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.

構造表象を用いた音声認識におけるパラメータ共有とその効果

松浦　良†　　齋藤　大輔††　　朝川　智†　　峯松　信明††　　広瀬　啓吉†††

† 東京大学大学院新領域創成科学研究科〒 277-8561 千葉県柏市柏の葉 5-1-5
†† 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
††† 東京大学大学院情報理工学系研究科〒 113-0033 東京都文京区本郷 7-3-1
E-mail: {matsuura,dsk_saito,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　性別、年齢、音響機器などの非言語的要因によって音声の音響特徴は不可避的に変動し、音声認識率を下げる要因となっているが、近年、これらの要因に影響されない音声の不変的・構造的表象が提案されている。これは音の絶対量ではなく相対量に基づく音声表象であり、音声ストリームを N 個の分布列へと変換した後に、${}_NC_2$ 個ある分布間距離、即ちエッジ長を計算する。語彙が L 語ある場合、合計 $L\times{}_NC_2$ 個のエッジ長を統計的にモデル化することになる。今回の報告では、不変表象であることを考慮し、学習話者一人の場合に構造的音声認識がどのような傾向を示すのかを実験的に検討する。構造表象は発話スタイルの影響を直接的に受ける。そのため、話者間の発話スタイル差異を吸収するためにパラメータ共有を導入し、その効果について検証した。また、従来の HMM を用いた枠組みでも同様の実験を行ない、パラメータ共有の効果について比較した。両実験結果を踏まえ、音声によって伝搬される言語情報は音の実体、或は、音の差異のいずれに符号化された情報なのか、について考察する。
キーワード　音声の構造的・不変的表象、音声認識、パラメータ共有、クラスタリング、音的実体と差異

Parameter sharing and its effects for structural speech recognition

R.MATSUURA†, D.SAITO††, S.ASAKAWA†, N.MINEMATSU†††, and K.HIROSE†††

†Grad. School of Frontier Sci., The Univ. of Tokyo, 5-1-5 Kashiwano-ha, Kashiwa-shi, Chiba 277-8561, Japan
†† Grad. School of Eng., The Univ. of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8658, Japan
†††Grad. School of Info. Sci. and Tech., The Univ. of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033,Japan
E-mail: {matsuura,dsk_saito,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Acoustic features in speech are affected inevitably by non-linguistic factors, which easily decrease speech recognition performance. Recently, we proposed a structural and speaker-invariant representation of speech, where speech substances are completely discarded and speech contrasts are only extracted. After converting an input speech stream into N feature distributions, every distance between any pair of the distributions is calculated. If a word has N distributions, its structure model comes to have ${}_NC_2$ parameters (edge length). If the vocabulary is comprised of L words, a recognition system will have $L\times{}_NC_2$ parameters. In this report, considering that the structural representation is speaker-invariant, we test it only with a single training speaker. To improve the robustness for variability of speaking styles, parameter sharing is introduced and its effect is examined. After a similar experiment with HMMs, we discuss whether linguistic information is encoded in speech substances or speech contrasts.
Key words　structural representation, speech recognition, parameter sharing, clustering, substance and contrast

1. はじめに

音声には、言語情報だけではなく、音響機器の特性、周囲の環境による雑音、話者の声道長の形状特性といった様々な歪みや雑音が混入しており、これらが音声認識率を下げる要因となっている。これらの要因に対処するためにケプストラム平均除去法、声道長正規化法といった手法が用いられているが [1]、どのような入力音声に対しても有効というわけではない。また、HMM を用いて不特定話者・環境の音響モデルを構築した場合、これらの歪み・雑音は平均化されるが、個々の利用環境・利用者は、某の方向性を持った歪みを有する。人間にとっては、平均声道長を持つ話者が最も明瞭な音声を発声する訳では無いが、

計算機にとってはこのような話者が最も明瞭な話者となる。

近年、これらの要因に影響を受けない、音声に内在する音響的に不変な構造を用いた音声認識の枠組みが提案されている [2], [3]。ここでは、個々の音を絶対的な特徴量では捉えず、音と音との関係（距離）をスカラー量として捉える。この枠組みに基づく音声認識は、不可避的な伝送歪みや声道長の形状特性に対して原理的に不変であり、頑健な音声認識が可能となる。日本語 5 母音を用いて構成された人工的な 120 単語セットに対してこれまで種々の検討を行なっており、少量の学習話者で高い認識精度が得られている [4]〜[6]。

本研究では、学習話者を 1 人まで落とした場合に本手法がどのような性能を示すのかについて実験的に検討する。本表象は発話スタイルには強い影響を受けるため、頑健性の向上を狙い、パラメータ共有の枠組みを導入する。HMM を用いた同様の実験も行ない、両者の比較を通して種々の考察を行なう。

2. 音声の構造的表象

2.1 音声に内在する非言語的特徴

音声に含まれる非言語的特徴は、大きく三種類ある。背景雑音、音響機器特性に対応する乗算性（畳み込み性）歪み、及び、話者の声道長・声道形状特性に対応する線形性歪みである。ここでは、不可避な歪みは後者の二種類であると考え、背景雑音については検討の対象としない。

乗算性歪みは、ケプストラムベクトル c に対するベクトル b の加算となる（$c'=c+b$）。また、声道長・形状の差異は、周波数領域では、スペクトルの周波数ウォーピング（フォルマントシフト）となるが、単調かつ連続なウォーピングは行列 A を掛ける演算へと変換可能である（$c'=Ac$) [7], [8]。例えば周波数ウォーピングを 1 次全域通過関数で実装した場合（図 1 参照）、行列 A（a_{ij}）は次のようになる。

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=m_0}^{j} \binom{j}{m} \frac{(m+i-1)!}{(m+i-j)!} (-1)^{(m+i-j)} \alpha^{(2m+i-j)}$$

但し、$|\alpha| \leqq 1.0$, $m_0 = \max(0, j-i)$ であり、また

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geqq m) \\ 0 & (j < m) \end{cases}$$

である。この行列は非零の要素が対角付近に集合する帯行列的な形態を示す。α は周波数ウォーピングのパラメータであり、$\alpha>0$ の場合フォルマント周波数は上がり、声道を縮める効果を持つ。$\alpha<0$ の場合は逆に作用する。

このように、乗算性歪みは c への加算、周波数ウォーピングは c に対する行列 A の掛算として表現されるため、不変表象を考える場合の必要条件は、ケプストラムに対するアフィン変換（線形変換）不変と考えることができる。

2.2 音声に内在する不変的な音響構造

音響事象空間において、音声の特徴量は時系列的にケプストラムの点軌跡で表される。HMM の学習アルゴリズムを用い、連続かつ類似したフレームをマージすることで、この軌跡を N 個の分布へと変換する。そして、${}_NC_2$ 個存在する全分布間距離

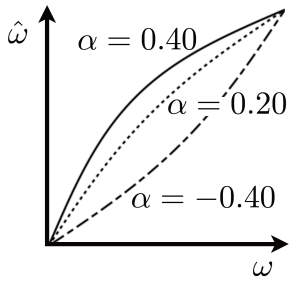

図 1　周波数ウォーピング（$-0.4 < \alpha < 0.4$）

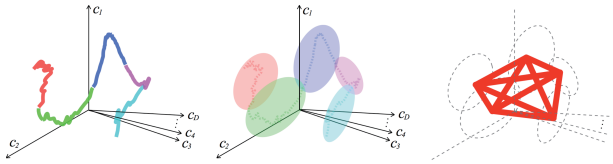

図 2　音と音のコントラストだけを用いた音声表象

を求めれば（距離行列）、音響構造が一意に定められる（図 2 参照）。分布間距離尺度としてはバタチャリヤ距離を用いる。

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

バタチャリヤ距離を初め、多くの分布間距離は変換不変性を有する。一対一、連続、かつ可逆な変換であれば、線形、非線形を問わず距離の不変性は成立する [9], [10]。筆者らが提唱する構造表象はこの不変性を利用している。なお、アフィン変換は一般に幾何学構造を歪める変換として使われるが、分布間距離は空間を歪めて距離を再定義していることに相当する。構造が歪まないように、空間を歪めて対象を観察している。

上記で構成された距離行列の上三角部分を抽出したものを構造ベクトルと呼ぶ。この特徴量は音の実体を捨象し、音と音の距離（配置関係）だけを表象したパラメータとなる。Δ パラメータ [11] や RASTA [12] など、音声の動的特徴を軌跡に対する速度ベクトルとして抽出する試みが広く行なわれているが、筆者等は、行列 A は n 次元空間において極めて高い回転性を有することを数学的に導出している [13]。動的特徴の方向成分には非言語的情報によるバイアスが強く含まれていると考えており、音群の配置関係のみを抽出する方法を提唱している。

2.3 音声の構造的表象を用いた音声認識の枠組み

図 3 に一発声から構造ベクトルを得る過程と、それを用いた孤立単語音声認識の枠組みを示す。ケプストラムストリームを N 個の分布へと変換し、距離行列化、構造ベクトル化が行なわれる。なお、1 発声から分布系列を推定するため、MAP 推定を導入している。各単語の音響モデルは、複数の発声より得られる複数個の構造ベクトルをガウス分布で近似する。この構造統計モデルは、入力音声に含まれる音と音の差異（コントラスト）を時間的に離れた 2 音間についても求め、それを統計的にモデル化していることになる。従来の音響モデルである HMM では多くの場合、スペクトル包絡相当の音の実体を統計的にモデル化しており、両者の違いは音の差異を捉えるのか、音の実体を捉えるのかに帰着される。認識は、入力された構造ベクトルに対して最高尤度を示すモデルを採択することで行なわれる。

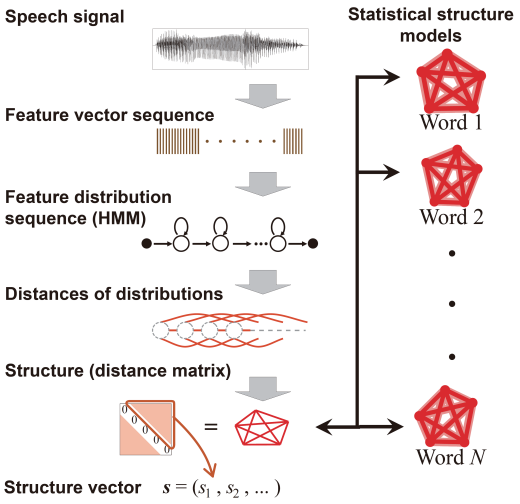

図 3　構造的音声認識の枠組み

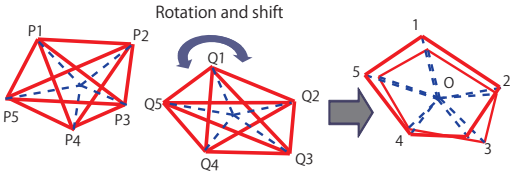

図 4　構造の回転とシフトを通した構造的音響照合

ここで、構造ベクトル間のユークリッド距離の物理的意味について示しておく。ユークリッド空間に、2 つの N 角形（$N\times N$ の距離行列対で表象される）が存在したとき、これら 2 つの構造ベクトルのユークリッド距離は両構造を回転、シフトさせて近づけた時の対応する 2 点間距離の総和の最小値の近似値を与える（図 4 参照）[14]。これはタンパク質の構造解析などでも使われる手法である。前節でも述べたように、非言語的音響歪みの最も簡素な数学モデルは線形変換 $c'=Ac+b$ である。アフィン変換の性質を考えると、A の乗算が構造の回転に、b がシフトに相当することになる。言うなれば、行列 A や b を明示的に求めることなく（音響モデルの適応を明示的に行なうことなく）、音響モデル適応後の照合スコアが近似的に求まることになる。これが、構造的音声認識の枠組みである。

2.4　強すぎる不変性と特徴量空間分割

構造の不変性は線形、非線形変換に依らず成立する強い不変性である。これは環境・話者による変動を取り除くだけでなく、異なる単語を「等しいもの」として扱う可能性がある。そのため、不変性を抑制する必要が生じる。ここでは行列 A の帯行列性に着目し、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 次元部分ベクトルを構成し、これを低次から 1 要素ごとにずらし、複数の特徴量ストリームを生成した。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て処理系を構築した。全次元から張られる一空間における構造不変性よりも、全部分空間における構造不変性の方がより強い制約となる。なお、この空間分割は、行列 A の帯行列性という制約条件を幾何学的に解釈して得られた方法論である [5]。以降、w をブロックサイズ（BS）と呼ぶ。図 5 に一例を示す。右図は BS=3 で部分空間を構成しており、左図は全次元数が 3 の時に、BS=2 で部分空間を構成している。なお、認識時の対数尤度計算は、各部分空間の対数尤度の加算で行なった。

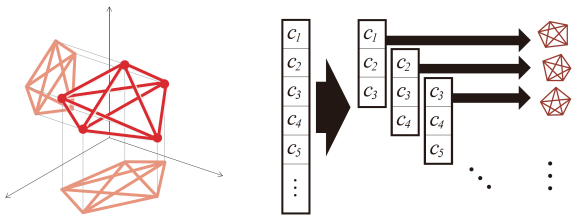

図 5　特徴量空間分割によるマルチストリーム化

3.　共有による推定パラメータ数の削減

3.1　K-means 法に基づくエッジクラスタリング

本研究では、話者不変表象としての性質を考慮し、学習話者を 1 名とした場合の構造的音声認識の特性を実験的に検討する。なお、N 個の音事象からなるストリームを、その実体を捉えてモデル化すれば、パラメータ（分布）数は N だが、構造化すると、${}_NC_2$ となる。即ち、構造表象はパラメータ数が $O(N^2)$ で増える。パラメータ数を削減するには、PCA や LDA などの方法が一般的であるが、逆に低次元化されたパラメータベクトルの物理的意味が不明瞭となる。本研究では、共有による低次元化を通して、構造的音声認識の特性を検討することとした。

本研究では日本語 5 母音を並びを変えて生成される 120 単語を認識タスクとしている。また一発声を 25 分布化して、構造を得ている。共有を導入しない場合、$120\times{}_{25}C_2$=36,000 だけエッジが存在し、その各々をガウス分布でモデル化することになる。構造表象は静的な歪みは抑制できるが、逆に、動的な歪み（発話スタイルなど）には直接的に影響を受ける [15]。この動的な歪みに基づく話者性による認識率低下があれば、これらはパラメータ共有によって緩和化されると予想される。

36,000 個のガウス分布の共有は、K-means 法で行なった。但し、セントロイド群の初期値として互いに適切に離れたセントロイド群を配置する目的で、下記のような初期セントロイド群の選択を行なった。まず第一のセントロイドを全ガウス分布からランダムに選び、ある閾値（半径に相当）よりも小さなバタチャリヤ距離を有する他ガウス分布をこのセントロイドに割付ける。残されたガウス分布から第二のセントロイドをランダムに選び、どのセントロイドにも割当てられていないガウス分布のうち閾値以下のものを、このセントロイドに割り当てる。以下この処理を繰り返し、閾値に（その数が）依存する形で、初期セントロイド群が構成される。これを用いて K-means 法を行ない、最終的なクラスタ群を得る。なお、ここまでの処理では複数のガウス分布を統合（平均化）する場合は、平均ベクトルの平均、及び、対角化分散・共分散行列の平均をとることで行なった。K-means 法による最終的なクラスタ群が決定すれば、36,000 だけ存在するエッジ群（ガウス分布群）の共有関係が決定されることになる。その共有関係を参照し、エッジ長の標本データにまで遡って、各クラスタの分布を再計算し、これを、「共有関係を有する統計的構造モデル」とした。

表 1　音響分析条件と認識実験条件

サンプリング	16bit / 16kHz / 25ms ハミング窓 / 10ms シフト
音響量	HMM：MCEP (1∼12)+E+Δ
	構造：MCEP (1∼12)+E による HMM よりエッジ計算
認識タスク	日本語 5 母音より構成される 120 単語の孤立単語認識
学習データ	成人男性 1 名、各単語 5 発声
評価データ	成人男性 7 名、女性 8 名、各単語 1 発声ずつ
単語 HMM	25 状態、対角化単一ガウス分布
単語構造モデル	25 状態、対角化単一ガウス分布 HMM からエッジ計算
	300 エッジ、各エッジは対角化単一ガウス分布で推定

3.2　特徴量空間分割処理を入れた分布共有

第 2.4 節で述べたように、強すぎる不変性を抑制するために、マルチストリーム化を行なっている。例えば本実験では、BS=n の時、ストリーム数は 14−n となる。これを前提として共有を考えた場合、前節の議論において、36,000 個のガウス分布は 1 次元ではなく、14−n 次元の多次元ガウス分布（対角化分散共分散行列）としてクラスタリングを行なえばよい。

3.3　HMM での状態共有

比較実験として、120 単語各々に対して単語単位の 25 状態 HMM を構築した。また、第 3.1 節と同様のクラスタリングを施し、状態共有を実装した。なお、音素単位 HMM の場合、音素コンテキストに基づくトップダウンクラスタリングが広く行なわれているが、ここでは比較実験のために上記の通りとした。

4.　実　　験

4.1　音響分析条件と認識実験条件

実験の音響分析・認識実験条件を表 1 に示す。HMM のパラメータはメルケプストラムとパワー、及びそのデルタを用いている（計 26 次元）。構造は（事前の予備実験の考察に基づき）、デルタ特徴量の無い（計 13 次元）HMM 学習を通して推定された分布からエッジ長を計算している。

認識タスクは日本語 5 母音を並び替えて構成される 120 単語（/uoaei/など）の孤立単語認識である。音響モデルは、HMM、構造モデルともに成人男性 1 名の音声から学習した。各単語モデルは 5 回の発声データから構築される。評価データは成人男性 7 名、成人女性 8 名の音声に加え、それらの音声を STRAIGHT [16] により、第 2.1 節に示した行列を用いてウォーピングを施した分析再合成音声も用いた。変換パラメータ α が 0.4 の時に身長が凡そ半分に、−0.4 の時に凡そ倍になる。

4.2　非共有モデルを用いた孤立単語認識結果

本実験でのベースラインとなる、非共有時の認識精度について HMM、構造モデルの結果を表 2、表 3 に示す。評価用音声はウォーピングを施していない元音声である。本実験では学習話者数を 1 とし、極めて話者性の強いモデルを構築している。そのため、評価話者に対する HMM の精度において、男女差が極めて大きい。一方構造モデルであるが BS が小さいほど、変換不変性は低くなる。[5] と同様、BS=2 において、最高精度を得た。HMM に比べ男女差は小さいが、それでも非常に大きな差がある。これは、1) 発話スタイルなどに起因する話者性はそもそも抑制できない。2) エッジ計算に用いられる HMM は対角分散共分散行列を用いており、分布の回転性に追従できていない、などの理由が考えられる。以降、パラメータ共有の枠組みを導入することで、HMM、構造の精度がどのように変化したのかについて結果を述べ、考察を行なう。

表 2　非共有 HMM を使用した音声認識結果

全評価話者	男性	女性	学習話者
36.4	68.3	8.4	100

表 3　非共有構造を使用した音声認識結果

BS	全評価話者	男性	女性	学習話者
1	33.4	55.6	14.1	100
2	**42.8**	**65.6**	**22.8**	100
3	38.8	59.3	20.4	100
4	34.4	58.7	13.2	100
5	33.7	57.6	12.7	100
6	33.1	55.8	13.2	100

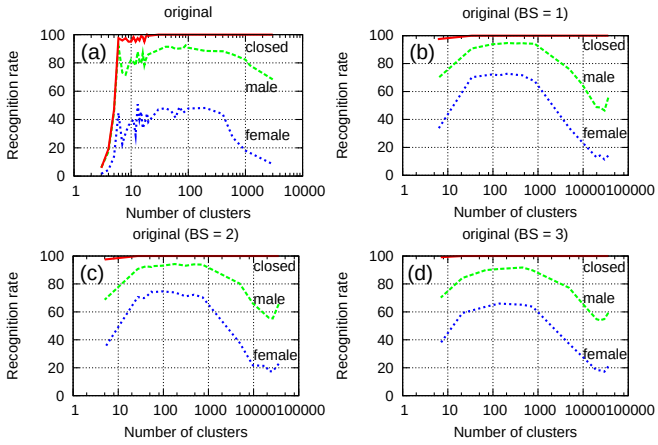

図 6　パラメータ共有による孤立単語認識精度の向上
(a)：共有 HMM　(b)–(d)：共有構造モデル（BS=1,2,3）

4.3　パラメータ共有による孤立単語認識率の変化

構造のパラメータ（分布）数は第 3.1 節で説明したように、36,000 個となる。但し各分布の次元数（即ち、一次元ガウス分布数）はストリーム数となり、BS=2 の時に 12 となる。一方で、HMM は 25 状態のものを用いているので、分布数は 25×120 = 3,000 個となる。但し、分布の次元数は 26 である。これを初期状態としてパラメータ共有により分布数を削減する。

まず、クラスタ数の削減による、全評価話者に対する認識率の変化を、HMM の場合を図 6(a) に、構造の場合を図 6(b)-(d) に示す。(b) は BS = 1、(c) は BS = 2 、(d) は BS = 3 の結果である。双方ともにクラスタ数削減（パラメータ共有）によって、認識率の向上が見られる。男性評価話者の場合、HMM と構造とでは認識率にあまり差がみられないが、女性評価話者に対する認識の向上率、及び精度は構造の方が良い。上記したように、構造モデルは、完全な話者不変性を実現している訳ではない。しかし分布共有により、パラメータ推定により多くの学習データ（但し同一話者）を割り当てられるようになり、話者不変量としての構造モデルの特性が顕在化したと考えている。全評価話者に対する最高精度は、BS=2、クラスタ数=約 180 の時に示され、42.8% から 83.4% まで向上した。

HMM の場合、クラスタ数が 6, 13 のとき局所的なピークが見られる。今回の認識実験では、各単語発声は前後の無音も含めて「発声」として捉えており、その結果、クラスタ群が無音区間（1 種類）と 5 種類の母音に各々対応した時に、特異的に精度が上がった結果となっている。これは、特定話者の連続発声の母音系列に対して、母音の渡りの存在を無視し、母音の数だけの分布種類数を仮定してモデル化すると、話者性に対する頑健性が特異的に向上することを意味する。言い換えれば、文字表記された音素列のイメージに従って、特定話者の音の実体を音響的にモデル化することが頑健性が向上する、とも解釈できる。しかしながら、例えば女声の認識率は特異的に向上した場合でも、構造モデルには大きく及んでいない。なお、全評価話者に対する最高精度はクラスタ数=約 100 の時に示された。この時、認識率は 36.4% から 67.8% まで向上した。

最高性能時の分布数は上記に示した通り、構造では約 180、HMM では約 100 であるが、構造は 1 分布が 12 の単一ガウス分布で構成され、HMM は 26 の単一ガウス分布で構成される。両者の性能差を考えると、今回の単語認識実験においては、単語識別を効率的に行なうために必要な音響量を、よりコンパクトにモデル化しているのは構造の方であると考察される。

4.4 ウォーピングを施した音声に対する認識結果

ウォーピングをかけた分析再合成音声（$\alpha = -0.3$~0.3）に対する認識率を、共有 HMM の場合を図 7 に、共有構造モデルの場合を図 8 に示す。構造モデルは BS = 2 の場合を示す。なお、α =0.0 は元音声ではなく、分析最合成音を意味する(注1)。

全体に共通して観測される傾向を考える。女声の場合、声道長を伸ばすことに相当する $\alpha < 0$ の変換をかけることで、音質がモデルに近づく。例えば HMM の場合、$\alpha = -0.1$ 辺りが認識率のピークになっていることが分かる。構造モデルにおいても、この傾向は見られ、$\alpha = -0.07$ 辺りがピークとなっている。

非共有の場合、双方とも同様の認識傾向を示すが、分布を共有し、パラメータを削減していくにつれ、HMM と構造モデルとの違いが明確に現れる。HMM は共有をかけたとしても（例えばクラスタ数 100）$|\alpha| > 0.2$ の時は学習話者に対しても、ほぼチャンスレベル (0.83 %) にまで性能は劣化する。一方で、構造モデル（例えばクラスタ数 180）での性能劣化は、HMM と比較すると極めて小さい。なお構造モデルにおける α の正負による性能の非対称性であるが、α を負、即ち声道長を長くするとフォルマント周波数が低域に集まり、各母音が音響的に類似してくることに起因する（人間が聞いても判別が困難になる）。

なお、クラスタ数 6 の HMM における女声の最高性能は、$\alpha = -0.1$ の時の約 70%であるが、クラスタ数 180 の構造モデルでは、ウォーピングをかけなくても、約 70%の性能を示す。構造モデルの音響照合は、話者・環境適応をかけた後の HMM による音響スコアを近似的に算出しているが、本実験の結果は、それを裏付ける結果となった。なお、クラスタ数 180 の構造モデルの結果より、ウォーピングによる率の低下が学習話者と評価話者で大きく異なる。今回採用したウォーピングが話者性の多様性を表現するにはまだ不十分であることを示唆しており、今後、更なる検討を加えたい。

4.5 言語情報は音的実体と音的差異のどこに在るのか？

今回の実験では、従来法である HMM（音の実体に対する統計モデリング）と、提案法である構造モデル（音の差異に対する統計モデリング）とを、学習話者数 1 という同条件で構築し、分布共有を通して両者の性能比較を行なった。両手法は音の全く異なる側面を統計的にモデル化しており、（母音連結単語という人工的なタスクではあるが）「単語を識別するためには単語音声の何を記憶しておくことが最も効率的か」「言語情報は音のどの側面に符号化されていると考えるべきなのか」という問いに対して、実験的に検討することができる。

構造モデルは音の差異のみをモデル化しているため、孤立音を提示されても一切識別は不可能である。一方 HMM は、$\alpha = 0.2$ のウォーピングを施すと、男声評価話者に対する性能はほぼ零になるが、構造モデルでは 70%ほどの識別力を残している。筆者らは本構造モデリングと幼児の言語獲得とを連携させて議論している [17]。幼児とは言え、発話できるようになると、自らの聞く声の約半数は自分の声となる。話者バランスのとれた音声の聴取は通常不可能である。また、自分を取り囲む成人の声と（よく聞く）自分の声とは大きな声道長差があるのも事実である。このような環境で、幼児は自らの（可愛い）「おはよう」という音声と、父親の太い「おはよう」という音声の同一性を感覚し、頑健な音声認識系を獲得する。発達心理学によれば、幼児の音韻（モーラ）意識は希薄なため、これらの発声を音韻単位に分割して、音韻の音響モデル（音韻意識）を持つことは困難である。幼児はまず発声の全体像（語ゲシュタルト）を捉える、と言われている。これらの事実・考察を考えるに「音声の言語情報は音的差異に符号化されている」と考える方が、事実をより良く説明できる、と筆者等は考えている。

5. ま と め

本稿では、音声の構造的表象と、HMM とでパラメータ共有をかけるこにより、両者の性能を比較した。その結果、話者性に対する、構造モデルの極めて強い頑健性を示すことができた。既に子音を入れた音韻バランス単語による認識実験も開始しており、これらの結果は別途報告する予定である。

(注1)：α=0.0 の合成音に対する認識精度は前節の結果と必ずしも一致しない。α=0.0 に対して最高性能を示した HMM はクラスタ数 6 であった。

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, “Mathmatical 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. SRIV*, pp.47–52, 2006.

[4] Y. Qiao, et al., “Random discriminant structure analysis for continous Japanese vowel recognition,” *Proc. ASRU*, pp.576–581, 2007.

[5] S.Asakawa *et. al.*, “Multi-stream parameterization for structural speech recognition,” *Proc. ICASSP*, pp.4097–4100, 2008.

[6] 鈴木雅之他、“スペクトル特徴量を用いた音声の構造的表象に関する実験的検討”，電子情報通信学会音声研究会, 2008-6.

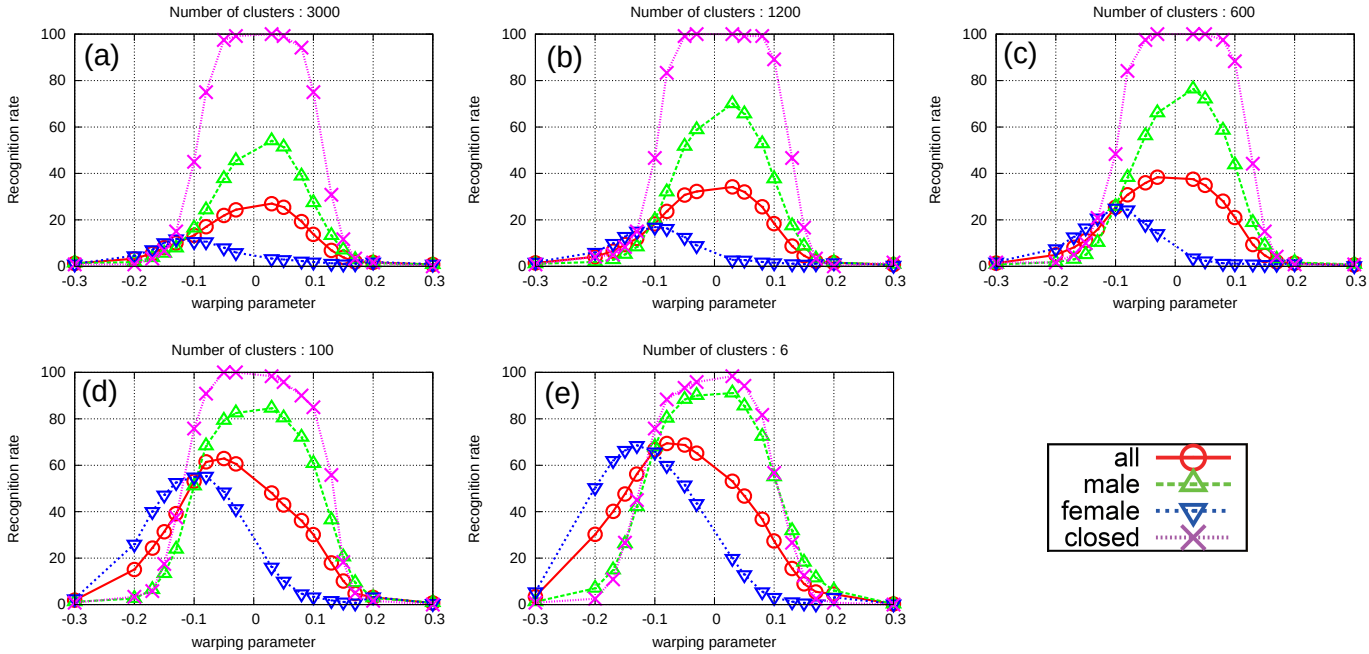

図 7　ウォーピングを施した評価音声に対する共有 HMM の認識結果

(a)：非共有（クラスタ数=3,000）　(b)：クラスタ数=1,200　(c)：クラスタ数=600　(d)：クラスタ数=100　(e)：クラスタ数=6

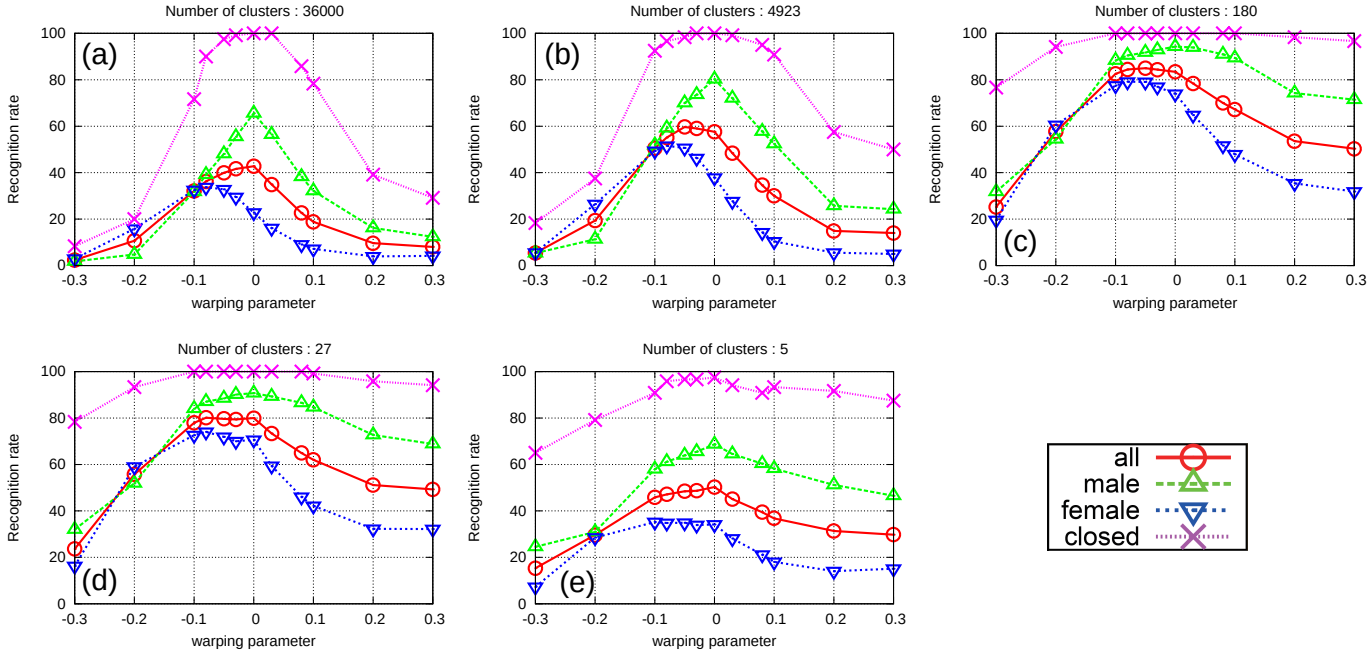

図 8　ウォーピングを施した評価音声に対する共有構造 (BS = 2) の認識結果

(a)：非共有（クラスタ数=36,000）　(b)：クラスタ数=4,923　(c)：クラスタ数=180　(d)：クラスタ数=27　(e)：クラスタ数=5

[7] M. Pitz et al, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech and Audio Processing, vol.13, no.5, pp.930–944, 2005.

[8] 江森正他、“音声認識のための高速最ゆう推定を用いた声道長正規化”、電子情報通信学会論文誌 vol.J83-D2, no.11, pp.2108-2117, 2000.

[9] 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.

[10] 喬宇他，“変換不変性を有するダイバージェンスとその一般形”，電子情報通信学会音声研究会, 2008-7.

[11] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” *IEEE Trans.*, ASSP, vol.34, no.1, pp.52–59, 1986.

[12] H. Hermansky, “RASTA processing of speech,” IEEE Trans. speech and audio processing, vol.2, no.4, pp.578–589, 1994.

[13] D. Saito et al. “Directional dependency of cepstrum on vocal tract length,” *Proc. ICASSP*, pp.4485–4488, 2008.

[14] 峯松信明他、“音声の構造的表象とその距離尺度”，電子情報通信学会音声研究会，SP2005-13，pp.9–12, 2005.

[15] N. Minematsu et al., “Para-linguistic information represented as distortion of the acoustic universal structure in speech,” *Proc. ICASSP*, pp.261–264, 2006.

[16] H.Kawahara, “STRAIGHT, Exploration of the other aspect of VOCODER,” *Acoustic Science and Technology*, vol.27, no.6, 2006.

[17] 齋藤大輔他、“音声の不変表象に基づく語ゲシュタルトの物理的解釈とそれに基づく幼児の音声模倣の実装”、人工知能学会全国大会講演論文集, 3F3-5, pp.1-4, 2008.