FAR-FIELD ASR USING LOW-RANK AND SPARSE SOFT TARGETS FROM PARALLEL DATA

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

ABSTRACT

Far-field automatic speech recognition (ASR) of conversational speech is often considered to be a very challenging task due to the poor quality of alignments available for training the DNN acoustic models. A common way to alleviate this problem is to use clean alignments obtained from parallelly recorded close-talk speech data. In this work, we advance the parallel data approach by obtaining enhanced low-rank and sparse soft targets from a close-talk ASR system and using them for training more accurate far-field acoustic models. Specifically, we (i) exploit *eigenposteriors* and *Compressive Sensing* dictionaries to learn low-dimensional senone subspaces in DNN posterior space, and (ii) enhance close-talk DNN posteriors to achieve high quality soft targets for training far-field DNN acoustic models. We show that the enhanced soft targets encode the structural and temporal interrelationships among senone classes which are easily accessible in the DNN posterior space of close-talk speech but not in its noisy far-field counterpart. We exploit enhanced soft targets to improve the mapping of far-field acoustics to close-talk senone classes. The experiments are performed on AMI meeting corpus where our approach improves DNN based acoustic modeling by 4.4% absolute ($\sim$8% rel.) reduction in WER as compared to a system which doesn't use parallel data. Finally, the approach is also validated on state-of-the-art recurrent and time delay neural network architectures.

Index Terms— far-field ASR, soft targets, low-rank, sparsity, deep neural networks

1. INTRODUCTION

Training accurate DNN acoustic models using far-field speech is difficult not only due to presence of reverberation and additive noise in the acoustic inputs, but also due to the poor quality of framewise senone alignments available with it. For example, a distant microphone might pick up strong background speech or other additive noise and align spoken words in the transcription with these unintended regions [1]. These effects degrade the quality of target senone alignments which in turn results in poor DNN based acoustic modeling. A common way to tackle this problem is to parallelly record speech data using close-talk microphones and use close-talk speech to generate better quality senone alignments for training ASR models [2] (as shown in Figure 1). DNN acoustic models trained with alignments from parallel close-talk speech have been consistently shown to outperform models which use alignments from the far-field data [2, 3]. In this work, we extend the approach of exploiting parallel data for far-field ASR by exploring the use of low-rank and sparse soft targets for training far-field acoustic models.

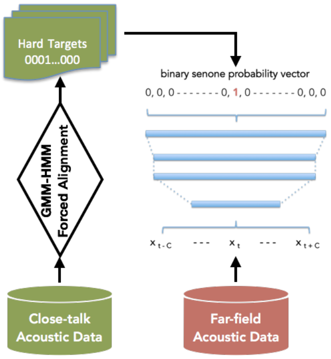

Fig. 1: Using hard alignments from parallel close-talk speech data to train DNN acoustic models for far-field speech.

Soft targets have been previously used for DNN knowledge distillation by model compression [4, 5] and knowledge transfer [6, 7]. In current work, they refer to the senone posteriors probabilities generated by an already trained DNN model on close-talk speech data. These posterior probabilities are used as targets for training a DNN which takes far-field speech features as input. As argued in our previous work [8], soft targets have high information content about underlying senone classes, but they are also prone to local unstructured and high-dimensional errors. On the other hand, the information content is manifested in the form of structured patterns visible among a large population of training data posterior probabilities. In [9] and [10], we showed that low-rank reconstruction using *eigenposteriors* (class specific principal components in DNN posterior space) and sparse reconstruction using compressive sensing (CS) dictionaries [11] are principled ways of preserving the global low-dimensional structures in soft targets while discarding the random high-dimensional noise. Soft targets enhanced in this way were successfully employed to improve close-talk ASR performance in our previous works [8, 9]. We will revisit more details in Section 2, but in summary, low rank and sparse soft targets are high quality substitutes to replace hard senone alignments for training DNN acoustic models as they provide an improved mapping of the acoustic features to the underlying senone classes for hybrid DNN-HMM ASR.

Prior research in far-field ASR using parallel data can be categorized into front-end and back-end based approaches. In front-end approaches [2, 12, 13] , the far-field acoustic features are first enhanced by mapping them to parallel close-talk features and then these enhanced acoustic features are used to train DNN based ASR system. In contrast, the back-end approaches focus on employing stronger acoustic models like CNNs [14], LSTM-RNNs and their variants [15, 16], or adapting the back-end model by knowledge sharing [2, 17] with a parallel close-talk based acoustic model. Another common approach, as discussed earlier, is to use hard alignments from clean speech data and has been explored successfully in [1, 3, 18].

1.1. Motivation and Contribution

Our motivation for using enhanced soft targets for learning far-field DNN acoustic models is twofolds. Firstly, in a reverberated speech signal, the acoustic realization of a senone would be continuously smudged by the presence of neighboring senones. Hence, any acoustic feature frame of reverberated speech can possibly have evidence of multiple senones which would actually appear in a comparatively more discrete sequence if the speech was captured using a close-talk microphone. This suggests an increased amount of temporal correlation exhibited by senones in the acoustic feature space of far-field speech. As argued in [8], such temporal correlation among senones is better characterized by soft targets as they are obtained by processing a context of neighboring acoustic feature frames at the input of the close-talk DNN.

Secondly, as shown in [1], far-field acoustic features might lead to a choice of different pronunciations for the same word transcription. In such a case, it will be more preferable to have soft targets as DNN outputs so as to support possibilities of multiple phonetic sequences rather than hard alignments which enforce one particular pronunciation of the underlying word sequence. Finally, we need the soft targets not to associate with unstructured local noise in the far-field acoustic features. This motivated us to work with enhanced low-rank and sparse soft targets which essentially focus on the intra-class global patterns and the inter-class correlations rather than local erroneous probability estimates present in the original DNN posteiors.

Experimental evaluations are conducted on AMI corpus [19] which has a collection of multi-party meeting recordings with unconstrained conversational speech. AMI corpus provides audio recordings which were parallely recorded using close-talk and distant microphones. This provides a perfect use case for our experiments on improving far-field ASR using low-rank and sparse soft-targets from close-talk data. We show in Section 3.2 that low-rank and sparse soft targets lead to improved ASR performance using DNN, LSTM as well as TDNN acoustic models. We achieve nearly 4-5% absolute WER reduction as compared to the traditional far-field data based DNN baseline.

In the rest of the paper, the proposed approach is described in Section 2 and the experiments on far-field ASR are explained in Section 3. Section 4 presents the concluding remarks and directions for future work.

2. LOW-RANK AND SPARSE SOFT TARGETS

This section provides a brief summary of eigenposteriors and CS dictionary based approach to enhance DNN posteriors. We also discuss the existence of senone specific low-dimensional subspaces and their importance for far-field ASR.

A large vocabulary ASR system typically works with senones in the order of $10^3 - 10^4$. On the contrary, a speech utterance is composed of a union of words which in turn consist of phonetic components and subphonetic attributes. Each acoustic component is produced through activation of a few highly constrained articulatory mechanisms leading to generation of speech data in union of low-dimensional subspaces [20, 21, 22]. In terms of CS theory [11, 23], while we take measurements in a very high dimensional DNN posterior space, the actual subspace where each senone belongs is very low-dimensional. Our earlier works [8, 9, 10, **?**] on acoustic modeling explicitly took into account this multi-subspace structure of the speech data.

In [8], it was shown that given a matrix of of DNN posteriors of a particular senone class, the actual rank of the matrix is $\sim\ 1\%$ of the overall DNN posterior dimension - thus rendering a very low-dimensional senone subspace. Moreover, a senone subspace is usually never 1-dimensional but multidimensional suggesting that it has correlations with other senone classes. These correlations can arise either from (i) sequential correlations among senones which usually appear together in a context or (ii) structural correlations among senones which are context dependent variations of the same triphone HMM state as they all share the same root in senone decision tree [24]. While hard targets are unable to capture this low-dimensional senone subspace information due to their binary nature, soft targets can easily encode the sequential and the structural correlations among senones. For a reverberated speech signal, soft targets can better capture the transition of senones over neighboring acoustic feature frames as well as accomodate alternative pronunciations of far-field speech as discussed in Section 1.1. Although soft targets provide a better mapping from the input far-field speech to the output senone classes, they may still suffer from the presence of unstructured high-dimensional errors due to inaccuracies in DNN training or erroneous local estimates. Thus,

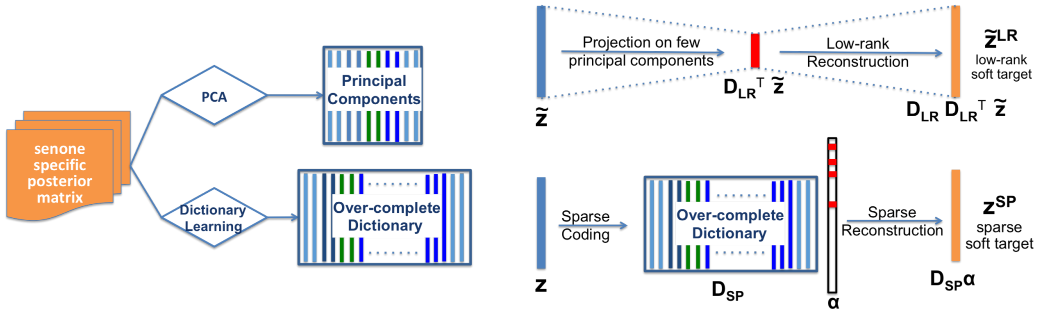

Fig. 2: Reconstruction of senone DNN posteriors to achieve more accurate soft targets for far-field DNN acoustic model training. PCA based approach (top) projects DNN posteriors on a low-dimensional intermediate representation to extract the senone subspace information whereas dictionary based sparse coding (bottom) uses l_1 norm sparsity constraints to access the the same subspace.

we propose the use of PCA and sparse coding to enhance the original soft targets so as to use them for training more accurate DNN acoustic models.

Section 2.1 and 2.2 below provide details of our enhancement processes. Figure 2 shows it visually and more details about this process can be found in [8].

2.1. PCA Based Enhancement of DNN Posteriors

Let matrix $M_k \in \mathcal{R}^{K\times N}$ consist of N training data DNN posteriors $\tilde{z}$'s which belong to the senone class s_k (K is the total number of senones i.e. the dimension of $\tilde{z}$). Here, *tilde* symbol on $\tilde{z}$ refers to the posterior z in logarithmic domain so as to avoid skewed distribution of posterior probabilities in subsequent principal component analysis. The baseline DNN acoustic model which generates these posteriors is trained using hard labels obtained from a GMM-HMM forced alignment and we segregate senone s_k posteriors by referring to the same GMM-HMM forced alignment on the training data.

We first compute the principal components of matrix M_k as $P_k \in \mathcal{R}^{K\times K}$. Next, based on the singular values, we pick the first l_k (typically $<< K$) principal components from P_k such that they preserve $\sigma\%$ variability of the space. We assume here that $\sigma\%$ variability, that quantifies the low-rank regularities in senone spaces, is a parameter independent of the senone class whereas the actual number of principal components retained, l_k, is class dependent. These principal components which contain the most important dynamics (correponding to a high value of σ) of the senone subspace are termed as the *eigenposteriors* of senone s_k, denoted as

$$D_k^{\text{LR}} = P_k^{l_k} \in \mathcal{R}^{K\times l_k} \tag{1}$$

where $P_k^{l_k}$ refers to first l_k columns of P_k. Now, we can project any training data DNN posterior $\tilde{z}$ belonging to senone s_k on the space of P_k^l and get a low-dimensional representation. This representation can be projected back to the original DNN posterior dimension as $\tilde{z}^{LR}$ such that only the low-rank senone specific information survives through this process. Note that the reconstruction results in a soft posterior and not a binary posterior. This reconstructed posterior can then be used as an enhanced soft target for training more accurate acoustic models.

2.2. Dictionary Based Enhancement of DNN Posteriors

In PCA, the principal components act as a set of orthogonal basis vectors for a given subspace such that their linear combinations can span the whole subspace. On the other hand, a dictionary as per CS theory is defined as an over-complete set of basis vectors for the subspace. *Over-completeness* here refers to having more basis vectors in the dictionary than the rank of the space itself. This property of the dictionary enables expressing any datapoint in the subspace as a sparse linear combination of the vectors already present in the dictionary. Choosing a sparse linear combination over dictionary columns enables modeling of the underlying subspace as an *union of low-dimensional manifolds* (which is non-linear) as compared to the linear subspace assumption of PCA.

Given a matrix M_k of N DNN posteriors of senone class s_k as in Section 2.2, we use online dictionary learning algorithm [25] to learn an over-complete dictionary D_k^{SP} as per the following optimization problem

$$D_k^{SP} = \arg\min_{D,A} \sum_{z\in M_k} \|z - D\,\alpha\|_2^2 + \lambda\|\alpha\|_1, \text{ s.t. } \|d_j\|_2^2 \leq 1 \forall j \tag{2}$$

where d_j is j_{th} column of D_k^{SP}, and λ in the second term is a regularization factor which controls the sparsity of α by regularizing its l_1 norm. Once the dictionary has been learned, any DNN posterior of senone class s_k can be expressed as a sparse linear combination of columns of D_k^{SP} by solving the l_1 norm based Lasso optimization problem [26] as

$$\hat{\alpha} = \arg\min_{\alpha} \|z - D_k^{\text{SP}}\,\alpha\|_2^2 + \lambda\|\alpha\|_1. \tag{3}$$

First term in equation (3) controls the accuracy of the reconstruction whereas the second term enforces sparse solutions. The reconstruction given by $z^{SP} = D_k^{SP}\ \hat{\alpha}$ is used as the enhanced sparse soft target for training acoustic models later. In the next section, we delve into experiments on far-field ASR using enhanced soft targets.

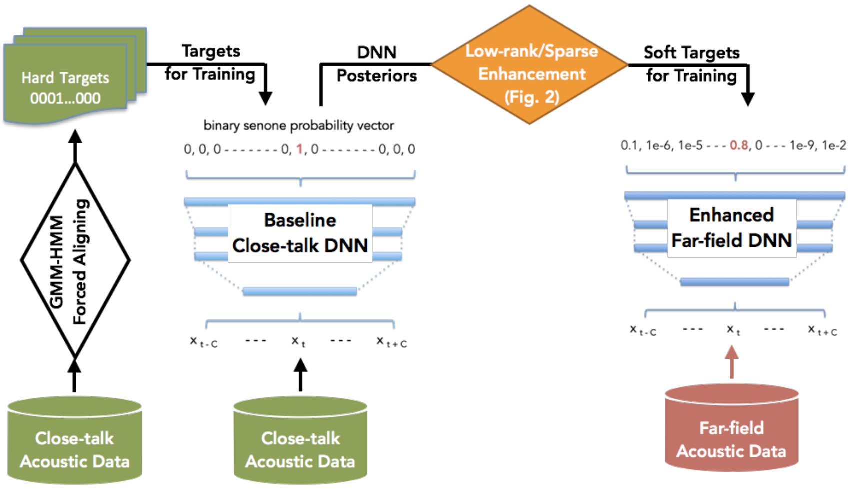

Fig. 3: Schematics of our system which uses low-rank and sparse soft targets for training the far-field DNN acoustic models. Required soft targets are obtained by PCA or dictionary based enhancement of close-talk speech DNN posteriors.

3. FAR-FIELD ASR USING PARALLEL DATA

We present our system for far-field acoustic modeling in Figure 3. Instead of using hard targets from parallelly recorded close-talk speech data, we propose to use low-rank and sparse soft-targets. For this, first a close-talk acoustic model is trained in a traditional way (shown as baseline close-talk DNN in Figure 3) with hard targets obtained from GMM-HMM forced alignments. The close-talk DNN thus trained is used to generate soft targets from the close-talk speech features. These soft targets are then passed through a PCA or dictionary based enhancement process as explained in Section 2 to generate enhanced soft targets. The enhanced soft targets are used with far-field speech to train more accurate DNN acoustic models (shown as enhanced far-field DNN in Figure 3). Below we describe the details of our ASR experiments and the subsequent analysis to evaluate the performance of our approach on far-field ASR.

3.1. Database and Experimental Specifications

We demonstrate our approach on AMI corpus [19]. Single distant microphone (SDM) data with mic-id 1 is used in our experiments for far-field speech and individual headset microphone (IHM) as the source of close-talk speech. AMI corpus has nearly 100 hours of recordings divided approximately as 80 hours train set, 10 hours *dev* and 10 hours *eval* set. 10% of training data is used for cross-validation during DNN training in all cases, whereas the *dev* set is used to tune the σ and λ parameters discussed in Section 2.

Kaldi toolkit [27] is used for training DNN-HMM systems. The input features to DNN have a dimension of 1320 (40 dimensional filterbank energies +Δ+$\Delta\Delta$ features $\times$ 11 frame context). Senone set generated using IHM data consists of 3992 senones which is the also the dimension of DNN posteriors. All DNNs have 6 hidden layers with 2048 neurons each. The experiments are based on Kaldi *tri3b* system where the senone set and the subsequent GMM-HMM forced alignment are learned after LDA+MLLT transforms [28]. All DNNs are randomly initialized and trained using cross-entropy (CE) loss backpropagation followed by sequence discriminative training to minimize the sMBR objective. For sequence training using sMBR loss, the alignments and denominator lattices are generated using the CE trained DNNs. AMI pronunciation dictionary has $\sim$47K words and a trigram model for decoding. All the results reported in this paper are reproducible using the standard AMI Corpus [19] setup, Kaldi toolkit [27] and scripts provided in [29].

For generating low-rank and sparse soft targets, a value of $\sigma = 95\%$ and $\lambda = 0.1$ was found to be optimal while optimizing WER on *dev* set. Setting $\sigma = 95\%$ results in different number of principal components being retained for different senone classes. The average number of retained principal components over all classes was found to be $\sim$40 as compared to the overall dimension of 3992 senones. This confirms the presence of low-dimensional senone subspaces underlying the DNN posterior matrices. After reconstructing DNN posteriors using PCA or dictionary based sparse coding, we preserved precision only upto first two decimal places in soft targets, followed by normalizing each vector to sum 1 before storing the data on the disk. We do the precision based thresholding to avoid memory issues as soft targets for the entire training data require large amount of storage space. The normalization of soft targets to sum to 1 ensures that they act as probability vectors under CE loss based DNN training.

Finally, the experiments based on long-short term memory (LSTM) and time-delay neural network (TDNN) (Section 3.2) are based on standard recipes and parameter settings from Kaldi *nnet3* scripts. Both LSTM and TDNN are trained using the same input fbank features and output senone labels as the baseline DNN acoustic model. Some architectural details of these models follow here. The LSTM and bidirectional(Bi-)

Table 1: ASR performance on AMI SDM eval set (in WER%) when soft targets are derived from eigenposteriors and dictionaries learned using SDM senone set and corresponding baseline DNN.

Sys #	Training Targets	Network Type	
		CE	CE+sMBR
1.1	SDM (hard)	58.6	54.4
1.2	SDM (PCA95)	57.9	53.1
1.3	SDM (SP0.1)	60.8	55.7

LSTM use a recurrence of 20 time steps for back-propagation and have 3 hidden layers each of size 1024. Splicing is done at the input to include a left and right context of 2 frames each and delta features are not appended. TDNN acoustic model is based on [30] and uses layerwise splicing of {-2,2 ; -1,2 ; -3,3 ; -7,2 ; -3,3 ; 0 ; 0}. Each ';' separated pair of numbers gives the left (with '-' symbol) and right context for splicing at each successive layer of the TDNN model. Similar to LSTM models, we do not append delta features to fbank features at the input of TDNN. Our LSTM and TDNN setup is more comparable to the previous works presented in [31] and [30]. In contrast to the system in [1], we do not employ speaker-adaptive training for acoustic modeling.

3.2. Experimental Analysis

We consider the whole SDM eval set for evaluation. Scoring is done using NIST asclite tool [32] for upto 4 overlapping speakers.

Our initial results shown in Table 1 are based completely on SDM data. ASR word error rates (WER %) are provided for DNN acoustic models which are first trained with CE loss and then subsequently sequence discriminatively trained with sMBR loss. First row depicts a traditional baseline system (Sys 1.1) trained using far-field acoustic features with hard aligments from SDM which works at 54.4% WER with sequence training. We enhance the soft targets generated from Sys 1.1 using PCA and sparse coding to train Sys 1.2 and 1.3 respectively. While PCA based Sys 1.2 gives a small (1.3% red. in WER) performance improvement, sparse coding based Sys 1.3 turns out to perform even worse than the hard target based baseline itself. We noticed this performance degradation over a range of values for σ and λ for ASR on the *dev* set as well. This experiment confirms the poor quality of SDM based senone alignments as well as the DNN in Sys 1.1 which generated the soft targets. We conclude that our approach is not able to learn meaningful senone subspace information with SDM senone set and SDM based DNN posteriors. Next, we do experiments with soft targets from IHM data.

The baseline for experiments with parallel data uses hard targets from IHM close-talk data with acoustic features from SDM data as shown in Figure 1. Table 2 depicts this as Sys 2.1. In these experiments, we also provide results for ASR on IHM data to compare how ASR performance improvements on IHM data relate to those on SDM data. As expected, Sys 2.1 with IHM hard targets performs better than Sys 1.1 which uses SDM hard targets. We use Sys 2.1 (CE loss based IHM system) to generate soft targets and perform low-rank and sparse reconstution to enhance them in order to train Sys 2.2 and 2.3. The original soft targets didn't bring any significant improvements on IHM data in [8] and we do not consider them here. On IHM data, we notice that PCA soft targets based Sys 2.2 performs the best at 26.8% WER with sequence training. Although sparse soft targets based Sys 2.3 outperforms the IHM hard target baseline, the improvements are still lower than PCA based Sys 2.2. However, on far-field SDM data, both Sys 2.2 and 2.3 give significant WER reductions and Sys 2.3 with sparse soft targets outperforms Sys 2.2 trained using PCA based soft targets. Compared to the SDM hard target based sequence trained baseline, the overall improvement by using Sys 2.3 is 4.4% absolute (∼8% rel.) and compared to IHM hard targets, it is 2.1% absolute (∼4% rel.).

An interesting obseration here is that the sparse soft targets result in better acoustic modeling than their low-rank counterparts for SDM data, whereas we observe the contrary on IHM data. The success of sparse soft targets for SDM shows that the non-linear low-dimensional modeling of senone subspaces, enabled by dictionaries, is highly beneficial for mapping reverberated noisy speech acoustic features to underlying senone classes. We also note that the performance improvements using enhanced soft targets are observed in both CE and sMBR loss based systems, and we conclude that the benefits of enhanced soft targets are complementary to those of sequence training, as shown previously in [9].

In Table 3, we further evaluate our approach on state-of-the-art recurrent and time-delay neural network architectures.

Table 2: ASR performance on AMI IHM and SDM eval set (in WER%) when soft targets are derived from eigenposteriors and dictionaries learned using IHM senone set and corresponding baseline DNN.

Sys #	Training Targets	Evaluation Condition			
		IHM Eval		SDM Eval	
		CE	CE+sMBR	CE	CE+sMBR
1.1	SDM (hard)	-	-	58.6	54.4
2.1	IHM (hard)	30.5	28.0	54.9	52.1
2.2	IHM (PCA95)	29.4	26.8	52.9	51.5
2.3	IHM (SP0.1)	30.4	27.3	52.1	50.0

Table 3: ASR performance using recurrent and time-delay NN architectures on AMI SDM eval set (in WER%) when soft targets are derived from eigenposteriors and dictionaries learned using IHM senone set and corresponding baseline DNN.

		Network Type		
Sys #	Training Targets	LSTM	Bi-LSTM	TDNN
1.1	SDM (hard)	54.9	54.2	55.0
3.1	IHM (hard)	51.3	49.7	51.0
3.2	IHM (PCA95)	50.1	49.3	49.8
3.3	IHM (SP0.1)	50.2	49.3	50.2

We observe in Table 3 that the enhanced soft targets are superior for training the LSTM and TDNN based acoustic models than IHM hard targets. The WER reductions are noticeably smaller for these strong baselines, but we consistently achieve $\sim$5% absolute improvement in WER as compared to the SDM hard targets baseline, and $\sim$1% absolute improvement as compared to IHM hard targets based systems. Bi-LSTM based Sys 3.2 and 3.3 with low-rank and sparse targets perform equally well and give the best WER of 49.3%. These experiments further confirm the importance of modeling low-dimensional senone subspaces for improving ASR. Note that the low-rank and sparse soft targets from the parallel IHM data were still obtained from CE loss based IHM System 2.1 depicted in Table 2. In future, we plan to explore modeling of low-dimensional senone subspaces from stronger IHM baseline systems based on LSTMs or TDNN instead of simple feed-forward DNN acoustic models.

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we presented the use of low-rank and sparse soft targets from parallelly recorded close-talk speech to improve ASR performance on far-field speech. PCA assumes a low-dimensional linear subspace underlying the population of a senone specific DNN posterior matrix. On the other hand, an over-complete dictionary models the senone subspace non-linearly as an union of low-dimensional manifolds. In context of far-field ASR using parallel data, we achieved improvemnts by using enhanced soft targets in place of hard targets from close-talk speech. Gains in ASR performance using sparse soft targets are particularly promising and suggest potential for exploring sparse modeling based techniques to improve far-field ASR. Specifically, we plan to investigate sparse modeling of far-field acoustic features for dereverberation and front-end enhancement of features before DNN based acoustic modeling.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Vijayaditya Peddinti, Vimal Manohar, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Far-field asr without parallel data," *Interspeech 2016*, pp. 1996–2000, 2016.

[2] Yanmin Qian, Tian Tan, and Dong Yu, "An investigation into using parallel data for far-field speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5725–5729.

[3] 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.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] William Chan, Nan Rosemary Ke, and Ian Lane, "Transferring knowledge from a rnn to a dnn," in *Interspeech*, 2015.

[6] Yifan Gong Jinyu Li, "Learning Small-Size DNN with Output-Distribution-Based Criteria," in *Interspeech*, September 2014.

[7] Ryan Price, Ken-ichi Iso, and Koichi Shinoda, "Wise teachers train better dnn acoustic models," *EURASIP Journal on Audio, Speech, and Music Processing*, , no. 1, pp. 1–19, 2016.

[8] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, "Low-rank and sparse soft targets to learn better dnn acoustic models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 5265–5269.

[9] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting eigenposteriors for semi-supervised training of dnn acoustic models with sequence discrimination," in *INTERSPEECH*, 2017.

[10] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, "Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2016, pp. 5690–5694.

[11] Emmanuel J Candès and Michael B Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[12] Jun Du, Qing Wang, Tian Gao, Yong Xu, Li-Rong Dai, and Chin-Hui Lee, "Robust speech recognition with

speech enhanced deep neural networks," *Interspeech 2014*, 2014.

[13] Khe Chai Sim, Yanmin Qian, Gautam Mantena, Lahiru Samarakoon, Souvik Kundu, and Tian Tan, *Adaptation of Deep Neural Network Acoustic Models for Robust Automatic Speech Recognition*, pp. 219–243, Springer International Publishing, 2017.

[14] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[15] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass, "Highway long short-term memory rnns for distant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 5755–5759.

[16] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "Residual lstm: Design of a deep recurrent architecture for distant speech recognition," *arXiv preprint arXiv:1701.03360*, 2017.

[17] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "Bridgenets: Student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5755–5759, 2018.

[18] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[19] Iain McCowan, Jean 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*, 2005, vol. 88.

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

[21] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, 2007.

[22] Leo J Lee, Paul Fieguth, and Li Deng, "A functional articulatory dynamic model for speech production," in *IEEE ICASSP*, 2001.

[23] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, "Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition," *Speech Communication*, 2015.

[24] Steve J Young, Julian J Odell, and Philip 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.

[25] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.

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

[27] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[28] Shakti P Rath, Daniel Povey, and Karel Veselỳ, "Improved feature processing for deep neural networks.," in *Interspeech*, 2013.

[29] Pranay Dighe, "Eigenposteriors and sparse dictionary codes: https://github.com/idiap/eigenposterior," .

[30] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[31] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 285–290.

[32] Jonathan G Fiscus, Jerome Ajot, Nicolas Radde, and Christophe Laprun, "Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech," in *LREC*, 2006.