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, no attention mechanisms have previously been applied to learn potentially asynchronous and non-stationary multiple inputs. We evaluate our neural attention model on the CHiME-3 challenge task, and 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 microphone location and the location of the target source. For downstream 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).

Recently, an ”attention mechanism” in 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. 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 the normal acoustic model and is jointly optimized towards phonetic state accuracy. Our attention module is unique in the way that we 1) deal with the problem of integrating different qualities and misalignment of multiple sources, and 2) exploit spatial information between multiple sources to accelerate learning of auditory attention. Our system plays a similar role to traditional multichannel preprocessing through deep neural network architecture, but bypasses the limitations of preprocessing, which requires an expensive, separate step and depends on prior information.

2 Model

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. Given a set of input sequences $\mathbf{X} = \{\mathbf{X}^{ch_1}, \cdots, \mathbf{X}^{ch_N}\}$, where $\mathbf{X}^{ch_i}$ is from the ith microphone, 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.

Our system consists of two subnetworks: $\mathrm{AttendMultiSource}$ and $\mathrm{LSTM\text{-}AM}$. $\mathrm{AttendMultiSource}$ is an attention-equipped Recurrent Neural Network (RNN) for learning to determine and focus on reliable channels and temporal locations among 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 LSTM-AM, which is a Long Short-Term Memory (LSTM) acoustic model to estimate the probability of the output HMM state $\mathbf{y}$.

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 of each signal differs according to the location of the microphone. This results in the misalignment of input features. At every output step t, the $\mathrm{AttendMultiSource}$ function produces 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. 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 mechanisms, 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 the (*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) additional spatial information between multiple microphones based on phase difference information $\mathbf{PD}_c$ corresponding to $\mathbf{X}_c$. The following three formulations describe the $\mathrm{AttendMultiSource}$ function:

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

$$\mathbf{A}_t^{time,ch} = \mathrm{softmax}(\mathbf{E}_t^{time,ch}) \tag{2}$$

$$\widehat{\mathbf{X}_c} = \mathbf{A}_t^{time,ch} \cdot \mathbf{X}_c \tag{3}$$

3 Experiments

We evaluated the performance of our architecture on the CHiME-3 task. The CHiME-3 (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. We use one layer of

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

MODEL (Input)	**DEV (WER %)**	**TEST (WER %)**
Baselines - Real + Simulated Data (18hrs)		
LSTM (Preprocessing 5 noisy-channel)	18.6	32.0
Proposed - Real + Simulated Data (18hrs)		
ALSTM	16.5	26.5

LSTM architecture with 512 cells. There is a mismatch between the Kaldi baseline (Povey et al., 2011) and our results because we did not perform sequence training (sMBR) or language model rescoring (5-gram rescoring or RNNLM). The inputs for all networks were log-filterbank features, with 5 channels stacking, and then with 7 frames stacking (+3-3).

In Table 1, we summarize word error rates (WERs) obtained on the CHiME3 task. ALSTM is our proposed model, which has an attention mechanism for multiple inputs. As our baseline, LSTM (Preprocessing 5 noisy-channel) was trained on the enhanced signal from 5 noisy channels. We obtained the enhanced signal from the beamforming toolkit, which was provided by the CHiME3 organizer (Barker, 2015; Loesch & Yang, 2010; Blandin et al., 2012; Mestre et al., 2003). Our model with the attention mechanism provided a significant improvement in WER compared to LSTM (5 noisy-channel). These results suggest that we can leverage the attention mechanism to integrate multiple channels efficiently.

4 Conclusions

We proposed an attention-based model (ALSTM) that uses asynchronous and non-stationary inputs from multiple channels to generate outputs. For a 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. We presented our results on the CHiME3 task and found that ALSTM showed a substantial improvement in WER. Our model achieved comparable performance to beamforming without any prior knowledge of the microphone layout or any explicit preprocessing. Our findings suggest that this 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 benefits of using inputs from multiple sources.

Acknowledgments

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

References

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

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

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

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, 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.

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

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

Xavier Mestre, Miguel Lagunas, 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.

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

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

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

Michael 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*, pp. 104–107. IEEE, 2008.

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

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