Vol.2010-SLP-83 No.1
2010/10/29

Hidden Conditional Neural Fields を用いた音声認識の検討

藤井康寿[†1]　山本一公[†1]　中川聖一[†1]

近年，識別モデルを用いた音声認識手法が注目を集めている．特に，Hidden Conditional Randam Fields(HCRF) を用いた音声認識手法は，HMM の自然な拡張となっており，今後の発展が期待できる．HCRF は有望なモデルであるが，仮説のスコアを特徴量の重み付き線形和によって計算するため，特徴量間の非線形な関係をうまくモデル化できないという問題があった．本稿では，HCRF にゲート関数を導入することで，特徴量間の非線形な関係をモデル化することができるように拡張した Hidden Conditional Neural Fields(HCNF) を用いた音声認識手法を提案する．HCNF は，一切の初期モデルを必要とせずに学習することが可能であり，種々の特徴量を使用することも容易である．TIMIT コーパスにおける core テストセット上での monophone を用いた音素認識実験の結果，HCNF による認識結果は，HCRF および，MPE 学習した HMM による認識結果よりもよく，提案法の有効性を示すことができた．

A Study of Automatic Speech Recognition using Hidden Conditionan Neural Fields

Yasuhisa Fujii,[†1] Kazumasa Yamamoto[†1]
and Seiichi Nakagawa [†1]

Recently, there has been increasing attention in automatic speech recognition using discriminative models. Especially, Hidden Conditional Random Fields(HCRF) is a natural extension of traditional HMM and therefore very promising. However, because HCRF computes the score of a hypothesis by summing up linearly weighted feature values, it cannot consider non-linearity between feature values that will be very crucial for speech recognition. In this paper, we extend HCRF by incorporating gate function used in neural networks and propose a new model called Hidden Conditional Neural Fields(HCNF). Differently with conventional approaches, HCNF can be trained without any initial model and incorporate any kinds of features. Experimental results of continuous phoneme recognition on TIMIT core test set using monophone showed that HCNF was superior to HCRF and HMM trained in MPE manner.

1. はじめに

現在，最先端の音声認識システムにおいては，音響モデルとして出力確率に GMM を用いた HMM が用いられることが一般的である．しかし，HMM は状態が与えられた上でのフレーム間の特徴量の独立性を仮定しているために，数フレームにまたがる特徴を十分に考慮できないことや，生成モデルであるがゆえに，識別性能が低いことがこれまで指摘されてきた．前者の問題を解決するために Δ 特徴量[1] やセグメント統計量[2]，変調スペクトルを用いた特徴抽出法[3] などが考案され，また，後者の問題を解決するために識別学習の研究が行われてきた[4]．

長時間の特徴の変化を考慮することができ，かつ識別的能力が高いモデルを HMM と組み合わせることで，これらの問題を解決する試みも行われてきた．例えば，Tandem システムは，Multi Layered Perceptron(MLP) を用いて予め音響特徴量系列を音素および弁別特徴の事後確率に変換し，これを HMM の入力とするモデルであり，MLP の特徴抽出能力および識別能力を HMM ベースの音声認識システムに組み込むことを可能にしている[5]．特徴抽出に MLP ではなく Conditional Random Fields(CRF)[6] を使用する試みもある[7]．

さらに最近では，計算機性能の向上も相まって，これまで実現が難しかった HMM を識別モデルで完全に置き換える試みが行われるようになってきた．特に，Hidden Conditional Randam Fields(HCRF) を用いた音声認識手法は，従来用いられてきた HMM の自然な拡張となっており，今後の発展が期待できる[8)–10]．HCRF は有望なモデルであるが，仮説のスコアを特徴量の重み付き線形和によって計算するため，特徴量間の非線形な関係をうまくモデル化できないという問題がある．明示的に特徴量を拡張することで SVM で使用されるカーネル関数を模擬する試みもあるが[11]，特徴量の次元数が増えた場合に問題が生じる．

Peng らは，CRF に対してゲート関数を導入し，特徴量間の非線形性を考慮できる枠組みである Conditional Neural Fields(CNF) を提案した[12]．CNF の考え方を HCRF に導入することで，非線形な関係をうまくモデル化できないという HCRF の欠点を補えると考えられる．本稿では，HCRF に MLP で使用されるようなゲート関数を導入することで，特徴量間の非線形な関係をモデル化することができるように拡張した Hidden Conditional Neural

†1 豊橋技術科学大学知能・情報工学系
Department of Computer Science and Engineering, Toyohashi University of Technology

Fields(HCNF) を用いた音声認識手法を提案する．HCRF に非線形性を導入したことで，数フレームにまたがる特徴を考慮でき，かつ識別能力の高いモデルとなることが期待できる．

2. Hidden Conditional Random Fields を用いた音声認識

2.1 定 式 化

HCRF を用いた音声認識では，音響特徴量の系列 $X = (x_1, x_2, \ldots, x_T)$ が与えられた上で，対応するラベル列が $Y = (y_1, y_2, \ldots, y_T)$ である確率を以下のように計算する．

$$P(Y|X) = \frac{1}{Z(X)} \sum_S \exp^{\sum_k \lambda_k F_k(X,Y,S)} \tag{1}$$

ここで $Z(X)$ は確率の総和を 1 にするための正規化項である．式 (1) 中の S は隠れ状態系列を表し，$\sum_S$ によって周辺化される．これは，HMM における隠れ状態の扱いと同等である．λ_k は特徴量 $F_k(X,Y,S)$ に対する重みを意味し，各特徴量の重要度を表す．特徴量 $F_k(X,Y,S)$ は音響特徴量系列 X，ラベル列 Y，隠れ状態系列 S の 3 つ組を引数にとる特徴関数として定義され，本来は系列間のいかなる関係性も定義できるが，効率的な学習および推論を可能にするために，通常以下の 2 つの特徴関数に限定する．

$$\Phi(X,Y,S) = \sum_k \Phi_k(X,Y,S) = \sum_k w_k \sum_t \phi_k(X, y_t, s_t) \tag{2}$$

$$\Psi(X,Y,S) = \sum_j \Psi_j(X,Y,S) = \sum_j u_j \sum_t \psi_j(X, y_t, y_{t-1}, s_t, s_{t-1}) \tag{3}$$

式 (2) で表される特徴を観測特徴，式 (3) で表される特徴を遷移特徴と呼ぶ．すなわち，観測特徴においては特徴抽出をフレーム t におけるラベル y_t および隠れ状態 s_t と観測特徴量系列 X からに限定し，遷移特徴においては特徴抽出を t および $t-1$ におけるラベル列および隠れ状態と観測特徴量系列 X からに限定する．例えば，観測特徴は MFCC や PLP などの音響特徴量の各状態に対するスコアとして定義され，遷移特徴は状態間の遷移スコアとして定義される．観測特徴および遷移特徴を用いると，式 (1) は以下のようになる．

$$\begin{aligned} P(Y|X) &= \frac{1}{Z(X)} \sum_S \exp^{\Phi(X,Y,S)+\Psi(X,Y,S)} \\ &= \frac{1}{Z(X)} \sum_S \exp^{\sum_t \left\{ \sum_k w_k \phi_k(X,y_t,s_t) + \sum_j u_j \psi_j(X,y_t,y_{t-1},s_t,s_{t-1}) \right\}} \end{aligned} \tag{4}$$

ここで，$Z(X)$ は以下のように定義される．

$$Z(X) = \sum_Y \sum_S \exp^{\sum_t \left\{ \sum_k w_k \phi_k(X,y_t,s_t) + \sum_j u_j \psi_j(X,y_t,y_{t-1},s_t,s_{t-1}) \right\}} \tag{5}$$

このように特徴量を限定することで，Forward-Backward アルゴリズムや，Viterbi アルゴリズムといった従来 HMM で用いられる効率的な計算手法を HCRF にも適用することができるようになる．

2.2 学　　習

学習データ $D = \{X^i, Y^i\}, i = 0, \ldots, N$ が与えられたとき，目的関数をパラメータ $\lambda = \{w_k, u_j\}$ における各データに対する負の対数確率の和として以下のように設定する．

$$\begin{aligned} l(\lambda; D) &= -\sum_i \log P(Y^i|X^i) \\ &= \sum_i \left\{ -\log \sum_S \exp^{\sum_t \sum_k w_k \phi_k(X,y_t^i,s_t) + \sum_j u_j \psi_j(X,y_t^i,y_{t-1}^i,s_t,s_{t-1})} + \log Z(X^i) \right\} \end{aligned} \tag{6}$$

パラメータ推定は，$l(\lambda; D)$ を最小化する λ を発見する最適化問題として定式化できる．$l(\lambda; D)$ の $\{w_k, u_j\}$ による偏微分を求めることができれば，勾配法や L-BFGS のような準ニュートン法を用いて最適化問題を解くことが可能になる．$l(\lambda; D)$ の w_k による偏微分は以下のように計算できる．

$$\begin{aligned} \frac{\partial l(\lambda; D)}{\partial w_k} &= -\sum_i \frac{\sum_S \exp^{\sum_k w_k \phi_k(X^i,y_t^i,s_t) + \sum_j u_j \psi_j(X,y_t^i,y_{t-1}^i,s_t,s_{t-1})} \phi_k(X^i, y_t^i, s_t)}{\sum_S \exp^{\sum_k w_k \phi_k(X^i,y_t^i,s_t) + \sum_j u_j \psi_j(X^i,y_t^i,y_{t-1}^i,s_t,s_{t-1})}} \\ &\quad + \sum_i \frac{\sum_Y \sum_S \exp^{\sum_k w_k \phi_k(X^i,y_t,s_t) + \sum_j u_j \psi_j(X^i,y_t,y_{t-1},s_t,s_{t-1})} \phi_k(X^i, y_t, s_t)}{Z(X^i)} \\ &= -\sum_i \sum_S P(S|Y^i, X^i) \phi_k(X^i, y_t^i, s_t) + \sum_i \sum_Y \sum_S P(Y, S|X^i) \phi_k(X^i, y_t, s_t) \\ &= -\sum_i E\left[\phi_k(X^i, y_t^i, s_t)\right]_{S|Y^i,X^i} + \sum_i E\left[\phi_k(X^i, y_t, s_t)\right]_{Y,S|X^i} \end{aligned} \tag{7}$$

ここで，$E[]_X$ は確率変数 X による期待値を表す．u_j に関する偏微分は同様に導出できるため省略する．

2.3 推　　論

文献 10) では，時間と状態で定義されるトレリス空間上において，各点に到達しうる系列を上位 N 個保持することで式 (4) における隠れ状態 S を周辺化しながら X が与えられた上で最尤の Y を求めるアルゴリズムを採用している．しかし，このアルゴリズムは，既存のデコーダに組み込むことが容易でなく，HCRF によるモデルを一種の音響モデルとみなして大語彙連続音声認識への発展を考える上では望ましくない．そのため，本稿では，HMM

を用いた音声認識で通常用いられる Viterbi アルゴリズムによって隠れ状態 S を周辺化することなく最尤の系列 S を求めることで Y を推論するアルゴリズムを採用する．

3. Hidden Conditional Neural Fields を用いた音声認識

3.1 定 式 化

HCNF は Peng らによって提案された Conditional Neural Fiedls(CNF)[12] を参考に，HCRF にゲート関数を導入することで，HCRF では捉えることのできない特徴量間の非線形な関係を考慮できるように HCRF を拡張したものである．HCRF と HCNF の構造を図 1 に示す．HCRF において式 (2) および (3) で定義された観測特徴および遷移特徴を HCNF では以下のように定義する．

$$\Phi(X,Y,S)=\sum_t\sum_g^K w_{y_t,s_t,g}h(\theta^T_{y_t,s_t,g}\phi(X,y_t,s_t)) \tag{8}$$

$$\Psi(X,Y,S)=\sum_j u_j\sum_t \psi_j(X,y_t,y_{t-1},s_t,s_{t-1}) \tag{9}$$

ここで，$\phi(X,y,s)$ は式 (2) の $\phi_k(X,y_t,s_t)$ をベクトルとして並べたもの，$\theta_{y,s,g}$ はこれに対応する重みベクトルである．$h(x)$ は以下に示すゲート関数である．

$$h(x)=\frac{1}{1+\exp(-x)}-0.5 \tag{10}$$

式 (8) が観測特徴，(9) が遷移特徴である．式 (3)，(9) からわかるように，遷移特徴は HCRF と HCNF で同一のものを用いる．一方で，観測特徴に K 個のゲート関数を導入し，特徴量間の非線形な関係をモデル化している．式 (8)，(9) を用いて，音響特徴量の系列 X が与えられた上で，対応するラベル列が Y である確率を以下のように計算する．

$$\begin{aligned}P(Y|X)&=\frac{1}{Z(X)}\sum_S \exp^{\Phi(X,Y,S)+\Psi(X,Y,S)}\\&=\frac{1}{Z(X)}\sum_S \exp^{\sum_t\sum_g^K w_{y_t,s_t,g}h(\theta^T_{y_t,s_t,g}\phi(X,y_t,s_t))+\sum_j u_j\psi_j(X,y_t,y_{t-1},s_t,s_{t-1})}\end{aligned} \tag{11}$$

ここで，$Z(X)$ は正規化項であり以下の様に定義される．

$$Z(X)=\sum_{Y'}\sum_S \exp^{\sum_t\sum_g^K w_{y'_t,s_t,g}h(\theta^T_{y'_t,s_t,g}\phi(X,y'_t,s_t))+\sum_j u_j\psi_j(X,y'_t,y'_{t-1},s_t,s_{t-1})} \tag{12}$$

HCRF と同様に，特徴量をフレーム t および $t-1$ のラベルから得られるものに限定しているため，Forward-Backward アルゴリズムや，Viterbi アルゴリズムといった効率的な計算手法を HCNF にも適用することができる．

3.1.1 学　　習

2.2 節に示した HCRF の学習と同様に，学習データ $D=\{X^i,Y^i\}, i=0,\ldots,N$ が与えられたとき，目的関数を $\lambda=\{w_{s,g},\theta_{s,g},\psi_j\}$ における各データに対する負の対数確率の和として設定する．

$$\begin{aligned}l(\lambda;D)&=-\sum_i \log P(Y^i|X^i)\\&=-\sum_i \log\sum_S \exp^{\sum_t\sum_g^K w_{y^i_t,s_t,g}h(\theta^T_{y^i_t,s_t,g}\phi(X,y^i_t,s_t))+\sum_j u_j\psi_j(X,y^i_t,y^i_{t-1},s_t,s_{t-1})}\\&\quad+\sum_i \log Z(X^i)\end{aligned} \tag{13}$$

パラメータ推定は，$l(\lambda;D)$ を最小化する $\{w_{y_t,s_t,g},\theta_{y_t,s_t,g},u_j\}$ を発見する最適化問題として定式化できる．$l(\lambda;D)$ の $\{w_{y_t,s_t,g},\theta_{y_t,s_t,g},u_j\}$ による偏微分を求めることができれば，勾配法や L-BFGS のような準ニュートン法を用いて最適化問題を解くことが可能になる．$l(\lambda;D)$ の $w_{y_t,s_t,g}$ および $\theta_{y_t,s_t,g}$ による偏微分は以下のように計算できる．

$$\begin{aligned}\frac{\partial l(\lambda;D)}{\partial w_{y,s,g}}&=-\sum_i\frac{\sum_S \exp^{\Phi(X^i,Y^i,S)+\Psi(X^i,Y^i,S)}\sum_t h(\theta^T_{y^i_t,s_t,g}\phi(X^i,y^i_t,s_t))\delta[s_t=s]}{\sum_S \exp^{\Phi(X^i,Y^i,S)+\Psi(X^i,Y^i,S)}}\\&\quad+\sum_i\frac{\sum_Y\sum_S \exp^{\Phi(X^i,Y,S)+\Psi(X^i,Y,S)}\sum_t h(\theta^T_{y_t,s_t,g}\phi(X^i,y_t,s_t))\delta[s_t=s]}{Z(X^i)}\\&=-\sum_i E\left[\sum_t h(\theta^T_{y^i_t,s_t,g}\phi(X^i,y^i_t,s_t))\delta[s_t=s]\right]_{S|X^i,Y^i}\\&\quad+\sum_i E\left[h(\theta^T_{y_t,s_t,g}\phi(X^i,y_t,s_t))\delta[s_t=s]\right]_{Y,S|X^i}\end{aligned} \tag{14}$$

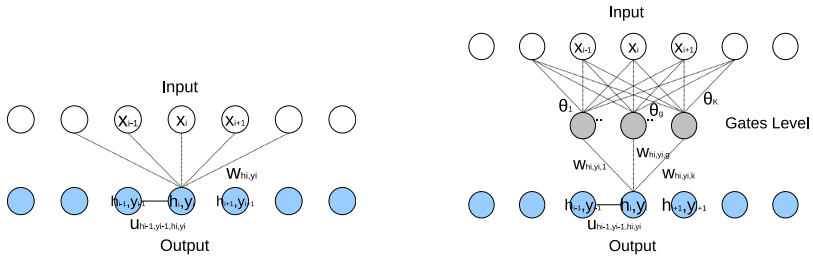
$$\begin{aligned}\frac{\partial l(\lambda;D)}{\partial \theta_{y,s,g}}&=-\sum_i E\left[\sum_t w_{y^i_t,s_t,g}\frac{\partial h(\theta^T_{y^i_t,s_t,g}\phi(X^i,y^i_t,s_t))}{\partial\theta_{y_t,s_t,g}}\delta[s_t=s]\right]_{S|X^i,Y^i}\\&\quad+\sum_i E\left[\sum_t w_{y_t,s_t,g}\frac{\partial h(\theta^T_{y_t,s_t,g}\phi(X^i,y_t,s_t))}{\partial\theta_{y_t,s_t,g}}\delta[s_t=s]\right]_{Y,S|X^i}\end{aligned} \tag{15}$$

式 (10) の微分は以下のように計算できる．

$$\frac{dh(x)}{dx}=(0.5+h(x))(0.5-h(x)) \tag{16}$$

u_j による偏微分は式 (7) と同様に導出できるため，省略する．

Vol.2010-SLP-83 No.1
2010/10/29



(a) HCRF (b) HCNF

図 1 HCRF と HCNF の構造

3.1.2 推　　論

HCNF の推論は，HCRF の場合と同様に Viterbi アルゴリズムによって行う．すなわち，隠れ状態 S を周辺化することなく最尤の系列 S を求めることで Y を推論する．

4. 学習手法の検討

4.1 正　則　化

HCRF の学習において，正則化が有効であることが知られている[9]．本稿では，HCRF，HCNF ともに正則化を行う．$l(\lambda; D)$ を元の目的関数とするとき，正則化付きの目的関数 $f(\lambda; D)$ は以下のようになる．

$$f(\lambda; D) = l(\lambda; D) + r(\lambda) \tag{17}$$

本稿では，以下に示す L1 正則化と L2 正則化を使用する．

L1 正則化

$$r(\lambda) = C||\lambda||_1 = C\sum_i |\lambda_i| \tag{18}$$

L2 正則化

$$r(\lambda) = C||\lambda||_2 = \frac{C}{2}\sum_i \lambda_i^2 \tag{19}$$

ここで C は正則化を考慮する度合いを表すパラメータである．L1 正則化を使用する場合には，零点における微分が計算できないため工夫が必要になる．

4.2 最適化アルゴリズム

本稿では，HCRF および HCNF の学習のために，勾配に基づく学習手法を採用する．勾配に基づく学習手法は，学習データ全体から勾配を求めて使用する通常のバッチ型学習と，学習データ全体のサブセットから勾配を求めて使用するオンライン型学習がある．文献 8) では，バッチ型学習手法として，準ニュートン法の一種である L-BFGS[13]，オンライン型学習手法として Stochastic Gradient Descent(SGD) を使用して両者を比較し，SGD による学習結果の方がやや良いという結果を得ていた．しかし，文献 8) は音素分類での実験結果であり，連続音声認識において両者を比較していない．そのため，本稿でも同様に L-BFGS による学習と SGD による学習を比較する．

L-BFGS は，パラメータ更新のために計算した勾配を過去数回分利用してヘッセ行列を推定し，ニュートン法による最適化を行う手法である．L2 正則化を用いる場合には，L-BFGS のアルゴリズムを変更する必要はないが，L1 正則化では零点での微分を計算できないため，アルゴリズムの変更が必要になる．本稿では，L-BFGS で L1 正則化を用いた最適化を行うために，Andrew らによって提案された OWL-QN を使用する[14]．

SGD は，学習データ全体のサブセットから計算した勾配 $\boldsymbol{g}_t$ を用いてパラメータ $\boldsymbol{\lambda}$ の更新を繰り返していく．

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \eta_t \boldsymbol{g}_t \tag{20}$$

ここで η_t は学習係数である．本稿では，$\boldsymbol{g}_t$ は 1 サンプルから計算するが，サンプルを学習データからランダムに抽出する代わりに，学習データ全体をランダムに並び替え，先頭から 1 つずつ取り出して使用していく．$\#sample$ 個の学習データについて，全ての学習データを使用したことでくり返し学習を 1 回行ったこととし，くり返し学習の回数を $\#iter$ とするとき，η_t を以下のように計算する．

$$\eta_t = \frac{\#sample \cdot \#iter - t}{\#sample \cdot \#iter} \tag{21}$$

本稿では，SGD で正則化を実現するために，Duchi らによって提案された FOBOS を利用する[15]．

5. 実　　験

5.1 実 験 条 件

本稿では，TIMIT コーパスを用いた連続音素認識タスクにおいて提案法の評価を行う．文献 16) に従い，TIMIT コーパスで定義される 61 音素について，学習時には 48 音素，評価時にはさらにそこから 39 音素にマッピングして評価を行った．学習データは 3696 文 (462 話者) からなり，テストデータには 192 文 (24 話者) の core テストセットを使用した．全

Vol.2010-SLP-83 No.1
2010/10/29

てのモデルにおいて，各音素について 3 状態を持つ left-to-right 型の monophone モデルを使用した．使用した特徴量は，観測特徴として各状態につき MFCC13 次元[*1]およびその Δ と $\Delta\Delta$，これらの値の 2 乗値，状態のユニグラム特徴でフレームあたり計 79 次元を使用し，遷移特徴として現在の状態と直前の状態とのペアで活性化する状態のバイグラム特徴を使用した．HCRF および HCNF で使用する特徴量は，学習データ全体の平均が 0，分散が 1 となるように正規化し，フレーム毎に抽出した計 79 次元の特徴量を，該当フレームに加えて前後 4 フレーム (計 9 フレーム) 結合して使用した．HCRF のパラメータは全て 0 で初期化し，HCNF のパラメータは-0.5 から 0.5 の間でランダムに初期化した．正則化パラメータ C は全て 1.0 を使用した．

比較のため HMM による認識も行った．HMM は，HTK を使用して対角共分散，32 混合のモデルを MLE 学習し，さらに，MMI，MPE 学習を 10 回行った．I-smoothing は 100，学習係数は 2.0 に設定した．使用した特徴量は MFCC13 次元+Δ+$\Delta\Delta$ の計 39 次元である．また，音素バイグラムを学習データから学習して使用した．

5.2 実験結果

HCRF および HCNF を最適化アルゴリズムおよび正則化手法を変えて学習した場合の実験結果を表 1 に示す．HCNF のゲート数は全て 4 である (式 (8) における $K = 4$)．最適化アルゴリズムに L-BFGS を用いる場合，HCRF，HCNF ともに L1 以外の正則化を用いたときには，学習の途中で目的関数の値が減少するパラメータを発見できずに，早期に学習が終了してしまった．これは，HCRF は隠れ状態を含んでおり，目的関数の形が凹ではないため，常に目的関数の値が減少する方向にパラメータを更新していく L-BFGS の学習方法ではうまく学習できないためと思われる．HCNF ではゲート関数によって HCRF よりもさらに目的関数の性質が悪くなるため，L1 正則化以外の場合にはほとんど学習ができないという結果になった．文献 8) において，L1 正則化を使用せずに L-BFGS を用いた学習が行えているが，文献 8) は音素分類の学習であり，本稿での連続音素認識の学習はより難しい問題である．今回の実験において，L1 正則化を用いた場合にうまく学習ができたのは，L1 正則化の性質によるものか，OWL-QN の特性によるものかについては調査が必要である．

一方で，最適化アルゴリズムに SGD を用いる場合には，正則化手法にかかわらず HCRF，HCNF ともにうまく学習ができていることがわかる．HCRF，HCNF ともに正則化を使用した方が 1 ～ 2% 程度正則化を使用しない場合に比べて認識率が良くなっており，正則化が重要であることがわかる．また，SGD は L-BFGS を使用した場合よりもかなり少ない学習回数で済んでおり，HCRF および HCNF を学習するために適した方法であるといえる．

表 1 学習手法の比較

モデル	最適化	正則化	Del.	Ins.	Subs.	Cor.	Acc.	PER	#iter(max)
HCRF	L-BFGS	なし	9.3	2.3	21.2	69.5	67.2	32.8	71(200)
		L1	8.6	2.3	19.9	71.5	69.2	30.8	200
		L2	9.1	2.2	20.3	70.6	68.3	31.7	77(200)
	SGD	なし	6.3	3.9	21.4	72.3	68.4	31.6	10
		L1	6.6	3.6	20.6	72.8	69.2	30.8	10
		L2	6.8	3.0	20.6	72.6	69.6	30.5	10
HCNF	L-BFGS	なし	97.2	0.0	0.0	2.8	2.8	97.2	12(200)
		L1	8.6	2.1	19.0	72.3	70.2	29.8	200
		L2	97.3	0.0	0.0	2.7	2.7	97.3	11(200)
	SGD	なし	6.3	3.5	20.9	72.8	69.2	30.8	30
		L1	7.8	2.6	18.3	73.9	71.3	28.7	30
		L2	7.2	2.5	19.2	73.6	71.1	28.9	30

HCRF による認識結果と HCNF による認識結果を比較すると，HCRF で最高の認識率が PER(音素誤り率)=30.5 だったのに対し，HCNF では PER=28.7 となっており，絶対値で 1.8%よくなっている．次節に示す通り，HCNF では最高で PER=27.9 を達成しており，HCRF に対して絶対値で 2.6%よい．HCRF にゲート関数を導入することでモデルの表現力が向上し，認識率が向上しているといえる．

5.3 従来法との比較

提案手法と従来法を比較した結果を表 2 に示す．表 1 において最高の認識率となったのは HCNF を SGD で L1 正則化を使用して学習した場合の PER=28.7 であったが，SGD で L2 正則化を使用する学習において文献 17) に示される state-flattening を $\kappa = 0.2$ で使用した場合，PER=27.9 を得た．この結果は，MLE，MMI，MPE 学習した HMM による認識結果よりもよく，また，本稿の実験条件と似ている文献 10)，18)，19) の結果と比較しても上回るかあるいは遜色ない結果となっており，提案法の有効性を示している．

6. おわりに

本稿では，HCRF にゲート関数を導入することで非線形な特徴量の組み合わせを考慮できる枠組みである HCNF による音声認識について述べた．TIMIT コーパスの core テストセットを用いた monophone での連続音素認識実験の結果，HCNF による認識結果はゲート関数を導入しない HCRF および MLE，MMI，MPE 学習した HMM による認識結果よ

⋆1 サンプリング周波数=16kHz，プリエンファシス=0.97，分析窓長=25ms，フレームシフト=10ms

表 2 従来法との比較結果

条件	Del.	Ins.	Subs.	Cor.	Acc.	PER
MLE-HMM(diag,32mix)	9.2	2.6	18.2	72.6	70.0	30.0
MMI-HMM(diag,32mix)	7.9	2.8	18.0	74.1	71.3	28.7
MPE-HMM(diag,32mix)	9.1	2.2	17.1	73.8	71.6	28.4
HCRF(SGD,L2)	6.8	3.0	20.6	72.6	69.6	30.5
HCNF(SGD,L1)	7.8	2.6	18.3	73.9	71.3	28.7
HCNF(SGD,L2,$\kappa = 0.2$)	9.2	1.5	17.3	73.6	72.1	27.9
HCRF(32mix)[10]	7.2	3.6	17.5	75.3	71.7	28.3
LM-HMM(8mix)[18]	-	-	-	-	71.8	28.2
DM(FullCov,32mix)[19]	-	-	-	-	72.2	27.8

りもよく，先行研究と比べても上回るか遜色ない結果となった．HCNF は初期モデルを必要とすることなく学習できるため，HMM を用いた音声認識手法に捉われない自由な特徴設計が可能である．

今後の課題としては，より大きなコーパスを用いた実験，大語彙連続音声認識への適用，さらにコンテキスト依存モデルの導入などが考えられる．また，本稿ではゲート関数として sigmoid 型の関数を使用したが，MLP の性能は使用するゲート関数に依存するという報告があり[20]，HCNF で使用するゲート関数の影響についても検討していきたい．

謝　　辞

本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援を受けた。

参 考 文 献

1) Furui, S.: Speaker-independent isolated word recognition using dynamic features of speech spectrum, *IEEE Transactions of Acoustics Speech and Signal Processing*, Vol.34, No.1, pp.52 – 59 (1986).
2) 山本一公中川聖一：セグメント統計量を用いた隠れマルコフモデルによる音声認識，電子情報通信学会論文誌， Vol.J79-D-II, No.12, pp.2032–2038 (1996).
3) Kanedera, N., Arai, T., Hermansky, H. and Pavel, M.: On the Relative Importance of Various Components of the Modulation Spectrum for Automatic Speech Recognition, *Speech Communication*, Vol.28, pp.43–55 (1999).
4) Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition, PhD Thesis, Cambridge University Engineering Dept (2003).
5) Hermansky, H., Ellis, D. and Sharma, S.: Tandem connectionist feature stream extraction for conventional HMM systems, *Proc. ICASSP* (2000).
6) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the 18th International Conference on Machine Learning* (2001).
7) Fosler.-L., E. and Morris, J.: Crandem systems: Conditional Random Field Acoustic Models for Hidden Markov Models, *Proc. ICASSP*, pp.4049–4052 (2008).
8) Gunawardana, A., Mahajan, M., Acero, A. and Platt, J.: Hidden Conditional Random Fields for Phone Classification, *Proc. Interspeech*, pp.1117 – 1120 (2005).
9) Sung, Y.-H., Boulis, C., Manning, C. and Jurafsky, D.: Regularization, Adaptation, and Non-Independent Features Improve Hidden Conditional Random Fields For Phone Classification, *Proc. ASRU*, pp.347–352 (2007).
10) Sung, Y.-H. and Jurafsky, D.: Hidden Conditional Random Fields for Phone Recognition, *Proc. ASRU*, pp.107 – 112 (2009).
11) Heigold, G., Rybach, D., Schluter, R. and Ney, H.: Investigations on Convex Optimization Using Log-Linear HMMs for Digit String Recognition, *Proc. ASRU*, pp. 216–219 (2009).
12) Peng, J., Bo, L. and Xu, J.: Conditional Neural Fields, *Proc. Advances in Neural Information Processing Systems 22*, pp.1419–1427 (2009).
13) Liu, D. and Nocedal, J.: On the Limited Memory Method for Large Scale Optimization, *Mathematical Programming B*, Vol.45, No.3, pp.503–528 (1989).
14) Andrew, G. and Gao., J.: Scalable Training of L1-regularized Log-linear Models, *Proc. ICML*, pp.33–40 (2007).
15) Duchi, J. and Singer, Y.: Efficient Learning using Forward-Backward Splitting, *Proc. NIPS* (2009).
16) Lee, K.-F. and HON, H.-W.: Speaker-Independent Phone Recognition Using Hidden Markov Models, *IEEE Transactions of Acoustics Speech and Signal Processing*, Vol.37, No.11, pp.1641 – 1648 (1989).
17) Mahajan, M., Gunawardana, A. and Acero, A.: Training algorithms for hidden conditional random fields, *Proc. ICASSP*, pp.I–273–I–276 (2006).
18) Sha, F. and Saul, L.K.: Comparison of Large Margin Training to Other Discriminative Methods for Phonetic Recognition by Hidden Markov Models, *Proc. ICASSP*, pp.IV313–316 (2007).
19) Watanabe, S., Hori, T., McDermott, E. and Nakamura, A.: A discriminative model for continuous speech recognition based on weighted finite state transducers, *Proc. ICASSP*, pp.4922–4925 (2010).
20) Siniscalchi, S.M., Svendsen, T., Sorbello, F. and Lee, C.-H.: Experimental Studies On Continuous Speech Recognition Using Neural Architectures With “Adaptive” Hidden Activation Functions, *Proc. ICASSP*, pp.4882–4885 (2010).