音声の構造的表象に基づく 音声認識における パラメータ 共有に関する 考察*

◎松浦良, 齋藤大輔, 朝川智, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声には、環境雑音の他に、音響機器特性、話者の声道長の形状特性といった歪みが混入する。特に、後者の2 つの歪みは不可避的であり 、音声認識率を下げる要因となっている。一般的には、多数の学習データを収集して不特定話者・ 環境の音響モデルを構築し 、さらにケプストラム平均除去や声道長正規化を施すことによって歪みに対処している。しかし 、個々の利用環境・ 利用者は某の方向性を持った歪みを有しているため、モデルとのミスマッチが大きい場合、所望の性能を得ること は難しい。

近年、これらの歪みに影響を受けない、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 [1, 2]。ここでは、個々の音を絶対的な特徴量では捉えず、音と 音との差異(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よる歪み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 [3, 4, 5, 6]。

そこで本研究では、学習話者を 1 人まで落としたとき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]。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。

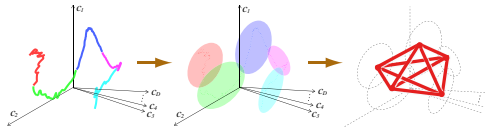

Fig. 1 音と 音のコントラスト だけを用いた音声表象

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を求めれば(距離行列) 、音響構造が一意に定められる (Fig. 2.1 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。

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

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]。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。

2.3 音声の構造的表象を用いた音声認識の枠組み

Fig. 2.2 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。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。

*Experimental study on parameter sharing in structure-based speech recognition. by MATSUURA.R, SAITO.D, ASAKAWA.S, MINEMATSU.N, HIROSE.K (The Univ. of Tokyo)

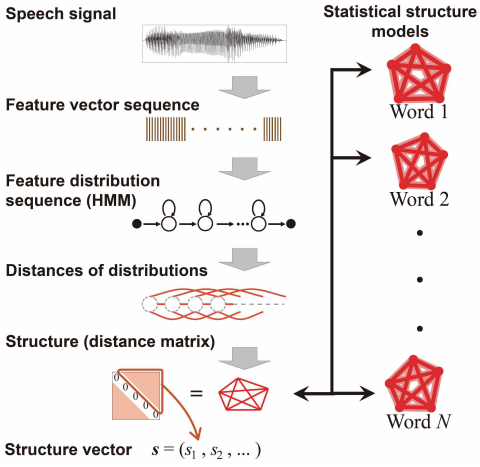

Fig. 2　構造的音声認識の枠組み

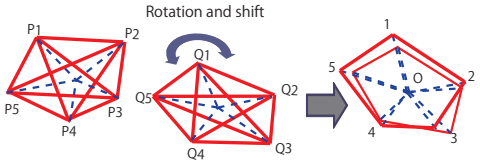

Fig. 3　構造の回転と シフト を通した構造的音響照合

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（ Fig. 2.2 参照）[10]。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

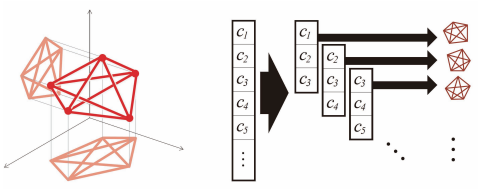

Fig. 4　特徴量空間分割によるマルチストリ ーム化

呼ぶ。Fig. 4 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。

3　共有による推定パラメータ数の削減

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 、${}_N C_2$ 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 [4, 5]、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 [11]。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。なお、特徴量空間分割のため、各エッジはストリ ーム数個のガウス分布を持つことになる。共有は data-driven のクラスタリングを用いて実装した。HMM も 同様のクラスタリングを行い、比較実験を行った。

4　実験

4.1　非共有モデルを用いた孤立単語認識結果

音響分析· 認識実験条件を Table. 1 に示す。HMM の音響特徴量はメルケプスト ラム、パワー、Δ の 26 次元であり 、メルケプスト ラム、パワーの 13 次元から分布を推定し、構造を得ている。学習データは成人男性 1 名である。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 [12] 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。

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 Table. 2、Table. 3 に示す。評価用音声はウォーピングを施して

Table 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 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	300 エッジ、各エッジは 対角化単一ガウス分布

Table 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

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、BS=2 と固定し、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。

4.2　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 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）はストリーム数であり、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 $25\times120 = 3,000$ 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 Fig. 5(a) に、構造 (BS = 2) の場合を Fig. 5(b) に示す。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

Table 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

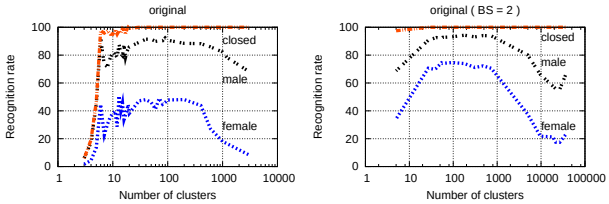

Fig. 5　パラメータ共有による認識精度の向上
(a)：共有 HMM　(b)：共有構造モデル（BS=2）

ピークが見られる。これは、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ためと考えられる。安定的に認識率の向上が得られた場合は、クラスタ数=約 100 の時に示され、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 つのエッジの分布が 12 個の単一ガウス分布で構成され、HMM は 26 次元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。

4.3　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、図 6 に示す。変換パラメータ α が 0.4 のときに身長が凡そ半分に、-0.4 のとき凡そ倍となる。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、分析再合成音を意味する[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（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。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。非ウォー

[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のた HMM はクラスタ数 6 であった。

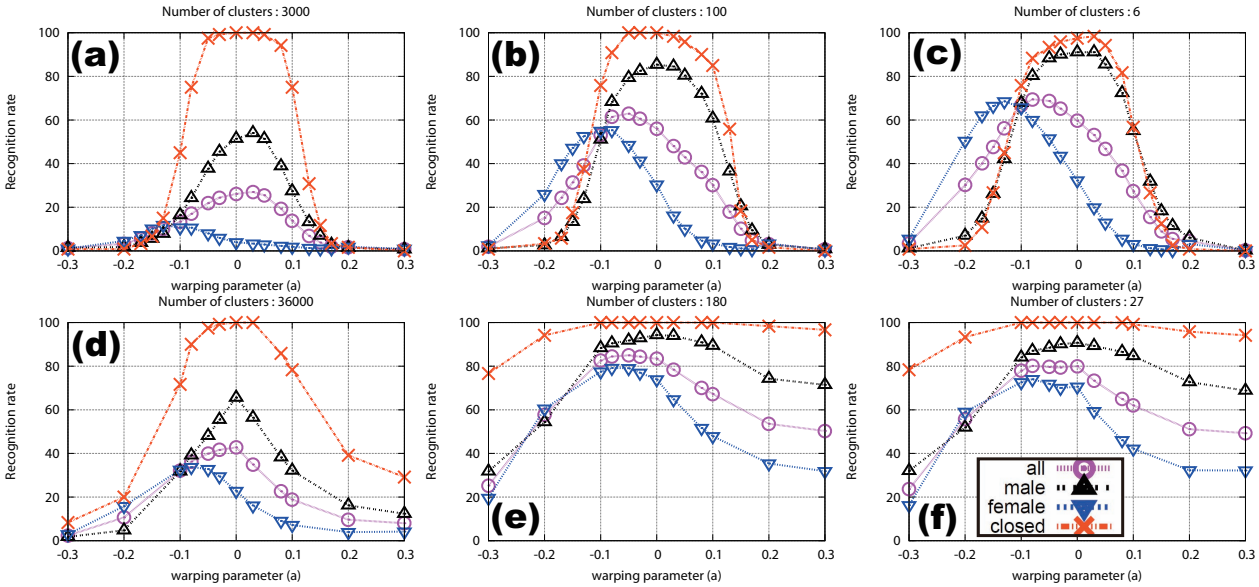

Fig. 6　ウォーピングを施した評価音声に対する共有モデルの認識結果
(a)：非共有 HMM(クラスタ数 3,000)　(b)：共有 HMM(クラスタ数 100)　(c)：共有 HMM(クラスタ数 6)
(d)：非共有構造 (クラスタ数 36,000)　(e)：共有構造 (クラスタ数 180)　(f)：共有構造 (クラスタ数 27)

ピング時は HMM では約 30% である。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。

5　まとめ

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、「語彙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、$\alpha = 0.2, 0.3$ 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。また、タスクオープンで厳密な比較はできないが、学習話者 4130 名による tied-mixture HMM [13] を音響モデルとして用いた場合、$\alpha = 0.2$ のとき 75% の率を得たが、$\alpha = 0.3$ では、ほぼ零になる。学習話者数に大きな差があるにも関わらず（構造は学習話者 1 名）、構造は約 70%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 [14]。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「音声の語彙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。

参考文献

[1] N.Minematsu, *Proc. ICASSP*, pp.889–892, 2005.
[2] N. Minematsu, *et al.*, *Proc. SRIV*, pp.47–52, 2006.
[3] 喬宇他, 音講論 (秋), 2-P-1, 2008 (発表予定).
[4] 喬宇他, 音講論 (秋), 2-P-2, 2008 (発表予定).
[5] 朝川智他, 音講論 (秋), 2-P-2, 2008 (発表予定).
[6] 鈴木雅之他, 音講論 (秋), 1-R-26, 2008 (発表予定).
[7] M. Pitz, *et al.*, IEEE Trans. Speech and Audio Processing, vol.13, no.5, pp.930–944, 2005.
[8] 江森正他, 電子情報通信学会論文誌 vol.J83-D2, no.11, pp.2108-2117, 2000.
[9] 喬宇他, 電子情報通信学会音声研究会 SP2008, 51, 2008.
[10] 峯松信明他, 電子情報通信学会音声研究会, SP2005-13, pp.9–12, 2005.
[11] N. Minematsu, *et al.*, *Proc. ICASSP*, pp.261–264, 2006.
[12] H.Kawahara, *Acoustic Science and Technology*, vol.27, no.6, 2006.
[13] T.Kawahara, *et al.*, *Proc. ICASSP*, pp.3069–3072, 2004.
[14] 齋藤大輔他, 人工知能学会全国大会講演論文集, 3F3-5, pp.1-4, 2008.