音素識別誤り情報に基づく平滑化制御を用いた VFS 話者適応化*

大倉 計美**

VFS Speaker Adaptation by a Control Method for Smoothing Parameters Based on Classification Error*

Kazumi OHKURA**

1. はじめに

音声認識性能を向上させるために、少量の音声データを用いてモデルを話者の音声特徴に適應する技術である話者適応化の研究が活発に行われている¹⁾⁻⁶⁾。Hidden Markov Model (HMM) における話者適応化では、学習に大量の音声データを必要とする HMM を、少量の音声データにより再学習することによって生じる「未学習モデル問題」と「モデルの推定誤差問題」を解決する必要がある。筆者は、この問題に対処するために、混合連続分布型 HMM に基づく、移動ベクトル場平滑化話者適応化方式 (transfer Vector Field Smoothing method: VFS) を提案した¹⁾。

VFS は、写像関数である移動ベクトルが張る場の平滑化に基づく手法であるため、部分空間ごとに平滑化の強さを制御することが必要である。筆者は、音素の正解系列が既知である教師あり話者適応化では、適応用音声を認識する場合の拘束条件の違いから生じる情報 (識別誤り) が、適応時のパラメータ制御の情報として有効であると考え、適応用音声を認識した場合に起こる識別誤りを最小化するように⁷⁾、部分空間ごとの平滑化を制御する平滑化制御法を提案する²⁾。また、提案の手法を、最大事後確率推定 (MAP 推定)^{3),4)} と VFS とを組み合わせた MAP-VFS 話者適応化手法⁵⁾ に適用した評価実験により、HMM の識別誤り情報が、部分空間ごとの平滑化制御を実現するために有効な情報であることを示す。

本速報では、まず、MAP 推定の概要を、つぎに、VFS の基本的な考え方とアルゴリズムを示し、最後に、平滑化制御法のアルゴリズムと評価結果について述べる。

2. 適応化アルゴリズム

2.1 MAP 推定

適応用単語音声データを $O = \{o_1, \dots, o_t, \dots, o_T\}$ と表す。 O に対してビタビアルゴリズムを適用し、音素 HMM を連結した単語モデル (λ_{word}) の各状態と、 o_t との対応関係 (Θ_c) を求める。MAP 推定^{3),4)} を Θ_c に基づいて行い、平均ベクトル μ_k^{map} を求める²⁾⁻⁴⁾。

$$\mu_k^{map} = \frac{\sum_{t \in \Psi_k} n_{k,t} o_t + \tau_k \mu_k}{n_k + \tau_k} \quad (1)$$

$$n_k = \sum_{t \in \Psi_k} n_{k,t} \quad (2)$$

ここで、 μ_k は、初期音素 HMM セット (Λ^{init}) 中の k 番目の平均ベクトルを表す。 Ψ_k は、 μ_k に対応付けられた学習データ (o_t) のフレーム番号 (t) の組である。 $n_{k,t}$ は、 μ_k から o_t が生起する確率である。 n_k は、 μ_k の学習に寄与した学習データの実質的な個数を表す。 τ_k ($\tau_k \geq 0$) は、学習データに対する事前情報の信頼度を制御する定数である。 τ_k の値を大きくするに従い、与えられた学習データに対する学習速度は、緩やかになる。

2.2 VFS 話者適応化方式

MAP 推定で学習できる音素 HMM は、 λ_{word} に含まれるモデルのみであるため、学習できない音素 HMM が存在するという問題がある。また、 O が表現できる音声特徴空間が、話者の音声特徴空間の一部であるため、 O から推定したモデルが表す特徴空間が、必ずしも大量の音声データから推定したモデルが表す話者の真の特徴空間を表現してはいないという問題がある。筆者は、この問題を「未学習モデル問題」および「モデルの推定誤差問題」と定義した¹⁾。VFS は、この二つの問題に対処する手法であり、初期モデルが表現している音声特徴空間を部分空間に分割した場合に、部分空間内に存在する各分布の相対的な位置関係が、再学習したモデル中においても保存されるという考え方に基づいている。

* 原稿受付 1997年8月2日

** 三洋電機 (株) ハイパーメディア研究所 SANYO Electric Co., Ltd. Hypermedia Research Center; 1-18-13 Hashiridani, Hirakata city, Osaka 573, JAPAN

Key Words: HMM, speaker adaptation, word recognition, maximum a posteriori estimation, VFS.

VFS の概念図を Fig. 1 に示す。Fig. 1 は、 k -近傍則で制限された部分空間における移動ベクトル t_k と μ_k^{map} が近傍の移動ベクトル (t_1, t_2, t_3) に拘束されることにより、平滑化移動ベクトル v_k と適応後の平均ベクトル μ_k^{vfs} が得られることを表している。以下に、VFS のアルゴリズムを説明する。

Fig. 1 Concept of VFS

ステップ 1 移動ベクトルの計算

μ_k^{map} および μ_k から、移動ベクトル t_k を求める。

$$t_k = \mu_k^{map} - \mu_k, \quad \forall k \in \Omega \quad (3)$$

ここで、 Ω は、初期音素 HMM セット (A^{init}) に含まれる、すべての平均ベクトルの添え字 (k) の組である。

ステップ 2 平滑化移動ベクトル v_k の計算

$$v_k = \frac{X_k + t_k}{Y_k + 1.0}, \quad \forall k \in \Omega \quad (4)$$

$$X_k = \sum_{k' \in E_k} \omega_{k,k'}(f) t_{k'} \quad (5)$$

$$Y_k = \sum_{k' \in E_k} \omega_{k,k'}(f) \quad (6)$$

X_k と Y_k は平滑化にかかわる項であり、 Y_k の値が大きいほど、平滑化の影響は強くなる。 $\omega_{k,k'}(f)$ は、 μ_k と $\mu_{k'}$ との距離に基づく重みであり、距離の逆数に比例した値をとる。 $\omega_{k,k'}(f)$ は、平滑化係数 f ($f \geq 0$) により制御できる。 f の値を 0 に近づけることにより、 μ_k の近傍にある $\mu_{k'}$ に重みをおいた平滑化移動ベクトル v_k が求められる。 E_k は、 μ_k の近傍にある K 個の学習された $\mu_{k'}$ の添え字の組である。部分空

間の大きさと拘束の強さは、 k -近傍則の K の値と重み $\omega_{k,k'}(f)$ により、それぞれ、制御できる。

ステップ 3 平均ベクトルの適応化

MAP-VFS に基づく適応後の平均ベクトル μ_k^{vfs} を参考文献 1), 2), 5) より、次式で求める。

$$\mu_k^{vfs} = \mu_k + v_k, \quad \forall k \in \Omega \quad (7)$$

また、MAP-VFS により得られる適応後音素 HMM セットを、 A^{vfs} と表す。

2.3 識別誤り情報に基づく平滑化制御

本速報で提案する手法は、(4) 式における X_k と Y_k を平滑化制御係数 α_k^n を用いて、平滑化を制御するものである²⁾。平滑化を制御するためには、平滑化係数 (f) を制御する方法がある。しかし、 f を制御した場合、 X_k と Y_k を再計算する必要があるため、本速報では、(8) 式に示すように平滑化制御係数 α_k^n を導入し、 α_k^n により平滑化の量を制御する方法を用いる。 α_k^n は k -近傍則で制限される部分空間ごとに求められるため、部分空間ごとに X_k と Y_k の制御を行うことが可能となる。ここでは、 α_k^n の決定問題を、単語辞書等の制約条件 $\mathcal{G}(\mathcal{O})$ のもとで、適応音声データ ($\mathcal{O}$) を認識した結果から得られる状態系列 (θ_g) の対数尤度 ($\ln(P(\mathcal{O}|\theta_g))$) と、正解状態系列 (θ_c) の対数尤度 ($\ln(P(\mathcal{O}|\theta_c))$) との 2 乗誤差 (d^2) を最小化する問題としてとらえ、最急降下法を用いて d^2 を最小化する α_k^n ($\alpha_k^n \geq 0$) を求める。

本速報では、制約条件 ($\mathcal{G}(\mathcal{O})$) として、 $\mathcal{G}_1(\mathcal{O})$ および $\mathcal{G}_2(\mathcal{O})$ の評価を行う。 $\mathcal{G}_1(\mathcal{O})$ は、 $\mathcal{G}_1(\mathcal{O})$ よりも、識別誤りが発生しやすい条件である。

$\mathcal{G}_1(\mathcal{O})$: あらゆる音素 HMM が、任意に接続可能。

$\mathcal{G}_2(\mathcal{O})$: すべての HMM の状態が、任意に接続可能。

$$v_k = \frac{\alpha_k^n X_k + t_k}{\alpha_k^n Y_k + 1.0}, \quad \forall k \in \Omega \quad (8)$$

ここで、 n は最急降下法の繰返し番号である。

以下に、平滑化制御アルゴリズムについて述べる。

ステップ 1 不特定話者音素 HMM セット (A) を初期音素 HMM セット (A^{init}) とする。

ステップ 2 初期音素 HMM セット (A^{init}) に対して MAP-VFS を適用し、音素 HMM セット (A^{vfs}) を求める。 $n=0$, $\alpha_k^n=1.0$ ($k \in \Omega$) とする。

ステップ 3 A^{vfs} を用いて、 θ_c , θ_g , 尤度 $P(\mathcal{O}|\theta_c)$, $P(\mathcal{O}|\theta_g)$ と、対数尤度の差 d を求める。

$$d = \ln(P(\mathcal{O}|\theta_g)) - \ln(P(\mathcal{O}|\theta_c)) \quad (9)$$

$d=0$ の場合 (θ_g が θ_c と一致する場合) は、ステップ 7 へ進む。

ステップ 4 α_k^{n+1} を, 最急降下法により求める.

$$\alpha_k^{n+1} = \alpha_k^n - e(n) \frac{\partial d^2}{\partial \alpha_k^n}, \quad \forall k \in \Omega \quad (10)$$

ステップ 5 $n \rightarrow n+1$ とする. (8) 式に従い, 平滑化移動ベクトル $\mathbf{v}_k$ を求める. $\alpha_k^n < 0$ の場合は, $\alpha_k^n = 0$ とする. さらに, 平滑化を制御した平均ベクトル (μ_k^{vfs}) を求める.

$$\mu_k^{vfs} = \mu_k^{init} + \mathbf{v}_k, \quad \forall k \in \Omega \quad (11)$$

ステップ 6 Λ^{vfs} を用いて, θ_c , θ_g , $P(\mathbf{O}|\theta_c)$ および, $P(\mathbf{O}|\theta_g)$ を求め, 下記の条件判定を行う.

条件 1 $\ln(P(\mathbf{O}|\theta_g)) - \ln(P(\mathbf{O}|\theta_c)) = d$

条件 2 $\ln(P(\mathbf{O}|\theta_g)) - \ln(P(\mathbf{O}|\theta_c)) = 0$

条件 3 $n = N$

条件 1 は, 学習が収束したこと, 条件 2 は, $\theta_g = \theta_c$ となったことを示している. N は学習の繰返し番号の最大値を表す. 本速報では, $N = 600$ とした. 上記のいずれかの条件を満たす場合は, ステップ 7へ進む. 満たさない場合は, $d = \ln(P(\mathbf{O}|\theta_g)) - \ln(P(\mathbf{O}|\theta_c))$ として, ステップ 4へ戻る.

ステップ 7 Λ^{vfs} を適応後音素 HMM セット (Λ^{adapt}) とし, 話者適応化処理を終了する. 適応化のための次の単語が存在する場合は, 適応後音素 HMM セット (Λ^{adapt}) を初期音素 HMM セット (Λ^{init}) とし, ステップ 2へ戻る.

3. 話者適応化アルゴリズムの評価

初期音素 HMM セットである不特定話者音素 HMM セットの作成には, 日本音響学会連続音声データベースに含まれる男性話者 30 名が発声した 2,554 文章を用いた²⁾. 評価用音声データには, 電子協日本語共通音声データ (JS-WRD-89-0031) に含まれる男性話者 70 名が発声した地名 100 単語を用いた. 特徴量には, 16 次 LPC ケプストラム, 16 次 Δ ケプストラム, Δ 対数パワーの 33 次元ベクトルを用いた. 音素 HMM は 4 状態 3 ループ, 対角共分散行列の混合ガウス分布型である. 各状態の混合数は 4, モデル数は 39 種とした. 適応用音声データには, 100 地名のうちの一部 (1~20 地名) を使い, 認識用音声データには, 適応用音声データ以外の評価用音声データを用いた. 評価は適応用単語セットを変え 3 回行い, 単語認識は, 音素 HMM を連結した単語 HMM を用いて行った. K および f の値は, 男性話者 5 名を評価話者とし, すべての部分空間に対する K および f の値を同一値とした条件で MAP-VFS を行い, 最も高い音声認識性能を示した K および f の値を用いることにした⁶⁾.

Table 1 に, 初期音素 HMM セットである不特定話者音

素 HMM セット (SID), MAP 推定により得られた音素 HMM セット (MAP), 平滑化制御なし (MAP-VFS), 2.3 で述べた制約条件 $\mathcal{G}_1(\mathbf{O})$ および, $\mathcal{G}_2(\mathbf{O})$ を用いて平滑化制御を行った場合の音声認識率を示した. Table 1 から, 識別誤り情報に基づく平滑化制御法の有効性がわかる. また, 制約条件に関しては, $\mathcal{G}_2(\mathbf{O})$ を用いた場合に, 若干ではあるが, $\mathcal{G}_1(\mathbf{O})$ を用いた場合よりも高い認識結果が得られることがわかる.

Table 1 Recognition rates [%]

	適応用単語数				
	1	3	5	10	20
SID	89.1	89.1	88.8	90.2	89.4
MAP	90.0	91.0	91.3	89.4	92.3
MAP-VFS	90.5	92.4	93.2	94.4	95.6
$\mathcal{G}_1(\mathbf{O})$	90.4	93.0	93.8	94.8	96.2
$\mathcal{G}_2(\mathbf{O})$	90.5	93.3	94.0	94.8	96.2

4. おわりに

本速報では, 識別誤り情報による平滑化制御法の提案と, 識別誤り情報が, 平滑化を制御するための情報として有効であることを示した. また, 識別誤りを得るための制約条件に関しては, 「あらゆる音素 HMM が, 任意に接続可能」という条件よりも, 「すべての HMM の状態が, 任意に接続可能」という条件を用いた場合に, 若干ではあるが, 高い認識結果が得られることがわかった. 今後は, 制約条件に関して, 詳細な検討を行う必要がある.

参考文献

- 1) 大倉, 杉山, 嵯峨山: 混合連続分布 HMM 移動ベクトル場平滑化話者適応方式; 信学論, Vol. J76-D-II, No. 12, pp. 2469-2476 (1993)
- 2) 大倉, 飯田: 音素識別誤りに基づく平滑化制御を用いた移動ベクトル場平滑化話者適応; 信学技報, SP96-23, pp. 23-30 (1996)
- 3) J-L Gauvain and C. H. Lee: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains; *IEEE Trans. Speech & Audio Process.*, Vol. 2, No. 2, pp. 291-298 (1994)
- 4) 中川, 越川: 最大事後確率推定法を用いた連続分布型 HMM の適応化; 音響学会誌, Vol. 49, No. 10, pp. 721-728 (1993)
- 5) J. Takahashi and S. Sagayama: Vector-field smoothed bayesian learning for incremental speaker adaptation; *Proc. of ICASSP95*, pp. 696-699 (1995)
- 6) 大倉, 飯田: MAP-VFS 話者適応化における移動ベクトル選択の効果; 電気関係学会関西支部連合大会, G20-14, pp. G404 (1996)
- 7) K. Ohkura, D. Rainton and M. Sugiyama: Noise-robust HMMs based on minimum error classification; *Proc. of ICASSP93*, pp. II-75-II-78 (1993)