社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

話者不変な相対関係特徴を音響単位とする音響モデリングに関する実験的検討

齋藤　大輔† 　松浦　良†† 　峯松　信明††† 　広瀬　啓吉†††

† 東京大学大学院工学系研究科 〒113–8656 東京都文京区本郷 7–3–1
†† 東京大学大学院新領域創成科学研究科 〒113–8656 東京都文京区本郷 7–3–1
††† 東京大学大学院情報理工学系研究科 〒113–0033 東京都文京区本郷 7–3–1
E-mail: {dsk_saito,matsuura,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　音声は年齢，性別，声道長や音響機器などの非言語的要因によって変形し，多様性に富んでいる．筆者らはこれらの非言語的な音響変形におよそ不変な音声の構造的・抽象的表象を提案してきた．この表象は音声の動きのみに着眼した物理表象である．先行研究において，音声の構造的表象に基づく音声認識について種々の検討を行ってきた．構造に基づく音声認識の問題点として，特徴量に対する次元の呪いと，大語彙音声認識や任意語彙への対応が難しい，という点が挙げられる．本報ではこれらの問題に対処するため，パラメータ共有学習に基づく学習の効率化と得られた共有モデルを用いた単語音響モデリングを提案する．相対関係を記述するエッジベクトルを考え，エッジベクトル空間におけるボトムアップクラスタリングによって共有関係をモデル化する．一方，新規単語をモデル化する際には，登録用発声に対して共有エッジモデルを割り当てることでモデル化する．日本語 5 母音連結発声による孤立単語音声認識実験を行い，提案手法の有効性について確認した．

キーワード　構造的表象, 話者不変特徴, 相対関係, クラスタリング, 音響モデリング

Experimental study of acoustic modeling using speaker-invariant speech contrast as modeling unit

Daisuke SAITO†, Ryo MATSUURA††, Nobuaki MINEMATSU†††, and Keikichi HIROSE†††

† Graduate School of Engineering, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
†† Graduate School of Frontier Sciences, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
††† Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan
E-mail: {dsk_saito,matsuura,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Speech acoustics vary due to differences in age, gender, vocal tract length, microphone, and so on. The authors recently proposed a structural and abstract representation of speech, where these variations were effectively removed. This representation captures only dynamics of speech. In our previous studies, using this abstract representation, an ASR framework, which we call the structure-based ASR, was proposed and examined. However, there are two problems for the structure-based ASR; the curse of dimensionality and the large size of modeling unit. As a solution for these problems, this report proposes a new acoustic modeling based on parameter sharing of statistical structure models and efficient reuse of the shared models. In the proposed method, edge vectors, which represent speech contrasts between any two acoustic events, are considered and parameter sharing is carried out based on clustering in the parametric space of edge vectors. To construct an acoustic model for a new word, the most likely edge models are selected and allocated for each edge vector in the structure of that new word. Experiments of recognition using continuous utterances of Japanese vowels show the validity of the proposed method.

Key words　structural representation, invariant features, speech contrasts, clustering, acoustic modeling

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，これらの情報が統合的に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これらの諸要因の変化をカバーするような大量データを用いた音響モデルの学習が行われる．さらに上記の歪みによって発生する，音響モデルと入力音声とのミスマッチに対して，ケプストラム平均除去法や声道長正規法，雑音抑圧等の技術により性能向上が図られてきた [1]．従来の音声認識では，音声に対する様々な歪みによって，音声を表す特徴量が変化する事を前提に技術が構築されており，その変化を緩和する目的で種々の技術が提案されてきたといえる．

一方近年，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 [2], [3]．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 [4], [5]．この枠組みは従来とは異なり，歪みが存在しても変化しない音声の特徴記述を提案しており，人間の頑健な知覚との対応が様々な観点から議論されている [6], [7]．

音響的不変構造による音声認識は，相対関係をモデル化するというその特徴量記述の特性から，次元の呪いの問題がより顕著になる [8]．例えば N 個の音響的実体があった場合，その相対関係は ${}_NC_2$ 個となり，特徴量の次元は $O(N^2)$ のオーダーで増加する．一方でこれらの相対関係には高い相関関係が存在し，そのため認識においては冗長な記述となっていることが考えられる．一般にパターン認識の分野においては，このようなデータの高次元性に起因する問題に対して，主成分分析（Principal Componet Analysis; PCA）や線形判別分析（Linear Discriminant Analysis; LDA）による次元削減が有効であることが示されている．我々の先行研究においても PCA と LDA を組み合わせた手法 [9] や LDA を 2 段階で適用する手法 [10] によって次元圧縮を行うことで認識率が大幅に向上する事が示されている．しかし線形変換に基づくこれらの次元圧縮手法では，個々の相対関係のもつ物理的，音声的意味づけが不明確になる問題点があった．

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である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．一方で不変構造に基づく音声認識ではこのような柔軟な音響モデリングが困難であり，単語認識のためのモデル学習にはそれぞれの単語の発声を必要としていた．

そこで本研究では，上述の問題点に対応する音響モデリングを検討する．物理的意味の明確な次元圧縮手法として，学習すべきパラメータを共有することによって，特徴量の意味付けを保持したまま特徴の冗長性を削減する方法を提案する．また，パラメータ共有に基づく冗長性削減を積極的に利用することで，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導出する．これにより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，提案手法の詳細を説明し，日本語 5 母音単語を用いた認識実験によって，従来の不変構造を用いた音声認識とほぼ同等の性能で，より柔軟な音響モデリングが可能である事を示す．

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 [11]．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．

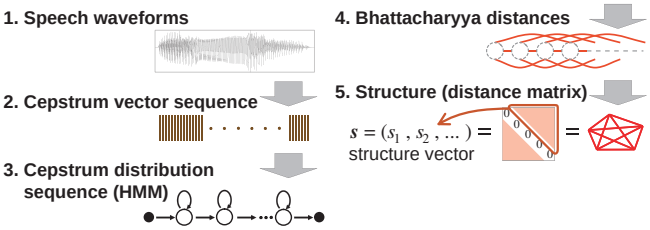
2.2 音声の構造的表象

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．

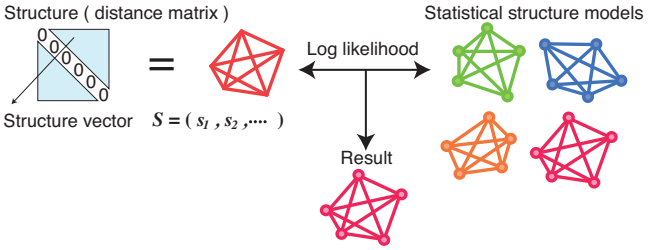
「**分布間距離**」の一つである Bhattacharyya 距離 (以下 BD と記述) を考えた場合，任意の二つの分布の確率密度関数を $p_1(x), p_2(x)$ として以下で表される．

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

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，



(a): 一発声からの構造抽出 (Structure extraction)



(b): 構造的音声認識の枠組み (Framework of the structural ASR)

図 1　発声の構造化と構造的音声認識の枠組み

Fig. 1　Structure extraction and structure-based ASR.

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 [12](注1)．

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．

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．図 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 (図 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

(注1)：この性質は分布間距離として f-ダイバージェンスと呼ばれる特徴量を用いた場合，変換の線形性に依らず成立する [13]．上述の Bhattacharyya 距離は f-ダイバージェンスの一種である．

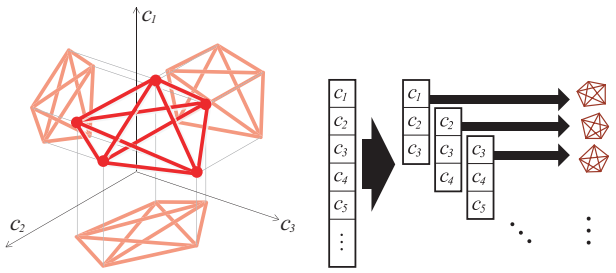


図 2　特徴量空間分割によるマルチストリーム化

Fig. 2　Multiple stream structuralization by parameter division.

ついて，その帯行列性に着目する [5]．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 [5]．以降 w をブロックサイズ (BS) と呼ぶ．図 2 は特徴量空間分割の一例を示している．図 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る．一般に次元の呪いへの対応として，PCA や LDA に代表される，線形変換に基づく次元圧縮が広く用いられる．しかしこれらの手法では変換後の座標系における基底の物理的な意味付けが明確でない場合も多い．

一方，音声認識におけるデータスパースネスの問題は次元の呪いの一種と考える事ができる．音声認識研究においては，トップダウンなコンテキストクラスタリングを用いる事で triphone モデルの複雑性を緩和し，モデル学習を効率化させる手法が広く一般に用いられている [14]．この観点を構造に基づく音声認識に導入する事で，筆者らは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を可能にする手法を検討してきた [15]．本稿ではこの枠組みを音響単位の導出過程として再定式化する．

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．このとき一つの発声は，s 個の構造ベクトルによって表現される事になる．あるストリームに着目すると，部分空間における幾何構造に対応するこれらの構造ベクトルの次元数は ${}_NC_2$ となる．一方で，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各ストリームにおける事象間距離を用いて，s 個の距離値に

よって表現できる．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．このエッジベクトルは事象間の相対関係を記述する多次元特徴量と考える事ができる．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．ただし，従来の HMM におけるコンテキストクラスタリングと異なり，相対関係に対する明示的なラベルが存在しないため，トップダウンクラスタリングによるパラメータ共有を行うことができない．そこで今回，パラメータ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，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．このときストリーム数は s とする．

（2） 学習データの単語数を W とする．構造統計モデルの各エッジの平均エッジベクトル（計 $W \times {}_NC_2$ 個）に対して，クラスタ数 K で K-means クラスタリングを行う．これはエッジベクトル空間上の類似度に基づいて，構造統計モデルのモデルパラメータを共有していることに相当する．

（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，各クラスタ毎にエッジベクトルの分布を学習する．このように構成されたエッジベクトルの統計モデルを以下では共有エッジモデルと呼ぶ．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） 再構築した単語モデルを用いて，従来の構造音声認識と同様に単語認識を行う．

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 [16]．この際，登録したい単語に対してその音素表記が得られていれば，細かな単位の音響モデルを繋ぎ合わせることで当該単語のモデルを構築できる．また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 [17]．

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おいて，単語辞書全体で相対関係の情報を共有する枠組みについて提案した．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図 3 に示す．

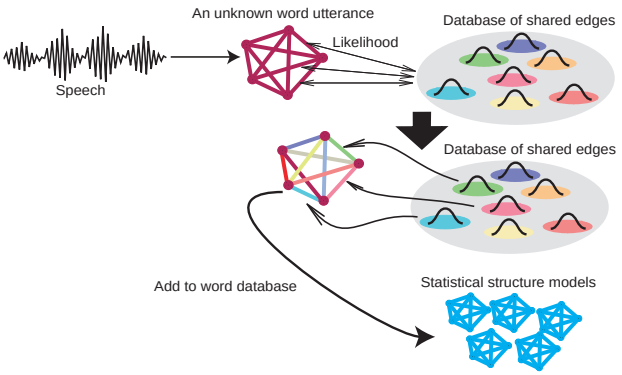


図 3 共有エッジに基づく未知語モデリング
Fig. 3 Efficient reuse of the shared edge models.

表 1 構造化のための音響分析条件
Table 1 Conditions of acoustic analysis for structuralization.

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 ($= N$) 状態 HMM
BS	2（$s = 12$ 次元エッジベクトル）

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．総発声数は $16 \times 120 \times 5 = 9,600$ となる．以下では学習に用いる異なり単語数を W，共有モデルのクラスタ数を K で表し，これらのパラメータ及び未知語登録に用いる発声数，学習話者数を変化させて実験を行った．

まず，パラメータ共有による次元削減の効果を確かめる実験（以下実験 A とする）を行った．単語数は $W = 120$ とし，女性話者 1 名の各単語 5 発声を学習データとした．クラスタ数 K を 50 から 3,000 まで変化させて認識率を調べた．評価は男性話者 8 名，女性話者 7 名の各 5 発声を用いた．

さらに未知語のモデリングに関して，本手法の有効性を確かめる実験（以下実験 B）を行った．共有エッジモデルの学習には男女各 4 名の $W = 100$ 単語，各 5 発声ずつを用い，クラス

表 2　実験 B の認識結果（値は認識率 (%)）．$PP = 120$ の時の未知語 20 単語に対する認識率．

Table 2　Recognition rates for the 20 new words with word perplexity of 120.

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

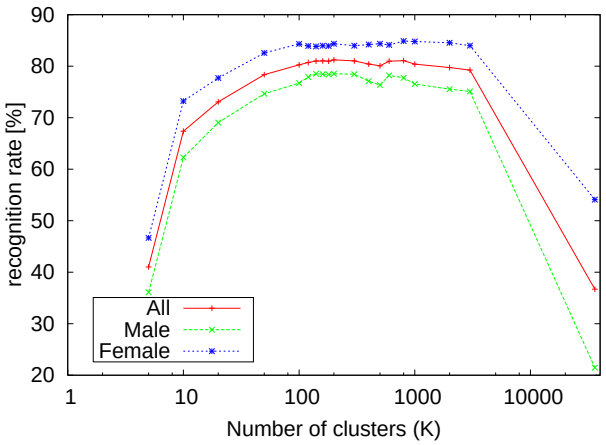


図 4　共有エッジモデルによる認識率の向上

Fig. 4　The performance of the shared edge models.

表 3　実験 C の認識結果（値は認識率 (%)）

Table 3　Recognition rates of experiment C.

発声数	all	male	female
1	59.33	57.25	61.71
2	73.60	69.50	78.29
3	77.00	74.12	80.29

タ数 $K = 50$ として共有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，登録用発声の学習話者数を 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．

最後に学習話者数を減少させた条件で未知語登録の有効性を確認する実験（以下実験 C）を行った．共有エッジモデルの学習に女性話者 1 名の $W = 100$ 単語，各 5 発声を用いた．クラスタ数は $K = 50$ とした．エッジモデルの分布形状として 2 混合ガウス分布を用いた．残り 20 単語の未知語に対して，登録用発声の発声数を 1, 2, 3 と変化させた．なお登録用発声は学習時の女性話者 1 名のものを用いた．

4.2　実 験 結 果

まず，実験 A の結果を図 4 に示す．図 4 はクラスタ数に対する認識率の変化を評価話者の性別毎にプロットしたものである．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．パラメータ共有を行わない場合，その認識率は 38.05 % であり，提案するパラメータ共有の枠組みによって大幅に性能向上することが分かる．なお HMM を音響モデルとし特徴量を MCEP とした単語音声認識において，学習話者 1 名の同一条件で状態共有を行ってモデル化した場合，その認識率は 67.8 % に留まった．さらに今回の結果では評価話者の性別差による認識率の差が 10 % 程度であるが，HMM を用いた場合はその差は 30 % ほどになった．すなわち構造に基づく音声認識によって，話者の違いが効果的に取り除かれていることが分かる．

次に実験 B の結果を表 2 に示す．表 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．

最後に，実験 C の結果を表 3 に示す．登録のための発声数を 1 から 2 に増やした所で，大幅な認識率の向上が確認できる．実験 B の場合と比べて，各登録発声の話者性の違いは非常に小さいため，登録発声を増やした事によって，発話スタイル等の動的揺らぎが緩和される効果が表れたと考えられる．性能としては登録発声数が 3 の場合に 77 % の認識率となっており，実験 A の結果と比較すれば，同等の性能でより柔軟なモデルを構築できているといえる．

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 [18]〜[20]．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 [21]．

6. 謝　　辞

本研究は東京大学グローバル COE「セキュアライフ・エレクトロニクス」の若手研究ファンドの支援を受けて行われた(注2)．

文　　献

[1] 松本弘, “雑音環境下の音声認識手法”, 情報科学技術フォーラム FIT2003, 2003.
[2] N. Minematsu: “Mathematical evidence of the acoustic universal structure in speech,” Proc. ICASSP, pp. 889–892, 2005.
[3] N. Minematsu *et al.*: “Theorem of the invariant structure and its derivation of speech Gestalt,” Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp. 47–52, 2006.
[4] Y. Qiao *et al.*: “Random discriminant structure analysis for continous Japanese vowel recognition,” Proc. ASRU2007, pp. 576–581, 2007.
[5] S. Asakawa *et al.*: “Multi-stream parameterization for structural speech recognition,” ICASSP2008, pp. 4097–4100, 2008.
[6] N. Minematsu *et al.*: “Implementation of robust speech recognition by simulating infants’ speech perception based on the invariant sound shape embedded in utterances,” Proc. Speech and Computer, pp. 35–40, 2009.
[7] N. Minematsu: “A consideration of ASR based on animal evolution and human development – what should A of ASR stand for? –,” Proc. Int. Workshop on Computational Models of Language Evolution, Acquisition and Processing, 2009.
[8] A. K. Jain: “Statistical pattern recognition: A review,” IEEE Trans. PAMI, vol. 22, no. 1, pp. 4–37, 2000.
[9] Y. Qiao *et al.*: “On invariant structural representation for speech recognition: theoretical validation and experimental improvement,” Proc. INTERSPEECH, pp. 3055–3058, 2009.
[10] 朝川智他: “音声の構造的表象と判別分析を用いた単語音声認識”, 電子情報通信学会音声研究会, SP2008-113，pp. 203–208, 2008.
[11] M. Pitz *et al.*: “Vocal tract normalization equals linear transformation in cepstrum space,” IEEE Trans. Speech and Audio Processing, vol.13, pp. 930–944, 2005.
[12] 峯松信明他: “線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察”, 日本音響学会春季講演論文集, 1-P-12, pp. 147–148, 2007.
[13] Y. Qiao *et al.*: “f-divergence is a generalized invariant measure between distributions,” Proc. INTERSPEECH, pp. 1349–1352, 2008.
[14] S. Young *et al.*: “The HTK Book version 3.4.1,” `http://htk.eng.cam.ac.uk`, 2009.
[15] 松浦良他: “音声の構造的表象を用いた音声認識におけるパラメータ共有に関する考察”, 日本音響学会秋季講演論文集, 2-P-4, pp. 117–120, 2008.
[16] 田中和世他: “音声の音素片ネットワーク表現と時系列のセグメント化法を用いた自動ラベリング手法”, 日本音響学会誌, vol. 42, No. 11, pp. 860–868, 1986.
[17] 中川聖一他: “任意語彙の追加登録可能な単語音声認識システム”, 電気学会論文誌 C, vol. 118-C, No. 6, pp. 865–872, 1998.
[18] 天野清: “子どものかな文字の習得過程”, 秋山書店, 1986.
[19] 早川勝廣: “言語獲得と育児語”, 月刊言語, vol. 35, No. 9, pp. 62–67, 2006.
[20] N. S. トルベツコイ: “音韻論の原理”, 岩波書店, 1958.
[21] D. Saito *et al.*: “Optimal event search using a structural cost function – improvement of structure to speech conversion –,” Proc. INTERSPEECH, pp. 2047–2050, 2008.

(注2)：`http://www.ee.t.u-tokyo.ac.jp/gcoe`