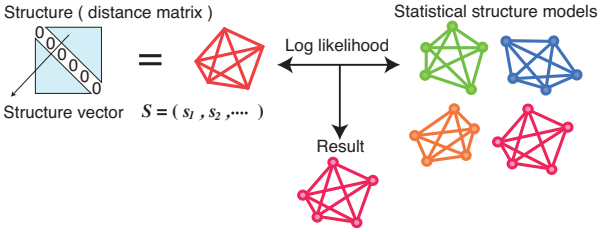
すなわちケプストラム空間において音響事象を分布として捉え，音響事象群を「**分布間距離**」のみによって定義することで，変換不変，すなわち非言語性歪みにおよそ不変な構造を求める事ができる．

---

*Experimental study of acoustic models of OOV words using speech contrast as a unit of acoustic modeling.
by SAITO Daisuke, MATSUURA Ryo, MINEMATSU Nobuaki, and HIROSE Keikichi (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

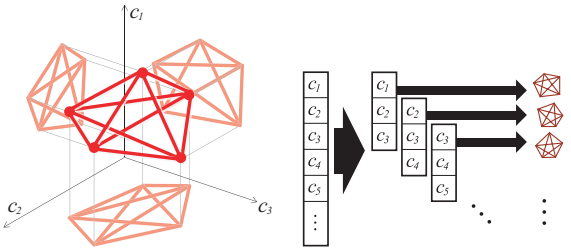


Fig. 2　特徴量空間分割によるマルチストリーム化

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 [4]．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．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 [4]．以降 w をブロックサイズ（BS）と呼ぶ．Fig. 2 は特徴量空間分割の一例を示している．Fig. 2 の左図のように射影された部分空間での構造も制約条件として含めることで，構造の回転に制約を与えることになる．

3　相対関係による音響単位の導出

3.1　クラスタリングによるパラメータ共有

特徴量空間分割による構造表現は単語識別力を大幅に向上させる一方で，パラメータ数の増大に伴う次元の呪いの問題が顕著になる．また本研究における認識タスクは日本語 5 母音によって構成される 120 単語の孤立単語音声認識となっている．このため個々の構造ベクトルの要素間には，冗長性が存在することが示唆され，構造統計モデルのパラメータを共有しコンパクト化することで効率的な学習が可能になることが考えられる．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響事象間の相対関係（エッジ）について，全単語モデルに渡って適切に共有することで認識率の向上を図ってきた [7]．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．これを s 次元のベクトル（エッジベクトル）と考えれば，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．この共有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実装した．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，

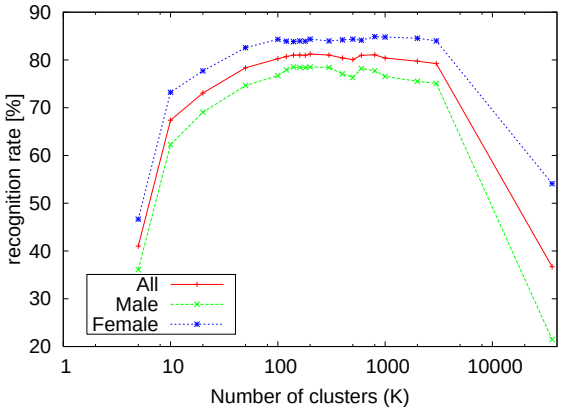


Fig. 3　共有エッジモデルによる認識率の向上

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. 2. で得られた共有関係に基づいて，1. で用いた全ての学習データを 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．

Fig. 3 は上記の枠組みによる認識率とクラス数の関係について図示したものである．ここでは女性話者 1 名の日本語 5 母音による 120 単語について，各単語 5 発声を学習データとしている．分布数は $N=25$，ストリームの個数については $BS=2$ で，12 次元のエッジベクトルによりモデル化した．共有前のエッジモデル数は $120 \times_{25} C_2 = 36000$ であり，これを上記の枠組みでパラメータ共有する．評価は男性話者 8 名，女性話者 7 名の各 5 発声を用いている．図中の “All” が全体の認識率であり，クラス数 200 のときに 81.24 % の認識率を示している．非共有時は 38.05 % の認識率であり，上記の枠組みによって大幅に性能向上することが分かる．

3.2　共有エッジモデルによる未知語のモデリング

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 [8]．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 [9]．

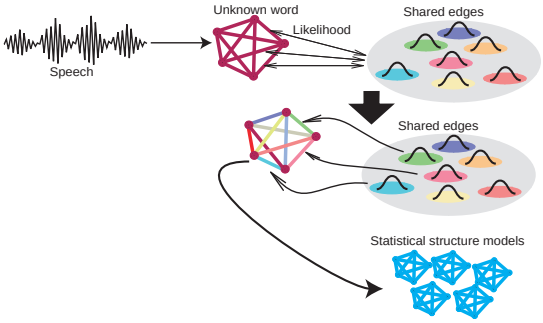


Fig. 4　共有エッジに基づく未知語モデリング

今，構造に基づく音声認識においても同様の枠組みを検討する．前節における議論により，単語辞書全体で相対関係の情報を共有する事が可能であることを示した．そこで共有エッジモデルを最小の音響単位と考え，未知語をこの単位で表現することを考える．すなわち未知語の登録用発声を構造化し，それぞれのエッジについて，最尤な共有エッジモデルによって単語の構造統計モデルを構成する．共有エッジモデルの学習時に適切な相対関係がモデル化できれば，構造に基づく音声認識においても拡張性の高い枠組みが実現可能となる．提案する未知語モデリングの概念図を Fig. 4 に示す．

3.3　混合数増加による歪みの緩和

音響的不変構造は話者性や伝送系といった音声に対する静的歪みに対して不変であるが，発話スタイルの違いと言った動的な歪みによって影響を受ける．また特徴量空間分割による構造化は単語識別力を向上させる一方で，話者性に対する不変性をわずかながら低下させている可能性もある．そこでこれらの歪みによる共有エッジモデルの揺らぎを緩和するため，共有関係から統計モデルを構築する際，混合ガウス分布を用いてモデル化することで，これらの問題に対処することを検討した．

4　実験

4.1　実験条件

共有エッジモデルによる未知語モデリング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．共有エッジモデルの学習には男女各 4 名の 100 単語，各 5 発声ずつを用い，クラス数 50 として共有

Table 2　認識結果（値は認識率 (%)）

話者数	対角			全角			混合数 2			混合数 4		
	all	male	female	all	male	female	all	male	female	all	male	female
2	45.75	27.25	64.25	46.75	32.75	60.75	**65.50**	**63.50**	**67.50**	61.50	57.25	65.75
4	77.62	72.00	83.25	66.75	55.50	78.00	**87.12**	**89.25**	**85.00**	86.50	87.25	85.75
8	85.62	86.75	84.50	68.88	58.75	79.00	**91.50**	**95.50**	**87.50**	89.12	93.00	85.25

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）

エッジモデルを作成した．エッジモデルの分布形状として，1) 対角分散共分散ガウス分布，2) 全角分散共分散ガウス分布，3) 2 混合ガウス分布，4) 4 混合ガウス分布 の 4 通りを用いた．なお 3) 4) において各ガウス分布は対角分散共分散とした．未知語として残り 20 単語を想定し，登録用発声の話者数を 2, 4, 8（1 話者 1 発声）と変化させて共有エッジモデルから 20 単語の単語モデルを作成した．認識時には，学習時とは異なる成人男女各 4 名の 20 単語，各 5 発声を評価音声とした．ただし単語パープレキシティは共有エッジモデル学習に用いた単語を含めた 120 とした．

4.2　実験結果

認識実験の結果を Table 2 に示す．Table 2 において，最左の列は単語登録に用いた話者数を表している．登録のための話者数を増やすにつれて，登録用発声の揺らぎが解消され，単語モデルが安定するため認識率が向上している．一方，対角分散共分散，全角分散共分散による共有エッジモデルに比べ，混合ガウス分布を用いた場合に大幅に性能向上していることがわかる．また混合ガウス分布によるモデルの場合，性別の違いによる認識率にあまり差がない．これは，発話スタイルなどの動的歪みの緩和に加えて，特徴量分割による話者不変性の低下に対してもある程度対処できていることを示唆している．全体的には 2 混合ガウス分布による共有エッジモデルの性能が最も高く，4 話者の発声による登録で 87.12 %，8 話者での登録で 91.50 % の認識率となった．なお，共有エッジモデルの学習に用いた 100 単語に対する認識率は約 95 % であった．このことから共有エッジモデルを用いることで，認識率をある程度保ったまま，構造に基づく音声認識をより柔軟に拡張しうることがわかった．

5　おわりに

本稿では，構造的表象に基づく音声認識における，大語彙音声認識や語彙の拡張が困難であるという問題に対して，音響事象の相対関係を基本単位とする音響モデリングを提案した．提案したモデリングでは，エッジベクトル間に存在する共有関係をクラスタリングによって抽出し，共有エッジモデルとしてモデル化する．このモデルは効率的に学習データを用いる事で認識率を大幅に向上させることができる．さらに，共有エッジモデルを基本単位として扱い，登録用発声に各モデルを割り当てることで新しい単語を登録して認識可能である事を示した．さらに発話の静的，動的歪みに対処するため，共有エッジモデルを混合ガウス分布でモデル化し，その有効性を示した．

今後の課題として，より少ない登録発声によって頑健な単語モデルを構築する事があげられる．そのためには，特徴量の次元削減や適切な登録手法を検討する必要がある．また，今回は教師なしの枠組みである「クラスタリング」によって共有情報を抽出したが，テキストとの対応関係などを念頭に，教師有りの枠組みによる相対関係のモデル化も検討可能である．また本稿で議論した共有関係の抽出は，幼児の発達における語ゲシュタルトの獲得と音素獲得との間をつなぐ上で重要な役割を果たすとも考えられる [10, 11, 12]．これらの知見を活かして，構造からの音声合成における利用についても検討したい [13]．

参考文献

[1] N. Minematsu et al., Proc. ICASSP, pp.889–892, 2005.
[2] N. Minematsu et al., Proc. SRIV'2006, pp.47–52, 2006.
[3] Y. Qiao et al., Proc. ASRU2007, pp.576–581, 2007.
[4] S. Asakawa et al., ICASSP2008, pp.4097–4100, 2008.
[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] 松浦他, 音講論（秋）, 2-P-4, 2008.
[8] 田中他, 日本音響学会誌, vol. 42, No. 11, pp.860–868, 1986.
[9] 中川他, 電気学会論文誌 C, vol.118-C, No.6, pp.865–872, 1998.
[10] 天野, "子どものかな文字の習得過程", 秋山書店, 1986.
[11] 早川, 月刊言語,35,9,pp.62–67,2006.
[12] N. S. トルベツコイ, "音韻論の原理", 岩波書店, 1958.
[13] D. Saito et al., INTERSPEECH, pp.1837–1840, 2008.