LOW-RANK AND SPARSE SOFT TARGETS TO LEARN BETTER DNN ACOUSTIC MODELS

Pranay Dighe[⋆◦] *Afsaneh Asaei*[⋆] *Hervé Bourlard*[⋆◦]

[⋆]Idiap Research Institute, Martigny, Switzerland
[◦]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Conventional deep neural networks (DNN) for speech acoustic modeling rely on Gaussian mixture models (GMM) and hidden Markov model (HMM) to obtain binary class labels as the targets for DNN training. Subword classes in speech recognition systems correspond to context-dependent tied states or senones. The present work addresses some limitations of GMM-HMM senone alignments for DNN training. We hypothesize that the senone probabilities obtained from a DNN trained with binary labels can provide more accurate targets to learn better acoustic models. However, DNN outputs bear inaccuracies which are exhibited as high dimensional unstructured noise, whereas the informative components are structured and low-dimensional. We exploit principal component analysis (PCA) and sparse coding to characterize the senone subspaces. Enhanced probabilities obtained from low-rank and sparse reconstructions are used as soft-targets for DNN acoustic modeling, that also enables training with untranscribed data. Experiments conducted on AMI corpus shows 4.6% relative reduction in word error rate.

Index Terms— Soft targets, Principle component analysis, Sparse coding, Automatic speech recognition, Untranscribed data.

1. INTRODUCTION

DNN based acoustic models have been state-of-the-art for automatic speech recognition over the past few years [1]. While DNN input consists of multiple frames of acoustic features, the target output is obtained from a frame level GMM-HMM forced alignment corresponding to the context dependent tied triphone states or senones [2]. This procedure results in inefficiency in DNN acoustic modeling [3, 4]. Unlike the conventional practice, the present work argues that the optimal DNN targets are probability distributions rather than Kronecker deltas (hard targets). Earlier studies on optimal training of a neural network for HMM decoding provides rigorous theoretical analysis that supports this idea [5]. Here, we propose a DNN based data driven framework to obtain accurate probability distributions (soft targets) for improved DNN acoustic modeling. The proposed approach relies on modeling of low-dimensional senone subspaces in DNN posterior probabilities.

Speech production is known as the result of activations of a few highly constrained articulatory mechanisms leading to generation of linguistic units (e.g. phones, senones) on low-dimensional non-linear manifolds [6, 7]. In the context of DNN acoustic modeling, low-dimensional structures are exhibited in the space of DNN senone posteriors [8]. Low-rank and sparse representations are found promising to characterize senone-specific subspaces [9, 10]. The senone-specific structures are superimposed with high-dimensional unstructured noise. Hence, projection of DNN posteriors on their underlying low-dimensional subspaces enhances the DNN posterior accuracies. In this work, we propose a new application of *enhanced* DNN posteriors to generate accurate soft targets for DNN acoustic modeling.

Earlier works on exploiting low-dimensionality in DNN acoustic modeling focus on exploiting low-rank and sparse representations to modify DNN architectures for small footprint implementation. In [11, 12] low-rank decomposition of the neural network's weight matrices enables reduction in DNN complexity and memory footprint. Similar goals have been achieved by exploiting sparse connections [13] and sparse activations [14] in hidden layers of DNN. In another line of research, soft targets based DNN training has been found effective for enabling model compression [15, 16] and knowledge transfer from an accurate complex model to a smaller network [17, 18]. This approach relies on soft targets providing more information for DNN training than the binary hard alignments.

We propose to bring together the advantage of higher information content of soft targets with the accurate model of senone space provided by low-rank and sparse representations to train superior DNN acoustic models. Soft targets enable characterization of the senone-specific subspaces by quantifying the correlations between senone classes as well as sequential dependencies (details in Section 2.1). This information is manifested in the form of structures visible among a large population of training data posterior probabilities. Potential of these posteriors to be used as soft targets for DNN training is reduced due to presence of unstructured noise. Therefore, to obtain reliable soft targets, we perform low-rank and sparse reconstruction of training data posteriors to preserve the global low-dimensional structures while discarding the random high-dimensional noise. The new DNNs trained with low-rank or sparse soft targets are capable of estimating the test posteriors on a low-dimensional space which results in better ASR performance. We consider PCA (Section 2.2) and dictionary based sparse coding (Section 2.3) for generating low-rank and sparse representations respectively. Strength of PCA lies in capturing the linear regularities in the data [19] whereas an over-complete dictionary used for sparse coding learns to model the non-linear space as a union of low-dimensional subspaces. Dictionary based sparse reconstruction also reduces the rank of the senone posterior space [9].

Experimental evaluations are conducted on AMI corpus [20], a collection of recordings of multi-party meetings for large vocabulary speech recognition. We show in Section 3 that low-rank and sparse soft targets lead to training of better DNN acoustic models. Reductions in word error rate (WER) are observed over the baseline hybrid DNN-HMM system without the need of explicit sparse coding or low-rank reconstruction of test data posteriors. Moreover, they enable effective use of out-of-domain untranscribed data by augmenting AMI training data in a knowledge transfer fashion. DNNs trained with low-rank and sparse soft targets yield upto 4.6% relative improvement in WER, whereas a DNN trained with non-enhanced soft targets fails to exploit any further knowledge provided by the untranscribed data. To the best of our knowledge, significant benefit of DNN generated soft targets for training a more accurate DNN

Fig. 1: Correlation among senones due to: (a) long input context and (b) acoustically similar root in decision trees. In (c), we show examples of DNN posterior probabilities for a particular senone class (in blue barplots) which highlight low-dimensional patterns (green boxes) super-imposed with unstructured noise. PCA and sparse coding enable recovery of the underlying patterns by discarding the unstructured noise, and provide more reliable soft targets for DNN training. K denotes the size of DNN outputs which is equal to total number of senones.

acoustic model has not been shown in the prior work.

In the rest of the paper, the proposed approach is described in Section 2. Experimental analysis is carried out in Section 3. Section 4 presents the concluding remarks and directions for future work.

2. LOW-RANK AND SPARSE SOFT TARGETS

This section describes the novel approach towards reliable soft target estimation. We study reasons for regularities among senone posteriors and investigate two systematic approaches to obtain more accurate probabilities as soft targets for DNN acoustic modeling.

2.1. Towards Better Targets for DNN Training

Earlier works on distillation of the DNN knowledge show the potential of soft targets for model compression and the sub-optimal nature of the hard alignments [15, 21]. Although hard targets assign a particular senone label to a relatively long sequence of ($\sim$10 or more) acoustic frames, senone durations are usually shorter. A long context of input frames may lead to presence of acoustic features corresponding to multiple senones in the input (Fig. 1(a)), so the assumption of binary outputs renders inaccurate.

In contrast, soft outputs quantify such sequential information using non-zero probabilities for multiple senone classes. Contextual senone dependencies arising in soft targets can be attributed to the ambiguities due to phonetic transitions [21]. Furthermore, the procedure of senone extraction leads to acoustic correlations among multiple classes corresponding to the same phone-HMM states [2], as they all share the same root in the decision tree (Fig. 1(b)).

These dependencies can be characterized by analyzing a large number of senone probabilities from the training data. The frequent dependencies are exhibited as regularities among the correlated dimensions in senone posteriors. As a result, a matrix formed by concatenation of class-specific senone posteriors has a low-rank structure. In other words, class-specific senones lie in low-dimensional subspaces with a dimension higher than unity [9], that violates the principal assumption of binary hard targets.

In practice, inaccuracies in DNN training lead to the presence of unstructured high-dimensional errors (Fig. 1(c)). Therefore, the initial senone posterior probabilities obtained from the forward pass of a DNN trained with hard alignments are not accurate in quantifying the senone dependency structures. Our previous work demonstrates that the erroneous estimations can be separated using low-rank or sparse representations [10, 9]. In the present study, we consider application of PCA and sparse coding to obtain more reliable soft targets for DNN acoustic model training.

2.2. Low-rank Reconstruction Using Eigenposteriors

Let $z_t = [p(s_1|x_t) \dots p(s_k|x_t) \dots p(s_K|x_t)]^\top$ denote a forward pass estimate of the posterior probabilities of K senone classes $\{s_k\}_{k=1}^K$, given the acoustic feature x_t at time t. DNN is trained using the initial labels obtained from GMM-HMM forced alignment. We collect N senone posteriors which are labeled as class s_k in GMM-HMM forced alignment and mean-center them in the *logarithmic* domain as follows:

$$\tilde{z}_t = \ln(z_t) - \mu_{s_k} \tag{1}$$

where μ_{s_k} is mean of the collected posteriors in log-domain. Due to skewed distribution of the posterior vectors, the logarithm of posteriors fits better the Gaussian assumption of PCA. We concatenate the N senone s_k posterior vectors after operation shown in (1) to form a matrix $M_{s_k} \in \mathcal{R}^{K \times N}$. For the sake of brevity, the subscript s_k is dropped in the subsequent expressions. However, all the calculations are performed for each of the senone classes individually.

Principal components of the senone space are obtained via eigenvector decomposition [22] of covariance matrix of M. The covariance matrix is obtained as $C = \frac{1}{N-1} M M^\top$. We factorize the covariance matrix as $C = PSP^\top$ where $P \in \mathcal{R}^{K \times K}$ identifies the eigenvectors and S is a diagonal matrix containing the sorted eigenvalues. Eigenvectors in P which correspond to the large eigenvalues in S constitute the frequent regularities in the subspace, whereas others carry the high-dimensional unstructured noise. Hence, the low-rank projection matrix is defined as

$$D_{\text{LR}} = P_l \in \mathcal{R}^{K \times l} \tag{2}$$

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 low-rank reconstruction of original senone matrix M.

The eigenvectors stored in the low-rank projection P_l are referred to as *"eigenposteriors"* of the senone space (in the same spirit as *eigenfaces* are defined for low-dimensional modeling of human faces [23]).

Low-rank reconstruction of a mean-centered log posterior $\tilde{z}_t$, denoted by $\tilde{z}_t^{\text{LR}}$ is estimated as

$$\tilde{z}_t^{\text{LR}} = D_{\text{LR}} {D_{\text{LR}}}^\top \tilde{z}_t \tag{3}$$

Finally, we add the mean μ_{s_k} to $\tilde{z}_t^{\text{LR}}$ and take its exponent to obtain a low-rank senone posterior z_t^{LR} for the acoustic frame x_t. Low-rank posteriors obtained for the training data are used as soft targets for learning better DNNs (Fig.2). We assume that $\sigma\%$ variability, that quantifies the low-rank regularities in senone spaces, is a parameter independent of the senone class.

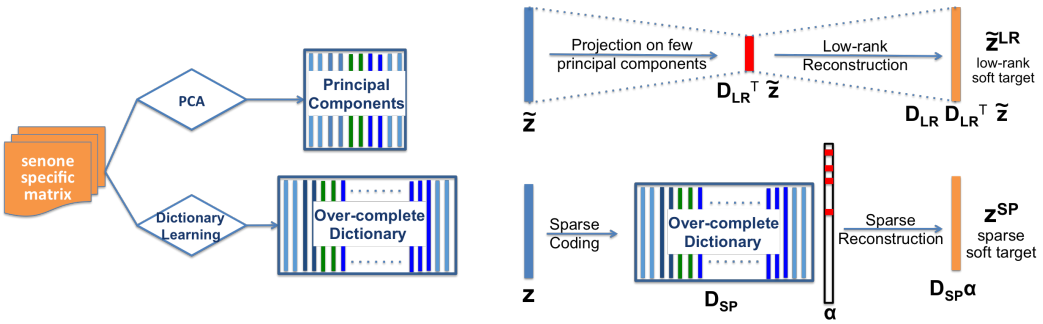

Fig. 2: Low-dimensional reconstruction of senone posterior probabilities to achieve more accurate soft targets for DNN acoustic model training: PCA is used to extract principal components of the linear subspaces of individual senone classes. Sparse reconstruction over a dictionary of senone space representatives is used for non-linear recovery of low-dimensional structures.

2.3. Sparse Reconstruction Using Dictionary Learning

Unlike PCA, over-complete dictionary learning and sparse coding enables modeling of non-linear low-dimensional manifolds. Sparse modelling assumes that senone posteriors can be generated as sparse linear combination of senone space representatives, collected in a dictionary D_{SP}. We use online dictionary learning algorithm [24] to learn an over-complete dictionary for senone s_k using a collection of N training data posteriors of senone s_k, such that

$$D_{\text{SP}} = \arg\min_{D,A} \sum_{t=t_1}^{t_N} \|z_t - D\,\alpha_t\|_2^2 + \lambda\|\alpha_t\|_1 \qquad (4)$$

where $A = [\alpha_{t_1} \ldots \alpha_{t_N}]$ and λ is a regularization factor. Again we have dropped the subscript s_k, but all calculations are still senone-specific. Sparse reconstruction (Fig.2) of senone posteriors is thus obtained by first estimating the sparse representation [25] as

$$\alpha_t = \arg\min_{\alpha} \|z_t - D_{\text{SP}}\,\alpha\|_2^2 + \lambda\|\alpha\|_1. \qquad (5)$$

followed by reconstruction as

$$z_t^{SP} = D_{\text{SP}}\,\alpha_t \qquad \forall t \in \{t_1, \ldots, t_N\}. \qquad (6)$$

Sparse reconstructed senone posteriors have been previously found to be more accurate acoustic models for DNN-HMM speech recognition [9]. In particular, it was shown that the rank of senone-specific matrices is much lower after sparse reconstruction. In the present work, we investigate if they could also provide more accurate soft targets for DNN training Regularization parameter λ in (4)-(5) controls the level of sparsity and the level of noise being removed after sparse reconstruction. Fig. 2 summarises the low-rank and sparse reconstruction of senone posteriors.

3. EXPERIMENTAL ANALYSIS

In this section we evaluate the effectiveness of low-rank and sparse soft targets to improve the performance of DNN-HMM speech recognition. We also investigate the importance of better DNN acoustic models to exploit information from untranscribed data.

3.1. Database and Speech Features

Experiments are conducted on AMI corpus [20] which contains recordings of spontaneous conversations in meeting scenarios. We use recordings from individual head microphones (IHM) comprising of around 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, whereas *dev* set is used for tuning regularization parameters σ and λ. For experiments using untranscribed additional training data, we use ICSI meeting corpus [26] and Librispeech corpus [27]. Data from ICSI corpus consists of meeting recordings (around 70 hours). Librispeech data is read speech from audio-books and we use a 100 hour subset of it.

Kaldi toolkit [28] 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. Input features are 39 dimensional MFCC+Δ+$\Delta\Delta$ (39×9=351 dimensional input) and output is 4007 dimensional senone probability vector. AMI pronunciation dictionary has ∼23K words and a bigram model for decoding. For dictionary learning and sprase coding, SPAMS toolbox [29] is used.

3.2. Baseline DNN-HMM using Hard and Soft Targets

Our baseline is a hybrid DNN-HMM system trained using forced aligned targets (IHM setup in [30]). WER using baseline DNN is 32.4% on AMI *test* set. Another baseline is a DNN trained using non-enhanced soft targets from the baseline. This system gives a WER of 32.0%. All soft-target based DNNs are randomly initialized and trained using cross-entropy loss backpropagation.

3.3. Generation of Low-rank and Sparse Soft Targets

We group DNN forward pass senone probabilities for the training data into class-specific senone matrices. For this, senone labels from the ground truth based GMM-HMM hard alignments are used. Each matrix is restricted to have $N = 10^4$ vectors of $K = 4007$ senone probabilities to facilitate computation of principal components and sparse dictionary learning. We found the average rank of senone matrices, defined as the number of singular values required to preserve 95% variability, to be 44. Dictionaries of size 500 columns were learned for each senone, making them nearly 10 times over-complete. The procedure as depicted in Fig. 2 is implemented to generate low-rank and sparse soft-targets.

We also encountered memory issues while storing large matrices of senone probabilities for all training and cross-validation data. It requires enormous amounts of storage space (similar to [16]). Hence, we preserve precision only upto first two decimal places in soft targets, followed by normalizing the vector to sum 1 before storing on the disk. We assume that essential information might not be in dimensions with very small probabilities. Although such thresholding can be a compromise to our approach, we did some experiments with higher precision (upto 5 decimal places), but there was no significant improvement in ASR. Both low-rank and sparse reconstruction were still computed on full soft-targets without any rounding; we perform thresholding only when storing targets on the disk.

First we tune the variability preserving low-rank reconstruction parameter σ and sparsity regularizer λ for better ASR performance in AMI *dev* set. When σ=80% of variability is preserved in the principal components space, the most accurate soft targets are achieved for DNN acoustic modeling resulting in the smallest WER. Likewise, $\lambda = 0.1$ was found the optimal value for sparse reconstruction. It may be noted that in both low-rank and sparse reconstruction, there is an optimal amount of enhancement needed for improving ASR.

Table 1: Performance of various systems (in WER%) when additional untranscribed training data is used. System 0 is hard-targets based baseline DNN. In paranthesis, SE-0 denotes supervised enhancement of DNN outputs from system 0 and FP-n shows forward pass using system n.

System #	Training Data	PCA(σ=80)	Sparsity(λ=0.1)	Non-Enhanced Soft-Targets
0	AMI (Baseline WER **32.4%**)	-	-	-
1	AMI(SE-0)	31.9	31.6	32.0
2	ICSI(FP-1) + AMI(SE-0)	31.2	31.6	32.5
3	LIB100(FP-1) + AMI(SE-0)	31.2	31.6	32.4
4	LIB100(FP-2) + AMI(SE-0)	31.0	31.8	32.4
5	LIB100(FP-2) + ICSI(FP-2) + AMI(SE-0)	**30.9**	31.7	32.4

While less enhancement leads to continued presence of noise in soft targets, too much of it results in loss of essential information.

3.4. DNN-HMM Speech Recognition

Speech recognition using DNNs trained with the new soft targets obtained from low-rank and sparse reconstruction is compared in Table 1). System-0 is the baseline hard target based DNN. System-1 is built by supervised enhancement of soft outputs obtained from system-0 on AMI training data as shown in Fig. 2. As expected, training with the soft targets yields lower WER than the baseline hard targets. We can see that both PCA and sparse reconstruction result in more accurate acoustic modeling, where sparse reconstruction achieves 0.8% absolute reduction in WER.

Sparse reconstruction is found to work better than low-rank reconstruction for ASR. It can be due to the higher accuracy of sparse model in characterizing the non-linear senone subspaces [8]. Unlike previous works [9, 10] which required two stages of DNN forward pass and explicit low-dimensional projection, a single DNN is learned here that estimates the probabilities directly on a low-dimensional space.

3.5. Training with Untranscribed Data

Given an accurate DNN acoustic model and some untranscribed input speech data, we can obtain soft targets for the new data through forward pass. Assuming that the initial model can generalize well on unseen data, the additional soft targets thus generated can be used to augment our original training data. We propose to learn better DNN acoustic models using this augmented training set. This method is reminiscent of the *knowledge transfer* approach [15, 16] which is typically used for model compression. In this work, we use the same network architecture for all experiments.

DNNs trained with low-rank and sparse soft targets are used to generate soft targets for ICSI corpus and Librispeech (LIB100) which are sources of untranscribed data. Table 1 shows interesting observations from various experiments using data augmentation. First, system-2 is built augmenting enhanced AMI training data with ICSI soft targets generated from system-1. We consider ICSI corpus, consisting of spontaneous speech from meeting recordings, as in-domain with AMI corpus. While PCA based DNN successfully exploits information from this additional ICSI data showing significant improvement from system-1 to system-2, the same is not observed using sparsity based DNN.

Next, system-3 is built by augmenting enhanced AMI data with Librispeech(LIB100) soft targets obtained from system 1. Read audio book speech data from Librispeech is out-of-domain as compared to spontaneous speech in AMI. Still, system-3 achieves similar reductions in WER as observed in system-2 which was built using in-domain ICSI data.

System 4 and 5 were built to further explore if we could extract even more information from the out-of-domain Librispeech data by using soft targets from system-2 instead of system-1. Note that system-2, trained using soft targets from both AMI and ICSI spontaneous speech data, is a more accurate model than system 1. Indeed, both system 4 and 5 perform better than previous systems using PCA based DNNs where system 5 outperforms the hard target based baseline by 1.5% absolute reduction in WER.

Surprisingly, DNN soft targets obtained from sparse reconstruction are not able to exploit the unseen data in all the systems. We speculate that dictionary learning for sparse coding captures the non-linearities specific to AMI database. These nonlinear characteristics may correspond to channel and recording conditions which vary over different databases and can not be transcended. On the other hand, the local linearity assumption of PCA leads to extraction of a highly restricted basis set that captures the most important dynamics in the senone probability space. Such regularities mainly address the acoustic dependencies among senones which are generalizable to other acoustic conditions. Hence, the eigenposteriors are invariant to the exceptional effects due to channel and recording conditions.

Sparse reconstruction is able to mitigate the undesired effects as long as they have been seen in the training data. Given the superior performance of sparse reconstruction of AMI posteriors (in system-1), we believe that sparse modeling might be more powerful if some labeled data from unseen acoustic conditions is made available for dictionary learning.

It may be noted that training with additional untranscribed data is not effective if non-enhanced soft targets are used. In fact, systems 2-5 without low-rank or sparse reconstruction, perform worse than system-1 although they have seen more training data.

4. CONCLUSIONS AND FUTURE DIRECTIONS

We presented a novel approach to improve DNN acoustic model training using low-rank and sparse soft targets. PCA and sparse coding were employed to identify senone subspaces, and enhance senone probabilities through low-dimensional reconstruction. Low-rank reconstruction using PCA relies on the existance of eigenposteriors capturing the local dynamics of senone subspaces. Although, sparse reconstruction proves more effective to achieve reliable soft targets when transcribed data is provided, low-rank reconstruction is found generalizable to out-of-domain untranscribed data. DNN trained on low-rank reconstruction achieves 4.6% relative reduction in WER, whereas DNN trained using non-enhanced soft targets fails to exploit additional information from additional data. Eigenposteriors can be better estimated using robust PCA [31] and sparse PCA [32] for better modeling of senone subspaces. Furthermore, probabilistic PCA and maximum likelihood eigen decomposition can reduce the computational cost for large scale applications.

This study supports the use of probabilistic outputs for DNN acoustic modeling. Specifically, enhanced soft targets can be more effective in training small footprint DNNs based on model compression. In future, we plan to investigate their usage in cross-lingual knowledge transfer [33]. We will also study domain adaptation based on the notion of eigenposteriors.

5. ACKNOWLEDGMENTS

Research leading to these results has received funding from SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507.

6. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "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] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994.

[3] N. Jaitly, V. Vanhoucke, and G. Hinton, "Autoregressive product of multi-frame predictions can improve the accuracy of hybrid models," 2014.

[4] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "Gmm-free dnn acoustic model training," in *IEEE ICASSP*, 2014.

[5] H. Bourlard, Y. Konig, and N. Morgan, *REMAP: Recursive Estimation and Maximization of a Posteriori Probabilities; Application to Transition-based Connectionist Speech Recognition*. ICSI Technical Report TR-94-064, 1994.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] D. Yu, F. Seide, G. Li, and L. Deng, "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," in *IEEE ICASSP*, 2012.

[14] J. Kang, C. Lu, M. Cai, W.-Q. Zhang, and J. Liu, "Neuron sparseness versus connection sparseness in deep neural network for large vocabulary speech recognition," in *ICASSP*, April 2015, pp. 4954–4958.

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

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

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

[18] 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.

[19] B. Hutchinson, M. Ostendorf, and M. Fazel, "A sparse plus low-rank exponential language model for limited resource scenarios," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 494–504, 2015.

[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] D. Gillick, L. Gillick, and S. Wegmann, "Don't multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[22] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[23] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, pp. 519–524, 1987.

[24] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[26] 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.

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

[28] 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.

[29] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *arXiv preprint arXiv:1411.3230*, 2014.

[30] I. Himawan, P. Motlicek, D. Imseng, B. Potard, N. Kim, and J. Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *IEEE ICASSP*, 2015, pp. 4540–4544.

[31] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 99, pp. 1–1, 2013.

[32] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[33] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr," in *IEEE Spoken Language Technology Workshop (SLT)*, 2012.