論文 / 著書情報
Article / Book Information

論題(和文)	隠れモードベイズ分類器を用いた音響モデルの適応学習
Title(English)	
著者(和文)	篠崎 隆宏, 古井 貞熙
Authors(English)	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会 2003年秋季講演論文集, Vol. , No. 2-6-2, pp. 63-64
Citation(English)	, Vol. , No. 2-6-2, pp. 63-64
発行日 / Pub. date	2003, 9

隠れモードベイズ分類器を用いた音響モデルの適応学習 *

◎篠崎 隆宏　古井 貞熙　(東工大)

1　はじめに

話し言葉音声では様々な要因により音声の性質が短い単位で変動する。この様な変動に対応するために、発話をクラスタリングし、各クラスに対して音響モデルを作成することが有効と考えられる。これまで音声のクラスタリング法としては k-means 法や尤度行列を用いる方法などが試みられている [1, 2]。本稿ではベイジアンネットの隠れ変数を発話クラスに対応させた、尤度を基準とした分類器の提案を行う。提案手法では音声のクラスを学習データから EM アルゴリズムにより統一的な方法で求めることが出来、また音声に関する様々な情報を組み込むことも容易である。提案手法に基づき講演音声のクラスタリングを行い、評価実験を行った。

2　隠れモードベイズ分類器

図 1 – 4 にベイジアンネットを用いた提案分類器の構成を示す。これらの分類器は尤度を基準として、音声を任意のクラス数に分類することが出来る。図において、離散変数に対応するノードは四角、連続変数は丸で示した。また観測値が与えられるノードは塗りつぶし、隠れ変数として使用するノードは透明としてある。分類を行う際の基本単位としては話者や個々の発話などが考えられるが、本稿では発話毎にクラスタリングを行った。

図 1 に、音響特徴量のみに基づき学習セットからクラスの学習、および発話に対応するクラスの決定を行うネットワーク Hidden Mode Bayesian Classifer: **HMBC(a)** を示す。図において離散隠れ確率変数 **Class** が発話のクラスを表現する。クラス数を N とするとき、**Class** がとり得る値は 0 から N-1 の N 通りであり、各値がクラスのインデックスである。ノード **Observation** は音響特徴量ベクトルを表す連続確率変数に対応し、離散値をとるノード **Class** と **Mixure** を親として持つ。連続確率変数の確率密度関数は離散値をとる親ノードの値の組み合わせごとのガウス分布とした。ノード **Observation** は **Class** を親として持つことから、音響特徴量ベクトルの観測確率は発話クラスに依存したものとなる。また、**Observation** はノード **Mixure** も親として持ち、全体としては発話クラスに応じた混合ガウス分布により、音響観測確率のモデル化を行っている。離散確率変数 **Mixure** の離散確率分布が、混合ガウス分布の混合重みに相当する。ノード **Observation** および **Mixure** は、発話のフレーム数に応じて複数回用いるが、発話のクラスを表す確率変数 **Class** は各発話に対してただ 1 つのみ用いる。

クラスの学習は、学習セット中の各発話を対象に、ネットワーク全体の尤度が最大となるように EM アルゴリズムを用いて各ノードのパラメタを求めることにより行う。各発話のクラスへの分類は、発話が与えられた条件でネットワークの尤度を最大とするよう、隠れ変数 **Class** へ値を割り当てることにより行う。なお、最尤割り当ての代わりに **Class** の各値の事後確率分布を分類結果とすることで、発話とクラスの対応関係が多対多となるようなクラスタリングを行うことも出来る。

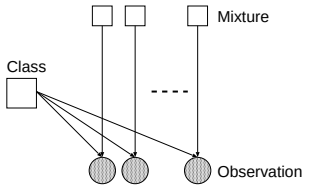

図 1. HMBC (a) using a set of Gaussians for the Observation node.

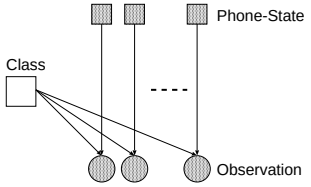

図 2. HMBC (b) using phone state information.

ベイジアンネットを用いた分類器では、音声中の各フレームに対応する音素の種類など、音声の内容に関する情報を組み入れることも可能である。図 2 に音素状態の情報を利用するネットワーク **HMBC(b)** を示す。このネットワークでは図 1 のネットワークと比較してノード **Mixure** が音素状態を表す離散確率変数 **Phone-State** に置き換わっている。音響特徴量の観測確率は発話クラスと音素状態に応じたガウス分布によりモデル化される。発話クラスの学習および学習セット中の発話のクラス分けにおいては、音素状態の情報は HMM を用いた音声と書き起こしの強制アライメントにより求められる。テストセット中の発話に関しては、不特定話者音響モデルを用いた認識仮説を近似的に用いるか、または、確率変数 **Phone-State** を隠れ変数として扱い音響特徴量のみを用いることでクラス分けを行うことが出来る。

図 1 のネットワークではノード **Mixure** は発話クラスと直接の関係は無く、混合重みは発話クラスに独立である。図 3 **HMBC(c)** に示すように、ノード **Class** から **Mixure** へ依存関係を表すリンクを追加することで、ガウシアンコンポーネントに加えて混合重みも発話クラスに応じた値とすることが出来る。連続確率変数の確率密度関数に、離散値をとる親ノードの値ごとの混合ガウス分布を仮定する場合、この分類器は図 4 のように実現することも出来る。

3　クラス分割および除去

発話クラスの学習において、クラス数は予め任意に設定することが出来る。しかし EM 学習による局所的な最大値への収束を抑制する目的で、少数のクラス数から出発し、順次クラス数を増加させる方法をとった。また学習過程において、選択される確率が極端に小さくなったクラスについては除去を行った。クラスの分割および除去の手順は以下の通りである。

∗ Hidden mode Bayesian classifier for acoustic modeling

By Takahiro Shinozaki and Sadaoki Furui (Tokyo Institute of Technology)

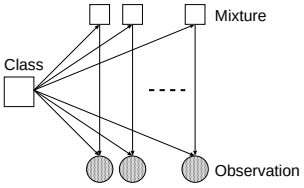

図 3. HMBC (c) using mixture weights depending on classes.

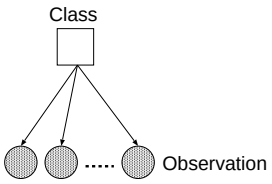

図 4. HMBC (c') equal to classifier (c) but having simpler network structure by using a set of Gaussian mixtures for the Observation node.

(1) 離散確率変数 **Class** の要素数を 2 にセット。

(2) ネットワークのパラメタを EM 学習を 5 回行い推定。クラス数が十分であれば終了。

(3) クラスの要素数を N とするとき、クラスの事前確率が $1.2/N$ より大きければクラスを分割、$1/3N$ より小さければクラスを除去。ステップ 2 へ。

クラスの分割・除去を行ってもネットワークの構造は不変であり、影響を受けるのはクラスを表す離散確率変数 **Class** および **Class** の子ノードの確率分布である。クラスの分割を行った場合、元の事前クラス確率の半分の確率を持つ 2 つのクラスが作成され、**Class** の要素数が 1 つ増加する。音響観測ノード **Observation** においては、分割前のクラスに対応するガウス分布を 2 つにコピーした後、パラメタに小さなランダム値を付加して用いる。クラスの除去では **Class** の要素数が 1 つ減少する。対応するパラメタの除去を行った後、必要に応じて確率分布の正規化を行う。

4 実験結果

発話分類器および音響モデルの学習セットとして、日本語話し言葉コーパス CSJ よりランダムに抽出した 40 時間の発話を用いた。抽出された音声は、学会講演・摸擬講演および男性・女性の音声を含む。提案ネットワークを用いてクラスの学習を行った後、学習セットの発話の分類を行った。離散変数 **Mixture** および **Phone-State** の要素数は 126 とした。学習セット全体を使用して学習した音響モデルを元に、各クラスの音声を用いて MAP 推定を行うことにより、クラスに対応する音響モデルを作成した。音響モデルは予備実験より、1000 状態 32 混合のトライフォンを用いた。

テストセットには、CSJ において選択された testset 1,2,3 の 30 講演 [3] 全ての発話を用いた。提案ネットワークを用いて発話のクラス分けを行った後、各クラスに対応する音響モデルおよび CSJ より学習した 3 万語彙のトライグラムを用いて認識を行った。音素状態情報を用いた分類器 **HMBC(b)** を用いる場合、テストセット発話の分類では、確率変数 **Phone-State** を隠れ変数として扱った。分類器のネットワークは Dynamic Bayesian Network (DBN) として実装し、パラメタ学習および発話の分類は GMTK [4] を用いて行った。音響モデルの学習および MAP 推定には HTK を用いた。

表 1. Word accuracy (%)

	#of classes	Average Acc
GID	1	69.8
GD	2	70.2
HMBC(a)	2	70.3
HMBC(b)	2	70.2
HMBC(c)	2	70.2
HMBC(a)	4	70.7
HMBC(b)	4	70.3
HMBC(c)	4	70.6
HMBC(a)	9	70.2
HMBC(b)	9	70.7
HMBC(c)	9	70.6

表 1 に各分類器を用いた場合の単語認識率を示す。クラス数は 2,4 および 9 である。**HMBC(a)** および **HMBC(c)** ではクラスの分割・除去を 6 回、**HMBC(b)** では 7 回行った時点でクラス数が 9 となった。表では学習セット全体を用いて学習したモデル **GID** 、および性別によるクラス分けに基づき MAP 推定を行ったモデル **GD** を用いた結果も合わせて示した。提案分類器を用いた場合、性別非依存モデル **GID** と比較して、クラス数を 2 としたときでは性別依存モデル **GD** を用いた場合と同程度の効果となった。クラス数を 4 または 9 とした場合には **GD** モデルを用いた場合より高い認識率が得られた。分類器 **HMBC(b)** および **HMBC(c)** ではクラス数が 9 のとき認識率が最大となったが、**HMBC(a)** ではクラス数を 9 に増やすと認識率が低下した。認識率の最大値については、3 種類の提案分類器の間で大きな差はみられなかった。性別非依存モデル **GID** と比較して最大で 1.0% の単語正解精度の改善がみられた。

5 まとめ

ベイジアンネットを用い、隠れ変数を発話クラスに対応させた、尤度を基準とした音声の分類器の提案を行った。提案分類器では音響情報のみを用いてクラスタリングを行うことに加え、音素情報などを組み込むことも可能である。CSJ の講演音声を用いた大語彙連続音声認識実験において本手法が有効であることを示した。今後の課題としては音素以外に講演種別や年齢情報を組み込むなどのネットワークの改良、他のクラスタリング手法との比較、教師なし適応化と組み合わせた実験などが挙げられる。

参 考 文 献

[1] 加藤他、“多数話者電話音声データベースを用いた話者クラスタリング,” SP-2000-10, 2000, pp. 1–8.

[2] 張他、“尤度最大化規準による雑音適応,” SLP-40-27, 2002, pp. 157–162.

[3] T. Kawahara et.al., “Benchmark test for speech recognition using the corpus of spontaneous Japanese,” in *Proc. SSPR*, 2003, pp. 135–138.

[4] J. Bilmes et.al., “The graphical models toolkit: An open source software system for speech and time-series processing,” in *Proc. ICASSP*, 2002, vol. 4, pp. 3916–3919.