MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Deep Long Short-Term Memory Adaptive Beamforming Networks for Multichannel Robust Speech Recognition

Meng, Z.; Watanabe, S.; Hershey, J.R.; Erdogan, H.

TR2017-012 March 2017

Abstract

Far-field speech recognition in noisy and reverberant conditions remains a challenging problem despite recent deep learning breakthroughs. This problem is commonly addressed by acquiring a speech signal from multiple microphones and performing beamforming over them. In this paper, we propose to use a recurrent neural network with long short-term memory (LSTM) architecture to adaptively estimate real-time beamforming filter coefficients to cope with non-stationary environmental noise and dynamic nature of source and microphones positions which results in a set of timevarying room impulse responses. The LSTM adaptive beamformer is jointly trained with a deep LSTM acoustic model to predict senone labels. Further, we use hidden units in the deep LSTM acoustic model to assist in predicting the beamforming filter coefficients. The proposed system achieves 7.97% absolute gain over baseline systems with no beamforming on CHiME-3 real evaluation set.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

DEEP LONG SHORT-TERM MEMORY ADAPTIVE BEAMFORMING NETWORKS FOR MULTICHANNEL ROBUST SPEECH RECOGNITION

Zhong Meng[1,2*], *Shinji Watanabe*[1], *John R. Hershey*[1], *Hakan Erdogan*[3]

[1] Mitsubishi Electric Research Laboratories, Cambridge, MA
[2] Georgia Institute of Technology, Atlanta, GA
[3] Microsoft Research, Redmond, WA

ABSTRACT

Far-field speech recognition in noisy and reverberant conditions remains a challenging problem despite recent deep learning breakthroughs. This problem is commonly addressed by acquiring a speech signal from multiple microphones and performing beamforming over them. In this paper, we propose to use a recurrent neural network with long short-term memory (LSTM) architecture to adaptively estimate real-time beamforming filter coefficients to cope with non-stationary environmental noise and dynamic nature of source and microphones positions which results in a set of time-varying room impulse responses. The LSTM adaptive beamformer is jointly trained with a deep LSTM acoustic model to predict senone labels. Further, we use hidden units in the deep LSTM acoustic model to assist in predicting the beamforming filter coefficients. The proposed system achieves 7.97% absolute gain over baseline systems with no beamforming on CHiME-3 real evaluation set.

Index Terms— beamforming, multichannel, speech recognition, LSTM

1. INTRODUCTION

Although extraordinary performance has been achieved in automatic speech recognition (ASR) with the advent of deep neural networks (DNNs) [1, 2], the performance still degrades dramatically in noisy and far-field situations [3, 4]. To achieve robust speech recognition, multiple microphones can be used to enhance the speech signal, reduce the effects of noise and reverberation, and improve the ASR performance. In this scenario, an essential step of the ASR front-end processing is multichannel filtering, or *beamforming*, which steers a spatial sensitivity region, or "beam," in the direction of the target source, and inserts spatial suppression regions, or "nulls," in the directions corresponding to noise and other interference.

Delay-and-sum (DAS) beamforming is widely used for multichannel signal processing [5], in which the multichannel inputs of an microphone array are delayed to be aligned in time and then summed up to be a single channel signal. The signal from the target direction is enhanced and the noises and interferences coming from other directions are attenuated. Filter-and-sum beamforming applies filters to the input channels before summing them up [6]. Minimum variance distortionless response (MVDR) [7] and generalized eigenvalue (GEV) [8] are filter-and-sum beamforming methods which solve for filter coefficients using different derivations.

Although these methods have achieved good performance in beamforming, their goal is to optimize only the signal-level objective (e.g., SNR). In order to achieve robust speech recognition, it is more important to jointly optimize beamforming and acoustic model with the objective of maximizing the ASR performance. In [9], the parameters of a frequency-domain beamformer are first estimated by a DNN based on the generalized cross correlation between microphones. Conventional features are extracted from the beamformed signal before passing through a second DNN for acoustic modeling. Instead of filtering in the frequency domain, [10] performs spatial and spectral filtering through time-domain convolution over raw waveform. The output feature is then passed to a convolutional LSTM DNN (CLDNN) acoustic model to predict the context-dependent state output targets. In [11], the beamforming and frequency decomposition are factorized into separate layers in the network. These approaches assume that the speaker position and the environment are fixed and estimate constant filter coefficients for either beamforming or spatial and spectral filtering.

However, in real noisy and far-field scenarios, as the position of the source (speaker), noise and room impulse response keep changing, the time-invariant filter coefficients estimated by these neural networks may fail to robustly enhance the target signal. Therefore, we propose to adaptively estimate the beamforming filter coefficients at each time frame using an LSTM to deal with any possible changes of the source, noise or channel conditions. The enhanced signal is generated by applying these time-variant filter coefficients to the short-time Fourier transform (STFT) of the array signals. Log filterbank like features are obtained from the enhanced signal and then passed to a deep LSTM acoustic model to predict the senone posterior. The LSTM beamforming network and the LSTM acoustic model are jointly trained using truncated back-propagation through time (BPTT) with a cross-entropy objective. STFT coefficients of the array signals are used as the input of the beamforming network. In ASR systems of [12, 13], the speech signal is enhanced by NMF and LSTM before fed into the acoustic model. But speech enhancement module and the acoustic model are not jointly optimized to minimize the WER and the input is only single channel signal.

Previous work [14] has shown that the speech separation performance can be improved by incorporating the speech recognition alignment information within the speech enhancement framework. Inspired by this, we feed the units of the top hidden layer of the LSTM acoustic model at the previous time step back as an auxiliary input to the beamforming network to predict the current filter coefficients. Note that our work is different from [15] in that: (1) we perform adaptive beamforming over 5 different input channels, but their system works only on 2 input channels; (2) our adaptive LSTM beamformer predicts only the frequency domain filter coef-

[*] Zhong Meng performed the work while he was an intern at Mitsubishi Electric Research Laboratories, Cambridge, MA.

ficients and performs frequency domain filter-and-sum over STFT coefficients, while their work majorly focuses on the time-domain filtering with raw waveforms as the input; (3) the log Mel filter bank like features are generated with fixed log Mel transform over the beamformed STFT coefficients for acoustic modeling in our work, while time/frequency domain convolution is performed with trainable parameters on the beamformed features in their work; (4) no additional gate modulation is applied to the feedback to reduce the system complexity for our much smaller dataset. In the experiments, we show that this feedback captures high-level knowledge about the acoustic states and increases the performance. The experiments are conducted with the CHiME 3 dataset. The joint training of LSTM adaptive beamforming network and deep LSTM acoustic model achieves 7.75% absolute gain over the single channel signal on the real test data. The acoustic model feedback provides an extra gain of 0.22%.

2. LSTM ADAPTIVE BEAMFORMING

2.1. Adaptive Filter-and-Sum Beamforming

As a generalization of the delay-and-sum beamforming, filter-and-sum beamformer processes the signal from each microphone using a finite impulse response (FIR) filter before summing them up. In frequency domain, this operation can be written as:

$$\hat{x}_{t,f} = \sum_{m=1}^{M} g_{f,m} x_{t,f,m}, \tag{1}$$

where $x_{t,f,m} \in \mathcal{C}$ is the complex STFT coefficient for the time-frequency index (t, f) of the signal from channel m, $g_{f,m} \in \mathcal{C}$ is the beamforming filter coefficient and $\hat{x}_{t,f} \in \mathcal{C}$ is the complex STFT coefficient of the enhanced signal. In Eq. (1), $t = 1, \ldots, T, f = 1, \ldots, F$ and M, T, F are the numbers of microphones, time frames and frequencies. To cope with the time-variant source position and room impulse response, we make the filter coefficients time-dependent and propose the adaptive filter-and-sum beamforming:

$$\hat{x}_{t,f} = \sum_{m=1}^{M} g_{t,f,m} x_{t,f,m}, \tag{2}$$

where $g_{t,f,m} \in \mathcal{C}$ is time-variant complex filter coefficient.

2.2. Adaptive LSTM Beamforming Network

The LSTM network is a special kind of recurrent neural network (RNN) with purpose-built memory cells to store information [16]. The LSTM has been successfully applied to many different tasks [17, 18] due to its strong capability of learning long-term dependencies. The LSTM takes in an input sequence $x = \{x_1, \ldots, x_T\}$ and computes the hidden vector sequence $h = \{h_1, \ldots, h_T\}$ by iterating the equation below

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{3}$$

We implement the LSTM in Eq. (3) with no peep hole connections.

In this work, we apply *real-value* LSTM to the adaptive filter-and-sum beamformer to predict the real and imaginary parts of the complex filter coefficients at time t and channel m. That is, we introduce the following real-value vectors for complex values $g_{t,f,m}$ and $x_{t,f,m}$ in Eq. (2):

$$g_{t,m} \triangleq [\Re(g_{t,f,m}), \Im(g_{t,f,m})]_{f=1}^{F} \in \mathcal{R}^{2F}$$
$$x_t \triangleq [\Re(x_{t,f,m}), \Im(x_{t,f,m})]_{f=1,m=1}^{F,M} \in \mathcal{R}^{2FM}.$$

With this representation, the real-value LSTM predicts $g_{t,m}$ as follows:

$$p_t = W_{x,p} x_t \tag{4}$$
$$h_t = \text{LSTM}^{BF}(p_t, h_{t-1}) \tag{5}$$
$$g_{t,m} = \tanh(W_{h,m} h_t), \quad m = 1, \ldots, M, \tag{6}$$

where $W_{x,p}$ and $W_{h,m}$ are projection matrices. We use $\tanh(\cdot)$ function to limit the range of the filter coefficients within $[-1, 1]$.

The real and imaginary parts of the STFT coefficient $\hat{x}_{t,f}$ of the beamformed signal are generated by Eq. (2) as follows

$$\begin{cases} \Re(\hat{x}_{t,f}) & = \sum_{m=1}^{M} \Re(x_{t,f,m})\Re(g_{t,f,m}) - \Im(x_{t,f,m})\Im(g_{t,f,m}) \\ \Im(\hat{x}_{t,f}) & = \sum_{m=1}^{M} \Re(x_{t,f,m})\Im(g_{t,f,m}) + \Im(x_{t,f,m})\Re(g_{t,f,m}). \end{cases} \tag{7}$$

More sophisticated features can be extracted from the beamformed STFT coefficients and are passed to the LSTM acoustic model to predict the senone posterior. In our experiments, the log Mel filterbank like feature is generated from Eq. (7) by

$$z_t = \log\left(\text{Mel}(P_t)\right) \tag{8}$$
$$P_t = \left[\Re(\hat{x}_{t,f})^2 + \Im(\hat{x}_{t,f})^2\right]_{f=1}^{F} \in \mathcal{R}^{F} \tag{9}$$

where Mel$(\cdot)$ is the operation of Mel matrix multiplication, and P_t is F dimensional real-value vector of the power spectrum of the beamformed signal at time t. Global mean and variance normalization is applied to this log Mel filterbank like feature. Note that all operations in this section are performed with the *real-value* computation, and can be easily represented by a differentiable computational graph.

2.3. Deep LSTM Acoustic Model

Recently, LSTMs are shown to be more effective than DNNs [1, 2] and conventional RNNs [19, 20] for acoustic modeling as they are able to model temporal sequences and long-range dependencies more accurately than the others especially when the amount of training data is large. LSTM has been successfully applied in both the RNN-HMM hybrid systems [21, 22] and the end-to-end system [23, 24].

In this work, the deep LSTM-HMM hybrid system is utilized for acoustic modeling. A forced alignment is first generated by a GMM-HMM system and is then used as the frame-level acoustic targets which the LSTM attempts to classify. The LSTM is trained with cross-entropy objective function using truncated BPTT. In this paper, to connect the deep LSTM with the adaptive LSTM beamformer, we compute log Mel filterbank z_t from the beamformed STFT coefficients.

$$q_t = W_{z,p} z_t \tag{10}$$
$$s_t = \text{LSTM}^{AM}(q_t, s_{t-1}) \tag{11}$$
$$y_t = \text{softmax}(W_{s,y} s_t) \tag{12}$$

q_t is the projection of z_t into a high-dimensional space and y_t is the senone posterior.

2.4. Integrated Network of LSTM Adaptive Beamformer and Deep LSTM Acoustic Model

In order to achieve robust speech recognition by making use of multichannel speech signals, LSTM beamformer in Section 2.2 and the deep LSTM acoustic model in Section 2.3 need to be jointly optimized with the objective of maximizing the ASR performance. In other words, the beamforming LSTM needs to be concatenated with the LSTM acoustic model to form an integrated network that takes multichannel STFT coefficients as the input and produces senone posteriors as illustrated in Fig. 1. The deep LSTM has three hidden layers in our experiments but only one is shown here for simplicity.

To train the integrated LSTM network, we connect the beamforming network (2) – (6), log Mel filtering (8), and the acoustic model (10) – (12) as a single feed forward network, and back-propagate the gradient of the cross-entropy objective function through the network so that both the adaptive beamformer and the acoustic model are optimized for the ASR task by using multichannel training data.

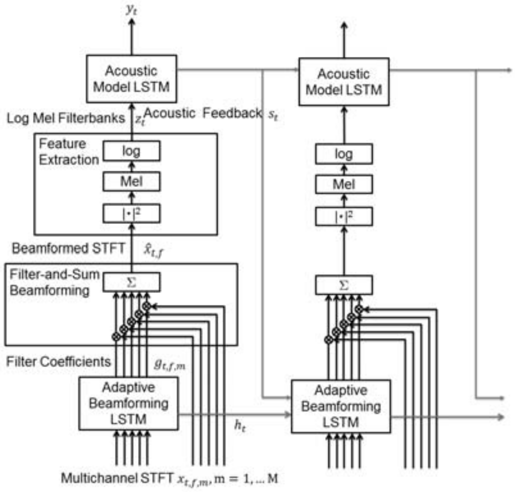

Fig. 1. The unfolded integrated network of an LSTM adaptive beamformer and an LSTM acoustic model. The acoustic feedback (in blue) is introduced to allow the hidden units in LSTM acoustic model to assist in predicting the filter coefficient at current time.

On top of that, we feed the hidden units of the top hidden layer of the deep LSTM acoustic model back to the input of the LSTM beamformer as the auxiliary feature to predict the filter coefficients at next time. By introducing the acoustic model feedback, the Eq. (5) is re-written as

$$h_t = \text{LSTM}^{BF}((p_t, s_{t-1}), h_{t-1}) \tag{13}$$

where (p_t, s_{t-1}) is the concatenation of the acoustic feedback from previous time s_{t-1} and the current projection p_t.

Direct training of the integrated network easily falls into a local optimum as the gradients for the LSTM beamformer and the deep LSTM acoustic model have different dynamic ranges. For a robust estimation of the model parameters, the training should be performed in sequence as shown in Algorithm 1.

Algorithm 1 Train LSTM adaptive beamformer and deep LSTM acoustic model

```
1: Train a deep LSTM acoustic model with log Mel filterbank feature extracted from the speech of all channels to minimize the cross-entropy objective.
2: Initialize the integrated network with the deep LSTM acoustic model in Step 1.
3: Train the integrated network with the ASR cross-entropy objective, update only the parameters in the LSTM beamformer.
4: Jointly train the integrated network in Step 3 with the ASR cross-entropy objective, updating all parameters in the LSTM beamformer and deep LSTM acoustic model.
5: Introduce the acoustic feedback and re-train the integrated network with the ASR objective, updating all the parameters.
```

3. EXPERIMENTS

3.1. CHiME-3 Dataset

The CHiME-3 dataset is released with the 3rd CHiME speech Separation and Recognition Challenge [25], which incorporates the Wall Street Journal corpus sentences spoken by talkers situated in challenging noisy environment recorded using a 6-channel tablet based microphone array. CHiME-3 dataset consists of both real and simulated data. The real data is recorded speech spoken by actual talkers in four real noisy environments (on buses, in cafés, in pedestrian areas, and at street junctions). To generate the simulated data, the clean speech is first convoluted with the estimated impulse response of the environment and then mixed with the background noise separately recorded in that environment [4]. The training set consists of 1600 real noisy utterances from 4 speakers, and 7138 simulated noisy utterances from the 83 speakers in the WSJ0 SI-84 training set recorded in 4 noisy environments. There are 3280 utterances in the development set including 410 real and 410 simulated utterances for each of the 4 environments. There are 2640 utterances in the test set including 330 real and 330 simulated utterances for each of the 4 environments. The speakers in training set, development set and the test set are mutually different (i.e., 12 different speakers in the CHiME-3 dataset). The training, development and test data are all recorded in 6 different channels. The WSJ0 text corpus is also used to train the language model.

3.2. Baseline System

The baseline system is built with Chainer [26] and Kaldi [27] toolkits. 40-dimensional log Mel filterbank features extracted by Kaldi from all 6 channels are used to train a deep LSTM acoustic model using Chainer. The LSTM has 3 layers and each hidden layer has 1024 units. The output layer has 1985 units, each of which corresponds to a senone target. The input feature is first projected to a 1024 dimensional space before being fed into the LSTM. The forced alignment generated by a GMM-HMM system trained with data from all 6 channels is used as the target for LSTM training. During evaluation, only the development and test data from the 5th channel is used for testing (only for the baseline system). The LSTM is trained using BPTT with a truncation size of 100 and a learning rate of 0.01. The batch size for stochastic gradient descent (SGD) is 100. The WER performance of the baseline system is shown in Table 1.

3.3. LSTM Adaptive Beamformer

The 257-dimensional complex STFT coefficients are extracted for the speech in channels $1, 3, 4, 5, 6$. The real and imaginary parts of STFT coefficients from all the 5 channels are concatenated together to form $257 \times 2 \times 5 = 2570$ dimensional input of the beamforming LSTM. The input is projected to 1024 dimensional space before being fed into the LSTM. The beamforming LSTM has one hidden layer with 1024 units. The hidden units vector is projected to 5 sets of $257 \times 2 = 514$ dimensional filter coefficients for adaptively beam-

System	Input Feature	Simu Dev	Real Dev	Simu Test	Real Test
AM (baseline)	Fbank	16.15	19.24	23.02	32.88
BeamformIt+AM	STFT	14.32	12.99	24.36	21.21
BF+AM (fixed)	STFT	15.23	15.01	23.14	25.64
BF+AM	STFT	14.43	15.19	22.40	25.13
BF+AM+Feedback	STFT	14.28	15.10	22.23	24.91

Table 1. The WER performance (%) of the baseline LSTM acoustic model (AM), BeamformIt-enhanced signal as the input of the AM, joint training of LSTM beamformer and LSTM acoustic model (BF+AM) with or without acoustic feedback.

forming signals from 5 channels using Eq. (2). The MSE objective is computed between the beamformed signal and BeamformIt [28]. The beamforming LSTM is trained using BPTT with a truncation size of 100, a batch size of 100 and a learning rate of 1.0.

3.4. Joint Training of the Integrated Network

The baseline LSTM acoustic model trained in Section 3.2 and the LSTM adaptive beamformer trained in Section 3.3 are concatenated together as the initialization of the integrated network. A feature extraction layer is inserted in between the two LSTMs to extract 40-dimensional log Mel filterbank features with Eq. (8). The integrated network is trained in a way described in Steps 3, 4 and 5 of Section 2.4. BPTT with a truncation size of 100 and a batch size of 100 and a learning rate of 0.01 is used for training. The data from all 5 channels in the development and test set is used for evaluating the integrated network. The WER performance for different cases are shown in Table 1.

3.5. Result Analysis

From Table 1, the best system is the integrated network of an LSTM adaptive beamformer and a deep LSTM acoustic model with the acoustic feedback, which achieves 14.28%, 15.10%, 22.23%, 24.91% WERs on the simulated development set, real development set, simulated test set and real test set of the CHiME-3 dataset respectively. The joint training of the integrated network without updating the deep LSTM acoustic model achieves absolute gains of 0.92%, 4.23% and 7.24% over the baseline system on the simulated development set, real development set and real test set respectively. The joint training of the integrated network with all the parameters updated achieves absolute gains of 1.72%, 4.05%, 0.62% and 7.75% respectively over the baseline systems on the simulated development set, real development set, simulated test set and real test set respectively. The large performance improvement justifies that the LSTM adaptive beamformer is able to estimate the real-time filter coefficients adaptively in response to the changing source position, environmental noise and room impulse response with the LSTM acoustic model jointly trained to optimize the ASR objective. Further absolute gains of 0.15%, 0.09%, 0.17% and 0.22% are achieved with the introduction of acoustic feedback, which indicates that the high-level acoustic information is also helpful in predicting the filter coefficients at the next time step.

Note that although the proposed system with acoustic feedback achieves 0.04% and 2.13% absolute gains over the beamformed signal generated by BeamformIt on the simulated development and test sets, it does not work as well as the BeamformIt on the real development and test sets. In BeamformIt implementation, the two-step time delay of arrival Viterbi postprocessing makes use of both the past and future information in predicting the best alignment of multiple channels at the current time, while in our system, only the history in the past is utilized to estimate the current filter coefficients. This may explain the differences in WER performance and can be alleviated by using bidirectional LSTM as part of the future work.

3.6. Beamformed Feature

The LSTM beamformer adaptively predicts the time-variant beamforming coefficients and performs filter-and-sum beamforming over the 5 input channels. The log Mel filter bank feature is obtained from the STFT coefficients. From Fig. 2, we see that the log Mel filter bank feature obtained from the LSTM adaptive beamformer is quite similar to the log Mel filter bank feature extracted from the STFT coefficients beamformed by BeamformIt for the same utterance. The SNR is not high but matches the LSTM acoustic model well for maximizing the ASR performance.

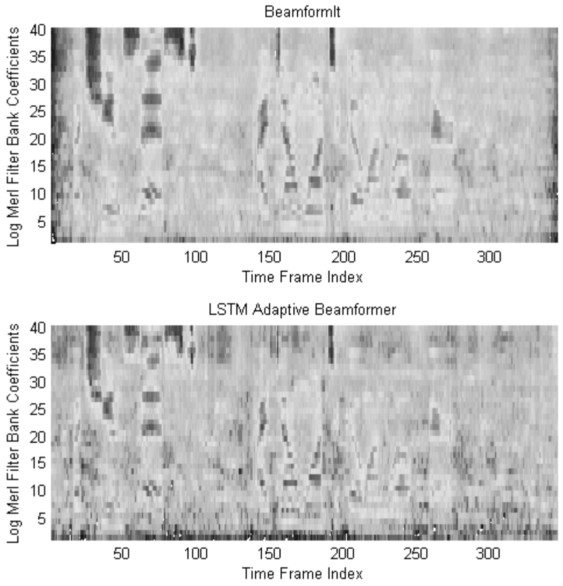

Fig. 2. The comparison of the log Mel filter bank coefficients of the same utterance extracted from STFT coefficients beamformed by BeamformIt (upper) and LSTM adaptive beamformer (lower) .

4. CONCLUSIONS

In this work, LSTM adaptive beamforming is proposed to adaptively predict the real-time beamforming filter coefficients to deal with the time-variant source location, environmental noise and room impulse response inherent in the multichannel speech signal. To achieve robust ASR, the LSTM adaptive beamformer is jointly trained with a deep LSTM acoustic model to optimize the ASR objective. This framework achieves absolute gains of 1.72%, 4.05%, 0.62% and 7.75% over the baseline system on the CHiME-3 dataset. Further improvement is achieved by introducing the acoustic feedback to assist in predicting the filter coefficients. However, our approach does not work as well as the BeamformIt on real data and we will look into this in the future.

5. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "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, Nov 2012.

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

[3] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Espi, Shoko Araki, Takaaki Hori, and Tomohiro Nakatani, "Strategies for distant speech recognitionin reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, 2015.

[4] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 475–481.

[5] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[6] Barry D Van Veen and Kevin M Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[7] H Erdogan, JR Hershey, S Watanabe, M Mandel, and J Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," 2016.

[8] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, July 2007.

[9] 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 *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5745–5749.

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

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

[12] J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wllmer, B. Schuller, and G. Rigoll, "Memory-enhanced neural networks and nmf for robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1037–1046, June 2014.

[13] Felix Weninger, Jrgen Geiger, Martin Wllmer, Bjrn Schuller, and Gerhard Rigoll, "Feature enhancement by deep {LSTM} networks for {ASR} in reverberant multisource environments," *Computer Speech & Language*, vol. 28, no. 4, pp. 888 – 902, 2014.

[14] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 708–712.

[15] Bo Li, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. Interspeech*, 2016.

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

[17] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[18] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "Lstm neural networks for language modeling.," in *Interspeech*, 2012, pp. 194–197.

[19] C. Weng, D. Yu, S. Watanabe, and B. H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5532–5536.

[20] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust asr," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4085–4088.

[21] 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*, Dec 2013, pp. 273–278.

[22] Hasim Sak, Andrew W Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling.," in *INTERSPEECH*, 2014, pp. 338–342.

[23] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[24] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[25] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 504–511.

[26] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[27] 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*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[28] 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, Sept 2007.