論文 / 著書情報
Article / Book Information

論題(和文)	情報量基準を用いた音声認識単位の自動生成
Title(English)	
著者(和文)	篠田浩一, 渡辺隆夫
Authors(English)	Koichi Shinoda
出典(和文)	日本音響学会平成8年度秋季研究発表会講演論文集, Vol. 2-3-11, No. , pp.
Citation(English)	, Vol. 2-3-11, No. , pp.
発行日 / Pub. date	1996,

2—3—11　情報量基準を用いた音声認識単位の自動生成 *

⊙ 篠田 浩一　渡辺 隆夫
(NEC 情報メディア研究所)

1. はじめに

音声認識の認識単位として、先行音素や後続音素などのコンテキストを考慮したコンテキスト依存単位を用いると性能が向上することが知られている。もし、十分な量の学習データがあれば、認識単位の種類が多いほど認識性能が高くなることが予想される。しかし、現実には学習データの量は有限であり、コンテキスト依存単位の種類を増やすと、各コンテキストに対応する学習データの量が減少し、性能が劣化する。この問題に対しては、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。

しかしながら、従来法では、認識単位の統合・分割の手続きの停止基準が与えられていない。認識単位の種類数は、テストデータに対する認識実験、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が、これらは、計算量が多く、最適性に対する定性的な説明がない。本稿では、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、停止基準として用いる方法を提案する。

2. MDL 基準

MDL 基準 [8, 9] は情報量基準の 1 つであり、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。

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 l を与えるモデルを最適なモデルとする基準である。確率モデル i に対する記述長 $l(i)$ は以下の式で与えられる。

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

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。上式における第 1 項は、データに対する対数尤度 (以下、尤度と記す) 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。モデルが複雑になるにつれ、第 1 項の値は減少し第 2 項の値は増加する。記述長 $l(i)$ は適当な複雑さをもつモデルで最小になる。

* Automatic Generation of Speech Recognition Units Using Information Criterion, by Koichi SHINODA and Takao WATANABE (NEC Corporation)

3. MDL 基準を用いた認識単位の自動生成

ここでは、音素を基本単位とし、単一ガウス分布連続 HMM の各状態をトップダウンに分割する [6, 7]。まず、音素 HMM の各状態に対し、複数の分割条件から最適な分割条件を選択し、状態を分割する。そして、分割された状態に対しさらに分割を行なう処理を繰り返し状態を細分化していく。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。最適な分割条件の選択、および、分割の停止の決定に MDL 基準を用いる。分割された結果は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)。

今、分割の過程で、状態 S が $S_1,\ldots,S_M$ の M 個の状態へと分割されているとする。このとき、単一ガウス分布を出力確率分布とする状態 M 個から構成されるモデル U に対する記述長 $l(U)$ を計算する。ここで、状態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し、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。まず、状態 S_m に対する学習データの尤度 $L(S_m)$ を、

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

と近似する。ここで、K は平均ベクトルおよび分散の次元数であり、

$$\gamma_t(S_m) = \frac{\alpha_t(S_m)\beta_t(S_m)}{\sum_s \alpha_t(S_m)\beta_t(S_m)} \quad (3)$$

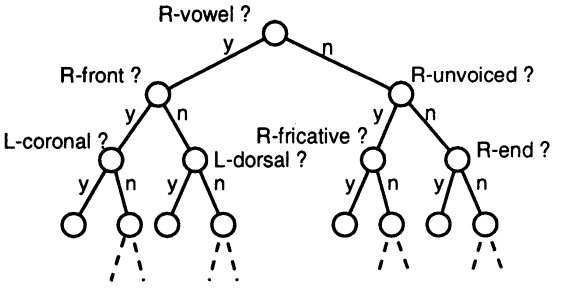

図 1: Phonetic decision tree

$$\Gamma(S_m) = \sum_{t=1}^{T} \gamma_t(S_m) \quad (4)$$

である。この$\Gamma(S_m)$は学習データ中に状態S_mが出現する頻度を表す。前向き確率 $\alpha_t(S_m)$、後ろ向き確率$\beta_t(S_m)$、平均ベクトルμ_{S_m}、分散Σ_{S_m}は、学習データより求める。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。

$$l(U) = \frac{1}{2}\sum_{m=1}^{M} \Gamma(S_m)\log(|\,\Sigma(S_m)\,|) + KM\log\sum_{m=1}^{M}\Gamma(S_m) \quad (5)$$

この記述長$l(U)$を最小にするモデル(状態の組)が、MDL基準の意味で最適な分割を表す。

今、ある分割条件qを用いて、ある状態Sを2分割したときの記述長の増分をΔ_qとする。まず、すべての分割条件について、状態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、$\Delta_{q'} < 0$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。

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。学習データとして男性46名の音素バランスを考慮した250単語1回発声を用い、テストデータとして男性5名の学習単語とは異なる250単語1回発声を用いた。

提案法の話者5名の平均認識率を表1に示す。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によって状態S_0をS_{q+}とS_{q-}に2分割したときの、式(2)で表される尤度Lの増分をδ_qとする。まず、式(4)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、状態を分割する。これは、分割後の各状態に対し閾値D以上の量の学習データが対応し、かつ、分割に伴う尤度の増分がある閾値V以上である、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) 表1に示す通り、提案法は参照実験よりも高い認識性能を示している。また、実験により最適化する必要のあるパラメータがないため、計算量も少ない。

表 1: 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

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。今後は、分割に用いる素性の最適化を行なうとともに、学習データ、認識データの規模がより大きい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)