音声言語情報処理 14−11
(1996. 12. 13)

情報量基準を用いた状態クラスタリングによる音響モデルの作成

篠田 浩一　　渡辺 隆夫

NEC 情報メディア研究所
〒216 川崎市 宮前区 宮崎 4-1-1
044-856-2304
{shinoda,watanabe}@hum.cl.nec.co.jp

あらまし:
近年、隠れマルコフモデル (HMM) を用いた大語彙音声認識システムにおいて、コンテキスト依存サブワード単位がしばしば用いられてきた。その場合すべての認識単位のパラメータを十分な精度で学習するためには、一般に学習データ量が不足しているため、これらのシステムのほとんどは、モデルの自由度を下げるために様々な方法でパラメータのクラスタリングを行なっている。しかしながら、これらのクラスタリングの手法は停止基準を内包していなかった。本稿では、情報量基準の１つである MDL 基準を停止基準として用いる方法を提案する。評価実験の結果、提案法は少ない計算量で従来の発見的な方法と同等以上の性能をもつことが明らかになった。

キーワード:
音声認識、隠れマルコフモデル、認識単位、情報量基準、MDL 基準

Acoustic Model Generation Using State Clustering by Infomation Criterion

Koichi SHINODA, Takao WATANABE

NEC Information Technology Research Laboratories
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, 216 JAPAN
044-856-2304

Abstract:
In recent years, context-dependent subword units have been often employed for large-vocabulary speech recognition systems using HMMs. Since the amount of training data available is usually not large enough to estimate the parameters of all the units with sufficient accuracy, most of these systems use clustering methods to reduce the degree of freedom. However, none of these clustering methods have given a stop criterion of the clustering. In this paper, we propose one clustering method which use MDL criterion as the stop criterion. The evaluation experiment proved that the proposed method achieved as high recognition accuracy as conventional heuristic methods with less computational amount.

key words:
speech recognition, Hidden Markov Model, recognition unit, information criteria, Minimum Descritption Length Principle

1 はじめに

現在、音声認識においては、隠れマルコフモデル (Hidden Markov Model; HMM) が広く用いられている。音声の音響的特徴は確率モデルのパラメータとして保持され、それらは安定して局所解を与える Baum-Welch アルゴリズムと呼ばれる学習アルゴリズムで推定される。十分な学習データを与えれば、高い性能を示すことが知られている。さて、認識対象の規模がある程度以上大きい場合、認識対象の発声 (単語あるいは文など) それぞれに対し別々に HMM を用意するのは、必要な学習データ量が多くなり、現実的ではない。そこで、一般にこのような大語彙の認識においては、音素などのサブワード単位ごとに HMM を用意し、認識時にはそれを連結して用いることで 1 つの HMM 当りの学習データ量を十分に確保している。以下、サブワードとして音素を例にとって説明する。

一般に音素の音響的特徴は、そのコンテキストにより大きく変化することが知られている。ここでコンテキストとは、先行音素、後続音素など、その音素の環境一般を指す。そのため、これらコンテキストの違いを無視して学習するよりも、コンテキストの違いにより音素をいくつかの単位に分けた方が認識性能が高くなると考えられる。そこでコンテキストに独立に HMM を用意するコンテキスト独立認識単位のかわりに、コンテキストを考慮したコンテキスト依存認識単位が多く用いられ、コンテキスト独立単位よりも良好な性能を得ている。

もし、十分な量の学習データがあれば、考慮するコンテキストの種類が多いほど、すなわち、認識単位の種類が多いほど、認識性能が高くなる。しかし、現実には学習データの量は有限であるため、コンテキスト依存単位の種類を増やすと、次の 2 つの問題が起こる。まず第 1 に学習データ中に現れないコンテキストに対応する認識単位が学習されない。これら未知のコンテキストに対するモデルは作成することができない。第 2 に、考慮するコンテキストの種類が増加するに従い、各コンテキストに対応する学習データの量が減少し、したがって、各認識単位のパラメータの推定精度が低くなる。これらの問題に対しては、従来、コンテキスト依存単位に対しクラスタリングを行なう、あるいは、コンテキスト独立単位を分割するなどの手段で、認識単位の種類数を調節する方法 [1-7] が用いられてきた。

これらは、分割の対象となる単位として、認識単位、HMM の各状態、各状態の出力確率分布、などがあり、また、分割あるいはクラスタリングする単位の選択の基準としては、尤度、エントロピーなどが用いられてきた。

しかしながら、これら従来法に共通した欠点として、認識単位の統合・分割の手続きの停止基準を内包していない点があげられる。例えばデータに対する尤度が大きくなる方向にコンテキスト独立単位の分割を進める方法 (尤度最大基準を用いる状態分割) では、当然、認識単位数を増加させるほど尤度が増加するため、そのままでは、認識単位が極限まで細分化されてしまう。認識単位あたりのデータ量は少なくなり、従ってパラメータの推定精度が劣化する。この問題の回避のために、多くの方法は、認識単位や状態などの種類数に対する閾値を設定し、その閾値に達した時点で分割をやめるという手続きをとっている。今のところこの閾値を定性的に決定する有効な方法はなく、閾値は、テストデータに対する認識実験の繰り返し、あるいは、学習データを分割し一部をテストデータとする方法 (クロスバリデーション法) を用いて最適化される。しかし、これらの方法では、すべての閾値に対し実験をすることは不可能であり、また、最適な値を求めようとするとテストデータ量、計算量がより多く必要になるという欠点がある。

本稿では、この問題を解決する一手法として、情報量基準の一つである MDL(Minimum Description Length) 基準を認識単位の分割・統合基準、分割・統合の停止基準として用いる方法を提案する。

次章で簡単に MDL 基準を紹介し、第 3 章で MDL 基準を用いてコンテキスト依存単位を自動的に生成する方法を説明する。最後に評価結果を述べる。

2 MDL 基準

MDL 基準 [8, 9] は情報量基準の 1 つであり、情報理論における情報源符号化の研究から生まれた。与えられたデータに対し最適なモデルを選択する問題において有効であることが知られている。AIC(Akaike Information Criterion; 赤池情報量

基準) などと同様、なるべく簡単で、しかも、与えられたデータをよく表現できるモデルが良いモデルである、という理念を具現化した基準の一つである。

MDL基準は、確率モデル $i = 1,\ldots,I$ の中で、データ $x^N = x_1,\ldots,x_N$ に対し、最も小さい記述長を与えるモデルを最適なモデルとする基準である。ここで、確率モデル i に対する記述長 $l_{MDL}(i)$ は以下の式で与えられる。

$$l_{MDL}(i) = -\log P_{\hat{\theta}(i)}(x^N) + \frac{\alpha_i}{2}\log N + \log I. \quad (1)$$

ここで、α_i はモデル i の次元数 (自由パラメータの個数) 、$\hat{\theta}^{(i)}$ はデータ X^N を用いて推定されたモデル i の自由パラメータ $\theta^{(i)} = (\theta_1^{(i)},\ldots,\theta_{\alpha_i}^{(i)})$ の最尤推定量である。上式において、第1項はデータに対する対数尤度 (以下、尤度と記す) に負符号を付けた量であり、第2項は、モデルの複雑さを表す量である。第3項は、モデル i を選択するために要する記述長である。モデルがより複雑なほど、データに対する尤度が大きくなり、したがって第1項の値は減少する。一方、モデルが複雑になれば、自由パラメータ数が増加するため、第2項の値は増加する。このように、第1項と第2項の間にはトレードオフの関係があり、記述長 $l_{MDL}(i)$ は適当な複雑さをもつモデルで最小値をとることが期待される。

ここで、従来から良く用いられている情報量基準であるAICとの相違を述べる。AICは、以下の l_{AIC} が最小のモデルを選択する。

$$l_{AIC}(i) = -\log P_{\hat{\theta}(i)}(x^N) + \alpha_i, \quad (2)$$

式(1)と式(2)を比較すると、第1項は同一であり、第2項が、式(1)ではデータ量 N に依存した量であるのに対し、式(1)ではデータ量から独立な量になっている (式(1)の第3項は最小化に無関係な量なので無視する)。データ量の変化に対する選択されたモデルの次数の変動の仕方が異なるが、多くの現実の問題に対しては両者の選択するモデルの次数に大きな相違はないという報告がある。

3 MDL基準を用いた認識単位の自動生成

3.1 音素決定木を用いた状態クラスタリング

本稿では、音素を基本単位とし、コンテキスト独立音素HMMの各状態を音素コンテキストの違いによりトップダウンに分割する枠組を採用する (e.g. [6, 7])。分割するときの質問 (分割条件) としては、先行音素及び後続音素の素性情報を用いる。例としては、先行音素が無声音であるかどうか (L-unvoiced ?)、あるいは、後続音素が摩擦音かどうか (R-fricative ?)、などがある。ここで、Lは先行音素、Rは後続音素を表す。

まず、コンテキスト独立音素HMMの各状態に対し、予め用意された複数の分割条件から1つの条件を選択し、状態を分割する。そして、分割された状態に対しさらに同様の分割処理を行なう。この手続きを繰り返すことにより状態を細分化していく。分割された結果は分割された状態をノードとする2分木 (音素決定木) で表すことができる (図1)。最適な分割条件の選択、および、分割を実行するかあるいは停止するかを決定するための基準としてMDL基準を用いる。

この枠組を採用した理由として以下の点が挙げられる。まず、第一にボトムアップの手法と比べると、未知コンテキストの問題によりよく対処できる。第2に、状態ではなく認識単位を分割する方法に比べると、最尤推定量の算出と同時に最尤推定量に対する尤度を計算可能であり計算量が少ない (次節参照)。認識性能の優劣は今回特に比較しなかった。

処理の流れは以下の通りである。

1. コンテキスト独立HMMの学習を行なう。
2. コンテキスト依存HMMをコンテキスト独立HMMのパラメータをコピーすることに

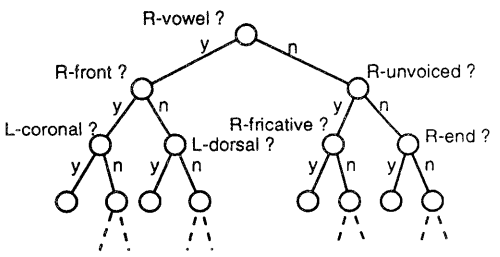

図 1: Phonetic decision tree

より作成する。コンテキスト依存 HMM はコンテキスト独立 HMM と同じ構造をもつ。

3. コンテキスト依存 HMM の学習を行なう。

4. コンテキスト独立 HMM の各状態のパラメータをコンテキスト依存 HMM の対応する状態の共有パラメータとして求める。

5. すべての音素のすべての状態について以下の処理を行なう。

   (a) コンテキスト依存 HMM の状態の集合を、コンテキストに関する質問で、複数のクラスタに分割する。これらのクラスタのパラメータはクラスタに属する状態のパラメータの共有パラメータとして求められる。分割結果は、各クラスタをノードとする音素決定木で表される。

   (b) この音素決定木のリーフノードのパラメータをそのノードに属するコンテキスト依存 HMM の状態にコピーする。

3.2 状態クラスタリングにおける記述長の計算

状態クラスタリングの枠組に対し MDL 基準を適用するために、まずモデルを定義し、その記述長を計算する。

今、ある音素の状態分割の過程で、コンテキスト独立音素 HMM の状態 S_0 に対応するルートノードが、$S_1,\ldots,S_M$ の M 個のリーフノードへと分割されているとする。このとき、この状態 S_0 に対するモデル U として、無相関単一ガウス分布を出力確率分布とする M 個のカテゴリから構成されるモデルを考える。各々のカテゴリは各リーフノードに対応する。このモデル U に対する記述長 $l(U)$ を計算する。

HMM の記述長をそのまま計算することは難しい。そこで、ここでは以下のような近似を行ない、問題を簡単化する。まず、あるノードの分割の前後でセグメンテーションは変わらないと仮定する。また、遷移確率は出現確率に比べ影響が小さいと仮定し尤度計算の際には無視する。そのとき、学習データ $\mathbf{o}_1,\ldots,\mathbf{o}_T$ が与えられたときのノード S_m の尤度 $L(S_m)$ は、

$$\begin{aligned} L(S_m) &= \sum_{t=1}^{T} \log(N(\mathbf{o}_t,\mu_{S_m},\Sigma_{S_m}))\gamma_t(S_m) \\ &= -\frac{1}{2}(\log((2\pi)^K \mid \Sigma_{S_m} \mid) + K)\Gamma(S_m) \end{aligned} \tag{3}$$

と近似できる。ここで、

$$\gamma_t(S_m) = \frac{\alpha_t(S_m)\beta_t(S_m)}{\sum_s \alpha_t(S_m)\beta_t(S_m)} \tag{4}$$

$$\Gamma(S_m) = \sum_{t=1}^{T} \gamma_t(S_m) \tag{5}$$

であり、K は特徴ベクトルの次元数、$N(\mu,\Sigma)$ は、平均が μ、共分散が Σ である正規分布を表す。また、$\alpha_t(S_m)$ は時刻 t、クラスタ S_m における前向き確率 、$\beta_t(S_m)$ は時刻 t、クラスタ S_m における後ろ向き確率、$\gamma_t(S_m)$ は、時刻 t においてクラスタ S_m に存在する確率を表し、その全時刻における和 $\Gamma(S_m)$ は学習データ中にクラスタ S_m が出現する頻度 (出現フレーム数) を表す。これらの S_m に関するパラメータは、クラスタ S_m に属するコンテキスト依存単位の状態のパラメータの共有パラメータとして学習データから推定される。

通常の場合、最尤パラメータに対する尤度を計算する際には、学習データに対する最尤パラメータを求めたのちに、再度その最尤パラメータを用いて学習データに対する尤度を計算する必要があるが、ここでは、尤度を式 (3) のように近似することにより、最尤パラメータに対する尤度 $L(S_m)$ を、最尤パラメータの推定と同時に求めることができる。

この $L(S_m)$ を用いて式 (1) の記述長 $l(U)$ は、以下のように表される。ここでは分割により変化しない項を除いている。

$$\begin{aligned} l(U) &= \frac{1}{2}\sum_{m=1}^{M} \Gamma(S_m)\log(\mid \Sigma(S_m) \mid) \\ &\quad + KM\log\Gamma(S_0), \end{aligned} \tag{6}$$

$$\Gamma(S_0) = \sum_{m=1}^{M} \Gamma(S_m). \tag{7}$$

式 (7) における $\Gamma(S_0)$ はコンテキスト独立 HMM の状態 S_0 の学習データ中の出現フレーム数に相当する量であり、分割の方法によらない。式 (6)

で与えられる記述長 $l(U)$ を最小にするモデル (クラスタの組) が、MDL 基準の意味で最適な分割を表す。

3.3 MDL 基準を用いた状態クラスタリング

音素決定木を用いた状態クラスタリングにおいて前節で求めた記述長を用いる。

ここでは以下の手続きで状態の分割を行なう。分割の過程でさらに分割を進めるかどうかは、以下のように分割前後の記述長の差分を計算して決定する。今、分割の過程で、ある分割条件 q を用いて、あるクラスタ S を S_{q+}、S_{q-} に分割することを考える。このとき、記述長の増分は、同じ決定木における他のクラスタの分割とは独立に計算することが可能である。すなわち、記述長の増分 Δ_q は以下のように表すことができる。

$$\Delta_q = \frac{1}{2}(\Gamma(S_{q+})\log|\Sigma_{S_{q+}}|+\Gamma(S_{q-})\log|\Sigma_{S_{q-}}| -\Gamma(S)\log|\Sigma_S|)+K\log\Gamma(S_0) \quad (8)$$

モデルの記述長を減少させるためには、この増分 Δ が 0 未満の時に分割を行なえばよい。なお、式 (8) より、MDL 基準を用いた分割では、尤度の増分の閾値が $K\log\Gamma(S_0)$ という値に設定され、増分がその閾値を超えた場合に分割が行なわれると見ることもできる。

今回分割に用いたアルゴリズムは以下の通りである。まず、すべての分割条件について、クラスタ S の 2 分割を行ない、Δ_q を最小とする条件 q' を求める。そして、$\Delta_{q'} < 0$、すなわち記述長が減少するならば、2 分割を行ない、$\Delta_{q'} > 0$ ならば分割を行なわない。この分割の手続きをコンテキスト独立音素 HMM の状態を出発点として繰り返すことにより、状態の分割を行なう。なお、この方法では、ルートノードから順にノードを展開していく形になるため、必ずしも最適な分割が行なわれる保証はない。

4 評価実験

評価実験として、日本語 5000 単語認識をシミュレートした類似 100 単語認識実験 [10] を行なった。入力音声は、標本周波数 16kHz、分析周期 10ms、分析窓長 32ms、周波数帯域 0.1–7.2kHz の条件で分析し、特徴量として、メルケプストラム 10 次元、メルケプストラム差分 10 次元、およびパワー差分を用いた。HMM は対角分散行列をもつ単一ガウス分布 HMM であり、音素の種類数は 37、各音素の状態数は 4 とした。また、状態の分割条件の種類数は 106 である。学習データとして 2 セット用意した。

データ A 男性 46 名の音素バランスを考慮した 250 単語 1 回発声

データ B 男性 36 名の音素バランスを考慮した 2150 単語 1 回発声

データ B はデータ A の 7–8 倍のデータ量となっている。テストデータとして学習データベースに含まれない男性 5 名の学習単語とは異なる 250 単語 1 回発声を用いた。

まず、データ A を用いて学習した場合の提案法の結果と参照実験の結果を比較した。参照実験として、次に示す尤度最大基準による状態分割法 (e.g., [6]) の認識実験を行なった。分割条件 q によってクラスタ S を S_{q+} と S_{q-} に 2 分割したときの、式 (3) で表される尤度 L の増分を δ_q とする。まず、式 (5) で表される頻度 $\Gamma(S_{q+})$、$\Gamma(S_{q-})$ がともに一定値 D 以上であるという条件を満たす分割条件 q のうちで、δ_q を最大とする分割条件 q' を選ぶ。そして、$\delta_{q'}$ が一定値 V 以上のとき、クラスタを分割する。これは、分割に伴う尤度の増分がある閾値 V 以上であり、かつ分割後の 2 つのクラスタそれぞれに対し閾値 D 以上の量の学習データが対応するという条件を満たした分割条件のうち、尤度の増分が最大になる分割条件で分割を行なう方法である。この方法を 12 通りの D、V について評価した (Ref 1–12) 話者 5 名について認識評価実験を行なった結果の平均認識率を表 4に示す。表 4に示す通り、提案法は 12 通りの参照実験よりも高い認識性能を示している。参照実験としてすべての場合を尽くしてはいないが、本手法を用いることにより、少ない計算量で、比較的高い認識性能を得ることが可能であることがわかる。また、提案法で得られた HMM の総状態数は 2000 程度であることがわかる。

次に、学習データベースとして、データ B を用いたときの実験結果を表 4に示す。データ量が増加することにより MDL 基準で選ばれた HMM

表 1: データ A を用いたときの認識実験結果

	D	V	状態数	認識率 (%)
提案法	–	–	2069	80.4
Ref 1	60	0	3739	75.4
Ref 2	100	0	3000	76.4
Ref 3	200	0	2001	76.7
Ref 4	300	0	1943	75.4
Ref 5	400	0	1200	73.4
Ref 6	500	0	1018	71.9
Ref 7	1000	0	591	66.6
Ref 8	60	200	2777	76.2
Ref 9	60	400	2034	77.0
Ref 10	60	600	1488	77.8
Ref 11	60	800	1248	77.9
Ref 12	60	1000	751	77.4

表 2: データ B を用いたときの認識実験結果

		データ A	データ B
状態数		2069	6223
認識率 (%)	話者 1	72.8	84.8
	話者 2	76.8	84.4
	話者 3	89.2	92.4
	話者 4	81.6	83.6
	話者 5	81.6	84.8
	平均	80.4	86.0

の状態数も増加している。また認識性能も向上し、誤りが約 30% 減少していることがわかる。

さらに MDL 基準が最適な基準かどうかを調べるため以下のような実験を行なった。式 (6) の代わりに以下の式を用いる。

$$l'(U) = \frac{1}{2}\sum_{m=1}^{M}\Gamma(S_m)\log(|\,\Sigma(S_m)\,|) + cKM\log\Gamma(S_0), \qquad (9)$$

係数 c を変えることによりデータに対する尤度を表す第 1 項に対するモデルの複雑さを示す第 2 項の重みを変化させることができる。データ B を用いて係数 c を 3 通りに変化させたときの実験結果を表 4に示す。係数 c の値の変動により認識性能はあまり変化しないことがわかる。$c = 2$ のときに認識性能がやや高い。

最後に 1 状態あたりのガウス数を増やしたときの効果を調べた。結果を表 4に示す。これは 1 ガウスで求められた認識単位をそのまま用いて、ガウス数のみを増やした場合の結果である。ガウス数を増加させることにより認識性能が 3% ほど向上する。対応するデータの分散が大きいクラスタにおいては、ガウス数の増加が分布当たりの分散を減少させる効果があったものと考えられる。MDL 基準は、分散を減少させる方向への状態分割にも用いることが可能であり、今後適用してみたい。

表 3: 重み係数 c と認識率 (%)

係数 c	0.1	0.5	1.0	2.0	4.0	10.0
状態数	13927	9798	6223	3949	2418	1341
話者 1	84.0	84.4	84.8	83.6	82.4	79.6
話者 2	81.6	83.6	84.4	84.4	84.8	80.8
話者 3	92.0	92.0	92.4	92.8	92.4	91.2
話者 4	84.8	85.2	83.6	85.2	84.8	82.0
話者 5	84.4	84.4	84.8	87.6	85.2	86.8
平均	85.4	85.9	86.0	86.7	85.9	84.1

表 4: ガウス数を増加させたときの認識性能の変化

		1 ガウス	2 ガウス
認識率 (%)	話者 1	84.8	86.8
	話者 2	84.4	87.6
	話者 3	92.4	94.4
	話者 4	83.6	88.8
	話者 5	84.8	87.2
	平均	86.0	89.0

5 おわりに

MDL 基準の意味で最適な状態数をもつ HMM を生成する枠組を提案し、実験により効果を確認した。情報量基準を導入することにより、音声認識モデルの複雑さを決定する際の一つの目安を与えることができたと考える。

今後は、分割に用いる素性の最適化を行なうとともに、ガウス数を MDL 基準を用いて増減させる場合について評価を行ないたい。

参考文献

[1] K.-F.Lee: "Automatci Speech Recognition: The Development of the SPHINX System", Kluwer Academic Publishers, Boston (1989).

[2] K.-F.Lee *et. al.*: "Allophone Clustering for Continuous Speech Recognition", *Proc.ICASSP-90*, Albuquerque, pp.749-753 (1990).

[3] L.R.Bahl *et. al.*: "Decision Trees for Phonological Rules in Continous Speech", *Proc.ICASSP-91*, Toronto, pp.185-188 (1991).

[4] 鷹見: "逐次状態分割法による隠れマルコフ網の自動生成", 信学論 (D-II), J76-D-II,10, pp.2155-2164 (1993).

[5] M.-Y.Hwang *et. al.*: "Predictiong Unseen Triphones with Senones", *Proc.ICASSP-93*, Minneapolis, pp.II-311-314 (1993)

[6] S.J.Young *et. al.*: "Tree-Based State Tying for High Accuracy Acoustic Modelling", *Proc. of Human Language Technology*, pp.307-312 (1994)

[7] 堀他: "音素決定木に基づく逐次状態分割法による HM-Net の検討", 信学技報, SP96-22, pp.15-22 (1996).

[8] J.Rissanen:"Univarsal Coding, Information, Prediction, and Estimation", *IEEE Trans. IT*, vol.30, No.4, pp.629-636 (1984)

[9] 韓, 小林: "情報と符号化の数理", 岩波講座応用数学 13, 岩波書店 (1994).

[10] 渡辺他: "半音節を単位とした HMM を用いた不特定話者大語い認識", 信学論 (D-II), J75-D-II,8 (1992)