隠れ条件付確率場を用いた音声認識のためのアニーリングに基づく学習アルゴリズム

学籍番号　２３４１３５６８　氏名　真野 翔平
指導教員名　南角 吉彦

1　はじめに

近年の音声認識では，識別モデルに基づく音響モデリングが盛んに研究されている．従来の生成モデルを用いた音声認識では，音響特徴量を確率分布としてモデル化するため，全ての特徴量を優劣を考慮せず均等に用いて認識していたのに対し，識別モデルを用いた音声認識では，特徴量を選択的に用いた認識が可能であるという利点がある．しかし，モデル学習において局所最適性の問題がある場合，特徴選択の効果が十分に発揮できない可能性がある．そこで，本研究では識別モデルである隠れ条件付確率場 (Hidden Conditional Random Fields; HCRF) [1] を用いた音声認識における，アニーリング制御の適用による学習アルゴリズムの改善を提案する．

2　HMM の構造を持つ HCRF を用いた音声認識

本研究では HCRF に生成モデルである HMM の構造を適用するため，2 種類の素性関数 $f_{i,j}^{(a)}$, $f_{l,n}^{(b)}$ およびその重み $\lambda_{i,j}^{(a)}$, $\lambda_{l,n,d}^{(b)}$ を用いて状態遷移確率と出力確率を表現する．特徴量 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ が与えられたときに単語クラス C が出力される確率は，以下のように定義される．なお，特徴量 $\boldsymbol{X}$ に対する状態系列を $\boldsymbol{Y} = (y_1, y_2, \ldots, y_T)$ とする．

$$P(C \mid \boldsymbol{X}, \boldsymbol{\Lambda}) = \frac{1}{Z} \sum_{\boldsymbol{Y}} \exp\left\{ \sum_{t=1}^{T+1} \sum_{i=0}^{M+1} \sum_{j=0}^{M+1} \lambda_{i,j}^{(a)} f_{i,j}^{(a)}(y_{t-1}, y_t) + \sum_{t=1}^{T} \sum_{l=1}^{M} \sum_{n=0}^{N} \sum_{d=1}^{D} \lambda_{l,n,d}^{(b)} f_{l,n}^{(b)}(x_{t,d}, y_t) \right\} \quad (1)$$

ただし，$\boldsymbol{\Lambda}$ は素性関数の重みを表すモデルパラメータ，Z は正規化項，T は時刻，M は状態数，N は特徴量の次数の最大値，D は次元数である．また，$\boldsymbol{x}_t = (x_{t,1}, x_{t,2} \ldots, x_{t,D})^{\mathsf{T}}$ は時刻 t に観測された D 次元の特徴量，y_t は $y_t \in \{0, 1, \ldots, M+1\}$ であり，y_0 は初期状態 0, y_{T+1} は終了状態 $M+1$ を表す．

状態遷移に対応する素性関数 $f_{i,j}^{(a)}$ は，以下のように定義される．

$$f_{i,j}^{(a)}(y_{t-1}, y_t) = \begin{cases} 1 & ((y_{t-1}, y_t) = (i, j)) \\ 0 & (\text{otherwise}) \end{cases} \quad (2)$$

この関数は，y_{t-1} と y_t が遷移を起こすときのみ 1 となることを意味している．このとき，素性関数の重み $\lambda_{i,j}^{(a)}$ は HMM における状態遷移確率に対応している．また，出力に対応する素性関数 $f_{l,n}^{(b)}$ は，以下のように定義される．

$$f_{l,n}^{(b)}(x_{t,d}, y_t) = \begin{cases} x_{t,d}^{n} & (y_t = l) \\ 0 & (\text{otherwise}) \end{cases} \quad (3)$$

この関数は，y_t に対応する特徴量を出力することを意味している．ここで，n は任意の次数を設定でき，様々な出力分布を表現可能である．このとき，素性関数の重み $\lambda_{l,n,d}^{(b)}$ により，同じ特徴量を出力する場合でも，次元毎，状態毎に異なる重みを設定できるため，特徴選択を行うことが可能である．

HCRF の学習は，対数事後確率 $\mathcal{L}(\boldsymbol{\Lambda}) = \log P(C \mid \boldsymbol{X}, \boldsymbol{\Lambda})$ に L1 ノルムを用いた評価関数 $\mathcal{L}_1(\boldsymbol{\Lambda})$ を最大とするモデルパラメータ $\boldsymbol{\Lambda}$ を推定する．なお，L1 ノルムを $||\cdot||_1$ とする．

$$\mathcal{L}_1(\boldsymbol{\Lambda}) = \log P(C \mid \boldsymbol{X}, \boldsymbol{\Lambda}) - \Omega||\boldsymbol{\Lambda}||_1 \quad (4)$$

ただし，Ω は L1 ノルムの重みであり，値が大きいほどモデルパラメータである素性関数の重みが 0 付近で急峻な分布となる．よって，弱い素性の重みは 0 になるので，スパースなモデルを構築できる．また，Ω により素性関数の重みの 0 になりやすさを調整する．以上から，L1 ノルムを用いることで，パラメータ推定と特徴選択を同時に行うことが可能である．

しかし，L1 ノルムはモデルパラメータが 0 のときに微分が不可能であり，一般的な勾配法を用いてパラメータ推定を行うことができない．このような問題に対し，勾配法の一種である Rprop アルゴリズムにおいて，L1 ノルムを使用可能にした Orthantwise-Rprop アルゴリズムが提案されている [2]. このアルゴリズムでは，パラメータ更新を同じ象限で行うように制限し，pseudo gradient と呼ばれる勾配を用いる．パラメータ更新式および勾配は，以下のようになる．

$$\lambda^{(r+1)} = \lambda^{(r)} + \mathrm{sgn}(\nabla\mathcal{L}(\lambda^{(r)}))\eta^{(r)} \quad (5)$$

$$\nabla\mathcal{L}(\lambda^{(r)}) = \begin{cases} \partial^{+}\mathcal{L}(\lambda^{(r)}) & (\partial^{+}\mathcal{L}(\lambda^{(r)}) > 0) \\ \partial^{-}\mathcal{L}(\lambda^{(r)}) & (\partial^{-}\mathcal{L}(\lambda^{(r)}) < 0) \\ 0 & (\text{otherwise}) \end{cases} \quad (6)$$

$$\partial^{\pm}\mathcal{L}(\lambda^{(r)}) = \frac{\partial\mathcal{L}(\lambda^{(r)})}{\partial\lambda^{(r)}} - \begin{cases} \mathrm{sgn}(\lambda^{(r)})\Omega & (\lambda^{(r)} \neq 0) \\ \pm\Omega & (\lambda^{(r)} = 0) \end{cases} \quad (7)$$

ただし，r は学習回数，$\lambda^{(r)}$, $\lambda^{(r+1)}$ は更新前，更新後のモデルパラメータ，$\nabla\mathcal{L}(\lambda^{(r)})$ は勾配，$\eta^{(r)}$ はステップサイズである．また，$\mathrm{sgn}(\cdot)$ は符号を表し，$\mathrm{sgn}(\cdot) \in \{-1, 0, 1\}$ である．式 (6), 式 (7) では，勾配を計算するのにモデルパラメータが 0 かどうかで場合分けしており，モデルパラメータが 0 のときに微分不可能である問題を対処している．HCRF の学習の手順は，勾配を Forward-Backward アルゴリズムを用いて計算し，式 (5) で示されるパラメータ更新を繰り返すため，EM アルゴリズムとほぼ同様な手順で行われる．

3　アニーリング制御を適用した HCRF の学習

Orthantwise-Rprop アルゴリズムのような勾配型アルゴリズムでは，局所最適性が問題として挙げられる．そのため，HCRF の初期値が適切に与えられなかった場合，素性関数の重みの推定精度が低下し，特徴選択の効果が十分に発揮できない可能性がある．このような問題を改善する手法として，HCRF の初期値に微小値を掛け，初期モデルを一様分布に近づくように平滑化を行う Flattening が提案されている [1]．この手法では初期モデルに対してのみ平滑化を行っているが，学習の初期段階では推定される重みの信頼性は低いため，学習過程においても平滑化を行うことで，より局所最適性の問題を緩和できると考えられる．一方，生成モデルの学習全体において平滑化を行う手法として，確定的アニーリング EM (Deterministic Annealing Expection Maximization; DAEM) アルゴリズムが提案されている [3]．そこで，本研究では HCRF を用いた音声認識における，アニーリング制御の適用による学習アルゴリズムの改善を提案する．

DAEM アルゴリズムは，EM アルゴリズムにおいて尤度関数を温度パラメータを導入した自由エネルギー関数として再定式化し，アニーリング過程を制御することで局所最適性の問題を改善する．同様に，HCRF の学習アルゴリズムにアニーリング制御を適用した場合，式 (1) は以下のようになる．

表 1: Flattening あり/なしにおける単語正解率

	初期モデル	Flat なし	Flat あり
39-HCRF-RP	94.66	95.36	97.45
120-HCRF-RP	79.86	86.34	95.82
120-HCRF0-RP	94.66	95.09	96.68

$$P_\theta(C \mid \boldsymbol{X}, \boldsymbol{\Lambda}) = \frac{1}{Z} \sum_{\boldsymbol{Y}} \exp\left\{ \sum_{t=1}^{T+1} \sum_{i=0}^{M+1} \sum_{j=0}^{M+1} \theta \lambda_{i,j}^{(a)} f_{i,j}^{(a)}(y_{t-1}, y_t) + \sum_{t=1}^{T} \sum_{l=1}^{M} \sum_{n=0}^{N} \sum_{d=1}^{D} \theta \lambda_{l,n,d}^{(b)} f_{l,n}^{(b)}(x_{t,d}, y_t) \right\} \quad (8)$$

ここで，$\theta \simeq 0$ のときには $P_\theta(C \mid \boldsymbol{X}, \boldsymbol{\Lambda})$ は一様に近づけた分布となり，$\theta = 1$ のときには式 (1) と一致する．よって，θ は事後確率を平滑化する効果を有すると言える．そこで，θ を 0 に近い値から 1 に緩やかに近づけることで，初期値に依存しにくいパラメータ推定を行うことが可能である．

4　評価実験

HCRF の学習におけるアニーリング制御に基づく学習アルゴリズムの有効性を確認するために，AURORA-2 データベースの clean 音声を用いて孤立単語認識実験を行った．数字単語は 11 種類，学習データは 1210 発話，テストデータは学習データに含まれない 3257 発話を用いた．特徴量は次元数による違いを比較するために，12 次元，39 次元の MFCC とパワーにそれぞれの 1 次，2 次動的特徴量を連結した計 39 次元，120 次元の特徴量を用いた．HMM はスキップなしの 3 状態 left-to-right モデルとした．比較手法は，それぞれ用いる特徴量，初期モデル，アニーリング制御を適用したかどうかが異なり，39, 120 は用いる特徴量の次元数，HCRF, HCRF0 は初期モデル，RP, DA はアニーリング制御なし，アニーリング制御ありを表している．初期モデルは，HCRF は ML 基準を用いて学習した HMM，HCRF0 は 120 次元のモデルパラメータのうち，39 次元までの低次元は 39 次元の特徴量と ML 基準を用いて学習した HMM，残りの高次元は全て 0 にしたもので，高次元の特徴量は認識に有効でないという事前情報を用いている．また，アニーリング制御における温度スケジュールは以下のように設定した．

$$\theta_\varepsilon(r) = \left(\frac{r}{100}\right)^{2^\varepsilon} (1 - \theta^{init}) + \theta^{init} \quad (9)$$

これは，初期値が θ^{init} で，100 回学習すると 1 になることを意味している．なお，学習回数が 100 回以上の場合は θ は 1 になるように設定した．また，ε は温度スケジュールの調整を行うパラメータである．今回の評価実験では，Orthantwise-Rprop アルゴリズム，Flattening, L1 ノルムの重み，温度スケジュールにおける調整パラメータを変化させて，最も単語正解率が高いものおよびそのときのパラメータ数を評価した．

Flattening の有効性を確認するために，Flattening あり / なしにおける比較を行った．表 1 に Flattening あり/なしにおける単語正解率を示す．なお，120-HCRF0-RP は，学習基準と学習アルゴリズムの影響を切り分けるために用いた．120-HCRF0-RP の認識率が 39-HCRF-RP より下がれば学習基準，同等であれば学習アルゴリズムに問題があることが分かる．表 1 より，どの手法も初期モデルより認識性能が改善できていることから，識別モデルの有効性を確認できる．また，Flattening を用いることによって認識性能が改善できていることから，Flattening による平滑化の有効性を確認できる．しかし，120-HCRF-RP は 39-HCRF-RP の認識率より低くなっているため，高次元の特徴選択が不十分であることが分かる．さらに，120-HCRF0-RP は，認識率が 39-HCRF-RP より低く 120-HCRF-RP より高いことから，学習基準，学習アルゴリズム共に改善が必要であると考えられる．

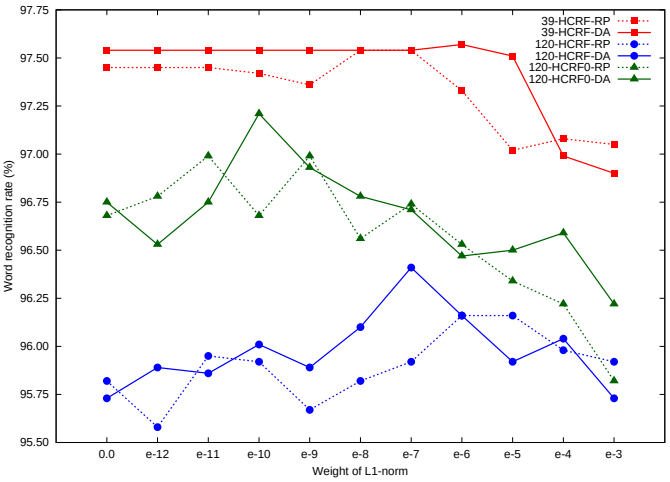

図 1: アニーリング制御あり，なしにおける単語正解率

表 2: 最高認識率のときのモデルパラメータ数

	初期	RP	DA
39-HCRF	2684 (94.66)	2646 (97.54)	2673 (97.57)
120-HCRF	8030 (79.86)	4942 (96.16)	4999 (96.41)
120-HCRF0	2684 (94.66)	4613 (96.99)	4710 (97.21)

学習基準，学習アルゴリズムを改善するために，L1 ノルムおよびアニーリング制御を適用した．図 1 に L1 ノルムの重みを変化させたときの単語正解率，表 2 に最高認識率のときのモデルパラメータ数を示す． 図 1，表 2 より，L1 ノルムを用いると，39 次元の特徴量を使用した手法は認識性能を改善できていないが，120 次元の特徴量を使用した手法は認識性能を改善できていることから，L1 ノルムによる特徴選択の改善を確認できる．120-HCRF-RP, 120-HCRF-DA は，どちらもモデルパラメータ数を約 4 割削減できているが，120-HCRF-DA の方が認識率が高いことから，アニーリング制御により局所最適性の問題をさらに改善することで，より特徴選択を改善できたことが確認できる．また，39-HCRF-DA と 120-HCRF0-DA の認識率の差より，120-HCRF-DA と 120-HCRF0-DA の認識率の差の方が大きくなったことから，学習アルゴリズムの改善による影響が大きいことが分かる．以上から，アニーリング制御を適用した学習アルゴリズムの有効性を確認できる．しかし，120-HCRF-DA, 120-HCRF0-DA は，39-HCRF-DA より低い認識率であるため，さらなる学習アルゴリズムや学習基準の改善が必要であると考えられる．

5　むすび

本研究では HCRF を用いた音声認識における，アニーリング制御の適用による学習アルゴリズムの改善を提案した．孤立単語認識実験により，特徴選択および認識性能の改善を確認した．今後の課題として，よりモデルに適した温度スケジュールの調査や連続音声認識実験による評価が挙げられる．

参考文献

[1] M. Mahajan, *et al.*, "Training Algorithms for Hidden Conditional Random Fields," Proc. of ICASSP, pp. 273–276, 2006.

[2] S. Wiesler, *et al.*, "Feature Selection for Log-Linear Acoustic Models," Proc. of ICASSP, pp. 5324–5327, 2011.

[3] N. Ueda and R. Nakano, "Deterministic Annealing EM Algorithm," Neural Networks, vol. 11, pp. 271–282, 1998.