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.

スペクトル領域特徴量を用いた音声の構造的表象に関する実験的考察

鈴木　雅之† 　朝川　智†† 　喬　宇† 　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 〒277–8561 千葉県柏市柏の葉 5–1–5
††† 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: †{suzuki,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が，近年，これら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，関係のみを捉えることによって得られる音声の構造的表象である．本稿では，音響的普遍構造を抽出する際に，従来用いられてきたケプストラム領域特徴量ではなく，スペクトル領域特徴量を用いる手法を提案し，それを考察する．提案手法は，スペクトル領域特徴量を用いているために，背景雑音の分離がしやすいという利点がある．また，ヒトの聴覚が周波数解析を行なっていることから，提案する表象はより聴覚系の音声情報表現に近いと考えることができる．本稿では，提案する表象が，従来のケプストラム領域特徴量を用いた構造的表象と同等の強い話者不変性を持つことを実験的に確認し，背景雑音環境下の音声認識における実用性について検討する．さらに，提案手法が雑音駆動音声に対し良好な認識性能を示すことを示し，ヒトの聴覚における音声情報処理との関係についても考察する．
キーワード　音声の構造的表象，スペクトル領域特徴量，音声認識，雑音駆動音声，聴覚特性

Experimental Study of Using Spectrum-based Features for Structural Representation of Speech

M. SUZUKI†, S. ASAKAWA††, Y. QIAO†, N. MINEMATSU†, and K. HIROSE†††

† Grad. School of Engineering, Univ. of Tokyo 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
†† Grad. School of Frontier Sciences, Univ. of Tokyo 5–1–5, Kashiwanoha Kashiwa, Chiba, 277–8561 Japan
††† Grad. School of Info. Sci. and Tech., Univ. of Tokyo 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
E-mail: †{suzuki,asakawa,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Non-linguistic factors such as morphological differences in vocal tracts inevitably affect acoustic features of speech. Recently, a new speech representation, called as structural representation, was proposed which is completely independent of these factors. In the representation, the absolute property of speech events is totally discarded and their relative property is only captured and modeled. In the previous studies, all the discussions on this new representation were done using cepstrum-based features. In this report, spectrum-based features are used for the structural representation and tested for speech recognition. Mathematical and experimental discussions show the followings. 1) The spectrum-based structural representation also has strong speaker-invariance. 2) It can show a better performance of noisy speech recognition compared to cepstrum-based structures. 3) It shows a rather similar performance to humans when noise vocoded speech samples are tested. Finally, we discuss the validity of the spectrum-based structural speech recognition as a model of human speech perception.
Key words　structural representation of speech, spectrum-based features, automatic speech recognition, noise vocoded speech, auditory characteristics

1. はじめに

音声は，生成の際には話者の声道形状の特性，収録や伝送の際には音響機器の特性や背景雑音，といった様々な非言語的特徴によって不可避的に変形する．そのため，音声認識のような音声から言語情報を抽出するタスクでは，非言語的特徴による

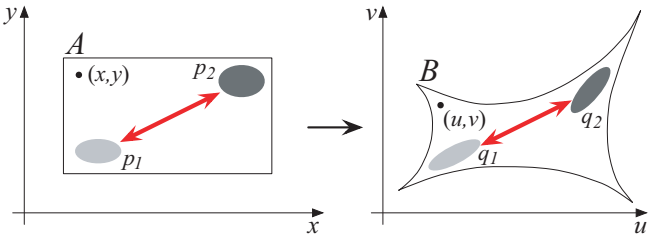


図 1　空間の線形／非線形写像

Fig. 1　Linear or non-linear mapping between two spaces

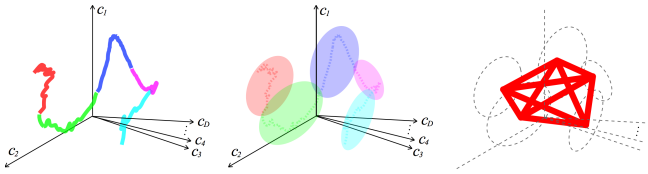


図 2　一発声に内在する音響的普遍構造

Fig. 2　Invariant structuralization of an utterance

変形に頑健であることが要求される．

従来の音声認識技術は，不特定話者モデルに代表されるように，データをたくさん集めることで頑健性を向上させようとしてきた．しかし，そもそも非言語的特徴によって変形したデータをたくさん集めたところで，頑健性の向上には限界がある．例えば Moore は，ヒトが生涯耳にする音声データをすべて機械の学習に用いたとしても，現在の音声認識技術では人間の認識性能に遠く及ばないことを指摘している [1]．

これに対して近年，声道形状の特性や音響機器の特性といった非言語的特徴を原理的に保有しない音声表象として，音響的普遍構造が提案された [2]．これは，音韻構造論に基づく音声の構造的表象であり，従来の音響音声学に基づく特徴量では捉えられていなかった，音声イベント間の関係の情報を表現するものである．この音声の構造的表象は音声認識に利用され，日本語 5 母音系列の連続発声の認識実験の結果，学習話者わずか 8 名で 99.0% の認識率を実現している [3]．

本稿では，音声の構造的表象の抽出過程において用いられる音響特徴量として，従来用いられてきたケプストラム領域特徴量の代わりにスペクトル領域特徴量を用いる手法を提案する．ケプストラム領域特徴量を用いると，ある周波数帯域のみに重畳した雑音の影響が全体に広がってしまうという欠点がある．一方，スペクトル領域特徴量を用いると，雑音の分離がしやすく有利になる [4]．また，ヒトの聴覚は周波数解析を行なっていることから，提案する表象はより聴覚系における情報表現に近いと考えることができる．

2.　音声の構造的表象

2.1　話者性に不変な音響的普遍構造

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，話者性を微分可能で可逆な音響特徴量の空間写像と仮定することが多い．図 1 に，空間 A と空間 B における微分可能で可逆な線形／非線形写像の例を示す．それぞれの空間において，事象は分布として与えられており，空間 A における事象 p_i は，空間 B における事象 q_i に写像される．ここで，それぞれの空間における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 [5]．すなわち，音声事象分布間の f-divergence は，話者性に不変である．

ここでは，f-divergence の一種である Bhattacharyya Distance（BD）を用いた場合の証明を行なう．空間 A，B の対応する 2 点を (x, y)，(u, v) とし，$x = f(u, v)$，$y = g(u, v)$ とすると，以下の式が得られる．

$$q_i(u,v) = p_i(f(u,v), g(u,v))|J(u,v)| \qquad (1)$$

ここで $J(u, v)$ はヤコビアンであり，ヤコビアンの存在は空間写像が微分可能で可逆なものであることから保証される．式 (1) を用いて，BD の不変性は以下のように証明できる．

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

$$= -\log \iint \sqrt{p_1(f(u,v),g(u,v)) \cdot p_2(f(u,v),g(u,v))}|J|dudv$$

$$= -\log \iint \sqrt{p_1(f(u,v),g(u,v))|J| \cdot p_2(f(u,v),g(u,v))|J|}dudv$$

$$= -\log \iint \sqrt{q_1(u,v)q_2(u,v)}dudv = BD(q_1,q_2) \quad \square$$

図 2 に，この話者不変量に基づく，一発声に内在する音響的普遍構造の抽出法を示す．まず，音声から音響特徴量ベクトル系列を抽出する．すると特徴量空間において，音声は時間軌跡を描くことになる．次に，特徴が似たものをマージさせて分布の系列をつくる．最後に，全ての 2 分布間の BD を計算する．ここで，全ての 2 分布間距離を計算することは，幾何学的には構造を決定することに他ならない．そしてこの構造は，変換不変量である BD によって張られているので，構造も変換不変である．よって，このように抽出された音声の構造的表象は，話者性に不変な，全ての話者の発声に普遍的に観測される音響的構造となる．

2.2　音声の構造的表象に基づく音声認識

音声の構造的表象に基づく単語音声認識の枠組みを図 3 に示す．まず，図 2 に示した方法で，音響的普遍構造を抽出していく．分布系列の推定には，Hidden Markov Model（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得る．ここで，少ないデータ量から頑健に HMM のパラメータを推定するために，最大事後確率推定を導入する [6]．分布推定後，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，上三角成分のみで情報がすべて表現される．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よぶ．

2 つの構造間の差異は，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定義する．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

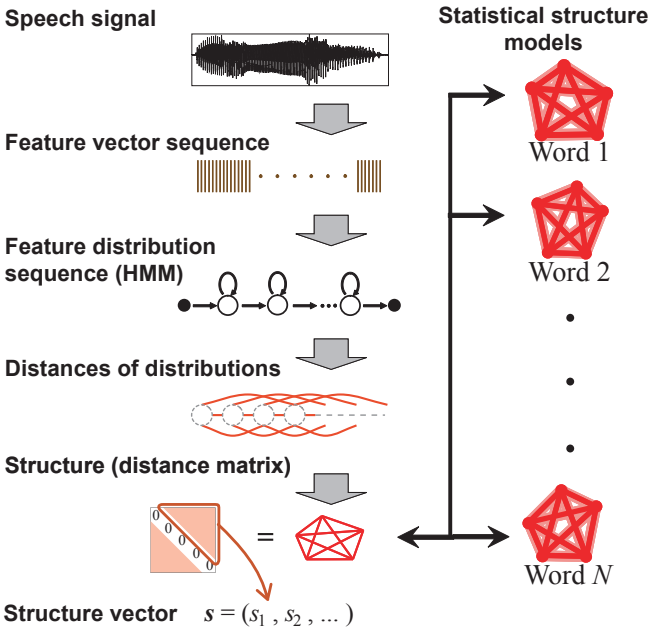


図 3　構造表象に基づく単語音声認識

Fig. 3　Speech recognition based on structural representation of speech

適応及び正規化処理に相当する．この構造間の差異は，構造ベクトル間のユークリッド距離でよく近似されることが示されている [7]．つまり，構造に基づく音声認識の枠組みは，話者正規化／適応化を明示的に行なっていないにもかかわらず，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．

最後に，入力音声からとりだした構造ベクトルに対する音響モデルの対数尤度を計算することによって認識を行なう．ここで音響モデルは，複数の構造ベクトルから得られる多次元ガウス分布を用いる．これを構造統計モデルとよぶ．

2.3　非言語的特徴によるケプストラムの変形

音響的普遍構造は，微分可能で可逆な空間写像に対し不変である．よって音響的普遍構造は，そのような空間写像として表現されるすべての非言語的特徴に不変となる．しかし実装の都合上，音響事象分布を多次元ガウス分布と仮定しているため，写像によってガウス分布がガウス分布でない分布に変換される場合，音響普遍構造の抽出時に歪みが生じる．そこで，音声に混入する非言語的特徴が，音響特徴量の一つであるケプストラムベクトル $\boldsymbol{c}$ を具体的にどう変形させるのかについて考える．

音声に混入する非言語的特徴による変形は，加算性変形・乗算性変形・線形変換性変形の 3 種類に分類される．加算性変形は，時間軸上の加算として表現される変形であり，背景雑音がその典型例である．これは，スペクトルサブトラクションの際に用いられる仮定のように，音声と雑音の独立性と雑音の定常性を仮定すれば，二乗スペクトルに対する加算として扱える．これは，$\boldsymbol{c}$ に対する非線形変換となる．すなわち，一般的にガウス分布を非ガウス分布へと写像する変換となる．

乗算性変形は，伝達関数の乗算で表現される変形であり，音響機器の特性がその典型例である．これは，$\boldsymbol{c}$ に対する定ベクトル $\boldsymbol{b}_c$ の加算 $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{b}_c$ に相当する．また，cepstral mean normalization によって話者性の違いの影響を軽減できることから，話者の声道形状の違いの一部も近似的に乗算性変形であると考えられる．これはガウス分布をガウス分布へと写像する変換である．

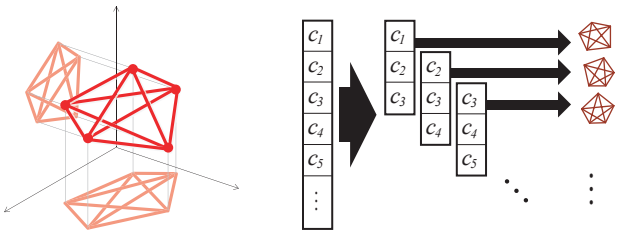


図 4　マルチストリーム構造化

Fig. 4　Multi-stream structuralization

線形変換性変形は，$\boldsymbol{c}$ に対する定行列 $\boldsymbol{A}_c$ の乗算 $\boldsymbol{c}' = \boldsymbol{A}_c\boldsymbol{c}$ で表現される変形である．話者の声道長の差異，聴取者の聴覚特性は，近似的にスペクトル領域特徴量に対する周波数ウォーピングとして表現されるが，単調増加かつ連続である周波数ウォーピングは，$\boldsymbol{c}$ に対する $\boldsymbol{A}_c$ の乗算で表されることが示されている [8]．即ち，声道長の差異，聴覚特性の差異は近似的に線形変換性変形として扱うことができる．これも，ガウス分布をガウス分布へと写像する変換である．

以上をまとめると，アフィン変換 $\boldsymbol{c}' = \boldsymbol{A}_c\boldsymbol{c} + \boldsymbol{b}_c$ で表現される乗算性変形と線形変換性変形に対し，ケプストラム空間において抽出した音響普遍構造は歪みが生じない(注1)．加算性変形は，$\boldsymbol{c}$ に対する非線形変換で近似されるため，理論的に音響普遍構造は不変であるが，音響普遍構造の抽出時に歪みが生じる．

2.4　ケプストラム空間でのマルチストリーム構造化

音響的普遍構造は，音響特徴量空間に対するあらゆる微分可能で可逆な空間写像に不変であった．しかし，あらゆる写像を許してしまうと，非言語的特徴を表す変換としてあり得ないような変換も許すこととなり，全く違う単語同士がマッチしてしまうことが起こりうる．そこで，構造を部分空間に射影し，部分空間における構造表象を用いることで変換に制約をかけるマルチストリーム構造化が提案されている [9]．

マルチストリーム構造化は，声道長の変化を表す行列 $\boldsymbol{A}_c$ の帯行列性に基づく手法である．江森らは，$\boldsymbol{A}_c$ が以下のように示されることを示している [10]．

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

ここで α は周波数ウォーピングの度合いを表すウォーピングパラメータであり，絶対値が大きいほど声道長を大きく変化させ

(注1)：実際は，多次元ガウス分布の分散共分散行列を対角と仮定して実装を行なっているため，歪みが生じないわけではない．

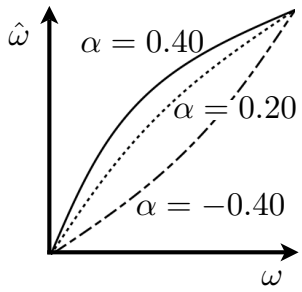


図 5　周波数ウォーピング関数
Fig. 5　Frequency warping function

ることになる．江森らは，一般的に α は十分小さく，α の 2 次以上の項は無視できるとしている．つまり，行列 $\boldsymbol{A}_c$ は，対角付近にのみ値を持つ帯行列として近似できる．

マルチストリーム構造化は，$\boldsymbol{A}_c$ の値が 0 の部分に相当する変換を禁止する手法である．図 4 に，マルチストリーム構造化の概念図を示す. [9] では，マルチストリーム構造化により，認識率が大幅に向上することを報告している．

2.5　非言語的特徴による対数スペクトルの歪み

音声の構造的表象に関する一連の研究では，ケプストラム空間における構造を用いてきた．本稿では，対数スペクトル空間において構造を抽出することを考える．

対数スペクトル空間は，ケプストラム空間と直交変換の関係にあるので，対数スペクトル空間にも音響的普遍構造が存在する．また，直交変換はガウス分布をガウス分布に変換するので，非言語的特徴によるガウス分布の変形についても，ケプストラム空間における議論とまったく同じ議論が可能である．すなわち，対数スペクトルベクトル $\boldsymbol{s}$ の変形について考えると，加算性雑音は $\boldsymbol{s}$ に対する非線形変換，乗算性変形は $\boldsymbol{s}' = \boldsymbol{s} + \boldsymbol{b}_s$，線形変換性変形は $\boldsymbol{s}' = \boldsymbol{A}_s \boldsymbol{s}$ で表現することができる．

ここで加算性雑音は，$\boldsymbol{s}$ に対する非線形変換となっているが，これは定ベクトル $\boldsymbol{n}$ を用いて $\boldsymbol{s}' = \log(\exp \boldsymbol{s} + \boldsymbol{n})$ として近似できる．よって，スペクトルのある領域のみに雑音が重畳した場合，その影響はスペクトル領域全体には広がらない．

以上をまとめると，アフィン変換 $\boldsymbol{s}' = \boldsymbol{A}_s \boldsymbol{s} + \boldsymbol{b}_s$ で表現される乗算性変形と線形変換性変形に対し，対数スペクトル空間において抽出した音響的普遍構造は歪みが生じない．加算性変形は，$\boldsymbol{s}$ に対する $\boldsymbol{s}' = \log(\exp \boldsymbol{s} + \boldsymbol{n})$ で近似されるため，音響的普遍構造の抽出時に歪みが生じる．しかし，スペクトルのある領域のみに雑音が重畳した場合，音響的普遍構造抽出時の歪みはその領域のみに限られることになる．

2.6　対数スペクトル空間でのマルチストリーム構造化

対数スペクトル空間においても，ケプストラム空間同様マルチストリーム構造化は有効であると考えることができる．図 5 に，江森らが式 (2) を導出する際に仮定している周波数ウォーピング関数を示す．ウォーピング関数が直線 $\omega = \hat{\omega}$ の近くに存在していることから，周波数ウォーピングは近い周波数帯域間のみでの変形しか与えないと仮定できる．すなわち，$\boldsymbol{A}_s$ は帯行列であると近似することができ，対数スペクトル空間のマルチストリーム構造化が有効であると考えられる．逆にいえば，対数スペクトル空間におけるマルチストリーム構造化は，ある帯域間での変換のみを許容するという物理的意味付けができる．

表 1　音響分析条件
Table 1　Acoustic analysis condition

Sampling	16bit / 16kHz
Window	25ms length and 10ms shift
Distribution	A gaussian with a diagonal matrix
#distributions	25 per utterance

3. 実　　験

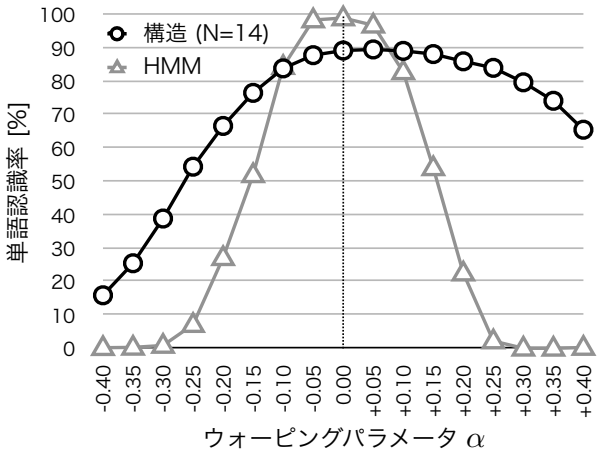
3.1　声道長を変化させた音声の認識

スペクトル領域特徴量を用いた音声の構造的表象（以下，スペクトル構造とよぶ）を用いた音声認識の枠組みが話者性の違いに頑健であることを確かめるために，擬似的に声道長を変化させた音声を，特徴量として 1) 対数スペクトル 2) ケプストラム，音響モデルとして a) 構造統計モデル b) 単語 HMM を用いて認識する実験を行なった．

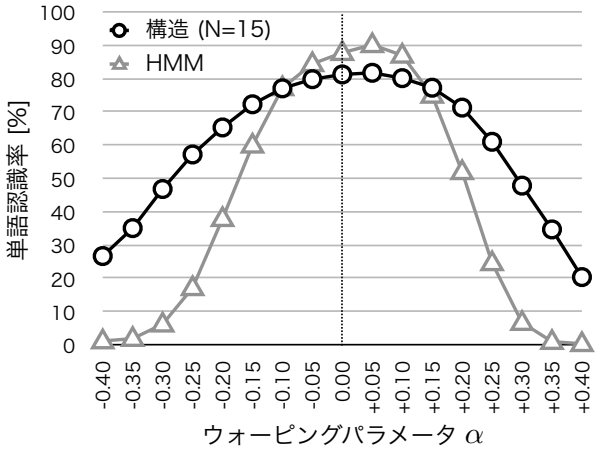
認識タスクは日本語 5 母音を重複を許さず 5 つ連結した/aiueo/など計 120 語の単語認識である．学習データには，日本人成人 8 名（男女 4 名ずつ）が，それぞれの単語を 5 回ずつ発声したデータを用いた．評価データには，学習データと異なる日本人成人 8 名（男女 4 名ずつ）が，それぞれの単語を 5 回ずつ発声したデータに対し，STRAIGHT [11] を用いて声道長変換に相当する周波数ウォーピングをかけた分析再合成音声を用いた．音響分析条件を表 1 に示す．音響特徴量には，スペクトル領域特徴量として MFCC を逆 DCT したもの（24 次元）及び FBANK [12]（24 次元），ケプストラム領域特徴量として MFCC（12 次元）の合計 3 種類を用いた．また，どの特徴量を用いた場合にも，分布推定時と HMM を用いた認識時には，Δ パラメータを付加した．マルチストリーム構造化の各ストリームの次元数は統一し，ストリーム数は実験的に決定した．

実験結果を図 6 に示す．ただし，α はウォーピングパラメータであり，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，$\alpha = \pm\, 0.40$ のときにおよそ声道長を 1/2 倍，2 倍する変換となる．また，N はストリーム数を表す．

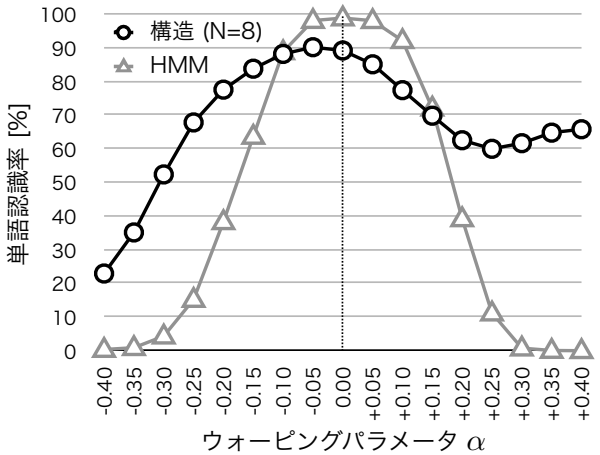
結果，どの特徴量を用いた場合も概ね似た傾向を示している．まず，声道長変換を大きくかけない場合は，構造より HMM を用いた認識率の方がよくなる．しかし，声道長変換を大きくかけた場合，HMM はチャンスレベル（= 0.83%）まで認識率が低下するが，構造は認識率の低下が小さい．よって，用いる特徴量に依存せず，構造音声認識は話者の違いに頑健な枠組みとなっているといえる．ただ，理論的に考えれば，構造音声認識の認識率は周波数ウォーピングに対して一定であることが予想される．それにもかかわらず認識率が低下している理由には，分布系列の推定時に多次元ガウス分布の分散共分散を対角と仮定しているために生じる歪みが関係していると考えている．また，α の正負における性能の非対称性は，母音を識別する情報量がスペクトルの帯域によって異なるためだと考えている．



(a) MFCC の逆 DCT の結果



(b) FBANK の結果



(c) MFCC の結果

図 6　声道長を変化させたときの認識率の変化

Fig. 6　Recognition rate vs. warping parameter

3.2　雑音重畳音声の認識

スペクトル構造を用いた音声認識の枠組みの雑音に対する頑健性を調べるため，雑音を重畳した音声をスペクトル構造／ケプストラム構造を用いて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音声を用い，評価データには様々な種類の雑音を 3 種類の SNR（0dB, 10dB，20dB）で重畳した音声を用いた．音響特徴量は，スペクトル特徴量として MFCC の逆 DCT を，ケプストラム特徴量として MFCC を用いた．次元分割のストリーム数は常に各ストリームの次元数が 2 となるように設定した．

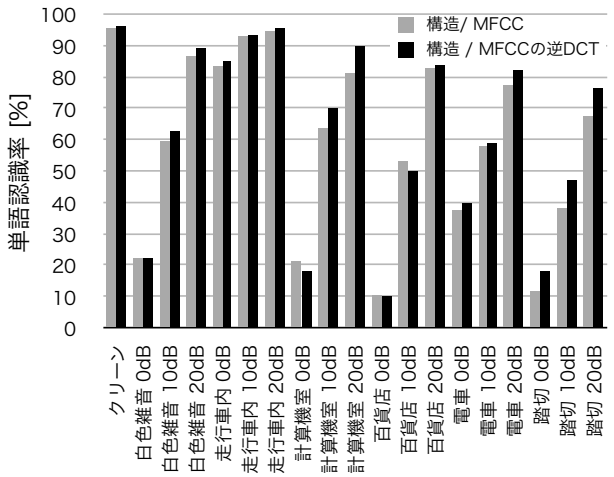


図 7　スペクトル構造とケプストラム構造の認識率の差

Fig. 7　Comparison of the recognition rates between spectrum-based structure and cepstrum-based structure

実験結果を図 7 に示す．スペクトル構造の方がケプストラム構造より加算性雑音に頑健であることがわかる．特に，踏切のような特定のスペクトル領域のみに雑音が重畳しているような雑音音声に対して，スペクトル構造の有意性が強い．

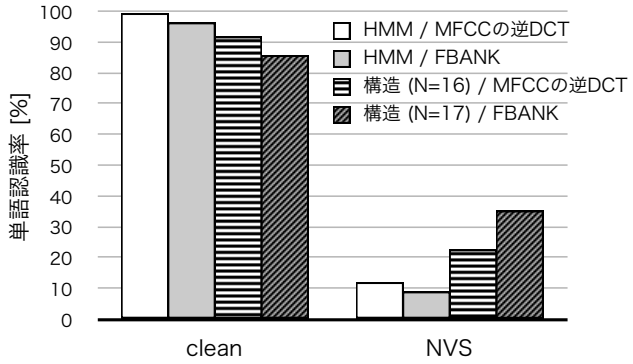
3.3　雑音駆動音声の認識

劣化雑音音声（Noise Vocoded Speech: NVS）という音声が知られている [13]．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．

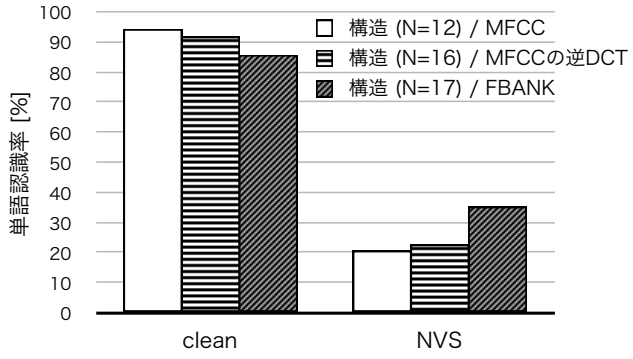
この NVS を，1) スペクトル構造，2) ケプストラム構造，3) 単語 HMM で認識する実験を行なった．認識タスク，学習話者，評価話者，音響分析条件は先の実験と同じである．学習データはオリジナルの学習話者の音声を用い，評価データには 0–600, 600–1500，1500–2100，2100–4000，4000–8000[Hz] の 5 帯域で分割した NVS を用いた．音響特徴量はスペクトル特徴量として MFCC の逆 DCT 及び FBANK，ケプストラム特徴量として MFCC の合計 3 種類を用いた．次元分割のストリーム数は実験的に決定した．

結果を図 8 に示す．ただし，N はストリーム数を表す．図 8 の (a) より，HMM より構造の方が NVS 化に頑健であることがわかる．図 8 の (b) より，ケプストラム構造よりスペクトル構造の方が NVS 化に頑健であることがわかる．

以下，この結果について考察していく．力丸は，NVS の知覚に関して，1) 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，2) 1 帯域 NVS はほとんど知覚できない，3) 話者性は男女の差を除き知覚できない，という性質をもつことを示し，NVS の知覚が通常と異なる知覚



(a) 周波数領域特徴量を用い，HMM と構造で認識率を比較した結果



(b) 構造音声認識で，特徴量を変えて認識率を比較した結果

図 8　clean 音声と NVS の認識率の比較

Fig. 8　Comparison of the recognition rates of cleen utterance and noise vocoded speech

方式によって実現されていることを考察している [13]．また上田らは，NVS の知覚にはパワーの振幅包絡，即ち動きの情報が重要であることを主張している [14]．これらの結果は，ヒトの聴覚は周波数の分解能を著しくて以下させても，パワーの振幅包絡に頼って音声知覚が可能であることを示している．つまり NVS の知覚は，孤立発声音 [あ] を/あ/と知覚するときのような静的なスペクトルの包絡特性に頼った音声知覚とは，まったく異なる音声知覚の枠組みであると考えられる [15]．

ここで，スペクトル特徴量を用いた構造音声認識の枠組みを考えると，1) 音声の動きの情報をとらえている，2) マルチストリーム化はスペクトルにある程度の解像度を要求する処理である，3) 話者不変性が非常に高い，という特徴がある．この特徴は，NVS に関するヒトの聴覚の特性と非常に類似性が高い．3 帯域以上に分割された NVS のようにスペクトルにある程度の解像度があれば，静的なスペクトル包絡を捨てて動的な振幅包絡に頼った音声認識が可能であり，そこには話者の情報はほとんど残っていない．これは，NVS に対するヒトの知覚にも，スペクトル構造を用いた音声認識にもあてはまる．

従来の HMM を用いた音声認識が，静的なスペクトル包絡に頼って孤立発声音 [あ] を/あ/と知覚するヒトの聴覚能力をモデル化しているのだとすれば，提案するスペクトル構造を用いた音声認識は，動的な振幅包絡に頼って NVS を知覚するヒトの聴覚能力をモデル化しているということができる．

4. ま と め

本稿では，音声の構造的表象を，スペクトル領域特徴量を用いて抽出し，それに関する実験的考察を行なった．まず，スペクトル領域の特徴量を用いた音声の構造的表象を用いて，声道長を変化させた音声を頑健に認識できることを実験的に示した．次に，雑音環境下における認識率が向上することを実験的に示した．最後に，雑音駆動音声の認識実験から，提案する表象がヒトの聴覚における音声情報処理のモデルとなることを考察した．

今後の課題としては，マルチバンド音声認識のように各ストリームに重み付けを行なって雑音頑健性を向上させることがあげられる．また，スペクトル包絡情報を用いる HMM と，振幅包絡情報を用いる提案手法とでは，音声知覚のまったく違う部分をモデル化しているという観点から，これら 2 つの手法を融合させることによってより頑健な音声認識手法を考案することがあげられる．

文　献

[1] R. K. Moore, “A Comparison of the Data Requirement of Automatic Speech Recognition Systems and Human Listeners,” *Proc. EUROSPEECH*, pp.2582-2584 (2003)

[2] N. Minematsu, “Mathmatical evidence of the acoustic universal structure in speech,” *Proc. ICASSP*, pp. 889-892 (2005)

[3] Y. Qiao *et. al.*, “Discriminant Analysis of Eigen-Structure for Speech Recognition,” *Proc. ICSLP’2008*, (submitted)

[4] 西村義隆 他, “周波数帯域ごとの重みつき尤度を用いた雑音に頑健な音声認識,” 信学技報, SP2003-103, pp.19–24 (2003)

[5] Y. Qiao *et. al.*, “f-Divergence is a Family of Invariant Measure Between Distributions,” *Proc. ICSLP’2008*, (submitted)

[6] 村上隆夫 他, “音声の構造的表象に基づく日本語孤立母音系列を対象とした音声認識,” 電子情報通信学会論文誌, vol. J91-A, no.2, pp.181–191 (2008)

[7] 峯松信明 他, “音声の構造的表象とその距離尺度,” 信学技報, SP2005-13, pp. 9–12 (2005)

[8] M. Pitz *et. al.*, “Vocal Tract Normalization Equals Linear Transformation in Cepstral Space,” *IEEE Transaction on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944 (2005)

[9] S. Asakawa *et. al.*, “Muti-stream Parametarization for Strucural Speech Recognition,” *Proc. ICASSP*, pp.4097–4100 (2008)

[10] 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-D-II, no.11, pp.2108–2117 (2000)

[11] H. Kawahara, “STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoustic Science and Technology*, Vol. 27, No. 6 (2006)

[12] S. Young *et. al.*, “The HTKBook,” Entropic Cambridge Research Laboratory, ver. 3.4 (2006)

[13] 力丸裕, “劣化雑音音声の聞こえ,” 日本音響学会誌, vol.61, no.5, pp.273–278 (2005)

[14] 上田和夫 他, “臨界帯域フィルターを用いたイギリス英語音声の主成分分析” 日本音響学会聴覚研究会資料, vol. 37, no. 10 (2007)

[15] 峯松信明, “「あ」という声を聞いて母音「あ」と同定する能力は音声言語運用に必要か？,” 日本語学 4 月号, pp. 187–197, 明治書院 (2008)