Far-field ASR without parallel data

Vijayaditya Peddinti[1], *Vimal Manohar*[1], *Yiming Wang*[1], *Daniel Povey*[1,2], *Sanjeev Khudanpur*[1,2]

[1]Center for Language and Speech Processing &
[2]Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD 21218, USA
{vijay.p,vmanoha1,yiming.wang,khudanpur}@jhu.edu, dpovey@gmail.com

Abstract

In far-field speech recognition systems, training acoustic models with alignments generated from parallel close-talk microphone data provides significant improvements. However it is not practical to assume the availability of large corpora of parallel close-talk microphone data, for training. In this paper we explore methods to reduce the performance gap between far-field ASR systems trained with alignments from distant microphone data and those trained with alignments from parallel close-talk microphone data. These methods include the use of a lattice-free sequence objective function which tolerates minor mis-alignment errors; and the use of data selection techniques to discard badly aligned data. We present results on single distant microphone and multiple distant microphone scenarios of the AMI LVCSR task. We identify prominent causes of alignment errors in AMI data.

Index Terms: far-field speech recognition, neural networks, parallel data

1. Introduction

In far-field speech recognition systems, alignments for training the acoustic models are typically generated using the distant microphone recordings. However, in some cases parallel close-talk microphone recordings are available for the training data. Typically, this scenario occurs when the training corpora have been recorded with both distant and close-talk microphones (*e.g.* AMI meeting corpus [1, 2, 3]) or in cases where far-field audio is simulated by distorting close-talk microphone recordings (*e.g.* ASpIRE [4], REVERB-2014 [5]). When such parallel recordings are available, the alignments used for training the acoustic models can be generated from close-talk microphone audio recordings. Empirical analysis shows that the use of these comparatively higher quality alignments leads to significant improvements ($\sim$8% relative improvement in word error rate). However in typical large data scenarios, where actual far-field audio is collected, assuming the availability of close-talk microphone recordings is not practical.

In this paper, we identify the possible reasons for the performance difference between the ASR systems that are trained using alignments generated from distant microphone recordings, and those trained with alignments generated from parallel close-talk microphone readings. Further, we propose a two pronged strategy to reduce this performance gap. Firstly, we use the lattice-free maximum mutual information (MMI) objective function [6], which is tolerant to minor mis-alignment errors, to train the neural networks from random initialization. Secondly, we propose a quality estimate which is used for selecting reliable utterances for training. The combination of these two techniques reduces the performance gap from $\sim$8% to $\sim$1.5%. We present results on both single distant microphone and multiple distant microphone scenarios of the AMI LVCSR task.

The paper is organized as follows. In Section 2, we describe the motivation for this work. In Section 3, we present an analysis of errors in alignments generated from distant microphone recordings available in AMI database. In Section 4.1, we describe the lattice-free MMI criterion. In Section 4.2, the proposed utterance quality metric and the data selection criteria are described. Section 5 describes the experimental setup, Section 6 presents the results and finally Section 7 presents our conclusions.

2. Motivation

There are three LVCSR tasks [7, 8] designed using the AMI meeting corpora [1]. These are the individual headset microphone (IHM), single distant microphone (SDM) and multiple distant microphone (MDM) tasks; named based on the type of audio used in the creation of the *train*, *dev* and *eval* sets. The AMI corpus, with parallel speech recordings from all these microphones, provides an opportunity to analyze the importance of alignment quality in far-field speech recognition systems. In addition to the three standard AMI LVCSR systems, which use alignments from the HMM-GMM systems trained using the corresponding audio, we also trained systems using alignments generated from the IHM audio. Table 1 summarizes the results of the 8 such LVCSR systems (see Section 5 for details). It can be seen that there is a significant reduction in word error rate (WER) (7.75% relative, on average) when using alignments from IHM audio. Further these relative improvements increase when using better acoustic models.

3. Analysis of alignment errors

Motivated by the observations in Table 1, we performed a comparison of alignments generated from IHM and SDM systems. We randomly sampled utterances from the AMI corpus and identified some prominent categories of errors.

3.1. Minor mis-alignment errors

A majority of the errors were minor mis-alignment errors. Figure 1 shows the log mel filter-bank coefficients from the IHM and SDM recordings; and compares the phone alignments generated by the corresponding HMM-GMM systems. It can be seen that there are just minor differences between these two

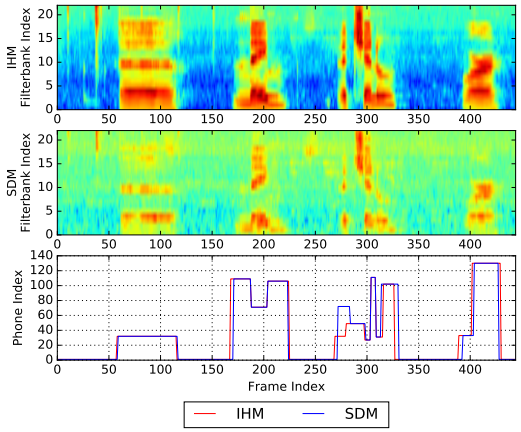
This work was partially supported by NSF CRI Grant No 1513128, DARPA LORELEI Contract No HR0011-15-2-0024, IARPA BABEL Contract No 2012-12050800010 and DARPA LORELEI Contract No HR0011-15-2-0027. .

Table 1: Comparison of AMI LVCSR systems trained with close-talk and distant microphone alignments

Model	LVCSR task	Alignments	WER (%)	
			dev	*eval*
TDNN	SDM	SDM	45.8	50.3
	SDM	IHM	41.8	46.6
	Rel. Change		8.7%	7.3%
	MDM	MDM	41	44.7
	MDM	IHM	38.2	42
	Rel. Change		6.8%	6.0%
BLSTM	SDM	SDM	42.5	45.6
	SDM	IHM	38.5	41.8
	Rel. Change		9.4%	8.3%
	MDM	MDM	38.6	41.0
	MDM	IHM	35.5	38.3
	Rel. Change		8.0%	6.6%

alignments. The significant difference in alignments, between frames 250 and 300, occurs due the choice of different pronunciations (*EY* vs *AH*) for the word *"a"* by the IHM and SDM systems.

Figure 1: TS3006b_MTD021PM (1668.61-1673.08 seconds)
"Uh name a channel or"

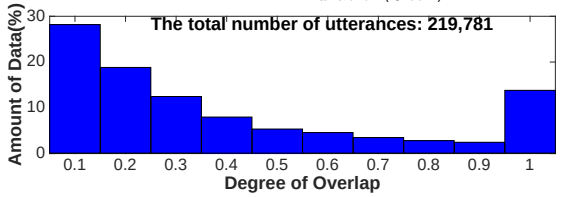


3.2. Speaker overlap errors

In a significant number of utterances, there were speaker overlaps in both IHM and SDM audio. For the IHM audio, the transcription corresponded to dominant speaker, as expected. However, this was not necessarily the case in SDM audio. These errors worsened the quality of the SDM alignments. Figure 3 represents one such utterance. In this plot the green line shows the alignment for the competing speaker using his IHM audio. It can be seen that the speech of this speaker, identified by the non-zero green line, distorts the SDM alignment. Figure 2 shows the amount of training data with different degrees of overlaps.

Figure 2: Histogram of data with different degree of overlaps

$$\text{DegreeOfOverlap}(Utt1) = \frac{\max_{Utt2} \text{Duration}(Utt1 \cap Utt2)}{\text{Duration}(Utt1)}.$$



3.3. Transcription errors

As in other databases, there were minor transcription errors. However, AMI corpora had errors where there were significant

Figure 3: TS3009d_MTD033PM (1991.99-1994.14 seconds) :
Transcribed speech : *"She already knows"*
Overlapping untranscribed speech: *"Who is she you're talking about"*

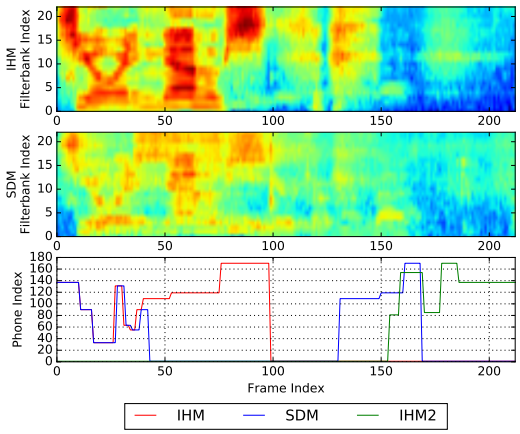
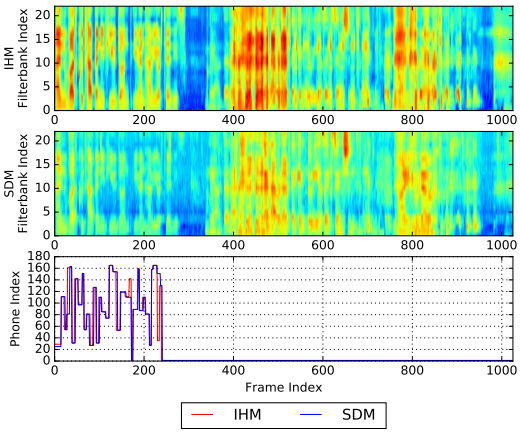


Figure 4: IN1002_MIO076 (686.66-696.94 seconds) :
Transcription:*"And what could happen if you don't even have your"*
Untranscribed trailing speech: *"Then I would have taken two year's extension"*



portions of untranscribed non-overlapping speech in the utterances. Figure 4 provides one such example. Significant portion of the signal corresponds to a second talker's response and it is untranscribed. Further, both the IHM and SDM systems align this trailing speech to silence.

4. Proposed techniques

We propose a two pronged strategy to tackle the errors described in Section 3. Firstly, to make the learning algorithm robust to minor mis-alignment errors, we use the lattice-free MMI objective function [6]. This approach is described in Section 4.1. Secondly, we filter the utterances that might have speaker overlap or transcription errors. To accomplish this, we propose a quality measure for utterances. This is described in Section 4.2.

4.1. Lattice-free MMI objective

Povey *et al.* [6] introduced a lattice-free version of the MMI criterion with modifications motivated from the recent efforts in CTC training ([9, 10, 11, 12, 13]).

The modifications to MMI-based training method are:

- Training from scratch without initialization from a cross-entropy system
- The use of a 3-fold reduced frame rate [13] (and a simpler HMM topology)

- Limiting the range of time frames where supervision labels can appear by using Finite State Acceptors [12]

This new method of training has been shown to provide significant improvements compared to conventional sequence discriminative training methods across different LVCSR tasks in [6]. However, our interest in this objective function arises from the fact that it is inherently tolerant to alignment errors, as we can specify a range of time frames for a particular context-dependent phone state using a desired tolerance. In this section, we highlight this particular aspect of the cost function. Readers are encouraged to refer to [6] for more details about this new training method.

The derivative computation for the MMI objective requires the computation of state occupancy statistics using the forward-backward algorithm on the numerator and denominator graphs [14]. The denominator graph is built using a phone n-gram language model. The numerator graph creation is of relevance to this paper.

Prior to training the neural net, a GMM-based system is used to generate lattices representing alternative pronunciations of the training utterances. These lattices are processed into phone graphs and then compiled into utterance-specific Finite State Acceptors (FSAs) as for conventional training. Separately, the lattices are also processed into frame-by-frame masks of what phones are allowed to appear on what frames: a user-specifiable tolerance allows a phone to appear slightly before or after where it appeared in the lattice. As the frame-by-frame phone mask built from the lattices has a tolerance, we expect the gradient computation to be tolerant to minor misalignment errors. We found a 50 ms tolerance to be optimal.

4.2. Lattice oracle WER

We use lattice oracle WER as a quality estimate of the transcript. Given a training utterance and its corresponding transcript, the procedure to find the lattice oracle WER is given in Algorithm 1. For step 4, we use the same algorithm as in [15] for finding the edit distance between a lattice and a reference, but replace the lattice forward-backward with a Viterbi search.

Algorithm 1 Procedure to compute lattice oracle WER

Input: Utterance u with transcript $\mathcal{R}$
Input: $\mathcal{W}$ = List of 100 most common words in the training set
1: **procedure** LATTICE ORACLE($u, \mathcal{R}$)
2: Build a biased unigram language model using the words in $\mathcal{R}$ and $\mathcal{W}$
3: Decode the utterance u using the language model to get a lattice $\mathcal{L}$
4: Find a path in the lattice $\mathcal{L}$ that has the minimum Levenshtein edit distance from $\mathcal{R}$. Let its score be d.
Output: $\frac{d}{|\mathcal{R}|}$, where $|\mathcal{R}|$ is the number of words in $\mathcal{R}$.
5: **end procedure**

We make use of the lattice oracle WER in a simple filtering scheme. Figure 5 shows the percentage of data covered under different oracle WER thresholds. It can be seen that $\sim$95% data can be preserved with WER thresholds around 50%. Utterances with larger oracle WER, including the 5% that has greater than 100% oracle WER, predominantly have either speaker overlaps (Section 3.2) or transcription errors (Section 3.3).

One drawback of this approach is that short segments that have a single word (*e.g.*, *"Yeah"*, *"Okay"*) in the reference will almost always be given an oracle WER of 0, because if that word is in the decoded lattice, it will be picked up by the Viterbi search. However, these amount to very little data.

Figure 5: Amount of data retained with each lattice oracle WER threshold

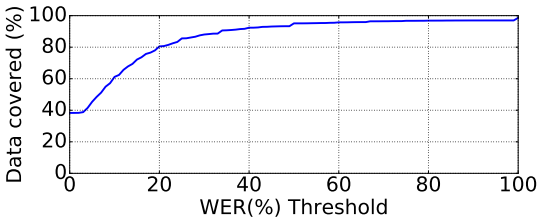


Table 2: Impact of alignment quality based filtering on TDNN acoustic models trained with lattice-free MMI critera for SDM LVCSR task

FER (%) threshold	Data retained	WER (%)	
		dev	*eval*
50	82	44.2	48.1
60	95	43.1	46.9
70	96	42.8	46.6
80	97	43.6	47.2
All	All	43.2	47.3

4.2.1. Filtering Based on Frame-Level Alignment Quality

As use of IHM alignments reduced the WERs significantly, we treated these as ground truth labels and measured the duration normalized Levenshtein distance between the per-frame phone alignments of SDM and IHM systems. This frame error rate (FER) was used to filter out utterances in the SDM task (see Table 2), resulting in similar improvements as with lattice oracle WER.

The utterance in Figure 1 had a lattice oracle WER of 20.00% and an FER of 7.2%. The utterance in Figure 3 had a lattice oracle WER of 0.00 and an FER of 46.48%. The utterance in Figure 4 had a lattice oracle WER of 20.00% and and FER of 7.02%.

5. Experimental Setup

In this paper, we use HMM-DNN hybrid neural network acoustic models. The training recipes for IHM, SDM and MDM LVCSR tasks, for which results are reported in this paper, are very similar. This common recipe is described here briefly. The experiments in this paper were performed using the Kaldi speech recognition toolkit [16]. In particular, the code to reproduce the results in this paper is available at [17].

The HMM-GMM systems for generating the alignments and lattices, used to train the neural network acoustic models, are as described in [8]. However, unlike in [8], we perform speaker-adaptive training of the HMM-GMM systems for all the three tasks, as we found the alignments from SAT HMM-GMM systems to be beneficial for neural network training on all three tasks.

The MDM LVCSR systems have an additional stage of beam-forming to combine the audio captured from different channels of the distant microphone. The BeamformIt toolkit [18] was used for delay-sum beamforming.

To train the SDM and MDM LVCSR systems with alignments generated from IHM data, we identified parallel segments in IHM audio corresponding to the utterances in SDM/MDM

Table 3: Comparison of rel. changes in WER(%) when using alignments from IHM and SDM/MDM data to train TDNN acoustic models

LVCSR task	Alignments	Cross-entropy		Lattice-free MMI		Lattice-free MMI + Data filtering	
		dev	*eval*	*dev*	*eval*	*dev*	*eval*
SDM	SDM	45.8	50.3	43.2	47.3	42.8	46.1
SDM	IHM	41.8	46.6	41.3	45.3	41.6	45.4
Rel. Change		8.7%	7.3%	4.4%	4.2%	2.8%	1.5%
MDM	MDM	41	44.7	40.5	43.2	38.5	41.5
MDM	IHM	38.2	42	38.1	42	38.1	41.5
Rel. Change		6.8%	6.0%	5.9%	2.78%	1.0%	0%
IHM	IHM	24.4	25.1	22.6	22.5	22.4	22.4

data. The IHM SAT HMM-GMM system was used to generate alignments and lattices from these parallel utterances.

5.1. HMM-DNN acoustic models

We use the *speed-perturbation* data augmentation technique ([19]) to simulate synthetic speakers; and we used iVectors to perform *instantaneous adaptation* of the neural network ([20]). The online iVector extraction procedure is described in [21]. The *nnet3* toolkit, by Povey *et al.,* ([22]) in Kaldi speech recognition toolkit [16] was used to perform neural network training. It uses model averaging based distributed optimization algorithm in described in [23].

5.1.1. Neural network architectures

In this paper, we report results on three different neural network architectures – the sub-sampled time delay neural networks (TDNNs, [24]), the long short term memory networks (LSTMs, [25]) and bidirectional LSTMs (BLSTMs). The TDNNs trained with the lattice-free MMI technique have smaller number parameters compared to their cross-entropy trained counterparts. The cross-entropy trained TDNNs are similar to those used in [21], while the the lattice-free MMI trained TDNNs are same as those used in [6]. We use (B)LSTM layers with recurrent and non-recurrent projections as suggested in [25].

The lattice-free MMI technique uses fixed length chunks of 1.5 seconds to perform sequence training. As nearly 50% of the utterances in the AMI corpus were less than 1.5 seconds long we combined neighboring utterances to reach the 1.5 second minimum utterance length.

6. Results

Table 3 contrasts the WER for various training+test conditions with different training criteria/data-sets.

First, compare across the row for the SDM task with IHM training alignments to note that the MMI training and data filtering have only a modest impact when parallel clean+noisy recordings are available: minimal difference in *dev* WER and small improvement in *eval* WER. The same is also true in the MDM task with IHM training alignments.

More importantly, compare across the row for SDM task with SDM alignments to note that MMI training results in a significant reduction in WER relative to cross-entropy training, and the data filtering step yields further gains. The same observation holds for the MDM task with MDM alignments.

Finally contrasting the IHM training alignments with the SDM training alignments for the SDM task, note that while the cross-entropy training criterion suffered a $7\% - 8\%$ degradation in WER relative to IHM alignments, the MMI criterion by itself limits the WER degradation to about 4%, and the data filtering brings down this difference to about 2%. The same trend holds even more strongly for the MDM task – the relative degradation from IHM alignments to MDM alignments is reduced from $6\% - 7\%$ to $0\% - 1\%$.

This last set of results supports the main claim of the paper, namely that the proposed method – using an alignment-tolerant MMI training objective after filtering out the most problematic part of the training data – mitigates strongly against degradation in WER when parallel clean+noisy speech is not available for training acoustic models.

Table 4 compares the impact of using different lattice oracle WER thresholds on the acoustic model quality, as measured in WER on *dev* and *eval* sets. It can be seen that at 45% lattice oracle WER threshold we see the maximum gains. This preserves $\sim 95\%$ of the data in the train set of the corpora. It can be seen that the same data filtering step does not have a significant impact on the acoustic models trained with the cross-entropy criterion.

Table 4: Impact of data filtering on TDNN acoustic models trained with cross-entropy or lattice-free MMI criteria, for SDM LVCSR task

WER threshold (%)	WER (%)			
	Cross-entropy		Lattice-free MMI	
	dev	*eval*	*dev*	*eval*
40	45.4	50.3	43.1	46.9
45	45.5	50.1	42.8	46.1
50	45.5	50.1	42.8	46.6
All	45.8	50.3	43.2	47.3

Our initial experiments with the lattice-free MMI objective function did not result in gains with BLSTM models on this task, though [6] suggests that gains should be expected. We attribute this so the small amount of data in the AMI task, and are currently investigating hyperparameter settings for BLSTM training that are most suitable for this task.

7. Conclusion and future work

In this paper, we proposed a two pronged strategy to reduce the performance gap in far-field ASR systems, when using alignments from close-talk microphone (IHM) and distant microphone (SDM/MDM) audio – using a lattice-free MMI objective function which is tolerant to minor mis-alignment errors; and a data filtering technique based on lattice oracle WER. We reduced the relative change in WER, on using IHM alignments, from 7.2% to 1.3%, on average.

8. References

[1] 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.

[2] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[3] S. Renals, T. Hain, and H. Bourlard, "Recognition and interpretation of meetings: The AMI and AMIDA projects," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, 2007.

[4] M. Harper, "The automatic speech recognition in reverberant environments (aspire) challenge," in *Proceedings of ASRU*, 2015.

[5] J. T. Geiger, E. Marchi, B. Schuller, and G. Rigoll, "Reverb workshop 2014," 2014.

[6] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proceedings of Interspeech*, 2016. [Online]. Available: http://www.danielpovey.com/files/2016_interspeech_mmi.pdf

[7] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The ami meeting transcription system: Progress and performance," in *Machine learning for multimodal interaction*. Springer, 2006, pp. 419–431.

[8] P. Swietojanski, A. Ghoshal, and S. 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.

[9] 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.

[10] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4280–4284.

[11] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[12] A. Senior, H. Sak, F. de Chaumont Quitry, T. N. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *ASRU*, 2015.

[13] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2005.

[15] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[16] 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.

[17] *"Code to reproduce results of the experiments in this paper."*, "2016 (accessed March 23, 2016)", "https://github.com/vijayaditya/kaldi/blob/chain_ami/egs/ami/s5/README_AMI_PAPER".

[18] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[19] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*, 2015.

[20] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 55–59.

[21] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] D. Povey, *"Nnet3: Neural network toolkit for generic acyclic computation graphs"*, "2016 (accessed March 23 , 2016)", http://www.danielpovey.com/kaldi-docs/dnn3_code.html.

[23] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *CoRR*, vol. abs/1410.7455, 2014. [Online]. Available: http://arxiv.org/abs/1410.7455

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of INTERSPEECH*, 2015.

[25] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Feb. 2014. [Online]. Available: http://arxiv.org/abs/1402.1128