Interspeech 2018
2-6 September 2018, Hyderabad

Multi-channel Attention for End-to-End Speech Recognition

Stefan Braun[1], *Daniel Neil*[1], *Jithendar Anumula*[1], *Enea Ceolini*[1], *Shih-Chii Liu*[1]

[1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

brauns@ethz.edu, daniel.l.neil@gmail.com, anumula@ini.uzh.ch, enea.ceolini@ini.uzh.ch, shih@ini.ethz.ch

Abstract

Recent end-to-end models for automatic speech recognition use sensory attention to integrate multiple input channels within a single neural network. However, these attention models are sensitive to the ordering of the channels used during training. This work proposes a sensory attention mechanism that is invariant to the channel ordering and only increases the overall parameter count by 0.09%. We demonstrate that even without re-training, our attention-equipped end-to-end model is able to deal with arbitrary numbers of input channels during inference. In comparison to a recent related model with sensory attention, our model when tested on the real noisy recordings from the multi-channel CHiME-4 dataset, achieves a relative character error rate (CER) improvement of 40.3% to 42.9%. In a two-channel configuration experiment, the attention signal allows the lower signal-to-noise ratio (SNR) sensor to be identified with 97.7% accuracy.

Index Terms: end-to-end speech recognition, multi-channel, attention mechanism

1. Introduction

In recent years, end-to-end models are being considered for automatic speech recognition (ASR) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] as they present a simplification in both the model architecture and training process over conventional DNN-HMM hybrids [11, 12]. End-to-end models transcribe speech to text with a single neural network, replacing the combination of separate deep neural network (DNN) acoustic models and hidden Markov models (HMMs). The simplified model learns the mapping from acoustic feature to character sequences in a single training process, thereby avoiding the disjoint multi-stage training procedures for hybrid ASR systems.

While most ASR research with end-to-end models focused on single-channel scenarios, the multi-channel scenario is less explored. Many real-world ASR applications (e.g. Amazon Echo, voice-control systems in cars etc.) deal with speech from multiple microphones in noisy environments, and their accuracy relies on methods that robustly pre-process multi-channel inputs, ideally avoiding noise corruption and generating a cleaner, enhanced signal. In this context, conventional beamforming algorithms are widely used to extract a single enhanced channel from multi-channel setups, but they introduce a separate beamforming processing stage which is typically optimized independently from the ASR objective. Alternate approaches for multi-channel integration are based on methods that leverage convolutional neural networks (CNNs) for channel combination [13, 14, 15], that learn a beamforming function with neural networks [16, 17, 18, 19, 20, 21], and attention mechanisms that focus on higher signal-to-noise ratio (SNR) channels [22]. While these methods are differentiable and suitable for joint optimization in an end-to-end model, they were usually combined with conventional hybrid ASR approaches.

To the best of our knowledge, only two recent studies [23, 24] have examined multi-channel ASR and meet the criteria of a strict end-to-end scenario, i.e. training a single neural network model towards the ASR objective only and testing without a separate lexicon or language model. In both studies, inputs from the multiple channels are combined into a single representation that is used for the classification task. In one case, a neural beamformer is used to combine the channels [24] and in the second case, a sensory attention mechanism [23] is used instead. While both approaches show promising performance compared to conventional beamforming, the neural beamformer shows benefits such as invariance to channel reordering and robustness to channel configurations that were not used during training.

In this work, we propose a sensory attention mechanism that follows a similar, but not identical design strategy as in [23]. Our proposed design shows invariance to channel re-ordering and the design is simplified by using long short-term memory (LSTM) and dense units instead of a custom-designed neural network cell. We evaluate the use of this sensory attention mechanism in an end-to-end ASR model and compare our results with related models [23, 24] on the CHiME-4 dataset. We demonstrate that our attention-equipped end-to-end model can process new channel configurations without re-training, and that the sensory attention signal is strongly correlated to the channel SNR.

2. End-to-end multi-channel ASR model

We first present the two main components of our end-to-end model for multi-channel ASR, that is, the sensory attention mechanism described in Subsection 2.1, and the acoustic model described in Subsection 2.2. The block diagram of this model is given in Figure 1.

2.1. Sensory attention mechanism

The attention mechanism combines multiple input channels into a single representation by summing the dynamically weighted frames from individual channels.

We consider a multi-channel setup with $c = 1, ..., N$ microphone channels. We assume that the input time series of every channel is binned into $t = 1, ..., T$ frames and that each channel c produces a D-dimensional feature vector $f_t^c \in \mathbb{R}^D$ for every frame t. The merged representation $m_t \in \mathbb{R}^D$ is generated as follows:

$$z_t^c = Z(f_{1...t}^c) \tag{1}$$

$$\alpha_t^c = \frac{\exp(z_t^c)}{\sum_{j=1}^{N} \exp(z_t^j)} \tag{2}$$

$$m_t = \sum_{c=1}^{N} \alpha_t^c f_t^c \tag{3}$$

10.21437/Interspeech.2018-1301

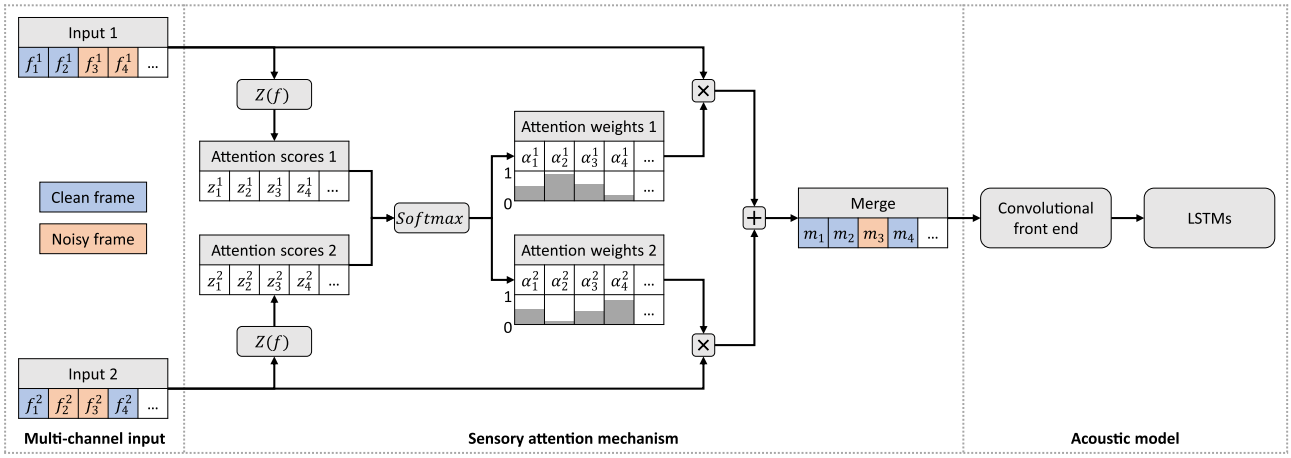

Figure 1: *Architecture of our proposed model for end-to-end ASR from multi-channel inputs. Depicted is the case of two input channels. Input feature vectors f_t^c are weighted and summed to create the merged representation m_t, which is then used for classification. The attention mechanism suppresses noisy frames irrespective of the channel, resulting in a cleaner merged representation.*

The scoring function Z produces attention scores $z_t^c \in \mathbb{R}^1$ based on the feature frames of channel c (Equation 1). The attention weights $\alpha_t^c \in \mathbb{R}^1$ are computed by performing a softmax operation on the attention scores $z_t^c \in \{z_t^1, ..., z_t^N\}$ (Equation 2), and thus $\sum_{c=1}^{N} \alpha_t^c = 1$. Finally, the individual feature frames f_t^c are weighted by the corresponding attention weights α_t^c and summed into the merged representation m_t (Equation 3) which is then presented to the acoustic model.

The scoring function Z is arbitrary and can be modelled using neural networks. In our experiments, we implemented Z using 10 LSTM units [25] followed by a single dense unit (weight W, bias b) with a SELU non-linearity [26] (Equation 4):

$$Z(f_{1..t}^c) = \mathrm{SELU}(W \cdot \mathrm{LSTM}(f_{1..t}^c) + b) \quad (4)$$

The use of LSTM units is convenient because past history is automatically considered.

By design, our sensory attention mechanism has the following useful properties. First, it is a soft attention mechanism which is fully differentiable and therefore, suitable for end-to-end optimization. Second, the attention weights a_t^c at each frame t indicate the contribution of single channels to frame t of the merged representation. Third, because the attention weights are re-computed on every frame, they can dynamically adjust for temporal changes in signal quality (e.g. temporary noise corruption) of each channel. Fourth, as the same scoring function Z is used for all input channels, the attention mechanism is invariant to channel re-ordering. Finally, because the scoring function Z evaluates each channel independently from other channels, channels may be removed or new channels may be added after training.

2.2. Acoustic model

The acoustic model receives as input, the merged representation generated by the sensory attention mechanism. It is composed of a convolutional front-end (CFE) followed by a stack of LSTM units.

The CFE is made of three convolutional blocks. Each block performs a function f that includes a 2D convolution, a 2D instance normalization [27] and a clipped ReLU non-linearity $\sigma(x) = \min\{\max(x, 0), 20\}$ [28]:

$$f(x) = \sigma(\mathrm{InstanceNorm2d}(\mathrm{Conv2d}(x))) \quad (5)$$

The CFE operates on spectrogram features. As the CFE uses a temporal stride of 2 in the first layer, it effectively halves the sequence length and reduces training time. Our CFE implementation is closely related to the DeepSpeech2 CFE, where a similar configuration helped to improve error rates especially in noisy conditions [3]. Therefore, the proposed acoustic model should provide a noise robust baseline model. The main difference of our implementation from that of DeepSpeech2 is that we use instance normalization (sample-wise normalization) instead of batch normalization (batch-wise normalization) and we do not keep mean and variance statistics from training to be applied during normalization at test time. On the four different noise environments of CHiME-4, using mean and variance statistics computed across samples from different environments for normalization, decreases our model performance therefore we use instance normalization.

The CFE is followed by a stack of bidirectional LSTM units and the final output layer is an affine transform to the class labels. We use the Connectionist Temporal Classification (CTC) [29] objective to automatically learn the mapping and alignment between input features and label sequences. The model is tested with strict end-to-end criteria and without use of external lexicons or language models. We use greedy decoding on the CTC output: at each time step, the most likely label is selected.

2.3. Related work

We compare our model to related work on multi-channel end-to-end ASR without additional lexicons or language models. The `ATTMULTI-E2E` model [23] combines multiple input channels into a single representation with a sensory attention mechanism based on weighted summation. Their attention mechanism has three main differences to our work: (1) it operates on filterbank features while ours operates on spectrogram features, (2) it uses a custom designed neural network cell to compute attention scores while we use generic LSTM and dense units, (3) by design, it is not invariant to channel re-ordering in contrast to ours which is invariant. The `MASK_NET(ATT)` model [24] applies an attention mechanism to select the reference microphone for a neural beamformer. In contrast to the `ATTMULTI-E2E` and our proposed model, the channels are not combined by a sensory attention mechanism but rather by using a neural beam-

former. The neural beamformer is also able to exploit spatial information, which is not considered by ATTMULTI-E2E and our model. Both ATTMULTI-E2E and MASK_NET(ATT) use a CTC+Encoder/Decoder hybrid model that is trained with a joint CTC-attention multi-task objective [10], while our model is trained with an encoder (i.e. the acoustic model) and standard CTC objective only.

3. Experiments

3.1. Dataset

All experiments are carried out as ASR tasks on the CHiME-4 data-set [30] which provides real and simulated noisy speech data from a tablet device with 6 microphones. Recordings were done in four noisy environments: a cafe, a street junction, public transport and a pedestrian area. The real data was recorded with the tablet device, while the simulated data was obtained by mixing clean utterances from WSJ0 [31] with environment background recordings. The tablet device provided 5 microphones facing the speaker and 1 microphone facing away from the speaker (backward channel #2, the noisiest of all). For training we use both real data ('tr05_real', 1600 samples) and simulated data ('tr05_simu', 7138 samples).

The audio samples are pre-processed into 161-dimensional spectrogram features with the short-time Fourier transform (STFT). First, the STFT-coefficients are computed (20 ms frame length, 10ms frame shift, Hamming window) and then the log of the magnitude of the STFT-coefficients is kept. The features are further normalized to zero mean and unit variance per sample. The output labels consist of 59 distinct labels such as characters and digits and are obtained with the EESEN preprocessing routines [5].

3.2. Models

In total, 5 different models are evaluated: NOISY, BEAMFORMIT, MVDR, MC-AVG and MC-ATT. The NOISY model is trained and evaluated only on channel 5. It provides a baseline for a model optimized on the best-performing channel. All other models are trained and tested on the front channels 1/3/4/5/6, but differ in their channel combination strategies. The BEAMFORMIT model uses a delay-and-sum beamformer [32], while the MVDR model uses a minimum variance distortionless response (MVDR) beamformer based on the implementation provided by the CHiME authors [30]. Both beamformers produce enhanced waveforms in a separate pre-processing stage that is not optimized towards the ASR objective, and so their corresponding models are not considered as end-to-end models. The MC-ATT model uses our proposed sensory attention mechanism (Subsection 2.1) to merge the input channels. In order to assess the effectiveness of this attention mechanism, we compare the MC-ATT model against an averaging model, MC-AVG, that assigns fixed attention weights $\alpha_t^c = 1/5$ for the five input channels. We do not include the simple channel concatenation strategy, because it is not inherently invariant to channel re-ordering (see [23]) and it complicates channel addition or removal after training because the acoustic model expects a fixed input dimensionality. We include results from both ATTMULTI-E2E [23] and MASK_NET(ATT) [24] models for comparison.

Table 1: *2D convolution filters of the CFE. First dimension is frequency and second dimension is time.*

Layers	Channels	Kernel	Stride
L1, L2, L3	32, 32, 96	41x11, 21x11, 21x11	2x2, 2x1, 2x1

3.3. Training parameters

All our models are optimized separately, but use the same acoustic model architecture presented in Subsection 2.2: a CFE with 3 layers of convolutional blocks (Table 1) followed by 5 layers of bidirectional LSTMs with 256 units in each direction. The final output layer is an affine transform to the 59 output classes. The MC-ATT model uses 10 LSTM units followed by a single dense unit with a SELU non-linearity to implement the attention scoring function Z (Equation 1), resulting in 7k additional parameters. The models were trained in an end-to-end fashion with the CTC objective [29] and the ADAM optimizer [33] for 150 epochs. The model with the lowest character error rate (CER) on the development set was used for evaluation.

3.4. Results

The CER obtained on the CHiME-4 development and evaluation sets are reported in Table 2. All reported models do not make use of external lexicons or language models.

3.4.1. Real noisy data

The MC-ATT model achieves the lowest overall error rates on both real noisy subsets 'et05_real' and 'dt05_real'. MC-ATT shows a relative CER improvement of 4.4% to 8.1% over MC-AVG and 22.7% to 23.3% over NOISY. Seemingly, MC-ATT benefits from the automatically learned channel weighting. The BEAMFORMIT model shows error rates that are similar to that of the MC-ATT model. The MVDR model shows better results than the single channel NOISY model, but is not competitive with the other approaches on real noisy data.

Results from related work report higher error rates. Our MC-ATT model shows a relative CER improvement of 15.3% to 15.9% over MASK_NET(ATT) and 40.3% to 42.9% over ATTMULTI-E2E. The higher error rates of MASK_NET(ATT) and ATTMULTI-E2E may originate from their hybrid CTC+Encoder/Decoder acoustic model unlike our simple CTC model. The number of parameters of the MC-ATT model (8.031M) also compares favorably against those of ATTMULTI-E2E ($\sim$8M) and MASK_NET(ATT) ($\sim$18M). Note that the latter implements the neural beamformer part with an estimated $\sim$10M parameters, while our sensory attention mechanism uses only 7k parameters.

3.4.2. Simulated noisy data

The MVDR model clearly achieves the lowest CER on both simulated noisy subsets 'et05_simu' and 'dt05_simu' and yields significantly lower error rates than it did on the real noisy data. For MVDR beamforming, better performance on simulated data was also reported by the CHiME-4 authors and explained with the absence of reverberation in the simulated data [30]. The MASK_NET(ATT) model performs significantly better than MC-ATT on 'dt05_simu', but worse on 'et05_simu'. The BEAMFORMIT model performed worse than the single channel NOISY model on 'et05_simu'. This result may be explained by the separate optimization of the beamforming and acoustic

Table 2: *CER [%] results on CHiME-4 ASR experiments. No language models are used. The best results are printed in bold. Related work did not give parameter counts, thus they were estimated to the best of our knowledge.*

Model	Parameters	dt05 simu	dt05 real	et05 simu	et05 real
NOISY	8.024M	20.1	19.8	25.3	29.6
MC-AVG	8.024M	18.2	16.0	24.8	24.7
MC-ATT	8.031M	17.5	**15.3**	22.5	**22.7**
BEAMFORMIT	8.024M	17.7	**15.3**	26.2	23.5
MVDR	8.024M	**13.0**	18.6	**17.4**	28.6
ATTMULTI-E2E [23]	∼8M	26.5	26.8	32.9	38.0
MASK_NET (ATT) [24]	∼18M	15.3	18.2	23.7	26.8

model components. We further hypothesize that the simulated noisy data itself could explain the unexpected findings: at times, the simulation process introduces residual speech artifacts on channels 1/3/4/6 but produces a cleaner channel 5 signal[1].

3.4.3. *New channel configurations*

The flexibility and interpretability of the sensory attention mechanism is demonstrated through additional experiments on 'dt05_real'. We test the CER of the MC-ATT and MC-AVG models for the cases of channel re-ordering, channel addition and channel removal. The models are not re-trained for these new channel configurations. The CER results are reported in Table 3 along with the average attention weight $(\bar{\alpha}^c = \frac{1}{T}\sum_{t=1}^{T} \alpha_t^c)$ of every channel c of MC-ATT, computed over all $T = 989608$ frames of 'dt05_real'. Note that the way we report the attention weights, corresponds to the CHiME-4 channels, and does not reflect the channel order. The MC-AVG model assigns equal attention weights to all N channels, i.e. $\alpha_t^c = \frac{1}{N}$.

As expected, both models are invariant to channel reordering and yield identical CER for channel orders 6/5/4/3/1 and 1/3/4/5/6. Adding the noisy channel 2 (1/2/3/4/5/6) leads to a smaller increase in CER for MC-ATT. In fact, MC-ATT suppresses channel 2 as seen by the lower attention weight α_t^2 of this channel when compared to the other channels. This indicates a good generalization of the sensory attention mechanism because it was not trained on the data from channel 2. For all channel configurations, channel 2 has the lowest attention weight and channel 5 has the highest attention weight whenever either one is present. When removing channels, MC-ATT has an increased advantage and shows a relative CER improvement of up to 12.7% over MC-AVG in the channel configuration 2/5. For this configuration, the results show that the attention mechanism is quite accurate: $\alpha_t^5 > \alpha_t^2$ holds true for 97.7% of all frames. In other words: by comparing attention weights alone, we can identify the higher SNR channel 5 with 97.7% accuracy. The high interpretability of the attention weights is further confirmed by the plots of the input features and attention weights for the channel configuration 2/5 in Figure 2.

4. Conclusion

In this work we presented an end-to-end model that embeds a sensory attention mechanism for noise-robust multi-channel ASR. The attention mechanism uses no prior assumptions on microphone configurations, and therefore enables our end-to-end model to deal with arbitrary channel ordering without retraining. The attention weights are dynamically decreased on channels with more noise, and the model is able to deal with the addition or removal of input channels. These are useful properties for real-world systems, as the attention weights could help to identify failing sensors that need replacement or suboptimal sensors which can then be removed to save computation and hardware resources.

The attention mechanism is implemented by a simple network consisting of generic LSTM and dense units. Even though the total parameter count of our end-to-end model increases by only 0.09% from the addition of this mechanism, it allows the model to achieve performance that, on real noisy data, is on par or better than using separate beamforming pre-processing stages. Compared to a related model which also uses a sensory attention mechanism, our end-to-end model showed a relative CER improvement of 40.3% to 42.9% on the real-world noisy recordings of the CHiME-4 data-set.

[1] e.g. sample 'M06_447C0216_STR' from 'et05_simu'

Table 3: *CER [%] results on the 'dt05_real' subset of CHiME-4 for new channel configurations. The attention weights for MC-ATT are an average over all frames of this subset. Lowest CER and highest attention weight are printed in bold.*

Channels	CER [%] MC-AVG	CER [%] MC-ATT	MC-ATT attention weights $\bar{\alpha}^1$	$\bar{\alpha}^2$	$\bar{\alpha}^3$	$\bar{\alpha}^4$	$\bar{\alpha}^5$	$\bar{\alpha}^6$
1/3/4/5/6	16.0	**15.3**	0.19	-	0.18	0.22	**0.23**	0.18
6/5/4/3/1	16.0	**15.3**	0.19	-	0.18	0.22	**0.23**	0.18
1/2/3/4/5/6	17.1	**16.1**	0.16	0.12	0.16	0.19	**0.21**	0.16
2/3/4/5	18.3	**16.9**	-	0.18	0.23	0.29	**0.30**	-
2/3/5	19.8	**18.0**	-	0.25	0.32	-	**0.43**	-
2/5	22.8	**19.9**	-	0.37	-	-	**0.63**	-
2	46.4	**45.8**	-	**1.00**	-	-	-	-
5	18.3	**17.9**	-	-	-	-	**1.00**	-

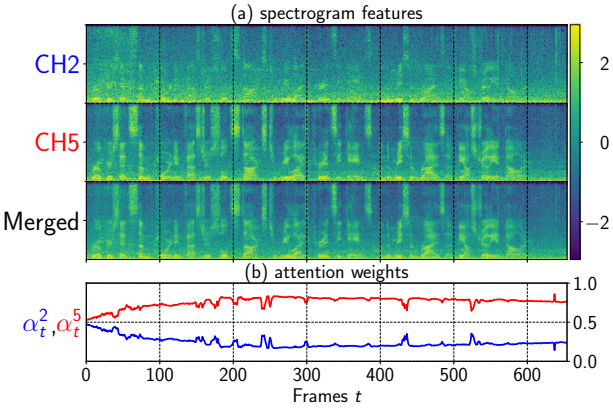

Figure 2: *Operation of MC-ATT on a sample with channel configuration 2/5. (a) Spectrogram features for the two input channels and the merged representation. (b) Attention weights for the two input channels. The merged representation is dominated by channel 5, as evident by the higher attention weight values of this channel which has less noise.*

5. Acknowledgements

This work was partially supported by Samsung Advanced Institute of Technology and the European Union's Horizon 2020 research and innovation program under grant agreement No 644732.

6. References

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

[2] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016, pp. 173–182.

[4] E. Battenberg *et al.*, "Exploring neural transducers for end-to-end speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 206–213.

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

[6] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015, pp. 577–585.

[7] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

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

[10] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4835–4839.

[11] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*. Kluwer Academic Publishers, 1994.

[12] G. Hinton *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.

[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] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. W. Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 30–36.

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

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

[18] H. Erdogan *et al.*, "Multi-channel speech recognition: Lstms all the way through," in *CHiME-4 workshop*, 2016.

[19] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 1976–1980.

[20] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 271–275.

[21] J. Heymann, L. Drude, C. Böddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5325–5329.

[22] S. Kim and I. R. Lane, "Recurrent models for auditory attention in multi-microphone distant speech recognition," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 3838–3842.

[23] S. Kim and I. Lane, "End-to-end speech recognition with auditory attention for multi-microphone distance speech recognition," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, 2017, pp. 3867–3871.

[24] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

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

[26] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 972–981.

[27] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 315–323.

[29] 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 (ICML)*, 2006, pp. 369–376.

[30] 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.

[31] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete, LDC93S6A," *Linguistic Data Consortium, Philadelphia*, 2007.

[32] 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.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.