特　集「ニューラルネットワーク研究のフロンティア」

ニューラルネットワークによる音声認識の進展

Speech Recognition with Deep Neural Net: After Its First Introduction

久保　陽太郎　Amazon Development Center Germany GmbH.
Yotaro Kubo　yotaro@ieee.org

Keywords: speech recognition, deep learning, automatic speech recognition, acoustic models.

1. ま　え　が　き

音声認識器の一部に深層学習に基づくニューラルネットワーク，すなわち深層ニューラルネット（Deep Neural Network，以下DNN）が利用されることが一般的になってからすでに数年が経過しようとしている．DNNは膨大な計算資源と学習データを要する扱いづらい統計モデルであるが，近年の音声インタフェース技術の普及に後押しされ，精力的に応用技術の検討が進められてきた．

深層学習の導入による効果は非常に大きい．深層学習登場以前の最先端技術で構成された音声認識器の認識誤りの数を100とすると，深層学習による統計モデルを導入することで，誤りの数をおおよそ70以下まで減らせる．その当時，すでにさまざまな最先端技術の集合となっていた精緻な音声認識器が，それでも克服できないエラーを，これだけ減らせるというのは大変な驚きであった．

残念ながら深層学習の導入以降，この最初のインパクトに匹敵するブレークスルーは起きていない．しかしながら，要素技術の向上は着実に進み，70個の誤りをさらに60個まで減らすような技術が研究され続けている．また，認識精度のみではなく，深層学習の一つの弱点である，計算資源の問題を軽減する技術に関してもさまざまな進歩が起こっている．

本稿では，こうした深層学習導入以降の着実な進歩の流れにおいて，どのようなことが検討され，進歩してきているのかを解説したい．本稿では紙幅の都合から深層学習の基礎については触れない．逆伝播や基本的な事前学習の方法など，深層学習技術そのものについては，本学会誌の深層学習に関する連載解説[久保14]や，書籍[麻生15]が詳しい．

本解説ではまず，音声認識の問題設定と，基本技術であるDNN-HMMハイブリッド方式について2章で解説した後，3章で学習技術の進展について述べる．続けて，4章ではニューラルネットの構造を拡張する研究事例について紹介する．5章では音声認識の認識時における計算効率を向上させるためのテクニックについて紹介する．最後に，6章で今後の展望を述べる．

2. 音声認識とDNN-HMMハイブリッド方式

本章では音声認識と，音声認識への深層学習の基本的な適用法を簡潔に述べる．音声認識技術についての詳しい解説は他の専門書を参照されたい．

図1に音声認識問題の入力とラベルの関係を示す．音声認識は，音声を短時間ごとに区切り，それぞれ分析した結果得られるベクトルを並べたベクトル系列から，出力ラベル系列（単語列）を得るパターン認識の一種である．ほぼ同様の枠組みがジェスチャ認識や，手書き文字認識にも利用可能なことから，本稿で紹介する技術の一部はこうした問題にも適用可能であることが期待されている．

音声認識のモデリングにおけるユニークな点は，長さの異なる二つの系列間のマッピングを考える点である．また，図1では区間と音素・単語の対応を図示したが，一般的に，こうしたセグメンテーション情報は未知であり，モデルによって識別とセグメンテーションの両

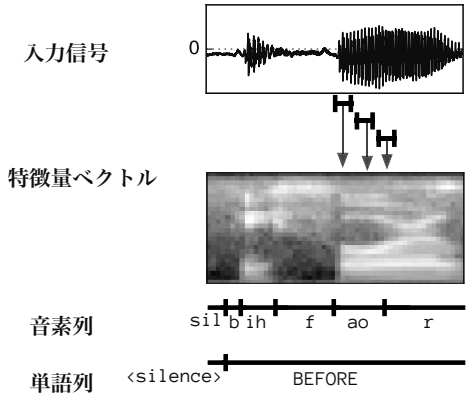

図1　音声認識における多対多系列ラベリング

方を実現しなければならない．深層学習技術によって，ニューラルネットの構成の自由度が向上した今日では，こういったセグメンテーション問題もニューラルネットのみで自然に表現できるようになってきているが，従来のニューラルネットは基本的に，一つの入力ベクトルから一つの出力ラベルを予測するモデルであった．そのため音声認識では，こうしたセグメンテーション問題に有効な統計モデルである，隠れマルコフモデル（Hidden Markov Model，以下 HMM）を援用し，DNN-HMM ハイブリッド方式 [Bourlard 94] と呼ばれる方法で DNN を HMM と組み合わせて利用する．

以降では，この DNN-HMM ハイブリッド方式を紹介する．入力ベクトル系列を $\boldsymbol{X} \stackrel{\text{def}}{=} \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_t \cdots | \boldsymbol{x}_t \in \mathbb{R}^D\}$ とし，出力ラベル系列を $\boldsymbol{l} \stackrel{\text{def}}{=} \{l_1, l_2, \cdots l_m, \cdots | l_m \in \mathcal{V}\}$ とする．ここで $\mathcal{V}$ は出力となり得る単語すべての集合である．音声認識は入力ベクトル系列 $\boldsymbol{X}$ が与えられたときに，最も確率の高いラベル系列 $\hat{\boldsymbol{l}}$ を求める問題として，以下のように定式化できる．

$$\hat{\boldsymbol{l}} = \underset{\boldsymbol{l}}{\operatorname{argmax}} \log p(\boldsymbol{l}|\boldsymbol{X}) = \underset{\boldsymbol{l}}{\operatorname{argmax}} \log p(\boldsymbol{X}|\boldsymbol{l}) p(\boldsymbol{l}) \tag{1}$$

ここで表われる音声信号と単語の関係を表現する確率分布 $p(\boldsymbol{X}|\boldsymbol{l})$ を音響モデルと呼び，人の話す言語の特徴を表現する確率分布 $p(\boldsymbol{l})$ を言語モデルと呼ぶ．深層学習の言語モデルへの適用に関しては，本特集のほかの解説や解説書 [麻生 15] が詳しい．本稿では主に音響モデルへの適用についての技術を紹介したい．

音響モデルは，入力ベクトルと同じ要素数の離散潜在変数の系列 $\boldsymbol{q} \stackrel{\text{def}}{=} \{q_1, q_2 \cdots q_t \cdots\}$ と，音素を表す離散変数の系列 $\boldsymbol{m}$[*1] を導入して，以下のようにモデル化される．

$$p(\boldsymbol{X}|\boldsymbol{l}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{m},\boldsymbol{q}} \left(\prod_t p(\boldsymbol{x}_t|q_t) p(q_t|q_{t-1}, \boldsymbol{m}) \right) p(\boldsymbol{m}|\boldsymbol{l}) \tag{2}$$

ここで $p(\boldsymbol{m}|\boldsymbol{l})$ は単語と音素を関連付けることから発音辞書と呼ばれ，人手で定義するのが一般的である．導入された潜在系列変数 $\boldsymbol{q}$ は音素列 $\boldsymbol{m}$ に依存したマルコフ連鎖 $p(q_{t-1}|q_t, \boldsymbol{m})$ によって生成されていることから，音響モデルは隠れマルコフモデルの一種であるといえ，$\boldsymbol{q}$ を HMM 状態変数と呼ぶ．

システムの構成にもよるが，多くの音声認識システムは，HMM 状態系列から音素列へのマッピングがほぼ一意になるように設計される．この場合，HMM 状態変数が正しく識別されれば，同音異義語や未知語など言語モデルに起因する誤り以外は起こらない．すなわち，音響モデルの学習則を後述するクロスエントロピー学習のような HMM 状態変数の識別問題に置き換えてしまっても大きな性能劣化は起こらない．

深層学習導入以前の音響モデルでは，上式の $p(\boldsymbol{x}_t|q_t)$ を混合正規分布とし，状態遷移確率を含め全体を EM アルゴリズムで学習する方法が一般的であった．DNN-HMM ハイブリッド方式では，この $p(\boldsymbol{x}_t|q_t)$ をソフトマックス出力層をもつ DNN によって推定した確率分布 $p(q_t|\boldsymbol{x}_t)$ を用いて以下の形で表現する．

$$p(\boldsymbol{x}_t|q_t) = \frac{p(q_t|\boldsymbol{x}_t)}{p(q_t)} p(\boldsymbol{x}_t) \tag{3}$$

ここで $p(\boldsymbol{x}_t)$ は認識時には利用されない．また $p(q_t)$ はカテゴリカル分布とし，別の手段を用いてあらかじめ学習しておく．

ここで登場する DNN 識別器 $p(q_t|\boldsymbol{x}_t)$ は，画像認識などの応用事例と同様，クロスエントロピー学習によって得ることができる．ただし，HMM 状態変数 q_t は潜在変数であるため，クロスエントロピー学習を行う前に何らかの手段で確定しておく必要がある．観測変数を条件とした確率分布 $p(\boldsymbol{q}|\boldsymbol{X}, \boldsymbol{l})$ の最尤推定値

$$\hat{\boldsymbol{q}} \stackrel{\text{def}}{=} \underset{\boldsymbol{q}}{\operatorname{argmax}} p(\boldsymbol{q}|\boldsymbol{X}, \boldsymbol{l})$$

（Viterbi 系列とも呼ばれる）がラベルとして用いられている．また HMM 状態遷移確率は，このようにして得た Viterbi 系列から最尤推定する方法で得ることができる．

この DNN-HMM ハイブリッド方式は，全体を一つの生成モデルとして隠れマルコフモデルでモデル化し，EM アルゴリズムで学習していた従来手法と比べ，学習規準の面での一貫性に欠けるが，DNN の高度な識別性能を従来の枠組みに導入し，ルールベースの辞書モデル $p(\boldsymbol{m}|\boldsymbol{l})$ や言語モデル $p(\boldsymbol{l})$ と統合する技術として有用である．

3. 学習技術の進展

DNN-HMM ハイブリッド方式によって，一つの入力ベクトルから一つのラベルを予測するタイプのニューラルネットを発音辞書や言語モデルと組み合わせることができるようになった．これによって，音声認識のパターン認識問題としての特殊性はある程度解消されたといえ，ニューラルネットの学習は画像認識や他の応用分野と同様，クロスエントロピー規準に基づく目的関数を確率的勾配降下法（Stochastic Gradient Descent，以下，SGD）で最小化することで実現できる．また，音声認識の分野で広く使われているような，全結合かつ階層の深い DNN を学習するには，事前学習手法の援用も重要である．

本章では，事前学習技術について経験的にわかってきたことと，クロスエントロピー学習のデータ並列学習法の解説，および系列ラベリングに特化した学習技術の解説を行う．

---

*1 典型的なシステムでは音素列 $\boldsymbol{m}$ の要素数は特徴量ベクトル系列の要素数以上であり，単語列の要素数以下である．

3·1 事 前 学 習

音声認識のためのDNNに関する最初期の検討は，事前学習を併用して行われた．そのため，以降の研究でも，多くの事前学習手法が検討され，どの事前学習手法が良いかについてもさまざまな検討がなされた [Plahl 12, Seide 11]．しかし反面，画像認識の分野では，事前学習を使わない畳込みニューラルネット（Convolution Neural Network，以下CNN）[LeCun 98] が広く用いられているということもあり，事前学習がどの程度重要であるかという問題に対する答えは，いまだ明確にはわかっていない．

事前学習には二つの効果が期待されている．一つは，初期値を適切に設定することによる勾配消失問題の緩和である．勾配消失問題は，1でない偏微分係数がDNNの勾配計算の過程で何度も乗算されることで，勾配のスケールが大きく変化してしまうことによって起こる．確率的勾配降下法をはじめ，最急勾配法を基本としたアルゴリズムは，勾配のスケールが要素ごとに大きく異なる目的関数の最適化を現実的な時間で終えることはできない．

図2に各種の活性化関数の導関数を示す．バックプロパゲーション時には，最終層のバイアスパラメータの偏微分係数に，重みパラメータの値，そしてこの導関数の値が隠れ層の数だけ乗算されていく．sigmoid活性化関数やtanh活性化関数の導関数は絶対値が大きい範囲が非常に狭く，ランダムに決めた初期値に基づいてこの活性化関数の導関数を計算しても，多くの場合0に近い値をとってしまう．隠れユニットの活性化関数の導関数が0に近い値を取るということは，そのユニットに入力する結合の重みパラメータはほとんど更新されないということを示しており，この場合，学習が正しく進行しない．活性化関数の導関数を小さくしないためには，重みパラメータを0に近い値にすることが考えられるが，その場合，重みパラメータの乗算によって勾配が減衰してしまう．この問題に関してはRectified Linear Unit（ReLU）が有効である．ReLUは半空間において，勾配を全く減衰させずに係数1で伝える [Dahl 13, Nair 10]．このため，重み係数の初期値の悪影響によって起こる，勾配消失問題の緩和に有効である．

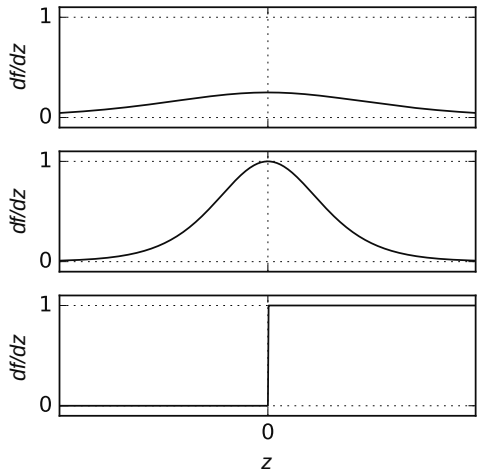

図2 活性化関数の導関数．
上からsigmoid関数の導関数，tanh関数の導関数，rectified linear関数の導関数である

事前学習に期待されているもう一つの効果は，Early Stoppingと組み合わせた際に得られる，正則化と類似の効果である [Erhan 10]．Early Stoppingは推定値が初期値から離れ過ぎる前にクロスバリデーションの結果を参照しながら最適化を中断するため，初期値の性質が学習結果に残る．ReLUを用いた場合のように，勾配消失問題が深刻ではない場合においてなお，事前学習が効果を発揮する理由の一つはこの正則化効果であろう．識別的事前学習 [Seide 11] は，事前学習と実際の学習で同じ目的関数を用いるため，正則化としての効果は小さいと予想される．しかし，勾配消失問題の緩和には識別的事前学習も十分に効果があると考えられる．

以上のことを踏まえると，事前学習手法の選び方に一定の直感が働く．すなわちデータが少なく，正則化の重要性が高いと考えられる場合は事前学習が重要である．またその場合，識別的事前学習ではなく，本来の目的と異なる制約を与えるため，生成的な規準を利用することが有効であると予想がつく．データが十分にあり，正則化の効果が薄い場合でかつ，活性化関数がReLUやmaxout [Zhang 14] のように勾配消失問題の影響を受けづらい場合，事前学習は不要なこともある．また，データが十分にあっても，sigmoid活性化関数の場合，勾配消失問題が大きいため事前学習は良く働くことが多い．このような場合，事前学習用のプログラムを別途必要としない識別的事前学習も良い選択肢となる．

3·2 データ並列学習

DNNのデータ並列学習は難しい．一般的には，パラメータ数が多く，勾配ベクトル・パラメータベクトルともに密な非凸のモデルの並列最適化が難しいともいえる．なぜなら，既存の多くの並列SGD手法は，勾配やモデルパラメータを頻繁に計算ノード間で交換する必要があるか，凸関数の性質に依存しているためである（例えば，[McDonald 10, Recht 11]）．DNNは一般的に3 000万次元程度の密なパラメータをもつ．この場合，各要素を4 Byteで表現するとして120 MByte程度のベクトルを交換しなければいけないこととなり，頻繁なパラメータや勾配ベクトルの共有は難しい．

データ並列の一つの方策は，フルバッチ学習，すなわちすべてのデータに関する勾配の和を計算した後，パラメータを勾配方向に移動させる方法である．フルバッチ学習であれば，各計算ノードで計算した勾配の和を，すべての学習データを処理した後で，マージするだけで済むため，通信量の問題は軽減される．ただし，フル

バッチ学習はミニバッチを用いたSGDと比べ，収束に至るまでの計算量が非常に多いため，効率の良い最適化アルゴリズムを利用する必要がある．Hessian-Free法[Martens 10]は二次勾配の情報を，明示的にそれを計算することなく最適化を行うテクニックであり，バッチ学習での高速な収束が期待されている．[Kingsbury 12]では，このHessian-Freeを用いることで，後述する系列識別学習を効率的に行う方法が示された．

ミニバッチ型のSGDを利用したまま並列学習を可能にする手法も多く検討されている．DistBelief[Dean 12]は各計算ノードがDNNの一部のパラメータのみを担当するモデル並列のテクニックを併用し，担当部分以外の更新頻度を下げることで，全体のパラメータ共有の頻度を下げながらも，精度を落とさず最適化する手法を示した．より直接的な通信量の削減の手段として，勾配ベクトルの量子化を行うことも提案されている[Seide 14]．特にsigmoidやtanh活性化関数を用いた場合は，勾配ベクトルの要素は多くの場合，0近辺に集中する．またReLUの場合，勾配ベクトルは疎となる．こうした場合，1-bit量子化を行い，一定以上の勾配がある要素に関してのみ，他の計算ノードのパラメータを更新することで通信量を大幅に削減することができる．

[Strom 15]はこの1-bit量子化を[Recht 11]のような非同期SGD技術と組み合わせて利用する手法で，複数のGPUを用いてGPUの数に対して性能がほぼ線形スケールするデータ並列学習を実現している．DistBeliefと異なり，モデル並列を仮定しない．すなわち，各ノードはすべての勾配を計算する．しかし，1-bit量子化が効率良く働き，他のノードと共有すべきデータの量が極めて小さく抑えられるため，他の非同期SGD技術を応用できる．

3·3 系列識別学習

DNN-HMMハイブリッドアプローチでは音声認識の問題をHMM状態の識別問題に簡略化する．しかし，HMM状態変数をラベルとしたクロスエントロピー規準は，音声認識の単語エラー率を直接的に表現する学習規準ではない．そのため，音声認識の性能を最大化するために単語エラー率や音素エラー率を考慮した最適化を行う系列識別学習技術は非常に重要である．

系列識別学習は音声認識の単語エラー率を近似的に表現した損失関数を最小化することで，DNNを直接的に最適化する技術である．系列識別学習そのものについては，深層学習が一般的になる以前から混合正規分布の枠組みで精力的に検討されてきた[He 08]．

以降では，学習データ中の入力ベクトル系列を$\boldsymbol{X}_n$とし，対応するラベル系列を$\boldsymbol{l}_n$と定義する．また，最適化の目的関数のn番目の学習データに関する項をL_nと置き，実際の目的関数は$L \stackrel{\text{def}}{=} \sum_n L_n$とする．

MMI規準は，正解ラベル列が出力される確率を直接的に最大化することを目指し，以下のMMI目的関数を用いる[Kapadia 93]．

$$L_n^{\text{MMI}}(\Theta;\boldsymbol{l}_n,\boldsymbol{X}_n) \stackrel{\text{def}}{=} -\log\frac{p(\boldsymbol{X}_n|\boldsymbol{l}_n,\Theta)p(\boldsymbol{l}_n)}{\sum_{\boldsymbol{l}'}p(\boldsymbol{X}_n|\boldsymbol{l}',\Theta)p(\boldsymbol{l}')} \tag{4}$$

この目的関数は言語モデル$p(\boldsymbol{l}_n)$によるバイアスを考慮したConditional Random FieldもしくはConditional Neural Fieldを構成しているともいえる．定義からわかるように，HMM状態系列変数を一致させるのではなく，出力ラベル列と正解ラベル列を一致させるようにパラメータを調整するために導入される目的関数である．

音声認識の問題において，ラベル列全体の一致は系列長が長くなるほど難しい．言い換えれば，一文の長さが長くなるほど，単語の誤りが含まれている可能性は高くなるため，MMI目的関数を用いた場合のように，文レベルの正解率を追求することの意義に関しては疑問が残る．そのため，実際の音声認識の評価尺度である単語エラー率や音素エラー率に即した学習も望まれてきた．bMMI学習は，正解を$\boldsymbol{l}_n$としたときの仮説$\boldsymbol{l}'$の単語エラー数（もしくは音素エラー数など）の近似関数$\boldsymbol{D}(\boldsymbol{l}_n;\boldsymbol{l}')$を用いて，その上界となる目的関数を最小化することによって実現される[Povey 08]．

パラメータがΘのとき，n番目の学習データに関する予測ラベル系列を

$$\hat{\boldsymbol{l}}_n = \operatorname*{argmax}_{\boldsymbol{l}'} \log p(\boldsymbol{l}'|\boldsymbol{X}_n,\Theta)$$

と置くと，以下が成立する．

$$\begin{aligned}
&D(\boldsymbol{l}_n;\hat{\boldsymbol{l}}_n)\\
&\le -\log p(\boldsymbol{l}_n|\boldsymbol{X}_n,\Theta) + D(\boldsymbol{l}_n;\hat{\boldsymbol{l}}_n) + \log p(\hat{\boldsymbol{l}}_n|\boldsymbol{X}_n,\Theta)\\
&\le -\log p(\boldsymbol{l}_n|\boldsymbol{X}_n,\Theta) + \log\sum_{\boldsymbol{l}'} e^{D(\boldsymbol{l}_n;\boldsymbol{l}')+\log p(\boldsymbol{l}'|\boldsymbol{X}_n,\Theta)}\\
&= -\log\frac{p(\boldsymbol{X}_n|\boldsymbol{l}_n,\Theta)p(\boldsymbol{l}_n)}{\sum_{\boldsymbol{l}'}p(\boldsymbol{l}'|\boldsymbol{X}_n,\Theta)p(\boldsymbol{l}')e^{D(\boldsymbol{l}_n;\boldsymbol{l}')}} \stackrel{\text{def}}{=} L_n^{\text{bMMI}}
\end{aligned} \tag{5}$$

したがってbMMIの目的関数L_n^{bMMI}を減少させることによって，予測ラベルに関するエラー尺度$\boldsymbol{D}(\boldsymbol{l}_n;\hat{\boldsymbol{l}}_n)$を間接的に減少させることができる．

Minimum Phone Error（MPE）[*2]規準[Povey 02]は，直接的に$\boldsymbol{D}$の期待値を以下の目的関数で表現し，最小化する手法である．

$$\begin{aligned}
L_n^{\text{MPE}} &= \Big\langle D(\boldsymbol{l}_n;\hat{\boldsymbol{l}}_n)\Big\rangle_{p(\boldsymbol{l}'|\boldsymbol{X}_n,\Theta)}\\
&= \frac{\sum_{\boldsymbol{l}'}p(\boldsymbol{X}_n|\boldsymbol{l}',\Theta)p(\boldsymbol{l}')D(\boldsymbol{l}_n;\boldsymbol{l}')}{\sum_{\boldsymbol{l}'}p(\boldsymbol{X}_n|\boldsymbol{l}',\Theta)p(\boldsymbol{l}')}
\end{aligned} \tag{6}$$

ここで$\langle f(\cdot)\rangle_{p(\cdot)}$は$f$の分布$p$に関する期待値である．

---

*2 Dの設定によってはState-level Minimum Bayes Risk（sMBR）とも呼ばれる[Kingsbury 09]．

系列識別学習の特に難しい点は，目的関数に登場する全単語列に関する総和（$\sum_{l'}$）である．これは，音声認識の結果得られる仮説候補のリスト，もしくはそれを効率良く表現する有向グラフ（ラティスと呼ばれる）上での総和で近似されるが，ラティスの再計算は非常に高コストであるため，SGDの各パラメータ更新の度にこれを再計算することは難しい．そこで，実用上はこのラティスは一定量のデータを処理するたびに再計算するという方法で現実的な学習を実現している．しかし，このラティスの更新頻度の低さが，目的関数の近似精度に悪影響を及ぼし，最適化が正しく進行しない場合も多い．

[Su 13]は，系列識別学習の目的関数とクロスエントロピー学習の目的関数の重み付き和を最小化することで，このような問題を避けることができることを示した．[Heigold 14]は系列識別目的関数の勾配ベクトル計算のうち，計算にラティスが必要な変数のみを非同期的に更新しながら最適化を続けるという枠組みを提示した．[Veselý 13]はラティス中で，正解HMM状態が表れていない区間に関しては無視することで，ラティスの近似が粗い箇所を避ける手法を示した．

このようなさまざまな安定化テクニックを駆使して，系列識別学習は広く利用されている．これらの導入によるエラー数の削減は，だいたい5～10%であると報告されている．すなわち従来100個の単語エラーがあったところを，深層学習の導入で70に減らすことができ，さらに系列識別学習の導入で65程度に減らすことができる．

系列識別学習はHMM状態変数を用いたクロスエントロピー学習を直接置き換えるものではない．系列識別学習の目的関数はクロスエントロピー学習の目的関数よりさらに最適化しにくい形状をしていることが予想されており，系列識別学習の前に初期値を正しく設定することが重要なためである．通常はクロスエントロピー学習の結果を初期値として系列識別学習処理をさらに追加することで実装される．

4. モデルの拡張

学習技術の進展と並行して，モデルの表現能力そのものを向上させる検討も行われてきた．

ニューラルネットに基づくモデルの一つの利点は，ブラックボックス化された処理の中に，さまざまな情報を統合できる点である．この利点を生かして従来以上にさまざまな特徴量が統合されはじめている．さらに，基本特徴量の抽出，すなわち図1における入力信号から特徴量ベクトルを計算するステップもニューラルネットの一部として組み入れ，目的関数に沿った最適化を適用するという試みがなされており，これも興味深い．

4·1 さまざまな特徴の統合

ニューラルネットは，明示的に関係性をモデル化することが難しい変数を追加の入力変数として導入し，副次的な情報源として活用することができる．

こうした方法は大別して，ラベルの識別に直接関係していると考えられる特徴量を導入する方法と，逆に，ラベルの識別には関係ないが，入力特徴の補正に有効であると考えられる特徴量を導入する方法に分けられる．ここでは特に，後者の，入力特徴の補正を行う補助特徴量について，話者識別の情報を統合する手法を解説する．このアプローチで最も広く用いられている話者識別特徴量はi-vectorと呼ばれている．i-vector [Dehak 11]はある話者の音声特徴量の分布と，すべての話者の音声特徴量の分布の違いを，混合正規分布の平均ベクトルの差で表し，それを次元縮約したものと解釈できる[小川 14]．

i-vectorは以下のような混合正規分布を用いて，話者sから得られた特徴量ベクトル$\boldsymbol{x}_t^{(s)}$をモデル化した際に現れるベクトル$\boldsymbol{\lambda}^{(s)}$のMaximum-a-Posteriori（MAP）推定値として定義される．

$$p(\boldsymbol{x}_t^{(s)}|\bar{\Theta},\boldsymbol{T},\boldsymbol{\lambda}^{(s)}) = \sum_k \bar{\alpha}_k \mathcal{N}(\boldsymbol{x}_t^{(s)}; \bar{\mu}_k + \boldsymbol{T}_k\boldsymbol{\lambda}_k^{(s)}, \bar{\boldsymbol{\Sigma}}_k^{-1}) \tag{7}$$

ここで$\mathcal{N}(\boldsymbol{x};\boldsymbol{m},\boldsymbol{S}^{-1})$は平均$\boldsymbol{m}$，共分散行列$\boldsymbol{S}$の多変量正規分布密度関数である．また$\bar{\Theta} \stackrel{\text{def}}{=} \{\bar{\alpha}_k, \bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\Sigma}}_k | \forall k\}$はすべての話者を用いて学習された音声特徴量の混合正規分布パラメータであり，k番目の要素についての混合重み$\bar{\alpha}_k$，平均ベクトル$\bar{\boldsymbol{\mu}}_k$，共分散行列$\bar{\boldsymbol{\Sigma}}_k$から構成される．$\boldsymbol{T}_k$はi-vectorから平均ベクトルの差への写像を表す行列である$\boldsymbol{\lambda}^{(s)}$の事前分布は$p(\boldsymbol{\lambda}^{(s)}) \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$と置く．

この設定で$\boldsymbol{\lambda}^{(s)}$とすべての$k$についての$\boldsymbol{T}_k$を推定することによって得たi-vector $\boldsymbol{\lambda}^{(s)}$を通常の音響特徴量に加えてニューラルネットに入力することで話者の情報，より正確には同じ話者ラベルsをもつ音響特徴量の傾向を吸収した識別を行うことができる．

4·2 特徴統合プロセスの統合

本章では，人手によってデザインされた特徴ではなく，より原始的な特徴量を用いて認識を行うことで，よりタスクに適した特徴量抽出を学習によって獲得することを目的とした手法について述べる．

図3に，混合正規分布による音響モデルとともに使われてきた，メル周波数ケプストラム係数（MFCC）特徴ベクトルの抽出処理を示す*3．MFCCは声道の共鳴特性を表現するために，音声の周波数スペクトルの対数の概形（対数スペクトル包絡）を聴覚と同様，低い周波数領域の違いをより細かく表現するように変換したものである．

*3 単にMFCCといった場合，時間微分処理（図3中の(f)）は含まれないが，多くの応用例で時間微分情報も同時に使われるため，本稿ではその処理も含めてMFCCの抽出プロセスとして扱う．

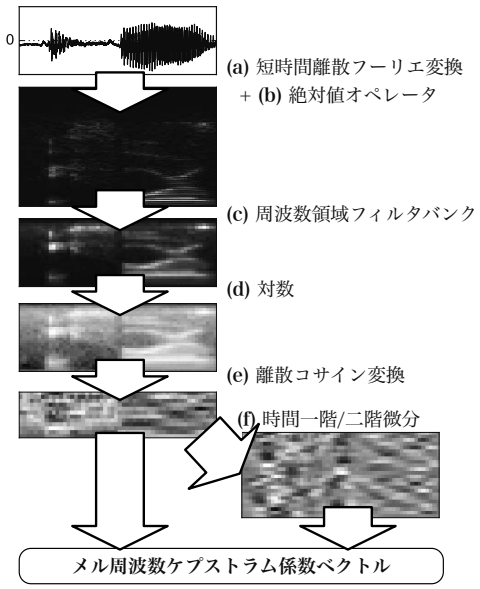

図 3　MFCC 特徴量ベクトルの計算

図 3 中の（a），（c），（e），（f）は入力の一部に対して一定の線形変換を行う処理であり，（b），（d）は線形変換の結果として得られたベクトルの要素ごとに何らかの非線形変換を行う処理である．このように，音声の特徴抽出プロセスは DNN と同様，線形変換と要素ごとの非線形変換という階層構造をもっているとみなせる．ただし，従来，これらの線形変換パラメータは学習によって得るものではなく，先験的な知識によって設定されるものであった．

[Sainath 14] では，この構造をほぼそのまま DNN とみなし，（c）以降の線形変換パラメータの再調整を行う手法である．単純に SGD によって線形変換パラメータを再学習するのではなく，入力スケールをそろえるための正規化処理を（b）の直後に追加し，（c）の係数を対数でもつことによって最適化の安定性を向上している．

この考えをさらに推し進めた手法に，CNN を用いて時間信号から直接学習を行うアプローチがある [Sainath 15a, Tüske 14]．このアプローチでは，（a），（b），（c）の処理を周波数領域のフィルタバンクで行うのではなく，時間信号に直接インパルス応答を CNN を用いて畳み込むことで時間領域のフィルタバンクを構成する．時間領域のフィルタバンクは，元の信号と同じだけの時間分解能をもち，音声認識の問題には細かすぎるので，畳込み層の直後に 10 ms ごとに一つの要素を出力するような max プーリング層を追加し，従来の音声認識と同じ時間分解能に縮約している．こうして，従来人手でデザインされてきた特徴抽出プロセスをタスクに合わせて直接推定する枠組みが確立されつつある．

加えてマイクロフォンアレーから得られるマルチチャネル信号への拡張を実現した手法もある [Sainath 15b]．複数のマイクロフォンを用いて録音した音声から目的の音声を見つける手法の一つである Delay-and-Sum ビームフォーマは，位相の異なる複数のフィルタを各チャネルに適用し，出力を足し合わせることに相当する．よって，この複数のフィルタが CNN 上に再現されることを期待して，マルチマイクロフォンの入力を直接マルチチャネルのフィルタをもつ CNN で処理することもできる．最先端のビームフォーミング技術と異なり，フィルタの特性を適応的に変化させることはできないが，十分に多くのフィルタを用意したうえで，後段の DNN にどの組合せを用いるかを選択させることで，擬似的に話者方向に適応した音声認識を実現することができる．

4·3　再帰結合ニューラルネットの利用

音声認識は時系列ラベリング問題であるため，再帰結合ニューラルネット（Recurrent Neural Net，以下 RNN）を用いて自然に表現することができる．しかし，[Hochreiter 97] で触れられているように RNN の学習は勾配消失問題によって難しい．音声特徴が時間的に影響を及ぼし合う範囲は 250 ms，すなわち典型的な設定では 25 フレームを超えるともいわれており [Morgan 05]，単純な RNN ではこのような長時間の依存関係は，実質的に学習できない．

Long Short-Term Memory（LSTM）は，このような問題に有効である [Graves 13, Hochreiter 97]．以降に紹介する技術では LSTM が主に用いられる．RNN や LSTM も基本的には一つの入力ベクトルを，時系列を考慮しながらも一つの出力ラベルと関連付けるニューラルネットであるが，DNN-HMM ハイブリッド方式を通して単語辞書や言語モデルと統合することができる [Graves 13]．

Connectionist Temporal Classification（CTC）[Graves 06] は，本来出力するラベルに加えて，「何もない」を意味するラベル ϕ を加えることで，音声認識のように出力ラベルのほうが入力ベクトルより少ない場合の系列マッピングを実現する方法である．これは予測モデルとしては，各音素ごとに一様分布をもつ HMM 状態を導入した DNN-HMM ハイブリッド方式とみなすこともできる．従来の HMM-DNN ハイブリッドによる方法では，すべての入力が対応するラベルをもち，その推定を行うようにパラメータが推定されるが，CTC においては，正しい出力系列を出すラベルであればどこに ϕ が挿入されてもよいという基準で学習される．そういった意味で，CTC は HMM の構造の工夫と言語モデル重みなどを考慮しない系列識別学習の複合技術であるともいえる．

近年では，CTC モデルの学習を，系列識別学習を用いて行う場合もある [Sak 15]．この場合，学習規準は先述した系列識別学習規準と同じであるため，CTC モデルと通常の HMM-DNN ハイブリッドモデルの差異は一様分布を出力するラベル ϕ の導入によって起こる特殊なパラメータ構造のみである．この特殊なパラメータ構造によって決まる識別モデルでは，時間的に局所的な情報

のみが識別に関与し，他の情報は捨てられる．このような時間的に局所的な情報のみが識別に寄与するという考えは画像認識のアテンションモデルに通じる．音声認識の分野でも初期的な検討が行われてきたが，確率モデルの枠組みとの統合が課題であった [Hasegawa-Johnson 05, Okawa 95]．深層学習の進展により，この仮説がDNNを通して復活した点は非常に興味深い．

5. 予測の高速化とパラメータサイズの削減

学習の高速化については上に述べたが，音声認識技術はモバイルデバイスなどに実装されることも多いため，認識段階の高速化も重要である．音響モデルの評価速度は，GPUを用いて実装する限りはあまり問題にならないが，低速なプロセッサを用いてシステムを構築する際には大きな問題となる．

ニューラルネットによる予測で最も時間を要する部分は行列積の計算である．この部分をベクトル整数演算が可能なCPUにおいて高速に処理することを目的として，重み行列と入力ベクトルの量子化を行う手法が提案されている [Vanhoucke 11]．この検討は，高速化の面でも興味深いが，学習後の重みパラメータを比較的粗く量子化してしまっても，精度があまり損なわれないという面でも興味深い．加えて，一度に複数フレームの予測を行うことでニューラルネットを駆動する頻度を減らすという試みも見られる [Vanhoucke 13]．

ニューラルネットのサイズ削減も予測の高速化に重要である．[Xue 13] では特異値分解（Singular Value Decomposition）を用いた重み行列の縮約を行う．

重み行列 $\boldsymbol{W}\in\mathbb{R}^{D\times D'}$ の特異値分解は以下のように表される．

$$\boldsymbol{W}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}} \tag{8}$$

ここで，$\boldsymbol{\Sigma}\in\mathbb{R}^{D\times D'}$ は対角要素以外はすべて0で埋められた行列であり，i 番目の対角要素を $\sigma_i=[\boldsymbol{\Sigma}]_{i,i}$ と表す．この σ_i は $i<j$ のとき $\sigma_i\geq j$ が成り立ち，また $\boldsymbol{U}\in\mathcal{R}^{D\times D}$ と $\boldsymbol{V}^{\mathsf{T}}\in\mathcal{R}^{D'\times D'}$ の各行は正規直交基底をなす（すなわち $\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}}=\boldsymbol{I}$ かつ $\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}=\boldsymbol{I}$）．

ここで $\boldsymbol{\Sigma}$ の大きいほうから R 個の要素を残し，残りを0で置き換えた行列 $\tilde{\boldsymbol{\Sigma}}$ を用いた $\tilde{\boldsymbol{W}}\stackrel{\text{def}}{=}\boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{V}^{\mathsf{T}}$ は，元の重み行列 $\boldsymbol{W}$ をフロベニウスノルム最小で近似する R-rank の行列となる．$\tilde{\boldsymbol{\Sigma}}$ が対角要素のみをもつことと，対角要素が大きい順に並んでいることから，この近似行列は以下のように書くこともできる．

$$\tilde{\boldsymbol{W}}=\underbrace{\boldsymbol{U}_{0:D,0:R}\boldsymbol{\Sigma}'}_{\boldsymbol{U}'}\underbrace{\boldsymbol{\Sigma}'\left(\boldsymbol{V}_{0:D,0:R}\right)^{\mathsf{T}}}_{\boldsymbol{V}'^{\mathsf{T}}} \tag{9}$$

ここで，$\boldsymbol{\Sigma}'\stackrel{\text{def}}{=}\mathrm{diag}(\sqrt{\sigma_1},\sqrt{\sigma_2},\cdots\sqrt{\sigma_R})$ は σ_1 から σ_R までの正の平方根を対角要素に並べた $R\times R$ 行列である．また，$\boldsymbol{U}_{0:D,\,0:R}$ および，$\boldsymbol{V}_{0:D',\,0:R}$ はそれぞれ元の行列 $\boldsymbol{U}$, $\boldsymbol{V}$ の最初の $\boldsymbol{R}$ 列で構成される部分行列である．

この行列分解による近似を重み行列として用いることで，識別結果を大きく変えることなく，計算コストを減らすことができる．しかし，この行列近似は元の行列をなるべく正確に近似することを目指しており，分解された後の識別性能を最大化するように設計されてはいないため，さらに性能向上の余地があると考えられる．そこで，この分解された行列表現をニューラルネットの一部として組み入れ，それを初期値とした再学習を行うことで，近似を行うことで生じた性能の劣化を軽減することができる．

蒸留（Distillation）と呼ばれるテクニックがモデルの縮約に用いられることもある [Hinton 14, Li 14]．蒸留は，より大きなニューラルネットで得られたエラー傾向に関する情報を用いて，簡潔なニューラルネットを学習するテクニックである．具体的には，大きなネットワークの出力分布を $q(y|\boldsymbol{x}_t)$ と置いたとき，出力確率値を $1/\tau$ 乗し，再度正規化した確率分布 $q^{(1/\tau)}(y|\boldsymbol{x}_t)$ を用いて，以下の最適化を行うことに相当する．

$$\text{minimize}\ \ L(\Theta)+\lambda\sum_t \mathrm{KL}[q^{(1/\tau)}(y|\boldsymbol{x}_t)||p(y|\boldsymbol{x}_t,\Theta)] \tag{10}$$

ここで λ および τ はハイパーパラメータ，$L(\Theta)$ は本来の学習の目的関数，例えばクロスエントロピー目的関数である．また $\mathrm{KL}[q||p]$ は p から q へのKLダイバージェンスである．この最適化は，学習しようとしているモデル Θ を，L を小さくするだけではなく，p が $q(1/\tau)$ に十分に近いように学習することを目指している．

この加えられた正則化項が，どうして元のモデルと同じ精度をより小さいモデルで達成できるのかは，まだ明確にはわかっていない．しかし，これを用いることで，大きなニューラルネットのアンサンブルの出力を小さな単一のニューラルネットで表現することが可能であると報告されている [Hinton 14]．また，上の定式化は半教師あり学習と組み合わせることもできる．正則化項の定義にはラベル情報が必要ないため，正則化項にラベルなしの学習データを加えることもできる．

6. 今 後 の 展 望

音声認識分野におけるニューラルネットは深層学習の導入以降も着実に進歩し，より高精度な認識を実現する礎となっている．本稿では残念ながら触れることはできなかったが，まだエンコーダ・デコーダネット [Lu 15] やアテンションモデル [Ba 15] など，大きな改善幅を期待できる拡張も提案され続けており，今後も目が離せない．学習・評価の計算量効率改善と，このように複雑なモデルの導入によって，これからも精度の向上が続くだろう．

最後に音声合成，音声言語理解など他の音声関連分野

においても深層学習が進展していることも触れておきたい．音声認識の精度がまた1段階上がったことにより，機械による本格的な音声コミュニケーションシステムの実現がまた一歩近づいた．より高度な音声コミュニケーションを実現するにあたり，これらの分野での深層学習技術も広く望まれており，また実際にさまざまな成果が出始めている．

音声認識は，問題設定を難化させながら，少しずつ進歩してきた．深層学習によって，多少ノイズの入った環境で，人が自然に話した音声でも，かなりの精度で認識できるようになってきたと考えている．多人数会話の状況理解，世界中の言語に対応した認識器の効率的な構築法，話者プライバシーと認識精度の両立など，音声認識技術の精度が上がることによって見えてきた音声技術全体の新たな展開にこれからも期待したい．

◇ 参 考 文 献 ◇

[麻生 15] 麻生英樹，安田宗樹，前田新一，岡野原大輔，岡谷貴之，久保陽太郎，ボレガラ ダヌシカ：深層学習― Deep Learning ―，近代科学社（2015）

[Ba 15] Ba, J., Grosse, R., Salakhutdinov, R. and Frey, B.: Learning wake-sleep recurrent attention models, *Advances in Neural Information Processing Systems*（2015）

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

[Dahl 13] Dahl, G. E., Sainath, T. N. and Hinton, G. E.: Improving deep neural networks for LVCSR using rectified linear units and dropout, *Proc. ICASSP,* pp. 8609-8613（2013）

[Dean 12] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Senior, A., Tucker, P., Yang, K. and Le, Q. V., et al.: Large scale distributed deep networks, *Advances in Neural Information Processing Systems,* pp. 1223-1231（2012）

[Dehak 11] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet,P.: Front-end factor analysis for speaker verification, *IEEE Trans. on Audio, Speech, and Language Processing,* Vol. 19, No. 4, pp. 788-798（2011）

[Erhan 10] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P. and Bengio, S.: Why does unsupervised pre-training help deep learning?, *J. Machine Learning Research,* Vol. 11, pp. 625-660（2010）

[Graves 06] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, *Proc. ICML,* pp. 369-376, ACM（2006）

[Graves 13] Graves, A., Jaitly, N. and Mohamed, A.-R.: Hybrid speech recognition with deep bidirectional LSTM, *Proc. IEEE Workshop on ASRU,* pp. 273-278（2013）

[Hasegawa-Johnson 05] Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K. and Mohan, S., et al.: Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop, *Proc. ICASSP,* Vol. 1, p. 1213（2005）

[He 08] He, X., Deng, L. and Chou, W.: Discriminative learning in sequential pattern recognition, *IEEE Signal Processing Magazine,* Vol. 25, No. 5, pp. 14.36（2008）

[Heigold 14] Heigold, G., McDermott, E.,Vanhoucke,V., Senior, A. and Bacchiani, M.: Asynchronous stochastic optimization for sequence training of deep neural networks, *Proc. ICASSP,* pp. 5587-5591（2014）

[Hinton 14] Hinton, G., Vinyals, O. and Dean, J.: Distilling the knowledge in a neural network, *Proc. Deep Learning and Representation Learning Workshop NIPS*（2014）

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

[Kapadia 93] Kapadia, S., Valtchev, V. and Young, S.:MMI training for continuous phoneme recognition on the TIMIT database, *Proc. ICASSP,* Vol. 2, pp. 491-494（1993）

[Kingsbury 09] Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling, *Proc. ICASSP,* pp. 3761-3764（2009）

[Kingsbury12] Kingsbury, B., Sainath,T. N. and Soltau, H.: Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization, *Proc. INTERSPEECH*（2012）

[久保 14] 久保陽太郎：音声認識のための深層学習，人工知能，Vol. 29, No. 1, pp. 62-71（2014）

[LeCun 98] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proc. IEEE,* Vol. 86, No. 11, pp. 2278-2324（1998）

[Li 14] Li, J., Zhao, R., Huang, J.-T. and Gong, Y.: Learning small-size DNN with output-distribution-based criteria, *Proc. INTERSPEECH*（2014）

[Lu 15] Lu, L., Zhang, X., Cho, K. and Renals, S.: A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition, *Proc. INTERSPEECH*（2015）

[Martens 10] Martens, J.: Deep learning via Hessian-free optimization, *Proc. ICML,* pp. 735-742（2010）

[McDonald 10] McDonald, R., Hall, K. and Mann, G.: Distributed training strategies for the structured perceptron, *Proc. NAACL,* pp. 456-464, Association for Computational Linguistics（2010）

[Morgan 05] Morgan, N., Zhu, Q., Stolcke, A., Sönmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H. and Ellis, D., et al.: Pushing the envelope-aside, *IEEE Signal Processing Magazine,* Vol. 22, No. 5, pp. 81-88（2005）

[Nair 10] Nair,V. and Hinton,G. E.: Rectified linear units improve restricted boltzmann machines, *Proc. ICML,* pp. 807-814（2010）

[小川 14] 小川哲司，塩田さやか：i-vector を用いた話者認識，日本音響学会誌，Vol. 70, No. 10, pp. 332-339（2014）

[Okawa 95] Okawa, S. and Shirai, K.: Estimation of statistical phoneme center and its application to accurate phoneme modelling, *Proc. EUROSPEECH,* pp. 791-794（1995）

[Plahl 12] Plahl, C., Sainath, T. N., Ramabhadran, B. and Nahamoo, D.: Improved pre-training of deep belief networks using sparse encoding symmetric machines, *Proc. ICASSP,* pp. 4165-4168（2012）

[Povey 02] Povey, D. and Woodland, P. C.: Minimumphone errorand I-smoothing for improved discriminative training, *Proc. ICASSP,* Vol. 1, pp. I-105（2002）

[Povey 08] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah, K.: Boosted MMI for model and feature-space discriminative training, *Proc. ICASSP,* pp. 4057-4060（2008）

[Recht 11] Recht, B., Re, C., Wright, S. and Niu, F.: Hogwild: A lock-free approach to parallelizing stochastic gradient descent, *Advances in Neural Information Processing Systems,* pp. 693-701（2011）

[Sainath 14] Sainath, T. N., Kingsbury, B., Mohamed, A.-R., Saon, G. and Ramabhadran, B.: Improvements to filterbank and delta learning within a deep neural network framework, *Proc. ICASSP,* pp. 6839-6843（2014）

[Sainath 15a] Sainath, T. N., Weiss, R. J., Senior, A.,Wilson, K. W. and Vinyals, O.: Learning the speech front-end with raw wave form CLDNNs, *Proc. INTERSPEECH*（2015）

[Sainath 15b] Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A., Bacchiani, M. and Senior, A.: Speaker location and microphone spacing invariant acoustic modeling from raw multi-channel wave forms, *Proc. ASRU*（2015）

[Sak 15] Sak, H., Senior, A., Rao, K., Irsoy, O., Graves, A., Beaufays, F. and Schalkwyk, J.: Learning acoustic frame labeling

for speech recognition with recurrent neural networks, *Proc. ICASSP,* pp. 4280-4284（2015）
[Seide 11] Seide, F., Li, G., Chen, X. and Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription, *Proc. IEEE Workshop on ASRU,* pp. 24-29（2011）
[Seide 14] Seide, F., Fu, H., Droppo, J., Li, G. and Yu, D.: 1-bit stochastic gradient descent and its application to data-parallel distributed trainingof speech DNNs, *Proc. INTERSPEECH*（2014）
[Strom 15] Strom, N.: Scalable distributed DNN training using commodity GPU cloud computing, *Proc. INTERSPEECH*（2015）
[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,* pp. 6664-6668（2013）
[Tüske 14] Tüske, Z., Golik, P., Schlüter, R. and Ney, H.: Acoustic modeling with deep neural networks using raw time signal for LVCSR, *Proc. INTERSPEECH*（2014）
[Vanhoucke 11] Vanhoucke, V., Senior, A. and Mao, M. Z.: Improving the speed of neural networks on CPUs, *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop,* Vol. 1（2011）
[Vanhoucke 13] Vanhoucke, V., Devin, M. and Heigold, G.: Multi-frame deep neural networks for acoustic modeling, *Proc. ICASSP,* pp. 7582-7585（2013）
[Veselý 13] Veselý, K., Ghoshal, A., Burget, L. and Povey, D.: Sequence-discriminative training of deep neural networks, *INTERSPEECH,* pp. 2345-2349（2013）
[Xue 13] Xue, J., Li, J. and Gong,Y.: Restructuring of deep neural network acoustic models with singular value decomposition, *INTERSPEECH,* pp. 2365-2369（2013）
[Zhang 14] Zhang, X., Trmal, J., Povey, D. and Khudanpur, S.: Improving deep neural network acoustic models using generalized max-out networks, *Proc. ICASSP,* pp. 215-219（2014）

2016年1月18日 受理

著 者 紹 介

久保 陽太郎（正会員）

2010年早稲田大学大学院基幹理工学研究科博士課程修了．博士（工学）．同年，ドイツRWTHアーヘン工科大学客員研究員を経て，日本電信電話株式会社に入社，NTTコミュニケーション科学基礎研究所に配属．2014年よりAmazon Development Center Germanyにて Speech Scientistとして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，IEEE SPS Japan Chapterより Student Journal Paper Award，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．