Exploiting Eigenposteriors for Semi-supervised Training of DNN Acoustic Models with Sequence Discrimination

Pranay Dighe[1,2], *Afsaneh Asaei*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{pranayd,aasaei,bourlard}@idiap.ch

Abstract

Deep neural network (DNN) acoustic models yield posterior probabilities of senone classes. Recent studies support the existence of low-dimensional subspaces underlying senone posteriors. Principal component analysis (PCA) is applied to identify eigenposteriors and perform low-dimensional projection of the training data posteriors. The resulted enhanced posteriors are applied as soft targets for training better DNN acoustic model under the student-teacher framework. The present work advances this approach by studying incorporation of sequence discriminative training. We demonstrate how to combine the gains from eigenposterior based enhancement with sequence discrimination to improve ASR using semi-supervised training. Evaluation on AMI meeting corpus yields nearly 4% absolute reduction in word error rate (WER) compared to the baseline DNN trained with cross entropy objective. In this context, eigenposterior enhancement of the soft targets is crucial to enable additive improvement using out-of-domain untranscribed data.

Index Terms: soft targets, eigenposteriors, automatic speech recognition, semi-supervised training, sequence discrimination

1. Introduction

Deep neural network (DNN) acoustic models rely on large amounts of accurately labeled training data to learn predicting senone posterior probabilities correctly for the input acoustic features. It was demonstrated in our recent works [1, 2, 3] that these posterior probability estimates can be improved by low-rank and sparsity based projection leading to enhanced posteriors and better ASR performance. Further, low-rank enhancements on DNN posteriors can be used to augment the labeled training data with high quality soft targets from untranscribed data for semi-supervised training of even more accurate DNN acoustic models.

It was shown in [3] that DNN posteriors live in low-dimensional senone-specific subspaces. A senone subspace encodes the inter-class correlations of the given senone with other senone classes. These correlations may arise from sequential dependencies among senones or from the structural acoustic relationships due to common roots or state tying in the decision tree. Senone specific subspaces were shown to be very low-dimensional - nearly 1% of the senone-dimension for ASR on AMI corpus. In practice however, inaccuracies in DNN training lead to the presence of unstructured high-dimensional errors in the posteriors. In [4, 1, 2, 3], low-rank and sparsity based approaches were exploited to characterize the senone subspaces in DNN posteriors. These approaches remove the high-dimensional noise in the DNN local estimate of the posteriors exploiting the global low-dimensional structure of the underlying senone classes.

In this work, we focus particularly on advancing further the principal component analysis (PCA) based approach proposed in [3]. This approach characterizes senone subspaces in terms of "*eigenposteriors*" (senone-specific principal components) which are used to enhance DNN posteriors by PCA based low-rank reconstruction. Under the student-teacher framework, this approach can incorporate untranscribed additional data for semi-supervised training. In this paper, we evaluate this approach under stronger baseline DNN acoustic models in this work and explore how to integrate it with sequence discriminative training of DNNs.

The present work contributes to the studies dedicated on exploiting the property that high-dimensional speech features lie on a low-dimensional manifold [5, 6, 4]. Low-dimensionality in DNN acoustic modeling has been used to achieve small footprint and manifold regularization in DNNs [7, 8, 9, 10]. In contrast to the earlier applications, our goal is to characterize the hidden subspace structure of the big data towards better representation of the local observations.

We exploit the framework of student-teacher DNN training that has been recognized promising for knowledge transfer and distillation [11, 12, 13, 14]. The basic idea of the student-teacher DNN training is that a teacher DNN (often trained with hard targets) provides soft targets for training a student DNN. The intuition is that the soft targets encode the teacher DNN information through the inter-dependencies among the output classes. Earlier studies investigate student-teacher framework to enable model compression and encapsulating the information of multiple models into a single network [12, 15]. This approach was also found beneficial for semi-supervised training exploiting untranscribed training data [14], although the investigation is limited to the in-domain data matching the initial transcribed speech used for supervised training. Semi-supervised training [16, 17, 18, 19] has been popular for low-resource tasks where cheap-to-obtain untranscribed data is readily available. The present work is a novel attempt towards exploring combination of sequence discriminative training with eigenposterior based semi-supervised training of DNN acoustic models to tackle a large vocabulary continuous speech recognition task on AMI meeting corpus [20].

In the rest of the paper, the procure of using eigenposteriors to obtain enhanced soft targets is outlined in Section 2. A student DNN is trained using enhanced posteriors and then used to generate soft targets for semi-supervised training as described in Section 3. Experimental analysis is carried out in Section 4. Section 5 presents the concluding remarks.

2. Eigenposteriors for Reliable Soft Targets

We perform principal component analysis of posteriors and low-rank reconstruction to obtain enhanced posteriors. A teacher DNN trained on binary hard alignments provides the initial posteriors. Enhanced posteriors are used as the soft targets to train a student DNN.

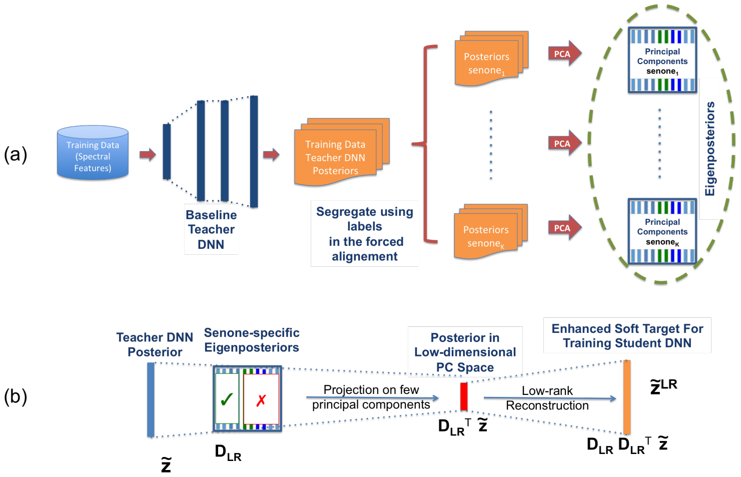
(a)
Training Data (Spectral Features)
Baseline Teacher DNN
Training Data Teacher DNN Posteriors
Segregate using labels in the forced alignement
Posteriors senone$_1$
Posteriors senone$_K$
PCA
PCA
PCA
Principal Components senone$_1$
Principal Components senone$_K$
Eigenposteriors
(b)
Teacher DNN Posterior
Senone-specific Eigenposteriors
Posterior in Low-dimensional PC Space
Enhanced Soft Target For Training Student DNN
Projection on few principal components
Low-rank Reconstruction
$\tilde{z}^{LR}$
$D_{LR}^{\top}\tilde{z}$
$D_{LR}D_{LR}^{\top}\tilde{z}$
$\tilde{z}$
D_{LR}
✓
✗

Figure 1: *(a) PCA is used to extract principal components (eigenposteriors) of the linear subspaces of individual senone classes. (b) Low-dimensional reconstruction of senone posterior probabilities is done to obtain more accurate soft targets for training improved student DNN acoustic models.*

2.1. Principal Component Analysis of Senone Subspaces

Let $z_t = [p(s_1|x_t) \dots p(s_k|x_t) \dots p(s_K|x_t)]^{\top}$ denote a DNN posterior vector for the acoustic feature x_t at time t, where we have K senone classes $\{s_k\}_{k=1}^{K}$. We collect N senone posteriors for each senone class in senone-specific matrices using labels from the GMM-HMM forced alignment. After mean-centering these matrices in *logarithmic* domain, principal components for each matrix are obtained (Fig. 1(a))) via eigenvector decomposition [21]. Due to skewed distribution of the posterior vectors, the logarithm of posteriors fits better the Gaussian assumption of PCA.

For each senone class, eigenvectors corresponding to the large eigenvalues characterize the frequent regularities in the subspace, whereas others represent the high-dimensional unstructured noise. If P denotes the set of principal components for a particular senone class, then the low-rank projection matrix is defined as

$$D_{\text{LR}} = P_l \in \mathcal{R}^{K \times l} \tag{1}$$

where P_l is truncation of P that keeps only the first l eigenvectors and discards the erroneous variability captured by other $K - l$ components. We select l such that relatively $\sigma\%$ variability is preserved in the PCA reconstruction of posteriors. The eigenvectors stored in the low-rank projection P_l are referred to as *"eigenposteriors"* of the senone subspace.

2.2. Enhancing DNN Posteriors using Eigenposteriors

Enhancing a DNN posterior relies on low-rank projection. The procedure is depicted in Figure 1(b) and involves the following steps:

Step 1: Identify the correct senone class of the posterior frame and pick the corresponding eigenposteriors.

Step 2: Projection on the underlying low-rank subspace using eigenposteriors to remove high-dimensional noise.

The low-rank reconstruction of the mean-centered log posterior $\tilde{z}_t$, denoted by $\tilde{z}_t^{\text{LR}}$ can be expressed as

$$\tilde{z}_t^{\text{LR}} = D_{\text{LR}} {D_{\text{LR}}}^{\top} \tilde{z}_t \tag{2}$$

Finally, the class mean is added to $\tilde{z}_t^{\text{LR}}$ and we take its exponent to obtain a low-rank senone posterior z_t^{LR} for the acoustic frame x_t.

To control the low-rank dimension of the space, we consider $\sigma\%$ variability in reconstruction of the posterior matrix. Tuning σ associates variable dimensions to the senone-specific subspaces. A lower σ indicates a higher level structural constraints in the form of low-rank regularities in the senone subspaces. We assume that σ is independent of the senone class.

2.3. Training a Student DNN using Soft Targets

The low-rank projection procedure described above requires ground-truth senone alignments. To enhance the (test) posteriors where the alignment is missing, we train a student DNN to estimate the posteriors on a globally characterized low-dimensional space. To that end, the training data posteriors are first enhanced using the supervised procedure of selecting the correct senone class. The enhanced posteriors are used as the reliable soft targets for training a student DNN to obtain low-rank posteriors.

3. Semi-supervised Training

In this section, we study integration of low-rank posterior estimation and semi-supervised training using sequence discrimination objective.

3.1. Reliable Soft targets for Untranscribed Data

If the ground truth senone alignment is missing, ASR decoding may be considered to obtain the alignments for supervised enhancement of the posteriors using senone-specific eigenposteriors (c.f. Section 2.2). In our experiments, we found that decoded alignments are too noisy for this task. In fact, inaccuracies in the decoded alignment further degrade the quality of posteriors instead of enhancing them.

To tackle this problem, we use a student DNN trained on the enhanced training data posteriors to generate the soft targets from untranscribed data. We perform the DNN forward-pass to obtain the posteriors from untranscribed data. These posteriors are used as soft-targets for augmenting our initial set of training data soft-targets obtained from supervised enhancement. We expect that an even better student DNN acoustic model can be achieved using this semi-supervised augmented training dataset. Experimental results presented in Section 4 demonstrate that the enhanced soft targets are indeed crucial to enable improvement using semi-supervised training.

3.2. Sequence Discriminative Training

We employ the sMBR objective for sequence discrimination which directly optimizes the DNN parameters to minimize the Bayes risk in state-level alignment [22].

Sequence discriminative training diminishes the inaccuracies in acoustic models due to contextual dependencies. On the other hand, the procedure of eigenposterior estimation relies on characterization of the global dependencies. Therefore, investigating a framework of combining both techniques requires further investigations. We study the following approaches:

Approach 1: We use the sequence discriminatively trained DNN as our baseline teacher acoustic model. We generate posteriors for training data using sMBR based teacher DNN and then use these posteriors to learn eigenposteriors for individual senone classes. Training data posteriors are then enhanced with supervision to obtain soft targets for training student DNN models. The task requires that the student model not only learns to generate enhanced low-rank posteriors, but also captures the sequence discrimination of the teacher DNN. Learning of eigenposteriors and posterior enhancement is done using forced alignment from the sMBR based teacher DNN in this case instead of GMM-HMM based alignments.

Approach 2: The sequence discriminative training is disentangled from eigenposterior estimation. The student DNN is trained using the enhanced soft targets obtained from low-rank projection of the teacher DNN posteriors that is trained based on cross-entropy objective. The student DNN is used to obtain the soft targets for semi-supervised training. The sequence discrimination in terms of sMBR is then applied to the student model for semi-supervised training. We evaluate and analyze the performance of the above two approaches in the next Section 4.

4. Experimental Analysis

In this section, we compare ASR performance using different systems exploiting eigenposterior based semi-supervised DNN training with sequence discrimination.

4.1. Database and Speech Features

ASR experiments are performed on AMI corpus [20] with the data recorded through individual headset microphones (IHM). AMI corpus contains recordings of spontaneous conversations in meeting scenarios, with 67 hours of *train* set, 9 hours of development, (*dev*) set, and 7 hours *test* set. 10% of training data is used for cross-validation during DNN training in all cases, whereas *dev* set is used for tuning the σ parameter. For experiments using untranscribed additional training data for semi-supervised training, we use ICSI meeting corpus [23] and Librispeech corpus [24]. In all experiments, ASR performance is evaluated on AMI test set.

Kaldi toolkit [25] is used for training DNN-HMM systems. All DNNs have 9 frames of temporal context at acoustic input and 4 hidden layers with 1200 neurons each. Our experiments consist of experiments on two different systems based on the source of initial ground truth alignments. First system is based on kaldi *tri2* scripts where the senone set and the subsequent GMM-HMM forced alignment are learned on MFCC+Δ+$\Delta\Delta$ features. Second system is based on kaldi *tri3b* scripts where the senone set and forced alignment are learned after LDA+MLLT+SAT transforms on MFCC+Δ+$\Delta\Delta$ features [26]. We call these setups *kaldi-a* and *kaldi-b* respectively hereafter. All DNNs are randomly initialized and trained using cross-entropy loss backpropagation except for sequence discriminative training experiments where 2 iterations are performed to minimize the sMBR objective.

While *kaldi-a* and *kaldi-b* setups provide different senone sets (4007 and 5000 senones respectively) and different forced alignments, the input features for training DNNs in both setups are kept the same. Input dimension for all DNNs is 351(39 dimensional MFCC+Δ+$\Delta\Delta$ features × 9 frame context) and output dimension is 4007 in *kaldi-a* and 5000 in *kaldi-b*. AMI pronunciation dictionary has ∼23K words and a bigram model for decoding.

4.2. Learning Eigenposteriors and Enhancing Posteriors

We collect at most $N = 10^4$ posterior frames for each senone class to learn the principal components for each class individually. Similar to experiments in [3], we encounter memory issues while storing soft targets for AMI corpus for learning eigenposteriors as well as during training of student networks. Hence, we preserve precision upto first two decimal places in soft targets, followed by normalizing the posterior frames to sum to 1. Eigenpoterior learning and low-rank reconstruction is still done on full soft-targets and rounding-off is done only when storing soft targets on disk. σ parameter which denotes the percentage of variability preserved was tuned on AMI dev set. Experiments on *dev* set shows that $\sigma = 90\%$ yields the best results in case of *kaldi-a* senones and $\sigma = 95\%$ is the optimal value for *kaldi-b* senones.

4.3. Eigenposterior based Semi-supervised Training

Table 1 lists the results for the experiments which utilise eigenposterior based enhancements to exploit untranscribed data. The DNN is trained using cross-entropy objective. The baseline System-1.0 corresponds to the hard target based DNNs for both *kaldi-a* and *kaldi-b* setups. *kaldi-b* baseline DNN with a WER of 30.9% is superior to *kaldi-a* baseline at 32.4% WER due to a superior set of senones and forced alignment coming from the LDA+MLLT+SAT based transforms on MFCC features.

Supervised PCA based reconstruction (exploiting the ground-truth label as Section 2.2) is done for posteriors from System-1.0 (teacher) to obtain enhanced soft targets to train System-1.1 (student) where SE denotes supervised enhancement. In both *kaldi-a* and *kaldi-b* setups, the student DNN in System-1.1 outperforms the corresponding teacher. As compared to [3], we achieved improved results for *kaldi-a* system by better tuning of the σ parameter. The scale of improvement is smaller in case of the stronger baseline system of *kaldi-b*. A noteworthy observation is that *kaldi-a* posteriors require 10% of variability to be discarded (i.e. $\sigma = 90\%$) whereas in *kaldi-b* posteriors, we needed to remove only 5% variability. This difference suggests that the weaker DNN acoustic model in *kaldi-a* setup results in more inaccuracies in estimating senone probabilities as compared to the *kaldi-b* setup. Thus, a lower amount of variability (i.e. using a lower value of σ) has to be preserved for PCA based reconstruction of *kaldi-a* posteriors to remove the high dimensional noise.

Next, we build System-1.2 by semi-supervised training using a joint set of supervised PCA enhanced soft targets for AMI train set and forward-passed posterior features for ICSI database from System-1.1. Again, *kaldi-a* setup improves significantly whereas we observe negligible improvements in *kaldi-b* setup. System-1.3 is built similar to System-1.2 with semi-supervised training using System-1.2 as the teacher model to generate soft targets for both ICSI and Librispeech(LIB) database. In both System-1.2 and 1.3, AMI train set soft targets are obtained from

Table 1: *Performance of ASR systems (in WER%) using eigenposterior based enhancements and semi-supervised training. System 0 uses hard-target based baseline DNN. In parenthesis, SE-1.0 denotes supervised enhancement of DNN posteriors from System-1.0 and FP-n shows forward-pass using system n to obtain the soft targets.*

Sys#	Training Data	*kaldi-a*	*kaldi-b*
1.0	AMI (Baseline)	32.4	30.9
1.1	AMI(SE-1.0)	31.6	30.2
1.2	ICSI(FP-1.1)+AMI(SE-1.0)	30.8	30.2
1.3	LIB(FP-1.2)+ICSI(FP-1.2)+AMI(SE-1.0)	30.7	30.2

supervised enhancement of DNN posteriors from System-1.0. In both *kaldi-a* and *kaldi-b* setups, System-1.3 student DNN performs better or no worse than the first student DNN learned in System-1.1. A possible reason for *kaldi-b* setup not being able to utilize information from additional untranscribed data is that the *kaldi-b* senone set is learned using spectral features modified by speaker adaptive feature transforms. While *kaldi-a* setup might be trying to learn speaker invariability using additional data from ICSI and Librispeech, *kaldi-b* setup has speaker invariance already encoded in its superior quality senone set. In addition, it may be noted that no gains from semi-supervised training are obtained when the System-1.1 is trained with non-enhanced soft targets from System-1.0 (similar to the the observation in [3]).

4.4. Integrating Sequence Discriminative Training

As discussed in Section 3.2, we evaluate two approaches to exploit untranscribed data in case of sequence discriminative training of DNNs. Table 2 provides results for Approach 1 when a sMBR trained network is used as a teacher DNN for generating soft targets.

System-2.0 is the sMBR objective based sequence discriminatively trained DNN baseline for *kaldi-a* and *kaldi-b* setups. We found that System-2.1, which is built in a fashion exactly similar to System-1.1 in Section 4.3, is unable to bring any improvements over System-2.0 using low-rank enhancement of DNN posteriors. Instead, the ASR performance degrades in both the cases significantly. When soft-targets are used without any PCA based reconstruction to build Systems-2.2, 2.3 and 2.4, we observe minor or no improvements at all. But the performance still doesn't degrade as it happens in the case of System-2.1. We conclude from this observation that although the soft-targets based training of student DNNs has potential in case of sequence discriminatively trained teacher DNNs (also confirmed in [15]), it is not possible to improve the student DNNs by eigenposteriors based low-rank enhancements. Sequence discriminative training essentially modifies the senone subspaces and underlying senone correlations in such a way that eigenposteriors are no longer capable of capturing them with linear PCA transformations.

Next, we evaluate Approach 2 using eigenposteriors based semi-supervised training prior to applying sequence discrimination. Table 3 provides the results for these experiments. First two columns simply represent the performance gains brought in by sequence discriminative training in the baseline System-1.0 from Table 1. We see an absolute reduction of 2.8% and 2.7% in WER for *kaldi-a* and *kaldi-b* setup respectively using sMBR sequence training on System-1.0. When we apply sMBR based sequence training on the best performing semi-supervised student models from System-1.3 in Table 1, we observe significant performance gains with 2.6% and 3.0% absolute WER reductions for *kaldi-a* and *kaldi-b* setups respectively. Compared to the baseline DNN (System-1.0), we achieve nearly an overall 4% absolute WER reduction for both the setups.

Table 2: *Performance of ASR systems (in WER%) using eigenposterior based enhancements as* ***Approach 1*** *described in Section 3.2. Baseline System 2.0 is trained on hard targets followed by sMBR based sequence discrimination. Other notations in parenthesis are similar as in Table 1.*

Sys#	Training Data	*kaldi-a*	*kaldi-b*
2.0	AMI (Baseline)	29.6	28.2
2.1	AMI(SE-2.0)	30.7	28.5
2.2	AMI(FP-2.0)	29.4	28.2
2.3	ICSI(FP-2.0)+AMI(FP-0)	29.6	28.2
2.4	LIB(FP-2.0)+ICSI(FP-2.0)+AMI(FP-2.0)	29.6	28.2

Table 3: *Comparison of ASR performance of baselines and the best semi-supervised DNNs from Table 1 when trained with sMBR sequence discriminative objective. The third raw corresponds to the results of* ***Approach 2*** *described in Section 3.2*

Sys (from Table 1)#	*kaldi-a*	*kaldi-b*
1.0	32.4	30.9
1.0+sMBR (2.0)	29.6	28.2
1.3	30.7	30.2
1.3+sMBR	**28.1**	**27.2**

Thus, this experiment demonstrates that Approach 2 is the suitable strategy for complimentary integration of both eigenposterior based enhanced acoustic modelling and sequence discrimination to improve ASR performance. First, we exploit the improvements from low-rank enhancement of DNN posteriors and semi-supervised training under the student-teacher framework. Then, we boost the performance using sMBR based sequence discriminative training. It may be noted that untranscribed data is only affecting the performance when eigenposteriors are exploited for training the student DNN. Eigenposterior estimation requires DNN to learn the regularities of contextual dependencies, thus the cross entropy objective is more suitable for DNN training.

5. Concluding Remarks

DNN posteriors live in low-dimensional senone-specific subspaces that can be characterized using principal component analysis. Eigenposteriors obtained through PCA enables enhancing the DNN posteriors via low-rank projection. Enhanced posteriors preserve the global structure of the senone posterior space and local inaccuracies are removed. Hence, they can be used as more reliable soft targets and training a student DNN using enhanced soft targets improves the acoustic model accuracy. Eigenposterior based enhancement is found to be crucial for exploiting untranscribed data and further improving the acoustic model performance using semi-supervised training.

The procedure of eigenposterior estimation relies on DNN learning the contextual dependencies as patterns in senone posterior space. sMBR based sequence discriminative training leads to significant reduction in WER when used as a cascade after eigenposterior based enhancement. We conclude that the performance gains from low-rank enhancement and sequence discrimination have different sources and they can be combined in a complementary way to improve ASR.

6. Acknowledgments

The research leading to these results has received funding from SNSF project PHASER-QUAD grant agreement number 200020-169398.

7. References

[1] G. Luyet, P. Dighe, A. Asaei, and H. Bourlard, “Low-rank representation of nearest neighbor phone posterior probabilities to enhance dnn acoustic modeling,” in *Interspeech*, 2016.

[2] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, “Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition,” in *IEEE ICASSP*, 2016.

[3] P. Dighe, A. Asaei, and H. Bourlard, “Low-rank and sparse soft targets to learn better dnn acoustic models,” in *ICASSP 2017*, 2017.

[4] P. Dighe, A. Asaei, and H. Bourlard, “Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition,” in *Speech Communication*. Elsevier, 2015.

[5] L. Deng, “Switching dynamic system models for speech articulation and acoustics,” in *Mathematical Foundations of Speech and Language Processing*. Springer New York, 2004, pp. 115–133.

[6] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition,” *The Journal of the Acoustical Society of America*, 2007.

[7] J. Xue, J. Li, and Y. Gong, “Restructuring of deep neural network acoustic models with singular value decomposition.” in *INTERSPEECH*, 2013.

[8] T. N. 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 *IEEE ICASSP*, 2013.

[9] V. S. Tomar and R. C. Rose, “Manifold regularized deep neural networks.” 2014.

[10] H. Chung, J. J. Kang, K. Y. Park, S. J. Lee, and J. G. Park, “Deep neural network based acoustic model parameter reduction using manifold regularized low rank matrix factorization,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 659–664.

[11] R. Z. J.-T. H. Y. G. Jinyu Li, “Learning Small-Size DNN with Output-Distribution-Based Criteria,” in *Interspeech*, 2014.

[12] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[13] W. Chan, N. R. Ke, and I. Lane, “Transferring knowledge from a rnn to a dnn,” in *Interspeech*, 2015.

[14] R. Price, K.-i. Iso, and K. Shinoda, “Wise teachers train better dnn acoustic models,” *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, pp. 1–19, 2016.

[15] J. H. Wong and M. J. Gales, “Sequence student-teacher training of deep neural networks,” in *Interspeech 2016*, 2016, pp. 2761–2765. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-911

[16] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, “Deep neural network features and semi-supervised training for low resource speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6704–6708.

[17] P. Zhang, Y. Liu, and T. Hain, “Semi-supervised dnn training in meeting recognition,” in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 141–146.

[18] Y. Huang, D. Yu, Y. Gong, and C. Liu, “Semi-supervised gmm and dnn acoustic model training with multi-system combination and confidence re-calibration,” in *Interspeech 2013*, August 2013.

[19] S. Li, X. Lu, S. Sakai, M. Mimura, and T. Kawahara, “Semi-supervised ensemble dnn acoustic model traning,” in *IEEE ICASSP*, 2017.

[20] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, “The ami meeting corpus,” in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[21] J. Shlens, “A tutorial on principal component analysis,” *arXiv preprint arXiv:1404.1100*, 2014.

[22] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks.” in *INTERSPEECH*, 2013.

[23] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, “The icsi meeting corpus,” in *IEEE ICASSP*, 2003.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *IEEE ICASSP*, 2015.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” 2011.

[26] S. P. Rath, D. Povey, and K. Veselỳ, “Improved feature processing for deep neural networks.” in *Interspeech*, 2013.