スペクトル領域特徴量を用いた音声の構造的表象による音声認識*

◎鈴木雅之, 朝川智, 喬宇, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声の物理的実体は，話者の声道形状によって不可避的に変形する．そのため，不特定話者音声認識において特定話者音声認識並の認識率を実現させるには，話者の違いになるべく影響を受けず，語彙情報をなるべく多く表すように音声をモデル化する必要がある．一方，ヒトの聴覚は，話者の違いにロバストな音声認識を実現している．よって，ヒトの聴覚に関する知見と，音声認識に用いられるモデルの関係について考察することは，不特定話者音声認識の発展のために有効であると考えられる．

近年，自動音声認識において最もよく利用されるモデルは，各時間フレームごとに計算した音響特徴量を，トライフォンごとに用意された隠れマルコフモデル（hidden Markov model: HMM）の出力とする音声の生成モデルである [1]．このように音声をモデル化すると，話者の違いによる音声の変形は，同じ確率密度関数をもつ分布から出力される値のゆれとして表現されることになる．しかし，このようなモデル化は，話者の違いを表すものとしては不十分であると考えられている．そのため，HMM のパラメータを入力音声に合わせて変換する話者適応技術や，HMM に合わせて認識したい音声の時間フレーム音響特徴量を変換する話者正規化技術が数多く研究されている [2].

これに対して近年，話者の違いによる音声の変形にほとんど影響を受けない音声モデルとして，音声の構造的表象が提案されている [3]．これは，ある程度の時間幅に含まれる複数の音の関係に注目し，その関係を構造として捉えるものである．この音声の構造的表象を用いて音声認識を行う場合，音声をある程度の時間幅ごとにモデル化している都合上，単語音声認識に相当するものしか実現することができない．しかし，HMM を用いた音声認識と異なり，話者適応や話者正規化に相当する処理をまったく行なわないまま，HMM のそれに相当する認識性能を得ることが可能である [4].

本稿では，この音声の構造的表象の実装の中で，特にスペクトル領域特徴量を用いた実装に注目する [5]．そして，人工的に作成した多様な音声に対して認識実験を行い，その頑健性を定量的に評価する．そして特に，ヒトの聴覚に関する研究においてよく知られている雑音駆動音声の認識実験から，ヒトの聴覚特性の関係について考察する．

2 音声の構造的表象

2.1 話者性に不変な音声の構造的表象

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関する研究においては，話者の違いを音響特徴量空間における一対一対応の空間写像と仮定することが多い．そこで，音響特徴量空間の空間写像に対する不変量を用いて音声をモデル化すれば，それは話者の違いに非常に頑健なものとなる．

ここで，空間における 2 つの分布間の距離として f-divergence を用いると，それは微分可能で可逆な空間写像に不変であることが証明されている [6]．すなわち，音響特徴量空間において各音声イベントを分布化し，分布間の f-divergence を求めると，話者性に対する不変量が抽出できることになる．今回は，f-divergence の一種である Bhattacharyya distance（BD）の平方根を用いることにする．2 次元の特徴量空間における BD の定義式は以下にようになる．

$$BD(p_1, p_2) \equiv -\log \iint \sqrt{p_1(x,y)p_2(x,y)}dxdy \quad (1)$$

ただし，p_1，p_2 は分布の確率密度関数である．

話者不変量である BD を用い，一発声から音声の構造的表象を抽出する枠組みを Fig.1 に示す．まず，音声に時間窓をかけて音響特徴量ベクトル系列を計算する．今回は，音響特徴量としてスペクトル領域特徴量を用いる．次に，HMM の学習アルゴリズムを用い，HMM の各状態における出力確率分布として分布系列を推定する．分布系列の推定後，任意の 2 分布間の BD の平方根を計算することで，分布間距離行列を得る．得られた距離行列は，対角成分を軸に対称であるので，上三角成分のみが情報をすべて表現している．そこで，この上三角成分をベクトルとして並べ替えたものを特徴量として用いる．これを構造ベクトルと呼ぶ．距離行列は一般的に，幾何学的に一つの構造を表現するため，構造ベクトルは一つの構造のパラメータ表現となっている．この音声の構造的表象の各エッジは，BD の平方根であるので，これは話者不変な表象である．

---

*Automatic speech recognition using spectrum-based speech structures. by M.Suzuki, S.Asakawa, Y.Qiao, N.Minematsu, and K.Hirose (The University of Tokyo)

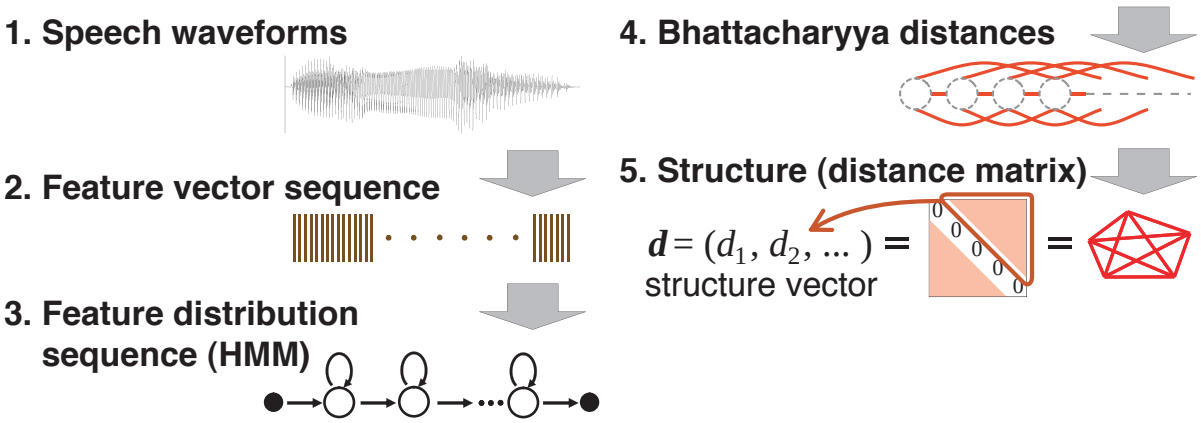

Fig. 1　Extraction of structural representation of an utterance

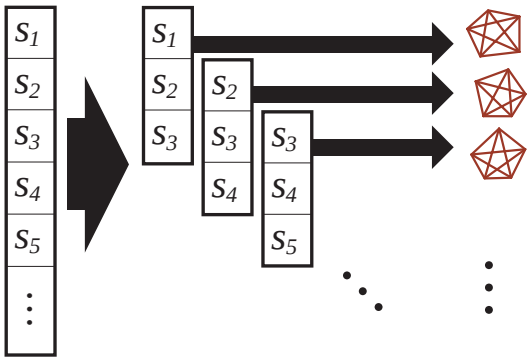

Fig. 2　Multi-stream structuralization

Table 1　Acoustic analysis condition and experimental condition

サンプリング	16bit / 16kHz
窓	25ms 幅 / 10ms シフト ハミング窓
特徴量 (HMM)	FBANK (24 次元) + Δ [10]
特徴量 (構造)	FBANK (24 次元) の分布間 BD
HMM	25 状態 / 対角共分散型ガウス分布
認識タスク	日本語 5 母音系列 120 単語の認識
学習データ	成人男女 4 名ずつ×各単語 5 発声
評価データ	成人男女 4 名ずつ×各単語 5 発声

2.2　音声の構造的表象に基づく音声認識

2 つの構造間の差異は，構造ベクトル間のユークリッド距離として表す．そこで，入力音声から構造ベクトルを抽出し，そのユークリッド距離に基づいて識別器を設計すれば，音声認識が実現できる．ここで識別器は任意のものを用いてよい（例えば [7]）．

構造ベクトル間のユークリッド距離は，それぞれの構造を回転・シフトさせたときの，各頂点間の距離の和の最小値という物理的意味を持つ [8]．ここで，構造のシフト・回転は，マイクの違い・声道長の違いに対する適応及び正規化処理に相当する．つまり，構造ベクトル間のユークリッド距離に基づく音声認識は，話者正規化／適応化を明示的に行なわいまま，これらを行なった後の音響マッチングスコアを算出することが可能な枠組みとなっている．

2.3　マルチストリーム構造化

音声の構造的表象は任意の空間写像に対して不変であるが，任意の空間写像を許してしまうと，異なる単語を同一にするような変換も許してしまう場合がある．この「強すぎる不変性問題」に対処するために，マルチストリーム構造化が提案されている [9]．

声道長の違いによるスペクトル特徴量の影響は，フォルマント周波数のシフトとして表現することができる．声道長の違いに起因するフォルマントシフトがあまり大きくないことを仮定すれば，スペクトル特徴量空間において声道長変換を表す空間写像は，隣り合う複数の次元によって張られる部分空間内での写像のみで表現できることになる．そこで，Fig.2 に示すように，音響特徴量を複数のストリームに分割し，それぞれのストリームにおいて構造を抽出する．これがマルチストリーム構造化である．これにより，話者の違いを表す変換としてあり得ない，遠い帯域間を表現する次元間での影響がなくなるよう，構造の不変性に制約条件をかけることができる．

3　多様な音声の認識実験

3.1　声道長を変化させた音声の認識

スペクトル領域特徴量を用いた音声の構造的表象を用いた音声認識の話者の違いに対する頑健性を定量的に評価するために，擬似的に声道長を変化させた音声を，1) 構造に基づく手法，2) 単語 HMM を用いる手法で認識する実験を行なった．

音響分析条件と実験条件を Table 1 に示す．認識

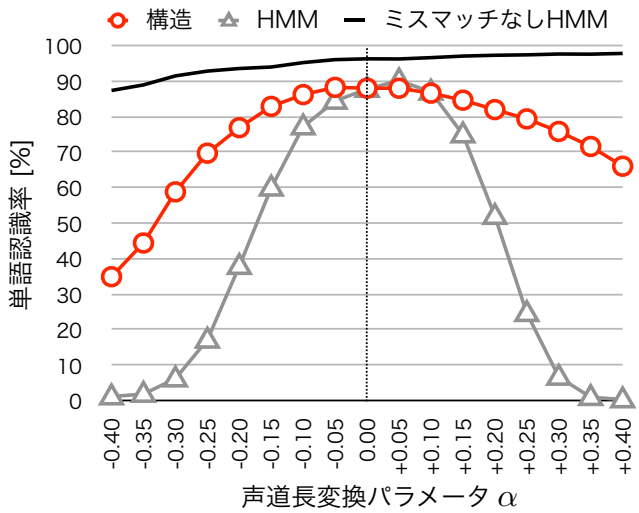

Fig. 3　Recognition rate vs. warping parameter

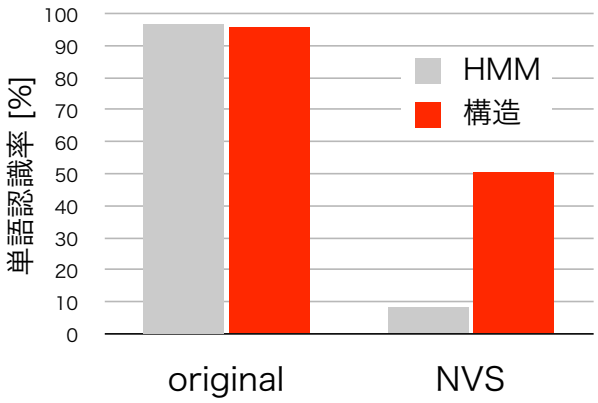

Fig. 4　Comparison of the recognition rates of clean utterances and noise vocoded speech

タスクには，日本語 5 母音を重複を許さず 5 つ連結した/aiueo/など計 120 語の単語認識を選んだ．学習データには，一単語につきのべ 40 発声分のデータを用いた．評価データには，学習データと異なる話者の発声に対し，STRAIGHT[11] を用いて声道長変換に相当する周波数ウォーピングをかけた分析再合成音声を用いた．また，各時間フレームごとの音響特徴量には，構造を用いた場合も HMM を用いた場合も対数スペクトルのメルフィルタバンク出力である FBANK[10] を用い，構造を用いた場合の分布推定時と HMM を用いる場合にはデルタパラメータを付加した．構造に基づく手法における識別器には，LDA に基づく識別器を用いた [4]．マルチストリーム構造化のストリーム数は，実験的に最も良好な結果を示した 7 を選択した．

実験結果を Fig.3 に示す．ここで，α は声道長変換の大きさを表すパラメータであり，α が正のとき声道長を短くする変換，負のとき声道長を長くする変換に対応し，+0.40 の時に身長がおよそ半分に，−0.40 の時におよそ倍になる [12]．また，参考のために，評価データにかけた周波数ウォーピングと同じ変換を学習データにもかけて，学習データと評価データのミスマッチをなくして単語 HMM に基づく手法で認識率を算出した結果ものせた．この参考実験は，学習データにも分析再合成を用いているため，声道長変換をかけない $\alpha = 0.00$ の場合においても，オリジナルのデータで HMM を学習した場合と結果が異なることに注意されたい．

結果から，声道長変換をあまり大きくかけない場合，構造を用いた場合も HMM を用いた場合もほぼ同等の認識率が得られていることがわかる．しかし，声道長変換を大きくかけた場合，HMM を用いた場合はチャンスレベル（= 0.83%）まで認識率が低下するのに対し，構造を用いた場合は認識率の低下が比較的小さい．よって，構造を用いた音声認識は，HMM を用いた場合と比較して話者の違いに頑健な枠組みとなっているといえる．ただ，学習データと評価データのミスマッチなしで HMM を用いた場合と比較から，構造を用いても，声道長のミスマッチに起因する認識率の低下は生じていることがわかる．その理由は，構造分析時における分布推定において，分布を対角共分散型の多次元ガウス分布と仮定することに起因する推定誤差が，声道長変換にしたがって大きくなっていくためだと考えられる．

3.2　雑音駆動音声の認識

雑音駆動音声（Noise Vocoded Speech: NVS）という音声が知られている [13]．これは、音声を少数の帯域に分割し，各帯域のパワーの時間変化を保存した形で雑音源と置換することで合成される音声である．ヒトは NVS を頑健に認識できることが知られている．

この NVS を，1) 構造に基づく手法，2) 単語 HMM を用いる手法で認識する実験を行なった．認識タスク，学習話者，評価話者，音響分析条件などは先の実験と同じである．学習データはオリジナルの学習話者の音声を用い，評価データには 0–600，600–1500，1500–2100，2100–4000，4000–8000[Hz] の 5 帯域で分割した NVS を用いた．マルチストリーム構造化のストリーム数は，実験的に最も良好な結果を示した 13 を選択した．

結果を Fig.4 に示す．結果から，オリジナル音声の認識時には構造を用いても単語 HMM を用いてもほとんど認識率の差はないが，NVS を認識させると，

HMM を用いた場合より構造を用いた方が頑健に音声認識できることがわかる．

4 考察

NVS に関しては，多くの知覚実験が行なわれており，さまざまな知見が蓄積されている [13, 14, 15]．多くの研究者が，NVS はパワーの振幅包絡を保存したまま，周波数の分解能を著しくて以下させた音声であることから，パワー振幅包絡情報が音声知覚に必要な最小条件らしいということを述べている．具体的に，例えば力丸らは，1) 話者性は男女の差を除き知覚できない，2) 簡単な訓練により 3 帯域以上に分割された NVS を 80%を超える了解度で知覚できる，3) 1 帯域 NVS はほとんど知覚できない，という性質をもつことを示している [13]．

ここで，音声の構造的表象は，任意の 2 分布間の距離を計算することから，ヒトの聴覚と同様パワー振幅包絡の情報を捉えているということができる．これは，構造を用いた音声認識が NVS に対する頑健性を持っていたことからも説明がつく．また，構造を用いた音声認識は，話者性対し強い頑健性を持っていることを示したが，これは，上記 1) に書いたように，NVS に含まれる情報からではヒトは個人性を知覚できないことと類似している．さらに，構造を用いた音声認識では，強すぎる不変性問題に対処するため，マルチストリーム化によって音声に対してある程度の周波数分解能を要求していたが，これは，上記 2), 3) に書いたように，NVS の知覚の際にもある程度の周波数分解能が要求されることと類似している．

一方，HMM が出力する音響特徴量としてデルタのような動的特徴量を用いるモデル化も，パワーの振幅包絡の情報を捉えているとはいえる．しかし，構造を用いた音声認識と異なり，HMM の出力としてデルタ特徴量を用いても，NVS 化や話者性に対する頑健性はあまり向上しない[1]．この違いは，パワーの振幅包絡を捉える方法の違いによるものと考えられる．デルタ特徴量では，パワーの振幅包絡成分を各帯域内のみでの動きと捉えているのに対し，構造は複数帯域にまたがる各部分空間内での動きとして捉えている．そのため，各帯域内のみでの変形である背景雑音や乗算性の変形に関してはデルタ特徴量は有効に働くが，NVS 化や声道長変換のような複数帯域にまたがる変換に関してはデルタ特徴量はあまり有効には働かない．

以上のことから，音声の構造的表象を用いた音声モデルは，ヒトが NVS を知覚する際に用いていると考えられるパワーの振幅包絡情報を，よりヒトに近い方法で捉えているモデルであると考えることができる．

[1] 3.2 節の単語 HMM を用いた NVS の認識実験を，FBANK とデルタの結合ベクトルではなく，デルタのみを用いて行うと，認識率は 14.65%に微増する．

5 まとめ

本稿では，スペクトル領域特徴量を用いた音声の構造的表象を用い，多様な音声の認識実験を行なった．その結果，HMM を用いた手法と構造を用いた手法を比較して，構造を用いた手法が話者性に頑健であること，雑音駆動音声化に頑健であることがわかった．そして，提案する表象が，ヒトの聴覚における音声情報処理の一つのモデルとなることを考察した．

今後の課題としては，今回得られた知見を，不特定話者音声認識システムの認識率向上にフィードバックすることが上げられる．

参考文献

[1] 中川, 信学論, J83-D-II(2), 433–457, 2000.

[2] 篠田, 信学論, J87-D-II(2), 371–386, 2004.

[3] Minematsu, *ICASSP*, 889-892, 2005.

[4] 朝川 他, 音講論（秋）, 2-P-3, 2008.

[5] 鈴木 他, 信学技報, SP2008-32, 73–78, 2008.

[6] 喬 他, 信学技報, SP2008-51, 49–54, 2008.

[7] Qiao *et.al.*, *ASRU*, 576–581, 2007.

[8] 峯松 他, 信学技報, SP2005-13, 9–12, 2005.

[9] Asakawa *et.al.*, *ICASSP*, 4097–4100, 2008.

[10] Young *et.al.*, http://htk.eng.cam.ac.uk/

[11] Kawahara, *Acoust. Sci. Tech.*, 27(6), 2006.

[12] 江森 他, 信学論, J83-D-II(11), 2108–2117, 2000.

[13] 力丸, 日本音響学会誌, 61(5), 273–278, 2005.

[14] 上田 他, 信学技報, H2008-1, 1–6, 2008.

[15] Shannon *et.al.*, *Science*, 270(5234), 303–304, 1995.