情報処理学会論文誌 Vol.58 No.5 1207–1217 (May 2017)

音声認識のための結合フレーム中心近傍を強調した Deep Neural Network の学習

倉田 岳人[1,a)] ヴィレット ダニエル[2]

受付日 2016年1月21日, 採録日 2017年2月9日

概要：音声認識における特徴抽出，および音響モデルに Deep Neural Network（DNN）を利用する場合，連続する複数フレームの音響特徴量を結合して利用することが一般的である．DNN は，結合された複数フレームの音響特徴量から，学習データ中で中心フレームに対応付けられた音素の隠れマルコフモデル（Hidden Markov Model: HMM）の状態を予測するように学習される．つまり，中心フレームやその近傍のフレーム，そして中心から離れたフレームすべてが同等に扱われ，結合された音響特徴量全体と，中心フレームの音素 HMM 状態が対応付けられることとなる．中心から離れたフレームは，中心フレームに対応する音素 HMM の状態と関係する情報を持つが，結合された音響特徴量全体がそのまま未知のデータにも出現するわけではない．本論文では，DNN は，中心フレームに対応する音素 HMM 状態を予測することを考慮し，結合されたフレームの中心フレームとその近傍を含む中心近傍部分を重視する DNN 学習方法を提案する．具体的には，最初に中心近傍部分のフレームのみを利用して DNN の Pre-training と Fine-tuning を行い，得られた DNN から，すべての結合されたフレームを利用してさらに Fine-tuning を行う方法を提案する．この方法を「2 段階 Fine-tuning」と呼ぶ．最初に英語の標準的な放送番組データセットを利用した実験により，提案手法による音声認識率の改善を確認した．次に，実際に運用されている音声認識システムから得られた 1,000 時間以上の学習データを利用して，その実運用システムに本手法を適用し，実運用システムに由来する制約の中でも，認識率の改善が得られることを確認した．

キーワード：Deep Neural Network，大語彙連続音声認識，結合フレーム

Deep Neural Network Training Emphasizing Central Frames for Speech Recognition

Gakuto Kurata[1,a)] Daniel Willett[2]

Received: January 21, 2016, Accepted: February 9, 2017

Abstract: It is a standard approach to concatenate several consecutive frames of acoustic features as input of a Deep Neural Network (DNN) for an acoustic model in speech recognition. A DNN is trained to map the concatenated frames as a whole to the Hidden Markov Model (HMM) state corresponding to the center frame while the side frames close to both ends of the concatenated frames and the remaining central frames are treated as equally important. Though the side frames are relevant to the HMM state of the center frame, this relationship may not be fully generalized to unseen data. Since a DNN predicts the HMM state of the central frame, we propose a new DNN training method to emphasize the central frames. We first conduct pre-training and fine-tuning with only the central frames and then conduct fine-tuning with all of the concatenated frames. We call this method "two-stage fine-tuning". We conducted experiments with the standard English Broadcast News data and confirmed that the proposed method improved the accuracy of speech recognition over the competitive baseline systems. Then, in large-scale experiments with more than 1,000 hours of training data collected in the real-life application, we confirmed that the improvement in speech recognition accuracy was obtained by the proposed method on the deployed system.

Keywords: Deep Neural Network, large vocabulary continuous speech recognition, concatenated frames

1. はじめに

Deep Neural Network（DNN）は，GMM-HMM（Gaussian Mixture Model-Hidden Markov Model）システムによる音声認識システムでの特徴量抽出器，および DNN-HMM シ

[1] 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research - Tokyo, IBM Japan Ltd., Chuo, Tokyo 103–8510, Japan
[2] Nuance Communications
Nuance Communications, Aachen, Germany
[a)] gakuto@jp.ibm.com

ステムよる音声認識システムでの音響モデルとして利用されている [1], [2], [3], [4], [5]．このような用途での DNN は，一般的には連結された複数フレームの音響特徴量を受け取る入力層，多段の隠れ層，そして連結されたフレームの中心フレームに学習データ中でアライメントされている音素 HMM 状態を予測する出力層から構成される．

音声認識において，連続する複数フレームを連結することは，一般的に利用されてきた．たとえば，fMPE（Feature-space Miminum Phone Error）に代表される特徴量空間での識別学習では，音響特徴量を高次元空間に写像する際に，対象とするフレームの前後複数フレームの音響特徴量の事後確率も利用される [6]．しかし，前後のフレームから得られる事後確率は，数フレームごとに平均化して利用されるため，すべてのフレームが同等に扱われるわけではない．

GMM-HMM システムでの特徴量抽出器，もしくは DNN-HMM システムにおける音響モデルとして DNN が利用される場合，**図 1** に示したように，固定長の連続したフレームの音響特徴量が結合され，入力層に入力される．ここで，連結されるフレームの中で，中心フレーム，およびその近傍フレームを中心近傍フレームと呼び，それ以外の両端に近い部分のフレームを遠隔フレームと呼ぶ．DNN の出力層では，中心フレームにアライメントされた HMM 状態が予測されるが，DNN の学習としてはすべてのフレームが同等に扱われ，結合されたフレーム全体が，中心フレームにアライメントされた音素 HMM 状態を予測することになる．しかし，学習データ以外の未知のデータにおいては，必ずしも学習データ内の連結されたフレームがそのまま出現するわけではない．遠隔フレームの音響特徴量と中心フレームにアライメントされた HMM 状態の間には関連があるが，無関係な情報も含まれることを考慮し，本論文では，結合されたフレームの中で，遠隔フレームよりも中心近傍フレームを重視する DNN 学習手法を提案する．具体的には，最初に中心近傍フレームのみを利用して DNN の Pre-training と Fine-tuning を行い，得られた DNN を開始点とし，すべての結合されたフレームを利用してさらに Fine-tuning を行う方法を提案する．中心近傍フレームのみで学習された DNN は，中心フレームに対応付けられた HMM 状態を，中心近傍フレームのみからある程度予測できるため，その後の学習で利用される遠隔フレームは補助的に働くことになる．結果として，中心近傍フレームが重視されることとなり，遠隔フレームへの依存を軽減することができる．

本論文の構成は以下のとおりである．2 章では，GMM-HMM システムにおける特徴量抽出，および DNN-HMM システムにおける音響モデルとして利用される DNN の一般的な手法について概説する．その後，3 章では，提案手法について詳細に説明する．4 章では，提案手法の効果を検証するための評価実験を，大語彙連続音声認識タスクで行い，さらに提案手法により学習された DNN について分析を加える．最後に 5 章で，本論文をまとめる．

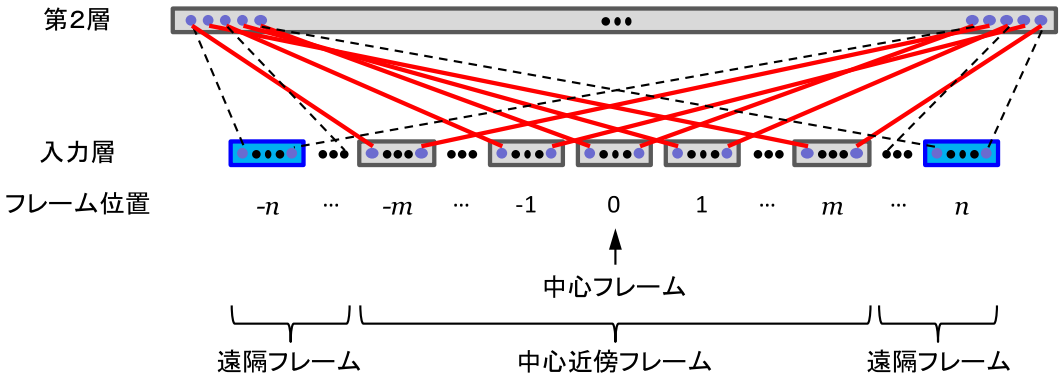

図 1 入力層と第 2 層の間の結合．結合された入力フレームの中心を中心フレームと呼び，このフレームにアライメントされた HMM 状態が，DNN の出力層で予測される．フレーム位置 $-m$ から m のフレームを中心近傍フレームと呼び，残りのフレームを遠隔フレームと呼ぶ．提案手法では，中心近傍フレームと第 2 層のすべてのユニットの間の結合を強調する．なお，図を煩雑にしないため，すべての結合を描いているわけではない

Fig. 1 Connections between input and second layers. We call the frame that is at the center of the concatenated frames the *center frame* whose associated HMM state is predicted in the DNN. We call the frames at positions from $-m$ to m the *central frames* and the remaining frames the *side frames*. The connections between the central frames and the units in the second layer are emphasized in the proposed method. Please note that only a fraction of the connections are depicted to reduce the complexity.

2. 従来手法

DNN は音声認識システムの特徴抽出器 [2], [3]，音響モデル [4], [5]，および言語モデルで利用されている [7], [8]．本論文では，GMM-HMM システムにおける特徴抽出器，および DNN-HMM システムにおける音響モデルでの DNN の活用を対象とする．DNN をこれらの方法に用いる場合，DNN 自身の学習方法は基本的には同じで，複数フレームを結合した音響特徴量を受理して，中心フレームにアライメントされた HMM 状態を識別するように学習される．GMM-HMM システムでは，学習された DNN の隠れ層の出力などを特徴量として利用し，DNN-HMM システムでは，出力層から音素状態の事後確率を得る．

DNN の学習は，一般的には Pre-training と Fine-tuning から構成される．最初の Pre-training では，層の数やユニット数などの DNN の構成を定義したうえで，その DNN の構成に従って，結合重みを初期化しながら層を積み上げる．近年の深層学習の進歩は，Pre-training における結合重み初期化手法の改善によるところが大きい．教師なし生成モデルを利用した手法として，Restricted Bolzmann Machines（RBMs）[9] が導入され，その後，autoencoder の利用も検討されている [10]．教師なし生成モデルは，音声認識以外の分野で提案された手法であり，それが音声認識でも効果が確認されている．しかし，音声認識では，音響特徴量系列に対して，HMM 状態系列が事前にアライメントされた学習データを利用することが一般的であり，教師なし生成モデルの代わりに，教師ありで識別的に Pre-training を行う手法も提案されている [11], [12], [13], [14]．なお，ここでの「教師あり」は，書き起こしデータが付与されている音声データを利用している，ということを意味し，その書き起こしデータが「教師あり」，つまり人手をともなって作成されたか，「教師なし」，つまり音声認識により自動生成されたかを区別しているわけではない．Pre-training により所定の構成の DNN が構築され，結合重みが初期化された後，その重みを微調整する Fine-tuning が行われる．Fine-tuning では，学習データ中でアライメントされた HMM 状態を参照し，誤差逆伝播により識別的に重みの調整を行う [15]．音声認識では，誤差関数として，フレームあたりのクロスエントロピーを利用することが広く行われている [1]．

音声認識は時系列データのラベリング問題と考えることができる．フレームあたりのクロスエントロピーを損失関数とする Fine-tuning では，出力層で扱う音素 HMM 状態を単位として識別しているが，時系列としての識別を行う Sequence Training を行うことで，さらに音声認識率が改善されることが報告されている [16], [17], [18]．ただし，Sequence Training に要する計算量は一般的に大きく，多くの実験条件での比較が困難であるため，本論文では，識別的な Pre-training [11] とクロスエントロピーを損失関数とする Fine-tuning で得られる DNN を基準に提案手法の効果を検討する．

3. 提案手法

DNN は結合されたフレームと，中心フレームにアライメントされた音素 HMM 状態の対応関係を学習するが，学習時にはすべてのフレームが完全に同等に扱われる．遠隔フレームの音響特徴量と中心フレームにアライメントされた HMM 状態の間には関連があるが，無関係な情報も含まれることを考慮し，本章では，DNN 学習時に，結合して入力されるフレームの中で，中心近傍フレームを強調する方法を提案する．これは，図 1 で，赤太線で示されている結合を強調することに相当する．この考え方を実現する方法として，本章では「**2 段階 Fine-tuning**」を提案する．

2 段階 Fine-tuning の処理の流れを，**図 2** に示した．この後の説明では，中心フレームの前後 n フレームを含めた $2n+1$ フレームを入力として受け取る DNN に対して，中心フレームの前後 m フレーム（$m < n$）を含めた $2m+1$ フレームを強調することとする．

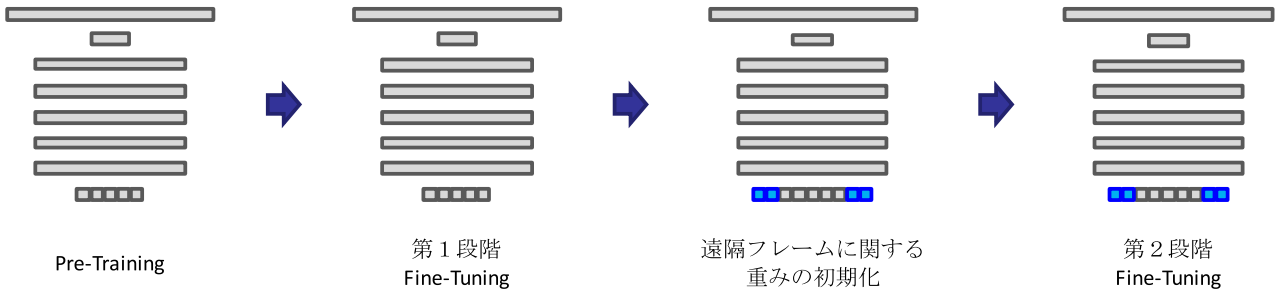

図 2 2 段階 Fine-tuning の流れ．最初に中心近傍フレームのみを利用して Pre-training と第 1 段階 Fine-tuning を行い，その後，すべての結合されたフレームを利用して第 2 段階 Fine-tuning を行う

Fig. 2 Flow of two-stage fine-tuning. We first conduct pre-training and fine-tuning with only the central frames and then conduct fine-tuning with all of the concatenated frames.

Pre-Training：フレーム位置が $-m$ から m の $2m+1$ フレームの中心近傍フレームを受け取り，入力層以外は所定の構成の DNN を構築し，Pre-training を行う．なお，Pre-training の際のパラメータの初期化は，文献 [19] の “Normalized Initialization” を利用してランダム値を割り当てた．その後，識別的 Pre-training でパラメータを調整する．識別的 Pre-trainig では，1 層追加するごとに，学習データ全体を 1 度走査し，フレームあたりのクロスエントロピーを目的関数として，誤差逆伝播により学習を行った [11].

第 1 段階 Fine-tuning：$2m+1$ フレームを利用して，Fine-tuning を行う．この段階を「第 1 段階 Fine-Tuning」と呼ぶ．

遠隔フレームに関する重みの初期化：フレーム位置が $-n$ から n までの $2n+1$ フレームを入力として受け取るように，DNN の構成を変更する．この構成の変更は，図 1 で破線で示されている，遠隔フレームから第 2 層のすべてのユニットへの結合を追加することに相当する．なお，追加された結合に対する重みは，Pre-training の最初の初期化と同様にランダム値を割り当て，他の結合重み，およびバイアスパラメータは，第 1 段階 Fine-tuning で推定されたものをそのまま利用する．2 層目以上の層に関しても，第 1 段階 Fine-tuning で推定されたパラメータをそのまま利用する．

第 2 段階 Fine-tuning：フレーム位置が $-n$ から n までの $2n+1$ フレームすべてを利用して，Fine-tuning を行う．この段階では，すべてのパラメータが誤差逆伝播により更新される．なお，ここで利用する学習データは，第 1 段階 Fine-tuning で利用する学習データと同一である．この段階を「第 2 段階 Fine-tuning」と呼ぶ．

第 1 段階 Fine-tuning で推定された DNN は，中心近傍フレームのみから，中心フレームに対応付けられた音素 HMM 状態をある程度予測できるため，第 2 段階 Fine-tuning で追加される遠隔フレームは，補助的に働くこととなる．

4. 評価実験

提案する DNN 学習手法により，音声認識精度が改善できるかを，2 種類の実験により検証した．最初に，様々な文献で利用されてきた英語の放送番組データセットを利用し，DNN-HMM ハイブリッドシステムを構築し，大語彙連続音声認識（Large Vocabulary Continuous Speech Recognition: LVCSR）実験を通じて，提案手法の効果を検証した．また，従来手法と提案手法により学習された DNN 内の入力層から第 2 層への結合重みを調べることで提案手法により，期待どおりに中心近傍フレームが重視されるようになっているかを検証した．次に，実運用されている音声認識システム上で，日本語の 1,000 時間以上の学習データを利用して，提案手法の優位性を検証した．

4.1 英語放送番組データでの評価

最初に，結合フレーム数，および利用する特徴量の比較を行い，ベースラインとする設定を決定する．次に，2 段階 Fine-tuning を様々な条件で行い，音声認識率の改善を確認する．最後に，学習された DNN に対して分析を加え，期待どおりに中心近傍フレームが重視されるようになっているかを検証する．

4.1.1 実験条件

音響モデルの学習には，英語の 50 時間の放送番組データセット（`1996, 1997 English Broadcast Speech corpora (LDC97S44, LDC98S71)` から番組単位で 50 時間を選択．話者数は 1359）を利用した．この中で，90%を学習用データに，残りの 10%をヘルドアウトデータとして利用した．評価には，2003 年 11 月の 6 番組からなる 3 時間の音声データ（`DARPA EARS Dev-04f set`）を利用した．評価時には，5,400 万エントリの n-gram を含む，語彙サイズ 8 万 4 千の単語 4-gram モデルを利用した．これらのデータに関しては，文献 [20] にさらに詳述されている．なお，音声認識精度の評価尺度として単語誤り率（WER: Word Error Rate）を利用した．

DNN に基づく音響モデルの構築の前に，GMM に基づく音響モデルを文献 [21] の手順に従って構築し，十分に精緻なアライメントを得た．得られたアライメントを利用して，識別的 Pre-training，Fine-tuning ともに，フレームあたりのクロスエントロピーを目的関数として，誤差逆伝播で学習を行った [11], [12]．DNN の学習中は，ヘルドアウトデータに対するフレームあたりのクロスエントロピーを確認して，改善が見られない場合には，学習率を 0.005 から順次減衰させていった [22]．Fine-tuning のエポック数は，従来手法の Fine-tuning，提案手法の第 1 段階 Fine-tuning，第 2 段階 Fine-tuning ともに 15 で固定した．

音声特徴量としては，声道長正規化された 40 次元の対数ログメルフィルタバンク（Log-mel）特徴量を利用した [23]．なお，結合フレーム数や，動的特徴量の利用に関しては，4.1.2 項で詳細に比較する．DNN の構成は，比較検討を行う入力層を除いてすべての実験で共通とし，512 個のユニットを持つ 6 層の隠れ層と 5,000 個のユニットを持つ出力層からなるようにした．出力層では，状態共有クインフォン HMM の状態を予測対象とした [24].

評価時には，学習された DNN から得られる各状態からの事後確率を利用する DNN-HMM ハイブリッドシステムを構築し，上記の言語モデル，語彙を利用して，LVCSR の実験を行った [22].

4.1.2 結合フレーム数と動的特徴量に関する比較

結合フレーム数を変化させ，また，静的な特徴量だけでなくデルタ特徴量（Δ），およびダブルデルタ特徴量（ΔΔ）を

表 1 入力フレーム数，および利用する動的特徴量を変化させた場合の単語誤り率 [%]
Table 1 Word Error Rate (WER) with changing number of input frames and type of delta features [%].

入力フレーム数	特徴量 Log-mel	Log-mel+Δ	Log-mel+Δ+ΔΔ
1 $(n=0)$	36.9	23.0	25.5
3 $(n=1)$	21.8	18.3	17.7
5 $(n=2)$	18.5	17.8	17.4
7 $(n=3)$	17.7	**17.5**	**17.2**
9 $(n=4)$	**17.4**	17.6	17.5
11 $(n=5)$	17.6	17.8	18.1
13 $(n=6)$	17.7	18.1	18.3

利用した場合について，音声認識率の比較を行った [25]．これ以降は，静的な特徴量を **Log-mel 特徴量**，デルタ特徴量も追加した場合を **Log-mel+Δ 特徴量**，ダブルデルタ特徴量まで考慮した場合を **Log-mel+Δ+ΔΔ 特徴量**と記述する．なお，Log-mel 特徴量だけの場合には，40 次元と入力フレーム数の積が DNN の入力ユニット数に，Log-mel+Δ 特徴量を利用する場合には，80 次元と入力フレーム数の積が DNN の入力ユニット数に，Log-mel+Δ+ΔΔ 特徴量を利用する場合には，120 次元と入力フレーム数の積が DNN の入力ユニット数となる．**表 1** に結果を示した．Log-mel 特徴量では，9 フレームを結合して入力とした場合に，単語誤り率が最低の 17.4%となった．それに対して，Log-mel+Δ 特徴量を利用した場合，7 フレームを結合して入力することで単語誤り率が 17.5%に，また，Log-mel+Δ+ΔΔ 特徴量を利用した場合，7 フレームを結合して入力することで，単語誤り率が 17.2%となった．これらを基準として，次項で，2 段階 Fine-tuning の効果を検証する．

4.1.3 2 段階 Fine-tuning の検証

各特徴量に対して，前項で最高性能を得た DNN の構成を最終的な構成とし，2 段階 Fine-tuning を行った．具体的には，Log-mel 特徴量のみを用いる場合には，最終的な DNN の構成として，前項で最高性能を得た 9 フレームを結合して入力する形式を利用することとし，最初に Pre-training と第 1 段階 Fine-tuning で利用するフレーム数を変化させて DNN の学習を行い，その後，9 フレームすべてを利用して第 2 段階 Fine-tuning を行った．図 1 で使っている n，m を利用して表現した場合，n を 4 で固定したまま，m を 0 から 3 に変化させて実験を行った．Log-mel+Δ 特徴量，および Log-mel+Δ+ΔΔ 特徴量を利用する場合には，n を 3 で固定したまま，m を 0 から 2 に変化させて実験を行った．

Log-mel 特徴量を利用した場合の結果を**表 2** に，Log-mel+Δ 特徴量を利用した場合の結果を**表 3** に，Log-mel+Δ+ΔΔ 特徴量を利用した場合の結果を**表 4** に示した．すべての特徴量の，すべての m と n の組合せで，2 段階 Fine-tuning により，単語誤り率が減少していることを確認した．最も単語誤り率が減少した場合について，各表では太字で示している．また，これらの改善は，有意水準 5%で統計的に有意であることを確認した．

また，表 1 と表 2，表 3，表 4 を比較すると，2 段階 Fine-tuning では，後半の第 2 段階 Fine-tuning で改善が得られ

表 2 2 段階 Fine-tuning を行った場合の単語誤り率と単語誤り率削減率（Log-mel）
Table 2 Word Error Rate (WER) and WER Reduction (WERR) with two-stage fine-tuning (Log-mel).

入力フレーム数	単語誤り率 (単語誤り率削減率) [%]
9 $(n=4)$ (ベースライン)	17.4
1 → 9 $(m=0, n=4)$	17.0 (2.30)
3 → 9 $(m=1, n=4)$	17.1 (1.72)
5 → 9 $(m=2, n=4)$	**17.0 (2.30)**†
7 → 9 $(m=3, n=4)$	17.1 (1.72)

† 入力フレーム数が 1→9 の場合よりも単語誤り数が小さいことを確認した．

表 3 2 段階 Fine-tuning を行った場合の単語誤り率と単語誤り率削減率（Log-mel+Δ）
Table 3 Word Error Rate (WER) and WER Reduction (WERR) with two-stage fine-tuning (Log-mel+Δ).

入力フレーム数	単語誤り率 (単語誤り率削減率) [%]
7 $(n=3)$ (ベースライン)	17.5
1 → 7 $(m=0, n=3)$	**16.9 (3.43)**
3 → 7 $(m=1, n=3)$	17.0 (2.86)
5 → 7 $(m=2, n=3)$	17.2 (1.71)

表 4 2 段階 Fine-tuning を行った場合の単語誤り率と単語誤り率削減率（Log-mel+Δ+ΔΔ）
Table 4 Word Error Rate (WER) and WER Reduction (WERR) with two-stage fine-tuning (Log-mel+Δ+ΔΔ).

入力フレーム数	単語誤り率 (単語誤り率削減率) [%]
7 $(n=3)$ (ベースライン)	17.2
1 → 7 $(m=0, n=3)$	**16.8 (2.33)**
3 → 7 $(m=1, n=3)$	16.9 (1.74)
5 → 7 $(m=2, n=3)$	17.0 (1.16)

ていることを確認することができる．たとえば，Log-mel 特徴量を利用した場合には，従来手法で 5 フレームを連結入力した場合の単語誤り率は 18.5%であった（表 1）が，第 2 段階 Fine-tuning を行うことで 17.0%に改善されている（表 2）．

なお，2 段階 Fine-tuning を行う場合，第 1 段階 Fine-tuning と第 2 段階 Fine-tuning で，各々のエポック数を 15 としたが，ヘルドアウトデータに対するクロスエントロピーは十分に飽和した．また，従来手法で DNN を学習する際にも，同様に Fine-tuning のエポック数を 15 としたが，クロスエントロピーは十分に飽和した．具体的に**図 3**，**図 4**，**図 5** に，Log-mel 特徴量，Log-mel+Δ 特徴量，および Log-mel+Δ+ΔΔ 特徴量を利用した場合について，従来手法による Fine-tuning の際のクロスエントロピーの推移と，2 段階 Fine-tuning で最も低い単語誤り率を得た場合の第 2 段階 Fine-tuning の際のクロスエントロピーの推移を示した[*1]．また，**表 5** に学習終了時のクロスエントロピーの値を示した．いずれのグラフにおいても，従来手法，提案手法ともに，十分にクロスエントロピーの観点で学習が飽和していることが確認できる．つまり，2 段階 Fine-tuning で得られた改善は，第 1 段階 Fine-tuning と第 2 段階 Fine-tuning での合計学習エポック数が増えたことに起因しているわけではない．また，表 5 からは，提案手法がクロスエントロピーの観点でも改善を得ていることを確認することができる．なお，エポック数が 0 の点は，初期化のみが完了した状態でのクロスエントロピーを示している．従来手法では，一番上の層がランダムに初期化されているため高いクロスエントロピーを示しているが，2 段階 Fine-tuning では，第 1 段階 Fine-tuning で中心近傍フレームを入力とした場合の DNN が学習されているため，最初から低いクロスエントロピーを示している．

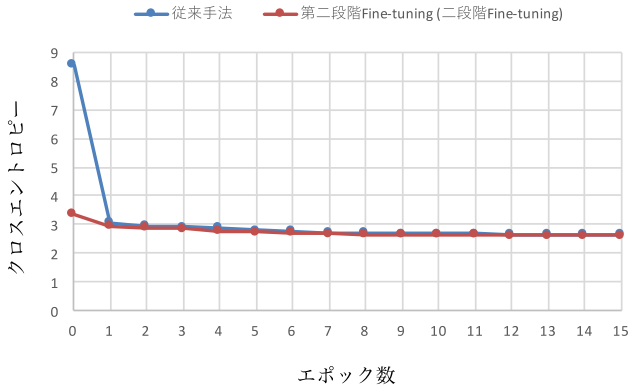

図 3 Log-mel 徴量を利用した場合のヘルドアウトデータに対するフレームあたりのクロスエントロピー

Fig. 3 Cross-entropy per frame over heldout data when using Log-mel feature.

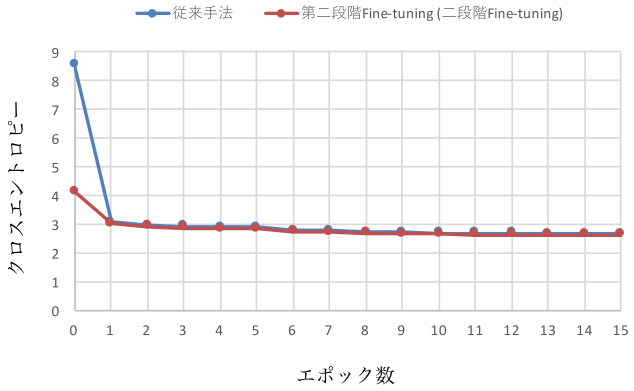

図 4 Log-mel+Δ 特徴量を利用した場合のヘルドアウトデータに対するフレームあたりのクロスエントロピー

Fig. 4 Cross-entropy per frame over heldout data when using Log-mel+Δ feature.

[*1] 表 2，表 3，表 4 のベースラインと太字の行に相当する．

4.1.4 結合強度に対する分析

提案手法を利用して学習された DNN が，期待どおりに中心近傍フレームを重視するように学習されているかを確認するために，従来手法と，提案する 2 段階 Fine-tuning で推定された DNN について，入力層と第 2 層の結合重みを調べた．

具体的には，入力層の位置 i（$i = -n, \cdots, n$. $i = 0$ が中心フレーム）のフレームから，第 2 層のすべてのユニットへの結合重みの絶対値（結合強度）の平均値 a_i を算出した．a_i は，位置 i のフレームの第 d 次元のユニットから第 2 層の u 番目のユニットへの結合重みを w_{du}^{i} と表すことで，以下の式 (1) で計算した．なお，D は特徴量の次元数を表し，Log-mel 特徴量を用いる場合には $D = 40$，Log-mel+Δ 特徴量を用いる場合には $D = 80$，Log-mel+Δ+ΔΔ 特徴量を用いる場合には $D = 120$ となる．そして第 2 層のユニット数 U_2 は 512 である．

$$a_i = \frac{\sum_{d=1,\cdots,D} \sum_{u=1,\cdots,U_2} |w_{du}^{i}|}{D \times U_2} . \tag{1}$$

表 2，表 3，表 4 の，各々のベースラインと太字で示した最も低い単語誤り率を示したモデルに対して分析を行い，

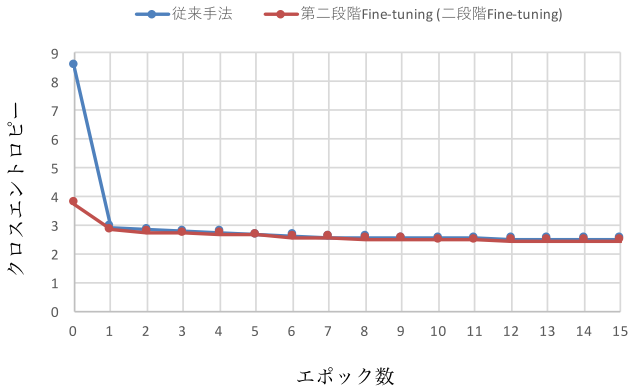

図 5 Log-mel+Δ+ΔΔ 特徴量を利用した場合のヘルドアウトデータに対するフレームあたりのクロスエントロピー

Fig. 5 Cross-entropy per frame over heldout data when using Log-mel+Δ+ΔΔ feature.

表 5 学習終了時のフレームあたりのクロスエントロピー
Table 5 Cross-entropy per frame after DNN training.

	特徴量		
	Log-mel	Log-mel+Δ	Log-mel+Δ+ΔΔ
従来手法	2.660	2.661	2.507
提案手法（第 2 段階 Fine-tuning）	2.633	2.615	2.448

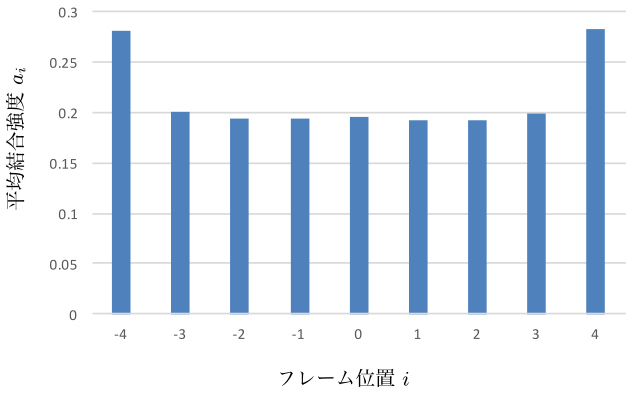

図 6 従来手法によって学習された DNN の入力層各フレームから第 2 層への平均結合強度（Log-mel 特徴量，$n = 4$）
Fig. 6 Averaged weight magnitudes between each frame in the input layer and the second layer after normal DNN training (Log-mel feature, $n = 4$).

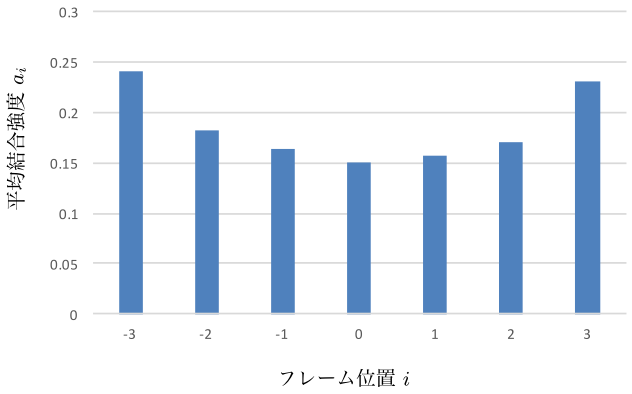

図 8 従来手法によって学習された DNN の入力層各フレームから第 2 層への平均結合強度（Log-mel+Δ 特徴量，$n = 3$）
Fig. 8 Averaged weight magnitudes between each frame in the input layer and the second layer after normal DNN training (Log-mel+Δ feature, $n = 3$).

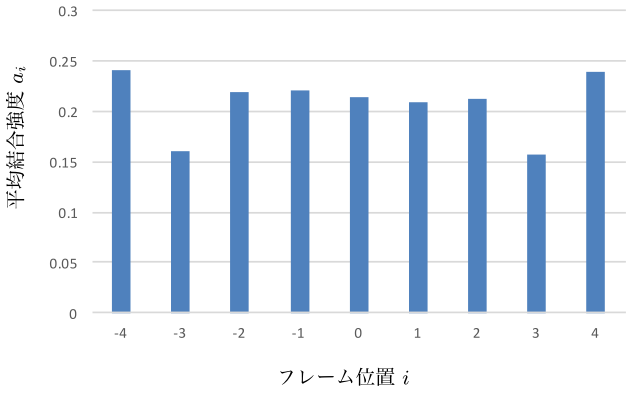

図 7 2 段階 Fine-tuning で学習された DNN の入力層各フレームから第 2 層への平均結合強度（Log-mel 特徴量，$m = 2$，$n = 4$）
Fig. 7 Averaged weight magnitudes between each frame in the input layer and the second layer after two-stage fine-tuning (Log-mel feature, $m = 2, n = 4$).

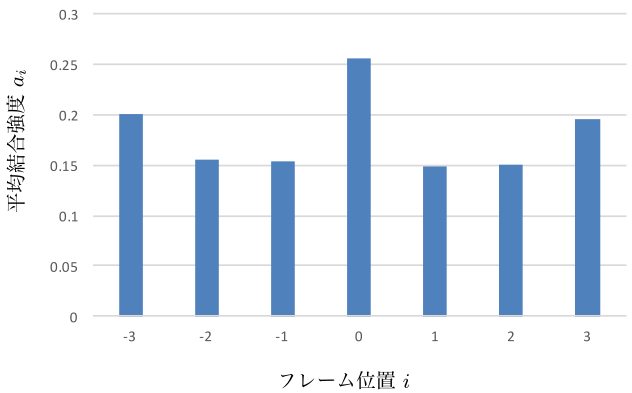

図 9 2 段階 Fine-tuning で学習された DNN の入力層各フレームから第 2 層への平均結合強度（Log-mel+Δ 特徴量，$m = 0$，$n = 3$）
Fig. 9 Averaged weight magnitudes between each frame in the input layer and the second layer after two-stage fine-tuning (Log-mel+Δ feature, $m = 0, n = 3$).

Log-mel 特徴量の結果を**図 6** と**図 7** に，Log-mel+Δ 特徴量の結果を**図 8** と**図 9** に，さらに Log-mel+Δ+ΔΔ 特徴量の結果を**図 10** と**図 11** に示した．

従来手法で学習したベースラインの DNN と比較すると，2 段階 Fine-tuning を利用した場合，Pre-training と第 1 段階 Fine-tuning で利用しているフレームの結合強度が大きくなり，遠隔フレームの結合強度が小さくなっていることが確認できた．たとえば，Log-mel 特徴量を用いて，従来手法で学習した DNN の結合強度を示した図 6 と，2 段階 Fine-tuning で最も良い認識率を得た DNN の結合強度を示した図 7 を比較すると，図 7 では強調している中心近傍 5 フレームの結合強度が大きくなっており，それ以外の遠隔フレームについては結合強度が抑えられている．つまり，提案手法により，従来手法と比較して，中心近傍フレームを遠隔フレームよりも重視する DNN の学習を実現することができた．

4.2 日本語実運用システムでの評価

次に，実運用されている日本語大語彙連続音声認識システムを利用して，提案する 2 段階 Fine-tuning の効果を検証した．このシステムでは，特徴量として，31 次元の

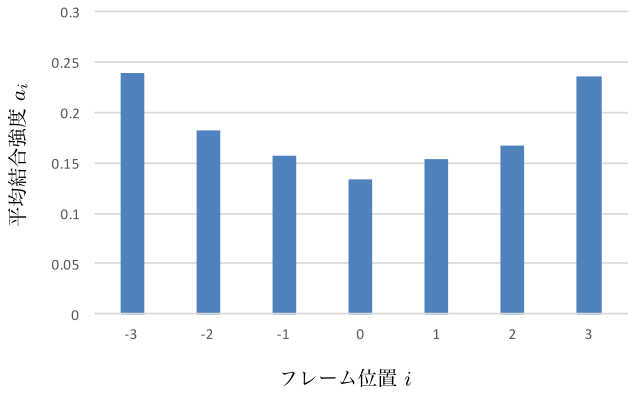

図 10 従来手法によって学習された DNN の入力層各フレームから第 2 層への平均結合強度（Log-mel+Δ+ΔΔ 特徴量, $n = 3$）

Fig. 10 Averaged weight magnitudes between each frame in the input layer and the second layer after normal DNN training (Log-mel+Δ+ΔΔ feature, $n = 3$).

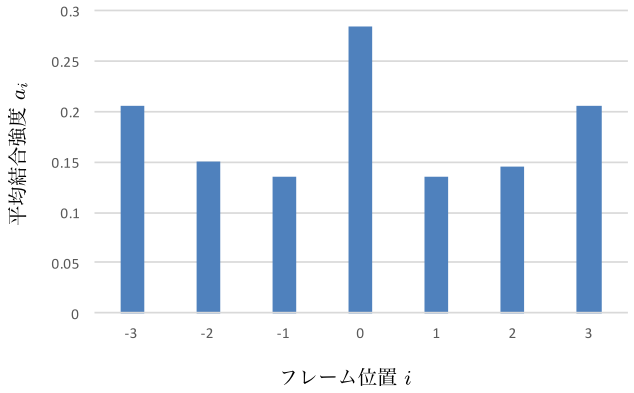

図 11 2 段階 Fine-tuning で学習された DNN の入力層各フレームから第 2 層への平均結合強度（Log-mel+Δ+ΔΔ 特徴量, $m = 0$, $n = 3$）

Fig. 11 Averaged weight magnitudes between each frame in the input layer and the second layer after two-stage fine-tuning (Log-mel+Δ+ΔΔ feature, $m = 0, n = 3$).

MFCC（Mel-Frequency Cepstral Coefficient）を 11 フレーム結合して利用している．また，DNN は GMM-HMM システムのための特徴量抽出器として利用されている[*2]．実運用システムでは，システム自身の構成変更は困難であるため，最終的にこの構成（MFCC 31 次元を 11 フレーム結合入力，DNN を特徴量抽出器として利用）を利用するということを前提として，提案手法によって音声認識率を改善することを検討した．

4.2.1 実験条件

DNN の構成は，すべての実験で共通とし，341（$= 11 \times 31$）個のユニットを持つ入力層，1,024 個のユニットを持つ 5 層の隠れ層，40 個のユニットを持つボトルネック層，そして 3,000 個のユニットを持つ出力層からなるようにした．出力層では，状態共有トライフォン HMM の状態を予測対象とした [24]．Pre-training と Fine-tuning ともに，フレームあたりのクロスエントロピーを目的関数として，誤差逆伝播で識別的に学習を行った [12]．DNN の学習中は，ヘルドアウトデータを別途用意し，それに対するフレームあたりのクロスエントロピーを確認して，学習率の調整を行った [22]．Fine-tuning のエポック数は，従来手法の Fine-tuning，提案手法の第 1 段階 Fine-tuning，第 2 段階 Fine-tuning ともに 15 で固定した．

DNN の学習が終了した後，推定した DNN を利用して，ボトルネック層から 40 次元のボトルネック特徴量（Bottleneck Feature: BNF）を抽出し [2], [3]，それに基づいて GMM の学習を行った．4.2.2 項の予備実験では，最尤（Maximum Likelihood: ML）法で GMM 学習を行い，4.2.3 項の大規模実験では，ML 法で GMM 学習を行った後，特徴量空間，およびモデル空間での識別学習を行った [21], [26], [27]．

評価データは，音声によってメールなどの文章入力を行った発話からなるタスク 1 と，音声検索の発話からなるタスク 2 の 2 種類からなり，各々のタスクは 10,000 以上の発話を含む．学習データと評価データとして，携帯電話を利用して，各々のタスクの実利用環境で収集された音声データを利用した．

言語モデルと認識用語彙は，タスクごとに用意した．各々のタスクの言語モデル学習用データのサイズは 10 億単語以上で，単語 4-gram とクラス 3-gram を線形補間して利用した [28]．また各々のタスクごとに，語彙サイズ約 100 万の認識用語彙を用意した．

音声認識精度の評価尺度として文字誤り率（Character Error Rate: CER）とカナ誤り率（Kana Error Rate: KER）を利用した．CER 算出時には，評価データに対する正解書き起こしと，音声認識結果を，それぞれ文字単位に分割し，文字ごとに比較を行った．KER 算出時には，評価データに対する正解書き起こしと，音声認識結果を，カナ表記に変換したうえで，カナ 1 文字ごとに比較を行った．日本語では単語分割に曖昧性があるため，日本語大語彙連続音声認識の評価尺度として，英語などの言語で一般的に利用される単語誤り率ではなく，CER および KER を利用した．なお，KER は CER よりも緩い評価基準となるが，日本語では無視することができない程度に出現する表記揺れに対して，頑健に対応することができる．

4.2.2 予備実験

予備実験として，50 時間の学習データを DNN と GMM 学習に利用して，2 段階 Fine-tuning を様々な条件で比較した．なお，従来手法として，11 フレームの音響特徴量を結合し，4.2.1 項で記述したように，識別的に Pre-training

[*2] このシステムの運用時，従来の音声認識デコーダからの変更を少なくするために，DNN-HMM によるハイブリッドシステムではなく，DNN による特徴量抽出を利用した．また，DNN の学習は，GMM の学習よりも計算量が大きいことを考慮すると，一定規模のデータで DNN による特徴量抽出器を学習し，さらにそれを大規模なデータに適用して GMM を学習することで，大規模な学習データを効率的に扱うことができる，という利点もある．

表 6 50 時間の学習データを利用した 2 段階 Fine-tuning の結果（文字誤り率（CER），従来手法からの文字誤り率削減率（CER Reduction: CERR），カナ誤り率（KER），従来手法からのカナ誤り率削減率（KER Reduction: KERR））

Table 6 Two-stage fine-tuning with 50 hours of training data. (Character Error Rate (CER), CER Reduction (CERR) from normal training, Kana Error Rate (KER) and KER Reduction (KERR) from normal training are shown).

DNN	入力	文字誤り率 (文字誤り率削減率) [%]			カナ誤り率 (カナ誤り率削減率) [%]		
学習方法	フレーム数	タスク 1	タスク 2	平均	タスク 1	タスク 2	平均
従来手法	11	20.45	31.67	26.06	12.25	15.21	13.73
2 段階 Fine-tuning	$1 \rightarrow 11$ $(m=0)$	20.44 (0.06)	32.22 (−1.71)	26.33 (−0.83)	12.25 (0.00)	15.35 (−0.93)	13.80 (−0.44)
	$3 \rightarrow 11$ $(m=1)$	20.17 (1.39)	31.12 (1.74)	25.64 (1.57)	12.11 (1.16)	14.95 (1.70)	13.53 (1.43)
	$5 \rightarrow 11$ $(m=2)$	20.12 (1.61)	31.01 (2.09)	**25.57 (1.85)**	12.09 (2.32)	14.71 (3.30)	**13.40 (2.29)**
	$7 \rightarrow 11$ $(m=3)$	19.96 (2.40)	31.44 (0.73)	25.70 (1.57)	11.97 (3.55)	14.97 (1.60)	13.47 (1.95)
	$9 \rightarrow 11$ $(m=4)$	20.21 (1.20)	31.42 (0.80)	25.81 (1.00)	12.15 (1.33)	14.92 (1.95)	13.53 (1.38)

表 7 大規模データによる 2 段階 Fine-tuning の結果（文字誤り率（CER），従来手法からの文字誤り率削減率（CER Reduction: CERR），カナ誤り率（KER），従来手法からのカナ誤り率削減率（KER Reduction: KERR））

Table 7 Two-stage fine-tuning with large amounts of training data. (Character Error Rate (CER), CER Reduction (CERR) from normal training, Kana Error Rate (KER) and KER Reduction (KERR) from normal training are shown).

DNN	文字誤り率 (文字誤り率削減率) [%]			カナ誤り率 (カナ誤り率削減率) [%]		
学習	タスク 1	タスク 2	平均	タスク 1	タスク 2	平均
従来手法	16.28	26.80	21.54	9.14	11.48	10.31
2 段階 Fine-tuning	16.12 (1.00)	26.42 (1.41)	**21.27 (1.20)**	9.05 (1.00)	11.20 (2.37)	**10.12 (1.68)**

と Fine-tuning を行って DNN を学習した．2 段階 Fine-tuning では，Pre-training と第 1 段階 Fine-tuning で利用するフレーム数を変化させて DNN の学習を行い，その後，11 フレームすべてを利用して第 2 段階 Fine-tuning を行った．図 1 で使っている n，m を利用して表現した場合，n を 5 で固定したまま，m を 0 から 4 に変化させて実験を行った．なお，基準となる DNN と，この予備実験で得られる DNN の層の数や入力，出力ユニット数などの構成は，すべて完全に同一となる．

DNN 学習が終了後，各々の DNN を利用して BNF を抽出し，GMM 学習を行い，評価データを利用して大語彙連続音声認識で評価を行った．**表 6** に結果を示した．$m=0$ の場合を除いたすべての場合において，2 段階 Fine-tuning を利用することで，認識率の改善を確認することができた．今回の実験では，$m=2$ の場合，つまり最初に 5 フレームを結合して DNN の学習を行い，その後 11 フレームすべてを利用した場合に，最も大きな改善が得られ，2 タスクの平均で，1.85%の CER の改善，2.29%の KER の改善を確認した．次の項では，この条件で大規模実験を行った結果を示す．

2 段階 Fine-tuning を行う場合，第 1 段階 Fine-tuning と第 2 段階 Fine-tuning でエポック数を 15 と固定したが，ヘルドアウトデータに対するクロスエントロピーは十分に飽和していた．また，従来手法で DNN を学習する際にも，Fine-tuning のエポック数は 15 に固定していたが，同様にクロスエントロピーは十分に飽和していた．つまり，2 段階 Fine-tuning で得られた改善は，第 1 段階 Fine-tuning と第 2 段階 Fine-tuning での合計学習エポック数が増えたことに起因しているわけではない．

4.2.3 大規模実験

次に，1,000 時間以上の学習データを DNN，GMM の学習に利用する大規模実験を行った．従来手法としては，予備実験と同様に，11 フレームの音響特徴量を結合して入力として利用し，識別的に Pre-training と Fine-tuning を行って，DNN の学習を行った．提案手法として，予備実験で最も良い認識率を得ることができた条件設定を利用した．具体的には，最初に中心近傍 5 フレームで Pre-training と Fine-tuning を行い，続いて 11 フレームすべてを利用して Fine-tuning を行う，2 段階 Fine-tuning を行った．

推定された DNN を利用して，BNF を抽出し，GMM 学習を行ったうえで，評価データを利用して大語彙連続音声認識で評価を行った．**表 7** に結果を示した．提案手法を利用することで，2 タスクの平均で，1.20%の CER の改善，1.68%の KER の改善を確認することができ，提案手法の有効性を示すことができた．また，これらの改善は，有意水準 5%で統計的に有意であることを確認した．

5. おわりに

本論文では，DNN 学習時に中心近傍フレームを重視する方法を提案した．最初に中心近傍フレームのみを利用

して Pre-training と Fine-tuning を行った後，すべてのフレームを利用して Fine-tuning を行う 2 段階 Fine-tuning を提案し，大語彙連続音声認識実験を通じて，音声認識精度を改善することができた．さらに，提案した 2 段階 Fine-tuning によって学習された DNN では，従来手法で学習された DNN と比較して，中心近傍フレームからの重みが大きくなり，遠隔フレームからの重みが抑制されていることを確認した．

今回の実験では，Fine-tuning を，入力フレーム数を変更して 2 度行った．しかし，多段階で Fine-tuning を行うことも可能であり，たとえば，1，3，5，7 フレームと徐々に利用するフレームを増やして Fine-tuning を行うことも，今後検討していきたい．

参考文献

[1] Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. and Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, Vol.29, No.6, pp.82–97 (2012).

[2] Sainath, T.N., Kingsbury, B. and Ramabhadran, B.: Auto-encoder bottleneck features using deep belief networks, *Proc. ICASSP*, pp.4153–4156 (2012).

[3] Gehring, J., Miao, Y., Metze, F. and Waibel, A.: Extracting deep bottleneck features using stacked auto-encoders, *Proc. ICASSP*, pp.3377–3381 (2013).

[4] Dahl, G.E., Yu, D., Deng, L. and Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.20, No.1, pp.30–42 (2012).

[5] Mohamed, A.-R., Dahl, G.E. and Hinton, G.: Acoustic modeling using deep belief networks, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.20, No.1, pp.14–22 (2012).

[6] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H. and Zweig, G.: fMPE: Discriminatively trained features for speech recognition, *Proc. ICASSP*, pp.961–964 (2005).

[7] Arisoy, E., Sainath, T.N., Kingsbury, B. and Ramabhadran, B.: Deep neural network language models, *Proc. NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pp.20–28 (2012).

[8] Kuo, H.-K., Arisoy, E., Emami, A. and Vozila, P.: Large scale hierarchical neural network language models, *Proc. Interspeech* (2012).

[9] Hinton, G., Osindero, S. and Teh, Y.-W.: A fast learning algorithm for deep belief nets, *Neural Computation*, Vol.18, No.7, pp.1527–1554 (2006).

[10] 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).

[11] Soltau, H., Kuo, H.-K., Mangu, L., Saon, G. and Beran, T.: Neural network acoustic models for the DARPA RATS program, *Proc. Interspeech*, pp.3092–3096 (2013).

[12] Sainath, T.N., Kingsbury, B. and Ramabhadran, B.: Improving training time of deep belief networks through hybrid pre-training and larger batch sizes, *Proc. NIPS Workshop on Log-linear Models* (2012).

[13] Seide, F., Li, G., Chen, X. and Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription, *Proc. ASRU*, pp.24–29 (2011).

[14] Larochelle, H., Bengio, Y., Louradour, J. and Lamblin, P.: Exploring strategies for training deep neural networks, *The Journal of Machine Learning Research*, Vol.10, pp.1–40 (2009).

[15] Rumelhart, D.E., Hinton, G.E. and Williams, R.J.: Learning representations by back-propagating errors, *Cognitive Modeling* (1988).

[16] Mohamed, A.-R., Yu, D. and Deng, L.: Investigation of full-sequence training of deep belief networks for speech recognition, *Proc. Interspeech*, pp.2846–2849 (2010).

[17] 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).

[18] Veselỳ, K., Ghoshal, A., Burget, L. and Povey, D.: Sequence-discriminative training of deep neural networks, *Proc. Interspeech*, pp.2345–2349 (2013).

[19] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proc. AISTATS*, pp.249–256 (2010).

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

[21] Soltau, H., Saon, G. and Kingsbury, B.: The IBM Attila speech recognition toolkit, pp.97–102 (2010).

[22] Sainath, T.N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P. and Mohamed, A.-R.: Making deep belief networks effective for large vocabulary continuous speech recognition, *Proc. ASRU*, pp.30–35 (2011).

[23] Sainath, T.N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.-R., Dahl, G. and Ramabhadran, B.: Deep convolutional neural networks for large-scale speech tasks, *Neural Networks*, Vol.64, pp.39–48 (2015).

[24] Yu, D. and Seltzer, M.L.: Improved bottleneck features using pretrained deep neural networks, *Proc. Interspeech*, pp.237–240 (2011).

[25] Furui, S.: Cepstral analysis technique for automatic speaker verification, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.29, No.2, pp.254–272 (1981).

[26] Chen, S.F., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Soltau, H. and Zweig, G.: Advances in speech transcription at IBM under the DARPA EARS program, *IEEE Trans. Audio, Speech and Language Processing*, Vol.14, pp.1596–1608 (2006).

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

[28] Chen, S.F. and Goodman, J.: An empirical study of smoothing techniques for language modeling, *Computer Speech & Language*, Vol.13, No.4, pp.359–393 (1999).

倉田 岳人 （正会員）

2002年東京大学工学部電子工学科卒業．2004年同大学大学院情報理工学系研究科電子情報学専攻修士課程修了．2013年同大学院情報理工学系研究科電子情報学専攻博士号取得．2004年日本アイ・ビー・エム株式会社入社．以来，同社東京基礎研究所にて，音声認識等の音声言語情報処理の研究に従事．同社リサーチ・スタッフ・メンバー，スピーチテクノロジー担当マネージャー．博士（情報理工学）．日本音響学会会員．

ヴィレット ダニエル

1995年ダルムシュタット工科大学計算機科学部卒業．2000年デュースブルク大学電子工学科博士課程修了．日本電信電話株式会社にて2年間の勤務ののち，ハーマンベッカー（ドイツ）にて車載組み込み音声認識の研究を担当．2006年ニュアンス（ドイツ）入社．以来，クラウド音声認識システムの音響モデル研究チームを担当．博士．