論文 / 著書情報
Article / Book Information

論題(和文)	弁別素性のグラフィカルモデリングによる音声認識
Title(English)	
著者(和文)	小林隆二, 篠田浩一, 古井貞熙
Authors(English)	Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2005年春季講演論文集, Vol. , No. 1-5-21, pp. 41-42
Citation(English)	, Vol. , No. 1-5-21, pp. 41-42
発行日 / Pub. date	2005, 3

弁別素性のグラフィカルモデリングによる音声認識*

◎小林 隆二　篠田 浩一　古井 貞熙 (東工大)

1 はじめに

音声認識の分野では、音響モデルとして HMM が広く用いられている。しかし、HMM を用いた音声認識は、雑音を含んだ音声や、会話などの話し言葉の認識において、その性能が著しく低下するという問題を抱えている。そこで近年、HMM のように一つの隠れ変数を用いて音声をモデル化するのではなく、音素の弁別素性を表す複数の隠れ変数を用いてモデル化する手法が提案されている [1]。また、そのようなモデルを実装するために、HMM よりも柔軟なフレームワークとしてダイナミックベイジアンネットワークが研究されている。

本研究では、従来の HMM に弁別素性を表す隠れ変数を導入し、弁別素性の組によって混合ガウス分布の重みを調節するモデルを提案する。弁別素性の値は音素名と、1 フレーム前の自身の値に依存して確率的に決めることとし、緩やかに変化するという弁別素性の特徴を反映させる。

2 グラフィカルモデリング

グラフィカルモデリング (GM) とは、確率変数間の依存関係をグラフによって表現する手法の総称である。GM の中で、確率変数間の条件付き独立性を有向非循環グラフによって表したものをベイジアンネットワーク (BN) という。BN ではグラフのノードが確率変数を表し、エッジが確率変数間の依存関係を表す。

例えば、確率変数 $X_1, X_2, \cdots, X_N$ の条件つき独立性を表したグラフにおいて、変数 X_i の親ノードの集合を π_{X_i} で表したとき、同時分布 $p(X_1, X_2, \cdots, X_N)$ は

$$p(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{N} p(X_i|\pi_{X_i}) \qquad (1)$$

となる。時系列データを対象とし、BN を時間軸方向に拡張したものをダイナミックベイジアンネットワーク (DBN) といい、本研究ではこれを用いて音響モデルを実装する。

3 弁別素性を用いた音響モデル

3.1 提案モデル

提案する DBN の構造を図 1 に示す。P は現在フレームの音素、$A_1, A_2, \cdots, A_N$ が弁別素性を表し、pS は HMM の状態番号を表す。そして、O は観測される特徴量ベクトルを表す。$P, A_1, A_2, \cdots, A_N, pS$ は離散確率変数で、O は連続確率変数である。弁別素性は、現在の音素と 1 フレーム前の自身の値に依存して決まる。これにより、弁別素性が緩やかに、そして非同期的に変化するために起こる調音結合をモデル化できると考えられる。O の分布として、$A_1, A_2, \cdots, A_N$ と pS の値の組ごとに異なる混合ガウス分布が用意される。以降では $A_1, A_2, \cdots, A_N, pS$ の値の組を「状態」と呼ぶことにする。弁別素性の値が音素によって決定的に決まるようにすると、この DBN は HMM と等価になる。本研究で使用した弁別素性を表 1 に示す。

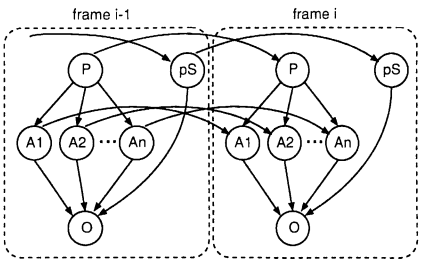

図 1. 弁別素性を用いたモデル

表 1. 使用した弁別素性

弁別素性	値
前/後舌性	後舌, 前舌
高/低舌性	低舌, 中舌, 高舌
音源	無声, 有声
調音位置	nil, 口唇, 歯茎, 口蓋, 声門
調音方式	nil, 摩擦音, 破擦音, 破裂音, 半母音, 鼻音
拗音	off, on

※ nil は母音を表す

3.2 弁別素性に対する制約

調音機構の物理的特性からくる制約により、弁別素性は任意の値をとれるわけではなく、その間の遷移確率には制限が加えられる。まず、弁別素性の値は緩やかに変化するという仮定から、高/低舌性の値については、低舌と高舌の間の遷移は必ず中舌を経由することとし、両者間の直接の遷移は起こらないものとする。次に、音素によって自由に動ける弁別素性を制限する。本研究で使用した弁別素性では、母音を区別するためには舌の位置が使われ、子音を区別するためにはそれ以外の弁別素性が用いられる。そこで、母音と子音について、音素を区別するために使われていない弁別素性のみを自由に動かせるように設定する。つまり、母音では舌の位置が決定的に決まり、音源、調音位置、調音方式、拗音が確率的に決まるようにし、子音では逆に音源、調音位置、調音方式、拗音が決定的に決まり、舌の位置が確率的に決まるようにする。さらに、母音と子音を区別できるようにするため、調音位置か調音方式の少なくとも一つは決定的に決まるようにする。これにより、各状態に対して一つの音素が対応するようになる。

各音素について、自由に動ける弁別素性の分だけ新しく状態が増えることになる。この増えた状態に対して新しく混合ガウス分布を追加すると、パラメータの数が著しく増大してしまう。そこで、対応する音素が同じである状態同士で混合成分を共有し、混合重みのみを別にもつという手法を用いる。このような混合重みのみを調節する手法としては、発話速度に応じて混合重みを調節するモデルが提案されている [2]。

---

* Speech recognition with graphical modeling of articulatory features

By Ryuji Kobayashi, Koichi Shinoda, and Sadaoki Furui (Tokyo Institute of Technology)

表 2. 弁別素性を一つ用いたときの音素正解精度 (%) (Baseline は 84.64 %)

弁別素性	前/後舌性	高/低舌性	音源
正解精度	84.94	84.97	84.74
弁別素性	調音位置	調音方式	拗音
正解精度	84.79	84.73	84.70

3.3 DBN の初期化

弁別素性の値は、実際の調音機構の動きを反映していることが望ましい。しかしながら、調音機構の動きを直接知ることはできないため、学習時に値を与えて学習することはできない。そこで、調音機構の動きを反映するように、混合ガウス分布や弁別素性の遷移確率の初期値を設定する。

混合ガウス分布は、予め HMM で学習を行い、各音素の混合ガウス分布を対応する状態の初期混合ガウス分布として用いる。弁別素性の値は緩やかに変化するという仮定を反映させるため、遷移確率の初期値は 1 つ前のフレームと同じ値である確率を 0.9 とし、別の値に遷移する確率は残りの 0.1 を等確率で分けあうように設定した。これにより、無声子音の後の母音が無声化するといった現象を反映できると考えられる。

4 評価実験

4.1 音声データ

DBN の学習と評価には ATR の孤立単語データベースを用いた。男性話者は MAU, MMS, MMY の 3 名、女性話者は FKS, FSU, FYN の 3 名で、特定話者で連続音素認識を行った。話者によって書き起こしの異なる単語は除外し、学習データとして 2806 単語、評価データとして 2800 単語を用いた。音声データのサンプリング周波数は 16 kHz である。音響特徴量は 0 次元を除いた MFCC 12 次元と、その差分 Δ と、パワー差分 ΔE の計 25 次元である。フレームシフトは 10 ms、フレーム幅は 25 ms とした。

学習データには予め SN 比 30 dB の白色雑音を付加した。また、評価データには SN 比 10, 15, 20, 30 dB の白色雑音を付加させ、それぞれについて実験を行った。

4.2 実験方法

DBN の学習と認識実験には GMTK [3] を用いた。混合ガウス分布の混合数は 4 とし、あらかじめ HTK [4] を用いて学習したものを初期値として用いた。条件を等しくするため、ベースラインとなる HMM も DBN でエミュレートし、提案 DBN と同様の条件で学習した。認識単位は monophone とし、 HMM は 3 状態の left-to-right HMM である。

弁別素性は、表 1 に示したもの全てを使うのではなく、性能を改善するものだけを使うのが望ましい。そのために、まず弁別素性を一つだけ使うモデルから始めて、SN 比 30 dB のテストセットで認識を行ったときの音素を単語とみなしたときの単語正解精度 (以降音素正解精度と呼ぶ) の話者平均が高いものを順に加えていくという方法を採った。

4.3 実験結果

弁別素性を一つだけ使った場合の音素正解精度を表 2 に示す。また、弁別素性を一つずつ加えて行ったときの音素正解精度の変化を図 2 に示す。最大で 0.9 % 音素正解精度が向上している。この結果から、調音方式以外の全ての弁別素性を使うことにした。その DBN を用いて話者と SN 比を変えて認識実験を行ったときの結果を表 3 に示す。 この表から、提案 DBN はベースラインと比べて 0.38 % から 3.18 % 高い音素正解精度を示していることがわかる。音

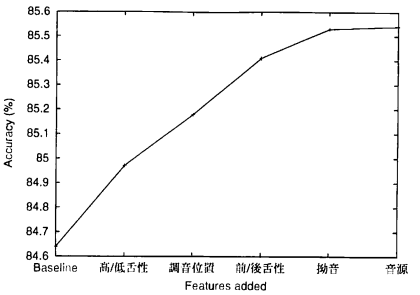

図 2. 追加した弁別素性と音素正解精度

表 3. 話者と SNR (dB) を変えたときの音素正解精度 (%) (BASE: Baseline, DBN: 提案 DBN)

話者	FKS		FSU		FYN	
SNR	BASE	DBN	BASE	DBN	BASE	DBN
30	82.92	84.24	84.85	85.90	82.73	83.35
20	69.02	72.20	75.30	77.16	76.55	77.49
15	57.61	60.47	64.72	66.29	65.22	65.98
10	48.11	50.35	54.44	55.99	48.93	49.99
話者	MAU		MMS		MMY	
SNR	BASE	DBN	BASE	DBN	BASE	DBN
30	85.87	86.64	85.90	86.89	85.57	86.20
20	79.02	79.84	74.96	75.98	76.76	78.11
15	69.73	70.46	64.32	65.03	66.67	68.37
10	57.88	58.41	54.46	54.84	55.30	56.50

素正解精度の改善率は話者によって差が見られるが、いずれの話者、ならびに SN 比においても音素正解精度はベースラインよりも優れていることがわかる。また、提案 DBN ではベースラインと比べて挿入誤りと置換誤りが大幅に減少し、削除誤りが若干増加するという傾向が見られた。

5 まとめ

弁別素性を導入した新たなモデルを提案し、ベイジアンネットを用いて実装し、ATR の孤立単語データベースを使った認識実験によってその有効性を確認した。提案モデルを用いることによって、パラメータ数の増加を抑えたまま性能を改善できることを確認した。

今後の課題としては、どの弁別素性が効果的かをテストセットを用いずに調べる方法の検討、それぞれの弁別素性の識別に有効な特徴量の検討、triphone との比較検討などが挙げられる。

謝辞

本研究は文部科学省科学研究費補助金萌芽研究 No.15650028 によるものである。また、21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の援助を受けた。

参考文献

[1] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," *Proc. of Eurospeech*, Geneva, pp. 2529-2532 (2003).

[2] 篠崎 隆宏, 古井 貞熙, "発話速度変動を考慮した隠れモード HMM による音声のモデル化," 電子情報通信学会技術研究報告, SP2003-41, pp. 37-42 (2003-6).

[3] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," *ICASSP*, Orland, pp. 3916-3919 (2002).

[4] S. Young et. al., "The HTK Book," Entropic Labs and Cambridge University, 1995-2002.