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

End-to-End Speech Recognition with Auditory Attention for Multi-Microphone Distance Speech Recognition

Suyoun Kim[1], *Ian Lane*[1]

[1]Electrical Computer Engineering
Carnegie Mellon University
suyoun@cmu.edu, lane@cmu.edu

Abstract

End-to-End speech recognition is a recently proposed approach that directly transcribes input speech to text using a single model. End-to-End speech recognition methods including Connectionist Temporal Classification and Attention-based Encoder Decoder Networks have been shown to obtain state-of-the-art performance on a number of tasks and significantly simplify the modeling, training and decoding procedures for speech recognition. In this paper, we extend our prior work on End-to-End speech recognition focusing on the effectiveness of these models in far-field environments. Specifically, we propose introducing Auditory Attention to integrate input from multiple microphones directly within an End-to-End speech recognition model, leveraging the attention mechanism to dynamically tune the model's attention to the most reliable input sources. We evaluate our proposed model on the CHiME-4 task, and show substantial improvement compared to a model optimized for a single microphone input.

Index Terms: end-to-end speech recognition, CTC, attention network, far-field speech recognition

1. Introduction

End-to-End speech recognition is a recently proposed approach that directly transcribes speech signal to text with a single neural network [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Unlike the traditional approach, Deep Neural Network - Hidden Markov Models (DNN-HMM) hybrid system [11, 12], the End-to-End model learns acoustic frames to character mappings in one step towards the final objective of interest, and attempts to rectify the suboptimal issues that arise from the disjoint training procedure. Also, these End-to-End models significantly simplify the pipeline of speech recognition system, without requiring the pre-defined frame-level alignment or segmentation. With enough large amount of training data, the End-to-End model has been shown to outperform HMM-based systems [13].

However, most of research on End-to-End speech recognition framework have focused on a single channel speech recognition, and the End-to-End model for multi-microphones distance speech recognition has not been studied actively. As many real-world speech recognition applications, including teleconferencing, robotics, and in-car spoken dialog systems, must deal with speech from distant microphones in noisy environments, exploiting the additional spatial information from multiple microphones for enhancing the signal is essential to achieve robust speech recognition in such noisy environments. Specifically, it is required to deal with the misaligned input channels because 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.

Signal processing techniques such as beamforming are widely used to extract an enhanced single channel from the misaligned multiple channels [14, 15, 16, 17]. These techniques, however, requires separate pre-processing step and are highly dependent on prior spatial information about the microphones and the environment in which the system is being used. Several recent studies [18, 19, 20, 21, 22, 23, 24] have explored an alternative way that processes multi-channels using a neural network. Some studies use Convolutional Neural Networks (CNNs) with a simply concatenated acoustic features from multiple microphones [19, 20, 21] to implicitly account for spatial relationships between channels. In more explicit way, the attention mechanism is used to estimate which channel at different time should be focused more on [23], or the neural beamformers within the acoustic model is used in[18, 25, 24]. Though these recent multi-channel processing methods based on the deep neural networks have been shown promising results, the multi-microphone processing within End-to-End speech recognition models have not been investigated yet.

In this work, we propose a multi-microphone End-to-End model for distant speech recognition. The key to our approach is that we use an additional attention mechanism within the End-to-End framework that enables to learn misaligned and non-stationary multiple input sources and automatically tune its attention to a more reliable input source among these sources. Our proposed method improves the performance by rectifying the misalignment issue between multiple channels. We evaluate our model on the CHiME-4 tasks, and show that our system outperforms the model optimized for a single microphone input.

The paper is organized as follows: in Section 2 we describe our proposed End-to-End model with auditory attention for multi-microphone processing. In section 3, we evaluate the performance of our model. Finally, in Section 4 we draw conclusions.

2. Model

In this section, our End-to-End model with auditory attention mechanism for multi-microphone processing. Figure 1 illustrates the overall architecture of our framework, where the attention mechanism for the multi-channel is embedded. In this work, we extend the joint CTC/Attention framework [9] and embed an additional attention model [23] to integrate misaligned multiple microphone sources. Our model addresses the (C) multiple channels and each channel has the variable (T) length input frames, $\boldsymbol{x} = (x_1^c, \cdots, x_T^c)$, and U length output characters, $\boldsymbol{y} = (y_1, \cdots, y_U)$, where $y_u \in \{1, \cdots, K\}$. K is the number of distinct labels.

http://dx.doi.org/10.21437/Interspeech.2017-1536

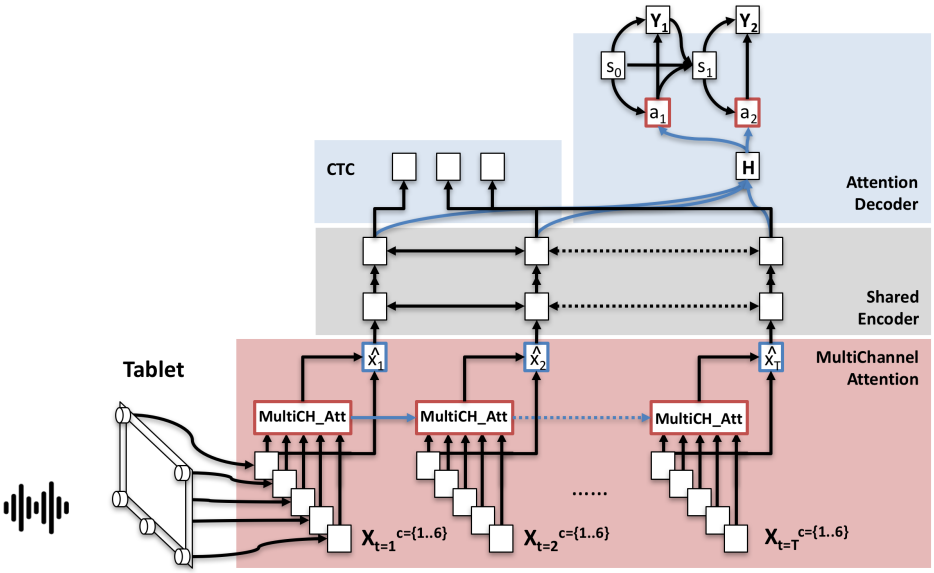

Figure 1: *Our proposed End-to-End with auditory attention for multi-microhpones. With an additional attention mechanism in* MultiChannelAttention *generates the enhanced input* $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ *from multi-microphones* $\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{1:C}}$*. The shared encoder is trained by both CTC and Attention objectives simultaneously. The shared encoder transforms the enhanced input* $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ *to high-level features* $\boldsymbol{h}$*, the localtion-based* AttentionDecoder *generates the character sequence* $\boldsymbol{y}$*.*

2.1. End-to-End Model

The key insight in joint CTC/Attention framework is that it can address the weaknesses of two main End-to-End models: Connectionist Temporal Classification (CTC) [26] and attention-based encoder-decoder (Attention) [7], by combining the two as they have the following complementary characteristics.

First, CTC can avoid *label bias problem* [27] since it is globally normalized model, however, Attention model is suffered from *label bias problem* since it is locally normalized model. Second, CTC is efficiently trained by forward-backward algorithm like hidden Markov Models (HMMs) and it can preserve the left-right order between input and output, however, the alignment of the Attention model does not preserve this order so that the alignment between input and output can be easily distorted. Unlike the machine translation task, the desired alignment between input and output in speech recognition task is monotonic, the CTC algorithm can be used to help naturally guide the alignment in Attention model to be monotonic. On the other side, the CTC still relies on the conditional independent assumption, it requires the separately trained language model like the hybrid framework [2, 3]. However, Attention model can jointly learn language model within a single network. The objective of each method and the details of how the joint CTC/Attention framework combines these two will be described in the followings.

CTC maximizes $P(\boldsymbol{y}|\boldsymbol{x})$, the probability distribution over all possible label sequences $\Phi(\boldsymbol{y}')$, allowing repetitions of labels and occurrences of a blank label $(-)$:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}\in\Phi(\boldsymbol{y})} P(\boldsymbol{\pi}|\boldsymbol{x}) \approx \sum_{\boldsymbol{\pi}\in\Phi(\boldsymbol{y})} \prod_{t=1}^{T} P(\pi_t|\boldsymbol{x}), \quad (1)$$

where $P(\pi_t|\boldsymbol{x})$ denotes the softmax activation of π_t label in Recurrent Neural Networks (RNNs) output layer at time t. The CTC loss to be minimized is defined as the negative log likelihood of the ground truth character sequence $\boldsymbol{y}^*$, i.e.

$$\mathcal{L}_{\text{CTC}} \triangleq -\ln P(\boldsymbol{y}^*|\boldsymbol{x}) = -\ln \sum_{\boldsymbol{\pi}\in\Phi(\boldsymbol{y})} P(\boldsymbol{\pi}|\boldsymbol{x}). \quad (2)$$

The Attention model consists of two sub-networks: *Encoder* and *AttentionDecoder*. The *Encoder* transforms $\boldsymbol{x}$, to high-level representation $\boldsymbol{h} = (h_1, \cdots, h_L)$ in Eq. (4), then *AttentionDecoder* generates the probability distribution over characters, y_u, conditioned on $\boldsymbol{h}$ and all the labels seen previously $y_{1:u-1}$ in Eq. (5) according to the following equations:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_u P(y_u|\boldsymbol{x}, y_{1:u-1}) \quad (3)$$

$$\boldsymbol{h} = \text{Encoder}(\boldsymbol{x}) \quad (4)$$

$$y_u \sim \text{AttentionDecoder}(\boldsymbol{h}, y_{1:u-1}). \quad (5)$$

Specifically, in this work, we use the location-based attention mechanism [5] in AttentionDecoder module as follows:

$$f_u = F * a_{u-1} \quad (6)$$

$$e_{u,l} = w^T \tanh(W s_{u-1} + V h_l + U f_{u,l} + b) \quad (7)$$

$$a_{u,l} = \frac{\exp(\gamma e_{u,l})}{\sum_l \exp(\gamma e_{u,l})} \quad (8)$$

$$c_u = \sum_l a_{u,l} h_l \quad (9)$$

where w, W, V, F, U, b are trainable parameters, s_{u-1} is the decoder state, γ is the sharpening factor [5], and * denotes convolution.

The loss function of the attention model is computed from:

$$\mathcal{L}_{\text{Attention}} \triangleq -\ln P(\boldsymbol{y}^*|\boldsymbol{x}) = -\sum_u \ln P(y_u^*|\boldsymbol{x}, y_{1:u-1}^*) \quad (10)$$

where $y_{1:u-1}^*$ is all the previous labels.

The joint CTC/Attention objective is represented as follows by combining two objectives in Eq. (2) and Eq. (10):

$$\mathcal{L} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Attention}}, \tag{11}$$

with a tunable parameter $\lambda : 0 \leq \lambda \leq 1$.

2.2. Auditory Attention for Multi-channel

The challenge we attempt to address with the neural attention mechanism is the problem of misaligned multiple input sources because the acoustic path length of each signal differs according to the location of the microphone. Similar to the recent work [23], we incorporate the $\mathrm{MultiChannelAttention}$ module that automatically tunes its attention to a more reliable input channel. Unlike traditional multi-microphone processing techniques, the model does not require any explicit signal preprocessing step.

At every input step t, the $\mathrm{MultiChannelAttention}$ produces an enhanced input representation $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$. For generating the input $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$, $\mathrm{MultiChannelAttention}$ estimates an attention weight vector over the channels $\text{MultiCH\_A}_t \in \mathbb{R}^C$ at each input step t. Attention weights are calculated based on two different information sources: 1) attention history MultiCH_A_{t-1}, and 2) content of input $\boldsymbol{x}_{\boldsymbol{t}}$:

$$\boldsymbol{e}_{\boldsymbol{t}} = w^T \tanh(W_a \text{MultiCH\_A}_t + W_x \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{1:C}}) \tag{12}$$

$$\text{MultiCH\_A}_t = \frac{\exp(\boldsymbol{e}_{\boldsymbol{t}})}{\sum_{c=1}^{C} \exp(\boldsymbol{e}_{\boldsymbol{t}}^{\boldsymbol{c}})} \tag{13}$$

$$\hat{\boldsymbol{x}}_{\boldsymbol{t}} = \sum_{c=1}^{C} \text{MultiCH\_A}_t^c \cdot \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{c}} \tag{14}$$

where $\boldsymbol{e}_{\boldsymbol{t}} \in \mathbb{R}^C$ is the energy at time t in Eq. (12), and MultiCH_A_t is an attention weights that normalized by softmax function in Eq. (13). Finally, the enhanced output $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ is generated by the weighted sum of the attention weights MultiCH_A_t and the multi-channel inputs $\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{1:C}}$ in Eq. (14). The subsequent procedure to generate the character sequence $\boldsymbol{y}$ from the integrated inputs $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ from the multi-channel inputs $\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{1:C}}$ is formalized as follows:

$$\hat{\boldsymbol{h}} = \text{Encoder}(\hat{\boldsymbol{x}}) \tag{15}$$

$$y_u \sim \text{AttentionDecoder}(\hat{\boldsymbol{h}}, y_{1:u-1}). \tag{16}$$

By combining the additional attention mechanism for processing multiple channels to the entire End-to-End framework, our system can be optimized jointly in one-step so that our framework attempts to rectify the suboptimal issues that arise from the disjoint training procedure.

3. Experiments

3.1. Data

We performed the experiments on the CHiME-4 corpus [28]. The CHiME-4 task is automatic speech recognition for a multi-microphone tablet device with 6 microphones in an everyday noisy environment - a cafe, a street junction, public transport, and a pedestrian area. There are two types of datasets: REAL and SIMU. The REAL data was recorded, and the SIMU data was generated by mixing clean utterance from WSJ0 into background recordings. The training set has 18 hours of speech data uttered by 83 speakers (3 hours REAL + 15 hours SIMU), the development set has 2.9 hours of speech data uttered by 3 speakers, and the evaluation set has 2.2 hours of data. The development set and the evaluation set consist of a 1:1 ratio of REAL and SIMU. We used the data from the 5 microphones except 2^{nd} microphone which is located on the backside of the tablet. None of our experiments used any language model or lexicon information.

As input features, we use 40 mel-scale filterbank coefficients, with their first and second order temporal derivatives to obtain a total 120 feature values per frame per each channel. As output label, we used only 59 distinct labels: 26 characters, digits, punctuation marks, apostrophe, period, dash, space, noise, sos/eos tokens, etc [3].

3.2. Training

The shared-encoder was a 4-layer Bidirectional Long Short-Term Memory (BLSTM) [29, 30] with 320 cells in each layer and direction, and linear projection layer is followed by each BLSTM layer. The bottom two layers of the encoder read every second hidden state in the network below, reducing the utterance length by the factor of 4, $L = T/4$. The decoder was 1-layer LSTM with 320 cells. The location-based attention mechanism with 10 centered convolution filters of width 100 were used to extract the convolutional features. We used the sharpening factor $\gamma = 2$. The AdaDelta algorithm [31] with gradient clipping [32] was used for optimization. All the weights are initialized with the range [-0.1, 0.1] of uniform distribution. For the CTC/Attention objective weights, we used 0.1 for λ.

3.3. Decoding

For decoding of the End-to-End model, we used a beam search algorithm similar to [33] with the beam size 20 to reduce the computation cost. The special *end-of-sentence(eos)* token is added to the target label, so that the decoder completes the generation of the hypothesis when *eos* is emitted. Since the model has a small bias for shorter utterances so we adjusted the score $\mathrm{s}(\boldsymbol{y}|\boldsymbol{x})$ by normalizing our probability with length penalty:

$$\mathrm{s}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\log P(\boldsymbol{y}|\boldsymbol{x})}{\delta \cdot |\boldsymbol{y}|_c}, \tag{17}$$

where $|\boldsymbol{y}|_c$ is the number of characters in the hypothesis and the tunable parameter $\delta = 0.3$. Note that we do not use any lexicon or language models.

3.4. Results

Table 1: *Character Error Rate (CER) on a noisy corpus CHiME-4. None of our experiments used any language model or lexicon information. Note that the training data consists of a 1:5 ratio of REAL and SIMU, whereas the development set and the evaluation set consist of a 1:1 ratio of REAL and SIMU.*

	Dev.		Eval.		
Model	**SIMU**	**REAL**	**SIMU**	**REAL**	**AVG**
`Single-E2E`	27.6	29.5	34.9	40.9	37.9
`ConcatMulti-E2E`	27.8	28.9	36.8	46.1	41.4
`BeamMulti-E2E`	25.0	23.1	35.8	34.2	35.0
`AttMulti-E2E`	26.5	26.8	32.9	38.0	35.5

In Table 1, we summarize character error rates (CERs) obtained on the CHiME-4 task. `AttMulti-E2E` is our proposed model, which has an attention mechanism for multiple inputs as

described in 2.2. As our baselines, we built three models with same joint CTC/Attention End-to-End model, but with three different inputs. `BeamMulti-E2E` was trained on the enhanced signal from 5 noisy channels. We obtained the enhanced signal from the beamforming toolkit, which was provided by the CHiME-4 organizer [34, 35]. `Single-E2E` was trained on a single noisy 5-th channel, and `ConcatMulti-E2E` used the concatenated 5 noisy channels.

The results in Table 1 show that our proposed model `AttMulti-E2E` significantly outperformed both `ConcatMulti-E2E` and `Single-E2E` in CER. Our model showed 5.6 - 7.4% relative improvements compared to the `Single-E2E` on evaluation and validation set, respectively. Also, our model showed 15.5 - 10.3% relative improvements compared to the `ConcatMulti-E2E` on evaluation and validation set, respectively. These results suggest that we can leverage the attention mechanism to integrate multiple channels efficiently. We also found that the model, which simply combined 5 features across microphones, did not perform very well. It showed poorer results than even the model trained with single microphone data. This result underscores the importance of integrating channels based on analysis of differences in arrival times. As expected, `BeamMulti-E2E` also provided a substantial improvement in CER compared to `Single-E2E` and `ConcatMulti-E2E`, showing a 15.4% and 7.3% relative improvement compared to `Single-E2E` on evaluation and validation set, respectively. Though the performance improvement of `BeamMulti-E2E` showed slightly greater than our model `AttMulti-E2E` (0.5 % better performed in CER in SIMU + REAL), `BeamMulti-E2E` requires separate pre-processing step with the manually defined beamforming parameter setting while our model bypasses these steps.

Table 2: *Character Error Rate (CER) on a noisy corpus CHiME-4. The order of microphone data is mismatched between training (1_3_4_5_6) and testing (6_5_4_3_1) modes. Note that the CER of `BeamMulti-E2E` is 35.0% (in Table 1)*

Model 1_3_4_5_6	**Eval. (SIMU+REAL)** **Matched** 1_3_4_5_6	**Mismatched** 6_5_4_3_1
`ConcatMulti-E2E`	41.4	46.9
`AttMulti-E2E`	35.5	36.6

To ensure the improvement of the system was coming from our time-channel attention mechanism, we evaluated on the modified test dataset that the order of microphones were shuffled. We changed the order from the one that we used for training (1_3_4_5_6). In Table 2 shows the CER between the models `ConcatMulti-E2E` and `AttMulti-E2E` in the different mismatched microphone order, (6_5_4_3_1). We observed that our model learned the desired alignment dynamically according to the content of the multiple channels. Our model showed that the performance degradation slightly compared to the matched order case, however, the `ConcatMulti-E2E` showed that 13.3 % relative performance degradation compared to the matched order case. Unlike that the performance of `ConcatMulti-E2E` was severely degraded in the mismatched case, our model showed more robust performance even in the mismatched case. This result indicates that the performance of our model less rely on the specific microphone setting. Figure 2 visualizes the log-Mel filterbank coefficients from each

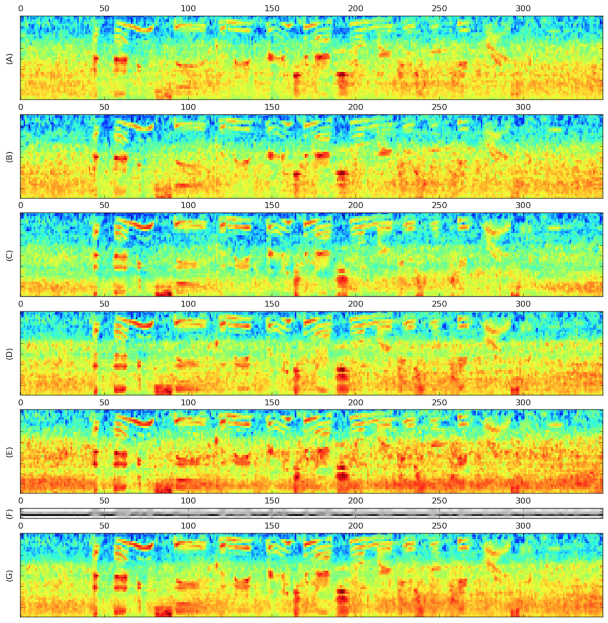

Figure 2: *Visualization of log-Mel filterbank coefficients and the multi-channel alignment over time. (A) 1^{st} channel, (B) 3^{rd} channel, (C) 4^{th} channel, (D) 5^{th} channel, (E) 6^{th} channel, (F) multi-channel alignment, and (G) integrated input by the alignment.*

channel (A) - (E), and the one from our enhanced feature (G) with the channel attention mechanism which is shown in subfigure (F). In general, the features generated from our model (G) were more clearly delineated.

We then analyzed the computational aspects of our system. As the multi-microphone processing is performed as part of the End-to-End model computation we have actually found it to be more computationally efficient than performing beamforming followed by the model. On our development machine (Intel(R) Xeon(R) CPU E5-2640 @ 2.50GHz) and with GeForce GTX TITAN X, the beamforming toolkit [34, 35] operated approximately 9.1 sec for 6.1 second of each utterance. However, our model does not require the beamforming computation which can save the time and achieve significantly faster decoding process.

4. Conclusions

We have introduced an End-to-End framework for far-field speech recognition that uses a novel attention mechanism for multiple microphone data. Our model improves performance by automatically tuning its attention to a more reliable input source. Moreover, it significantly speeds up the process of decoding without requiring any explicit preprocessing step. We presented our results on the CHiME-4 task and found that our far-field End-to-End achieved comparable performance to beamforming without any prior knowledge of the microphone layout or any explicit preprocessing.

5. Acknowledgements

This research was supported by LGE.

6. References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1764–1772.

[2] 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.

[3] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[4] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," *arXiv preprint arXiv:1412.1602*, 2014.

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

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[7] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *arXiv preprint arXiv:1508.04395*, 2015.

[8] L. Lu, X. Zhang, and S. Renals, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5060–5064.

[9] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," *arXiv preprint arXiv:1609.06773*, 2016.

[10] L. Lu, L. Kong, C. Dyer, and N. A. Smith, "Multi-task learning with ctc and segmental crf for speech recognition," *arXiv preprint arXiv:1702.06378*, 2017.

[11] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[12] 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," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[13] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[14] D. Van Compernolle, W. Ma, F. Xie, and M. Van Diest, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communication*, vol. 9, no. 5, pp. 433–442, 1990.

[15] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 489–498, 2004.

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

[17] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[18] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5542–5546.

[19] 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.

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

[21] 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.

[22] T. Yoshioka, S. Karita, and T. Nakatani, "Far-field speech recognition using cnn-dnn-hmm with convolution in time," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4360–4364.

[23] S. Kim and I. Lane, "Recurrent models for auditory attention in multi-microphone distance speech recognition," *arXiv preprint arXiv:1511.06407*, 2015.

[24] J. Heymann *et al.*, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP).*, 2017.

[25] 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.

[26] 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.

[27] J. Lafferty, A. McCallum, F. Pereira *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, vol. 1, 2001, pp. 282–289.

[28] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

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

[30] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[31] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[32] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *arXiv preprint arXiv:1211.5063*, 2012.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[34] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 4, pp. 1763–1773, 2004.

[35] C. Y.-K. Lai and P. Aarabi, "Multiple-microphone time-varying filters for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–233.