孤立音を対象とした構造的表象の理論的考察と実験的検討*

○甲斐常伸，鈴木雅之，齋藤大輔，峯松信明，広瀬啓吉 (東大)

1　はじめに

音声の物理的特徴量はたとえ言語情報が同じであっても，非言語的条件の違い（話者や収録条件の違い）によって変動してしまう。このような変動を考慮した音声モデルとして隠れマルコフモデル (Hidden Markov Model: HMM) がある。HMM は各時刻の特徴量を確率分布からの出力として捉え，その分布の時系列として音声の時間的変化を記述する。各分布を，多数の話者，多数の環境の音声より推定することで，非言語的条件の違いによる特徴量の変動は，出力特徴量の確率的な変動に対応することになる。しかし，確率的変動では表しきれない音声変動が入力されることも多く，その場合，HMM パラメータを修正する適応技術や特徴量を修正する正規化技術を適用することになる。

これに対して，非言語的変動に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 [1, 2]。これは，非言語的条件の違いによる音声変動を写像として捉え，任意の連続かつ可逆な写像に対して不変な f-divergence のみを特徴量として音声を表現することで，変動に不変な音声モデルを導出している。構造的表象を利用し，適応や正規化を施すことなく，頑健な単語音声認識や発音評価を実現している [3, 4]。

従来の構造的表象は，発声を分布（事象）群に変換し，事象間距離を計測することで構成していたため，孤立音を扱うことは原理的に不可能であった。しかし，f-div. の写像不変性は数学的には，分布間距離の不変性を主張しているだけであり，孤立音を分布群として表現できれば，応用することができる。本稿では，ある時刻のスペクトル包絡を分布群として表現することで，構造表象の考えを導入することを試みる。また，孤立母音認識実験を通してその有効性を検証する。

2　音声の構造的表象

ケプストラム空間における一軌跡（発声）を考える。これを，HMM 学習などを通して分布列へと変換し，全ての二分布間距離を f-div. で計測する（Fig. 1 参照）。

$$f_{div}(p_i, p_j) = \oint p_j(\boldsymbol{x}) g\left(\frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\right) d\boldsymbol{x} \qquad (1)$$

f-div. は任意の連続かつ可逆な写像に不変であり，また，任意の連続かつ可逆な写像に不変な事象間距離は f-div. のみである（Fig. 2 参照）[5]。このようにして

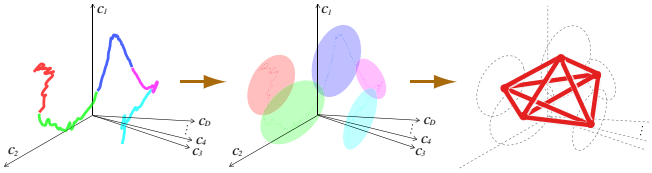

Fig. 1　Utterance to structure conversion

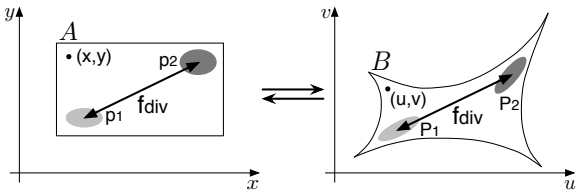

Fig. 2　Transform-invariance of f-divergence

得られた写像不変な距離行列を構造的表象と呼んでいる。より具体的には，(1) 式において，$g(t) = \sqrt{t}$ とした時の $-\ln(f\text{-div.})$ として定義されるバタチャリヤ距離（以下 BD）を用いる。この時，各分布が単一正規分布に従うとすると BD は次のようになる。

$$\mathrm{BD}(p_i, p_j) = -\ln\left(\oint \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})} d\boldsymbol{x}\right) \qquad (2)$$

$$= \frac{1}{8}\mu_{ij}^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} \mu_{ij} + \frac{1}{2}\ln\frac{|(\Sigma_i + \Sigma_j)/2|}{|\Sigma_i|^{\frac{1}{2}}|\Sigma_j|^{\frac{1}{2}}} \qquad (3)$$

ここで $\mu_{ij} = \mu_i - \mu_j$ であり，Σ_i は分散共分散行列である。T は転置を，$|\cdot|$ は行列式を表す。

発声の中に事象が n 個あれば，${}_nC_2$ 個の BD を求めることができる（対称型距離行列）。この上三角成分を構造ベクトルと呼ぶ。また，構造ベクトルの絶対値は距離行列を幾何学的な形態と考えた時の，重心と各節点との平均距離を近似できるため（構造の半径に相当），これを構造サイズと呼んでいる。

構造表象を単語音声認識に応用する場合は，各発声から構造を抽出し，その尤度を，統計的な構造モデルより算出する [3]。発音評定に応用する場合は，ある学習者の複数発声から話者依存音素 HMM を構成し，音素単位で距離行列を得る。これを教師距離行列と（形態的に）比較することで発音習熟度を推定する [4]。

いずれの応用もパラメータ空間内の複数の事象間の距離（関係）を捉えており，個々の事象（音）を不変に扱うことはしていない。本稿では孤立音を構造的に扱う方法について検討する。

---

*Theoretical and experimental investigation of structure-based modeling of isolated speech sounds.
By KAI Tsunenobu, SUZUKI Masayuki, SAITO Daisuke, MINEMATSU Nobuaki, HIROSE Keikichi
(The University of Tokyo)

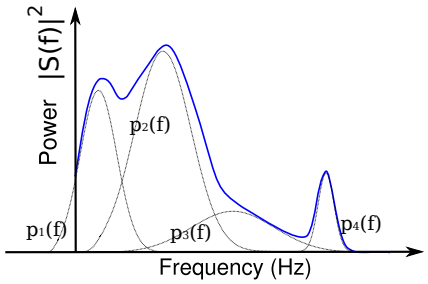

Fig. 3　Spectrum modeling using $\{p_i(f)\}$

3　構造的表象を用いた孤立音のモデル化

3.1　スペクトル包絡の GMM 近似

f-div. の不変性は，分布間距離の写像不変性でしかない。例えばスペクトル包絡を，周波数 f に対する一次元分布群 $\{p_i(f)\}$ の和 $\sum p_i(f)$ でモデル化できれば (Fig. 3 参照)，その分布間の全 f-div. は，如何なる f の連続かつ可逆な変形に対して不変性を有する。しかし，全積分が 1.0 という制約を満たしつつ，任意のスペクトル形状を近似することは困難と考えられるため，各 $p_i(f)$ に対して適切な重みをかけることで，柔軟な近似を考える。即ち，GMM である。

音声のスペクトル包絡に対する GMM 近似[*1]は，分析合成や音声合成の分野で検討されている [6, 7]。本稿でもこれを利用し，時刻 t の包絡を次式で近似する。

$$\frac{|S(f)|^2}{Z} = \sum_{m=1}^{M} w_m \mathcal{N}_m(f; \mu_m, \sigma_m) \tag{4}$$

$S(f)$ は時刻 t の線形スペクトル包絡，Z はエネルギー ($\oint |S(f)|^2 df$) であり，二乗包絡を確率密度分布化するための正規化項である。w_m, μ_m, σ_m は各正規分布の，重み，平均値，標準偏差であり，M は総分布数である。各種パラメータは EM アルゴリズムにより推定される。初期分布としては，全帯域を $N(> M)$ 個に等分割し，各帯域の平均周波数，及び $\frac{F_s}{2N}$ (F_s は標本化周波数) を標準偏差とした N 個の初期分布を用意し，その後，EM アルゴリズムを用いて平均，分散の再推定を繰り返した。推定後，分布数を $N-1$ に減らして再推定を行う。この場合，いずれかの 2 分布を 1 分布にマージした $N-1$ 個の分布を初期分布とする。マージ対象の分布の決定，及び，各包絡毎に決定される最終的な混合数 (M) の決定は [7] を参照して戴きたい。

3.2　各分布の中心周波数とフォルマント周波数

Fig. 3 より明らかなように，BD は，二つのスペクトルピーク間の距離を，f の写像に不変に計測することを狙いとしている。これは，音声科学の世界で古くから議論されている話者不変な特徴量であるフォルマント周波数比を計測することと類似している [8, 9]。声道長の成長による周波数軸変換を $\hat{f} = \alpha f$ と仮定すれば，$\frac{\alpha F_i}{\alpha F_j} = \frac{F_i}{F_j}$ であるため，不変特徴量となる。周波数軸を対数化すれば $\ln(F_i) - \ln(F_j)$ が不変量となる。

しかし，このためには GMM 近似における各正規分布の中心周波数がフォルマント周波数に対応している必要がある。しかし最尤推定で得られた GMM の各分布は必ずしもフォルマント周波数に対応する訳ではない。Nguyen らは各正規分布を，スペクトルピーク，スペクトルトレンドを表現する分布に自動分類し，フォルマント周波数に凡そ対応する分布群の選別法を提案しており，ここでもそれに従った [7] 。

3.3　低周波数帯域の GMM 近似

ピークに対応する正規分布を自動選別したところ，低周波数帯域において不適切な正規分布が生起することが観測された。GMM による近似は Fig. 3 に示すように，$f \to f_{\max}, f_{\min}$ において，滑らかに減少するパターンを前提としているが，音声スペクトルの場合，これは $f \to f_{\min}(= 0)$ では一般に成立しない。このような不適切な形状近似による影響が出ているものと考えられたため，ここでは，スペクトルを負の周波数領域まで考え ($-F_s/2 \le f \le F_s/2$，F_s は標本化周波数)，$f = 0$ において対称形を成したスペクトルを GMM 近似の対象とした。GMM 近似の後，$100 \le f \le F_s/2$ の帯域に中心周波数を持つ，スペクトルピークに対応する正規分布を選択して BD の計算に用いた。100[Hz] 以上としたのは，基本周波数の領域に生起した正規分布を除去することが目的である。

3.4　BD を基本とした BD 的距離尺度の計算

第 2 節で述べたように本来 BD は (3) 式で表され，p_i，p_j が一次元正規分布であれば各々の平均値と分散によって算出される距離である。しかし，GMM の各正規分布には平均値，分散に加えて重み w_m があるため，これを BD にどう組み込むかによって複数の計算法を考えることができる。例えば重み付き分布に対する BD として，以下の BD_1 を定義する。

$$\begin{aligned}
&\mathrm{BD}_1(w_m\mathcal{N}_m, w_n\mathcal{N}_n) \\
&= -\ln\left(\oint \sqrt{w_m\mathcal{N}_m(f) w_n\mathcal{N}_n(f)}df\right) && (5)\\
&= -\ln\left(\sqrt{w_m w_n}\right) - \ln\left(\oint \sqrt{\mathcal{N}_m(f)\mathcal{N}_n(f)}df\right) && (6)\\
&= -\ln\left(\sqrt{w_m w_n}\right) + \mathrm{BD}(\mathcal{N}_m, \mathcal{N}_n) && (7)
\end{aligned}$$

BD_1 は，本来の BD に対して，(7) 式第一項を加算した距離となる。この第一項は $0 \le \sqrt{w_m w_n} \le 1$ であるため，常に正値をとる。つまり，本来の BD に対して常により大きな距離を返すこととなる。この付加項は，二分布の重みの相乗平均が小さいほど大きくなるため，重みの小さな分布間の距離は，重みの大きな分布間の距離に比べて，より広げて計算される。

[*1] 但し，対数スペクトルではなく，$|S(f)|^2$ を縦軸とした場合のスペクトル包絡である。

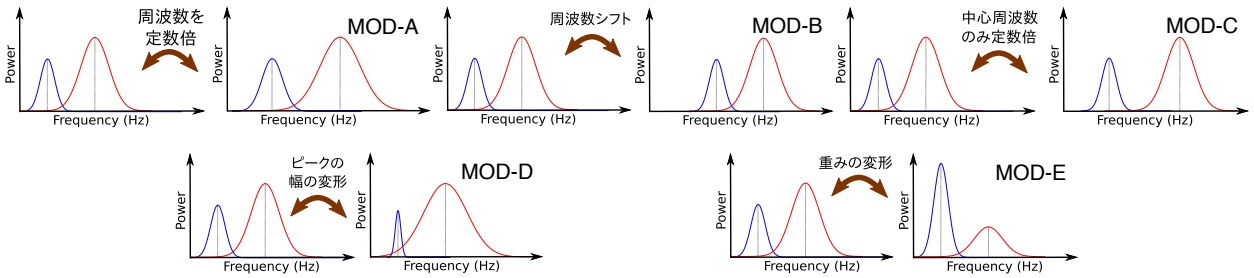

Fig. 4　Several kinds of spectrum modifications

次に以下に示す，BD_2 を定義する。

$$\mathrm{BD}_2(w_m\mathcal{N}_m, w_n\mathcal{N}_n) = \mathrm{BD}(\mathcal{N}_m, \mathcal{N}_n) \tag{8}$$

これは，GMM の重みを一切無視して，純粋に正規分布群だけに着目した距離定義となる。

さらに，(4) 式における Z を右辺の移項し，$|S(f)|^2$ を関数近似した場合の重み付き分布群を考えれば，以下の BD_3 が定義できる。

$$\begin{aligned}&\mathrm{BD}_3(Zw_m\mathcal{N}_m, Zw_n\mathcal{N}_n)\\ &= -\ln\left(\oint \sqrt{Zw_m\mathcal{N}_m(f)Zw_n\mathcal{N}_n(f)}df\right) &(9)\\ &= -\ln(Z) + \mathrm{BD}_1(w_m\mathcal{N}_m, w_n\mathcal{N}_n) &(10)\end{aligned}$$

BD_1 に対して $-\ln(Z)$ が加わる。多くの場合 $Z > 1.0$ となるため，この第一項は負の値になる。Z が大きい場合は，BD_3 は負の値をとることもある。

3.5　不変性の制御に関する定性的考察

構造表象を使った単語音声認識では，f-div. の高すぎる不変性が問題となる。例えば，声道長の変化を近似する演算はケプストラムに対する帯行列の乗算として近似される。即ち「帯行列乗算のみに対する不変性」が実現すべき不変性となるが，これは，次元分割を通して得られる部分空間において構造を構成することで解決できる [3]。このように，構造表象を用いる場合，不変性の制御が一つの技術的検討事項となる。本稿の場合，f の次元分割は不可能であるが，BD を基本として複数の距離尺度を構成することができる。構造表象は音声以外にもロボットの視覚制御などにも応用されているため [10]，ここでは，特徴量の物理次元を限定せず，スペクトル密度関数によって孤立事象が表現された場合に，どのようなスペクトル変形に対して不変性を有するのか否かを BD_i に対して考察する。

スペクトル変形としては，1) 周波数 f の線形変換 $f' = af + b$ による変形[*2]，2) スペクトルピークの中心周波数のみの線形変換による変形，3) スペクトルピークの幅の変換による変形，4) 重み w の変換による変形，を考え（Fig. 4 参照）。各々の変形による不変性について Tab. 1 に示す。なお，ここではスペクトルの物理的対象物を限定してないため，個々の変形の物理的意味（聴感上の意味）については不問にしている。

[*2] f-div. や BD は，本来非線形変換でも不変であるが，正規分布を非線形変換すると非正規分布となるため，ここでは正規性が保たれる線形変換を主な対象としている。

Table 1　各 BD_i が有する写像不変性

写像	F_i 比	BD_1	BD_2	BD_3
MOD-A	不変	不変	不変	非不変
MOD-B	非不変	不変	不変	不変
MOD-C	不変	非不変	非不変	非不変
MOD-D	不変	非不変	非不変	非不変
MOD-E	不変	非不変	不変	非不変

4　孤立母音を対象とした自動認識実験

4.1　ミスマッチ状況下での認識実験

孤立音の構造的表象が話者の違いに対して頑健性を示すか否かを実験的に評価するために，ミスマッチ状況下での孤立母音の認識実験を行った。

実験に用いる音声は，男性 5 名，女性 5 名が孤立母音を各々 5 回ずつ発声した計 250 発声である。各母音に対する HMM を構築する学習データと評価データは，1）男性 5 名で学習，女性 5 名で評価，2）女性 5 名で学習，男性 5 名で評価，というミスマッチの起こる組合せとした。

TANDEM-STRAIGHT[11] で得られたスペクトル包絡を 0 Hz で折り返して対称形にし，GMM の推定をする。音響分析条件を Tab. 2 に示す。推定された正規分布の中から中心周波数が 100 Hz 以上であり，かつフォルマントに対応していると判断された 5 個，あるいは 4 個を自動抽出して BD を計算し構造ベクトルを得る（それぞれ次元数は ${}_5C_2$ 次元，${}_4C_2$ 次元になる）。ここではエネルギー項を考慮した BD_3 を分布間の距離として計算した。これと MFCC 12 次元を組み合わせたものを特徴量として認識を行った。またサイズ正規化を行った構造ベクトル，及び比較のためにフォルマント比（正規分布の中心周波数比）を特徴量とした実験を行った。

実験結果を Tab. 3 に示す。MFCC と組み合わせて用いた特徴量のうちでは，男 5→ 女 5 で + 正規化構造 10 次元 の，女 5→ 男 5 では + 正規化構造 6 次元の認識率が最も高くなっており，MFCC と比べて向上している。しかし，相対的な特徴量を組み合わせても

Table 2　音響分析条件

サンプリング	16bit / 16kHz
窓	Hamming 窓
窓長, シフト長	25msec, 10msec
HMM	1 状態 / 対角共分散行列正規分布
分析対象区間	各母音の中心 200msec

Table 3　実験結果 1

特徴量	男 5→ 女 5	女 5→ 男 5
MFCC 12 次元のみ	39.2 %	60.8 %
＋ 構造 10 次元	39.5 %	49.6 %
6 次元	39.5 %	55.2 %
＋ 正規化構造 10 次元	47.5 %	64.8 %
6 次元	33.0 %	69.6 %
＋ F_i 比 10 次元	43.5 %	63.2 %
6 次元	45.9 %	54.4 %

Table 4　実験結果 2

特徴量	男 4 → 男 1	男 4 → 1.5×男 1
構造 10 次元	76.80 %	53.6 %
6 次元	71.20 %	57.6 %
正規化構造 10 次元	72.80 %	34.4 %
6 次元	73.60 %	17.6 %
F_i 比 10 次元	85.60 %	62.4 %
6 次元	84.00 %	56.0 %

認識率が低下している場合もあり，必ずしも孤立音の構造的表象の頑健性は十分にはみられなかった。

4.2　構造の不変性の確認

第 4.1 節では構造の十分な頑健性が確認できなかったため，より単純な変換である周波数の定数倍という周波数ウォーピングに対して構造が不変に保たれているかを確認する実験を行った。

構造ベクトルの抽出の仕方は第 4.1 節と同様で，MFCC とは組み合わせずに認識を行った。学習データと評価データを，1）男性 4 名で学習，男性 1 名で評価，2）男性 4 名で学習，周波数 1.5 倍の変換を施した男性 1 名で評価，という組合せとした。

実験結果を Tab. 4 に示す。第 3.5 節で述べたように，数式上では BD やフォルマント比は周波数の定数倍という変換に対しては不変である。しかし，変換をかけた音声の認識では認識率が軒並み低下している。これは，GMM で推定した分布がうまくフォルマントに対応していないということを示唆している。実際に変換の前後で，フォルマントに対応すると判断された正規分布の中心周波数をプロットしたものを Fig. 5 に示す。変換前後で，対応する中心周波数がずれていることが確認できる。

5　まとめ

本稿では，従来，特徴量空間の軌跡を分布系列化することで得られる事象群を対象にした構造的表象を，孤立事象を分布群として表現することで適応し，その理論的考察及び実験的検討を行った。音声知覚の恒常性

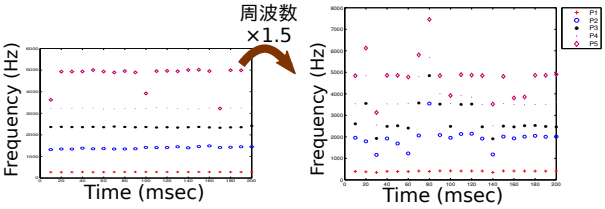

Fig. 5　Means of normal distributions

に対しては，例えば，言語学的観点から [12]，及び，発達心理学的観点から [13]，音声刺激の中の関係性への着眼が古くから報告されている。音響音声学的には，この関係性を複数の刺激群の関係性に求めたり，あるいは，孤立音を複数の成分に分解し，成分間の関係性に求めるなどの検討がある [14, 15]。例えば [14] では，前者を extrinsic な正規化，後者を intrinsic な正規化と呼んでいる。本稿は，この両者の正規化を，f-div. の不変性を用いて一つの枠組みで実装する試みである。実験の結果，構造的表象の不変性が保たれておらず，その有効性は一部の結果においてのみ示された。この原因にはフォルマントと対応しない分布が得られたり，分布の平均周波数が時間的に不連続となるなど，分析系における不具合があると考えられる。今後，これらの不安定性を解消するなどして再度検討する予定である。また，今回，フォルマント比のみを比較対象としたが，スペクトル包絡を周波数の定数倍に対して不変に扱う方法は他にも幾つか提案されており [16, 17, 18]，これらとの比較も検討したい。

参考文献

[1] N. Minematsu, *Proc. ICASSP*, 585–588, 2004.
[2] 峯松他, 電子情報通信学会論文誌, vol.J94-D, 1, 12–26, 2011.
[3] N. Minematsu *et al.*, *J. New Generation Computing*, 28, 3, 299–319, 2010.
[4] M. Suzuki *et al.*, *Proc. INTERSPEECH*, 586–589, 2010.
[5] Y. Qiao *et al.*, *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.
[6] P. Zolfaghari *et al.*, *Proc. ICSLP*, 1229-1232, 1996.
[7] B. P. Nguyen, “Studies on Spectral Modification in Voice Transformation,” 北陸先端科学技術大学院大学博士論文, 2009.
[8] G.E. Peterson *et al.*, *JASA*, 24, 175–184, 1952.
[9] G.E. Peterson, *J. Speech and Hearing Research*, 4, 10–29, 1961.
[10] 郷古他, 人工知能学会全国大会論文集, 2E3-1, 2010.
[11] H. Kawahara *et al.*, *Proc. APSIPA*, 111–120, 2009.
[12] R. Jakobson *et al.*, *The sound shape of language,* Mouton De Gruyter, 1987.
[13] P.K. Kuhl, *Child Phonology*, vol.2, chap.4, Academic Press, 1980.
[14] W. Ainsworth, in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham, 103–113, Academic, 1975.
[15] T. M. Nearey, *JASA*, 85, 5, 2088–2113, 1989.
[16] T. Irino *et al. Speech Communication*, 36, 181–203, 2002.
[17] A. Mertins *et al. Proc. ASRU*, 308–312, 2005.
[18] 益子, 秋音講論, 3-7-2, 105–106, 2005.