RECURRENT MODELS FOR AUDITORY ATTENTION IN MULTI-MICROPHONE DISTANCE SPEECH RECOGNITION

Suyoun Kim
Electrical Computer Engineering
Carnegie Mellon University
suyoun@cmu.edu

Ian Lane
Electrical Computer Engineering
Carnegie Mellon University
lane@cmu.edu

ABSTRACT

Integration of multiple microphone data is one of the key ways to achieve robust speech recognition in noisy environments or when the speaker is located at some distance from the input device. Signal processing techniques such as beamforming are widely used to extract a speech signal of interest from background noise. These techniques, however, are highly dependent on prior spatial information about the microphones and the environment in which the system is being used. In this work, we present a neural attention network that directly combines multi-channel audio to generate phonetic states without requiring any prior knowledge of the microphone layout or any explicit signal preprocessing for speech enhancement. We embed an attention mechanism within a Recurrent Neural Network (RNN) based acoustic model to automatically tune its attention to a more reliable input source. Unlike traditional multi-channel preprocessing, our system can be optimized towards the desired output in one step. Although attention-based models have recently achieved impressive results on sequence-to-sequence learning, however, no attention mechanisms have been applied to learn multiple inputs which may be asynchronous and non-stationary. We evaluate our neural attention model on a subset of the CHiME-3 challenge task, and we show that the model achieves comparable performance to beamforming using a purely data-driven method.

1 INTRODUCTION

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. When a human voice is captured with far-field microphones in these environments, the audio signal is severely degraded by reverberation, and background noise. This makes the distant speech recognition task far more challenging than near-field speech recognition, which is commonly used for voice-based interaction today.

Acoustic signals from multiple microphones can be used to enhance recognition accuracy due to the availability of additional spatial information. Many researchers have proposed techniques to efficiently integrate inputs from multiple distant microphones. The most representative multi-channel processing technique is the beamforming approach (Van Compernolle et al., 1990; Seltzer et al., 2004; Kumatani et al., 2012; Pertilä & Nikunen, 2015), which generates an enhanced single output signal by aligning multiple signals through digital delays that compensate for the different distances of the input signals. However, the performance of beamforming is highly dependant on prior information about the microphone location and the location of a target source. For down-stream tasks such as speech recognition this preprocessing step is suboptimal because it is not directly optimized towards the final objective of interest: speech recognition accuracy (Seltzer, 2008).

Over the past few years, deep neural networks (DNNs) have been successfully applied to acoustic models in speech recognition (Seide et al., 2011; Mohamed et al., 2012; Hinton et al., 2012). Other works (Liu et al., 2014; Renals & Swietojanski, 2014; Swietojanski et al., 2014; Yoshioka et al., 2015; Himawan et al., 2015) have shown that DNNs can learn suitable representations for distant speech recognition by directly using multi-channel input. These approaches however, simply concatenated acoustic features from multiple microphones without considering the spacial properties of acoustic signal propagation.

Recently, an ”attention mechanism” of neural networks has been proposed to address the problem of learning variable-length input and output sequences (Bahdanau et al., 2014). At each output step, the previous output history is used to generate an attention vector over the input sequence. This attention vector enables models to learn to focus attention on specific parts of their input. Their attention-equipped frameworks have shown very promising results on many challenging tasks involving inputs and outputs which may have variable length, including machine translation (Bahdanau et al., 2014), parsing (Vinyals et al., 2014), image captioning (Xu et al., 2015) and conversational modeling (Vinyals & Le, 2015). Specifically, for the speech recognition tasks, (Chorowski et al., 2014; Chan et al., 2015; Bahdanau et al., 2015) attempted to align between the input features and the desired character sequence using an attention mechanism. However, no attention mechanisms have been applied to learn to integrate multiple inputs.

In this work, we propose a novel attention based model that enables to learn misaligned and non-stationary multiple input sources for distant speech recognition. We embed an attention mechanism within a Recurrent Neural Network (RNN) based acoustic model to automatically tune its attention to a more reliable input source among misaligned and non-stationary input sources at each output step. The attention module is learned with normal acoustic model and it is jointly optimized towards phonetic state accuracy. Our attention module differs in the way that we 1) deal with the problem of the integration of the different qualities and misalignment of multiple sources, and we 2) exploit spatial information between multiple sources to accelerate learning the auditory attention. While playing a similar role to traditional multichannel preprocessing through deep neural network architecture, at the same time, our system can bypass the limitation of preprocessing that requires separate expensive step and depending on prior information.

Through a series of experiments on the CHiME-3 (Jon Barker, 2015) dataset, we show that our proposed approach improves recognition accuracy in various types of noisy environments. In addition, we also compare our approach with beamforming technique(Jon Barker, 2015; Loesch & Yang, 2010; Blandin et al., 2012; Mestre et al., 2003). The paper is organized as follows: in Section 2 we describe our proposed attention based model. In section 3, we evaluate the performance of our model. Finally, in Section 4 we draw conclusions.

2 Model

In this section, we describe our neural attention model that allows neural networks to focus more on reliable input sources across different temporal locations. We formulate the proposed framework with applications in multi-channel distant speech recognition. While there has been some recent work on end-to-end neural Large-vocabulary continuous speech recognition (LVCSR) systems - from speech directly to transcripts (Graves et al., 2006; Graves & Jaitly, 2014; Hannun et al., 2014; Chorowski et al., 2014), our model is based on typical hybrid DNN-HMM frameworks (Morgan & Bourlard, 1994; Hinton et al., 2012), wherein the acoustic model estimates hidden Markov model (HMM) state posteriors, because we focus on dealing with re-weighted input representation of misaligned multiple input sources.

Given a set of input sequences $\mathbf{X} = \{\mathbf{X}^{ch_1}, \cdots, \mathbf{X}^{ch_N}\}$, where $\mathbf{X}^{ch_i}$ is an input sequence $(x_1^{ch_i}, \cdots, x_T^{ch_i})$ from the ith microphone, $i \in \{1, \cdots, N\}$, our system computes a corresponding sequence of HMM acoustic states, $\mathbf{y} = (y_1, \cdots, y_T)$. We model each output $\mathbf{y}_t$ at time t as a conditional distribution over the previous outputs $y_{<t}$ and the multiple inputs $\mathbf{X}_t$ at time t using the chain rule:

$$P(\mathbf{y}|\mathbf{X}) = \prod_t P(y_t|\mathbf{X}, y_{<t}) \tag{1}$$

Our system consists of two subnetworks: $\mathrm{AttendMultiSource}$ and LSTM-AM. $\mathrm{AttendMultiSource}$ is an attention-equipped Recurrent Neural Network (RNN) for learning attention to determine and focus on reliable channel and temporal locations over the candidate multiple input sequences. $\mathrm{AttendMultiSource}$ produces re-weighted inputs, $\widehat{\mathbf{X}}$, based on the learned attention. This $\widehat{\mathbf{X}}$ is used for the next subnetwork $\mathrm{LSTM\text{-}AM}$, which is a Long Short-Term Memory (LSTM) acoustic model to estimate the probability of the output HMM state $\mathbf{y}$. Figure 1 visualizes our overall model with these two components. We describe more details of each component in the following subsections 2.1 and 2.2.

$$\hat{\mathbf{X}} = \mathrm{AttendMultiSource}(\mathbf{X}, \mathbf{y}) \tag{2}$$

$$P(\mathbf{y}|\mathbf{X}) = \text{LSTM-AM}(\hat{\mathbf{X}}, \mathbf{y}) \tag{3}$$

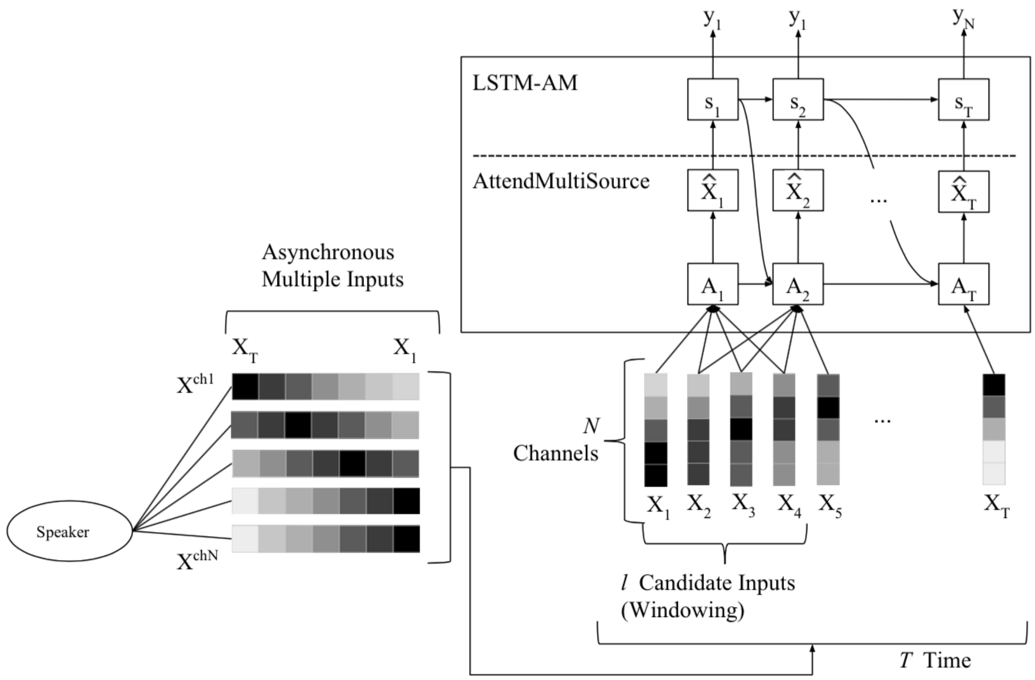

Figure 1: Schematic representation of our neural attention model.

2.1 Attention mechanism for multiple sources

The challenge we attempt to address with the neural attention mechanism is the problem of misaligned multiple input sources with non-stationary quality over time. Specifically, in multi-channel distant speech recognition, the arrival time of each channel is different because the acoustic path length from the source signal is different according to the location of the microphone. This results in misalignment of input features and the difference of arrival time would be greater when the space between microphones is longer. Even worse, quality of signal across channels can themselves vary over time because the speaker and interfering noise sources may keep changing. Figure 1 describes this environment and asynchronous multiple inputs due to the acoustic path length differences.

We now introduce an attention mechanism to cope with the misaligned input problem and formulate the $\mathrm{AttendMultiSource}$. At every output step t, the $\mathrm{AttendMultiSource}$ function produce a re-weighted input representation $\widehat{\mathbf{X}_c}$, given cth candidate input set $\mathbf{X}_c$. $\mathbf{X}_c$ is a subsequence of time frames. As proposed by (Bahdanau et al., 2015), we perform similar windowing to limit the exploring temporal location of inputs for computational efficiency and scalability. We limit range of attention to l=7 time frames ($\pm$3). In our experiments, longer time steps had little impact on the overall performance and they would rather benefit from microphones placed further apart from each other.

For re-weighting the input $\mathbf{X}_c$, $\mathrm{AttendMultiSource}$ predicts an attention weight matrix $\mathbf{A}_t^{time,ch}$ at each output step t. Unlike previous attention mechanism, we produce a weight matrix rather than a vector, because our attention mechanism additionally identifies which channel, in a given time step, is more relevant. Therefore, $\mathbf{A}_t^{time,ch}$ is (*number of channels*) by (*number of candidate input frames*) matrix, here it is N x l matrix. Attention weights are calculated based on four different information sources: 1) attention history $\mathbf{A}_{t-1}^{time,ch}$, 2) content in the candidate sequences $\mathbf{X}_c$ 3) decoding history $\mathbf{s}_{t-1}$, and 4) an additional spatial information between multiple microphones based on phase difference information $\mathbf{PD}_c$ corresponding to $\mathbf{X}_c$. The following three formulations describes the $\mathrm{AttendMultiSource}$ function:

$$\mathbf{E}_t^{time,ch} = \mathrm{MLP}(\mathbf{s}_{t-1}, \mathbf{A}_{t-1}^{time,ch}, \mathbf{PD}_c, \mathbf{X}_c) \tag{4}$$

$$\mathbf{A}_t^{time,ch} = \mathrm{softmax}(\mathbf{E}_t^{time,ch}) \tag{5}$$

$$\widehat{\mathbf{X}_c} = \mathbf{A}_t^{time,ch} \cdot \mathbf{X}_c \tag{6}$$

Specifically, MLP (in equation 4) computes an energy matrix $\mathbf{E}_t^{time,ch}$ (N x l) by following equation:

$$\mathbf{E}_t^{time,ch} = tanh(\mathbf{W}_s \cdot \mathbf{s}_{t-1} + \mathbf{W}_a \cdot \mathbf{A}_{t-1}^{time,ch} + \mathbf{W}_p \cdot \mathbf{PD}_c + \mathbf{W}_x \cdot \mathbf{X}_c + b) \tag{7}$$

where $\mathbf{W}_s$, $\mathbf{W}_a$, $\mathbf{W}_p$, $\mathbf{W}_x$ are parameter matrices, and b is a parameter vector. Once we compute the energy $\mathbf{E}_t^{time,ch}$ at time t, then we obtain $\mathbf{A}_t^{time,ch}$ by normalizing $\exp(\mathbf{E}_t^{time,ch}) / \sum_{time,ch} \exp(\mathbf{E}_t^{time,ch})$, such that, $\forall t$, $\mathbf{A}_t^{time,ch} \geq 0$, and $\sum_{time,ch} \mathbf{A}_t^{time,ch} = 1$ (in equation 5). Finally, re-weighted output $\widehat{\mathbf{X}_c}$ is generated by calculating the dot product of the attention weights $\mathbf{A}_t^{time,ch}$ and candidate input $\mathbf{X}_c$ (in equation 6). Typically, the selection of elements from input candidates is a weighted sum. However, we only calculate the dot product in order to avoid losing information.

To accelerate the learning of the attention mechanism, we use additional spatial information based on analysis of differences in arrival time. It is generally assumed that the human auditory system can localize multiple sounds and attend to the desired signal using information from the interaural time difference (ITD) (Stern et al., 2008; Park & Stern, 2009). A previous work (Kim et al., 2009) attempts to emulate human binaural processing and estimate ITD indirectly by comparing phase difference from the two microphones at each frequency domain. They identify "close" time-frequency component to the speaker based on the estimated ITD. Similarly we also use phase difference between two microphones for inferring spatial information. The following equations are used to compute phase difference between two microphones i and j, where $i \neq j, i, j \in \{1 \cdots N\}$:

$$pd^{ch_i - ch_j} = \min |\angle x^{ch_i} - \angle x^{ch_j} - 2\pi r| \tag{8}$$

$$\mathbf{PD}^{ch_i - ch_j} = (pd_1^{ch_i - ch_j}, \cdots, pd_T^{ch_i - ch_j}) \tag{9}$$

$$\mathbf{PD} = \{\mathbf{PD}^{ch_1 - ch_2}, \cdots, \mathbf{PD}^{ch_4 - ch_5}\} \tag{10}$$

From these equations, we calculate the phase differences of each pair of multiple microphones, then the MLP network accepts the $\mathbf{PD}_c$ corresponding to the input candidates, with $\mathbf{X}_c$ as an additional input.

2.2 LSTM ACOUSTIC MODEL

Our next subnetwork LSTM-AM serves as a typical RNN-based acoustic modeling, except it accepts the re-weighted input $\widehat{\mathbf{X}_c}$ instead of the original input $\mathbf{X}_c$. The LSTM-AM uses a Long Short-Term Memory RNN (LSTM)(Hochreiter & Schmidhuber, 1997), which has been successfully applied to the speech recognition tasks due to its ability to handel long-term dependencies. The LSTM contains special units called memory blocks in the recurrent hidden layer, and each block has memory cells c_t with special three gates (input i_t, output o_t, and forget f_t) to control the flow of information.

In our work, we use a simplified version of an LSTM without peephole connections and biases to reduce the computational expense for learning the standard LSTM models. Although LSTMs have many variations for enhancing their performance, such as BLSTM (Graves et al., 2013), LSTMP (Sak et al., 2014), PBLSTM (Chan et al., 2015), in our work, we focus on verifying an additional attention mechanism with a simple LSTM architecture, instead of improving LSTM acoustic modeling overall.

LSTM-AM maps a re-weighted input sequence based on the attention mechanism $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}^{ch_1}}, \cdots, \widehat{\mathbf{x}^{ch_N}}\}$, where $\widehat{\mathbf{x}^{ch_i}} = (\widehat{x_1^{ch_i}}, \cdots, \widehat{x_T^{ch_i}})$, to an output sequence $\mathbf{y}_t = (y_1, \cdots, y_T)$ by calculating the network unit activations using the following equations iteratively from $t = 1$ to T:

$$i_t = \sigma(\widehat{\mathbf{x}_c} W_{xi} + h_{t-1} W_{hi}) \tag{11}$$
$$f_t = \sigma(\widehat{\mathbf{x}_c} W_{xf} + h_{t-1} W_{hf}) \tag{12}$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(\widehat{\mathbf{x}_c} W_{xc} + h_{t-1} W_{hc}) \tag{13}$$
$$o_t = \sigma(\widehat{\mathbf{x}_c} W_{xo} + h_{t-1} W_{ho}) \tag{14}$$
$$s_t = o_t \cdot \tanh(c_t) \tag{15}$$

where W terms denote weight matrices, and σ the logistic sigmoid function. i_t, f_t, o_t, and c_t are the input gate, forget gate, output gate and cell activation vectors, respectively. Finally, the output s_t are used to predict the current HMM state label by softmax (in equation 16). s_t is also used to predict the next $t + 1$ attention matrix as well as next c_{t+1} hidden state of LSTM-AM.

$$y_t = argmax_i P(y = i|s_t) \tag{16}$$

3 Experiments

3.1 Dataset

We evaluated the performance of our architecture on a subset of the CHiME-3 task. The CHiME-3 (Jon Barker, 2015) task is automatic speech recognition for a multi-microphone tablet device in an everyday environment, a cafe, a street junction, public transport, and a pedestrian area. There are two types of datasets: `REAL` and `SIMU`. The `REAL` data consists of 6-channel recordings. 12 US English speakers were asked to read the sentences from the WSJ0 corpus (Garofalo et al., 2007) with using the multi-microphone tablet. They were encouraged to adjust their reading position, so that the target distance kept changing over time. The simulated data was generated by mixing clean utterances from WSJ0 into background recordings. To verify our method in a real noisy environment, we chose not to use the simulated dataset but rather to use only the `REAL` dataset, with 5-channels from the five microphones which were located in each corner of tablet, about 10cm to 20cm away from each other (we excluded one microphone, which faced backward in tablet device).

3.2 System Training

All the networks were trained on a 1,600 utterance (about 2.9 hours) dataset. The dataset was represented with 25ms frames of 40-dimensional log-filterbank energy features computed every 10ms. We produced 1,992 HMM state labels from a trained GMM-HMM system using near field microphone data, and these state labels were used in all subsequent experiments. We use one layer of LSTM architecture with 512 cells. The weights in all the networks were initialized to the range (-0.03, 0.03) with a uniform distribution, and the initial attention weights were initialized to $1/n$ in n dimension. We set the configuration of the learning rate to 0.4 and after two epochs it decays during training. All models resulted in a stable convergence range from 1e-05 to 5e-05. To avoid the vanishing gradient problem, we limited the norm of gradient to 1 (Pascanu et al., 2012). Apart from the gradient clipping, we did not limit the activations of the weights.

During training, we evaluated frame accuracies (i.e. phone state labeling accuracy of acoustic frames) on the development set of 1,640 utterances. The trained models were evaluated in a speech

Table 1: Comparison of WERs(%) on development and evaluation set of the subset (REAL) of the CHiME-3 task between the three baseline systems, and our proposed framework, ALSTM

MODEL (Input)	**DEV (WER %)**	**TEST (WER %)**
Baselines		
LSTM (Preprocessing 5 noisy-channel)	35.2	52.1
LSTM (single noisy-channel)	39.1	57.1
LSTM (5 noisy-channel)	43.0	60.1
Our models		
ALSTM	35.9	52.3
ALSTM (with phase)	33.9	50.0

recognition system on a test set of 1,320 utterances. For all the decoding experiments, we used a size 18 of beam and size 10 of lattices. We did not perform sequence training. The input for all networks was log-filterbank features, with 5 channels stacking, and then with 7 frames stacking (+3-3).

3.3 RESULTS

In Table 1, we summarize word error rates (WERs) obtained on the subset of the CHiME3 task. ALSTM is our proposed model that has an attention mechanism for multiple inputs as described in 2.1, and ALSTM (with phase) used phase information in addition to ALSTM.

As our baselines, we built four models and used the same simple version of LSTM architecture that we described in Section 2.2 with four different inputs. LSTM (Preprocessing 5 noisy-channel) trained on the enhanced signal from 5 noisy-channels. We obtained the enhanced signal from the beamforming toolkit, which is provided by CHiME3 organizer (Jon Barker, 2015; Loesch & Yang, 2010; Blandin et al., 2012; Mestre et al., 2003). LSTM (single noisy-channel) trained on a single noisy-channel, and LSTM (5 noisy-channels) used the concatenated 5 noisy channels.

As expected, LSTM (Preprocessing 5 noisy-channel) provides a substantial improvement in WER compared to LSTM (single noisy channel) and LSTM (5 noisy-channel), and it showed 13.3% and 5.0% relative improvement in WER, respectively. Also we found that the model, which simply combined 5 features across microphones, does not perform very well. It showed poorer results than even the model trained with single microphone data. This result indicates that integrating channels based on analysis of differences in arrival time is necessary.

Our model with the attention mechanism provided a significant improvement in WER compared to LSTM (5 noisy-channel). Compared to LSTM (5 noisy-channel), a 17% reduction in relative error rate at evaluation set was obtained by ALSTM (with phase), and a 13% relative error rate by ALSTM. These results show that we can leverage the attention mechanism to integrate multiple channels efficiently. We also found that the additional phase information can help to learn attention and WER improved in 4.6% relatively. In comparison with LSTM (Preprocessing 5 noisy-channel), we also found our proposed model achieved comparable performance to beamforming without any preprocessing. Although ALSTM shows a slightly lower performance as compared to LSTM (Preprocessing 5 noisy-channel), a 4.0% relative error rate was obtained by ALSTM (with phase).

4 CONCLUSIONS

We proposed an attention based model (ALSTM) that uses asynchronous and non-stationary inputs from multiple channels to generate outputs. For the distant speech recognition task, we embedded a novel attention mechanism within a RNN based acoustic model to automatically tune its attention to a more reliable input source. Our framework is significantly flexible and fast. Because it works without any prior knowledge of the microphone layout or any explicit multi-channel preprocessing.

We presented our results on the subset of the CHiME3 task and ALSTM showed a substantial improvement in WER in comparison with the model using channel concatenation. Also, our model achieved comparable performance to beamforming without any preprocessing.

The implications of this work are significant and far-reaching. It would allow the microphones to be placed more flexibly because our framework does not dependent on prior information about microphone arrays. Also, our work suggests the possibility of building a highly efficient ASR system through bypassing preprocessing. These finding suggest that our approach will likely do well on tasks that need to exploit misaligned and non-stationary inputs from multiple sources, such as multimodal problems, and sensory fusion. We believe that our attention framework can greatly improve these tasks by maximizing the benefit of inputs from multiple sources.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions made by Richard M. Stern, and William Chan for their valuable and constructive suggestions. This research was supported by LGE.

REFERENCES

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bahdanau, Dzmitry, Chorowski, Jan, Serdyuk, Dmitriy, Brakel, Philemon, and Bengio, Yoshua. End-to-end attention-based large vocabulary speech recognition. *arXiv preprint arXiv:1508.04395*, 2015.

Blandin, Charles, Ozerov, Alexey, and Vincent, Emmanuel. Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, 92(8):1950–1960, 2012.

Chan, William, Jaitly, Navdeep, Le, Quoc V, and Vinyals, Oriol. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

Chorowski, Jan, Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.

Garofalo, John, Graff, David, Paul, Doug, and Pallett, David. Csr-i (wsj0) complete. *Linguistic Data Consortium, Philadelphia*, 2007.

Graves, Alan, Jaitly, Navdeep, and Mohamed, Abdel-rahman. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 273–278. IEEE, 2013.

Graves, Alex and Jaitly, Navdeep. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1764–1772, 2014.

Graves, Alex, Fernández, Santiago, Gomez, Faustino, and Schmidhuber, Jürgen. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376. ACM, 2006.

Hannun, Awni, Case, Carl, Casper, Jared, Catanzaro, Bryan, Diamos, Greg, Elsen, Erich, Prenger, Ryan, Satheesh, Sanjeev, Sengupta, Shubho, Coates, Adam, et al. Deepspeech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Himawan, Ivan, Motlicek, Petr, Imseng, David, Potard, Blaise, Kim, Namhoon, and Lee, Jaewon. Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, number EPFL-CONF-207946, 2015.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Jon Barker, Ricard Marxer, Emmanuel Vincent Shinji Watanabe. The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. *Submitted to IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

Kim, Chanwoo, Kumar, Kshitiz, Raj, Bhiksha, and Stern, Richard M. Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain. In *INTERSPEECH*, pp. 2495–2498, 2009.

Kumatani, Kenichi, McDonough, John, and Raj, Bhiksha. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *Signal Processing Magazine, IEEE*, 29(6):127–140, 2012.

Liu, Yulan, Zhang, Pengyuan, and Hain, Thomas. 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*, pp. 5542–5546. IEEE, 2014.

Loesch, Benedikt and Yang, Bin. Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions. In *Latent Variable Analysis and Signal Separation*, pp. 41–48. Springer, 2010.

Mestre, Xavier, Lagunas, Miguel, et al. On diagonal loading for minimum variance beamformers. In *Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on*, pp. 459–462. IEEE, 2003.

Mohamed, Abdel-rahman, Dahl, George E, and Hinton, Geoffrey. Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):14–22, 2012.

Morgan, N and Bourlard, H. Connectionist speech recognition: a hybrid approach, 1994.

Park, Hyung-Min and Stern, Richard M. Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero-crossings. *Speech Communication*, 51(1):15–25, 2009.

Pascanu, Razvan, Mikolov, Tomas, and Bengio, Yoshua. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2012.

Pertilä, Pasi and Nikunen, Joonas. Distant speech separation using predicted time–frequency masks from spatial features. *Speech Communication*, 68:97–106, 2015.

Renals, Steve and Swietojanski, Pawel. Neural networks for distant speech recognition. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, pp. 172–176. IEEE, 2014.

Sak, Haşim, Senior, Andrew, and Beaufays, Françoise. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

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

Seltzer, Michael L. 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*, pp. 104–107. IEEE, 2008.

Seltzer, Michael L, Raj, Bhiksha, and Stern, Richard M. Likelihood-maximizing beamforming for robust hands-free speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 12 (5):489–498, 2004.

Stern, Richard M, Gouvêa, Evandro, Kim, Chanwoo, Kumar, Kshitiz, and Park, Hyung-Min. Binaural and multiple-microphone signal processing motivated by auditory perception. In *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, pp. 98–103. IEEE, 2008.

Swietojanski, Pawel, Ghoshal, Arnab, and Renals, Steve. Convolutional neural networks for distant speech recognition. *Signal Processing Letters, IEEE*, 21(9):1120–1124, 2014.

Van Compernolle, Dirk, Ma, Weiye, Xie, Fei, and Van Diest, Marc. Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication*, 9(5):433–442, 1990.

Vinyals, Oriol and Le, Quoc. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Vinyals, Oriol, Kaiser, Lukasz, Koo, Terry, Petrov, Slav, Sutskever, Ilya, and Hinton, Geoffrey. Grammar as a foreign language. *arXiv preprint arXiv:1412.7449*, 2014.

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Courville, Aaron, Salakhutdinov, Ruslan, Zemel, Richard, and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

Yoshioka, Takuya, Karita, Shigeki, and Nakatani, Tomohiro. Far-field speech recognition using cnn-dnn-hmm with convolution in time. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4360–4364. IEEE, 2015.