LEARNING FEATURE MAPPING USING DEEP NEURAL NETWORK BOTTLENECK FEATURES FOR DISTANT LARGE VOCABULARY SPEECH RECOGNITION

Ivan Himawan[1], *Petr Motlicek*[1], *David Imseng*[1], *Blaise Potard*[1], *Namhoon Kim*[2], *Jaewon Lee*[2]

[1]Idiap Research Institute, Martigny, Switzerland
{ihimawan,motlicek,dimseng,bpotard}@idiap.ch
[2]Samsung Electronics Co. Ltd, Suwon, South Korea
{namhoon.kim,jwonlee}@samsung.com

ABSTRACT

Automatic speech recognition from distant microphones is a difficult task because recordings are affected by reverberation and background noise. First, the application of the deep neural network (DNN)/hidden Markov model (HMM) hybrid acoustic models for distant speech recognition task using AMI meeting corpus is investigated. This paper then proposes a feature transformation for removing reverberation and background noise artefacts from bottleneck features using DNN trained to learn the mapping between distant-talking speech features and close-talking speech bottleneck features. Experimental results on AMI meeting corpus reveal that the mismatch between close-talking and distant-talking conditions is largely reduced, with about 16% relative improvement over conventional bottleneck system (trained on close-talking speech). If the feature mapping is applied to close-talking speech, a minor degradation of 4% relative is observed.

Index Terms— Deep neural network, bottleneck features, distant speech recognition, meetings, AMI corpus

1. INTRODUCTION

Most of today's speech recognition applications require speakers to talk to a microphone located spatially close to the talker's mouth. These applications will benefit tremendously if the automatic speech recognition (ASR) performance can be improved for distant-talking microphones. For instance, applications such as meetings, multi-party teleconferencing, hands-free interfaces for controlling consumer-products will benefit from distant-talking operation without constraining the users of speaking closer to a microphone or wearing a headset microphone.

ASR from a distant microphone is a difficult task since the speech signals to be recognized are degraded by both interfering sounds (e.g., other speakers) and reverberation caused by the large speaker-to-microphone distance. Approaches to noise-robust speech recognition can generally be classified into two classes: front-end based and back-end based [1]. The front-end based approaches aim at removing distortions from the observations prior to recognition, and can either take place in time domain, spectral domain, or directly from the corrupted feature vectors [2, 3]. The back-end approaches on the other hand aim at adjusting parameters of the existing acoustic model to reduce the mismatch between the training and testing conditions [4].

The acoustic models based on DNN have recently been shown to significantly improve the ASR performance on variety of tasks compared to the state-of-the-art GMM/HMM systems. The DNN is a conventional multi-layer perceptron (MLP) with multiple hidden layers. A DNN architecture allows layers to be shared between tasks, while others are allocated to specific problems. This property was exploited in multilingual ASR tasks where the hidden layers are trained on multiple languages [5, 6]. In similar fashion, for noise robustness, the DNN-based acoustic models can be trained using multi-condition data (e.g., clean and noisy speech) via multistyle training to improve ASR accuracy when dealing with different acoustic channels and various environmental noises [7, 8].

This paper investigates the use of DNN-based acoustic modeling for distant speech recognition in the context of a meeting recognition task using AMI corpus [9]. The objective is to study how deep architectures can reduce the mismatch between systems trained on clean speech from close-talking microphones (also called individual head microphone (IHM)) and noisy and reverberant speech from single distant microphone (SDM) (i.e., to improve the distant ASR performance by also using IHM data). We investigate using condition specific layers on DNN/HMMs systems (e.g., the softmax layers are made to be a channel specific), similar to what was employed in multilingual speech recognition tasks for adapting existing multilingual DNN for a new language [5, 6]. We then propose to transform the reverberant speech to a feature space close to clean speech where DNN is used to learn the mapping between the two conditions.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the experimental setup followed by meeting recognition and model adaptation experiments. The proposed approach is presented in Section 4. Results are discussed in Section 5 and the study is concluded in Section 6.

2. RELATED WORK

The popular method for adapting hybrid ASR systems is to augment the existing network with an extra input layer with a linear activation function (as in the case of linear input network (LIN) [10]) to transform the testing features into the training feature space before forwarding them to the original network. The transformation layer could also be employed before the final output activation functions (i.e., softmax) as in the case of linear output network (LON) approach [11]. Different multi-condition learning architectures within the context of the DNN-based acoustic modelling framework to ex-

This work was supported by Samsung Electronics Co. Ltd, South Korea, under the project "Acoustic Model Adaptation toward Spontaneous Speech and Environment". This work was also partially supported by Speaker Identification Integrated Project (SIIP), funded by the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607784.

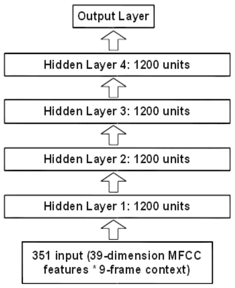
Output Layer
Hidden Layer 4: 1200 units
Hidden Layer 3: 1200 units
Hidden Layer 2: 1200 units
Hidden Layer 1: 1200 units
351 input (39-dimension MFCC features * 9-frame context)

Fig. 1. DNN architecture with four hidden layers.

plicitly model different conditions were explored in [12]. For example, the channel specific layer can be trained for model adaptation while keeping the top layers fixed, already trained with mixed channels data. This was shown to reduce word error rate (WER) over the baseline multi-condition model. Another alternative is to model the phone posterior probabilities given the speech observation and the acoustic scene, where both are added to the neural network as an input such as noise aware training in [7]. This work investigates ASR performance when condition specific layer is trained for model adaptation on DNN/HMMs systems. Rather than inserting an extra transformation layer, the input or output layer from the existing DNN/HMMs models is adapted using a channel specific data.

Motivated by a recent work which uses DNN as an inverse filter for dereveberation in [13, 14], this work proposes supervised learning approach to learn the mapping between features affected by reverberation and its clean version, where the clean speech features are extracted from the bottleneck layer. We hypothesize that the DNNs could be used as nonlinear transformation layers to transform features from one condition to another (e.g., reverberant to anechoic). Denoising input features rather than bottleneck features were proposed recently in [14], but the the noisy speech are obtained by corrupting the clean speech with different type of additive noises and channel distortions rather than true distant microphone recordings. We use bottleneck features since ASR accuracy increases when using context-dependent target labels for GMM/HMM training [15]. The bottleneck features used in combination with traditional features such as MFCCs or PLPs as input features to ASR systems was shown to capture information that is complementary to conventional features derived from the short-time spectra [15, 16].

3. EXPERIMENTS

3.1. Data and system setup

Experiments in this paper used AMI corpus which contain meetings recorded in equipped instrumented meeting rooms at three sites in Europe (Edinburgh, IDIAP, TNO) [9]. The meetings captured natural spontaneous conversations between participants who play different roles in a fictitious design teams (i.e., project manager, designer) for scenario meetings, as well as non-scenario meetings in a range of topics. Perfectly synchronized (i.e., on a frame-level) recording devices include individual head microphones (IHM), lapel microphones, and one or more microphone arrays. For distant speech recognition experiments in this work, the single distant microphone (SDM) of the first microphone of the primary array is used.

The ASR experiments employ both headset recordings (IHMs) and their corresponding distant microphones (SDMs). The speech makes about 67 hours for each audio stream (after performing voice activity detection) available for training, and holds around 7 hours for evaluation sets. The experiments use the suggested AMI corpus partitions for training and evaluation sets [17], even though some of the meeting recordings were discarded from the original corpus when arrays recordings were missing to ensure both headset and the corresponding synchronized array recordings are available for training and testing. This work considers the overlapping speech segments. Preliminary experimental results reveal that training and test on non-overlapping speech reduces WERs.

Table 1. WERs[%]: Results using MFCC features with GMM/HMM systems on the IHM and SDM test sets.

	Test sets	
Trained on	IHM	SDM
IHM	42.8	89.2
SDM	85.4	71.5
WSJ	58.3	91.5

Table 2. WERs[%]: Results using MFCC features with DNN/HMM systems on the IHM and SDM test sets. The DNNs are trained using four hidden layers (Figure 1).

	Test sets	
Trained on	IHM	SDM
IHM	32.3	76.0
SDM	49.7	61.4
SDM (alignment from IHM)	46.7	58.0
WSJ	52.2	83.9

For both IHM and SDM configurations, the baseline GMM/HMM systems are trained on 39-dimensional MFCC features including their delta and acceleration versions. The acoustic models for the GMM/HMM systems have the number of tied-states roughly 4K in both IHM and SDM configurations, and each of the GMM/HMM system have a total of 120K Gaussians. The state alignments for training the DNNs are obtained from the GMM/HMM systems.

The Kaldi toolkit is used for training both GMM/HMM and DNN/HMM systems [18]. The DNNs were trained using 9-frame temporal context, employing four 1200-neuron hidden layers, and with cepstral mean and variance normalized per speaker. Figure 1 shows DNN/HMM architecture used in this study. The AMI pronunciation dictionary of approximately 23K words is used in the experiments, and the Viterbi decoding is performed using a 2-gram language model, previously built for NIST RT'07 corpora [9].

3.2. Distant speech recognition on matched and mismatched conditions

Tables 1 and 2 show the WER results of MFCC features on both IHM and SDM test sets for GMM/HMM and DNN/HMM systems respectively. The performance of GMM/HMMs systems trained and tested on the matched condition perform significantly better than those trained and tested on mismatched condition. In case of DNN/HMM, the system yields about 10% absolute WER reduction over the GMM/HMM for both IHM and SDM models tested on matched condition. On the mismatched condition, for a model trained on IHM and tested on SDM, the ASR performance is significantly reduced (by about 44% absolute WER). It is surprising that for DNN/HMM model trained on SDM (that is noisy and reverberant speech) performs better in mismatched condition (recognizing

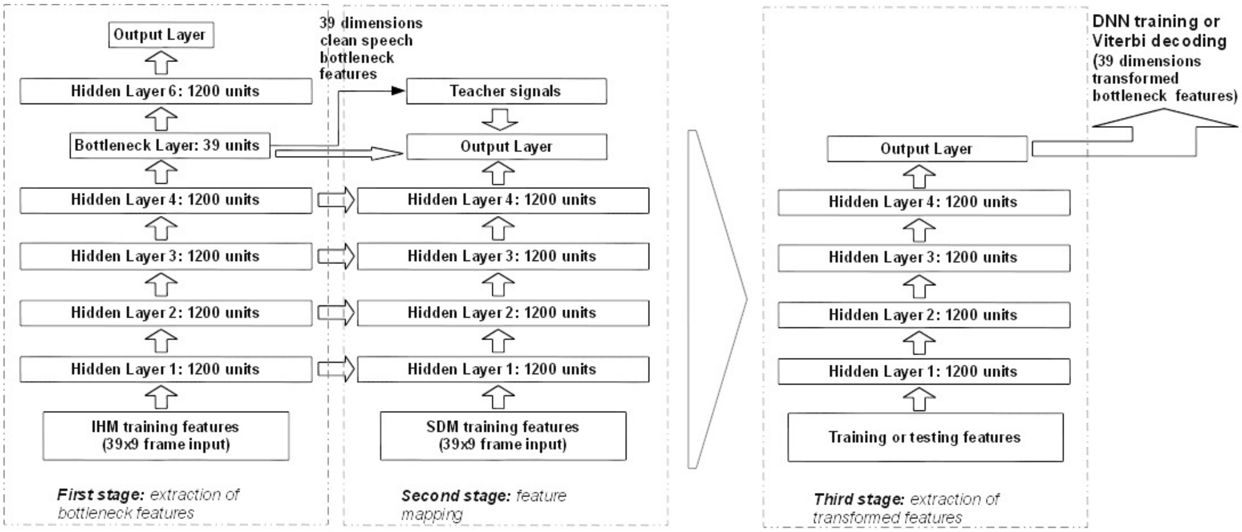
Output Layer
Hidden Layer 6: 1200 units
Bottleneck Layer: 39 units
Hidden Layer 4: 1200 units
Hidden Layer 3: 1200 units
Hidden Layer 2: 1200 units
Hidden Layer 1: 1200 units
IHM training features (39x9 frame input)
First stage: extraction of bottleneck features
39 dimensions clean speech bottleneck features
Teacher signals
Output Layer
Hidden Layer 4: 1200 units
Hidden Layer 3: 1200 units
Hidden Layer 2: 1200 units
Hidden Layer 1: 1200 units
SDM training features (39x9 frame input)
Second stage: feature mapping
DNN training or Viterbi decoding
(39 dimensions transformed bottleneck features)
Output Layer
Hidden Layer 4: 1200 units
Hidden Layer 3: 1200 units
Hidden Layer 2: 1200 units
Hidden Layer 1: 1200 units
Training or testing features
Third stage: extraction of transformed features

Fig. 2. Flowchart of creating the transformed bottleneck features.

Table 3. WERs[%]: Performance of the adaptation on DNN/HMM systems with condition specific layers.

	Test sets	
Trained on	IHM	SDM
IHM, adapt output layer on SDM	43.7	57.0
IHM, adapt input layer on SDM	44.9	58.1

IHM) compared to the same condition. It also achieves a much lower WER compared to GMM/HMM system (degradation from 85.4% to 49.7%). To the authors's knowledge, these results have not been reported before. Using IHM for HMM state alignment further improve the performance of DNN/HMM system trained on SDM data [19].

To clarify these results further, standard Wall Street Journal (WSJ) corpus (about 81 hours of read-speech from a close-talking microphone) was used for the training while an evaluation was done using IHM and SDM test sets. We found similar trends with either GMM/HMM or DNN/HMM models trained on AMI IHM data. In the case of GMM/HMM, since features are directly modelled by a mixture of Gaussians, the Gaussians simply model the additional variability introduced by the noise. Hence, we assume that GMM/HMM trained with noisy data will not perform well for classifying clean speech. In discriminative training, DNN may be able to extract some pertinent information for classification through multiple processing of nonlinear layers with noisy speech as an input [7]. Based on these findings, using the SDM model for improving ASR recognition on clean and reverberant conditions may be a promising research direction in the future. The current efforts however are focused on improving distant speech recognition using IHM model.

3.3. Model adaptation using condition specific layers

These experiments attempt to reduce the performance gap between the DNN system trained and tested on mismatch conditions by performing model adaptation using condition specific DNN layer. We borrowed a similar idea from multilingual DNN works in which the hidden layers are shared while the softmax layers could be made language specific [6].

Starting with the existing model trained on IHM data, we replace the output layer with a new layer and initialise the $\mathbf{W}_L$ which is the matrix connection weights between $L-1-th$ layer and output layer L with random weights, and then retrain the network with SDM data (with IHM used for the HMM state alignment). Another alternative is to place the condition specific layer in the input layer of the network. In similar fashion, we replace the input layer with a new layer that is initialised with random weights, and then retrain the network with SDM data (with IHM used for the HMM state alignment). The evaluation of the resulting models using the condition specific layers on the matched and mismatched test sets are shown in Table 3.

Adapting the output layer gives better performance on both matched and mismatched conditions rather than adapting the input layer. We obtain 19% absolute WER reduction (from 76.0% to 57.0%) with respect to the baseline DNN/HMM when tested on SDM data, but the performance decreases by 11.4% absolute WER (from 32.3% to 43.7%) when tested on IHM data as shown in Tables 2 and 3.

4. NEURAL NETWORK BASED FEATURE MAPPING

A recent study used DNNs to learn a spectral mapping from the reverberant speech to the anechoic speech in order to reconstruct the raw clean data using the noisy data [13]. Using a similar idea, to reduce the mismatch between clean and reverberant conditions, this paper proposes a feature based transformation using DNN by learning the mapping between reverberant speech feature and clean speech bottleneck feature to improve the distant ASR performance. The proposed approach is presented in three stages as shown in Figure 2 and explained in each subsection below.

4.1. First stage: extraction of bottleneck features

The bottleneck features are extracted from the bottleneck layer of a DNN structure. There are 8 layers in total and 6 of them are hidden including the bottleneck layer. The bottleneck layer is placed just between the 5^{th} and the 7^{th} layer, and has 39 dimensions with

Table 4. WERs[%]: Results using transformed feature space with DNN/HMM systems on the matched and mismatched conditions.

	Test sets	
Trained on	IHM	SDM
IHM	33.8	63.4
IHM, adapt output layer on SDM	40.6	56.7
SDM (alignment from IHM)	41.8	57.3

linear activation functions. The output layer has 4K output units for the IHM configuration. The bottleneck structure is shown on the left of Figure 2, and the activations of the units yield the bottleneck features.

4.2. Second stage: feature mapping between SDM features and IHM bottleneck features

The bottleneck features extracted from the DNN are then used as a teacher input for a nonlinear feature transformation of distant-talking speech input (towards SDM data). The bottleneck features trained from IHM data is used to transform the reverberant speech features to a new feature space close to clean speech features. The original DNN network for training the bottleneck layer in the previous step was cut up to the bottleneck layer, and then used to learn the mapping between the SDM data as an input and the IHM bottleneck features as the teacher signal. To learn this mapping, the network (as shown in the middle of Figure 2) is trained using the standard error backpropagation procedure and the optimization is done through stochastic gradient descent by minimizing a mean square error cost function (i.e., not cross-entropy, since the target features are floating points and not posterior probabilities).

4.3. Third stage: extraction of transformed features

Once the mapping is completed, the trained network structure as shown on the right side of Figure 2 is used to generate new features from the activations of the units of the output layer for training new acoustic model. This yields 39-dimensional transformed features (to compare with 39 MFCCs) for subsequent DNN training with four hidden layers similar to what is illustrated in Figure 1. Note that, unlike systems in [15, 16], we do not train GMM/HMM models using single pass retraining on the bottleneck features. For recognition using Viterbi decoding, test sets for both IHM and SDM are fed to the trained network and the transformed features are extracted from the activations of the output layer.

4.4. Experimental Results

The baseline performance from DNN/HMM system with bottleneck layer trained on IHM data in the first stage network achieves 32.6% and 75.8% WERs when tested on IHM and SDM test sets respectively. These results are roughly comparable (less than 0.5% WER difference) to what is obtained using the typical DNN/HMM system with four hidden layers as shown previously in Table 2.

Table 4 shows results of DNN/HMM systems trained on transformed training features for training IHM and SDM models. Note that, for training SDM model, IHM is used for generating the HMM state alignment. Furthermore, we perform an adaptation on IHM model by replacing the output layer with a new layer and then retrain the network using transformed SDM features as training data and IHM is used for generating the HMM state alignment.

5. DISCUSSION

Supervised learning for mapping the SDM features to the IHM bottleneck feature space using DNN is effective for improving the distant speech recognition, when trained on IHM data. Compared to the baseline DNN/HMM performance (trained with bottleneck layer), on the mismatched condition (recognizing SDM), about 12% absolute (16% relative) WER reduction is achieved. A minor degradation of about 1% absolute (4% relative) is observed on the matched condition. Table 2 shows that retraining the model with SDM condition output layer degrades the performance on the IHM condition but improves on the SDM condition as expected. Using trained model from the proposed approach, we achieve better performance compared to condition specific layer experiment in Table 3 by 3.1% absolute WER (from 43.7% to 40.6%) on the matched condition and by 0.3% absolute WER (from 57.0% to 56.7%) on the mismatched condition.

Training SDM model with transformed SDM features performs better on matched and mismatched conditions compared to SDM model trained with MFCCs and IHM used for state alignment. Large improvement is observed on mismatched condition compared to the baseline DNN/HMM performance in Table 2, 4.9% absolute (10.5% relative) WER reduction (from 46.7% to 41.8%) is achieved. This suggests that the SDM features have more discriminant classification ability close to IHM condition after being transformed by the trained network.

Other approaches to noise robustness include multi-style training. The preliminary results show that training on 134 hours of combined IHM and SDM data yielded WERs of 33.3% and 59.5% for decoding IHM and SDM test sets respectively. Compared to the proposed approach, no performance gain is observed on the IHM test. When decoding SDM, the proposed approach performs better by 2.9 % absolute WER (59.5% compared to 56.7%) if output layer is adapted to SDM condition. The performance obtained by the multi-condition model is promising, and our preliminary findings using the SDM data (with IHM is used for HMM state alignment) to construct the deep bottleneck feature based DNN system achived WERs of 36.8% and 56.8% for decoding IHM and SDM test sets respectively. This suggests that features extracted from multi-condition system is inherently robust to noise.

6. CONCLUSIONS

This paper studied the problem of distant ASR task using DNN on the AMI meeting corpus by first conducting recognition on the matched and mismatched conditions. The performance on DNN/HMM systems are significantly better than GMM/HMM systems. The DNN/HMM model trained on the SDM performs better on IHM test compared to the matched condition. This result is suprising and we will investigate this further in future. Adapting IHM model with SDM data using condition specific layers degrades the performance on IHM but improves on SDM condition. Adapting the output rather than the input layer to be a channel specific gives better results on matched and mismatched conditions.

Utilizing DNN to learn the mapping between SDM and IHM condition is shown to be effective for improving distant speech recognition using IHM model on the AMI meeting corpus. Using the condition specific layers on the transformed model gives further improvement on both matched and mismatched conditions. In future, we plan to experiment the proposed approach with multiple distant microphones and to study their impact on WER. Experiments with other distant speech recognition tasks are also planned.

7. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, Nov. 2012.

[2] Dong Yu et al., "Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 1061–1070, 2008.

[3] Takuya Yoshioka, Xie Chen, and Mark J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[4] C. Legetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[5] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[6] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[7] Michael L. Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[8] Chao Weng, Dong Yu, Michael L. Seltzer, and Jasha Droppo, "Single-Channel Mixed Speech Recognition using Deep Neural Network," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[9] Thomas Hain et al., "Transcribing Meetings With the AMIDA Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[10] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *Proceedings Eurospeech*, 1995, pp. 2183–2186.

[11] Bo Li and Khe Chai Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Interspeech*, 2010.

[12] Yan Huang, M. Slaney, M. L. Seltzer, and Y. Gong, "Towards Better Performance with Heterogeneous Training Data in Acoustic Modeling using Deep Neural Networks," in *Interspeech*, 2014.

[13] Kun Han, Yuxuan Wang, and DeLiang Wang, "Learning Spectral Mapping for Speeech Dereverberation," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[14] Jun Du and Qing Wang and Tian Gao and Yong Xu and Lirong Dai and Chin-Hui Lee, "Robust Speech Recognition with Speech Enhanced Deep Neural Networks," in *in Proceedings of Interspeech*, 2014.

[15] Dong Yu and Michael L. Seltzer, "Improved Bottleneck Features Using Pretrained Deep Neural Networks," in *Interspeech*, 2011.

[16] Yulan Liu, Pengyuan Zhang, and Thomas Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[17] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid Acoustic Models for Distant and Multichannel Large Vocabulary Speech Recognition," in *Automatic Speech Recognition and Understanding*, 2013.

[18] Daniel Povey et al., "The Kaldi speech recognition toolkit," in *Automatic Speech Recognition and Understanding*, 2011.

[19] Marc Delcroix, Yotaro Kubo, Tomohiro Nakatani, and Atsushi Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modelling?," in *Interspeech*, 2013.