情報処理学会論文誌 Vol.57 No.3 1080–1088 (Mar. 2016)

低ランクDNN音響モデルの騒音下音声認識での評価と系列の識別学習

太刀岡 勇気[1,a)] 渡部 晋治[2,b)] ルルー ジョナトン[2,c)] ハーシー ジョン[2,d)]

受付日 2015年5月25日, 採録日 2015年12月7日

概要：深層神経回路網（DNN）音響モデルは従来のガウス混合モデル（GMM）を上回る性能を達成しているが，パラメータ数がGMMより多くなる傾向にある．これにより，計算コストがGMMよりも増大する．DNNのモデルサイズを縮減するために，特異値分解（SVD）を用いた重み行列の低ランク近似が提案されている．従来の検討はクリーン音声のみであるが，騒音下音声はより複雑であり，モデル化が難しくなる可能性がある．よってこのSVD手法の有効性を騒音残響下音声認識タスクで検証する．加えて，低ランク近似と系列の識別学習を併用する．系列の識別学習はフレームごとの識別的基準により構築されたDNNの性能を向上させることが知られている．また低ランク近似と系列の識別学習の適用順の影響を調査した．実験により，低ランク近似は騒音下音声認識に有効であり，低ランク近似を先にモデルに適用し，その後に低ランクモデルに対して識別学習を行うと最も効果的であることが分かった．この識別学習した低ランクモデルは，モデル縮減せずに識別学習したモデルの性能を上回った．

キーワード：音声認識・理解，深層神経回路網，特異値分解，識別学習

Evaluation of Noisy Speech Recognition and Sequence Discriminative Training for Low-rank Deep Neural Network Acoustic Models

Yuuki Tachioka[1,a)] Shinji Watanabe[2,b)] Jonathan Le Roux[2,c)] John R. Hershey[2,d)]

Received: May 25, 2015, Accepted: December 7, 2015

Abstract: Deep neural network (DNN) acoustic models outperform conventional Gaussian mixture model (GMM) but the number of parameters tends to be larger. This leads to higher computational costs than those of GMM. To reduce DNN model size, by using singular value decomposition (SVD) have previously been applied for low-rank approximations of weight matrices. Previous studies only focused on clean speech, whereas because noisy speech is more complicated and its modeling could be difficult. Thus we investigate the effectiveness of this SVD method on noisy reverberated speech recognition task. Furthermore, we combine the low-rank approximation with sequence discriminative training, which further improved the performance of the DNN, which was constructed using a frame-by-frame discriminative criterion. We also investigated the effect of the order of application of the low-rank and sequence discriminative training. Our experiments show that low rank approximation is effective for noisy speech recognition and the most effective combination of discriminative training with model reduction is to apply the low rank approximation to the base model first and then to perform discriminative training on the low-rank model. This discriminatively low-rank trained model outperformed the low-rank discriminatively trained low-rank model.

Keywords: Automatic speech recognition and understanding, Deep neural networks, Singular value decomposition, Discriminative training

[1] 三菱電機株式会社情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Kanagawa, 247–8501, Japan
[2] Mitsubishi Electric Research Laboratories, Cambridge, MA, 02139–1955, US

[a)] Tachioka.Yuki@eb.MitsubishiElectric.co.jp
[b)] watanabe@merl.com
[c)] leroux@merl.com
[d)] hershey@merl.com

1. はじめに

深層神経回路網（Deep neural networks; DNN）による音響モデルは，音声認識の分野で成功を収めた [1]．DNN 音響モデルは，従来のガウス混合分布モデル（Gaussian mixture model; GMM）に基づく音響モデルの性能を多くの場合上回るが [1], [2]，DNN モデルのパラメータ数は GMM のそれよりも多くなる傾向にある．大語彙連続音声認識タスクを例とすると [2]，GMM に基づく音声認識システムの隠れマルコフモデル（hidden Markov model; HMM）の状態数は 3 k 程度，状態あたりのガウス分布の混合数は 32 程度で，合計のパラメータ数は 10 M より少ない程度となる．一方で，DNN に基づく音声認識システムでは，HMM 状態数は同じであっても，一般的に用いられる設定として，各隠れ層のノード数を 2 k とし，隠れ層の数を 7 層とすると，パラメータ数は 30 M を超える．さらに HMM の状態数を，GMM の場合より多くすることもある．この場合，DNN モデルには GMM の 3 倍の数のパラメータがあるので，これにより必要なメモリ量・計算量が増加する．

この問題に対処するため，DNN のモデルサイズを縮減する手法がいくつか提案されている [3], [4]．ここでは，通常の方法で学習した全パラメータを持つモデルを「完全モデル」，パラメータを縮減したモデルを「低ランクモデル」と呼ぶことにする．ヒューリスティックな方法としては，ある一定の閾値より小さい重みに関しては 0 とする方法が考えられる．DNN の重み行列は密結合で冗長であるので，この方法によって重みパラメータ数を減らすことはできる．ただしこの方法では，結合されているか否かの情報を表すベクトルが新たに必要となるほか，行列積に特化したライブラリを使う場合には積和演算が並列化されるため，この方法では高速化されない可能性がある．Xue らは，特異値分解（singular value decomposition; SVD）を DNN モデルに適用し，総パラメータ数を削減する方法を提案した．彼らの方法では，SVD により重み行列のランクを低減させた低ランクモデルを初期値として，ファインチューニングを併用することで，モデルサイズを小さくしながらも，認識性能を維持できることを実験的に示した [4]．実験には，彼ら独自のデータベースを用いているため，音声データの収録仕様は不明であるが，大語彙連続音声認識のタスクに使う音声データは接話マイクを使ったデータであることが多いため，比較的クリーンなデータであることが想定される．そのような状況では，音素を弁別すればよいため，比較的単純なモデルで済み，低ランク近似が有効であると考えられる．これに対して，マイクと発話者の距離が離れた騒音・残響環境下においては，騒音の多様性に対処するため，これをモデル化する DNN モデルはより複雑になる傾向にある．このような使用状況では，低ランク近似が有効に働かず，モデルサイズの縮減は性能に悪影響を与えることもありうる．ゆえに，この手法が，騒音・残響環境下音声認識タスクにおいて有効であるかは自明ではなく，別に検証される必要がある．

さらに，従来のモデル縮減の検討は，クロスエントロピー（cross-entropy; CE）学習によるもので，フレームレーベルでの識別基準に基づいたものに限られている．音声認識はフレームごとの音素ラベルを当てることを最終的な目的としているわけではなく，文単位で正解単語列を出力することが期待される系列の識別学習問題なので，文単位での識別基準に基づいて音響モデルを改善する必要がある．GMM では，たとえば相互情報量最大化（maximum mutual information; MMI）基準に基づいて音響モデルの系列における識別学習を行うことで，最尤モデルの性能を向上させることができる [5], [6], [7]．DNN の場合も同様である [8], [9], [10], [11], [12], [13], [14]．上述のモデルサイズ低減手法と系列の識別学習を併用する際には，モデル縮減と系列の識別学習の適用順序より性能差が生じることが考えられるため，効果的な適用順序を調査する必要がある．直観的には，低ランク近似による性能低下を回復するためには，モデル縮減を行った後に識別学習を行うことが重要であると考えられる．本報では 3 つの組合せを検討する．第 1 は，SVD に基づく低ランク近似とファインチューニングを CE 完全モデルに対して行い，その後，CE 低ランクモデルに対して MMI 識別学習[*1]を行う方法である．第 2 は，低ランク近似とファインチューニングを MMI 完全モデルに対して行う方法である．第 3 は，第 2 の方法で得られた MMI 低ランクモデルに再び識別学習を行う方法である．本報ではこれらの 3 種の組合せに関して，騒音残響環境下音声認識タスクにより，SVD と系列の識別学習の組合せを実験的に検討することとする．

2. DNN-HMM ハイブリッド音声認識システム

2.1 DNN 音響モデル

DNN-HMM ハイブリッド音声認識システムは，さまざまな条件で従来の GMM-HMM システムを上回る性能を発揮している．これは従来の GMM による尤度計算を DNN による疑似尤度計算に置き換えるもので，尤度計算部のみの変更で済むため，既存の音声認識システムとの親和性が高い．本節では，DNN 音響モデルの概要について述べる．ここで，DNN 音響モデルのパラメータ θ は，L 層の隠れ

[*1] DNN の識別学習には，系列ベイズリスク最小化（sequence Minimum Bayes Risk; sMBR）もよく使われるが，同様に MMI もよく使われる．実際 Kaldi ツールキット [15] でも nnet1 は sMBR が標準，nnet2 では MMI が標準になっている（2014 年 10 月現在）．ここでの実験では nnet2 を使ったので，MMI 識別学習を採用した．方式の差異による性能の差は出ると思われるが，sMBR でも大まかな傾向は同じであると考えられる．

層からなることとし，0 番目の層が入力層であり，$(L+1)$ 番目の層が出力層であるとする．DNN 音響モデルの l 番目の層（$0 \le l \le L+1$）に入力される n 次元の入力特徴量を，$\boldsymbol{x}^l$ で表す．出力特徴量は m 次元であり，これは同時に $(l+1)$ 番目の層の入力特徴量となる．よってこれは，$\boldsymbol{x}^{l+1}$ と表される．NN では一般的に，非線形操作 f が線形操作に加えて用いられる．隠れ層には，シグモイド関数が f として使われる一方で，最終層にはソフトマックス関数が使われる．重み行列 $\boldsymbol{A}^l_{m\times n}$ とオフセット項 $\boldsymbol{b}^l$ のパラメータは，誤差逆伝搬によるファインチューニング（fine tuning; FT）により調整される．ここで，行列における下付きの記号は，その行列の次元を表すこととする．学習の際に確率的勾配法を使い，低位層から高位層に向かって，特徴量 $\boldsymbol{x}$ を以下のように伝搬させる．

$$\boldsymbol{x}^{l+1} = f\left(\boldsymbol{A}^l_{m\times n}\boldsymbol{x}^l + \boldsymbol{b}^l\right) \tag{1}$$

本報では，制約付きボルツマンマシンを用いた初期化の代わりに，DNN を 1 層ずつ積み重ねる識別的プレトレーニングにより，DNN を構築した．

フレーム t，HMM 状態 j に対応する事後確率を算出するために，DNN を用いる．DNN-HMM ハイブリット音声認識システムでは，疑似音響尤度 p を以下のように求める．

$$p\left(\boldsymbol{x}^0_t|j\right) \propto \frac{p\left(j|\boldsymbol{x}^0_t\right)}{p_0\left(j\right)} \tag{2}$$

ここで，$p_0\left(j\right)$ は学習データの個数から算出される事前分布の確率である．DNN の入力特徴量 $\boldsymbol{x}^0_t$ は，連続 $(2s+1)$ フレームの特徴量をつなげた $[\boldsymbol{x}_{t-s},\ldots,\boldsymbol{x}_t,\ldots,\boldsymbol{x}_{t+s}]$ である．DNN の出力は，各コンテキスト依存 HMM 状態に対応する出力確率である．出力層では，ソフトマックス活性化関数が使われる．

$$p(j|\boldsymbol{x}^0_t) = \frac{\exp a(j|\boldsymbol{x}^0_t)}{\sum_{j'}\exp a(j'|\boldsymbol{x}^0_t)} \tag{3}$$

ここで，a は出力層ノード j での活性化前の値であり，DNN への入力 $\boldsymbol{x}^0_t$ の関数となっている．

2.2 低ランク近似による DNN モデルサイズの縮減

前述のとおり，DNN-HMM システムは多くの場合，従来の GMM-HMM システムを上回る性能を発揮するが，DNN のパラメータ数は GMM のそれよりも多くなりやすいという欠点がある．それゆえ，文献 [4] はパラメータの総数を削減するために，SVD を用いてある層 l における重み行列 $\boldsymbol{A}^l_{m\times n}$ のランクを縮退させる方法を提案した．特異値分解（式 (4)）により，行列 $\boldsymbol{A}^l_{m\times n}$ は

$$\boldsymbol{A}^l_{m\times n} = \boldsymbol{U}_{m\times n}\boldsymbol{\Sigma}_{n\times n}\boldsymbol{V}^\top_{n\times n} \tag{4}$$

のように 3 つの行列の積に分解される．ここで $\boldsymbol{\Sigma}$ は対角行列であり，その要素は特異値である．特異値 σ は降順に並べ替えられているとする（$\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n$）．

$$\begin{aligned}\boldsymbol{\Sigma}_{n\times n} &= \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{k\times k} & & & \\ & \sigma_{k+1} & 0 & \\ & & \ddots & \\ & 0 & & \sigma_n \end{pmatrix}\end{aligned} \tag{5}$$

行列 $\boldsymbol{U}$ と $\boldsymbol{V}$ は，直交正規化された列ベクトルを持つ．$\top$ は転置を表す．行列 $\boldsymbol{A}_{m\times n}$ のパラメータ数を減らすため，k 番目までに大きな特異値とそれに対応する $\boldsymbol{U}$ と $\boldsymbol{V}$ の特異値ベクトルが低ランク近似に使われる．その際に，部分行列 $\boldsymbol{U}_{k\times k}$，$\boldsymbol{\Sigma}_{k\times k}$，$\boldsymbol{V}^\top_{k\times n}$ により，$\boldsymbol{A}_{m\times n}$ は以下のように低ランク近似できる．

$$\begin{aligned}\boldsymbol{A}^l_{m\times n} &\approx \boldsymbol{U}_{m\times k}\boldsymbol{\Sigma}_{k\times k}\boldsymbol{V}^\top_{k\times n} \quad (k<n) \\ &= \left[\boldsymbol{U}_{m\times k}\sqrt{\boldsymbol{\Sigma}_{k\times k}}\right]\left[\sqrt{\boldsymbol{\Sigma}_{k\times k}}\boldsymbol{V}^\top_{k\times n}\right] \\ &= \boldsymbol{A}^{l+\frac{1}{2}}_{m\times k}\boldsymbol{A}^l_{k\times n}\end{aligned} \tag{6}$$

重み行列と入力特徴量の行列積の計算コスト $\boldsymbol{Ax}$ は，$O(mn)$ に比例する．低ランク近似の後，これは $O\left((m+n)k\right)$ に比例し，k が以下に示す条件を満たすときに，計算コストが元のモデルに比べて小さくなる．

$$k < \begin{cases} \frac{m}{2},\frac{n}{2} & (m \simeq n) \\ \min(m,n) & (m \gg n,\ m \ll n) \end{cases} \tag{7}$$

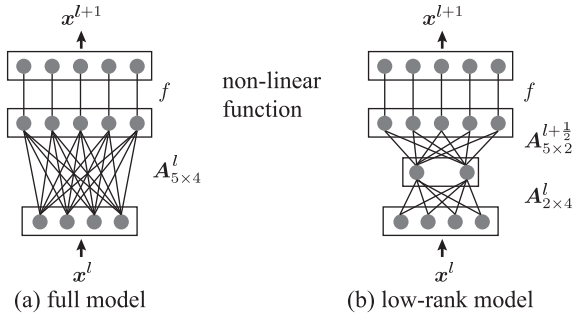
低ランク近似は l 番目の層を，重み行列 $\boldsymbol{A}^l_{k\times n}$ を持つ第 1 の線形変換層と，重み行列 $\boldsymbol{A}^{l+\frac{1}{2}}_{m\times k}$ を持つ第 2 のシグモイド層の 2 層に分解していると考えることもできる[*2]．**図 1** に示すように第 1 の層と第 2 の層の間には，シグモイドユニットのような非線形ユニットは存在しない．オフセット項を追加すると[*3]，新しい層は

$$\begin{aligned}\boldsymbol{x}^{l+\frac{1}{2}} &= \boldsymbol{A}^l_{k\times n}\boldsymbol{x}^l + \boldsymbol{b}^l \\ \boldsymbol{x}^{l+1} &= f\left(\boldsymbol{A}^{l+\frac{1}{2}}_{m\times k}\boldsymbol{x}^{l+\frac{1}{2}} + \boldsymbol{b}^{l+\frac{1}{2}}\right)\end{aligned} \tag{8}$$

のようになる．ここで $\boldsymbol{b}^l$ は k 次元のベクトルであり，初期値は 0 である．$\boldsymbol{b}^{l+\frac{1}{2}}$ は元のモデルの $\boldsymbol{b}^l$ である．

[*2] このような構成の DNN をランダム初期化すると，層数が多いため局所最適に陥る可能性が高くなる．

[*3] 通常 DNN の学習ツールは重み行列とオフセットを拡大係数行列の形で 1 つの行列として扱うので，オフセット $\boldsymbol{b}^l$ を付け加えた方がツールを使う際にコードの変更が必要ない．モデル削減後にファインチューニングをしない場合はこの $\boldsymbol{b}^l$ を 0 に初期化しておけば元のモデルと等価となり，ファインチューニングをする場合は，オフセット項が必要なければ 0 に近い値をとるはずである．さらにオフセット項のパラメータ数は重み行列に比べれば無視できるほど小さいので，モデル削減という観点からも $\boldsymbol{b}^l$ を導入することは問題ないと考えられる．



(a) full model (b) low-rank model

図 1 低ランク分解による DNN モデルパラメータの削減．例では (a) $5\times4=20$ から (b) $5\times2+2\times4=18$ にパラメータ数が減少している

Fig. 1 DNN model parameter via low-rank factorization. Example shows that model parameters are reduced from (a) $5\times4=20$ to (b) $5\times2+2\times4=18$.

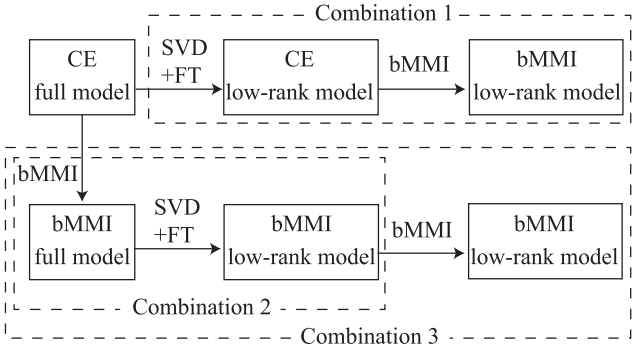


図 2 識別学習とファインチューニング（FT）と組合せて，MMI 低ランクモデルを生成する 3 種の組合せ法

Fig. 2 Three types of combinations to generate MMI low-rank model by combining discriminative training with fine tuning (FT).

2.3 DNN の CE 学習

CE 基準では，評価関数は

$$\mathcal{F}_{\mathrm{CE}}(\theta)=\sum_r\sum_t\sum_j \hat{p}(j,t)\log\frac{\hat{p}(j,t)}{p(j|\boldsymbol{x}_t^0)} \tag{9}$$

のようになる．ここで $\hat{p}(j,t)$ は，時刻 t，クラスラベル j に対応する正解の分布である．これをアクティベーション a で微分すると，時刻 t での勾配は

$$\left.\frac{\partial\mathcal{F}_{\mathrm{CE}}}{\partial a(j)}\right|_t=p(j|\boldsymbol{x}_t^0)-\hat{p}(j,t) \tag{10}$$

のようになり，誤差逆伝搬として知られる連鎖規則に基づく勾配法により，DNN のモデルパラメータ θ が最適化される．

2.4 DNN のための系列 MMI 学習

DNN は識別モデルであり，上述のとおり，そもそもフレームごとに識別的な基準（すなわち CE 基準）に基づき学習されている．音声認識のタスクは，系列での識別問題を扱うため，フレームごとに識別的な基準で学習しただけでは不十分である．そこで，CE モデルに対して，系列の識別学習を行う方法が提案されている．実際，DNN に対する系列の識別学習により，CE モデルから性能が向上することが広く知られている [10], [11], [12], [13], [14]．系列の識別学習では，DNN の活性化関数は，文全体での誤差を最小化する基準に従って識別的に学習される．たとえば，MMI 基準を用いると，DNN 音響モデルのパラメータ θ は，以下のようにして最適化される．

$$\mathcal{F}_{\mathrm{MMI}}(\theta)=\sum_r\log\frac{p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_s\right)^\kappa p_L(s)} \tag{11}$$

ここで，$\boldsymbol{x}_{1:T_r}$ は r 番目の発話の音響特徴量系列で，特徴量の長さは T_r である．$\mathcal{H}_{s_r}$ は正解ラベル s_r に対する HMM の状態系列，$\mathcal{H}_s$ は認識仮説 s に対する HMM の状態系列である．κ は音響スケールであり，p_L は言語モデル尤度である．

MMI の拡張である boosted MMI [7] では，評価関数は式 (11) のようになる．

$$\mathcal{F}_{\mathrm{bMMI}}(\theta)=\sum_r\log\frac{p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_s\right)^\kappa p_L(s)e^{-bA(s,s_r)}} \tag{12}$$

ここで A は s の s_r に対する音素正解率である．b は増幅係数であり，これが大きいほど正解率に応じてマージンを広くとることを意味している．最終層のソフトマックス活性化関数 a に関する時刻 t での勾配を求めると，以下のようになる [8], [9], [10], [13].

$$\left.\frac{\partial\mathcal{F}_{\mathrm{bMMI}}(\theta)}{\partial a(j)}\right|_t=\kappa\left(\gamma_{j,t}^{num}-\gamma_{j,t}^{den}\right) \tag{13}$$

ここで，$\gamma_{j,t}^{num}$ と $\gamma_{j,t}^{den}$ は式 (11) もしくは式 (12) の分子もしくは分母の事後確率である．すべての DNN のパラメータは式 (13) から誤差逆伝搬の手順に基づいて導出される．

2.5 系列の識別学習と SVD の併用

識別学習により，CE モデルに対して性能を向上させることができる．ただし，モデルサイズ縮減と識別学習の適用順は重要であり，自明なものではなく，そもそも低ランクモデルにおいて識別学習が有効であるかも明らかでない．**図 2** には，識別的に学習された低ランクモデルを構築するための手順を示す[*4]．この場合，識別学習とモデルサイズ縮減の前後によって以下の 3 つに手順が考えられる．この手順に沿って，以下で実験を行っている．すべての組合せに共通で，初期モデルは CE 学習された完全モデル（CE 完全モデル）である．第 1 の組合せ（Combination 1）は，SVD と FT を CE 完全モデルに適用し，その後 CE 低ランクモデルに対して識別学習を行う方法．第 2 の組合せ

[*4] ここでは MMI 基準で識別学習を行うが，他の基準であっても，手順は同じである．

（Combination 2）は，MMI 完全モデルに SVD と FT を適用する方法．第 3 の組合せ（Combination 3）は，第 2 の組合せによって得られた MMI 低ランクモデルに対して，さらに識別学習を行う方法である．

3. 騒音下音声認識実験

3.1 実験の設定

モデル縮減による効果を検討するために，第 2 回 CHiME チャレンジトラック 2 で性能を評価した．これは，中程度語彙の音声認識タスク（*Wall Street Journal* (WSJ0)）で，残響・非定常騒音環境下での音声認識性能を単語誤り率（word error rate; WER）の観点から評価するために設計されたタスクである [16]．言語モデルのサイズは 5 k（basic）である．開発セット（si_dt_05）は 10 話者の 409 発話であり，評価セット（si_et_05）は 12 話者の 330 発話（Nov'92）である．音響モデルは学習セットから学習した．学習セットは 83 話者の 7,138 発話よりなる．音響スケール κ は，開発セット（si_dt_05）により調整した．これらの音声データは，実際に起こりうる環境を模擬している．騒音は他話者による発話や家庭内の騒音，音楽といった非定常性のもので，孤立（'isolated'）発話に対して SN 比（signal-to-noise ratio; SNR）$\{-6, -3, 0, 3, 6, 9\}$ dB で騒音を重畳したものである．データベースは 2 チャネルの音声データであるが，事前分布に基づくバイナリマスク [17] により騒音抑圧された単一チャネルの音声データを用いた[*5]．

音響特徴量と特徴量変換の設定は，以下のとおりである [18]．DNN の学習には，Kaldi ツールキット [15] の Povey の実装を用いている．ベースラインの特徴量は，0 次〜12 次元の MFCC とその動的特徴量（Δ および $\Delta\Delta$）である．特徴量変換手法（線形判別分析（linear discriminant analysis; LDA）[19] と最尤線形変換（maximum likelihood linear transformation; MLLT））[20] および適応手法（話者適応化学習（speaker adaptive training; SAT）[21] および特徴量空間最尤線形回帰（feature-space maximum likelihood linear regression; fMLLR））[22] を用いて，40 次元からなる話者適応した特徴量により認識を行った．DNN の入力特徴量はこれらの各フレーム 40 次元の特徴量を連続 9 フレーム分連結した 360 次元の特徴量とした．

音響モデルの学習手順と特徴量変換の設定は文献 [17], [18] に詳しい記載がある．コンテキスト依存 HMM の状態数は 1,989 であり，これが最後のソフトマックス層の出力ノード数となる．DNN モデルでは HMM の状態数を GMM よりも多くすることが多いが，ここではそれの代わりに，最終層の出力に重み行列を掛け 8,000 ノードに拡張し，それを重みづけてソフトクラスタリングして本来の状態数に削減する層を追加している．これはもともと，GMM の混合分布の重みづけに想を得たもので，Kaldi における Povey のレシピに実装されている処理である．隠れ層数は 3 である．単層の神経回路網から始めて，2 回の繰返しごとに層を 1 層ずつ追加していくことで，多層神経回路網を構築した．1 回の繰返しには，400,000 サンプルを用いた．

[*5] 騒音抑圧により，3–5 dB 程度の SNR の改善効果が得られる．

表 1 SVD {1,2,3} に対応する DNN の構造．下線を引いたノードは低ランク近似により追加されたノードである

Table 1 DNN structure corresponding to SVD {1,2,3}. Underlined nodes are added nodes by low-rank approximation.

	lower layer ⟶ higher layer
CE-full (2.85 M)	$360\times331+331^2\times2+331\times8000$
SVD1 (1.47 M)	$360\times\underline{100}+\underline{100}\times331+(331\times\underline{96})\times2\times2$ $+331\times\underline{162}+\underline{162}\times8000$
SVD2 (1.52 M)	$360\times331+(331\times\underline{96})\times2\times2$ $+331\times\underline{162}+\underline{162}\times8000$
SVD3 (1.59 M)	$360\times331+331^2\times2+331\times\underline{160}+\underline{160}\times8000$
SVD3' (1.91 M)	$360\times331+331^2\times2+331\times\underline{200}+\underline{200}\times8000$
(CE-full (1.54 M)	$360\times184+184^2\times2+184\times8000$)

この学習データに対して最適になるように CE 完全モデル（CE-full）の構造を決め，SVD はおおむねパラメータ数を半分に削減するように設定した．参考までにパラメータを半分にして CE 完全モデルを学習した場合も実験を行った．パラメータの総数は**表 1** にまとめたとおりである．完全モデルに SVD を適用する手順は，以下の 3 通り検討した．第 1 には，SVD をすべての隠れ層に適用した（SVD 1）．第 2 には，最下層は特徴量抽出という重要な役割を担っているため低ランク近似せず，最下層を除くすべての層に SVD を適用した（SVD 2）．第 3 には，最もパラメータ数の多い最終層にのみ SVD を適用した（SVD 3）．SVD 3 に関してはパラメータ数の影響を調査するために，パラメータ削減のレベルを 2 段階（1.59 M の SVD3 と 1.91 M の SVD3'）用意した．

CE 学習では，初めの 15 エポックの間は学習率を低減させながら学習を行い，最後の 5 エポックに関しては，学習率を固定して学習を行った．CE の完全モデルに対する学習率は，初期値が 0.01 であり，これを学習の終盤には 0.001 まで低減させた．ミニバッチサイズは 128 である．SVD を CE 完全モデルもしくは MMI 完全モデルに適用後，ファインチューニングを行った．初めの 3 エポックは学習率を 0.001 から 0.0005 に低減させながら学習を行い，続けて 2 エポックは固定の学習率（0.0005）で学習を行った．識別学習（boosted MMI 学習）を行う際には，学習率は CE 完全モデルに対して 0.001 とし，低ランクモデルに対しては 0.0001 である．識別学習の繰返し回数は 4 回とした．これは事前の実験により，4 回以上モデル更新を繰り返しても性能が向上せず，逆に過学習の影響で性能が低下することを確認したためである．確率勾配法は，低ラン

表 2 CHiME チャレンジトラック 2 開発セット（si_dt_05）での WER [%]．DNN モデルを用いて特異値分解（SVD）とファインチューニング（FT）の騒音残響環境下音声認識における効果を示している．初期モデルはクロスエントロピー（CE）完全モデルである．3 種の SVD をこのモデルに適用し，SVD{1,2,3} が得られた．入力特徴量は MFCC + LDA+MLLT + SAT+fMLLR である（40 次元 × 連続 9 フレーム）

Table 2 WER [%] on the CHiME challenge track 2, development set (si_dt_05), using DNN model showing the effectiveness of singular value decomposition (SVD) and fine-tuning (FT) on noisy reverberated speech recognition. Initial model was CE-full model. Applying three types of SVD to this model, SVD {1,2,3} models were obtained. Input features were MFCC + LDA+MLLT + SAT+fMLLR (40 dimension × contiguous 9 frames).

	FT	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
CE-full (2.85 M)	✓	53.4	42.4	34.5	27.9	24.8	20.5	**33.9**
SVD1 (1.47 M)	-	59.0	48.5	40.2	34.3	31.1	26.1	39.9
	✓	53.8	42.9	35.6	28.7	25.4	21.9	34.7
SVD2 (1.52 M)	-	59.1	48.6	40.1	34.3	31.1	26.1	39.9
	✓	52.7	42.1	34.3	28.3	25.0	20.6	33.8
SVD3 (1.59 M)	-	58.9	48.1	40.0	34.0	30.5	25.4	39.5
	✓	51.8	41.0	32.6	26.4	23.6	19.9	32.6
SVD3' (1.91 M)	-	57.1	46.7	38.9	33.0	29.1	24.1	38.2
	✓	51.8	40.7	32.9	26.2	23.8	19.9	**32.5**
(CE-full (1.54 M)	✓	55.2	44.3	35.8	30.2	26.4	21.9	35.6)

クモデルに対して安定性を欠く傾向にあるので，低ランクモデルに対する学習率は完全モデルのそれよりも小さくしなければならない．

3.2 結果と考察

3.2.1 最良のタイプの SVD

表 2 には，開発セット（si_dt_05）での WER を示す．これらのモデルはすべて CE モデルであり，系列の識別学習は行っていない．SVD を行った後に，FT をしないと，すべての低ランクモデルの性能は顕著に低下している．FT により，すべてのモデルの性能が著しく向上する．これは文献 [4] の結果とも一致する．それらのモデルの中でも，SVD3 のように最終層にのみ SVD を適用したタイプの分解が最良であった．全層に SVD を適用した SVD1 タイプのモデルの性能は，入力層以外の層に SVD を適用した SVD2 タイプのモデルの性能よりも劣った．両者でモデルパラメータの数にあまり差がないのに性能に差が出たことから，これは最下層の重み行列が上層の重み行列よりも実効的なランクが高く，低ランク近似により精度が低下したと推測できる．参考までに初めからパラメータ数を半分程度にした完全 CE モデル（CE-full (1.54 M)）の性能を表 2 の最下段に示している．この場合はベースラインの CE-full (2.85 M) よりも性能が劣り，SVD によるモデル縮減の有効性が示された．

表 3 CHiME チャレンジトラック 2 開発セット（si_dt_05）での WER [%]．DNN モデルを用いて系列の識別学習の効果を示している．初期モデルは CE モデルであり，3 種の手法を評価している

Table 3 WER [%] on the CHiME challenge track 2, development set (si_dt_05), using DNN model showing the effectiveness of sequence discriminative training. Initial model was CE model and three types of combinations were evaluated.

	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
bMMI-full (2.85 M)	48.4	36.7	30.2	24.2	20.7	17.3	**29.6**
– Combination 1 (with CE low-rank model) –							
SVD1 (1.47 M) bMMI	47.9	37.6	30.6	24.4	21.2	18.1	30.0
SVD2 (1.52 M) bMMI	47.4	36.5	29.3	24.0	20.6	17.3	29.2
SVD3 (1.59 M) bMMI	46.4	35.1	28.1	23.0	19.4	16.5	**28.1**
SVD3' (1.91 M) bMMI	47.0	35.3	28.4	22.8	19.5	16.8	28.3
– Combination 2 (with bMMI full model) –							
SVD1 (1.47 M) wo FT	54.6	43.3	35.8	30.8	27.3	22.4	35.7
w FT	53.3	42.5	34.9	28.8	25.3	21.7	34.4
SVD2 (1.52 M) wo FT	54.8	43.3	35.9	30.8	27.3	22.5	35.8
w FT	52.8	41.6	34.4	27.7	24.6	21.0	33.7
SVD3 (1.59 M) wo FT	54.1	42.1	34.6	29.4	25.9	21.8	34.7
w FT	51.7	41.3	33.2	26.6	23.6	19.7	32.7
SVD3' (1.91 M) wo FT	53.0	41.1	33.9	27.7	24.7	20.8	33.5
w FT	51.6	40.6	33.1	26.6	23.5	19.7	32.5
– Combination 3 (with Combination 2 model) –							
SVD1 (1.47 M) bMMI	48.6	37.8	30.8	25.2	21.5	18.5	30.4
SVD2 (1.52 M) bMMI	48.1	37.0	30.5	23.8	21.2	17.5	29.7
SVD3 (1.59 M) bMMI	47.7	36.9	29.4	23.4	20.6	17.0	29.1
SVD3' (1.91 M) bMMI	47.7	37.1	29.3	23.3	20.6	17.1	29.2

3.2.2 最良の識別学習の手順

表 3 には，MMI 識別学習したモデルの WER を示す．表 2 の CE-full と表 3 の bMMI-full を比較すると，識別学習により，完全モデルの性能が 4.3%（絶対値，以下同様）向上したことが分かる．組合せ 1（Combination 1）では，CE 低ランクモデルからの識別学習による性能向上は，4.2–4.7%で完全モデルの性能向上を上回った．これは文献 [23] での，一般的な音声認識のための識別学習の検討において，モデルが小さいほど識別学習の効果が大きいという知見と一致する．最終的に，bMMI 低ランクモデルは bMMI-full の性能を 1.5%上回った．

組合せ 2（Combination 2）では，FT なしで比較すると，表 3 の FT なしの bMMI 低ランクモデルの性能は，表 2 の FT なしの CE 低ランクモデルの性能を上回っていた．しかしながら，bMMI 低ランクモデルでは，FT は 1.3–2.1%程度の性能向上にとどまった．これは識別的基準によって学習されたモデルが最後に CE 基準で FT をかけることによって，識別学習の効果が薄れてしまったためと

表 4 CHiME チャレンジトラック 2 開発セット（si_dt_05）での GMM モデルを用いた場合の WER [%]

Table 4 WER [%] on the CHiME challenge track 2, development set (si_dt_05), using GMM model.

	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
GMM (2.43 M) bMMI	58.0	46.3	37.1	30.3	25.9	21.5	36.5
GMM (2.45 M) f-bMMI	54.7	43.9	35.9	28.3	24.2	20.6	34.6

考えられる．

組合せ 3（Combination 3）では，bMMI 低ランクモデルに識別学習を再び行うことで性能は 3–4%向上した．このことから，学習の最後に識別学習を行うことが有効であることが分かる．ただし，CE 低ランクモデルに識別学習を適用した場合ほどの性能向上は得られなかった．これは過学習によるものと考えられる．

全体的に見て，組合せ 1 が最も良いと結論づけることができる．必要な計算量も Combination $1 < 2 \ll 3$ の順[*6]なので，Combination 1 を採用するのがよいと考えられる．また前節での検討同様，SVD1，2，3 の中では SVD3 が最良であることが分かる．

3.2.3 GMM との比較

表 4 には，同じ特徴量を用いた場合の GMM の WER を示している．GMM-HMM の総ガウス分布数は 30,000 である[*7]．この場合，各特徴量次元に関して平均と分散のパラメータがあるため，平均と分散を表すパラメータ数は $30,000\,[\text{分布}] \times 2 \times 40\,[\text{次元/分布}]$ となる．これに混合重みのパラメータが 30,000 あるため，総パラメータ数は 2.43 M となる．

識別学習（bMMI）やそれに加えて特徴量領域での識別学習（feature-space bMMI; f-bMMI）を行った場合[*8]でも，DNN モデルに比べて認識性能は大幅に低いことが分かる．また SVD によるパラメータ数縮減により，GMM のパラメータよりも大幅に少ないパラメータ数にすることができている．

3.2.4 評価セット

表 5 には，評価セット（si_et_05）の WER を示す．開発セットでの場合と傾向は同じである．識別学習との組合せは，3.2.2 項での検討から，組合せ 1 を選択した．SVD 3 のタイプの分解が効果的であり，これによる性能は元の bMMI 完全モデルの性能を絶対値で 1%上回った．また GMM の最良の結果を 6.5%上回っていることから，SVD を行った DNN の有効性が確かめられた．

[*6] Combination 3 は 1，2 に対して，識別学習が 1 回多いので，ラティスの生成・識別学習の計算量が 1，2 に比べて大きくなる．1 と 2 は SVD と識別学習が 1 回ずつで同じであるが，1 は小さいモデルに対して識別学習を行うので，2 に比べると若干計算量が少なくて済む．

[*7] HMM の状態共有構造は DNN の場合と同じである．

[*8] 特徴量領域の識別学習を行うと，特徴量変換行列分（400×40）パラメータ数が増加する．

表 5 CHiME チャレンジトラック 2 評価セット（si_et_05）での WER [%]．DNN モデルを用いて系列の識別学習の効果を示している．初期モデルは CE モデルである

Table 5 WER [%] on the CHiME challenge track 2, evaluation set (si_et_05), using DNN model showing the effectiveness of sequence discriminative training. Initial model was CE model.

	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
CE-full (2.85 M)	44.5	35.7	29.5	22.0	16.6	15.3	27.3
bMMI-full	39.0	28.9	23.4	18.3	13.9	12.0	**22.6**
– Combination 1 (with CE low-rank model) –							
SVD1 (1.47 M) bMMI	40.0	29.7	24.0	18.5	14.4	12.7	23.2
SVD2 (1.52 M) bMMI	39.5	28.4	23.1	18.2	13.5	12.1	22.5
SVD3 (1.59 M) bMMI	37.9	27.6	22.5	17.4	12.9	11.0	**21.6**
SVD3' (1.91 M) bMMI	37.5	27.7	22.4	17.5	12.8	11.5	**21.6**
(GMM (2.45 M) f-bMMI	45.5	37.4	30.0	22.2	17.9	15.9	28.1)

4. おわりに

DNN モデルのパラメータを削減するために，SVD による低ランク近似を用いたモデル縮減手法を検討した．従来はクリーン音声での評価であったが，本報では多様性が増し，音響モデルがより複雑になると想定される騒音残響環境下音声認識に適用した．まずは，SVD を適用する層の検討を行った．実験により，パラメータ数にそれほど差がない設定であっても，DNN の最終層だけに低ランク近似を適用する，もしくは第 1 層を除くすべての層に低ランク近似を適用する方が，すべての層に低ランク近似を適用するよりも性能が高いことが示された．

さらにモデル縮減と系列の識別学習を組み合わせる方法を検討した．識別学習とモデルサイズ削減を組み合わせる際は，まず元のモデルを縮減し，それから識別学習を低ランクモデルに対して行うことが最も効果的であるということが明らかになった．このように識別学習された低ランクモデルは，パラメータ数の多い GMM モデルを単語誤り率で 6.5%（絶対値）上回り，識別学習されたモデル縮減しないモデルをも 1%上回る性能を示した．

参考文献

[1] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Processing Magazine*, Vol.28, pp.82–97 (2012).

[2] Kanda, N., Takeda, R. and Obuchi, Y.: Elastic Spectral Distortion for Lowresource Speech Recognition with Deep Neural Networks, *Proc. ASRU*, pp.309–314 (2013).

[3] Sainath, T., Kingsbury, B., Sindhwani, V., Arisoy, E. and Ramabhadran, B.: Low-Rank Matrix Factorization for Deep Neural Network Training with High-Dimensional Output Targets, *Proc. ICASSP*, pp.6655–6659 (2013).

[4] Xue, J., Li, J. and Gong, Y.: Restructuring of Deep

Neural Network Acoustic Models with Singular Value Decomposition, *Proc. INTERSPEECH*, pp.2365–2369 (2013).
[5] Povey, D. and Woodland, P.: Minimum Phone Error and I-smoothing for Improved Discriminative Training, *Proc. ICASSP*, Vol.I, pp.105–108 (2002).
[6] McDermott, E., Hazen, T., Le Roux, J., Nakamura, A. and Katagiri, S.: Discriminative Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.15, pp.203–223 (2007).
[7] 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).
[8] Bridle, J. and Dodd, L.: An Alphanet Approach to Optimising Input Transformations for Continuous Speech Recognition, *Proc. ICASSP*, pp.277–280 (1991).
[9] Kingsbury, B.: Lattice-based Optimization of Sequence Classification Criteria for Neural-network Acoustic Modeling, *Proc. ICASSP*, pp.3761–3764 (2009).
[10] Wang, G. and Sim, K.: Sequential Classification Criteria for NNs in Automatic Speech Recognition, *Proc. INTERSPEECH*, pp.441–444 (2011).
[11] Kingsbury, B., Sainath, T. and Soltau, H.: Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization, *Proc. INTERSPEECH*, pp.485–488 (2012).
[12] Jaitly, N., Nguyen, P., Senior, A. and Vanhoucke, V.: Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Rrecognition, *Proc. INTERSPEECH* (2012).
[13] Veselý, K., Ghoshal, A., Burget, L. and Povey, D.: Sequence-discriminative Training of Deep Neural Networks, *Proc. INTERSPEECH* (2013).
[14] Kubo, Y., Hori, T. and Nakamura, A.: Large Vocabulary Continuous Speech Recognition Based on WFST Structured Classifiers and Deep Bottleneck Features, *Proc. ICASSP*, pp.7629–7633 (2013).
[15] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Petr, M., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. and Veselý, K.: The Kaldi Speech Recognition Toolkit, *Proc. ASRU*, pp.1–4 (2011).
[16] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F. and Matassoni, M.: The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines, *Proc. ICASSP*, pp.126–130 (2013).
[17] Tachioka, Y., Watanabe, S., Le Roux, J. and Hershey, J.: Discriminative Methods for Noise Robust Speech Recognition: A CHiME Challenge Benchmark, *Proc. 2nd CHiME Workshop on Machine Listening in Multisource Environments*, pp.19–24 (2013).
[18] Tachioka, Y., Watanabe, S. and Hershey, J.: Effectiveness of Discriminative Training and Feature Transformation for Reverberated and Noisy Speech, *Proc. ICASSP*, pp.6935–6939 (2013).
[19] Haeb-Umbach, R. and Ney, H.: Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition, *Proc. ICASSP*, pp.13–16 (1992).
[20] Gopinath, R.: Maximum Likelihood Modeling with Gaussian Distributions for Classification, *Proc. ICASSP*, pp.661–664 (1998).
[21] Anastasakos, T., McDonough, J., Schwartz, R. and Makhoul, J.: A Compact Model for Speaker-adaptive Training, *Proc. ICSLP*, pp.1137–1140 (1996).
[22] Gales, M.: Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, *Computer Speech and Language*, Vol.12, pp.75–98 (1998).
[23] McDermott, E.: *Discriminative Training for Speech Recognition*, Doctrorial dissertation, Waseda University (1997).

太刀岡 勇気 （正会員）

2006 年東京大学工学部建築学科卒業．2008 年同大学大学院修士課程修了．同年三菱電機（株）入社．以来，音声認識の研究開発に従事．現在，同社情報技術総合研究所音声・言語処理部研究員．2008 年日本建築学会優秀修士論文賞，2014 年日本音響学会粟屋潔学術奨励賞．日本音響学会，日本建築学会，計量国語学会，IEEE 各会員．

渡部 晋治

1999 年早稲田大学理工学部物理学科卒業，2001 年同大学大学院修士課程修了．同年 NTT コミュニケーション科学基礎研究所入社．2012 年より Mitsubishi Electric Research Laboratories（MERL）senior principal member．2009 年ジョージア工科大学客員研究員．博士（工学）．音声認識を中心とした音声言語処理の研究に従事．2003 年日本音響学会粟屋潔学術奨励賞，2004 年電子情報通信学会論文賞，2006 年日本音響学会独創研究奨励賞板倉記念，電気通信普及財団テレコムシステム技術賞各受賞．2012 年より IEEE Transaction on Audio, Speech, and Language Processing の Associate Editor，2014 年より IEEE Signal Processing Society, Speech and Language Technical Committee，および APSIPA Speech, Language, and Audio Technical Committee 等を歴任．日本音響学会，電子情報通信学会各会員，IEEE シニア会員．

ルルー ジョナトン

パリ高等師範学校にて理学士および理学修士を取得後，2009 年東京大学およびパリ 6 大学博士課程修了，博士．同年 NTT コミュニケーション科学基礎研究所リサーチアソシエイト．現在，MERL Principal Researcher．信号処理技術および機械学習技術を用いた音声・音響処理の研究に従事．IEEE シニア会員．IEEE Audio and Acoustic Signal Processing Technical Committee 会員．

ハーシー ジョン

カリフォルニア大学サンディエゴ校にて博士号取得．博士論文のテーマは生成グラフィカルモデルの音声分離，顔追跡および両者への応用である．2004 年 Microsoft Research にて客員研究員．その後ニューヨークの IBM T. J. Watson 研究センターに移り，Speech Algorithms and Engines グループの研究員および耐騒音プロジェクトのチームリーダを務める．2010 年より MERL にて，Speech and Audio チームのリーダを務める．信号強調・分離，音声認識，言語処理，適応ユーザインタフェースのための機械学習の研究に従事．