構造的表象を用いた音声認識における状態数決定に関する実験的検討*

☆グェンドゥックズイ, 鈴木雅之, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声の音響的特徴は声道形状の特性，音響機器の特性などの非言語的要因によって様々に変形する．これによって，学習時と評価時において音響的なずれが生じ，音声認識においては不一致問題 (mismatch problem) を抱えることになる．従来の音声認識技術では，この不一致問題を解決するために 1) 大規模なデータを用いた統計モデルにより（不特定話者モデルなど）非言語的要因を隠れパラメータとして扱う，2) 音響モデルを話者や環境に対して適応・修正する，3) 特徴量を例えば声道長に対して正規化する，などして対処していた．しかし，データが少量しか得られない場合には，これらの手法は必ずしも効果的であるとは限らない．近年，上記の非言語的変形を特徴量空間の写像として捉え，写像不変量のみを用いて音声を表現する手法が提案され (音声の構造的表象)，孤立単語音声認識 [1]，連続音声認識 [2] においてその有効性が示されている．

構造的表象で用いる写像不変量 f-divergence は分布間距離であるため，構造的表象を音声認識に用いる場合，音声ストリームをまず分布系列に変換する必要がある．そのため，分布数（隠れマルコフモデル (Hidden Markov Model; HMM) を用いて分布系列化するので分布数は状態数でもある）をいかにして決定するかが問題になる．本稿ではモデル選択問題によく使われるベイズ情報量規準 (BIC) を音声の構造的表象における状態数決定問題に適用し，その有効性を実験的に検証する．

2 音声の構造的表象

ある単語発声から短時間特徴量を抽出した後，N 状態の HMM をその発声のみから学習することで特徴量系列を分布化する．N 個の分布に対して $\binom{N}{2}$ 個の全ての 2 分布間距離を求めれば，その幾何学的構造を一意に規定することになる．本稿では，分布間距離として f-divergence の一つである以下の Bhattacharyya 距離を用いる:

$$BD(p_i, p_j) = -\log \int \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})} d\boldsymbol{x} \qquad (1)$$

Bhattacharyya 距離に基づく分布間距離は変換不変であるので，分布間距離により規定される構造も変換不変である．これを音声の構造的表象という．

3 ベイズ情報量規準

ベイズ情報量規準 (Bayesian information criterion; BIC) はパターン認識におけるモデル選択問題の 1 つの有効な解決法である．BIC は最尤法に基づいて，推定されたモデルを評価し，式 (2) のように定義される．

$$BIC(\alpha) = -2 \ln P(D|\lambda) + \alpha k \ln(n) \qquad (2)$$

ここでは，D は学習データ，$\ln P(D|\lambda)$ は推定されたモデル λ に対する D の対数尤度，k はモデルの自由度，n はサンプル数，α はモデル複雑度に対する重み係数である．BIC を用いたモデル選択では，以下のように $BIC(\alpha)$ を最小化する $\hat{k}$ を最適なモデル自由度とする．

$$\hat{k} = \arg\min_{k} BIC(\alpha) \qquad (3)$$

構造的表象を用いた音声認識における状態数決定問題では状態数（N）とモデルの自由度は次の関係：

$$k = N * (MFCC \text{ の次元数}) \qquad (4)$$

があるので，k が求まれば，N も導かれる．

4 実験

4.1 実験条件

本稿では構造的表象に基づく孤立単語音声認識実験を行なって，BIC による状態数決定の有効性を検証する．まず，入力発声をケプストラムベクトル系列に変換する．それを用いて，状態数 N の単語 HMM を学習し，各状態の出力分布を求める．得られた分布系列に対して，マルチストリーム構造化を行ってから線形判別分析（Linear Discriminant Analysis; LDA）を施して，入力発声の構造特徴ベクトルを計算する [3, 4]．構造的な統計モデルを構築する場合は，同一単語の異発声全てを各々構造ベクトル化し，それらを正規分布で近似する．認識時には入力発声を構造ベクトル化し，構造的統計モデルとの照合によって認識結果が得られる．

本実験には東北大・松下単語データベース [5] の音韻バランス単語 212 語の 60 人の話者による単語発話を用いた．単語長は 2 モーラから 7 モーラであった．データセットは 30 名ずつの 2 つのセットに分かれている．これらのデータのうち，セット 1 は構造統計モデルの学習データとして用いた．セット 2 は，さらに 10 人ずつの 3 つのサブセットに分割し，1/3 をバリデーションセット，2/3 を評価セットとする 3-fold クロスバリデーションによって実験を行った．最適な状態数はバリデーションセットを用いて決定し，決定された状態数のもとで評価セットを用いて認識実験を行った．

分析条件を表 1 に示す．マルチストリーム構造化に関してはブロックサイズを 2 にして行った．線形判別分析（LDA）は 2 段階の LDA を用いた．まず，

*An experimental study on determining the number of states per utterance for structural speech recognition. by NGUYEN Duc Duy, SUZUKI Masayuki, SAITO Daisuke, MINEMATSU Nobuaki, HIROSE Keikichi (The University of Tokyo)

Table 1　音響分析条件

サンプリング	16 bit / 16 kHz
窓	25 ms length / 10 ms shift
特徴量	メルケプストラム 12 次元
出力分布	対角共分散ガウス分布
分布推定	MAP・重み 1.0
状態数	5，10，15，20，25，30，35，40

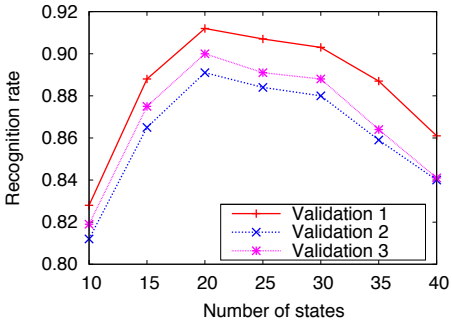

Fig. 1　状態数を固定した場合の各バリデーションセットにおける認識率

マルチストリーム構造化を行った後，それぞれのストリームでの構造ベクトルに対して LDA を適用して次元を削減する（1 段目 LDA）．そして，各ストリームの変換ベクトルを全ストリームで結合したベクトルに対して，更に LDA を行う（2 段目 LDA）．1 段目と 2 段目のそれぞれの LDA の削減次元数は，1 段目 LDA の削減次元数を 20，2 段目 LDA の削減次元数を 211 とした．

4.2　バリデーションセットを用いた認識実験

まず，BIC を用いず，バリデーションセットを用いて状態数を一つ選択して利用する手法を評価した．HMM の状態数を 5 から 40 まで刻み 5 で 1 つずつ固定して（単語長によらず固定の状態数 N を用いる），上記の各バリデーションセットに対して認識実験を行った．状態数 10 から 40 までの実験結果[1]を Fig. 1 に示す．Fig. 1 から分かるように，各バリデーションセットにおいて，最大認識率を出す状態数は 3 つとも 20 であった．そこで，対応する 3 つのテストセットに対して，HMM 状態数を 20 と固定して認識を行った．最終的にテストセットにおける平均認識率は 90.09% が得られた．

4.3　BIC を用いた認識実験

次に，BIC を用いて状態数を入力音声に依存させて変化させる手法を検討した．バリデーションセットを使って，以下のように BIC 係数 α を最適化した．α をある値の固定し，BIC を最小化するモデル自由度 k を入力発声毎に求め，これに基づいて入力発声の状態数を決定する．推定された HMM 状態数を使って，認識を行う．α の値を変化させ，各バリデーションセットに対して最大認識率を出す α を求める．得られた α を使い，対応するテストセット中の各発声の状態数を

[1]状態数 5 での認識率は各 Validation 1，2，3 ではそれぞれ 45%，43%，43%であった

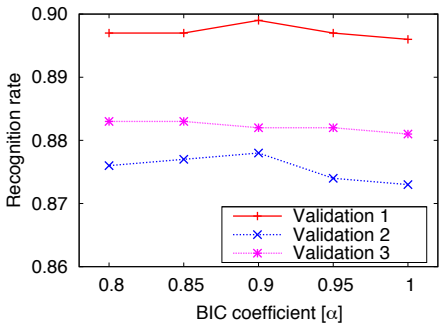

Fig. 2　BIC による状態数を推定した場合の各バリデーションセットにおける認識率

求めた上で分布系列化を行い，認識実験を行なった．バリデーションセットにおける実験結果を Fig. 2 に示す．Fig. 2 から読み取れる通り，Validation1 と 2 では $\alpha = 0.9$，Validation3 では $\alpha = 0.8$ で認識率が最大になる．これらを BIC のパラメータとして用いて，テストセットで認識を行うと，平均認識率 88.47% が得られた．

4.4　考察

今回行なった孤立単語認識実験では，BIC を用いて入力に応じて分布数を決定するよりも，（単語長に依らず）予め決められた分布数で分布系列化することが効果的であった．音声を構造化する場合に，発声の長さに応じて状態数を変化させた方がより良いモデルが得られると思われるが，今回の結果を見る限りにおいては，これは必ずしも成立しないと言える．

5　おわりに

本稿では，構造的表象による音声認識における状態数決定問題について検討を行った．しかし孤立単語認識実験の結果，状態数を固定して利用した方が高い認識率を得られることが分かった．この原因の 1 つとしては同じ単語の発声であっても BIC によって割り当てられた状態数が異なる場合が多発したためであると考えられる．今後，BIC による状態数決定の詳細を調査するほか，構造的表象を用いた音声認識において状態決定と認識を同時に行う枠組みについて検討する予定である．

参考文献

[1] Minematsu，N.，*et al.*，Speech Structure and Its Application to Robust Speech Processing，New Generation Computing，28，299-319，2010.

[2] Suzuki，M.，*et al.*，Continuous digits recognition leveraging invariant structure，in Proc. INTERSPEECH，993-996，2011.

[3] 朝川　智，音声の構造的表象に基づく単語音声認識に関する研究，博士論文，2008.

[4] C.M. ビショップ，パターン認識と機械学習　上下，Springer Japan，2010.

[5] 牧野　他，日本音響学会誌，vol.48，no.12，899-905，1992.