Sequence Discriminative Training for Low-Rank Deep Neural Networks

Yuuki Tachioka
Information Technology R & D Center
Mitsubishi Electric Corporation
Kanagawa, Japan 247-8501
Email: Tachioka.Yuki@eb.MitsubishiElectric.co.jp

Shinji Watanabe, Jonathan Le Roux, and John R Hershey
Mitsubishi Electric Research Laboratories
Cambridge, USA 02139
Email: {watanabe,leroux,hershey}@merl.com

***Abstract*—Deep neural networks (DNNs) have proven very successful for automatic speech recognition but the number of parameters tends to be large, leading to high computational cost. To reduce the size of a DNN model, low-rank approximations of weight matrices, computed using singular value decomposition (SVD), have previously been applied. Previous studies only focused on clean speech, whereas the additional variability in noisy speech could make model reduction difficult. Thus we investigate the effectiveness of this SVD method on noisy reverberated speech. Furthermore, we combine the low-rank approximation with sequence discriminative training, which further improved the performance of the DNN, even though the original DNN was constructed using a 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 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 low-rank discriminatively trained model outperformed the full discriminatively trained model.**

***Index Terms*—automatic speech recognition, deep neural networks, singular value decomposition, discriminative training**

I. Introduction

Deep neural network (DNN) have been very successful in the area of automatic speech recognition (ASR) [1]. Although DNNs outperform conventional Gaussian mixture model (GMM) in many cases [1], [2], the number of parameters in DNNs tends to be greater than that in GMM. For example, in the study of large vocabulary continuous speech recognition ASR task [2], for a GMM based system, the number of hidden Markov model (HMM) states is 3k and the mixture of Gaussian per state is 32; totally, the number of parameters is less than 10M. On the other hand, for DNN based system, the number of HMM states is the same, the number of nodes in each hidden layer is 2k, and the number of hidden layer is seven; totally, the number of parameters is over 30M. Thus the DNN model has three times larger number of parameters, which increases the computational cost.

There are some attempts to reduce a DNN model size [3], [4]. Xue *et al*. have proposed to apply singular value decomposition (SVD) to DNN models and reduce the total number of parameters. Their method reduces the rank of the weight matrices and they show that SVD combined with fine-tuning is effective experimentally [4]. In their experiments, the speech data properties are not clear because their experiments were performed on private data. However typical LVCSR data uses close-talking microphones and so is relatively clean. Under reverberant and noisy environments in far-field conditions, DNN acoutic models need to be more complex to handle the increased variability of the signal. In this scenario, model reduction may have a negative effect on performance. Thus, the effectiveness of this technique on noisy reverberated speech needs to be evaluated.

Previous experiments on model reduction have focused on frame-level discriminative criteria such as cross-entropy (CE). However, sequence-level discriminative training of acoustic models, using criteria such as maximum mutual information (MMI) has improved the performance of conventional maximum likelihood based GMM models [5], [6], [7], as well as DNNs [8], [9], [10], [11], [12], [13], [14]. When combining the model reduction technique above with a sequence discriminative training, we need to investigate the effect of the order in which model reduction and sequence discriminative training are applied. For example it may be important to perform discriminative training after after model reduction in order to recover from loss of performance due to the approximation. We evaluate three approaches: the first approach is to apply SVD-based rank-reduction and fine-tuning for a CE full model and to perform discriminative training on a low-rank CE model; the second approach is to apply rank-reduction and fine-tuning for a MMI full model; the third approach is to perform discriminative training on the MMI low-rank model obtained from the second approach. This paper investigates a several combinations of SVD reduction techniques with DNN sequence training experimentally for noisy reverberant speech recognition.

II. DNN-HMM Hybrid ASR Systems

DNN-HMM hybrid ASR systems have been shown to outperform conventional GMM-HMM systems in a wide variety of conditions. Let us assume that DNN acoustic parameters θ are composed of L hidden layer. Here, 0-th layer is the input layer and $(L+1)$-th layer is the output layer. For the l-th layer of DNN acoustic models ($0 \leq l \leq L+1$), n-dimensional input feature is denoted as $\boldsymbol{x}^l$. The output

978-1-4799-7088-9/14/$31.00 ©2014 IEEE

feature is m-dimensional and also an input feature of the $(l+1)$-th layer, thus, this can be denoted as $\boldsymbol{x}^{l+1}$. Non-linear operation f is used in addition to linear operation. For hidden layers, sigmoid function is used as f, whereas for the last layer, softmax function is used. Weight matrix of $\boldsymbol{A}^l_{m\times n}$ and offset $\boldsymbol{b}^l$ are trained using back propagation (fine-tuning) with stochastic gradient descent where the lower-suffix of the matrices represents their dimension. From the lower layer to the higher layer, feature $\boldsymbol{x}$ is propagated as

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

In this paper, the DNN was constructed by discriminatively training hidden layers one by one.

The DNN model provides posterior probabilities for the HMM state j at frame t. In the hybrid DNN approach, the pseudo acoustic likelihood p is obtained as

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

where $p_0\left(j\right)$ is the prior probability calculated from the count of training data. DNN input feature $\boldsymbol{x}^0_t$ is a spliced feature $[\boldsymbol{x}_{t-s},\ldots,\boldsymbol{x}_t,\ldots,\boldsymbol{x}_{t+s}]$ in contiguous $(2s+1)$ frames. The DNN output is an output probability of each context-dependent HMM state. A softmax activation function is used for the output layer

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

where a is the pre-activation value of the output layer node j, as a function of the input $\boldsymbol{x}^0_t$ to the DNN.

III. Reducing DNN Model Size Using Low-Rank Approximation

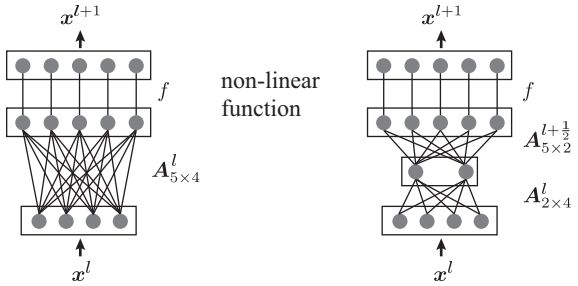
Although DNN-HMM systems outperforms conventional GMM-HMM systems, the number of parameters in DNN tends to be greater than that in GMM. Therefore, [4] proposed to use SVD to reduce the rank of the weight matrix $\boldsymbol{A}^l$ for a given layer l to reduce the total number of parameters. Eq. (4) factorizes matrix $\boldsymbol{A}^l_{m\times n}$ as

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

where $\boldsymbol{\Sigma}$ is a diagonal matrix, whose elements are singular values arranged in a descending order ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$), $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns, and $\top$ denotes transpose. To reduce the number of parameters of $\boldsymbol{A}_{m\times n}$, the k largest singular values and their corresponding left and right singular vectors are used to from the low-rank factorization,

$$\begin{aligned}\boldsymbol{A}^l_{m\times n} &\approx \boldsymbol{U}_{m\times k}\boldsymbol{\Sigma}_{k\times k}\boldsymbol{V}^\top_{k\times n} \quad (k<n),\\ &= \left[\boldsymbol{U}_{m\times k}\sqrt{\boldsymbol{\Sigma}_{k\times k}}\right]\left[\sqrt{\boldsymbol{\Sigma}_{k\times k}}\boldsymbol{V}^\top_{k\times n}\right], \\ &= \boldsymbol{A}^{l+\frac{1}{2}}_{m\times k}\boldsymbol{A}^l_{k\times n}.\end{aligned} \tag{5}$$

Originally, computational costs of the matrix multiplication $\boldsymbol{A}\boldsymbol{x}$ are proportional to $O(mn)$. After low rank approximation, this becomes $O\left((m+n)k\right)$, so that computation is reduced for $k < mn/(m+n)$. The low rank approximation can be viewed as decomposing the l-th layer into two layers, the first a linear layer with weight matrix $\boldsymbol{A}^l_{k\times n}$, and the second a sigmoid layer with weight matrix, $\boldsymbol{A}^{l+\frac{1}{2}}_{m\times k}$, as shown in Fig. 1. In [4], $\boldsymbol{A}^l_{m\times n}$ is decomposed into the alternative factorization $[\boldsymbol{U}_{m\times k}]\left[\boldsymbol{\Sigma}_{k\times k}\boldsymbol{V}^\top_{k\times n}\right]$ which is functionally equivalent to (5). With offsets, the new layers become:



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

Fig. 1. Reducing DNN model parameters via low-rank factoriztion, from (a) $5\times4=20$ to (b) $5\times2+2\times4=18$.

$$\begin{aligned}\boldsymbol{x}^{l+\frac{1}{2}} &= \boldsymbol{A}^l_{k\times n}\boldsymbol{x}^l + \boldsymbol{b}^l,\\ \boldsymbol{x}^{l+1} &= f\left(\boldsymbol{A}^{l+\frac{1}{2}}_{m\times k}\boldsymbol{x}^{l+\frac{1}{2}} + \boldsymbol{b}^{l+\frac{1}{2}}\right).\end{aligned} \tag{6}$$

where $\boldsymbol{b}^l$ is a k-dimensional vector initialized to zero, and $\boldsymbol{b}^{l+\frac{1}{2}}$ is the original $\boldsymbol{b}^l$. Fine tuning based on various discriminative objective functions can then be applied.

IV. Cross-Entropy Training for DNN

For the CE criterion, the objective function is

$$\mathcal{F}_{\mathrm{CE}}(\theta) = \sum_r\sum_t\sum_j \hat{p}(j,t)\log\frac{\hat{p}(j,t)}{p(j|\boldsymbol{x}^0_t)}, \tag{7}$$

where $\hat{p}(j,t)$ is the reference distribution for class label j at time t. The gradient with respect to a is

$$\frac{\partial\mathcal{F}_{\mathrm{CE}}}{\partial a(j)} = p(j|\boldsymbol{x}^0_t) - \hat{p}(j,t). \tag{8}$$

Gradient descent based on the chain rule, known as back-propagation, can then be used for optimization of the DNN parameters

V. Sequence MMI Training for DNN

Sequence discriminative training considers the full HMM state sequences. The DNN acoustic models parameter θ is optimized according to the MMI objective function:

$$\mathcal{F}_{\mathrm{MMI}}(\theta) = \sum_r \log\frac{p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_s\right)^\kappa p_L(s)}, \tag{9}$$

where $\boldsymbol{x}_{1:T_r}$ is the r-th utterance's acoustic feature sequence whose length is T_r, $\mathcal{H}_{s_r}$ is the state sequence for correct label s_r and $\mathcal{H}_s$ is the state sequence for recognition hypothesis s, κ is the acoustic scale, and p_L is the language model likelihood.

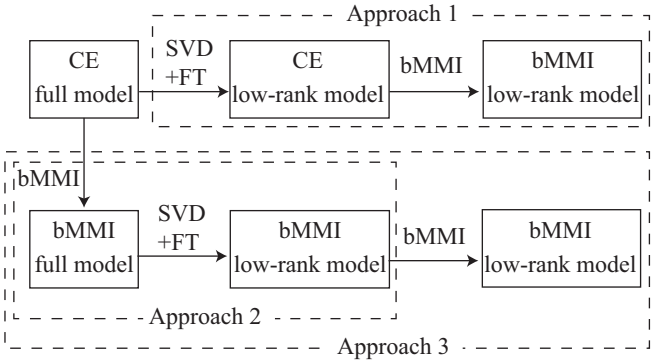


Fig. 2. Three approaches to generate MMI low-rank model with fine tuning (FT).

The boosted version of Eq. (9) also follows

$$\mathcal{F}_{\mathrm{bMMI}}(\theta) = \sum_r \log \frac{p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\theta\left(\boldsymbol{x}_{1:T_r}|\mathcal{H}_s\right)^\kappa p_L(s) e^{-bA(s,s_r)}}. \tag{10}$$

The gradient w.r.t. softmax activation a becomes [8], [9], [10], [13]:

$$\frac{\partial \mathcal{F}_{\mathrm{bMMI}}(\theta)}{\partial a(j)} = \kappa(\gamma_{j,t}^{num} - \gamma_{j,t}^{den}), \tag{11}$$

where $\gamma_{j,t}^{num}$ and $\gamma_{j,t}^{den}$ are the numerator and denominator posterior in Eq. (9) or (10). All DNN parameters are derived from Eq. (11) based on the back-propagation procedure.

VI. Combination of Sequence Discriminative Training with SVD

The order of discriminative training and model reduction is important and not trivial. Fig. 2 shows three approaches to generate discriminatively trained low-rank models, which we tested in this paper. For all approaches, the initial model is a cross-entropy (CE) trained full model. The first approach, approach 1, is to apply SVD and fine-tuning for a CE full model and to perform discriminative training on a low-rank CE model; the second approach, approach 2, is to apply SVD and fine-tuning for a MMI full model; the third approach, approach 3, is to perform discriminative training on the MMI low-rank model obtained from approach 2.

VII. Experiments

A. Setup

We evaluated the performance on the second CHiME challenge Track 2, which was designed for evaluating the word error rate (WER) of a medium vocabulary task (Wall Street Journal (WSJ0)) under reverberated and non-stationary noisy environments [15]. The language model size was 5 k (basic). The development set, si_dt_05, contained 409 utterances from 10 speakers; the evaluation data set, si_et_05 contained 330 utterances from 12 speakers (Nov'92). Acoustic models were trained on the training set, which contained 7,138 utterances from 83 speakers. The acoustic scale κ was tuned using si_dt_05. These data simulated realistic environments. Noise was non-stationary, such as other speakers' utterances, household noise, or music and was added to 'isolated' speech at SNR $= \{-6, -3, 0, 3, 6, 9\}$dB. Although the database provided two-channel data, we used noise-suppressed single-channel data obtained by prior-based binary masking [16].

TABLE I
DNN Structure Corresponding to SVD {1,2,3}.

	input $\longrightarrow$ output
CE-full (2.85M)	$360 \times 331 + 331^2 \times 2 + 331 \times 8000$
SVD1 (1.47M)	$360 \times \underline{100} + \underline{100} \times 331 + (331 \times \underline{96}) \times 2 \times 2$ $+ 331 \times \underline{162} + \underline{162} \times 8000$
SVD2 (1.52M)	$360 \times 331 + (331 \times \underline{96}) \times 2 \times 2$ $+ 331 \times \underline{162} + \underline{162} \times 8000$
SVD3 (1.59M)	$360 \times 331 + 331^2 \times 2 + 331 \times \underline{160} + \underline{160} \times 8000$
SVD3 (1.91M)	$360 \times 331 + 331^2 \times 2 + 331 \times \underline{200} + \underline{200} \times 8000$

The settings of the acoustic features and feature transformation were as follows [17]. We used Povey's implementation of neural network training in the Kaldi toolkit [18]. The baseline features were 0th$\sim$12th order MFCCs + Δ + $\Delta\Delta$. Feature transformation techniques (linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT)) and speaker adaptation techniques (speaker adaptive training (SAT) and feature space maximum likelihood linear regression (fMLLR)) were used to obtain 40-dimensional speaker-adapted features. The DNN input features were 9 consecutive frames of these feature concatenated into a 360-dimensional feature vector.

The procedure of training acoustic models and the setup of feature transformations are described in [16], [17]. The number of the context-dependent HMM states was 1,989, which is equal to that of the last softmax layer outputs. The number of hidden layer was three. The initial learning rate for a CE full model was 0.01 and was decreased to 0.001 at the end of training. Starting from single-layer neural networks, we added layers one by one in every two iterations. One iteration used 400,000 samples. The total number of parameters was summarized in Table I. In the CE training, the number of epoch was 15 for reducing learning rates and 5 for the constant final learning rate. Minibatch size was 128. After applying SVD to the CE full model or MMI full model, fine-tuning needed 3 epochs for reducing learning rates from 0.001 to 0.0005 and 2 epochs for the constant final learning rate. For boosted MMI training, the learning rate was 0.001 when starting with the full CE model and 0.0001 for the low-rank models. The learning rate must be smaller for low-rank models than for full models because stochastic gradient descent tends to be less stable for low-rank models.

We evaluated three ways of applying SVD to full models: the first one was applying SVD to the all hidden layers (SVD 1); the second one was applying SVD to the all hidden layers except the first hidden layer because the first hidden layer has an important role for extracting features (SVD 2); the third one was applying SVD to the last layers, which have the largest

TABLE II
WER [%] ON THE CHiME CHALLENGE TRACK 2 (SI_DT_05) USING DNN MODEL SHOWING THE EFFECTIVENESS OF 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).

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
CE-full (2.85M)	53.44	42.40	34.53	27.94	24.77	20.49	**33.93**
SVD1 (1.47M) wo FT	58.95	48.52	40.15	34.33	31.05	26.10	39.85
w FT	53.81	42.88	35.58	28.74	25.36	21.92	34.72
SVD2 (1.52M) wo FT	59.06	48.59	40.12	34.30	31.07	26.08	39.87
w FT	52.68	42.06	34.34	28.29	24.96	20.58	33.82
SVD3 (1.59M) wo FT	58.93	48.12	39.98	33.97	30.48	25.36	39.47
w FT	51.82	41.04	32.64	26.42	23.63	19.87	32.57
SVD3 (1.91M) wo FT	57.09	46.72	38.91	33.04	29.08	24.07	38.15
w FT	51.76	40.67	32.87	26.21	23.79	19.86	**32.53**

number of parameters (SVD 3).

B. Results and discussions

1) Which type of SVD is the best?: Table II shows the WER on si_dt_05. These models were all cross-entropy (CE) model without sequence discriminative training. After SVD, without fine-tuning (FT), every low-rank model degraded significantly. Fine-tuning greatly improved the performance of all models, consitent with the results of [4]. Among them, the SVD3 type of decomposition was the best. The performance of SVD1 was inferior to that of SVD2. This indicates that the weight matrices in the first layer had higher effective rank than those in the upper layers.

2) Which type of discriminative training approach is the best?: Table III shows the results of discriminatively trained models. Sequence discriminative training led to significant improvements for the full model. In approach 1, The performance improvement of low-rank CE model was larger than CE full model, which is reported in general discriminative training studies for speech recognition that smaller models have bigger improvement [19]. In approach 2, without FT, the performance of bMMI low-rank model was better than that of CE low-rank model without FT, however, for the bMMI low-rank model, FT was less effective. In approach 3, discriminative training on the bMMI low-rank model again improved the performance but was less effective than for CE low-rank model perhaps due to over-training. Overall approach 1 was the best.

3) Evaluation set: Table IV shows the results on evaluation set (si_et_05). Tendencies were the same to the development set. SVD 3 types of decomposition was effective and their performance was superior to that of the original bMMI full model by 1% absolute.

VIII. CONCLUSION

To reduce the number of DNN parameters, a model reduction technique using low-rank approximation has been applied to noisy reverberant speech recognition. Experiments demonstrate that low-rank approximation of the last layer of DNN or all layers except the first layer is more effective than rank reduction of all layers. Sequence discriminative training further improved performance. The most effective combination of discriminative training with model reduction was to reduce the base model first and then to perform discriminative training on the low-rank model. This discriminatively trained low-rank model outperformed the discriminatively trained full model.

TABLE III
WER [%] ON THE CHiME CHALLENGE TRACK 2 (SI_DT_05) USING DNN MODEL SHOWING THE EFFECTIVENESS OF SEQUENCE DISCRIMINATIVE TRAINING. INITIAL MODEL IS CROSS-ENTROPY (CE) MODEL. THREE TYPES OF APPROACHES WERE EVALUATED.

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
bMMI-full (2.85M)	48.37	36.66	30.15	24.18	20.71	17.27	**29.56**
*Approach 1 (from CE low-rank model)							
SVD1 (1.47M) bMMI	47.87	37.62	30.61	24.43	21.23	18.07	29.97
SVD2 (1.52M) bMMI	47.38	36.47	29.30	24.00	20.64	17.32	29.19
SVD3 (1.59M) bMMI	46.36	35.11	28.06	23.03	19.41	16.48	**28.08**
SVD3 (1.91M) bMMI	47.03	35.31	28.38	22.82	19.53	16.77	28.31
*Approach 2 (from bMMI full model)							
SVD1 (1.47M) wo FT	54.61	43.30	35.79	30.80	27.25	22.39	35.69
w FT	53.25	42.51	34.93	28.81	25.30	21.71	34.42
SVD2 (1.52M) wo FT	54.82	43.27	35.89	30.83	27.28	22.54	35.77
w FT	52.80	41.64	34.39	27.70	24.56	20.96	33.68
SVD3 (1.59M) wo FT	54.08	42.13	34.64	29.41	25.87	21.79	34.65
w FT	51.67	41.26	33.15	26.60	23.57	19.66	32.65
SVD3 (1.91M) wo FT	52.97	41.07	33.94	27.66	24.72	20.77	33.52
w FT	51.60	40.64	33.13	26.61	23.51	19.72	32.54
*Approach 3 (from Approach 2 model)							
SVD1 (1.47M) bMMI	48.61	37.81	30.82	25.20	21.52	18.47	30.41
SVD2 (1.52M) bMMI	48.10	36.95	30.54	23.82	21.20	17.54	29.69
SVD3 (1.59M) bMMI	47.71	36.91	29.36	23.35	20.56	16.96	29.14
SVD3 (1.91M) bMMI	47.74	37.14	29.34	23.31	20.55	17.14	29.20

TABLE IV
WER [%] ON THE CHiME CHALLENGE TRACK 2 (SI_ET_05) USING DNN MODEL SHOWING THE EFFECTIVENESS OF SEQUENCE DISCRIMINATIVE TRAINING. INITIAL MODEL IS CROSS-ENTROPY (CE) MODEL.

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
CE-full (2.85M)	44.48	35.72	29.46	21.99	16.63	15.34	27.27
bMMI-full	39.02	28.94	23.39	18.27	13.94	11.96	**22.59**
*Approach 1 (from CE low-rank model)							
SVD1 (1.47M) bMMI	40.03	29.67	23.97	18.51	14.40	12.74	23.22
SVD2 (1.52M) bMMI	39.47	28.41	23.11	18.16	13.53	12.11	22.47
SVD3 (1.59M) bMMI	37.94	27.59	22.53	17.39	12.87	10.97	**21.55**
SVD3 (1.91M) bMMI	37.51	27.65	22.42	17.52	12.82	11.47	21.57

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 28, pp. 82–97, 2012.

[2] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for lowresource speech recognition with deep neural networks," in *Proceedings of ASRU*, 2013, pp. 309–314.

[3] T. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *Proceedings of ICASSP*, 2013, pp. 6655–6659.
[4] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Proceedings of INTERSPEECH*, 2013, pp. 2365–2369.
[5] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of ICASSP*, 2002, pp. 105–108.
[6] E. McDermott, T. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 203–223, 2007.
[7] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proceedings of ICASSP*, 2008, pp. 4057–4060.
[8] J. Bridle and L. Dodd, "An alphanet approach to optimising input transformations for continuous speech recognition," in *Proceedings of ICASSP*, 1991, pp. 277–280.
[9] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proceedings of ICASSP*, 2009, pp. 3761–3764.
[10] G. Wang and K. Sim, "Sequential classification criteria for NNs in automatic speech recognition," in *Proceedings of INTERSPEECH*, 2011, pp. 441–444.
[11] B. Kingsbury, T. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Proceedings of INTERSPEECH*, 2012, pp. 485–488.
[12] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proceedings of INTERSPEECH*, 2012.
[13] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of INTERSPEECH*, 2013.
[14] Y. Kubo, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition based on WFST structured classifiers and deep bottleneck features," in *Proceedings of ICASSP*, 2013, pp. 7629–7633.
[15] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proceedings of ICASSP*, 2013, pp. 126–130.
[16] Y. Tachioka, S. Watanabe, J. Le Roux, and J. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proceedings of the 2nd CHiME Workshop on Machine Listening in Multisource Environments*, 2013, pp. 19–24.
[17] Y. Tachioka, S. Watanabe, and J. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proceedings of ICASSP*, 2013, pp. 6935–6939.
[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011, pp. 1–4.
[19] E. McDermott, *Discriminative training for speech recognition*. Doctrorial dissertation, Waseda University, 1997.