連載解説「Deep Learning（深層学習）」〔第5回〕

音声認識のための深層学習

Deep Learning for Speech Recognition

久保 陽太郎 NTTコミュニケーション科学基礎研究所
Yotaro Kubo NTT Communication Science Laboratories.
kubo.yotaro@lab.ntt.co.jp, http://www.kecl.ntt.co.jp/icl/signal/member/yotaro/

Keywords: deep learning, speech recognition, sequential classification, hidden Markov model.

1. は じ め に

2011年，まだ30％程度のエラーが出てしまうような難しい課題であった電話会話音声の認識において，深層学習技術による音声認識システムが20％以下のエラー率を達成したとして，関連研究者を大いに驚かせた[Seide 11]．この連載のほかの解説でも触れられているが，深層学習は層数の大きな多層パーセプトロン（Multilayer Perceptron：MLP）を学習するための手段である．音声認識の研究コミュニティでは，主流とはいえないながらも層数の少ない多層パーセプトロンに関する研究も進められていたが，このように多数の層を用いたパーセプトロンが，最も難しい音声認識問題の一つである電話会話音声で有効に利用可能であるというのは，多くの研究者にとって衝撃的だったのではないかと予想される．音声認識分野における深層学習技術の適用は，2009年にDengらのグループが“Deep Learning for Speech Recognition and Related Applications”と題したワークショップをNIPSと併催で行っていることから，ほかの応用分野に先駆けていたといえる．しかし，この時点では，音声認識技術の最先端で利用されているような，大語彙で複雑な依存関係を必要とする統計モデルに直接適用できるかどうか疑問が残っていた．

画像における深層学習にネオコグニトロンのような先行事例があるように，音声における深層学習にもTime Delay Neural Network（TDNN）と呼ばれる先行事例がある[Waibel 89]．TDNNは当時ATRに所属していたAlex Waibelによって発案された，一種のConvolution Neural Networkである．現在の音声認識のように音声特徴ベクトル系列を入力とし，ラベル系列を出力する系列ラベリングのための技術ではなく，多変量系列を入力として，その信号中に含まれる子音の候補を出力するという限定的なものであった．当時のことをよく知る研究者によると，音声信号の識別能力に関してはTDNNのほうが高いこともあったが，結局N-gram言語モデルと組み合わせて利用された隠れマルコフモデル（Hidden Markov Model：HMM）[Rabiner 93]に系列ラベリングへの拡張性や単語認識時での精度などの面で及ぶことができずに，音声認識はHMMを用いて行うことが主流となったといわれている．HMMが普及した後の音声認識分野では，機械翻訳や品詞タギングなど，そのほかの系列識別問題と同様，系列のモデルをいかにつくるかに重点が置かれ，MLPのような静的な，すなわち一つの事例を一つの音素に変換するような識別器の技術は，あまり積極的には活用されてこなかった．

本稿では，このような背景にあった深層学習技術が，どのようにして音声認識技術の中に入り込んでいったのか，また深層学習の適用によってどのような前進があったのかについて，事例を紹介しながら説明していく．

2. 音声認識の基本

本章では音声認識問題に対する定式化法を述べた後，深層学習登場以前に検討された深層学習につながる重要な技術について紹介する．

2·1 音声認識問題の定式化

本稿で対象とする音声認識器は，音響特徴分析の結果をベクトルの時系列として与えられたとき，その音声中で発話されている単語列を推定するシステムである．形式的には音声認識器はベクトル系列$\boldsymbol{X} \stackrel{\text{def}}{=} \{x_1, x_2, \cdots x_t \cdots\}$を入力とし，単語系列$\boldsymbol{l} \stackrel{\text{def}}{=} \{l_1, l_2, \cdots\}$を出力するアルゴリズムとして表現される．ほかの生成モデルに基づく識別器と同様，入力を観測したうえでの出力の条件付き確率$P(\boldsymbol{l}|\boldsymbol{X})$を最大にする出力候補$\boldsymbol{l}$を認識結果$\hat{\boldsymbol{l}}$として出力する．

$$\hat{\boldsymbol{l}} \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{l}} P(\boldsymbol{l}|\boldsymbol{X}) = \arg\max_{\boldsymbol{l}} \frac{P(\boldsymbol{X}|\boldsymbol{l})P(\boldsymbol{l})}{P(\boldsymbol{X})} \quad (1)$$

ここで，単語列の生起確率を示す$P(\boldsymbol{l})$を言語モデル，単語列が与えられたうえでの観測ベクトル系列の条件付き確率を示す$P(\boldsymbol{X}|\boldsymbol{l})$を音響モデルと呼ぶ．またこの

最大化は $\boldsymbol{l}$ についてのみ行うため，$P(\boldsymbol{X})$ は定数と考えることができる．

音響モデル $P(\boldsymbol{X}|\boldsymbol{l})$ は HMM を応用したモデルによって，HMM 状態系列変数 $\boldsymbol{q}$ を導入し，以下のように定義される．

$$P(\boldsymbol{X}|\boldsymbol{l}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{q}} P(\boldsymbol{X}|\boldsymbol{q})P(\boldsymbol{q}|\boldsymbol{l}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{q}} \left(\prod_t P(\boldsymbol{x}_t|q_t) \right) P(\boldsymbol{q}|\boldsymbol{l}) \quad (2)$$

すなわち，観測ベクトル系列の個々の要素 $\boldsymbol{x}_t$ は隠れ状態系列 $\boldsymbol{q}$（以降，この隠れ状態を HMM 状態と呼ぶ）が与えられたうえでの条件付き独立であると仮定され，また単語列 $\boldsymbol{l}$ には直接依存していないと仮定される．単語列が与えられたうえでの HMM 状態系列の確率 $P(\boldsymbol{q}|\boldsymbol{l})$ は単語と読みの関係といったようなルールに応じて制約されたマルコフ連鎖によって表される．深層学習登場以前の音声認識では，この音響モデルに登場する出力分布 $P(\boldsymbol{x}_t|q_t)$ を混合正規分布（Gaussian Mixture Model：GMM）を用いて以下のように表してきた．

$$P(\boldsymbol{x}_t|q_t) = \sum_k \pi_{q_t,k} \mathcal{N}(\boldsymbol{x}_t; \mu_{q_t,k}, \boldsymbol{S}_{q_t,k}) \quad (3)$$

ここで，k は混合要素を示すインデックス，$\pi_{q_t,k}$ は状態 q_t における k 番目の混合要素の混合重み，$\mathcal{N}(\boldsymbol{x}_t; \mu_{q_t,k}, \boldsymbol{S}_{q_t,k})$ は平均ベクトル $\mu_{q_t,k}$，共分散行列 $\boldsymbol{S}_{q_t,k}$ を用いて表される正規分布の確率密度関数である．これらのモデルの学習は正規分布のパラメータ $\mu_{q_t,k}, \boldsymbol{S}_{q_t,k}$，およびマルコフ連鎖の重みを EM アルゴリズムや変分ベイズ法を用いて推定することによって行われてきた．また，識別的な規準でパラメータを再学習する識別学習技術なども盛んに研究されてきた [He 08, Woodland 02].

音声認識の精度には，音響特徴ベクトルを元の信号からどのように抽出するかという点も重要である．深層学習登場以前より，音響特徴ベクトル系列 $\boldsymbol{x}_t$ としてはメル周波数ケプストラム係数（Mel-Frequency Cepstral Coefficients：MFCC）が広く用いられている．音声特徴の計算では一般的に，音声信号 10 ミリ秒ごとに 25 ミリ秒程度の長さの音声を切り出し，それを一つの単位（フレームと呼ぶ）として計算を行う．MFCC の計算ではフレームとして切り出された音声をまず対数メルフィルタバンク特徴ベクトルという特徴ベクトルに変換し，その後，MFCC ベクトルへの変換を行う．対数メルフィルタバンク特徴ベクトルは，各フレームとして切り出された音声信号に対応する短時間周波数スペクトルを算出した後，メル周波数と呼ばれる周波数尺度で等間隔となるように配置した複数の三角窓を適用し，各窓の通過エネルギーの対数を要素としたベクトルを計算することによって得られる．また，MFCC は，対数メルフィルタバンク特徴に離散コサイン変換（Discrete Cosine Transform：DCT）を適用し，周波数軸上での滑らかな変化に対応する低周波数の要素のみを取り出すことによって得られる．音声認識に MFCC を用いる場合，一般的には 12 次元の MFCC 特徴量に加え，フレーム内の音声信号の対数エネルギーを記述した一次元特徴量を追加し，さらにそれら 13 次元特徴量の一階 / 二階時間微分を加えた 39 次元の特徴量を用いることが多い．

HMM 状態と単語列を結ぶ確率分布 $P(\boldsymbol{q}|\boldsymbol{l})$ はルールベースのモデルとマルコフ連鎖を融合したものとなっている．単語系列 $\boldsymbol{l}$ に登場する単語は，すべて辞書によって読み（音素系列）が与えられていると考える．そうすると，単語列 $\boldsymbol{l}$ が与えられたときの音素系列 $\boldsymbol{m}$ の確率分布 $P(\boldsymbol{m}|\boldsymbol{l})$ はルールによって与えることができる．HMM 状態の確率 $P(\boldsymbol{q}|\boldsymbol{m})$ は，コンテキスト依存音素と呼ばれる概念を導入してモデル化することが有効であるといわれている．コンテキスト依存音素は各音素（日本語の場合，ローマ字表記した際の一文字とおおまかに対応する）が前後の音素に応じてその音響的特性を変化させるという現象を説明するために導入された概念であり，コンテキスト依存音素を用いた音響モデルでは，同一音素であっても，前や後の音素が異なる場合は異なるラベルをもつとしてモデル化される．コンテキスト依存音素を導入することで，全音素数の L 乗（L はコンテキスト長）通りのラベル *1 を考慮する必要があり過学習が予想されるが，これらのラベルは実際にはコンテキスト依存音素クラスタリングのテクニックによって，縮約されている [Young 94]．そのうえで縮約された各コンテキスト依存音素ごとに，3 状態 Left-to-Right 型 HMM を考える．3 状態 Left-to-Right 型の HMM では，各コンテキスト依存音素系列は，その前半部と中央部，後半部から成ると仮定され，それぞれに異なる HMM 状態が割り当てられる．普通の HMM と同様，隠れ変数はマルコフ連鎖で遷移していくが，各遷移は，同じ HMM 状態を繰り返す遷移か次の部分に移動する遷移かに限定される．各コンテキスト依存音素に対応する HMM の最終状態からは，コンテキストの制約を満たす，ほかのコンテキスト依存音素の先頭状態に遷移し得る．このような複雑な制約を考慮した遷移確率を一般的に重み付き有限状態トランスデューサ（Weighted Finite-State Transducer：WFST）で表現するが，本稿はこれについては詳述しない（詳細は，例えば [堀 04] を参照）．ここで特筆すべきは，このように制約をもった状態遷移をもつ HMM を利用することで，HMM 状態系列が定まることで少なくとも音素系列が一意に定まり，同様に単語列もほぼ一意に定まるという点である．すなわち音声認識の音響モデルの難しさの大部分は，この HMM 状態系列の推定であるといえる．

---

*1 システムのデザインにもよるが，英語，日本語ともに音素数 40 ～ 50 の音素体系を用いることが多い．また，コンテキスト長としては $L = 3$ を用いることが多い．

2·2 深層学習に影響を与えた技術

深層学習登場以前の音声認識研究の中で，現在の深層学習によるブレークスルーに関連する，最も重要な技術はTandemアプローチと呼ばれるMLPとGMMの複合アプローチであろう [Hermansky 00]．音声認識では，Fisher Discriminant Analysis（Linear Discriminant Analysisとしても知られる）で行われてきたような特徴量の識別分析とそれに基づく特徴変換（例えば，[Haeb-Umbach 92, Povey 05]）をMLPによって行うというアプローチが1990年代から行われてきた．Tandemアプローチに基づく研究では，MLPに有効な特徴を発見させることに主眼が置かれ，MFCCのみではなくさまざまな特徴が利用されてきた．その中には現在のような事前学習に基づくものではないものの，深い構造を有するMLPによる特徴抽出を行っている例もあり，現在の深層学習・表現学習の流れと近いものもある [Chen 05, Kubo 11]．

Tandemアプローチでは，MLPを特徴抽出のためのモデルとして利用する．図1に沿ってTandemアプローチについて説明する．Tandemアプローチでは音声特徴ベクトル $\boldsymbol{x}_t$ は，MLPの出力ベクトルを正規化したベクトル $\boldsymbol{\Psi}(\boldsymbol{y}_t)$ に変換される．この $\boldsymbol{\Psi}(\boldsymbol{y}_t)$ をGMMによってモデル化し，元の特徴ベクトルの出力分布 $P(\boldsymbol{x}_t \mid q_t)$ は $P(\boldsymbol{\Psi}(\boldsymbol{y}_t) \mid q_t)$ に比例するとして利用される．$\boldsymbol{\Psi}(\boldsymbol{y}_t)$ を用いる代わりに $\boldsymbol{h}_t$ を正規化したベクトルを用いることも多い．MLPは入力ベクトル $\boldsymbol{x}_t$ を例えば対応する音素に分類するように設計される．このように補助的な識別問題を解くようにMLPを学習し，その出力や隠れ層を特徴抽出の部品として用いることでMLPの非線形特徴分析能力をHMMベースの音声認識技術の枠に導入するということが行われてきた．

Tandemアプローチは深層学習との組合せにおいても注目を集めている．Tandemアプローチにおいては，ニューラルネットワークによる音響特徴抽出部分と，これまで深く検討されてきたGMM/HMMによる音声認識部分が明示的に分離されているため，これまでGMM/HMMのために検討されてきた各種技術がそのまま利用可能なことも多い．

Tandemアプローチに先駆けてMLP/HMM Hybridアプローチと呼ばれる研究があったことも重要であろう [Bourlard 94]．後述するDeep Neural Network（DNN）に基づくHMM音響モデル，すなわちDNN-HMMの基本構造は，実際のところMLP/HMM Hybridアプローチの時点ではすでに完成していたものであり，単に深いMLPを学習する手段を発見できなかったことから，現在のDNN-HMMのような大幅な精度向上を実現できなかったのであろうと著者は考えている．このように音声認識に向けて深層学習の導入に必要な技術の多くは，深層学習登場以前に音声認識の分野で検討されていた．しかし，こうしたアプローチでは主に浅いMLPのみが利用され，深層学習の本格的な普及には時間を要した．

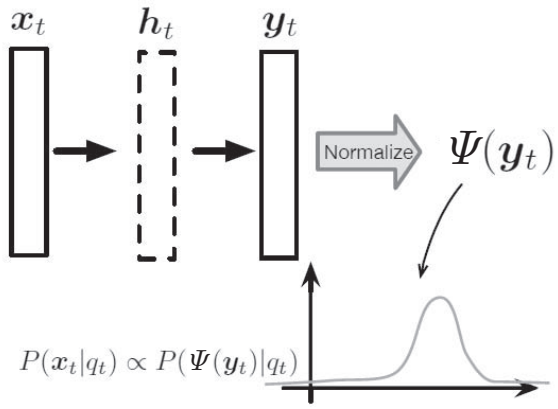

図1 Tandemアプローチ

3. DNN-HMMによる音声認識

DNNの音声認識への適用は先述したMLP/HMM Hybridアプローチに沿って実現される．本章ではMLP/HMM Hybridアプローチについて紹介するとともに，音声認識のためのDNNの学習法について紹介する．

3·1 MLP/HMM Hybridアプローチ

MLP/HMM Hybridアプローチは，ベイズ則によって示される以下の関係式を基本として，個々の確率分布を別のパラメータで表現することによって行われる．

$$P(\boldsymbol{x}_t \mid q_t) = \frac{P(q_t \mid \boldsymbol{x}_t)}{P(q_t)} P(\boldsymbol{x}_t) \tag{4}$$

ここで，$P(q_t)$ はHMM状態のユニグラム生起確率であり離散確率分布でモデル化する．$P(q_t \mid \boldsymbol{x}_t)$ はHMM状態の予測確率でありMLPを用いてモデル化する．$P(\boldsymbol{x}_t)$ は識別器を実際に利用するときには定数として扱うことができるため学習を行わない．この表現では，HMM状態が与えられたときの観測ベクトルの確率をHMM状態のユニグラム確率モデル $P(q_t)$ と，入力が与えられたうえでのHMM状態の予測確率モデル $P(q_t \mid \boldsymbol{x}_t)$ で表し，それに加えてHMM状態遷移確率 $P(\boldsymbol{q} \mid \boldsymbol{l})$ を導入することでHMMを構成している．個々の確率モデルは最尤規準（MLPは最小クロスエントロピー規準）によって推定され，推定されたモデルを組み合わせることによって利用される．

従来GMMを用いた音声認識を行う際には，入力ベクトル $\boldsymbol{x}_t$ としてMFCCをベースとした39次元の特徴量を用いることが一般的であった．しかし，DNNを用いる場合は，MFCCなどの特徴を $\boldsymbol{z}_t$ と置き $\boldsymbol{x}_t = [\boldsymbol{z}_{t-\tau}^{\mathsf{T}}, \boldsymbol{z}_{t-\tau+1}^{\mathsf{T}} \cdots \boldsymbol{z}_{t+\tau-1}^{\mathsf{T}}, \boldsymbol{z}_{t+\tau}^{\mathsf{T}}]^{\mathsf{T}}$ のように時間方向で前後にある特徴量を連結して用いることが一般的である．このようなアプローチの有効性は先述したTandemアプローチや特徴変換の研究ではよく知られていたが，GMMでは高次元の特徴をうまく扱えないという問題があったため直接

は用いられてこなかった．また，DNN では $\boldsymbol{z}_t$ として，MFCC を計算する前段階である対数メルフィルタバンク特徴を用いることもある．このように，DNN では高次元かつ相関を強くもつような特徴を直接扱うことができる点が優れていると考えられる．

3·2 DNN の学習

前節の Hybrid アプローチを用いることで，DNN を含む MLP を音響モデルに取り込むことができる．本節では，事前学習を援用した DNN の構成法について解説する．音声認識では畳込みニューラルネットワーク（Convolution Neural Networks：CNNs）を用いない場合も多いことから，事前学習は重要なステップとして認知されている．いくつかの検討結果は，データが十分にありモデル規模が十分に大きい場合の事前学習の不要性を示唆しているが（例えば [Seide 11]），依然として事前学習による性能向上は観測されており，実際に用いられる場合が多い．

音声データは一般的に音声特徴系列の集合として表されるが，本節の学習モデルでは系列としての情報を利用しないため，すべての系列内の要素を並べた $\mathcal{X} \stackrel{\text{def}}{=} \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_t \cdots\}$ のような音声特徴ベクトル集合を学習データとして考える．

音声認識のための DNN の事前学習には Restricted Boltzmann Machine（RBM）が用いられることが一般的である．RBM に関する詳細な解説は本連載の第 1 回を参照されたい．音声特徴量は実数値で示される信号であるため，最初の一層に対応するパラメータを事前学習する RBM として，以下で示される Gaussian-Bernoulli RBM（GB-RBM）が用いられる．

$$P(\boldsymbol{v}, \boldsymbol{h} | \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}) \stackrel{\text{def}}{=} \frac{1}{Z} \exp\left\{ -\|\boldsymbol{v} - \boldsymbol{c}\|_2^2 - \boldsymbol{h}^{\mathsf{T}} \boldsymbol{W} \boldsymbol{v} - \boldsymbol{b}^{\mathsf{T}} \boldsymbol{h} \right\} \tag{5}$$

ここで，$\boldsymbol{v} \in \mathbb{R}^D$ は RBM の観測変数（D 次元実数ベクトル），$\boldsymbol{h} \in \{0, 1\}^H$ は RBM の潜在変数（H 次元バイナリベクトル），$\boldsymbol{W}$, $\boldsymbol{b}$, $\boldsymbol{c}$ は RBM のパラメータである．この RBM は以下のような条件付き確率を内包する．

$$\left.\begin{aligned} P(\boldsymbol{v} | \boldsymbol{h}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}) &= \mathcal{N}(\boldsymbol{v}; \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, \boldsymbol{I}) \\ P(h_i = 1 | \boldsymbol{v}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}) &= \sigma\left(\sum_j w_{i,j} v_j + C \right) \end{aligned}\right\} \tag{6}$$

ここで，h_i はベクトル $\boldsymbol{h}$ の i 番目の要素，σ はシグモイド関数，$w_{i,j}$ は重み行列 $\boldsymbol{W}$ の（i, j）番目の要素，c_i は $\boldsymbol{c}$ の i 番目の要素，$\boldsymbol{v}_j$ は v の j 番目の要素，$\mathcal{N}(\boldsymbol{v}; \boldsymbol{W}\boldsymbol{h}+\boldsymbol{b}, \boldsymbol{I})$ は平均 $\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$ であり，共分散行列が単位行列である多変量正規分布である．GB-RBM も，Bernoulli-Bernoulli RBM（BB-RBM）と同様 Contrastive Divergence によって最適化可能であるが，GB-RBM の推定は BB-RBM の推定より数値的な安定性が悪く学習率などの設定には注意を要する [Hinton 10]．また，正規分布の分散が単位行列となっているように，この GB-RBM は入力ベクトルの要素ごとのスケールを適切にモデル化しないため，学習データ全体を平均 0，分散 1 となるように正規化して利用することが一般的である．

DNN の事前学習では，まず GB-RBM をトレーニングデータの特徴量集合から推定した後，GB-RBM に対応する単層のニューラルネットワークを用いて，以下のように各トレーニングデータ $\boldsymbol{x}_t$ に対応する特徴 $\boldsymbol{z}_t^{(1)} \stackrel{\text{def}}{=} [z_{t,1}^{(1)}, z_{t,2}^{(1)} \cdots z_{t,i}^{(1)} \cdots]^{\mathsf{T}}$ を得る．

$$z_{t,i}^{(1)} = P(h_i = 1 | \boldsymbol{v}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}) \tag{7}$$

続いて，このようにして得た $\boldsymbol{z}_t^{(1)}$ を学習データとして，2 層目の BB-RBM を計算し，同様に 2 層目の特徴 $\boldsymbol{h}_t^{(2)}$ を求める．これを繰り返すことによって所望の層数をもつ RBM の集合を得る（Deep Boltzmann Machine の貪欲学習，本連載の第 1 回 §7 を参照）．この RBM のパラメータを DNN パラメータの初期値として用いるのが DNN の事前学習である．具体的には，$\boldsymbol{l}$ 層目の DNN の重みパラメータを対応する RBM の重みパラメータ $\boldsymbol{W}$ で，バイアスパラメータを対応する RBM の潜在変数バイアス $\boldsymbol{b}$ で初期化する．最後の隠れ層から最終層への結合に関しては小さな正規乱数で初期化されることが一般的である．最終層のアクティベーション関数としてはシグモイド関数 σ の代わりにソフトマックス関数を用いる（本連載の第 4 回「画像認識のための深層学習」を参照）．そのほかの層に関してもシグモイドの代わりに Rectified Linear Unit や線形ユニットを用いたりするような試みもなされている．

これを初期値として確率的勾配降下法（Stochastic Gradient Descent：SGD）によって最小クロスエントロピー規準での最適化を行うことが一般的である．一般的には HMM 状態変数 q_t は隠れ変数であり，それを固定して学習を行うことはあまり有効ではないと予想されるが，先述したとおり，音声認識問題の場合 HMM 状態の系列 $\boldsymbol{q}$ が定まった場合，音素列 $\boldsymbol{m}$ は一意に定まる，また音素列 $\boldsymbol{m}$ が定まった場合の単語列もほぼ一意に定まる．このような性質から，逆に単語列 $\boldsymbol{l}$ が与えられたときに取り得る $\boldsymbol{q}$ を正確に再現することのできる $P(q_t | \boldsymbol{x}_t)$ を得ることができれば，そのようなモデルは十分に良い性能を出すことが期待される．よって音声認識ではほかの，例えば従来の最尤（ゆう）推定によって得た混合ガウス分布に基づく音声認識用のモデル $\tilde{\Theta}$ と正解単語列 $\mathcal{L}$ を用いて正解 HMM 状態系列 $\boldsymbol{q}$ を $P(\boldsymbol{q} | \mathcal{L}, \mathcal{X}, \tilde{\Theta})$ を最大にするよう推定し，そのうえで，その正解ラベルを用いて最小クロスエントロピー学習することが一般的である．

現在，さまざまな研究機関で網羅的に実験が行われている最中であるが，現時点での典型的な設定としては，各隠れ層のノード数が 2 048，層数としては MFCC を入力する場合は 5 層，対数メルフィルタバンクを入力する場合は 8 層というモデル構造がよく用いられる．また最適化にはミニバッチ，すなわち複数のトレーニングデー

タについての勾配を加算して利用するのが一般的である．

4. 学習法の進展

音声認識における深層学習の進展はさまざまな側面で見られる．音声認識分野で特に大きな流れとして表れつつあるのは，音声の性質や音声認識のマルチタスク性に着目した学習方法の進展であると考えられる．本章では，最初に音声の時系列性に着目した系列学習の進展を，次にそれに付随した最適化の進展，最後に音声認識の多面性に着目したマルチタスク学習について述べる．

4·1 系 列 学 習

深層学習が登場する以前，音声認識技術の研究分野において注目を集めていたのは，最適化が容易な系列ロス関数をどのように設計するかについてである．HMMは確率モデルであるため，最尤推定（や，ベイズ推論）に基づく推定は最も直接的な推定方法の一つであるが，HMMを単なる識別器のデザインの一つとみなし，ロス関数を最小化することによって行う「識別学習」と呼ばれる技術が実際上高い性能を得るための手段として利用されてきた [He 08, Woodland 02]．識別学習技術では，実際に音声認識器を言語モデルなどと組み合わせて駆動することで，どことどこが誤りやすいかを対立仮説候補の形で出力し，対立仮説に関連付いたスコア（対数尤度）を減少させ，正解単語列に関連付いたスコアを増加させるよう，パラメータを最適化する．最小クロスエントロピー規準によるMLPの学習は，各フレームが正しいHMM状態に割り当てられることを目標とした識別学習の一種であると考えられるが，系列の識別を直接最適化することによって，例えば，あるクラスとあるクラスの識別境界は曖昧に，すなわちほぼ等確率を割り当てるように調整しておいたほうが，前後のフレームでの識別率が向上し，結果として単語列の正解精度が上がるといったような，識別器の出力系列における相関関係および最終的な評価規準を考慮した学習ができる．

本節では，系列の学習を直接取り扱うことから，学習データとして入力系列ごとに区切られた集合を考える．学習データ中の音声特徴系列は $\mathcal{X} \stackrel{\text{def}}{=} \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)} \cdots\}$ のようにベクトル系列の集合として表す．また学習データ中の単語列も $\mathcal{L} \stackrel{\text{def}}{=} \{\boldsymbol{l}^{(1)}, \boldsymbol{l}^{(2)} \cdots\}$ のように単語列の集合として表すことにする．

MLPにおける最小クロスエントロピー規準をそのまま系列に拡張した規準が最大相互情報量規準（Maximum Mutual Information 規準：MMI 規準）である*2．以下に，MMI規準において最大化されるべき目的関数を示す．

$$F^{(\mathrm{MMI})}(\Theta) \stackrel{\text{def}}{=} \sum_n \log \frac{P\left(\boldsymbol{l}^{(n)} \middle| \boldsymbol{X}^{(n)}, \Theta\right)}{\sum_{l'} P\left(\boldsymbol{l}' | \boldsymbol{X}^{(n)}, \Theta\right)} \tag{8}$$

ここで，分母に登場する総和（$\Sigma_{l'}$）は考え得るすべての単語列についての総和である．勾配ベースの最適化法を適用するにあたり，この部分の偏微分を正確に計算するのは困難である．実際上は，ラティスと呼ばれる効率的なグラフベースの対立仮説集合の表現を利用することで，この部分の偏微分係数を計算することが一般的である．ラティスの代わりに逐次推定されたサンプル値で表現する構造パーセプトロンアルゴリズム [McDonald 10] も利用されることがあるが，非線形性をもつモデルに対し並列処理によって最適化することが難しいとの理由から，ラティスによる近似とすべてのトレーニングデータに関する勾配を加算した後，パラメータを更新するフルバッチ勾配降下法による学習が深層学習の登場以前より利用されてきた．

MMI規準に基づく学習は系列レベルの識別学習であり，系列識別器として高い精度を出すようにパラメータを調整するが，MMI規準による評価は必ずしも実際の音声認識器の評価と一致しない．その一つの理由はMMIの目的関数が系列ラベルを一体と見たときの一致・不一致しか見ていない点がある．音声認識では，ラベル系列の全体を正確に推定する識別器を得ることは困難であるため単語エラー率のような部分一致の尺度で評価を行う．こうしたことを受け，音声認識器の学習では，ラベル系列間のエラー尺度を利用した目的関数による最適化が行われてきた．例えば，以下に示すbMMI規準では，エラーを多く含む仮説の出現確率を上げることによって，積極的にそのような仮説の出現を抑える [Povey 08]．

$$F_{\mu}^{(\mathrm{bMMI})}(\Theta) \stackrel{\text{def}}{=} \sum_n \log \frac{P\left(\boldsymbol{l}^{(n)} \middle| \boldsymbol{X}^{(n)}, \Theta\right)}{\sum_{l'} P\left(\boldsymbol{l}' | \boldsymbol{X}^{(n)}, \Theta\right) \exp\left\{\mu E\left(\boldsymbol{l}^{(n)}, \boldsymbol{l}'\right)\right\}} \tag{9}$$

ここで $\boldsymbol{E}(\boldsymbol{l}^{(n)}, \boldsymbol{l}')$ は $\boldsymbol{l}^{(n)}$ と $\boldsymbol{l}'$ の間の距離（例えば編集距離）である．この $\boldsymbol{E}$ を用いた確率値のシフト $\exp\{\mu \boldsymbol{E}(\boldsymbol{l}^{(n)}, \boldsymbol{l}')\}$ を導入することによって，エラーが多い，すなわち正解 $\boldsymbol{l}^{(n)}$ からの距離が大きい仮説の出現確率を上げる．この操作はStructured SVM [Tsochantaridis 04] におけるマージンシフトに対応し，正解ラベル系列と比べた距離が大きな仮説ラベル系列ほど，大きなマージンをもって識別するような目的関数を導入することに相当する．より直接的な手段として，ラベル系列間のエラー尺度の期待値を直接的に最適化する以下のようなMPE目的関数も導入され効果を上げてきた [Povey 02]．

$$F^{(\mathrm{MPE})}(\Theta) \stackrel{\text{def}}{=} -\sum_n \sum_{l''} \frac{P\left(\boldsymbol{l}'' | \boldsymbol{X}^{(n)}, \Theta\right) \exp\left\{E\left(\boldsymbol{l}^{(n)}, \boldsymbol{l}''\right)\right\}}{\sum_{l'} P\left(\boldsymbol{l}' | \boldsymbol{X}^{(n)}, \Theta\right)} \tag{10}$$

*2 逆に，MMIからのアナロジーで，最小クロスエントロピー規準をFrameごとのMMI規準（Framewise MMI）と呼ぶこともある．

こうした学習の定式化は以上に示したように明快であるが，実際の最適化は多くの場合困難である．最も大きな理由はラティスによる総和Σ_l近似の不正確性であると考えられる．ラティス表現はその辺の数や頂点の数を増やすことによって，正確に近似対象の総和Σ_lの計算が可能であり，GMMを元にした音声認識ではそれがうまく働いていた．しかしDNNの学習では非常に表現力の高い，非線形性の高いモデルを調整するため，ラティス中に出現しないパターンに偶発的に高い確率が与えられることも多い．上述の目的関数のそうしたパターンに対する値はゼロに近づくことが知られており，これは本来不正解であるパターンが不正解としてみなされないのと同様であり，最適化が進んでも実際の音声認識器のエラー率は改善されないということが起こる．こうした問題に対処するため，さまざまなヒューリスティックを用いてラティスによる近似学習を成立させるといったことが行われてきた[Su 13, Veselý 13].

4·2 最適化の進展

音声認識への深層学習の応用とは直接関係しないが，上述の系列学習や後述するRecurrent Neural Networkでは入出力が系列となり，見掛けの学習サンプル数が減少し，各サンプルの学習にかかる時間が増える．このような場合，高速なアップデートを何回も繰り返すことで比較的良い解を高速に得るSGDの計算効率はほかの手法と比べて必ずしも良いといえない．また，系列学習に用いるラティスの扱いなどをGPU上で実装するのは困難であり，ネットワークによって接続された複数のマシンで並列最適化を行う手法が望まれている．こうした背景から，すべてのトレーニングデータを用いて計算した正確な勾配ベクトルを用いてモデルを更新していくフルバッチ型の最適化手法が望まれている．

準ニュートン法によるMLPの学習は古くから検討されてきたトピックの一つであるが，近年，深層学習の登場に従い，再度注目を集めている[Saito 97]．特にHessian Free法[Martens 10]と呼ばれる，Newton-CG法をDNNに適用した事例は近年多くの研究機関が評価している．Newton-CG法では，全パラメータを並べたベクトル$\boldsymbol{\theta} \in \mathbb{R}^N$の最適化に，ニュートン法を用いることを考え，以下の更新則を考える．

$$\boldsymbol{\theta}^{(e+1)} = \boldsymbol{\theta}^{(e)} - \eta^{(e)} \boldsymbol{H}^{-1} \nabla \boldsymbol{F}(\boldsymbol{\theta}^{(e)}) \qquad (11)$$

ここで，上付きの$^{(e)}$は最適化の繰返し回数を示す変数，$\boldsymbol{F}$は最適化の目的関数，$\eta^{(e)}$はステップサイズ，$\boldsymbol{H}$は目的関数$\boldsymbol{F}$のヘッセ行列である．ニュートン法は，ヘッセ行列が正定値でないと正しく収束しない．DNNの最適化の目的関数の凸性は保証されないため，Hessian Free法ではここで，ガウス・ニュートン行列を代わりに導入する．ガウス・ニュートン行列は，DNNの最終層の出力をベクトル空間からベクトル空間への関数を$\boldsymbol{N}: \mathbb{R}^D \rightarrow \mathbb{R}^O$としたときのヤコビ行列$\boldsymbol{J} \in \mathbb{R}^{O \times D}$と出力層を変数とした目的関数のヘッセ行列$\bar{\boldsymbol{H}} \in \mathbb{R}^{O \times O}$を用いて，$\boldsymbol{G} = \boldsymbol{J}^{\top} \bar{\boldsymbol{H}} \boldsymbol{J}$のように定義される．また，ヤコビ行列$\boldsymbol{J}$の$(d, n)$要素$j_{d,n}$は，最終層の$d$番目のユニットの出力として$z_d(\boldsymbol{x}_t)$をもつようなDNNに対して，$\partial z_d(\boldsymbol{x}_t) / \partial \theta_n$と定義される．このガウス・ニュートン行列は$\boldsymbol{F}$が最小クロスエントロピー規準やMMI，bMMI規準の目的関数であるとき，正定値である．

ガウス・ニュートン行列による近似を用いても，上述の更新式は$(N \times N)$行列の逆行列を求める必要があるため，現実的な時間では実行できない．そこで，Hessian Free法ではConjugate Gradient（共役勾配法）を用いて$\boldsymbol{G}^{-1} \nabla \boldsymbol{F}(\boldsymbol{\theta}^{(e)})$を直接求めることを考えるConjugate Gradient法の具体的なアルゴリズムについては言及しないが，Conjugate Gradient法は$\boldsymbol{G}^{-1} \nabla \boldsymbol{F}(\boldsymbol{\theta}^{(e)})$を$\sum_k \alpha_k \boldsymbol{G}^k \nabla \boldsymbol{F}(\boldsymbol{\theta}^{(e)})$として近似するための$\alpha_k$を高速に求める手法である．すなわち，Conjugate Gradient法を利用することで，逆行列を必要とした計算を行列積の計算の連鎖で近似することができる．

DNNのようにパラメータ数が多い場合，行列積の計算も明示的に行うのは困難である．[Pearlmutter 94]や[Schraudolph 02]の手法は，ヘッセ行列やガウス·ニュートン行列と任意のベクトルとの積を高速に計算する手法を提案しており，これを援用することで，先述した共役勾配法の計算，ニュートン法の更新計算が可能になる．このヘシアン行列やガウス・ニュートン行列の高速積算法はDNNのような非常に大量のパラメータをもつ識別器において，パラメータ空間における目的関数の性質を調べる数少ない方法であり，Hessian Free法以外の意味でも重要であると考えられる．

4·3 マルチタスク学習

マルチタスク学習は，目的とする問題以外の問題に対する統計モデルを情報を共有しながら同時に学習することで，共有知識を明示的に取り出し，学習されるモデルの汎化性能を向上させる手法である．音声には本質的にさまざまなラベリングの方法があり，DNN-HMMで用いられているラベルの表現，すなわちフレームごとに対応するHMM状態が最適であるとは考え難い．このような問題では，本来解くべき問題，すなわちHMM状態識別問題以外に，関連する問題を同時にマルチタスク学習の枠組みで学習し，ラベルの情報を増強することで，より汎化能力の高い識別器を得ることができる可能性がある．

先述したように現状スタンダードとなっている音声認識におけるDNNの適用は入力音声特徴ベクトルに対応するHMM状態を推定するだけという非常に部分的なものとなっている．入力音声特徴ベクトルは比較的長い音声区間の音響特徴量で構成されており，その入力には，前後の文脈に依存する変動と，推定対象であるHMM状態に依存する変動の双方が含まれていることが予想される．そこで，音素コンテキストを同時に推定することで

その双方の変動を明示的にモデル化するという試みが行われている．[Seltzer 13] では，ラベル上で前後の音素やフレームレベルでの前後の音素や HMM 状態を推定対象として追加することで，結果的に本当の推定対象である HMM 状態の識別精度が向上し，音声認識の精度が向上したと報告している．このように，DNN では出力層のデザインを変更することでさまざまなラベルの情報を取り入れることができる．

音声特有の興味深い問題を解いている事例としては，多言語問題への適用がある．国際音声記号と呼ばれるアルファベットで世界のさまざまな言語の音声を記述できるとされているように，音声信号は，ある程度は言語を超えて共有の現象の組合せで記述可能であることが推測される．こうしたことから，DNN の入力信号に近い部分では，言語を超えて，ある程度共通の特徴抽出プロセスが有効であることが推察される．このアイディアをマルチタスク学習の枠組みに導入し，学習データ中の各訓練事例の言語に応じて利用する出力層を取りかえていくことによってメインとなる言語の認識率を上げることができるという報告が複数なされた [Ghoshal 13, Heigold 13, Tuske 13]．これらの応用は，多言語に共通の特徴があるということや，DNN の入力に近い層は主に特徴抽出を行っているという仮説を支持しており，非常に興味深い．

5. Recurrent Neural Network の進展

本章では，深層学習技術の音声認識への応用におけるもう一つの側面，すなわちモデル構造の工夫に関して，特に音声の時系列性を表現するために導入された Recurrent Neural Network について詳述する．

Recurrent Neural Network は時間遅れ素子を介してユニットの出力が別のユニットの入力となっているようなニューラルネットワークの総称である．音声認識のコンテキストでは Elman Network と呼ばれる，2層多層パーセプトロンの隠れ層の出力が時間遅れを介して次の時刻の隠れ層への入力となっている図2(a)のようなパーセプトロンを用いることが一般的である．

Elman Network の学習は DNN と同様 SGD を用いた最小クロスエントロピー規準の最適化によって行うが，入出力の依存関係を系列データ間の依存関係として表現するため，入出力データは各系列ごとに与える必要がある．入力の系列データを受け取り，出力の系列データを出力するネットワークを図2（b）に示す．図のとおり，系列データを対象にすることで，Deep Neural Network と同様の深い構造を含んでいることがわかる．

RNN の学習は DNN の学習と同様に SGD を用いて行われる．隠れユニットの出力変数を決定してしまえば，図2の形の RNN は単なる隠れ層が1層の MLP として学習できるが，実際には図2（b）で示されるような DNN 表現にしたうえで途中で打ち切って学習をすることが多い（この学習法を Back Propagation Through Time（BPTT）と呼ぶ）．しかし，DNN の場合と異なり，音声認識における RNN の学習では事前学習を使わないことが一般的である．すなわち DNN の場合と異なり，深層学習技術の登場によって学習法が進展したということは，ほとんどない*3．しかし，DNN の流行に伴い，RNN が見直される過程において，さまざまな興味深い応用や新たな学習方法が登場してきている．

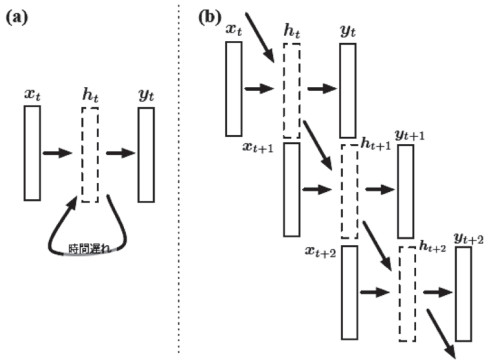

図2 Recurrent Neural Network.
（a）時間遅れ素子を用いた表示，（b）時系列入出力を展開した表示

5·1 RNN 言語モデル

音声認識における言語モデル（式（1）の $P(\boldsymbol{l})$）では，N-gram モデルが用いられることが一般的である．N-gram 言語モデルでは，各単語の生起確率が直前の数単語に依存した確率分布で定まると仮定し，以下の近似を置いたモデルを用いる．

$$P(\boldsymbol{l}) = \prod_m P(l_m \mid l_{m-1}, l_{m-2}, l_{m-3} \cdots)$$
$$\approx \prod_m P(l_m \mid l_{m-1}, l_{m-2}, \cdots l_{m-N+1}) \qquad (12)$$

単純な離散分布による実現では，この確率分布は語彙数 V に対して $V^{(N)}$ のパラメータを必要とする．実際には，N-gram 言語モデルでゼロ頻度となってしまう単語の連なりを（$N-1$）-gram 言語モデルの確率を用いて計算するバックオフの仕組みによってパラメータ数は $V^{(N)}$ より大幅に小さいものとなっているが，基本的に定められた N に対してパラメータの個数が決まるという構造となっている．

近年，こうした N-gram 言語モデルの代わるモデルとして，もしくは確率値を補完して同時に使うためのモデルとして RNN 言語モデルが注目を浴びている [Mikolov 10]．[Mikolov 10] では RNN 言語モデルを単独で使った場合においても，従来の N-gram 言語モデルの精度を超えることが報告されているほか，N-gram 言語モデルと

*3 Hessian Free 法などは RNN においても効果的であると報告されている [Martens 11].

確率値を補完することによってより高精度な音声認識ができることが報告されている．

RNN 言語モデルでは，上式中の単語履歴が与えられたときの条件付き確率 $P(l_m \mid l_{m-1}, l_{m-2}, l_{m-3}\cdots)$ を RNN によって直接計算することを考える．具体的には入力を直前の単語 l_{m-1} を 1-of-N 表現によって表現したベクトル $\boldsymbol{\delta}_{l_{m-1}}$ として，出力層によって単語の予測分布を表現する RNN を以下のように計算する．

$$P(l_m \mid l_{m-1}, l_{m-2}, l_{m-3}\cdots) = \frac{\exp\{\sum_i \lambda_{l_m,i} h_{m,i}\}}{\sum_{l'} \exp\{\sum_i \lambda_{l',i} h_{m,i}\}} \quad (13)$$

ここで $h_{m,i}$ は m 番目の単語の予測に用いる隠れ層の状態であり，以下の式で表される．

$$h_{m,i} = \sigma\left(\sum_j v_{i,j} h_{m-1,j} + w_{i,l_{m-1}}\right) \quad (14)$$

このモデルの学習において最適化されるパラメータは入力層から隠れ層への結合重み係数である $w_{i,l'}$，隠れ層同士の再帰的リンク結合重み係数である $v_{i,j}$，および隠れ層から出力層への結合重み係数である $\lambda_{l',i}$ である．

N-gram 言語モデルの場合と異なり，RNN 言語モデルでは隠れ層の素子数を調整することによってモデルの複雑さを制御する．単語の履歴はすべて RNN における再帰的リンクによって記憶されていると考えるため，より長いコンテキストをもつ RNN 言語モデルを作成したい場合は素子数を増やす必要がある．しかし，深層学習そのものの問題から素子数を増やしても，再帰的リンクで時間的に離れた二つの入出力の関係性を適切に表現するのは難しいとされている [Hochreiter 97]．RNN 言語モデルの学習においては，事前学習技術を使うことは一般的でなく，先述した BPTT を用いて学習することが多い．学習法の進展なしに，RNN 言語モデルが音声認識の高精度化に寄与できるようになった理由の一つとして，近年の計算機の進歩に伴う学習可能なデータ量の増大があげられる．音響モデルの深層学習においても，学習データが増加することで，従来事前学習なしでは困難であった DNN の学習が事前学習なしで精度良く行うことができることが報告されており，データの増大は深層学習の困難性を緩和することが示唆されている．

RNN 言語モデルの最大の問題はその計算量である．N-gram モデルは直前数単語のみが予測に影響するという点で，マルコフ連鎖で表現可能である．音声認識技術は潜在変数がマルコフ連鎖に従っていることを仮定しており，そのうえで高速な推定を行うように進歩してきたため，RNN のように連続値を状態変数として用いる一般的なマルコフ過程の場合は高速に認識結果を推定する手法が存在しない．現在は，従来の音声認識器で有望な単語列の仮説である N-best リストを出力し，そのうえでリスコアリングを行うことで RNN 言語モデルを音声認識と統合しているが，従来法に比べた際の計算効率の悪化は著しい．また，リスコアリング処理そのものにおいても RNN 言語モデルは，基本的な実装では，語彙数 V，隠れユニット数 H としたとき，$V \times H$ の行列の乗算を含むため，大規模な N-best リストに適用することは難しい．[Mikolov 10] ではさまざまなヒューリスティックを用いて計算時間を削減しているが，推定精度に与える影響が無視できないと考えられる．こうしたことから，計算効率の改善が今後の課題といえる．

5·2 Connectionist Temporal Classification

上記の RNN に関する進展を考えると，音響モデルのにおける HMM の隠れ状態も RNN で保持される記憶によって表現し，音響モデルと言語モデルの双方を一つの巨大な RNN で表現することで，音響特徴系列を入力とし単語系列を出力とするような RNN を構築することが考えられる．しかし，音響モデルと言語モデルでは系列の単位が違う，すなわち音響データは物理的な時刻に対応する単位で与えられるのに対し，予測するラベル列は単語の単位で出力しなければならないため，その間を吸収する仕組みと，そのための学習法が必要となる．Connectionist Temporal Classification（CTC）は，単純なアプローチでこの差を吸収することができる可能性を示した [Graves 06].

CTC では出力側のシンボルとして何も出力しないことを示す null シンボル（例えば ε と示す）を許容することによってこの系列長の違いを吸収する．すなわち，音響特徴ベクトル $\boldsymbol{x}_t$ に対応する RNN 出力 y_t のうち，一部の時刻 t のみが実際に単語もしくは音素を出力 $y_t \neq \varepsilon$ とし，それ以外の時刻では $y_t = \varepsilon$ を出力するような RNN を導入することで，単語数や音素数が入力音声特徴ベクトルの数に比べて少なく，固定長でないような出力が表現できる．特に音声認識の場合，入力の系列長，すなわち一発話に含まれるフレームの数に比べて出力の系列，すなわち単語の数が常に少ないということを仮定しても問題が起こりにくく，こうした単純なアプローチであっても有効に働くことが期待される．

モデル構造の変更は非常に小さいが，null シンボルを許容することによって，認識時や学習時にどこに null が挿入されるかについての不確定性を考慮する必要がある．まず学習時では，HMM 学習時の Forward-Backward アルゴリズムと類似の処理によって，null シンボルを含めた各ラベルの確率を最適化中のモデルを用いて計算する．CTC の最適化規準は，このようにして求めた各時刻にどのラベルにいるかを示す確率値（HMM の EM アルゴリズムにおいて得られる占有率と等価）に基づいて教師信号を変更した最小クロスエントロピー規準の一種として記述できる．教師信号の変更以外はバックプロパゲーションと同様であり，DNN の学習と同様に最適化を行うことができる．認識時はすべての null シンボルの取り得るパターンについて考え，それで周辺化した系列ラベルを導出しなければならないが，これに

関して厳密に高速に計算する手法はない．[Graves 06]では，最初にnullシンボルの取り得るパターンについて最尤推定を行い，その後に対応するラベルを抽出するという処理と，Forward-Backward処理によって得られた統計量を活用して効率的に探索をする手法の2通りが紹介されている．

CTCは現在，連続音素認識での利用が主となっており，大語彙連続音声認識に直接適用することは難しい．音素認識器としてCTCを用いたRNNというのも可能ではあるが，その場合，CTCがRNN言語モデルのように振る舞うことは期待できない．こうしたことから，深層学習の音響モデル技術，言語モデル技術をどのように統合するかについて，決定打が出ているとはまだ言い難い．

6. ま と め

音声認識の諸問題における深層学習の適用事例について紹介するとともに，現在国際会議などの場で活発に議論されている問題について紹介した．

本稿の最初で示したように，従来の音声認識器はメル周波数ケプストラム係数で表現した音声特徴ベクトルの生成モデルを導入することによって音声認識における音響事象をモデル化しようとしてきた．反面，現在の深層学習に基づく音声認識器では，例えば対数メルフィルタバンク特徴ベクトルを用いることで，周波数スペクトルをより高い解像度で表現し，さらに時間的に長いセグメントの情報を入力として音響事象をモデル化している．このようなリッチな入力情報の利用が，従来のアーキテクチャでは深層学習ほど効果的でなかったことから，深層学習によって，ある種の高次特徴抽出プロセスが最適化可能なモデルとして表現され，それが人手でデザインした特徴量より効率的であったため，爆発的な性能向上が実現できたと類推することもできる．また，深く（比較的）細いニューラルネットワーク構造は，その特徴抽出のプロセスにおいて，次元削減を導入していると考えることもでき，浅く太いモデルより有効であることが類推される．

現在のところ，音声認識における深層学習は比較的スタンダードな問題設定のもとで利用されている．すなわち，タスクとしては大語彙連続音声認識，学習データが100時間分くらいあり，音響環境が比較的クリーン（無雑音）に近い設定である．そして，このような問題に対する深層学習の各種ハイパーパラメータ，すなわち隠れユニットの数や層数，確率的勾配降下法における学習率や正則化係数は，この分野に深層学習を導入した研究者らによって比較的よく検討されており，現在ではそうした値をそのままもってくることである程度の精度を誰でも出せるようになってきている．しかし対象とするタスクの変化や，収音環境の悪化などによって問題の性質が変化してくると，モデル選択や各種ハイパーパラメータチューニングを再度行う必要がある．一般的に音声認識のモデル学習には非常に多くのデータを用いる必要があることから，こうしたチューニングは本連載第3回（Vol. 27, No. 5）にて紹介されたようなRandom Searchを用いても網羅的に行うことは難しい．新しいタイプの問題に新しい手法を応用する場合，その実験的な検討は常に必要となるが，現状の深層学習はその部分のコストがほかの手法と比べても非常に大きいと考えている．

音声認識における深層学習も，画像における深層学習と同様，性能の大幅な向上があるため，さまざまな研究機関が一斉に着目し，その実用可能性が短期間で広く議論された．今後もさまざまな音声認識関連技術において，深層学習技術の適用が進展していくと考えられるが，それをより一層大きな流れにしていくためには，多層パーセプトロンの採用によってブラックボックスとなってしまった部分の解明が重要であろうと考えられる．理論と応用の両面から深層学習がより深みのある技術となることを期待する．

◇ 参 考 文 献 ◇

[Bourlard 94] Bourlard, H. A. and Morgan, N.: *Connectionist Speech Recognition: A Hybrid Approach,*Vol. 247, Springer（1994）

[Chen 05] Chen, B. Y., Zhu, Q. and Morgan, N.: Tonotopic multi-layered perceptron: A neural network for learning long-term temporal features for speech recognition, *Proc. ICASSP-05*（2005）

[Ghoshal 13] Ghoshal, A., Swietojanski, P. and Renals, S.: Multilingual training of deep neural networks, *Proc. ICASSP-13*（2013）

[Graves06] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, *Proc. 23rd Int. Conf. on Machine Learning,* pp. 369-376, ACM（2006）

[Haeb-Umbach 92] Haeb-Umbach, R. and Ney, H.: Linear discriminant analysis for improved large vocabulary continuous speech recognition, *ICASSP-92, 1992 IEEE Int. Conf.,* Vol. 1, pp. 13-16（1992）

[He 08] He, X., Deng, L. and Chou,W.: Discriminative learning in sequential pattern recognition, *Signal Processing Magazine,* Vol. 25, No. 5, pp. 14-36, *IEEE*（2008）

[Heigold 13] Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M. and Dean, J.: Multilingual acoustic models using distributed deep neural networks, *Proc. ICASSP-13*（2013）

[Hermansky 00] Hermansky, H., Ellis, D. P. and Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems, *ICASSP-00, 2000 IEEE Int. Conf.,*Vol. 3, pp. 1635-1638（2000）

[Hinton 10] Hinton, G.: A pratical guide to training restricted Boltzmann machines, Technicar Report 2010-003, Machine Learning Group, University of Tronto（2013）

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Computation,*Vol. 9, No. 8, pp. 1735-1780（1997）

[堀 04] 堀 貴明，塚田 元：重み付き有限状態トランスデューサによる音声認識，*IPSJ Magazine,* Vol. 45, No. 10, pp. 1020-1026（2004）

[Kubo 11] Kubo, Y., Okawa, S., Kurematsu, A. and Shirai, K.: Temporal AM-FM combination for robust speech recognition,

Speech Communication, Vol. 53, No. 5, pp. 716-725（2011）
[Martens 10] Martens, J.: Deep learning via Hessian-free optimization, *Proc. 27th Int. Conf. on Machine Learning*（*ICML-10*）, pp. 735-742（2010）
[Martens 11] Martens, J. and Sutskever, I.: Learning recurrent neural networks with Hessian-free optimization, *Proc. 28th Int. Conf. on Machine Learning*（*ICML-11*）, pp. 1033-1040（2011）
[McDonald 10] McDonald, R., Hall, K. and Mann, G.: Distributed training strategies for the structured perceptron, *Human Language Technologies: 2010 Annual Conf. North American Chapter of the Association for Computational Linguistics,* pp. 456-464, Association for Computational Linguistics（2010）
[Mikolov 10] Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J. and Khudanpur, S.: Recurrent neural network based language model, *INTERSPEECH,* pp. 1045-1048（2010）
[Pearlmutter 94] Pearlmutter, B. A.: Fast exact multiplication by the Hessian, *Neural Computation,* Vol. 6, No. 1, pp. 147-160（1994）
[Povey 02] Povey, D. and Woodland, P. C.: Minimum phone error and I-smoothing for improved discriminative training, *ICASSP-02, 2002 IEEE Int. Conf.,* Vol. 1, pp. I-105（2002）
[Povey 05] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H. and Zweig, G.: fMPE: Discriminatively trained features for speech recognition, *Proc. ICASSP-05,* Vol. 1, pp. 961-964 Philadelphia（2005）
[Povey 08] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah, K.: Boosted MMI for model and feature-space discriminative training, *ICASSP-08, IEEE Int. Conf.,* pp. 4057-4060（2008）
[Rabiner 93] Rabiner, L. and Juang, B.-H.: *Fundamentals of Speech Recognition,* Prentice Hall（1993）
[Saito 97] Saito, K. and Nakano, R.: Partial BFGS update and efficient step-length calculation for three-layer neural networks, *Neural Computation,* Vol. 9, No. 1, pp. 123-141（1997）
[Schraudolph 02] Schraudolph, N. N.: Fast curvature matrix-vector products for second-order gradient descent, *Neural Computation,* Vol. 14, No. 7, pp. 1723-1738（2002）
[Seide 11] Seide, F., Li, G. and Yu, D.: Conversational speech transcription using context-dependent deep neural networks, *INTERSPEECH,* pp. 437-440（2011）
[Seltzer 13] Seltzer, M. L. and Droppo, J.: Multi-task learning in deep neural networks for improved phoneme recognition, *ICASSP-13, 2013 IEEE Int. Conf.,* pp. 6965-6969（2013）
[Su 13] Su, H., Li, G., Yu, D. and Seide, F.: Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription, *Proc. ICASSP*（2013）
[Tsochantaridis 04] Tsochantaridis, I., Hofmann, T., Joachims, T. and Altun,Y.: Support vector machine learning for interdependent and structured output spaces, *Proc. 21st Int. Conf. on Machine Learning,* p. 104, ACM（2004）
[Tuske 13] Tuske, Z., Pinto, J., Willett, D. and Schluter, R.: Investigation on cross-and multilingual MLP features under matched and mismatched acoustical conditions, *ICASSP-13, 2013 IEEE Int. Conf.,* pp. 7349-7353（2013）
[Veselỳ 13] Veselỳ, K., Ghoshal, A., Burget, L. and Povey, D.: Sequence-discriminative training of deep neural networks, *Proc. ICASSP-13*（2013）
[Waibel 89] Waibel, A., Hanazawa,T., Hinton, G., Shikano, K. and Lang, K. J.: Phoneme recognition using time-delay neural networks, *IEEE Trans. on Acoustics, Speech and Signal Processing,* Vol. 37, No. 3, pp. 328-339（1989）
[Woodland 02] Woodland,P. C. and Povey, D.: Large scale discriminative training of hidden Markov models for speech recognition, *Computer Speech and Language,* Vol. 16, No. 1, pp. 25-47（2002）
[Young 94] Young, S. J., Odell, J. and Woodland, P. C.: Tree-based state tying for high accuracy acoustic modelling, *Proc. Workshop on Human Language Technology,* pp. 307-312, Association for Computational Linguistics（1994）

2013 年 11 月 12 日　受理

著　者　紹　介

久保　陽太郎

2010 年早稲田大学大学院理工学研究科博士課程修了，博士（工学）．同年ドイツ RWTH アーヘン大学客員研究員を経て，日本電信電話株式会社に入社．音声認識の研究に従事．2010 年日本音響学会より粟屋潔学術奨励賞，2011 年情報処理学会山下記念研究奨励賞，2013 年日本音響学会独創研究奨励賞板倉記念，同年電子情報通信学会音声研究会奨励賞を受賞．日本音響学会，IEEE などの各会員．