IMPROVING MULTICHANNEL SPEECH RECOGNITION WITH GENERALIZED CROSS CORRELATION INPUTS AND MULTITASK LEARNING

Yu Zhang[1,2]*, Wenjie Li*[1,2]*, Pengyuan Zhang*[1,2]*, Yonghong Yan*[1,2,3]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China
[2]University of Chinese Academy of Sciences, China
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing,
Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China

ABSTRACT

Acoustic signals from microphone arrays are used to improve performance in distant speech recognition due to the availability of spatial information. And multichannel automatic speech recognition (ASR) systems often separate speech enhancement module from acoustic modeling, which may be not optimal for improving recognition accuracy. In this work, we propose to improve multichannel speech recognition by supplying the generalized cross correlation (GCC) between microphones, which encodes spatial information, as input features to a long short-term memory (LSTM) acoustic model in parallel with the regular acoustic features. Moreover, multitask learning architecture is incorporated and shows its ability to improve the robustness of the model. We performed experiments on the AMI and ICSI meeting corpora, with results indicating that the proposed model outperforms the model trained directly on the concatenation of multiple microphone outputs and the model trained on a beamformed channel.

Index Terms— speech recognition, microphone array, acoustic model, generalized cross correlation, multitask learning

1. INTRODUCTION

Deep neural networks (DNNs) based acoustic models [1] have driven tremendous improvements in automatic speech recognition (ASR) in recent years. Further improvements are achieved by using more complex models such as convolutional neural networks (CNNs) [2] and long short-term memory based recurrent neural networks (LSTMs) [3]. However, it still remains challenging to perform recognition when the speaker is distant from the microphone, because of the presence of background noise, reverberation, and competing acoustic sources [4]. In such cases, ASR systems often use signals from multiple microphones to enhance the speech signal and reduce the impact of noise and reverberation. And multichannel ASR systems often adopt a two-part architecture, in which a beamforming algorithm is applied to enhance the speech, followed by conventional acoustic modeling approaches [5]. Since the speech enhancement part is usually separate from the speech recognition part, the system fails to optimize towards the final objective, i.e. speech recognition accuracy, which leads to a suboptimal solution [6].

To obtain an optimal performance, joint training of speech enhancement and acoustic model was proposed to improve speech recognition accuracy. Sainath et al. [7] presented a multichannel neural network model trained directly from raw waveform input signal. The spatial and spectral filtering were performed through one layer of neural network. In [8], the beamforming and frequency decomposition were factored into two separate layers in the network. However, these approaches estimated fixed filter coefficients for decoding, potentially limiting the ability of models to adapt to unseen environments. A neural network adaptive beamforming technique was proposed in [9] to address this issue in which the filter coefficients are the output of the beamforming neural network. To reduce the computational complexity, these approaches can be implemented efficiently in the frequency domain [10]. Instead of filtering in the time domain, Xiao et al. [11] estimated the parameters of the frequency-domain beamformer from a generalized cross correlation (GCC) [12] between microphones. However, it requires simulated data to train the beamformer part of neural networks in advance and then pretrains the acoustic model using the features generated by the beamformer part.

Over the past few years, some works have shown that performance can be improved by supplying complementary features as inputs to the network in parallel with the regular acoustic features for ASR. Seltzer et al. [13] have shown that augmenting the inputs of a neural network with an estimate of background noise can improve the robustness of the network to background noise. In the meanwhile, Saon et al. [14] augment DNN inputs with speaker i-vector features, and demonstrate significant improvement on the speech recognition task.

Motivated by the above work, we propose the idea that the generalized cross correlation between microphones is considered as input features to improve the performance of multichannel speech recognition. Acoustic signals from microphone arrays can be used to improve the robustness in distant speech recognition due to the availability of additional spatial information. Therefore, exploiting the additional spatial information from multiple microphones is essential for robust speech recognition in distant-talking scenarios. Moreover, the generalized cross correlation between microphones is one of the representations that encode spatial information and typically computed for localization. Consequently, we propose augmenting the traditional acoustic features from microphone arrays with the generalized cross correlation features between microphones. On the other hand, it has been shown that multitask learning (MTL) architecture improves the generalization performance of a learning task by jointly learning multiple related tasks together [15]. The model in MTL architecture is able to transfer knowledge to others by sharing some internal representations. Therefore, it is incorporated to further improve the robustness, and the model learns to classify the

This work is partially supported by the National Natural Science Foundation of China (Nos. 11590770-4, U1536117), the National Key Research and Development Plan (Nos. 2016YFB0801203, 2016YFB0801200).

feature into sensones and performs feature enhancement at the same time. The simultaneously recorded close-talk feature is used to be the second reference output.

Experiments on the AMI [16] and ICSI [17] meeting corpora show that our proposed approach achieves improvements over the model trained directly on the outputs of multiple microphones. In addition, we also compare our model with the beamforming technique [18], in which GCC-PHAT is also used to compute the TDOA. On the whole, our model achieves 6.3% and 3.6% relative improvements over the model trained directly on the concatenation of multiple microphone outputs on the AMI and ICSI evaluation set, respectively. Moreover, it also performs better than the beamforming baseline model. The rest of this paper is organized as follows. Section 2 describes our proposed model. The experimental setup is discussed in Section 3. Section 4 and Section 5 present the results and conclusions, respectively.

2. MODEL

2.1. Generalized cross correlation between microphones

Generalized cross correlations have been used successfully to determine the time delay of arrival (TDOA) of propagating waves between two spatially separated microphones. And TDOA estimated from multiple microphone pairs can be used to parameterize the source location [19, 20]. Hence, GCC actually encodes the location of the speaker. In this work, the generalized cross correlation between each microphone pair is computed using the generalized cross correlation method with phase transform (GCC-PHAT) [21], that is more robust to reverberation.

Given two channel signals $x_i(n)$ and $x_j(n)$, the GCC vectors are computed as follows:

$$gcc_{ij}(n) = IFFT\left(\frac{X_i(f)\,X_j^*(f)}{|X_i(f)\,X_j^*(f)|}\right) \tag{1}$$

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals, and $*$ denotes the complex conjugate. The TDOA for these two microphones is estimated as:

$$\hat{d}(i,j) = \underset{n}{\operatorname{argmax}}\ gcc_{ij}(n) \tag{2}$$

Ideally, $gcc_{ij}(n)$ should exhibit a peak over a restricted range, which corresponds to the TDOA between microphone i and j. And the separation distance of the microphones physically limits the range of valid time delays. The acoustic path length of each signal differs according to the location of the microphone and these differences in arrival time are even greater when the space between microphones is larger. This finite range is determined by the distance between the microphones divided by the speed of sound.

In this work, our models are trained and evaluated on the AMI and ICSI meeting corpora. AMI used an 8-microphone 10cm radius uniform circular array, and ICSI used 4 boundary microphones placed about 1m apart along the tabletop. The maximum distance between any pair of microphones in the AMI corpus is 20cm, and the maximum delay between two microphones is $\tau = 0.2m/340m/s = 0.588ms$. It corresponds to a less than 10 sample delay at a sample rate of 16kHZ. Therefore, the center 21 correlation coefficients for each microphone pair are sufficient to predict the location of the speaker. There are totally 28 microphone pairs in the 8-microphone array. On the whole, 588-dimensional GCC vectors are used as auxiliary features for the neural network acoustic model at each time step. It encapsulates the relevant information about the location of the speaker in a vector representation. Similarly, we adopted the center 281 coefficients for the ICSI data.

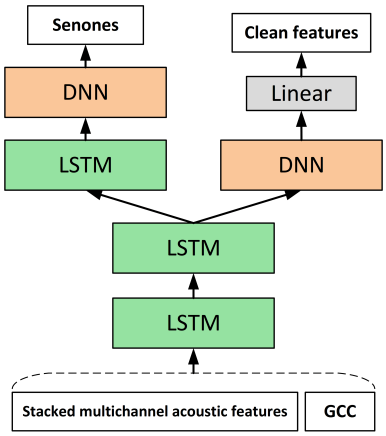


Fig. 1. Diagram of an LSTM acoustic model with augmented GCC inputs.

Figure 1 shows an overview of the proposed model. In this work, our model is evaluated on typical hybrid DNN-HMM frameworks, in which the acoustic model estimates context-dependent hidden Markov model (HMM) state posteriors. For both training and testing, the GCC features are concatenated to the acoustic features from the microphone array at each time step. Thus the neural network acoustic model is informed which speech segment comes from which location. Two sets of time-synchronous inputs should be created: one set of acoustic features which is the concatenation of the individual features from each microphone in the microphone array for phonetic discrimination and another set of GCC features that characterize the location of the speaker which provides the audio for the first set of features. The GCC features enhance the discrimination between different channels and enable the neural network acoustic model to make better use of acoustic signals from different channels.

2.2. Regularization with multitask learning

The multitask learning architecture is adopted to improve the robustness of our model. It is implemented by configuring the network with two outputs, one recognition output which predicts context-dependent states, and a second denoising output which reconstructs clean features derived from the close-talk speech. The MTL module branches off from the second LSTM layer of the acoustic model and is composed of one fully connected DNN layer followed by a linear output layer, as shown in Figure 1. In the recognition task, a discriminative model is learned to classify sensons by optimizing the cross-entropy (CE) criterion. Instead, the denoising model is optimized by minimizing the mean squared error (MSE). During training, the gradients back propagated from the two outputs are weighted by α and $1-\alpha$ for the recognition and denoising task respectively. The model parameters of the entire architecture are jointly learned to optimize the interpolated objective function

$$E(\theta) = \alpha E_{ce}(\theta) + (1-\alpha)\,E_{mse}(\theta) \tag{3}$$

The denoising output is only used in training to regularize the model parameters, and the associated layers are discarded during decoding.

3. EXPERIMENTAL SETUP

We perform experiments using the AMI and ICSI meeting corpora. The split recommended in the AMI corpus is used: a training set of 80 hours, a development set and a evaluation set, each of 9 hours. In the 72 hours ICSI corpus, 6 complete meetings are used for testing (Bed008, Bmr005, Bmr020, Bmr026, Bro015 and Bro016). Meeting speech recognition is characterised by speech overlap. For these experiments, the overlapping segments are not excluded from training sets, and results on the full set as well as the subset that only contains the non-overlapping segments are reported. The simultaneously recorded individual headset microphone (close-talk) data is used to be the second reference output. An interpolation weight $\alpha = 0.9$ is used to balance the two tasks. We use a 50000 word pronunciation dictionary [5] for the AMI and ICSI experiments. Two in-domain trigram language models are estimated using the AMI and ICSI training transcripts, respectively. They are further interpolated with the trigram language model estimated from Fisher transcripts.

Kaldi [22] is exploited for building speech recognition systems. The GMM model, which is used for generating the alignments to train the neural network acoustic model, is the same as that in [23]. 3 LSTM layers of 1024 memory cells with a 512-unit projection layer for dimensionality reduction [3] are used for acoustic modeling. In this work, 40-dimensional log-Mel filterbank features are extracted from every recording, and 5 frames (2 on each side of the current frame) of acoustic features are spliced as input for acoustic models to incorporate contextual information. During training, the network is unrolled for 20 time steps for training with truncated backpropagation through time (BPTT) and acoustic models are trained with cross-entropy (CE) criterion.

For comparison, the results of single distant microphone (SDM) and traditional beamforming are also shown. Experiments with SDM make use of the first microphone of the microphone array. For the beamforming experiments, the BeamformIt toolkit [18] is adopted to implement a weighted delay-and-sum beamforming, in which GCC-PHAT is used to compute the TDOA to create a single enhanced signal.

4. RESULTS

4.1. Analysis window size of GCC

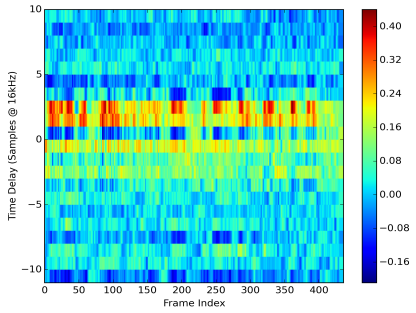
The computation of GCC between each microphone pair is repeated along the recording in order to respond to changes in the location of the speaker. And a big analysis window leads to a reduction in the resolution of changes in the location of speaker. On the other hand, using a very small analysis window reduces the robustness of the cross-correlation estimation, as less acoustic frames are used to compute it. Accordingly, there is a tradeoff between resolution and robustness. We begin by exploring the behavior of the proposed model as the analysis window size of GCC varies. To match the time-scale of acoustic features, GCC between microphones is also computed every 10ms.

The word error rate (WER) results for the AMI experiments are summarised in Table 1. It shows that we get improvements up to a window size of 105ms. It also can be seen that making the window size too large hurts performance because the estimation of GCC reduces responsiveness to changes in the location of speaker during an utterance. Thus an analysis window size of 105ms is used in the following experiments.

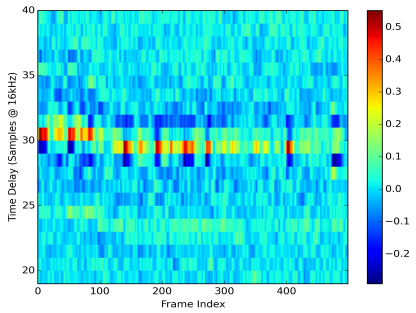
Figure 2 shows two examples of GCC features computed on a window size of 105ms between first two microphones for two utterances on the AMI and ICSI corpora. The vertical axis of this 2-D image plot is time-delay parameter, and the horizontal axis is the frame index of an utterance. The color of the image represents the amplitude of GCC. For comparison with AMI, the time-delay between 20 and 40 is ploted for ICSI, where the estimated TDOA is included. It can be observed that the delay values that correspond to these maxima on the vertical axis are the estimated TDOA .

Table 1. WER(%) for different window sizes on AMI.

Window size (ms)	25	55	75	105	155
dev	36.6	36.3	35.8	**35.7**	36.5
eval	41.7	41.2	40.5	**40.4**	41.5



(a) AMI



(b) ICSI

Fig. 2. Illustration of GCC features between first two microphones on AMI and ICSI.

4.2. Comparisons to baseline models

In this subsection we report on speech recognition experiments using the AMI and ICSI corpora. Three baseline models are considered: (1) training the LSTM acoustic model on the SDM data; (2) beamforming the multichannel signals into a single channel and following the standard acoustic modeling approaches used for the SDM case; (3) training the LSTM acoustic model directly on the concatenation of the individual 40-dimensional log-Mel filterbank features from the microphone array.

Since we have found that similar improvements were observed on both the development and evaluation set for the AMI dataset, results on the evaluation set are only reported for simplicity of exposition. Table 2 and 3 show the results for the AMI and ICSI corpora respectively. As expected, severe performance degradation was

observed with speech overlap. Compared with the results of SDM experiments, significant improvements were achieved by using multichannel data. It shows the benefit of additional spatial information in improving the performance of distant speech recognition.

Table 2. WER(%) on the AMI evaluation set.

Data	Model	with overlap	no overlap
SDM	-	48.4	39.8
MDM	beamformer	43.6	34.3
	8ch concatenated	42.7	34.8
	+ GCC-PHAT	40.4	33.0
	+ MTL	**40.0**	**32.2**

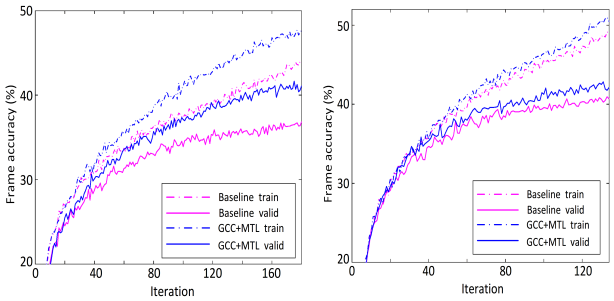
For the AMI experiments, the model trained on the beamformed signal performed slightly better than that trained directly utilising the multichannel features on the non-overlapping speech recognition task, but it showed a lower performance on all segments which include overlapping speech. That is probably because the competing acoustic source results in less accurate TDOA estimates. It indicates that using raw multiple input features in place of beamformed signal makes acoustic models learn better representations which take into account some factor such as speech overlap. We next evaluate the model in which GCC vectors are used as auxiliary features. In this case, the model obtained WERs of 40.4% and 33.0% on all segments and non-overlapping segments respectively, significantly outperforming the two MDM baseline models. Another 0.4-0.8% absolute reduction in WER was obtained by using multitask learning architecture. Compared with the multiple input baseline model, the proposed model achieved 6.3% and 7.4% relative improvements in WER for all segments and non-overlapping segments. In addition, it provided 8.2% and 6.1% relative improvements on all segments and non-overlapping segments over the beamforming baseline model. It is showed that our model performs well on both the overlapping and non-overlapping speech recognition task.

Table 3. WER(%) on the ICSI evaluation set.

Data	Model	with overlap	no overlap
SDM	-	40.1	34.7
MDM	beamformer	34.1	27.6
	4ch concatenated	30.3	25.7
	+ GCC-PHAT	29.6	24.7
	+ MTL	**29.2**	**24.2**

For the ICSI experiments, the multiple input baseline model obtained considerable improvements over the beamforming baseline model probably due to less accurate TDOA estimates from the microphone array which is characterised by large distances between microphones. And similar improvements were also obtained by the proposed model on the ICSI dataset. The results in Table 3 show that the model, in which the GCC vectors are considered as input features and MTL architecture is adopted, achieved 3.6% and 5.8% relative improvements on all segments and non-overlapping segments over the multiple input baseline model.

We observed that LSTM with GCC inputs were better than the ones trained on ASR features only. Small, but consistent reductions in WER can be further obtained by MTL architecture. These trends can be observed for both the AMI and ICSI dataset. On the whole, the GCC-augmented network outperforms the multiple input baseline model and beamforming baseline model. It suggests that additional spatial information is beneficial for the neural network acoustic model and could be utilized directly by the neural network. The GCC-augmented network could take advantage of the spatial information to improve performance for multichannel speech recognition. Moreover, consistent reductions in WER are further obtained by using multitask learning architecture. The denoising model, which performs feature enhancement, predicts clean features through the acoustic features from the microphone array and the GCC features between microphones. It shares hidden representations with the model that predicts acoustic states. The parameters of shared layers are regularized by the denoising model, which improves the robustness of the GCC-augmented network.



(a) AMI (b) ICSI

Fig. 3. Frame accuracy on the validation set and training set during training on AMI and ICSI.

Figure 3 shows the progress of validation set and training set frame accuracies during training for the multiple input baseline network and the model with GCC inputs and MTL architecture. The proposed model obtained improvements for frame accuracy on both the validation set and training set of the two datasets. It also suggests that our proposed model improves the ability to model the acoustic signal from the microphone array.

5. CONCLUSIONS

In this work, we proposed an architecture for multichannel speech recognition tasks, in which the GCC vectors between microphones are supplied as additional input features to acoustic models and multitask learning is employed to improve the robustness through regularizing the model parameters. The proposed model showed promising results on the AMI and ICSI meeting corpora. It performs well on both the overlapping and non-overlapping speech recognition task. Besides, beamforming computation is not required, speeding up the decoding process. In the future, other ways of training with GCC features for a streaming application will be investigated.

6. REFERENCES

[1] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks.," in *Interspeech*, 2011, pp. 437–440.

[2] Dong Yu, Wayne Xiong, Jasha Droppo, Andreas Stolcke, Guoli Ye, Jinyu Li, and Geoffrey Zweig, "Deep convolutional

neural networks with layer-wise context expansion and attention," in *Interspeech*, 2016.

[3] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014.

[4] Steve Renals, Thomas Hain, and Hervé Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, 2007, pp. 238–247.

[5] Thomas Hain, Lukáš Burget, John Dines, Philip N Garner, František Grézl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[6] Michael L Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, 2008, pp. 104–107.

[7] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Arun Narayanan, Michiel Bacchiani, et al., "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 30–36.

[8] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Arun Narayanan, and Michiel Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5075–5079.

[9] Bo Li, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech*, 2016.

[10] Tara Sainath, Arun Narayanan, Ron J. Weiss, Ehsan Variani, Kevin W. Wilson, Michiel Bacchiani, and Izhak Shafran, "Reducing the computational complexity of multimicrophone acoustic models with integrated feature extraction," in *Interspeech*, 2016, pp. 1971–1975.

[11] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5745–5749.

[12] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[13] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7398–7402.

[14] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 2013, pp. 55–59.

[15] Yu Zhang, Pengyuan Zhang, and Yonghong Yan, "Attention-based lstm with multitask learning for distant speech recognition," in *Interspeech*, 2017, pp. 3857–3861.

[16] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[17] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al., "The icsi meeting corpus," in *Acoustics, Speech, and Signal Processing (ICASSP), 2003 IEEE International Conference on*, 2003, vol. 1, pp. I–I.

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

[19] Ali Pourmohammad and Seyed Mohammad Ahadi, "Real time high accuracy 3-d phat-based sound source localization using a simple 4-microphone arrangement," *IEEE Systems Journal*, vol. 6, no. 3, pp. 455–468, 2012.

[20] Dongwen Ying and Yonghong Yan, "Robust and fast localization of single speech source using a planar array," *IEEE Signal Processing Letters*, vol. 20, no. 9, pp. 909–912, 2013.

[21] Michael S Brandstein and Harvey F Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, vol. 1, pp. 375–378.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[23] 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*, 2013, pp. 285–290.