# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 話し言葉音声認識へのベイジアンネットの適用 |
| Title | |
| 著者(和文) | 篠崎 隆宏, 古井 貞熙 |
| Author | Takahiro Shinozaki, SADAOKI FURUI |
| 出典(和文) | 国立国語研究所公開研究発表会 「話し言葉のデータベース ―『日本語話し言葉コーパス』―」 講演予稿集, Vol. , No. , pp. 47-48 |
| Journal/Book name | , Vol. , No. , pp. 47-48 |
| 発行日 / Issue date | 2003, 12 |

# 話し言葉音声認識へのベイジアンネットの適用

篠崎 隆宏† 古井 貞熙†
†東京工業大学情報理工学研究科計算工学専攻

## 1 はじめに

従来、音声のモデル化には一般に HMM が用いられて来た。HMM は状態を持ち、音声の特性を状態遷移確率および状態に応じた音響特徴量の観測確率で表すが、状態を複数の要因に分解してモデル化出来ないなど構造上の制限が存在する。しかし、例えば発話速度やイントネーションなどは認識に直接必要な音響特徴量とは異なる時間尺度で変動し、かつ音響特徴に影響を及ぼすことから、音素状態とは別にモデル化を行うことが望ましいと考えられる。これに対し、ベイジアンネットは HMM を含め様々な確率モデルを記述できる柔軟性があり、また、ベイジアンネットとして表現された確率モデルのための一連の学習/推論アルゴリズムが開発されている。このため、ベイジアンネットの枠組を用いることで、様々なモデルを容易に実現することが可能になる。本研究では従来の HMM が持つ音素状態と独立して、発話速度を表す状態を持つ新しい確率モデルの提案を行い、ベイジアンネットを用いて実現する。

## 2 ベイジアンネットによる音響モデル

デコーダによる音声認識誤りと関係の深い要因として発話速度が挙げられる。特に話し言葉音声では発話速度の変動が大きいことから、認識率向上のためには明示的なモデル化が重要であると考えられる。本節ではベイジアンネットについて簡単に復習した後、従来 HMM および発話速度変動に対応する提案モデルの、ベイジアンネットを用いた構成法について示す。

### 2.1 ベイジアンネット

ベイジアンネットは確率変数間の依存関係を有向グラフにより表したものであり、グラフ構造および各ノードに対応する条件付確率分布により定義される。図 1 にベイジアンネットの例として、CSJ からランダムに講演を 1 つ取りだしたときに観察される講演属性に関するモデル構造を示す。対象としている確率変数は講演種別 (P)、性別 (G)、F0 の高低 (F) である。確率変数 P がとり得る値は "学会講演" および "模擬講演"、G は "男性" および "女性"、F は "高い" および "低い" である。このモデルは性別の確率分布が講演種別に依存すること、F0 の高低が性別を経由して間接的に講演種別に依存していることなどを表現している。各確率変数の条件付確率分布は表 1 に示す条件付確率テーブル (CPT) により表すことが出来る。これらモデルのグラフ構造および確率分布のパラメタをもとに、例えば F0 が高いときに学会講演である確率などを計算することが出来る。ベイジアンネットでは有向グラフとして表された確率モデル一般に対して、効率的に様々な確率推論を行うアルゴリズムが開発されている [1]。

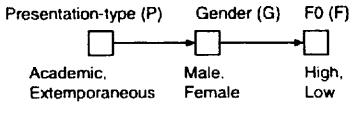


図 1: Example of a Bayesian network.

### 2.2 ベースラインモデル

ベースラインとして使用する音響モデルは図 2 に示す left-to-right 型の音素 HMM であり、ベイジアンネットとして実現する。図 3 に音素 HMM 系列をエンコードしたネットワークを示す [2]。ネットワークは音声のフレーム数に合わせて、一定の部分ネットワークを繰り返す構造をしている。図 4 (a) **BASE** に基本構造の一部を抜き出して示す。このベイジアンネットには音素の種類を表すノード **Phone** および音素 HMM 内での状態位置を示すノード **Phone-State** があり、音響特徴量の観測確率を表すノード **Observation** と HMM の状態遷移を表すノード **Phone-State-Transition** の親となっている。これは HMM において音響観測確率密度関数および状態遷移確率分布が音素の種類および音素内での状態位置に依存していることに対応している。**Observation** の条件付確率密度関数は混合ガウス分布集合により定義した。

表 1: Example of Conditional Probability Tables (CPTs)

(Artificial data)

| P | Academic | Extemporaneous |
|---|---|---|
| | 0.4 | 0.6 |

| G | Male | Female |
|---|---|---|
| P=Academic | 0.8 | 0.2 |
| P=Extemporaneous | 0.4 | 0.6 |

| F | High | Low |
|---|---|---|
| G=Male | 0.3 | 0.7 |
| G=Female | 0.8 | 0.2 |

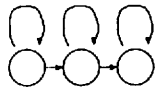

図 2: Baseline phone HMM.

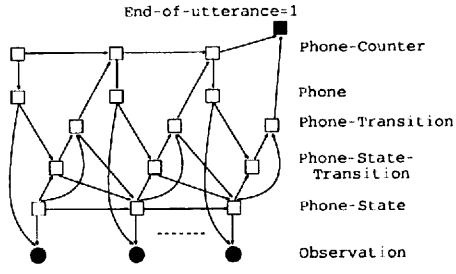


図 3: DBN representation of a phone HMM sequence.

### 2.3 隠れモード HMM による発話速度変動のモデル化

発話速度の変動が認識性能に悪影響を与える理由としては、調音結合の度合などスペクトルの変動や各音素継続時間の不均一な伸び縮みなどが考えられる。

発話速度に依存したスペクトルの変動を明示的にモデル化するため、図 4 (b) に示す隠れモード混合重みモデルの提案を行う。提案モデルでは、発話速度の状態に対応させた隠れ変数の値をもとに HMM の混合重みを調節する。このネットワークでは HMM と比較して、2 つのノードが追加されている。**Mode** は発話速度の状態を示す離散隠れ確率変数、**Speaking-Rate** は発話速度を示す連続確率変数である。**Speaking-Rate** の条件付確率密度関数はガウス分布集合により定義される。**Observation** の条件付確率密度関数は親ノードである **Phone**, **Phone-State**, および **Mode** の値の組み合わせ毎の混合ガウス分布となる。パラメタの大幅な増大を防ぐために、異なる **Mode** の値に対しては要素となるガウス分布を共有とした。すなわち、異なる **Mode** の値は異る混合重みを指定することになる。このモデルを **HM-MW** とする [3]。

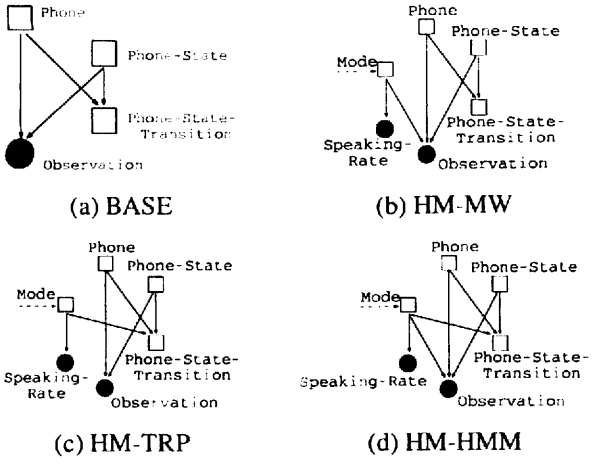


図 4: Portion of a time slice of each DBN-based acoustic model.

発話速度に依存した継続時間の変動に対応するため、図 4 (c) に示す隠れモード遷移確率モデルを提案する。このモデルでは、隠れ変数の値をもとに HMM の遷移確率を制御する。離散隠れ確率変数 **Mode**、連続確率変数 **Speaking-Rate** の定式化は **HM-MW** と同じである。このモデルを **HM-TRP** とする。

発話速度に依存したスペクトルおよび継続時間の変動に対応する隠れモード HMM を図 4 (d) に提案する。このモデルは隠れモード混合重みモデルおよび隠れモード遷移確率モデルを統合したモデルと見ることが出来る。このモデルを **HM-HMM** とする。

## 3 発話速度

提案手法では音響モデルの学習および認識処理において通常の音響特徴量に加え、発話速度を使用する。発話速度の計測法として、書き起こしテキストを用いた強制アライメント、最尤認識仮説用いた強制アライメント、および Enrate [4] を用いた。強制アライメントによる計測では HMM の状態滞留時間の逆数を求め、400ms の窓幅で平滑化した値として発話速度を求めた。Enrate 法は音声の書き起こしを仮定せずに、信号処理的手法により発話速度の推定を行う手法である。 Enrate を計算する際の窓幅は 400ms とした。

## 4 コーパスおよび実験条件

モデルの学習/評価には日本語話し言葉コーパス (CSJ) [5] を使用した。実験では男性話者 10 名による学会講演音声 116 分を用いて話者独立モデルを作成し、別の男性 5 名による学会講演から抽出した 16 分の音声を用いて評価を行った。

ベイジアンネットのパラメタ学習においては、先ず HTK を用いて作成した monophone HMM のパラメタを用いて音響観測確率および遷移確率の初期化を行い、次いで GMTK [6] を用いて EM/GEM 学習を 10 回繰りかえすことにより行った。ベースとする monophone HMM の混合数は予備実験よりテストセットの認識率を最大とする 28 混合に決定した。音響特徴量は窓幅 25msec、フレームシフト 10ms の MFCC、Δ MFCC および Δ エナジーの計 25 次元である。提案モデルにおいてはこの他に発話速度を用いる。

認識実験は monophone HMM を用いて HTK により作成した 100-best をリスコアすることにより行った。言語モデルには CSJ の講演約 6.7M 形態素より学習した、語彙サイズ 3 万の bigram を使用した。

## 5 実験結果

理想的な発話速度推定値を用いた場合の実験として、モデルの学習および評価双方に正解音素系列を用いた場合の結果を図 5 に示す。離散変数 **Mode** の要素数は 4 とした。HMM と比較して全ての提案モデルで高い認識率が得られている。提案手法の中では、混合重みと遷移確率両方の制御を行う **HM-HMM** の認識率が最大となった。

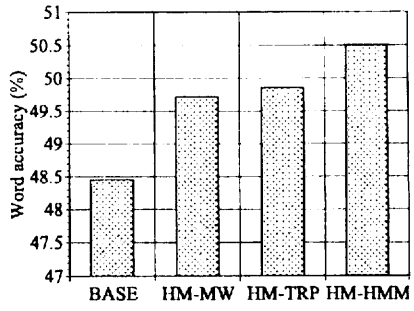


図 5: Recognition result using oracle speaking rate.

表 2: Recognition results using estimated speaking rate

| | HM-HMM(3) | HM-HMM(4) |
|---|---|---|
| **BASE** | 48.5 | |
| **HYP** | 49.3 | 49.7 |
| **ENRATE** | 48.8 | 48.7 |
| **ORACLE** | 50.0 | 50.5 |

認識時に正解音素系列を仮定しない場合の実験として、正解発話速度をモデルの学習時のみ使用し、認識時には認識仮説から求めた発話速度を使用した場合、および学習および認識双方に Enrate を用いた場合の結果を表 2 に示す。表で **HYP** は認識仮説を用いた場合、**ENRATE** は Enrate を用いた場合である。発話速度を使用しない従来 HMM を使用した **BASE**、および正解発話速度を使用した **ORACLE** の結果も合わせて示した。離散変数 **Mode** の要素数は 3 および 4 を試みた。提案手法では HMM と比較してどの条件においても高い認識率が得られた。**Enrate** と比較して **HYP** の方が認識率の向上が大きく最大で 1.3%の改善が得られた。

## 6 まとめ

ベイジアンネットを用いた認識システムは、音声認識に特化した従来システムと比べると計算量が多い欠点があるものの、様々な新しい確率モデルのプロトタイピングに適している。発話速度の状態を表す隠れ変数を持つ音響モデルの提案を行い、ベイジアンネットを用いて評価を行った。提案モデルは従来モデルと比較して高い性能を示した。今後の課題としては発話速度推定法の改良や、より大規模な認識対象に対応するための実装法の検討などが挙げられる。

## 参 考 文 献

1) K. Murphy: "A brief introduction to graphical models and Bayesian networks", http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html [Online], (1998).

2) G. Zweig: "Speech recognition with dynamic Bayesian networks", PhD thesis, University of California, Berkeley (1998).

3) T. Shinozaki and S. Furui: "Time adjustable mixture weights for speaking rate fluctuation", Proc. EUROSPEECH, Vol. 2, pp. 973–976 (2003).

4) N. Morgan, E. Fosler and N. Mirghafori: "Speech recognition using on-line estimation of speaking rate", Proc. Eurospeech, Vol. 4, pp. 2079–2082 (1997).

5) S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira: "Toward the realization of spontaneous speech recognition", Proc. ICSLP, Vol. 3, pp. 518–521 (2000).

6) J. Bilmes and G. Zweig: "The graphical models toolkit: An open source software system for speech and time-series processing", Proc. ICASSP, Vol. 4, pp. 3916–3919 (2002).