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.

音声の構造的表象と判別分析を用いた単語音声認識

朝川　智[†,*]　喬　宇[††]　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 〒113–8656 東京都文京区本郷 7–3–1
＊ 現在，ソニー株式会社
E-mail: †{asakawa,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　音声には話者の声道形状の特性，音響機器の特性などの非言語的特徴が不可避的に混入する．近年，これらの非言語的特徴を表現する次元を原理的に保有しない音響的普遍構造が提案されている．これは，音声事象の物理的実体を捨象し，音声事象間の相対量のみをとらえることによって得られる音声の構造的表象である．本論文では，構造的表象に基づく単語音声認識を検討する．その際に問題となる「強すぎる不変性」と「高すぎる次元数」の 2 つの問題に対処するため，それぞれ特徴量空間分割と線形判別分析による解決法を提案し，それらに基づく単語音声認識実験を行った．連続的に発声された日本語 5 母音系列および東北大松下単語音声データベースの音韻バランス単語をタスクとして認識実験を行い，音響特徴量の絶対量に基づく従来手法との比較を行った．さらに，意図的に話者性のミスマッチを生じさせた条件下で認識実験を行い，提案手法の頑健性を実験的に検証した．
キーワード　音声の構造的表象，非言語的特徴，Bhattacharyya 距離，線形判別分析，孤立単語音声認識

Isolated word recognition based on speech structures and discriminant analysis

Satoshi ASAKAWA[†,*], Yu QIAO[††], Nobuaki MINEMATSU[††], and Keikichi HIROSE[†††]

† Grad. School of Frontier Sciences, Univ. of Tokyo, 5–1–5, Kashiwanoha Kashiwa, Chiba, 277–8561 Japan
†† Grad. School of Eng., Univ. of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo 113–8656 Japan
††† Grad. School of Info. Sci. and Tech., Univ. of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo 113–8656 Japan
＊ Currently, Sony Corp.
E-mail: †{asakawa,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Non-linguistic factors of speech such as vocal tract sizes and recording devices easily change acoustic features of speech. Recently, a new representation of speech with complete cancelation of these changes has been proposed. This representation discards the absolute properties of speech events and captures only the contrasts among them. As a full set of the contrasts in the events can define a unique geometrical structure, the proposal can be regarded as structural representation. In this paper, the new representation is examined based on two kinds of isolated word recognition tasks, a five-vowel-sequence word set and a phonetically balanced word set. Here, two problems, too strong invariance and too high dimensionality, are solved by multiple stream structuralization and linear discriminant analysis. To compare the conventional method and the proposed one, frequency-warped utterances are also used for testing. The experimental results show the high robustness of our proposed method.
Key words　structural representation of speech, non-linguistic features, Bhattacharyya distance, linear discriminant analysis, isolated word recognition

1. はじめに

年齢，性別，個人性，音響機器など，音声には不可避的に非言語的情報が混入し，これが音声の音響的特徴を変形させ，音声記号としては同一音であったとしても，物理的には異なる特性を持った音響現象として観測される．従来の音声工学では，

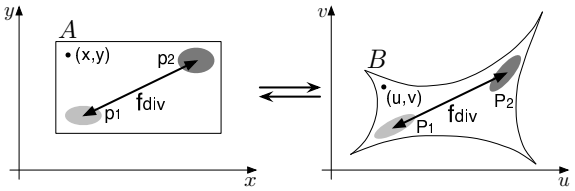


図 1　一対一対応関係を有する二つの空間 A と B
Fig. 1　One-to-one mapping between two spaces

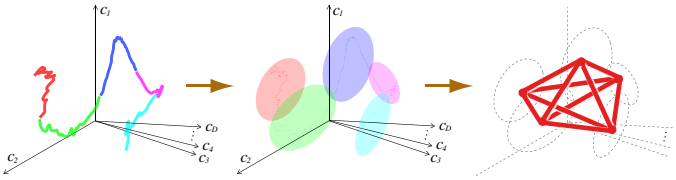


図 2　発声の構造的・全体的・話者不変的表象
Fig. 2　Structural representation of an utterance

数千・万の話者の音声から同一記号音を収集し，スペクトル包絡に代表される音の実体を統計的にモデル化することで多様性問題の解決を図って来た．非言語的要因は一般に時不変であり，静的なバイアス項として音響量を変形する．上記のような統計モデルでの解決は，非言語的要因による音声の変動を母集団からのサンプリングに伴うランダム雑音としてモデル化することとなり，例えば不特定話者音響モデルは，音声に含まれる話者性は時間軸上でランダムに変化することを前提としている．これは，明らかに，事実にそぐわない．

幼児の言語獲得は，両親の発声を真似ることで行なわれる．音声模倣と呼ばれるこの活動は，霊長類ではヒトのみに見られる．しかし，親の太い声を真似る幼児はいない．彼らは音は模倣しない．霊長類以外では，音声模倣は鳥やクジラに見られるが，この場合は音を模倣する [1], [2]．即ち，ヒトの幼児は個体サイズを超えた音声模倣をする．言い換えれば，静的なバイアスを超えて，父親の声と自らの声に同一性を感覚する．この場合注意すべきは，幼児は与えられた発声をモーラに分割することは困難であるため [3]〜[5]，両発声をモーラ列として比較し同一性を感覚することも困難であれば，親の声から抽出した各モーラを自らの声で再生することも困難となることである．発達心理学では，幼児が模倣する対象を語全体の音形／語ゲシュタルトなどの言葉で表現している [3]．

音声合成における話者変換や音声認識における話者適応において，話者の違いは音響空間の写像としてモデル化される．話者の違いが静的なバイアスであれば，この写像は静的となる．写像で対応づける二話者が変われば当然写像関数は変わるが，任意の写像を施しても不変なる音響量が定義できれば，それは話者不変量となり，それが語の全体的な様態を表象していれば，発達心理学の言う語ゲシュタルトの物理的定義として議論することができる [6], [7]．

本稿では筆者等が提唱する，微分可能かつ可逆な任意の写像に対して不変な音声の全体的・構造的表象 [8] が，孤立単語認識というタスクにおいて，話者の多様性に対して如何に頑健に機能するのか，を実験的に示す．この時，「強すぎる不変性問題」「高すぎる次元数問題」を解決する必要が生じるが，これらを特徴量空間分割及び線形判別分析の段階的適用により解決する．

2.　話者不変な音声の構造的表象とその照合

2.1　発声の構造化とそれに基づく話者不変表象

図 1 に示す写像を考える．微分可能かつ可逆な変換関数で対応付けられた二つの空間である．空間 A の点 (x, y) は空間 B の点 (u, v) へ写像される．空間 A における事象 p_1 と p_2 を考える．全ての事象は点ではなく確率密度関数として存在する．p_i に対応する空間 B の事象を P_i とする．P_i も確率密度関数となる．この時 f-divergence は両空間で常に等しい [13]．

$$f_{div}(p_1(x,y), p_2(x,y)) = f_{div}(P_1(u,v), P_2(u,v))$$

但し，f_{div} は下記で定義される．

$$f_{div}(p_1(x,y), p_2(x,y)) = \oint p_2(x,y) g\left(\frac{p_1(x,y)}{p_2(x,y)}\right) dxdy$$

Bhattacharyya 距離も f-divergence の一種であり，本稿では写像不変な分布間距離として，これを用いる．

写像不変量のみを用いて音声ストリームを表象すれば，それは話者不変量となる．図 2 にその手順を示す．ケプストラム時系列を分布系列へと変換し，任意の二分布間距離を Bhattacharyya 距離として計測し，距離行列として発声を表象する．当然この距離行列は話者不変となる．距離行列を求めることは，その幾何学的構造を一意に規定することであり（但し，鏡像の曖昧性は残る），この表象は構造的表象であるといえる．

話者の声道長の違いに対して簡単な数学モデルを仮定し，その枠組みの中で不変となるような特徴量を求める研究もいくつか行われている [9]〜[11]．これらの研究では，声道長の違いに起因するスペクトル変形を $f \to \alpha f$（f : 周波数）という形でモデル化し，この変換に対する不変量をそれぞれ異なるアルゴリズムにより導いている．これらの研究では，その不変性が数学的に保証されているものの，$f \to \alpha f$ という非常に単純なモデルを仮定しており，話者の違いを表現するには十分ではないと考えられる．また，これらの不変量は音そのものを不変にしようとするのに対して，提案法は音のコントラストを不変量として捉える枠組みである点が異なる．

2.2　構造表象における音響的照合

話者 A のある発声を構造化し，同様に話者 B のある発声を構造化する．この時，両発声を同一数の分布の系列としてモデル化し，構造化する．距離行列をベクトルとして考え（構造ベクトル），両構造ベクトルのユークリッド距離を計算すると，それは図 3 に示すように，両構造を近づけ（シフト），両者が重なるように回転させて得られる，「対応する二事象間距離の総和の最小値」を近似できることが示されている [8]．例えば音響空間としてケプストラム空間を考えれば，シフトは音響機器特性の違いを表現し，回転は声道長の違いを表現する [12]．即ち発声を構造化して得られるユークリッド距離は，両発声を音響機器や声道長の違いに関して適応／正規化して計算される音響スコアを近似することになる．音の実体をモデル化する従来の

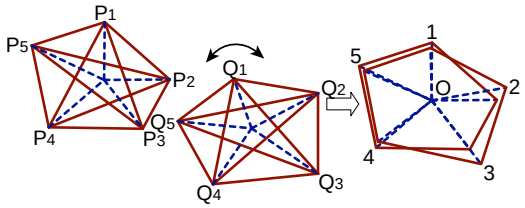


図 3　二発声間の構造的音響照合

Fig. 3　Structure-based acoustic matching between two utterances

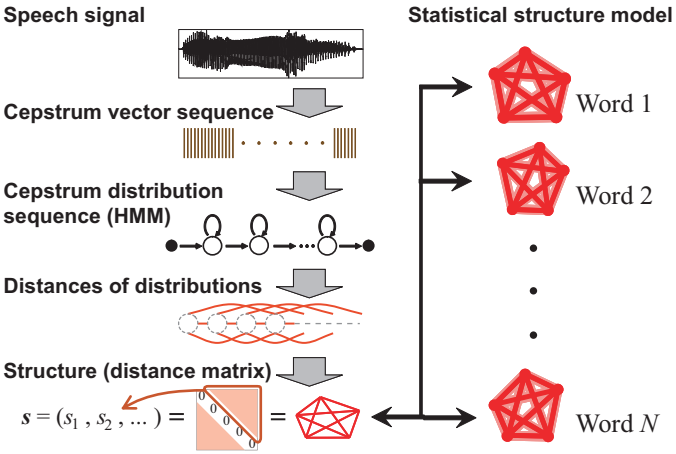


図 4　構造表象に基づく単語音声認識の枠組み

Fig. 4　Framework of structure-based word recognition

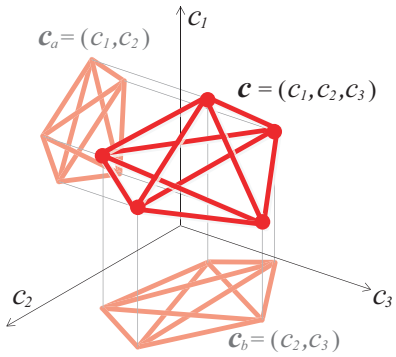


図 5　部分空間への射影

Fig. 5　Projection into sub-spaces

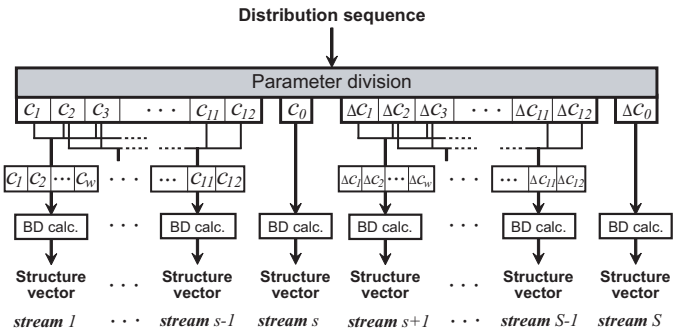


図 6　特徴量空間分割に基づく発話の構造化

Fig. 6　Structuralization based on parameter division

音響モデルでは，このシフトや回転を表現するパラメータを具体的に推定し，実際に音響モデル（あるいは入力音声）を変形した上で照合を行なう必要があるが，構造表象ではこれらの操作を明示的に行なう必要は一切ない．しかし，これらの操作を行なった後に計算される音響スコアが得られる．これが構造表象に基づく音響照合の枠組みであり，話者の違いを効果的に消失させた上での音響スコア計算が容易に可能となる [8].

以上の議論をまとめたものが図 4 である．一発声より得られる音響特徴量系列を HMM 学習により分布系列化し，推定された分布群を用いて構造ベクトルを得る．各単語のモデルは，学習データより構成される該当単語の複数の構造ベクトルより，多次元ガウス分布を求めることにより得られる．

3.　二つの問題とその解決策，及び改善策

3.1　強すぎる不変性問題とその解決

音声の構造表象では，音声の絶対的な物理特性は一切捨象され，音声のコントラストのみが抽出される．そして，任意の写像前後の二構造は，同一の発声として扱われる．これは極めて強い不変性を有することを意味しており，言語的に異なる二発声を同一と見なすことが容易に起きる．単語音声認識というタスクを考えると，異なる単語同士の音響的照合は不可避であり，任意写像に不変という構造表象の利点は，言語的要因による変形までを不変にする欠点となり得る．ここで，例えば声道長の違いによる変形のみに対して不変性を有する音響照合方式が必要となる．声道長の変化はケプストラムベクトル c に対して，行列 A を掛ける演算 Ac で近似できるが，この時 A は，例えば下記のような行列により表現される [14].

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ 0 & 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ 0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad (1)$$

α はウォーピングパラメータで，$|\alpha| < 1$ である．α が十分に小さい場合であれば α の高次項を無視することができ，対角およびその隣接項のみに非零の値をもつような行列となる．このような帯行列による写像に関してのみ不変性を有するような音響照合は，特徴量空間を適切に次元分割し，個々の部分空間で構造的照合を行なうことで実装できることを先行研究で報告している [15]．具体的には最低次から連続する w 個の次元で部分空間を張り（w をブロックサイズと呼ぶ），その空間で構造表象を構成し構造照合を行なう．同様の操作を次元を一つずらした w 次元の部分空間でも行ない，最高次までこれを繰り返す．最終的な音響照合スコアは各部分空間でのスコアの和として定義する．例えば，3 次元で構成される全空間に対して二つの 2 次元部分空間を考えた場合，それぞれの 2 次元平面に射影した構造に基づいて照合を行うことになる（図 5）．特徴量空間分割に基づく構造化の流れを図 6 に示す．ケプストラム及び Δ ケプストラムで構成される特徴量に対して空間分割する様子を示しており，Δ 項を使う場合は，ケプストラムとは別々に部分空間を構成している．

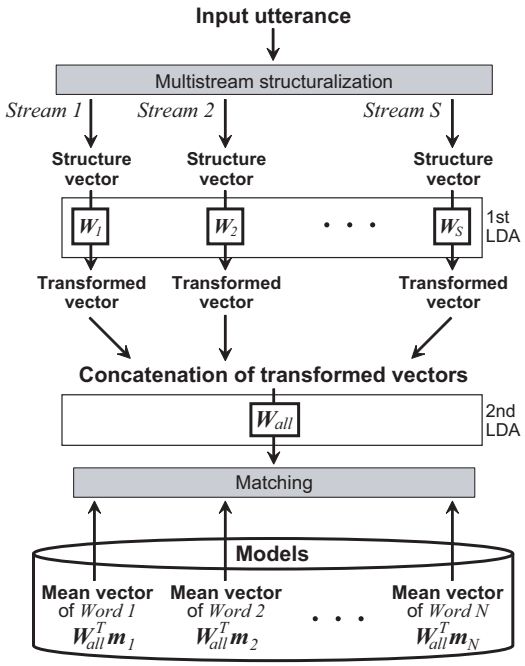


図 7　2 段階 LDA を通して行なう構造的音響照合
Fig. 7　Structural matching based on 2-phase LDA

3.2　高すぎる次元数問題とその解決

音声の実体をモデル化する場合，音声ストリームに N 個の事象があれば，モデル化対象数は N である．しかし，構造表象の場合 ${}_NC_2$ 個となり，モデル化対象数は $O(N^2)$ で増加する．さらに，前節での特徴量空間分割の導入によりパラメータ次元数は容易に増大する．その一方で，個々のパラメータの独立性が高いとは言えず，適切なパラメータ次元数の削減は，認識精度の向上に寄与することが期待される．既に先行研究にて PCA や LDA による次元圧縮を行なうことで，構造ベクトルをより低次元かつ識別的なパラメータへと変換する方法を提案しており [16]，本稿では図 6 において特徴量分割を施し，各ストリームで構造ベクトルを算出した時点でまず LDA を導入して次元削減し，次元削減された特徴ベクトルを結合してできる拡大構造ベクトルに対して再度 LDA を施してより識別的なパラメータとする方法を採択した（2 段階 LDA）．図 7 に 2 段階 LDA の様子を示す．

3.3　ストリーム間構造距離の導入

図 3 では個々のストリームは全く独立に扱われている．ここでは，各ストリームによって張られる構造間の距離の情報も，語彙同定に有効に寄与すると考えた．そして，S 本のストリームに対して ${}_SC_2$ 個だけ得られるストリーム間構造距離をベクトル化し，拡大構造ベクトルに連結する形で導入し，これを 2 段目の LDA の入力として学習される識別器についても，実験的に検討することとした．

なお，下記の実験では，単語モデルを構造ベクトルの平均ベクトルにより定義している．即ち，各単語の学習データに対して 1 段階目の LDA 後に得られる拡大構造ベクトルの平均ベクトルとして単語モデルを定義している．認識時には，入力される拡大構造ベクトルと，各単語モデル（ベクトル）とを 2 段階目の LDA（図 3 では W_{all}）を通して比較し，認識結果を得る．

表 1　音響分析条件（自然音声入力/構造モデル）
Table 1　Acoustic analysis conditions (natural speech/structure model)

サンプリング	16 bit / 16kHz（5 母音単語）
	12 bit / 16kHz（212 単語）
窓長	25 ms length / 10 ms shift
音響特徴量	MCEP (0〜12) + Δ (0〜12)
分布	対角分散行列による単一ガウス分布
分布数	20（5 母音単語）/ 25（212 単語）
分布推定	MAP 推定

表 2　音響分析条件（自然音声入力/単語 HMM）
Table 2　Acoustic analysis conditions (natural speech/word HMM)

サンプリング	16 bit / 16kHz（5 母音単語）
	12 bit / 16kHz（212 単語）
窓長	25 ms length / 10 ms shift
音響特徴量	MFCC (1〜12) + Δ (1〜12) + ΔP
分布	対角分散行列による単一ガウス分布
分布数	25（5 母音単語）/ 25（212 単語）
分布推定	ML 推定

4.　孤立単語認識実験

4.1　二つの孤立単語音声認識タスク

構造的表象に基づく孤立単語認識実験を行うにあたり，日本語 5 母音を並び替えて構成される母音単語セット（語彙数 120）と，東北大・松下単語音声データベースの音韻バランス単語（語彙数 212）の 2 種類のタスクを用意した．母音は話者の違いによってその音響的特徴が大きく変化するが，無声の摩擦音や破裂音は話者の影響が比較的小さい．これらは音の実体を捉えても話者の影響は少ないことを意味する．音のコントラストを捉えることで不変量を定義する提案手法は，前者のタスクには最適な方法であるが，後者のタスクには必ずしも最適な枠組みとは言えない．本稿では両タスクにおける構造的音声認識の性能を実験的に算出することで，その妥当性について検討する．

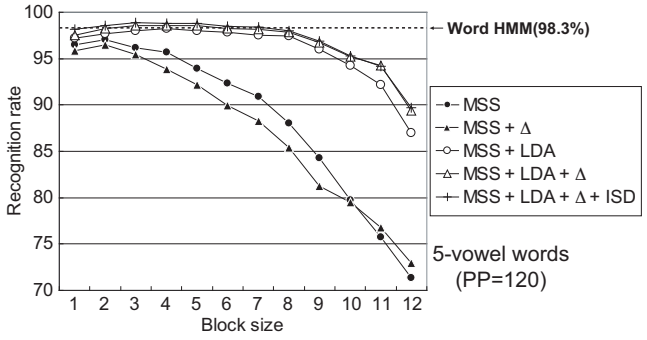
4.2　多様な話者性を伴う音声入力

本稿ではデータベースに収録されている音声をそのまま用いるだけでなく，これらに対して周波数ウォーピングを施し，声道長を伸縮することに等しい変換をかけた音声も入力音声として使用した．具体的な変換方法は [14] で示されている行列演算を用いてケプストラムを線形変換し，幅広い多様な話者性を人工的に生成した．

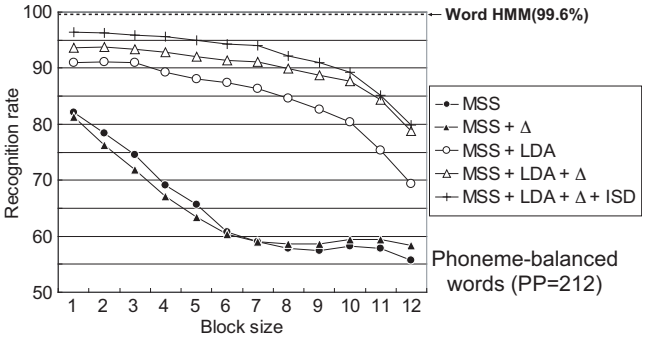
以降，変換を施す前の音声を自然音声，変換を施した後の音声を変換音声と呼ぶ．

4.3　音響分析条件と実験条件

自然音声を用いた場合の構造抽出及びモデリングに関する音響分析条件を表 1 に示す．また，比較対象として構築した単語 HMM 構築時の音響分析条件を表 2 に示す．なお，変換音声を対象とした実験では，特徴量として FFT ケプストラム（0〜16 次元）を用いた．これはメル化は周波数ウォーピングに相当す



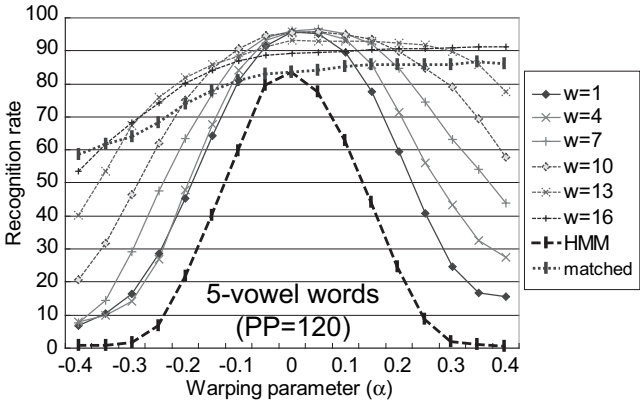
(a) 5 母音単語



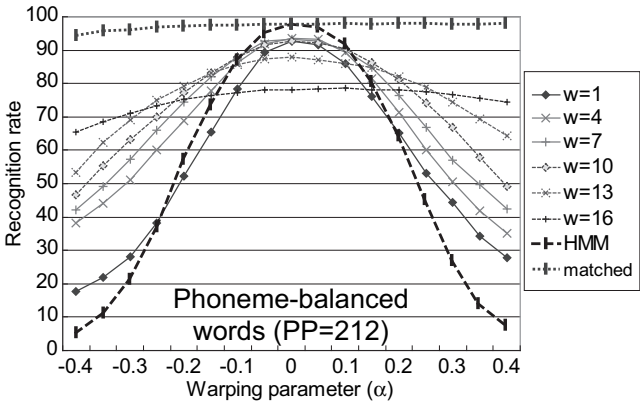
(b) 音韻バランス単語

図 8　自然音声を入力した場合の認識率

Fig. 8　Recognition results of natural speech



(a) 5 母音単語



(b) 音韻バランス単語

図 9　変換音声を入力した場合の認識率

Fig. 9　recognition results of warped speech

る操作であり，ある特定のウォーピングパラメータによる声道長変換とほぼ等しい操作となるためである．比較実験で用いる単語 HMM も同様，FFT ケプストラム（0〜16 次元）を用いて学習した．

なお，学習・評価話者数であるが，5 母音単語に関しては学習・評価ともに男女 4 名ずつであり，各話者が各単語を 5 回ずつ発声している．そのため，学習データ数は 40 発声／語であり，評価データ数も 40 発声／語となる．音韻バランス 212 単語は，学習・評価ともに男女 15 名ずつであり，各話者が各単語を 1 回ずつ発声している．そのため，学習データ数は 30 発声／語であり，評価データ数も 30 発声／語である．

4.4 実験結果

図 8 に，自然音声を入力した場合の認識結果を示す．上図が5母音単語，下図が音韻バランス単語である．共に，横軸はブロックサイズ w，縦軸が単語認識率である．MSS は特徴量ストリームの次元分割を意味し（Multiple Stream Structuralization），LDA は判別分析の導入を，Δ は Δ ケプストラムの導入を，ISD はストリーム間距離（Inter-Stream Distance）の導入を表す．Δ を伴わない場合ではストリーム数は S=$14-w$ であるが，伴う場合は S=$2(14-w)$ となる．各ストリームに対して構成される構造ベクトルの次元数は，5 母音単語の場合 ${}_{20}C_2$=190，音韻バランス単語の場合 ${}_{25}C_2$=300 である．2 段階 LDA は，前段で次元数を D まで削減し，後段では拡大構造ベクトル次元数がクラス数以上である時のみ，次元数は「クラス数 −1」へと削減される．前段での次元削減数については種々の条件にて検討を行ったが，図 8 は $D=20$ とした場合での結果を載せている．なお，単語 HMM の結果は 5 母音単語の場合が 98.3%，音韻バランス単語の場合が 99.6%であった．構造モデルの最高性能は 5 母音単語の場合，MSS+LDA+Δ+ISD の条件下で，w=3 の時に 98.9%となり，音韻バランス単語の場合は上記条件下で，w=1 の時に 96.4%となった．

評価データとして変換音声を入力した場合の結果を図 9 に示す．MSS+LDA+Δ+ISD を用いた結果である．横軸はウォーピングパラメータ α であり，値の負/正によって声道長を伸/縮させることに相当する．α=−0.4 で声道長が約 2 倍に，+0.4 で約半分になる．−0.4 から 0.05 ステップで増加させ，+0.4 までの 17 通りについて実験した．5 母音単語，音韻バランス単語両者とも，ブロックサイズ w=1,4,7,10,13,16 の場合を示している．なお HMM とは，構造モデルと同一学習データで構築した単語 HMM の性能であり，matched とは，各 α で学習データも変換させ，17 通りの HMM を構築した単語 HMM の場合での性能である．これは，学習／評価間にミスマッチが無い場合の性能であり，話者適応による理論的な最高性能である．

4.5 考　　察

図 8 より LDA 導入の効果が分かる．特に，音韻バランス単語の場合に極めて大きい．これは逆に言えば，無声子音など，構造表象と相性の悪い音声が音韻バランス単語には存在していることが一因であると考えられる．現在の実装では，無声子音までを含めて単一のアフィン変換を仮定することに無理が生じ

ていると考えられるが，アフィン変換ではないほかの単一の変換関数により話者の違いを表現し，適切な分布と距離尺度を設定することにより，無声子音まで含めた形での変換不変の表象を実現できる可能性もある．また，分布系列推定における MAP 推定時の事前分布として，それぞれのタスクに対して，一つの事前分布を与えている．音韻バランス単語の場合では無声子音も母音も同一の事前分布を用いることとなり，若干不適切な分布推定が行なわれた可能性もある．共鳴音・非共鳴音など，最低限の音実体カテゴリーを利用することにより，より適切な分布推定が可能となると考えられる．Δ 項は，LDA 導入前は性能を下げる方向に働いたが，導入後は上げる方向に働いている．構造的表象は相対量のみによる音声の表現であり，近隣フレームとの局所的な相対量である Δ ケプストラムはその多くが冗長な情報となるが，次元削減を導入することによって冗長性を削減し，単語識別において有用な情報のみを効果的に利用することが可能となると考えられる．

変換音声に対する認識率の結果（図 9）より，構造表象の話者性に対する極めて強い頑健性が分かる．話者の違いも音韻の違いも音色（スペクトル包絡）の違いであるため，前者に対する不変性を高めれば，後者に対する不変性も自ずと高まる．w がその制御パラメータとなっており，w が大きければ話者不変性が高くなり，単語同定率は落ちる．このトレードオフは，図からも窺える．本実験では FFT-ケプストラムを使用したことにより，特に 5 母音単語において HMM の性能が劣化してしまった．これとの比較となってしまうが，構造表象の場合は，例えば 5 母音単語認識における w=16 の場合では，ほぼ全ての条件で matched を超える性能が得られた．話者適応せずとも，話者適応したかのように機能する構造表象の特性が明確に現れた結果となった．一方，音韻バランス単語であるが，常に matched と同等の性能が得られた訳では無い．しかし，同一条件下で学習した HMM と比較すると（例えば w=10），無ミスマッチ時（α=0）の性能劣化が小さく，その分ミスマッチが存在する時の性能向上が非常に大きい．これも構造表象の特性が明確に現れた結果と言える．

構造的表象に基づく音声認識は，ケプストラム以外の特徴量を用いても可能である．例えば，ケプストラムを FFT することによって得られるスペクトル包絡でも実現可能であり，特徴量分割を用いたスペクトルに基づく構造的表象は，変調スペクトルや RASTA 等の手法と似た手法であるといえる．これらは音そのものではなく，音の時間的な動きに着眼する点が構造的表象と類似しているが，我々の手法は，その動きを話者不変の形で捉えようとする点においてこれらの手法とは異なるものであるといえる．

5. ま と め

本論文では，構造的表象に基づく単語音声認識を検討し，その頑健性を実験的に検証した．構造的音声認識において問題となる「強すぎる不変性」と「高すぎる次元数」の 2 つの問題に対処するため，それぞれ特徴量空間分割と線形判別分析による解決法を提案し，それらに基づく単語音声認識実験を行った．母音のみにより構成される単語および子音も含まれる音韻バランス単語をタスクとして認識実験を行い，音響特徴量の絶対量に基づく従来手法との比較を行った．さらに，ケプストラム線形変換に基づく声道長の伸縮により擬似的に生成した多様な話者性を含む音声を用いて，意図的に話者性のミスマッチを生じさせた条件下で認識実験を行い，提案手法の話者性の違いに対する頑健性を実験的に検証した．

今後の課題として，提案手法と従来手法との融合を検討する必要があると考えられる．本研究では，構造的表象のみを用いた音声認識を検討したが，音の実体に基づく手法と排反するものではなく，両者で異なる特性を持っているといえる．両者を適切に組み合わせることで，より高い頑健性を持つ音声認識が可能となると考えられる．更に，本研究では対象としなかった加算性雑音や残響に対する構造的表象の頑健性について，今後検証する必要があると考えられる．

文　献

[1] 岡ノ谷一夫，“小鳥の歌と言語：共通する進化メカニズム”，日本音響学会秋季講演論文集，1-7-15，pp.1555–1556，2008.

[2] 宮本健作，“声を作る・声を見る”，森北出版，1995.

[3] 早川勝廣，“言語獲得と育児語”，月刊言語，vol.35，no.9，pp.62–67，2006.

[4] 原恵子，“子どもの音韻障害と音韻意識”，コミュニケーション障害学，vol.20，pp.98–102，2003.

[5] サリー・シェイウィッツ，“読み書き障害のすべて”，PHP 研究所，2006.

[6] 峯松信明，西村多寿子，“多次元音楽としての音声モデリングと音声模倣”，人工知能学会全国大会講演論文集，1F2-3，pp.1–4，2007.

[7] 齋藤大輔，朝川智，峯松信明，広瀬啓吉，“構造的表象からの音声合成とそれに基づく音声模倣に関する検討”，電子情報通信学会音声研究会，SP2008-40，pp.115–120，2008.

[8] N. Minematsu, “Mathematical evidence of the acoustic universal structure in speech,” *Proc. ICASSP'2005*, pp.889–892, 2005.

[9] S. Umesh,L. Cohen,N. Marinovic, and D.J. Nelson, “Scale transform in speech analysis,” *IEEE Trans. Speech and Audio Processing*, vol.7, no.1, pp.40–45, 1999.

[10] T. Irino and R.D. Patterson, “Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform,” *Speech Communication*, vol.36，no.3，pp.181–203，2002.

[11] A. Mertins and J. Rademacher, “Vocal tract length invariant features for automatic speech recognition,” *Proc. ASRU'2005*, pp.308–312, 2005.

[12] 齋藤大輔，松浦良，朝川智，峯松信明，広瀬啓吉，“ケプストラムの声道長依存性に関する幾何学的考察”，電子情報通信学会音声研究会，SP2007-128，pp.189–194，2007.

[13] 喬宇，峯松信明，“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，SP2008-51，pp.49–54，2008.

[14] 江森正，篠田浩一，“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-D-II，no.11，pp.2108–2117，2000.

[15] S. Asakawa, N. Minematsu, and K. Hirose, “Multi-stream parameterization for structural speech recognition,” *Proc. ICASSP'2008*, pp.4097–4100, 2008.

[16] Y. Qiao, S. Asakawa, N. Minematsu, and K. Hirose, “Dimension reduction and discriminant analysis for Japanese connected vowel recognition,” *Proc. Autumn Meeting of Acoust. Soc. Jpn.*, 2-P-2, pp.109–112, 2008.