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

Attention-based LSTM with Multi-task Learning for Distant Speech Recognition

Yu Zhang[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
{zhangyu,zhangpengyuan,yanyonghong}@hccl.ioa.ac.cn

Abstract

Distant speech recognition is a highly challenging task due to background noise, reverberation, and speech overlap. Recently, there has been an increasing focus on attention mechanism. In this paper, we explore the attention mechanism embedded within the long short-term memory (LSTM) based acoustic model for large vocabulary distant speech recognition, trained using speech recorded from a single distant microphone (SDM) and multiple distant microphones (MDM). Furthermore, multi-task learning architecture is incorporated to improve robustness in which the network is trained to perform both a primary senone classification task and a secondary feature enhancement task. Experiments were conducted on the AMI meeting corpus. On average our model achieved 3.3% and 5.0% relative improvements in word error rate (WER) over the LSTM baseline model in the SDM and MDM cases, respectively. In addition, the model provided between a 2-4% absolute WER reduction compared to a conventional pipeline of independent processing stage on the MDM task.

Index Terms: distant speech recognition, long short-term memory, attention, multi-task learning

1. Introduction

Deep neural networks (DNNs) acoustic models [1, 2, 3] have driven tremendous improvements in large vocabulary continuous speech recognition in recent years. Further improvements are achieved by using more advanced models such as convolutional neural networks (CNNs) [4] and long short-term memory based recurrent neural networks (LSTM RNNs) [5]. Although these new techniques decrease the word error rate (WER) on distant speech recognition, performance in distant talking scenarios is still far-behind their close-talking equivalents [6]. Distant speech recognition remains a challenging task owning to background noise, reverberation, and speech overlap.

Distant speech recognition systems are usually configured to record audio data using a single distant microphone (SDM) or multiple distant microphones (MDM). It has been shown that the MDM system performs better than the SDM counterpart in terms of accuracy since it considers the additional spatial information. Many distant speech recognition systems have adopted a two-part architecture where signal processing techniques are applied to enhance the speech before being further processed by conventional acoustic modeling approaches [7]. Since the signal processing part is usually distinct from the speech recognition part, it fails to optimize towards the final objective (speech recognition accuracy), which leads to a suboptimal solution [8].

To obtain an optimal performance, joint training of speech enhancement and acoustic model were proposed to improve distant speech recognition accuracy [9, 10, 11, 12]. These approaches, however, has high computational complexity due to a complex network for beamforming. Some studies [13, 14, 15] have shown that the neural networks can learn discriminative representations of acoustic features from the simple concatenation of raw feature vectors. However, [5] proposed a deep CNNs acoustic model which introduced location-based attention by weighting the contribution from each frame according to their distance to the current frame. Kim et al. applied similar idea to distant speech recognition with multiple microphones where a neural attention network is adopted to combine temporal and spatial information of audio to predict acoustic states without explicit preprocessing for speech enhancement [16]. It is shown that the attention mechanism could automatically tune its attention to a more reliable input source, which brings a improvement in accuracy. However, the architecture proposed in [16] is limited to a small dataset which is CHiME-3 task with only 3 hours real data and 15 hours simulated data and does not work well on a larger dataset in our experiments, such as AMI meeting corpus [17].

In this paper, we propose a novel ALSTM-MTL model for distant speech recognition. Attention mechanism across time is embedded to capture the temporal information at input layer. And multi-task learning (MTL) architecture is incorporated to improve robustness, where the network learns to classify the observations into sensones and performs feature enhancement at the same time. Consistent reductions in WER were obtained by using MTL in both the SDM and the MDM case. The proposed architecture has better scalability on large dataset when compared with work [16]. Experiments show that our proposed architecture achieves a 3.3% relative improvement in word error rate over the LSTM baseline model in the SDM case. In the experiments with multiple microphones, our model provided significant improvements over the beamforming baseline on the overlapping speech recognition task. It indicates that using raw multiple channel input features in place of beamformed signal makes acoustic model learn better representations which take into account some factor such as speech overlap. On the whole, we achieved a 5.0% relative improvement in WER over the LSTM baseline model trained on the raw multiple inputs in the MDM case.

The rest of this paper is organized as follows. Section 2 briefly introduces LSTM acoustic model which is used as baseline in our experiments. Section 3 describes our proposed model. The experimental setup is discussed in Section 4. Section 5 and Section 6 present the results and conclusions, respectively.

http://dx.doi.org/10.21437/Interspeech.2017-805

2. LSTM-HMM hybrids for speech recognition

In a neural network-hidden Markov model (HMM) hybrid system, the neural network is trained to classify input features into classes corresponding to HMM states. After training, the output of the neural network is an estimate of the posterior probability. With the development of deep learning, more complex neural networks have been proposed. Dramatic improvements in acoustic modeling are achieved by LSTM due to its ability to handle long-term dependencies. Therefore, the baseline speech recognition system in our experiment uses LSTM based acoustic model. Several variants of the LSTM architecture for RNNs have been proposed. The LSTM architecture used in our work is described here briefly.

The LSTM contains memory blocks in the recurrent hidden layer, and each block has memory cells with three gates to control the flow of information. In addition, peephole connections from its internal cells to the gates in the same cell are contained to learn precise timing of the outputs. Given the input sequence $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$, the LSTM layer computes the gates and memory cells activation sequentially from $t = 1$ to T. The computation at the time step t can be formally written as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_{t-1} + b_o) \quad (4)$$

$$h_t = o_t \odot \phi(c_t) \quad (5)$$

where i_t, o_t, f_t, c_t are the outputs of the input gate, output gate, forget gate and memory cells respectively. The $W_{.x}$ weight matrices connect the inputs with the memory blocks, and the $W_{.h}$ matrices connect the previous hidden output with the memory blocks. The $W_{.c}$ terms are diagonal weight matrices for peephole connections. σ is the logistic sigmoid nonlinearity, ϕ is the hyperbolic tangent nonlinearity, and $\odot$ is the element-wise product of the vectors. Finally, the hidden output h_t from the current layer are treated as input into the next recurrent layer.

3. Attention-based LSTM with multi-task learning

The proposed attention-based LSTM model with multi-task learning (ALSTM-MTL) is depicted in Figure 1. The attention mechanism and multi-task learning architecture used are described, respectively.

3.1. Attention-based LSTM

Attention mechanism is a mechanism that the model iteratively processes its input by selecting relevant content at every time step, rather than a specific model implementation. The input of traditional neural network acoustic model $\boldsymbol{x}_t$ at time t is formed from a contextual window of L frames, which results in that the temporal information within L frames is ignored. However, the contribution from each frame at the input layer to the state prediction should be different. In this paper, attention mechanism

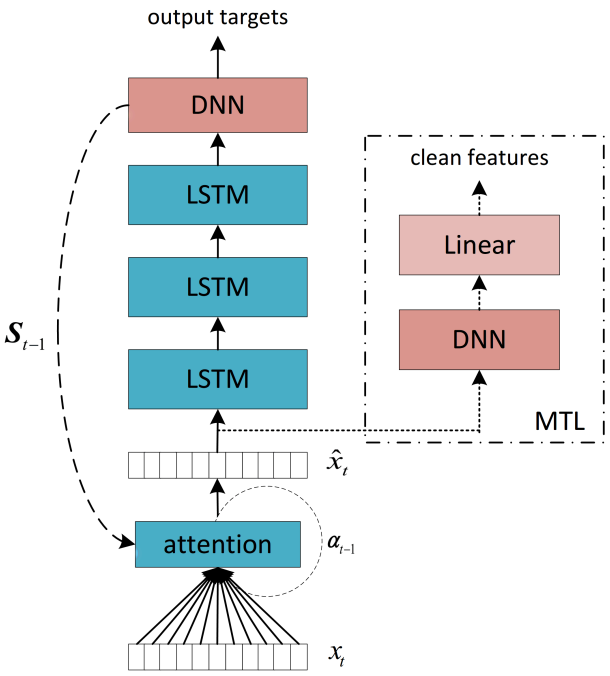

Figure 1: *Attention-based LSTM model with multi-task learning.*

across time is embedded to learn to focus more attention to more relevant frames at each time step. Therefore, the temporal information at input layer is captured by the attention mechanism.

As is shown in Figure 1, the weighted representation $\hat{\boldsymbol{x}}_t$, which is the input of traditional LSTM acoustic model to estimate the probability of context-dependent HMM state $p(\boldsymbol{s}|\boldsymbol{x}_t)$, is generated by scaling input $\boldsymbol{x}_t$ with attention weights $\boldsymbol{\alpha}_t$. And the attention weights $\boldsymbol{\alpha}_t$ enable the LSTM model to tune its attention to the input frames at each time step. The attention-based LSTM model in this work can be described by the following equations:

$$\boldsymbol{e}_t = Attend(\boldsymbol{x}_t, \boldsymbol{s}_{t-1}, \boldsymbol{\alpha}_{t-1}) \quad (6)$$

$$\alpha_{tl} = \frac{\exp(e_{tl})}{\sum_{l=1}^{L} \exp(e_{tl})} \quad (7)$$

$$\hat{\boldsymbol{x}}_{tl} = \alpha_{tl}\boldsymbol{x}_{tl} \quad (8)$$

$$p(\boldsymbol{s}|\boldsymbol{x}_t) = LSTM(\hat{\boldsymbol{x}}_t) \quad (9)$$

where $Attend(.)$ is a feed forward neural network that computes the attention scores $\boldsymbol{e}_t$, $LSTM(.)$ stands for LSTM model that predicts state labels. As illustrated by equation (6), the attention scores $\boldsymbol{e}_t$ depend on the input $\boldsymbol{x}_t$, the prediction from previous frame $\boldsymbol{s}_{t-1}$ and the attention weights history $\boldsymbol{\alpha}_{t-1}$. Equation (7) shows how the attention weights α_{tl} are obtained by normalizing the attention scores e_{tl}. As shown in equation (8), the weighted representation $\hat{\boldsymbol{x}}_t$ is generated by scaling each frame x_{tl} in $\boldsymbol{x}_t$ with the attention weights α_{tl}. Instead of the raw input $\boldsymbol{x}_t$, the weighted representation $\hat{\boldsymbol{x}}_t$ is regarded as input to the following LSTM acoustic model.

3.2. Joint training with multi-task learning

For speech recognition task, typical neural network acoustic model is trained by optimizing one criterion, such as cross-entropy (CE), state minimum Bayes risk (sMBR) and maximum mutual information (MMI). However, neural network model in multi-task learning framework jointly optimize more than one criterion. For example, [18] proposed to train acoustic models that jointly predict context-dependent and monophone targets. In [19], multilingual systems were trained to optimize for several languages simultaneously. Multi-task learning architecture aims at improving the generalization performance of a learning task by jointly learning multiple related tasks together. The model in MTL architecture is able to transfer knowledge to others by sharing some internal representations.

Noise robustness is always one of the critical issues for speech recognition task. Joint training of feature mapping and acoustic model was proposed in [20] for robust speech recognition. The feature mapping network was used to map the input noisy features to the desired clean acoustic features. In the distant talking scenarios, the speech signal is susceptible to distortion from noise. Thus, feature enhancement is used as the secondary task to improve the performance of acoustic model in the MTL architecture.

There are two outputs in the network, one recognition output which predicts context-dependent states and a second denoising output which reconstructs 40-dimensional filter-bank features derived from the close-talk speech. The denoising output is only used during training stage to regularize the model parameters and the associated layers are discarded during decoding. The MTL module 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 CE criterion. Instead, the denoising model is optimized by minimizing the mean squared error (MSE) between the DNN outputs $\hat{\boldsymbol{y}}_t$ and the referenced close-talk features $\boldsymbol{y}_t$. During training, the gradients back propagated from the two outputs are weighted by β and $1-\beta$ for the recognition and denoising task respectively. The model parameters of the entire architecture are jointly learned to optimize the interpolated objective function

$$E = \beta \sum_t p\left(\boldsymbol{s}|\boldsymbol{x}_t\right) + (1-\beta) \sum_t \left(\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t\right)^2 \qquad (10)$$

where the weight parameter β determines how much importance the secondary task should get.

4. Experimental Setup

We evaluated our models on the AMI corpus which contains around 100 hours of meetings recorded in specifically instrumented meeting room at three sites in Europe. Acoustic signal is captured by multiple microphones including individual head microphones, lapel microphones, and one or more microphone arrays. Each recording site uses a primary 8-microphone uniform circular array of 10cm radium. In this work the primary microphone array data is used, which is referred to as MDM in the following. Experiments with SDM make use of first microphone of the primary array. And the simultaneously recorded individual headset microphone (IHM, close-talk) data is used to be the second output in the multi-task learning module. Our models are trained and tested using the split recommended in the corpus release: a training set of 80 hours, a development set and a evaluation set each of 9 hours. The overlapping speech segments were not excluded during training.

In this work, we exploited Kaldi [21] for building speech recognition systems. The HMM-GMM system, which is used for generating the alignments to train the neural network, is as described in [6]. In the experiments, we used alignments from IHM data, since training acoustic models with alignments generated from the parallel close-talk microphone data provides significant improvements [22]. The LSTM baseline has 3 hidden layers with 1024 memory cells in each layer. The networks are trained using the stochastic gradient descent (SGD) based truncated back propagation through time (BPTT) algorithm. Each BPTT segment contains 20 frames and 100 segments are processed in a minibatch. The models were trained on 40-dimensional log Mel filterbank features in the SDM case. For the MDM case, the concatenation of the individual 40-dimensional log Mel filterbank features from 8 microphones at each time step was considered as a single input frame. An interpolation weight β = 0.9 is used to balance the two objectives.

5. Results

In this section we report results on speech recognition experiments using the AMI corpus with two distant speech recognition cases (SDM and MDM). The performance of models are evaluated using WER. Results on the development set and the evaluation set are reported. Since we do not exclude the overlapping segments during training stage, we show results on the full set as well as the subset that only contains the non-overlapping speech segments. The subset will be referred to as *dev** and *eval** in the following experiments.

5.1. Number of input frames

We begin by exploring the effect of input frame number on the attention-based LSTM model without MTL architecture. Table 1 reports the performance of the baseline 3-layer LSTM model and the attention-based LSTM model with different lengths of the input context. The attention-based LSTM model is denoted as ALSTM.

Table 1: *Performance at different number of input frames (WER,%)*

LVCSR task	Context	LSTM		ALSTM	
		dev	*eval*	*dev*	*eval*
SDM	[-3, 3]	43.0	47.5	43.0	47.6
	[-5, 5]	42.8	47.2	**41.7**	**46.2**
	[-7, 7]	43.1	47.3	42.1	46.7
MDM	[-3, 3]	37.5	42.4	36.7	41.7
	[-5, 5]	37.8	42.7	**36.0**	**41.4**
	[-7, 7]	38.0	43.3	36.4	41.5

The configuration in the second column of Table 1 stands for spliced context. For instance, splicing together frames from $t-3$ to $t+3$ at the input layer is written compactly as [-3, 3]. From Table 1, it can be seen that [-5, 5] is the optimal temporal context for the attention-based LSTM model in both the SDM and MDM case. The attention-based models achieve more than 1% absolute reduction in WER compared to the LSTM baseline models. This indicates that 11 frames are sufficient for the attention mechanism. Thus a window of 11 frames of input features are used in the following experiments.

5.2. Single Distant Microphone

The multi-task learning architecture is adopted to improve robustness by training part of the network to reconstruct 40-dimensional clean filter-bank features as a secondary objective to the primary task of context-dependent state prediction. The proposed model is denoted as ALSTM-MTL in the table. Our results on the SDM experiments are shown in Table 2.

Table 2: *Performance comparison on the development and evaluation set in the SDM case (WER,%)*

Model	*dev*	*dev**	*eval*	*eval**
LSTM	42.8	34.3	47.2	38.3
ALSTM	41.7	33.6	46.2	37.6
ALSTM-MTL	**41.3**	**33.1**	**45.8**	**37.2**

As shown, severe performance degradation can be observed with speech overlap. We see a 8-9% reduction in WER when only considering segments with non-overlapping speech. As expected, on average attention-based model achieved 2.3% and 2.0% relative improvements in WER for all segments and non-overlapping segments, respectively. Another 0.4-0.5% absolute reduction in WER was obtained by using multi-task learning architecture. It suggests that introducing MTL architecture is beneficial for the attention-based LSTM model. On the whole, our proposed model achieved 3.3% relative improvement in WER over LSTM baseline model on the SDM task.

5.3. Multiple Distant Microphones

For the MDM speech recognition task, we consider two baseline systems: (1) beamforming the multichannel signals into a single channel and following the LSTM acoustic model used for the SDM case; (2) training the LSTM acoustic model with the concatenated 8 microphone channels. For beamforming experiments, we used a delay-sum beamforming on 8 uniformly-spaced array channels through BeamformIt [23]. Table 3 shows the results on the MDM experiments.

Table 3: *Performance comparison on the development and evaluation set in the MDM case (WER,%)*

Model	*dev*	*dev**	*eval*	*eval**
LSTM (beamforming)	39.5	30.1	43.3	34.0
LSTM	37.8	30.7	42.7	34.5
ALSTM	36.0	29.7	41.4	33.6
ALSTM-MTL	**35.5**	**29.1**	**41.0**	**33.2**

The second row in Table 3 shows the results for the model trained on a single beamformed channel. And the results of the model trained directly on the outputs of multiple microphones are shown in the third row. Compared with the results of the baseline model in Table 2, MDM baseline systems obtained 3-5% absolute improvements in WER over the SDM system by using the multiple microphone data. Meeting speech recognition is characterised by speech overlap. Although the model trained on the beamformed signal performed better than that trained directly utilising the multi-channel features on the non-overlapping speech recognition task, it showed a lower performance on all segments which include overlapping speech. It suggests that using raw multiple channel input features in place of beamformed signal makes the acoustic model learn better representations which take into account some factor such as speech overlap.

We observed a 3.8% relative improvement in WER over the multiple input LSTM baseline model in Table 3 by embedding the attention mechanism across time. Additionally, further improvements can still be achieved by MTL architecture. On average our proposed model achieved 5.0% relative improvements in WER compared to the multiple input baseline model. In addition, it provided between a 2-4% absolute WER reduction compared to the beamforming baseline model. Overall, our proposed model performs best on both the overlapping and the non-overlapping speech recognition task in Table 3. And it is more computationally efficient than the beamforming baseline system which is a two-stage system: beamforming algorithm is applied to the multichannel speech, followed by conventional acoustic modeling approaches.

6. Conclusions

In this paper, we have presented an attention-based LSTM acoustic model with multi-task learning for distant speech recognition with single distant microphone or multiple distant microphones. The attention mechanism is embedded to utilize the temporal information at input layer. Furthermore, consistent improvements are achieved by incorporating multi-task learning architecture. The model trained on single distant microphone performed better than LSTM baseline model. In experiments with multiple microphones, we compared our model with those trained on the output of a delay-sum beamformer. The results presented here suggest that our model preforms well on both the overlapping and the non-overlapping speech recognition task.

7. Acknowledgements

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) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 2016A03007-1).

8. References

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

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[4] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention," in *Proc. Interspeech*, 2016.

[5] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Interspeech*, 2014, pp. 338–342.

[6] 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.

[7] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grézl, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[8] M. 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*. IEEE, 2008, pp. 104–107.

[9] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. 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*. IEEE, 2015, pp. 30–36.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5075–5079.

[11] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. Interspeech*, 2016.

[12] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5745–5749.

[13] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 2014, pp. 172–176.

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

[15] I. Himawan, P. Motlicek, D. Imseng, B. Potard, N. Kim, and J. Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4540–4544.

[16] S. Kim and I. Lane, "Recurrent models for auditory attention in multi-microphone distance speech recognition," in *Proc. Interspeech*, 2016.

[17] J. 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.

[18] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4290–4294.

[19] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[20] T. Gao, J. Du, L. Dai, and C. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4375–4379.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[22] V. Peddinti, V. Manohar, Y. Wang, D. Povey, and S. Khudanpur, "Far-field asr without parallel data," in *Proceedings of Interspeech*, 2016.

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