Vol.2009-SLP-77 No.6
2009/7/17

最小相対エントロピー識別学習に基づくカーネルマシンを利用した音声認識

久保 陽太郎[†1]　渡 部 晋 治[†2]　中 村 篤[†2]
エリック マクダーモット[†2]　小 林 哲 則[†1]

本稿ではカーネルマシンに基づく音声認識法を提案する．提案手法では，Log-Linear タイプの出力分布を持つ連続分布型隠れマルコフモデル (CD-HMM) を最小相対エントロピー識別によって学習させる時の目的関数，および学習されたモデルのパラメタにカーネル法に基づく表現を導入する．従来のカーネル法ベースの音声認識と異なり，提案法は隠れマルコフモデルの構造をそのまま利用しているため，音声認識問題をフレーム列，隠れ状態列，ラベル列の 3 つの系列の変換問題として取り扱うことができる．そのため，提案法には動的計画法を用いた各種探索技法を容易に組み込むことができる．これは現代的な音声認識システムの実装には必要不可欠な要素である．加えて，本稿では提案法を効率的に実現するために，Cutting Plane 法を拡張した最適化アルゴリズムについても提案する．提案法の評価は孤立音素認識タスクにて行なった．評価の結果として，提案モデルがテストデータを用いて十分に調整した CD-HMM と同等の性能を持つことを確認した．

A Kernel Machine Derived by Minimum Relative Entropy Discrimination For Automatic Speech Recognition

Yotaro Kubo,[†1] Shinji Watanabe,[†2]
Atsushi Nakamura,[†2] Erik McDermott[†2]
and Tetsunori Kobayashi [†1]

This article describes a novel method for automatic speech recognition (ASR) based on kernel-based nonlinear classification. The new approach is obtained by substituting kernel function into inner-product performed in the dual problem of a learning problems of CD-HMMs formulated using minimum relative entropy discrimination (MRED). Unlike earlier work on sequential pattern recognition using kernel methods, our method can accurately model the three kinds of hierarchical dynamic patterns in CD-HMMs. Sequences of discrete-valued labels (string-level) or hidden states (state-level), as well as sequences of continuous-valued speech-derived frames (frame-level) can all be represented within the same kernel-based framework. Therefore, many efficient sequential pattern recognition algorithms for CD-HMMs (e.g. dynamic programming, Viterbi decoding, and the forward-backward algorithm) can be integrated into the new approach. This is an essential requirement for state-of-the-art speech recognition systems. We also describe a practical and effective optimization procedure for the proposed model, based on a cutting plane algorithm. The performance of the proposed method was evaluated in isolated phoneme recognition experiments. The method was found to be comparable with well-tuned conventional models.

1. はじめに

現在，ほとんどの音声認識器では，出力分布として混合ガウス分布 (Gaussian Mixture Model; GMM) を用いた隠れマルコフモデル (Continuous Density Hidden Markov Model; CD-HMM) が用いられている．CD-HMM では一般的に，トレーニングデータと GMM の混合数を増やしていくことで最適な識別を行なうモデルが推定できると言われているが，著者らは GMM の持つ特異性，すなわち各種の最適化基準に対し局所解を複数持つという性質が，近年急速に整備されつつある膨大なトレーニングデータの有効利用を妨げている可能性があると考えている．そこで，この問題を軽減できるようなモデルを用いて音声認識を行なうことを目的とする．

指数分布族に属する分布 (ガウス分布など) は最尤推定法 (ML) や最小識別誤り学習法 (MCE) [⋆1]において局所最適解を持たないモデルである．CD-HMM がこれらの推定法に対し局所解を持つ原因は，状態系列と混合という二つの離散な隠れ変数を持つためであり，これらが適切に固定された上での推定であれば，やはり大域最適解を求めることができる．例えば，混合数を 1 とし状態系列を既知とした CD-HMM は ML であっても MCE であってもそれぞれの意味で最適なモデルを得ることができる．しかし，そのようにして生成したモデルは多くの場合，モデルの表現力が不足している．特に混合数を減じることによる表現力

†1 早稲田大学
Waseda University
†2 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
⋆1 一般的に用いられるロス関数 (シグモイド, 線形) のような性質の良いロス関数を用いた場合

の低下は著しく，混合数 1 の CD-HMM では，各フレームにおいて二次関数で表現される識別面しか構成できないことが知られている．

状態系列に関する隠れ変数は連続音声認識等，系列モデルが隠れマルコフモデルで表現されていることを必要とする技術の適用のためには必要不可欠なものであるが，混合数に関する隠れ変数は，混合をしなくても十分な表現力を持つ出力分布を利用することで消去することができる．本稿では混合数に関する隠れ変数を消去することで局所解に収束することを避けることを考える．そこで，局所解を持たない Log-Linear モデルを出力分布として考え，それをカーネル法によって任意の非線形識別が可能な形に拡張することで，大域最適性と非線形識別の両立を行なう．

既存のカーネル法を用いた音響モデル研究は，二つのアプローチに大別できる．一つ目のアプローチは，系列データを取り扱うカーネル関数を設計する手法である．下平らの Dynamic Time Alignment Kernel Support Vector Machine (DTAK-SVM) では系列データ間の内積を計算するカーネル関数を導入することにより，Support Vector Machine (SVM) で系列データの識別を可能にする手法であり，孤立単語認識で効果を挙げている [1]．しかし，DTAK-SVM は HMM を内包した枠組みではないため，動的計画法に基づく各種アプリケーション (連続音声認識やワードスポッティング等) に拡張することが容易ではない．二つ目のアプローチは，SVM と HMM を組み合わせて使うアプローチである．中井らは，Viterbi アラインメントによって得た HMM 状態とフレームの対応関係を SVM によって学習させ，入力系列を HMM 状態に識別する識別器を得た後，その識別器が導出するスコアをデコードすることで音声認識を行なう手法を提案した [2]．しかし，Left-To-Right 型 HMM における隣接した HMM 状態のように，本質的に区別をつける必要性の小さい識別もあり，フレームを HMM 状態に識別するアプローチは非効率的である．

本稿で提案する手法は，隠れマルコフモデルの構造を保ち，既存の音声認識技法の応用を可能にしながら，カーネル法を用いた非線形の識別を可能にする枠組みである．また，この手法は既存の静的パターン識別器 (SVM 等) を用いるのではなく，系列の識別を目的に定式化されているため，HMM 状態間の識別という非効率性を避けることができている．

本稿ではまず第 2 節においてモデルの定義を，第 3 節にて，そのモデルの学習法とカーネル法の導入を行なう．第 4 節では，学習法を実装する上で必要なテクニックについて述べる．最後に，第 5 節で実験とその結果に対する考察を述べる．

2. Log-Linear 出力分布を用いた HMM による識別

本稿では，N 個のデータを含むトレーニングデータに含まれる i 番目のをフレーム列を X^i，それに対応する正解ラベル列を W^i とおき，トレーニングデータ全体を $\{(X^i, W^i)|i \in [1, N]\}$ のようにあらわす．ここで X^i は T^i 個の D 次元ベクトルの系列であり，系列内の各フレーム (ベクトル) は $X^i(t)$ $(t \in [1, T^i])$ であらわす．また，サフィックスの付かない X，W はトレーニングデータに含まれないフレーム列，ラベル列をあらわす．

現在の音声認識デコーダでは，全体で S 個の状態を持つ HMM $\{\Lambda \stackrel{\text{def}}{=} \{\lambda_s|s \in [1, S]\}, \Pi, \Theta\}$ (それぞれ，出力分布パラメタ，状態遷移パラメタ，言語モデルパラメタ) を用い，以下の最適化問題 (最短経路問題) を解くことによって，音声データ X に対応する正解ラベル W を推定する．

$$\begin{aligned} W =& \operatorname*{argmax}_{\tilde{W}} P(X, \tilde{W}|\Lambda, \Pi, \Theta) \\ P(X, \tilde{W}|\Lambda, \Pi, \Theta) =& P(X|\tilde{W}, \Lambda, \Pi) P(\tilde{W}|\Theta) \\ \approx& P(X|\hat{q}_{\tilde{W}}, \Lambda) P(\hat{q}_{\tilde{W}}|\tilde{W}, \Pi) P(\tilde{W}|\Theta) \\ =& \prod_t P(X(t)|\lambda_{\hat{q}_{\tilde{W}}(t)}) P(\hat{q}_{\tilde{W}}|\tilde{W}, \Pi) P(\tilde{W}|\Theta) \end{aligned} \tag{1}$$

ここで $\hat{q}_{\tilde{W}}$ は Viterbi 系列，すなわち

$$\hat{q}_{\tilde{W}} = \operatorname*{argmax}_q P(q|\tilde{W}, X, \Lambda, \Pi) \tag{2}$$

であり，フレーム t における HMM 状態は $\hat{q}_{\tilde{W}}(t)$ で表わされる．

本稿では，上式であらわされる既存のデコーダで正解ラベルを推定できるような音響モデルを提案する．言語モデルおよび，状態遷移パラメタの推定は本稿のスコープから外れるため，以降，Θ と Π は適宜省略する．

本稿ではまず，トレーニングデータ中のフレーム列 X^i を誤り候補 W^{c} に対し，どれだけ余裕 (マージン) を持って識別が可能であるかを示す識別関数 $\mathcal{K}$ を導入する．

$$\mathcal{K}(X^i, W^{\mathrm{c}}|\Lambda) = \log \frac{P(X^i, W^i|\Lambda)}{P(X^i, W^{\mathrm{c}}|\Lambda)} \tag{3}$$

デコーダと同条件での評価を行なった上での識別関数を導出するため，デコーダ (式 (1)) で用いられている Viterbi 近似を導入することにより以下を得る．

$$\begin{aligned} \tilde{\mathcal{K}}(X^i, W^{\mathrm{c}}|\Lambda) \stackrel{\text{def}}{=}& \sum_t \left[\log P(X^i(t)|\lambda_{\hat{q}_{W^i}(t)}) - \log P(X^i(t)|\lambda_{\hat{q}_{W^{\mathrm{c}}}(t)})\right] + \log \frac{P(\hat{q}_{W^i}) P(W^i)}{P(\hat{q}_{W^{\mathrm{c}}}) P(W^{\mathrm{c}})} \\ \approx& \mathcal{K}(X^i, W^{\mathrm{c}}|\Lambda) \end{aligned} \tag{4}$$

この $\tilde{K}(X^i, W^{\mathrm{c}}|\Lambda)$ が全ての $W^{\mathrm{c}} \neq W^i$ に対して正になるようなモデルパラメタ Λ があれば，トレーニングデータにおける識別誤りをなくすことができる．

本手法では，各 HMM 状態の出力分布のモデルとして非線形写像関数 ϕ の写像先での Log-Linear モデルを用いる．

$$\begin{aligned} P(X^i(t)|\lambda_s) =& \frac{1}{Z_\phi(\lambda_s)} \exp\left\{\lambda_s^T \phi(X^i(t))\right\}, \\ Z_\phi(\lambda_s) =& \int_x \exp\left\{\lambda_s^T \phi(x)\right\} dx. \end{aligned} \tag{5}$$

一般に Z_ϕ の定義中にある積分を解析的に求めることはできない．本稿では，Z_ϕ の計算は省略し，正規化されていない疑似の確率分布を利用する．著者らは，パラメタが識別的に，つまり確率分布の推定問題の枠組みから外れて識別誤りを最小化するように推定される場合，この近似は問題になりにくいと考えている．この出力分布 (式 (5), $Z_\phi(\lambda_s) = 1$) を，識別関数 (式 (4)) に代入し，以下を得る．

$$\tilde{\mathcal{K}}(X^i, W^{\mathrm{c}}|\Lambda) = \sum_t \left[\lambda^T_{\hat{q}_{W^i}(t)} \phi(X^i(t)) - \lambda^T_{\hat{q}_{W^{\mathrm{c}}}(t)} \phi(X^i(t))\right] + \log \frac{P(\hat{q}_{W^i}) P(W^i)}{P(\hat{q}_{W^{\mathrm{c}}}) P(W^{\mathrm{c}})} \tag{6}$$

以降の節では，この識別関数が全てのトレーニングデータ X^i と誤り候補 W^{c} に対して十分に大きな値をとるようなパラメタ Λ の推定を考える．

3. 最小相対エントロピー識別による学習とカーネルトリックの導入

本節では，出力分布パラメタ Λ の推定について述べる．提案法は，隠れ変数を含む確率分布の識別学習を凸最適化の枠組みに沿って定式化することができる最小相対エントロピー識別 (Minimum Relative Entropy Discrimination; MRED) の枠組みに沿って定式化した．MRED は SVM の一般化であり，SVM と同様にパラメタ推定を制約付き最適化問題に帰着させているが，SVM と異なり，最適化中の全ての変数は確率変数であると仮定される．全ての変数を確率変数であると仮定すると，既存の正則化技法を事前分布との KL ダイバージェンスという形で記述することができ，従来の正則化手法の意味付けを行なうことができる．MRED を使うことで隠れ変数を含む分布の識別学習への拡張が容易である．本稿では局所解の問題を避けるため単一のカーネル関数でモデルを構築することを考えているが，[7] のように複数の特徴量を用いて複数のカーネル関数によるモデルの混合分布でモデル化するようなことも可能である．

最小相対エントロピー識別は，前節で定義した識別関数 $\tilde{\mathcal{K}}$ の期待値が一定以上であるという制約の上で，パラメタの事前分布に最も近いパラメタ分布を求める制約付き最適化問題として定式化される．制約の緩和を実現するスラック変数 ξ も確率変数として扱うことで，この問題はスラック変数とパラメタの事前分布 $P^0(\Lambda, \xi)$ から事後分布 $P(\Lambda, \xi)$ への KL ダイバージェンスを最小化する最適化問題となり，以下のように定式化される (主問題)．

$$\begin{aligned} &\underset{P(\Lambda,\xi)}{\text{minimize}}\ KL[P(\Lambda,\xi)||P^0(\Lambda,\xi)], \\ &\text{subject to} \int_\Lambda \int_\xi P(\Lambda,\xi)\left\{\tilde{\mathcal{K}}(X^i, W^{\mathrm{c}}|\Lambda) - \xi_{iW^{\mathrm{c}}}\right\} d\xi d\Lambda \geq 0, \qquad \forall i, \forall W^{\mathrm{c}} \neq W^i. \end{aligned} \tag{7}$$

この最適化問題は目的関数が凸汎関数 (KL ダイバージェンス) で，制約が線形汎関数 (期待値操作) で記述されている凸最適化問題である．

KKT 条件より，最適化問題の解はラグランジュ汎関数の鞍点にあることが知られている．変分法によって鞍点を求めることで，最適解におけるラグランジュ未定乗数 α と最適化問題 $P(\Lambda, \xi)$ に以下の関係があることを導出できる．

$$P(\Lambda,\xi) \propto P^0(\Lambda,\xi) \exp\left[\sum_{i, W^{\mathrm{c}} \neq W^i} \alpha_{iW^{\mathrm{c}}} \left\{\mathcal{K}(X^i, W^{\mathrm{c}}; \Lambda) - \xi_{iW^{\mathrm{c}}}\right\}\right], \qquad \alpha_{iW^{\mathrm{c}}} \geq 0. \tag{8}$$

ここで，α はラグランジュ未定乗数であり，制約ごとに一つずつ定義される．すなわち $\alpha_{iW^{\mathrm{c}}}$ は $i \in [1, T]$ と $W^{\mathrm{c}} \neq W^i$ において定義される．

凸最適化問題のラグランジュ関数の鞍点はラグランジュ未定乗数 α に関する最大点に存在する．得られた関係 (式 (8)) を主問題 (式 (7)) に代入し，$P(\Lambda, \xi)$ を消去し，α のみに関する最大化問題として定式化すると以下の双対問題を得る．

$$\begin{aligned} &\underset{\alpha}{\text{maximize}}\ J(\alpha) \qquad \text{subject to } \alpha_{iW^{\mathrm{c}}} \geq 0, \qquad \forall i, \forall W^{\mathrm{c}} \neq W^i \\ &\text{where } J(\alpha) = -\log \int_\Lambda \int_\xi P^0(\Lambda) P^0(\xi) \exp\left[\sum_{i, W^{\mathrm{c}}} \alpha_{iW^{\mathrm{c}}} \left\{\tilde{\mathcal{K}}(X^i, W^{\mathrm{c}}|\Lambda) - \xi_{iW^{\mathrm{c}}}\right\}\right] d\xi d\Lambda. \end{aligned} \tag{9}$$

双対問題の解と主問題の解は式 (8) で結ばれているため，この問題を解くことで得られた最適なラグランジュ未定乗数を式 (8) に代入することで，最適なパラメタ分布を求めることができる．

本手法では，ここで Soft-margin SVM と同様に，L2-norm 正則化を Log-Linear の重みベクトル λ_s に，L1-norm に相当するペナルティをスラック変数 ξ に与えることを考え，以下の事前分布を導入する．

$$\begin{aligned} P^0(\lambda_s) \stackrel{\mathrm{def}}{=}& \mathcal{N}(\lambda_s|0, I), \\ P^0(\xi_{iW^{\mathrm{c}}}) \stackrel{\mathrm{def}}{=}& \begin{cases} \frac{1}{C}\exp(-C|\Delta(W^i, W^{\mathrm{c}}) - \xi_{iW^{\mathrm{c}}}|) & \xi_{iW^{\mathrm{c}}} < \Delta(W^i, W^{\mathrm{c}}), \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{10}$$

Vol.2009-SLP-77 No.6
2009/7/17

ここで $\Delta(W^i, W^{\mathrm{c}})$ は正解ラベル W^i を W^{c} と誤認識した場合に与えるペナルティ尺度であり，実験の節で定義を行なう．

式 (10) で定義した事前分布を式 (9) に代入することにより，以下の目的関数を得る．

$$\begin{aligned} J(\alpha) =& \sum_{i,W^{\mathrm{c}}\neq W^i} J_{\xi_{iW^{\mathrm{c}}}}(\alpha) + \sum_s J_{\lambda_s}(\alpha) + \sum_{i,W^{\mathrm{c}}\neq W^i} J_{q_{iW^{\mathrm{c}}}}(\alpha), \\ J_{\xi_{iW^{\mathrm{c}}}}(\alpha) =& \Delta(W^{\mathrm{c}}, W^i)\alpha_{iW^{\mathrm{c}}} + \log(C - \alpha_{iW^{\mathrm{c}}}), \\ J_{\lambda_s}(\alpha) =& -||\hat{\lambda}_s(\alpha)||^2, \\ J_{q_{iW^{\mathrm{c}}}}(\alpha) =& -\alpha_{iW^{\mathrm{c}}}\left\{\log P(W^i) + \log P(\hat{q}_{W^i}) - \log P(W^{\mathrm{c}}) - \log P(\hat{q}_{W^{\mathrm{c}}})\right\}, \\ \hat{\lambda}_s(\alpha) =& \sum_{i,W^{\mathrm{c}}\neq W^i} \alpha_{iW^{\mathrm{c}}} \sum_t \Delta_s^\gamma(t; i, W^{\mathrm{c}})\phi(X^i(t)), \\ \Delta_s^\gamma(t; i, W^{\mathrm{c}}) =& I(\hat{q}_{W^i}(t), s) - I(\hat{q}_{W^{\mathrm{c}}}(t), s), \end{aligned} \tag{11}$$

ここで，$I(x; y)$ は指示関数であり，$x = y$ のときに 1 を，そうでない時に 0 を返す．

双対問題中では，目的関数の一部の項 ($J_{\lambda_s}(\alpha)$) は非線形変換をしたベクトル $\phi(x)$ 間の内積の重み付き和で表現される．カーネルトリックは，このような内積を半正定値カーネル関数 $K(x, y) \stackrel{\mathrm{def}}{=} \phi(x)^T\phi(y)$ で置換しカーネル関数 $K(x, y)$ を直接計算することで高次元空間への写像 $\phi(x)$ を現実的な時間で取り扱うテクニックである．式 (11) の $J_{\lambda_s}(\alpha)$ にカーネル関数を導入すると，以下のように変形できる．

$$J_{\lambda_s}(\alpha) = \sum_{i',W^{\mathrm{c}'}\neq W^{i'}} \sum_{i,W^{\mathrm{c}}\neq W^i} \alpha_{iW^{\mathrm{c}}}\alpha_{i'W^{\mathrm{c}'}} \sum_{t,t'} \Delta_s^\gamma(t; i, W^{\mathrm{c}})\Delta_s^\gamma(t'; i', W^{\mathrm{c}'})K(X^i(t), X^{i'}(t')). \tag{12}$$

上式の最適化の結果として得られるパラメタ分布は以下のようになる．

$$\hat{P}(\Lambda) = \prod_s \mathcal{N}(\lambda_s|\hat{\lambda}_s(\hat{\alpha}), I). \tag{13}$$

本稿では，この分布の最頻値 $\hat{\Lambda} = \mathrm{argmax}_\Lambda \hat{P}(\Lambda)$ をパラメタの推定値として利用する (MAP plug-in)．MAP plug-in は一般に分布表現の近似手法として導入されるが，提案法においては $E_\Lambda[\lambda_s^T x] = \hat{\lambda}_s^T x$ が常に成立するため，厳密な表現である．

出力分布のスコア $\lambda_s^T\phi(x)$ の計算も，$\hat{\lambda}_s$ が式 (11) のように表わされることを利用すると，以下のように直接 ϕ を計算することなく求めることができる．

$$\lambda_s^T\phi(x) = \sum_{i,W^{\mathrm{c}}\neq W^i} \alpha_{iW^{\mathrm{c}}} \sum_t \Delta_s^\gamma(t; i, W^{\mathrm{c}})K(X^i(t), x). \tag{14}$$

4. Modified Cutting Plane アルゴリズムを用いた実装法

音声認識の問題では，正解と異なるラベル $W^{\mathrm{c}} \neq W^i$ の数が組み合わせ爆発的に増加するため，前節で定義した最適化問題を既存の凸最適化ツールを用いて解くことは現実的ではない．そこで，Tsochantaridis らが提案した Cutting Plane アルゴリズム [8] に，Viterbi アラインメントの再構成を加えたアルゴリズムを提案し，実装した．表 1 に提案アルゴリズムの疑似コードを示す．

表 1 Modified Cutting Plane アルゴリズムの疑似コード
Table 1 Pseudo code of Modified Cutting Plane Algorithm

```
1: Λ̂(α) ≡ {λ̂_1(α), ··· , λ̂_s(α), ··· , λ̂_S(α)} (Eq. (11))
2: M(W^c|Λ) ≡ K̃(X^i, W^c|Λ) − Δ(W^i, W^c)
3: α_{iW^c} ← 0 for all i and W^c ≠ W^i
4: C_i ← φ for all i
5: loop
6:    i ← トレーニングデータを一つ選択する
7:    if λ̂_s(α) ≠ 0 for all s then
8:       Ŵ^c ← argmax_{W^c≠W^i} M(W^c|Λ̂(α)) /* 既存のデコーダで実行することができる */
9:    else
10:      Ŵ^c ← 誤りラベル列をランダムに生成する
11:   end if
12:   if λ̂_s(α) = 0 ∃s, or {min_{W^c∈C_i} M(W^c|Λ̂(α))} > min{0, M(Ŵ^c|Λ̂(α))} + ε then
13:      C_i ← C_i ∪ {Ŵ^c}
14:   end if
15:   q̂_{W^i} と q̂_{W̃} (∀W̃ ∈ C_i) をモデル Λ̂(α) と Viterbi アルゴリズムを用いて計算する．
      /* Viterbi 系列は 8 行目におけるデコーダの実行の中間出力として得られていることが多い．既に得られている場合はそれを利用し，計算をスキップできる */
16:   while α_{iW^c} converges for all W^c do
17:      W̃ ← C_i からランダムに選んだラベル W
18:      α_{iW̃} を与えられた q̂_{W^i} と q̂_{W̃} の上で最適化
19:   end while
20: end loop
```

提案アルゴリズムでは，集合 $\mathcal{C}_i$ が i 番目のトレーニングデータにおいて現在最適化の対象となっている変数 $\alpha_{iW^{\mathrm{c}}}$ のに対応する W^{c} を格納している．$\mathcal{C}_i$ の要素は，予測マージン $M(\hat{W}^{\mathrm{c}}|\Lambda)$ を最も小さくする $\hat{W}^{\mathrm{c}}$ が逐次追加されていく．これは主問題におい最も制約を満たしづらい $\hat{W}^{\mathrm{c}}$ を選択することに相当し，最適化に効果のありそうな変数を貪欲に選択し追加していくことに相当する．予測マージン $M(\hat{W}^{\mathrm{c}}|\Lambda)$ は識別関数の期待値 $E_\Lambda[\tilde{\mathcal{K}}(X^i, W^{\mathrm{c}}|\Lambda)]$ と最も望ましいスラック変数の値 $\mathrm{argmax}_\xi P^0(\xi)$ の差分として与えられる．

Axis Parallel Optimization [4] と同じく本手法でもラグランジュ未定乗数ベクトルを一要素ずつ最適化することを考えることで，各アップデートを解析的に導くことができ，最急

降下法より効率の良い最適化を行なうことができる．また，音声認識問題においては，18 行目で必要とされる，$\hat{q}_{\bar{W}}$ と $\hat{q}_{W^i}$ は 8 行目に行なわれるデコーディングの中間出力として得られている場合が多く，適宜計算をスキップさせることで効率的に解くことができる．

5. 実　　験

提案法の有効性を検証するため，孤立音素認識実験を行なった．先述の通り，提案法は CD-HMM と同様の構造を持つモデルであるため，容易に連続音声認識に拡張することができるが，本稿では系列識別器としての性能を評価するため孤立音素認識での性能を評価する．

比較対象としては，混合ガウス分布を出力分布として用いる HMM (CD-HMM) を最尤推定で学習させたものを用意した．また，提案モデル，CD-HMM ともに，3 状態の HMM を用いた．提案法のカーネル関数としては，3 次の多項式カーネルを用いた ($K(x,y)=(x^T y+1)^3$)．トレーニングセットおよびテストセットとしては，TIMIT データベースを元に作成した孤立音素データセットを用いた．データセットに含まれる音素カテゴリは 39 種類である．データセットの詳細およびモデルのハイパーパラメタの詳細を表 2 に示す．

特徴量は MFCC (12 次元), ΔMFCC, ΔΔMFCC, Energy, Δ Energy, ΔΔ Energy の計 39 次元を用いた．前処理としてトレーニングセット全体の平均と共分散を求め，トレーニングセットの平均が 0 分散が単位行列となるような線形変換を行なった (白色化)．白色化は識別モデルのトレーニングの前処理として一般的な手法であり，トレーニングデータから学習した線形変換であるため，Cepstral Mean Normalization (CMS) 等の正規化手法とは異なり，未知の入力系列に対し 1 フレーム毎に適用できる．

事前分布 (式 (10)) の調整に利用されている正解 W^i をラベル W^{c} に誤認識した際のペナルティ尺度としては，以下の式で定義される Viterbi 系列間ハミング距離を用いた．

$$\begin{aligned}\Delta(W^i, W^{\mathrm{c}}) =& \sum_{t=1}^{T^i}(1 - I(\hat{q}_{W^i}(t), \hat{q}_{W^{\mathrm{c}}}(t))) \\ =& T^i\end{aligned} \tag{15}$$

本実験は孤立音素認識タスクであるため，異なるラベルの Viterbi 系列間のハミング距離はフレーム数と一致する．

5.1 実 験 結 果

表 4 に比較した手法の音素誤り率を示す．

表より，提案法はトレーニングデータに対する誤り率を大幅に減らすことができることを確認した．また，オープンセットで調整した混合数を持つ CD-HMM に匹敵する性能を確認することができた．本実験ではクロスバリデーション等に基づくカーネルパラメータの調整は行なっておらず，結果としてトレーニングデータに対する誤り率とテストデータに対する誤り率の間に大きなギャップが生じてしまっているが，それでも適切にパラメータ調整が成された CD-HMM に匹敵する性能が達成できた．適切なハイパーパラメータの設定は重要な今後の課題である．

表 2 モデルとデータセット
Table 2 Dataset and model description

Dataset	
# categories	39
# seq. (train)	9,275
# seq. (test)	4,594
# frames (train)	77,463
# frames (test)	36,790
Model	
# states	3
Kernel type	Polynomial (order = 3)
C (Eq. (10))	5.0

表 3 音響分析条件
Table 3 Acoustical analysis configuration; Δ means time-domain derivative of feature sequence.

Sampling rate	16 kHz
Quantization	16 bits
Feature vector	MFCC (12 dims.), Energy Δ MFCC, Δ Energy ΔΔ MFCC, ΔΔ Energy (Total: 39 dims.)
Window len./ shift	25 ms / 10 ms

表 4 各手法の音素誤り率 (Closed: トレーニングセット/ Open: テストセットにおける誤り率； GMM: 一般的な CD-HMMs/ Poly. kernel: 提案法)
Table 4 Phoneme error rates of compared methods

State models	Closed	Open
GMM (1 mix.)	34.5	38.5
GMM (2 mix.)	29.2	35.0
GMM (4 mix.)	23.8	33.3
GMM (8 mix.)	17.0	31.0
GMM (16 mix.)	10.6	31.3
GMM (32 mix.)	7.8	32.5
Poly. kernel (order = 3)	1.8	31.4

5.2 解のスパースネス

SVM と同様に，ラグランジュ未定乗数 $\alpha_{iW^{\mathrm{c}}}$ が 0 とならない誤りラベル W^{c} を持つ ($\alpha_{iW^{\mathrm{c}}} \neq 0\ \exists W^{\mathrm{c}}$) X^i をサポートシーケンスと呼ぶ (サポートベクターと異なり系列全体でモデルを構成する)．サポートシーケンスの数は SVM におけるサポートベクターの数と同じ意味を持ち，計算効率と汎化能力の両面において重要な数である．

サポートシーケンスのみがモデルの構成に関係しているという事実から，サポートシーケンス以外の系列を一つ，トレーニングセットから除去してモデルを推定しても結果は全てのデータを使って推定したものと同一であることが予測される．また相補性定理より，サポートシーケンスでない系列は式 (7) で記述されている識別制約を満たすことが明かである．こ

れらのことから，トレーニングデータからサポートシーケンスでない系列を一つ除いて推定したモデルを用いて除いた系列を識別しても正しい結果が得られることが保証されている．これは Leave-One-Out 法 (LOO 法) として知られているクロスバリデーションの手順と一致しており，サポートシーケンスでない系列の数を増やす，つまりサポートシーケンスを減らすことで LOO 法で評価した際の性能を向上させることができる．LOO 法は汎化性能の予測に利用される手法であり，サポートシーケンス数の減少は汎化性能向上に寄与すると考えられる．

本実験で得られたサポートシーケンスの数は 6972 であった．従って，少なくとも 2303 個のデータ (25%のデータ) は LOO 法に基づくクロスバリデーションでも正しく識別されることが分かった．全てのデータがモデル構成に寄与していないというスパースネスの原理を提案法においても確認することができた．

スパースネスは計算効率の面でも重要である．計算効率で特に重要なのは，カーネル関数 $K(x, y)$ の評価回数であり，これはサポートシーケンスの中で，$\alpha_{iW^c}\Delta_s^\gamma(t; i, W^c) \neq 0\ \exists W^c$ が成り立つ (s, i, t) の組の数で決定される．HMM 状態 s について，上の条件を満たす (i, t) で示されるフレーム (i 番目のトレーニングデータにおける t 番目のフレーム) を，モデルの構成を支えるフレームであるという観点から HMM 状態 s のサポートフレームと呼ぶ．

本実験で得られたサポートフレームの総数は $280,184$ であった．サポートフレーム数の上限は，トレーニングデータのフレーム数 ((i, t) の数) に，競合する可能性のある状態の数 (競合が起こる s の数) を乗算したもの，すなわち，(トレーニングデータのフレーム数) $\times$ (HMM カテゴリ数 - 1) $= 2,943,594$ である．このことから，提案手法はカーネル法本来の計算コストの 90%を削減していることになる．提案法は系列識別問題の最適化であり，本質的にスパースになりやすい変数である α はトレーニングデータ中の系列毎に割り当てられる．フレーム毎の重みは系列毎の重み α_{iW^c} に Δ_s^γ を乗算することによって得られるが，Δ_s^γ は孤立音素認識タスクでは 0 にはならない．そのため，サポートフレームによる表現はいくらか冗長であることが考えられ，Reduced Set 法 [9] 等のより少ない数のベクトルでモデルを表現する手法が有効に働くことが期待される．

6. ま と め

本稿では，カーネル法で非線形の識別が可能になるように拡張した Log-Linear 出力分布を用いた隠れマルコフモデルと，その最小相対エントロピー識別基準による学習法，効率的な学習を行なうための最適化アルゴリズムを提案した．提案法では，入力系列中のベクトル (フレーム) はカーネル関数によって定義される特徴抽出関数によって高次元に写像され，写像先での Log-Linear モデルによってモデル化される．提案法の有効性は孤立音素認識タスクで評価を行なった．実験結果より，提案法はハイパーパラメタの調整を行なわなくても，テストセットでチューニングした従来法に匹敵する性能を持つことが示された．加えて，提案法においても SVM 同様スパースなモデル表現が得られていることがわかった．ハイパーパラメタの調整に関する検討は今後の課題である．

参 考 文 献

1) H. Shimodaira, K. Noma, M. Nakai, S. Sagayama, “Dynamic Time-Alignment Kernel in Support Vector Machine,” Advances in Neural Information Processing Systems (NIPS) 14, pp. 921–928 2002.
2) 中井，中井，下平，嵯峨山，“SVM を用いた時系列パターンの認識，” 信学技報, PRMU 99–167, pp. 122–144, 1999 年 12 月.
3) A. Ganapathiraju, J. Hamaker, J. Picone, “Hybrid SVM/ HMM Architectures for Speech Recognition,” Proc. 6th International Conference on Spoken Language Processing (ICSLP-2000), pp. 504–507, Beijing, China, 2000.
4) T. Jebara, “Machine Learning: Discriminative and Generative,” Kluwer Academic Publishers, 2004.
5) S. Reiter, B. Schuller, G. Rigoll, “Hidden Conditional Random Fields for Meeting Segmentation,” Proc. IEEE International Conference on Multimedia and Expo (ICME), pp. 639–642, Beijing, China, 2007.
6) G. Heigold, R. Schlüter, H. Ney, “On the Equivalence of Gaussian HMM and Gaussian HMM-like Hidden Conditional Random Fields,” Proc. Interspeech-2007, pp. 1721–1724, Antwerpen, Belgium, 2007.
7) D.P. Lewis, T. Jebara, W.S. Noble, “Nonstationary Kernel Combination,” Proc. 23rd International conference on Machine Learning (ICML), pp. 553–560, NY, USA, 2006.
8) I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, “Large Margin Methods for Structured and Interdependent Output Variables,” Journal of Machine Learning Research, Vol. 6, pp. 1453–1484, 2005.
9) C.J.C. Burges, B. Schölkopf, “Improving the Accuracy and Speed of Support Vector Machines,” Advances in Neural Information Processing Systems (NIPS-9), Vol. 9, pp. 375–381, 1997.