INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On multi-domain training and adaptation of end-to-end RNN acoustic models for distant speech recognition

Seyedmahdad Mirsamadi and John H.L. Hansen⋆

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3020, U.S.A.
mirsamadi@utdallas.edu, john.hansen@utdallas.edu

Abstract

Recognition of distant (far-field) speech is a challenge for ASR due to mismatch in recording conditions resulting from room reverberation and environment noise. Given the remarkable learning capacity of deep neural networks, there is increasing interest to address this problem by using a large corpus of reverberant far-field speech to train robust models. In this study, we explore how an end-to-end RNN acoustic model trained on speech from different rooms and acoustic conditions (different domains) achieves robustness to environmental variations. It is shown that the first hidden layer acts as a domain separator, projecting the data from different domains into different subspaces. The subsequent layers then use this encoded domain knowledge to map these features to final representations that are invariant to domain change. This mechanism is closely related to noise-aware or room-aware approaches which append manually-extracted domain signatures to the input features. Additionaly, we demonstrate how this understanding of the learning procedure provides useful guidance for model adaptation to new acoustic conditions. We present results based on AMI corpus to demonstrate the propagation of domain information in a deep RNN, and perform recognition experiments which indicate the role of encoded domain knowledge on training and adaptation of RNN acoustic models.

Index Terms: distant speech recognition, recurrent neural network, multi-domain training

1. Introduction

Deep neural network (DNN) acoustic models have led to significant improvements in speech recognition accuracy. Moreover, the application of recurrent neural networks (RNN) has fundamentally changed the design of speech recognition systems from complex DNN-HMM hybrids to simpler end-to-end models, where a single deep RNN maps the sequence of acoustic features to a sequence of phonemes or text characters. Along with this change, there has been a shift of focus in research on distant (far-field) speech recognition. Many existing solutions focus on front-end enhancement strategies to compensate for far-field distortions [1–3]. However, given the powerful modeling capabilities of deep networks, there is increasing interest to address the far-field problem from a robust modeling perspective by training a neural network on far-field data from multiple reverberant environments (domains) with diverse characteristics. This is achieved either by using actual distant speech recorded by far-field microphones [4], or by convolving existing speech corpora by recorded Room Impulse Responses (RIR) from different rooms [5]. Training on a diverse set of reverberant conditions can significantly reduce the mismatch between the resulting model and a new test environment. Such multi-domain trained DNN acoustic models provide very competetive accuracies, often outperforming most other approaches in benchmarks [6]. The provided robustness is attributed to the fact that the deeper hidden representations in the network become increasingly invariant to those variations in data which are not relevant to the classification task. For multi-domain acoustic models trained on speech from different environments, this means that deeper layers become increasingly insensitive to room and RIR characteristics, thus achieving robustness across a wide variety of conditions [7]. However, little is understood about how this invariance is achieved by the network. There has been some effort on analyzing the hidden features of a clean-trained model in order to understand how the final phoneme-discriminative features are generated by the network [8]. However, it is not known how domain invariance is achieved when the model is trained on data from different domains (e.g., speech data from different rooms with different reverberation charactetistics). Because of this lack of understanding, adaptation of such models to new room conditions has been difficult, with often heuristic and ad-hoc strategies used to determine how to tune model parameters towards a new test domain.

The goal of this study is to understand how neural network acoustic models learn to produce domain-invariant features from multi-condition data. We will show that with multi-domain training, although the only provided supervision is the output label sequence, the network implicitly learns to discriminate between the different domains in the training data as well. The basis for this analysis is the residual room (domain) information in different hidden layers of a deep network. We use the accuracy of a simple domain classifier on the hidden features as a proxy measure for the amount of discriminating information contained in each layer about the recording environment, or more generally, about the acoustic path (RIR) characteristics. Additionally, the results of this study provides insights on how adaptation should be carried out when data is available from a specific target environment. We will show how an understanding of the internal hidden representations can provide clues about effective adaptation strategies. We focus on an end-to-end model where a deep RNN is used to directly convert acoustic features to the corresponding seqeunce of characters. However, our analysis is independent of the particular network used, and the results can be extended to other model architectures.

⋆This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

http://dx.doi.org/10.21437/Interspeech.2017-398

2. RNN-CTC acoustic modeling

In conventional DNN-HMM acoustic models, a feed forward network is trained to predict the context-dependent HMM states (senones) from the input speech frames. Recurrent neural networks, in contrast, are able to model temporal dynamics internally, and thus do not depend on a separate HMM model to define the output space. In an end-to-end acoustic model, an RNN directly maps the sequence of speech features to the corresponding sequence of labels in the transcripts. There are two major network architectures which enable such direct transformation, namely encoder-decoder models with attention [9], and RNN-CTC models which use Connectionist Temporal Classification (CTC) objective for automatic alignment [10]. While encoder-decoder RNNs jointly model both acoustic and language information [11], RNN-CTC models use a simplifying conditional independence assumption between outputs, thus acting only as an acoustic model and requiring a separate language model to incorporate language information [12]. This study focuses on RNN-CTC models without any language component, thus focusing excusively on acoustic model robustness.

Given a sequence of feature vectors $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_T]$ from a speech utterance, a deep RNN applies multiple stages of nonlinear recurrent transformations of the form

$$\mathbf{h}_t^l = f(\mathbf{h}_{t-1}^l, \mathbf{h}_t^{l-1}, \boldsymbol{\theta}^l), \tag{1}$$

where $\mathbf{h}_t^0 = \mathbf{x}_t$ and $\boldsymbol{\theta}^l$ denotes the set of weights for each layer. Here we use $f(\cdot)$ to generally represent the layer transformation details in either of the two popular implementations of RNNs, namely Long Short-Term Memory (LSTM) networks [13], or Gated Recurrent Unit (GRU) networks [14] (we leave out the connection details of these recurrent layers as there is now general familiarity with such recurrent architectures).

The activations of the last recurrent layer are passed to a final softmax layer of size $(|S| + 1)$, where $|S|$ is the size of the label set $S = \{s_1, \cdots, s_{|S|}, blk\}$. The softmax outputs at each frame are interpreted as the probability of observing the corresponding label given the input feature sequence.

$$p(s_{i,t}|\mathbf{X}) = \frac{\exp(\mathbf{w}_i^T \mathbf{h}_t^L)}{\sum_{j=1}^{|S|+1} \exp(\mathbf{w}_j^T \mathbf{h}_t^L)}. \tag{2}$$

The first $|S|$ labels $s_1, \cdots, s_{|S|}$ are actual symbols in label sequences which can either be phonemes or text characters. Using character output space has the benefit of not requiring a phonetic dictionary, although it generally requires more data and a deeper network to learn the mapping. The extra label (blk) represents a blank, or *no output* for a particular time, which enables the algorithm to effectively align the label sequence with the features. The CTC objective is to maximize the overall probability of the label sequence given the feature sequence, using any possible alignment between them:

$$\boldsymbol{\theta} = \arg\max \sum_{a \in \mathcal{A}} \prod_{t=1}^{T} p\big(s_{a[t],t}|\mathbf{X}\big). \tag{3}$$

where $\mathcal{A}$ is the set of all possible alignments, $a[t]$ is one such alignment which gives a symbol index for every time frame t, and $\boldsymbol{\theta}$ represents all parameters in the network. A forward-backward recursion is used to efficiently compute the above objective for each utterance [10]. The resulting gradients are then back-propagated through time and over all hidden layers to tune parameters. To decode a test example, the simplest approach is to use a memoryless search by selecting the most active output at each frame, followed by removing blanks and label repetitions (often referred to as best-path decoding). Alternatively, a beam search can be used similar to [15] which tracks multiple candidate paths over the previous frames to yield the most likely overall label sequence.

3. System setup and data

This study uses the far-field recordings of AMI corpus [16] to examine how environmentally-robust representations are learned by RNN acoustic models. The AMI corpus consists of conversational speech recorded during meetings in 3 different rooms[1]. The meetings are recorded both by independent headset microphones (IHM) and a microphone array placed on the meeting table, from which we only use a single channel (single distant microphone, SDM). The SDM data provides the multi-domain far-field speech we need for our study, because it contains speech not only from different rooms, but also from multiple different speaker positions within each room (hence various RIRs and signal to reverberation ratios). Additionally, IHM data allows us to build clean-trained models for comparisons. We remove all utterances containing any overlapped speech frames from both train and test sets, resulting in 30 hours of data for train, and 3.5 hours for each of dev and test sets.

We define domains within SDM data in two different ways. We can consider each meeting room as a single domain, which leads to 3 domains each containing 10 hours of data. Alternatively, we can consider each RIR (corresponding to a speaker position) in each meeting as a separate domain, yielding a finer partitioning into 674 different domains, each with a few minutes of data. An important point to note about AMI corpus is that each source position (RIR) also corresponds to a different speaker. In other words, the above defined domains differ not only in terms of environmental characteristics, but also in terms of speakers within each domain.

We use an end-to-end RNN-CTC model consisting of 3 bidirectional LSTM (BLSTM) layers with 128 cells in each direction, followed by a final softmax layer with 79 outputs representing each of the symbols in our character set plus the blank symbol. We adopt an output space similar to [17], where instead of using a space character, capital letters are used as word delimters. The input features are 24 dimensional Mel filterbank coefficients extracted from 25 msec frames at a rate of 100 frames per second, and are mean and variance normalized across each speaker. The network parameters are optimized using RMSprop [18] with an inital learning rate of 0.001 and a minibatch size of 20 utterances. We use frame-skipping [19] with a context window of 3 frames to speed up training. Training iterations are stopped when no further improvement is observed on the development data. Beam search decoding uses a beam width of 10 paths in all cases. All decoding is based on acoustic scores only, using no language or lexicon information.

4. RNN hidden representations and domain-invariance

The multiple levels of transformation in a deep RNN are intended to extract final (last layer) features which are sensitive only to speech sounds, with minimal displacement caused by other variations in the data. In other words, ideally, the under-

[1]Three different collection sites: University of Edinburgh (U.K.), Idiap Research Institute (Switzerland), and the TNO Human Factors Research Institute (The Netherlands).

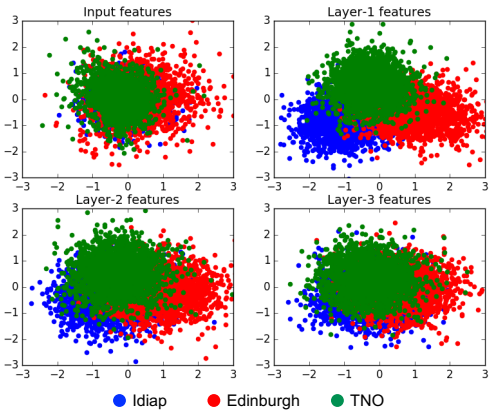

Figure 1: *RNN hidden features projected onto the 2D plane.*

lying domain should be indistinguishable from the final hidden layer representations. In the far-field ASR problem, the same speech recorded in different rooms should ideally produce identical last layer features.

We have observed that this environment invariance is achieved by automatically projecting the input features of different domains onto disjoint subspaces through the first hidden layer transformation. The subsequent layers then use this domain information to learn suitable mappings which compensate domain-specific distortions and yield invariant final representations. Note that the only supervising information provided during training is the label sequence, so this projection into different subspaces is achieved implicitly without presenting any explicit domain information to the network. Figure 1 shows the hidden representations of a 3-layer deep BLSTM trained on far-field (SDM) data, projected into the 2D plane by Linear Disrminant Analysis (LDA). It is observed that while the input filterbank features show almost no domain separation, the first layer has achieved to separate the data of each room into different support regions. From that stage forward, the subsequent layers gradually remove domain dependencies, trying to achieve final representations in which environmental dependencies have been fully removed.

To further quantify this observstion, we use a simple logistic regression domain classifier trained on each of the hidden layer features:

$$p(d_i|\mathbf{h}^l) = \frac{\exp(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_i^T\mathbf{h}_t^l)}{\sum_{j=1}^{D}\exp(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_j^T\mathbf{h}_t^l)}, \quad (4)$$

where D represents the total number of domains within the data ($D = 3$ for room domains and $D = 539$ for RIR domains). Here we use an average pooling of the logisitic regression outputs from each frame to obtain an utterance-level representation of the corresponding domain. The cross-entropy error between the above posteriors and the corresponding ground-truth domain label is used to train the classifier parameters $\mathbf{u}_i (i = 1, \cdots, D)$. Fig. 2 shows the final domain classificaion accuracy achieved in each case for the different hidden layers. Best accuracy is achieved at layer 1, which indicates maximum domain information at this stage. The subsequent layers gradually discover features that are increasingly insensitive to domain change, thus yielding lower classification accuracy [1]. This overall mechanism can be viewed as closely related to noise-aware [20] or room-aware [21] approaches where input features are augmented with manually extracted noise or reverberation signatures to provide the model with extra information about the environment. Our results here indicate that given sufficient depth and multi-domain training data, such domain-specific augmentations will be automatically learned in the intial layers of the network.

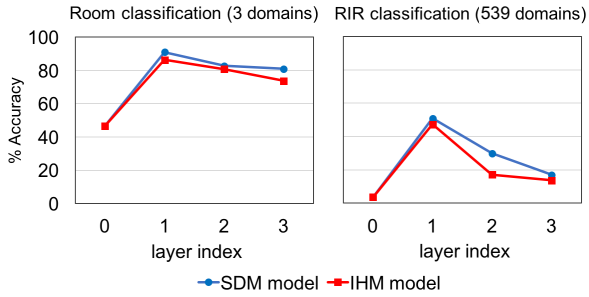

Figure 2: *Room classification accuracies based on hidden layer representations. The input to the network is SDM channel for all curves. Blue curves use SDM-trained model to produce hidden features, and red curves use IHM-trained model.*

As was mentioned in Section 3, the different domains in AMI corpus differ not only in terms of the recording environment, but also in terms of the underlying speakers. Thus, the SDM model (blue curves) actually learns discrepancies resulting both from environment and speaker characteristics. It is not clear how much of this acquired domain knowledge pertains to room (RIR) characteristics and how much to speaker differences. To quantify the role of each factor, Fig. 2 also shows domain classification results when the underlying model is trained using IHM data, i.e., a clean-trained model (red curves). Note that the input features in this case are still the same SDM features, but the hidden features have been obtained using IHM-trained model. The IHM model cannot contain any information about environment characteristics since it is trained only on close-talking data. Thus, any domain information in the hidden layers must represent speaker characteristics. Consequently, the accuracy improvement provided by the SDM model (i.e., the difference between red and blue curves) indicates the additional room or RIR information learned by the model from far-field data.

5. Model adaptation for deep RNNs

An important problem in acoustic modeling is the ability to adapt existing models to new acoustic conditions using a small number of utterances from a target environment. This has often been difficult with DNNs due to their large number of parameters. Specifically, it is not possible to tune all of the parameters in a deep network without suffering from excessive overfitting (removing previously learned information). A popular solution is to introduce additional domain-dependent layers in the network and tune only the new parameters based on adaptation

[1] Note that we have used a simple linear classifier. By using approporiate nonlinear feature transformations (i.e., an arbitrarily deep domain classifier), we can achieve higher accuracies with input filterbank features. However, the goal here is to measure the domain knowledge already encoded in the feature representations at each layer, without any supervised nonlinear transformations to extract those features.

Table 1: *Character Error Rates with clean-trained (IHM) and multi-domain far-field (SDM) models*

Train Data	Test Data	CER (best-path)	CER (beam-search)
IHM	IHM	37.1	35.8
IHM	SDM	65.8	64.8
SDM	SDM	52.7	51.6

These results use standard ASR partitions of AMI corpus [25].

data, keeping the rest of the network unchanged [22, 23]. Most often the new layer is a simple feed-forward linear layer which is initialized as an identity transform and trained using adaptation data. However, there is no conclusive study on where exactly in the network this additional transformation should be inserted. Most studies choose the position which provides the best empirical results.

Our discussion on multi-domain training of deep RNNs suggests that a linear transformation after the first hidden layer is the most appropriate choice for model adaptation to new rooms and acoustic conditions. This is in light of the observation that the goal in adaptation is to transform the intermediate features such that the resulting distribution resembles that of the training data. In other words, we are interested in a *domain switch* from the new test domain to the domain of training samples. As was pointed out in Section 4, domain information is maximum in the first hidden layer of a deep RNN. Therefore, it is reasonable to perform the domain switch at this level. We report results of adaptation experiments in Section 6 which verify this observation. Moreover, our previous results in [24] (on adaptation of clean-trained models to reverberant data) are also in agreement with this observation.

6. ASR experiments

6.1. Results on multi-domain training

In this section we report results of ASR experiments to assess the effectiveness of multi-domain training on end-to-end RNN-CTC acoustic models, using the setup details explained in Section 3. Table 1 shows the obtained Character Error Rates (CER) on the AMI far-field (SDM) test set. These results use the standard ASR train/dev/test data partitions in [25], which includes data from all 3 meeeting rooms in the train set, but uses separate meeting sessions (hence different speakers and RIRs) for train and test. The IHM model suffers a sharp performance degradation when presented with far-field SDM test data due to the large mismatch between the acoustic conditions of train and test. Training a model on multi-domain SDM data compensates a significant portion of this degradation, yielding 20% improvement relative to the clean-trained model.

To better assess the effect of having multiple different domains within the training data, we run a second set of experiments using train and test subsets of the AMI corpus that are different from the standard partitions. From the standard train set, we select three 10-hour subsets, each containing a different number of RIR domains. These subsets all contain the same amount of data (10 hours), but differ in terms of the number of RIR domains from which they are sampled. So any difference in the resulting performance can be attributed to the difference in domain diversity. The obtained error rates are shown in Table 2. It is observed that increasing the diversity of acoustic conditions in the train data consistently improves performance.

Table 2: *Character Error Rates with increasing number of domains within the train set. The total amount of train data in all cases is 10 hours.*

Train Data	Test Data	# domains	CER (best-path)	CER (beam-search)
SDM	SDM	200	59.5	58.4
		400	59.0	57.7
		539	57.9	57.1

These results use the standard AMI test set according to [25], but the train data is 10-hour subsets of the standard train set which sample from the specified number of domains in each case.

6.2. Results on adaptation

Here we use yet another custom partitioning of AMI data, in which all the data from Edinburgh and TNO meeting rooms ($\sim$25 hours) is used for training, and the Idiap room data ($\sim$10 hours) is equally split into three subsets for adaptation, development and test. The goal is to investigate adaptation of the original acoustic model trained on Edinburgh and TNO rooms to the acoustic conditions of the Idiap meeting room. The resuls are provided in Table 3. As expected from the discussion in Section 5, an intermediate transformation inserted after the first hidden layer is most effective for adaptation of the deep RNN to the acoustic properties of a new environment. This approach provides 3.5% relative improvement compared to the unadapted model. These results agree with our previous findings in [24]. However, the work in [24] studies adaptation of clean-trained models to reverberant data. Here, the original model is already trained on far-field speech (but from different rooms than the adaptation and test data).

Table 3: *Character Error Rates with different adaption methods*

Train Data	Test Data	position of adaptation layer	CER (best-path)	CER (beam-search)
SDM	SDM	None	56.5	55.6
		after input	55.4	54.5
		after layer-1	**54.5**	**53.5**
		after layer-2	55.2	54.2
		after layer-3	55.3	54.4

These results use the custom partitioning of AMI corpus explained in Section 6.2 (Train on data from Edinburgh and TNO rooms, and divide the Idiap room data into adaptation, dev, and test sets.)

7. Summary and Conclusions

We provided an investigative analysis of deep RNN acoustic models trained on far-field data for robust distant speech recognition. It was shown that the invariant representations acquired by such models at the last hidden layer is a result of a projection of training data from different environements onto different subspaces in the initial hidden layers. This encoded domain knowledge helps the subsequent layers to apply appropriate transformations to obtain final hidden representations that are insensitive to the characteristics of the recording environment. Based on this observation, we concluded that domain-dependent transformations used for model adaptation should be applied on the hidden features of the first RNN layer. These observations were verified by detailed ASR experiments based on AMI corpus.

8. References

[1] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, Nov 2012.

[2] T. Yoshioka and M. J. Gales, "Environmentally robust asr front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[3] S. Mirsamadi and J. H. L. Hansen, "A generalized nonnegative tensor factorization approach for distant speech recognition with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1721–1731, Oct 2016.

[4] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *ICASSP*. IEEE, 2016, pp. 5755–5759.

[5] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, March 2017.

[6] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASPIRE system: Robust LVCSR with TDNNs, ivector adaptation and RNN-LMs," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 539–546.

[7] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[8] S. Tan, K. C. Sim, and M. Gales, "Improving the interpretability of deep neural networks with stimulated learning," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 617–623.

[9] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[11] W. Chan, "End-to-end speech recognition models," Ph.D. dissertation, Carnegie Mellon University, 2016.

[12] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks." in *ICML*, vol. 14, 2014, pp. 1764–1772.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[15] A. Y. Hannun, A. L. Maas, D. Jurafsky, and A. Y. Ng, "First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs," *arXiv preprint arXiv:1408.2873*, 2014.

[16] 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.

[17] G. Zweig, C. Yu, J. Droppo, and A. Stolcke, "Advances in all-neural speech recognition," *arXiv preprint arXiv:1609.05935*, 2016.

[18] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[19] V. Vanhoucke, M. Devin, and G. Heigold, "Multiframe deep neural networks for acoustic modeling," in *ICASSP*. IEEE, 2013, pp. 7582–7585.

[20] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *ICASSP*. IEEE, 2013, pp. 7398–7402.

[21] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *ICASSP*. IEEE, 2015, pp. 5014–5018.

[22] Y. Miao and F. Metze, "On speaker adaptation of long short-term memory recurrent neural networks." in *INTERSPEECH*, 2015, pp. 1101–1105.

[23] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 366–369.

[24] S. Mirsamadi and J. H. Hansen, "A study on deep neural network acoustic model adaptation for robust far-field speech recognition." in *INTERSPEECH*, 2015, pp. 2430–2434.

[25] The AMI meeting corpus. [Online]. Available: http://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml