Multi-Channel Speech Recognition: LSTMs All the Way Through

Hakan Erdogan[2]*, Tomoki Hayashi*[1]*, John R. Hershey*[1]*, Takaaki Hori*[1]*, Chiori Hori*[1]
Wei-Ning Hsu[1]*, Suyoun Kim*[1]*, Jonathan Le Roux*[1]*, Zhong Meng*[1]*, Shinji Watanabe*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge MA, USA
[2]Sabanci University, Istanbul, Turkey

haerdogan@sabanciuniv.edu,

{hayashi,hershey,thori,chori,whsu,skim,leroux,zmeng,watanabe}@merl.com

Abstract

Long Short-Term Memory recurrent neural networks (LSTMs) have demonstrable advantages on a variety of sequential learning tasks. In this paper we demonstrate an LSTM "triple threat" system for speech recognition, where LSTMs drive the three main subsystems: microphone array processing, acoustic modeling, and language modeling. This LSTM trifecta is applied to the CHiME-4 distant recognition challenge. Our previous state-of-the-art ASR systems for the previous CHiME challenge employed LSTM mask estimation based beamforming, noise robust features, in addition to DNN/RNNLM based back end. The proposed system refines each module of the previous system including bidirectional LSTM (BLSTM) mask estimation based beamforming, BLSTM-DNN hybrid acoustic model, and language model rescoring based on LSTM. We perform constrained re-estimation based speaker adaptation, and also prepare several complementary systems by changing the beamforming strategy and the acoustic model configurations, and combine these systems based on word-posterior based system combination. The final system achieved 2.98% WER for the real test set in the 6-channel track, which reduces the WER from the baseline by 8.5% absolute, and also outperforms our previous CHiME-3 system by 6.1% absolutely.

1. Background

The MERL-Sabanci system, as shown in Figure 1, is a multi-channel ASR system that focuses on the CHiME-4 6ch track [1]. It is an extension of our CHiME-3 system [2], and improves upon it using the following methods:

- BLSTM mask estimation for Minimum Variance Distortionless Response (MVDR) and Generalized EigenVector (GEV) beamformers.
- BLSTM-DNN hybrid acoustic model via state posterior combination.
- Expanded noisy data training using all 6 channels of official training speech data (i.e., 6 times the amount of training data).
- Unsupervised speaker adaptation based on constrained retraining of DNN.
- Language model re-scoring based on LSTM.
- System combination across multiple methods and input features.

The authors are listed in the alphabetic order. Tomoki Hayashi, Wei-Ning Hsu, Suyoun Kim, and Zhong Meng performed the work during their internship programs at MERL.

These techniques steadily improve the performance from the baseline. Their technical details are explained in the following section.

2. Contributions

2.1. BLSTM mask estimation for beamformers

We train a unique BLSTM neural network for single-channel mask prediction using the simulated training data for all six channels. The network takes a single channel as input and predicts both speech and noise masks for that channel using sigmoid output activations and ideal binary masks as targets. The network is trained with the binary cross-entropy loss function [3]. During recognition, the network is applied separately to each channel, and the predicted masks for the six channels are combined to obtain a single mask by taking their median. The obtained speech and noise masks are then used to predict speech and noise spatial covariance matrices which are used in MVDR and GEV beamformers to perform beamforming-based enhancement on the multi-channel signal to be recognized.

2.2. Beamforming

We perform MVDR and GEV beamforming. The version of MVDR beamforming we use only uses spatial covariance estimates of speech and noise. To obtain these spatial covariances, we make use of the masks predicted by the network. The covariances are estimated as follows:

$$\hat{\mathbf{\Phi}}_x(f) = \frac{\sum_t \hat{M}_x(t,f)\boldsymbol{Y}(t,f)\boldsymbol{Y}^H(t,f)}{\sum_t \hat{M}_x(t,f)},$$

where $\hat{M}_x$ is the predicted mask for speech or noise and $\boldsymbol{Y}(t,f)$ is the received multi-channel signal's spatial vector corresponding to time-frequency bin (t,f). For GEV beamforming [3], we form the beamforming filters by maximizing the SNR for each frequency by solving the generalized eigenvalue problem for the spatial filter $\boldsymbol{h}$:

$$\hat{\mathbf{\Phi}}_{\text{speech}}\boldsymbol{h} = \lambda\hat{\mathbf{\Phi}}_{\text{noise}}\boldsymbol{h}.$$

For MVDR beamforming, we first choose a reference microphone and then find the direction of minimum noise variance while keeping the speech signal distortionless. Using one possible formulation [4], the solution can be found as:

$$\hat{\boldsymbol{h}} = \frac{1}{\text{trace}(\hat{\mathbf{\Phi}}_{\text{noise}}^{-1}\hat{\mathbf{\Phi}}_{\text{speech}})}\hat{\mathbf{\Phi}}_{\text{noise}}^{-1}\hat{\mathbf{\Phi}}_{\text{speech}}\boldsymbol{e}_{\text{ref}},$$

where $\boldsymbol{e}_{\text{ref}}$ is a standard unit vector in direction ref.

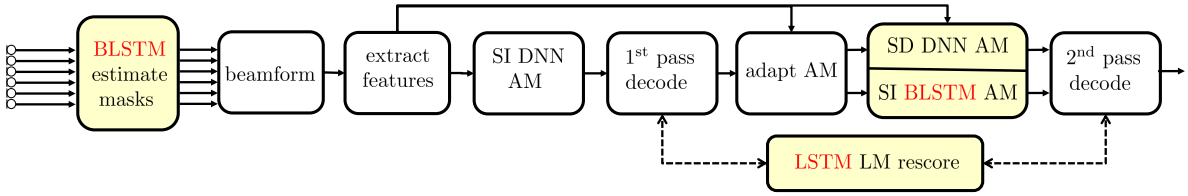

Figure 1: A flow chart of the proposed system for decoding.

2.3. Acoustic modeling and adaptation

Although RNNs (especially LSTMs) have been shown to be very effective for noise robust speech recognition [5, 6], our preliminary attempt at applying LSTMs/BLSTMs to the CHiME-4 task was not successful probably due to the limited amount of training data and the difficulty of obtaining correct state alignments from noisy speech. Instead, the acoustic models we used in our experiments are hybrid BLSTM-DNN systems. BLSTM and DNN models are separately trained with augmented training data by using the noisy speech training data from all 6 channels [7]. The DNN model configuration is the same as that of the official baseline acoustic model [1]: a 7 hidden layer sigmoid DNN with 2048 activations per layer trained by using state-level Minimum Bayes Risk (sMBR) criterion in the kaldi nnet1 module [8]. The BLSTM acoustic model has 3 layers, where each layer consists of forward and backward unidirectional LSTMs with 512 cell states and one linear bottleneck layer to combine the outputs of both unidirectional LSTMs outputting 512 activations. The BLSTM was trained based on the cross entropy criterion by using stochastic gradient descent. We used a state alignment obtained by using the DNN as a target.

On top of the training, we adapt the speaker-independent DNN to the data of each speaker in an unsupervised way. We used a constrained re-training (CRT) adaptation method where we re-estimate the DNN parameters of only a subset of layers while holding the remaining parameters fixed with the cross entropy criterion. The optimal subset of layers to be estimated is selected according to the development set performance. Since we cannot use any prior knowledge about the environment according to the CHiME-4 regulation, we train each speaker-dependent DNN with the speaker's speech from all different environments. We also use KL divergence adaptation [9] by using the speaker-independent DNN to regularize the speaker-dependent DNN. The adaptation target (1-best alignment) was obtained at the first-pass decoding, and the second-pass decoding is performed using this speaker-adapted DNN, as shown in Figure 1.

The BLSTM acoustic model and DNN model adaptation are implemented by using chainer deep learning toolkit [10].

2.4. Language model re-scoring

We train an LSTM-based RNN language model (LSTMLM) using the official training data for language modeling in CHiME-4.

RNN language models (RNNLMs) [11] robustly estimate word probability distributions by representing the contextual information in a continuous space, which are kept in the hidden layer with recurrent connections. Compared to N-gram models, RNNLMs can exploit more long-distance interword dependencies to predict the next word, and yield better performance in many tasks. However, RNNs are not able to keep very long histories because the contextual information at a certain time exponentially decays by doing recurrent propagations through time. Accordingly, we introduce LSTMLM [12, 13] to improve the system performance. The LSTM RNN has a memory cell in each hidden unit instead of a regular network unit, which can remember the contextual information for an arbitrary length of time. By expoiting the longer contextual information, LSTMLM can predict the next word more accurately than the standard RNNLMs.

In the decoding phase, word lattices are first generated using the baseline language model for CHiME-4, which is the standard 5k WSJ trigram downsized with an entropy pruning technique [14]. After that, N-best lists are generated from the lattices using a 5-gram language model with a modified Kneser-Ney smoothing [15, 16]. Finally, the N-best lists are rescored using a linear combination of the 5-gram and LSTMLM probabilities in the log domain, i.e.,

$$\log P(W) = \sum_{i=1}^{L} \{\lambda \log P_{lstm}(w_i|h_i) + (1-\lambda) \log P_{5gkn}(w_i|h_i)\}, \quad (1)$$

where $W = w_1, w_2, \ldots, w_L$ denotes each sentence hypothesis, λ the interpolation weight, and h_i the history of w_i.. The best-rescored hypothesis is selected as the result of each single system. The N-best lists are also used for system combination.

For the challenge, LSTMLM was designed as an RNN with one projection layer of 1000 units and one hidden layer of 1500 LSTM cells. We set the interpolation weight λ in Eq. (1) to 0.9 and the number of N-best hypotheses to 100, which were selected based on word error rate for the development set.

2.5. System combination

In the proposed system, multiple feature vector sequences are obtained for different pairs of beamforming and feature extraction methods, and they are separately processed by a WFST-based decoder to output word lattices. After rescoring with the LSTMLM, multiple lists of N-best hypotheses are obtained and then used for system combination.

System combination is a technique to improve recognition accuracy by combining different recognition hypotheses from different systems [17]. First, the multiple hypotheses are combined by taking their union after reweighting each hypothesis with its posterior probability. After that, minimum Bayes risk (MBR) decoding is performed on the combined hypotheses using an algorithm in [18]. With this decoding, we can find the hypothesis with the minimum expected word error rate from among all the hypotheses obtained by the multiple systems.

3. Experimental evaluation

3.1. Mask prediction network and beamforming setup

For mask prediction and beamforming, we used windows of length 1024 samples with a frame shift of 256 samples. The non-redundant FFT vector dimension was 513. Magnitude FFT was used as an input to the mask prediction network for each frame. The mask prediction network had a single BLSTM layer with 256 nodes. After the BLSTM layer, we used two feedforward layers with rectified linear unit activations with an output dimension of 513. The output layer predicted predicted 513 dimensional masks for speech and noise separately with a sigmoid activation for each output. The target ideal binary masks did not sum to one for each time-frequency bin. Ideal binary masks were chosen to be one when the corresponding source was significantly larger than the other source.

For beamforming, we pass each channel's input through the network, take the median of each channel's outputs for each time-frequency bin and use the value as a mask directly. For the MVDR beamformer, we chose microphone CH5 as the reference microphone.

3.2. Experimental Results

The first set of experiments compare the baseline script (BeamformIt [19], DNN sMBR, and 3-gram) with two beamforming techniques. Table 1 summarizes results for three types of beamforming, and both methods using the BLSTM based masks greatly improve the performance from BeamformIt. The training utilizes noisy data from channel 5 only. Also we observed that the GEV beamformer yields similar performance on simulated versus real data, both for the development and for the test sets, whereas the MVDR beamformer has systematically better performance on the simulation data. Because these properties are complementary, both beamformers are included in the final system combination.

Table 1: Average WER (%) for the front-end systems with fixed DNN sMBR, 3-gram back-end.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	Baseline: BeamformIt	8.14	9.07	15.00	14.23
	BLSTM-Mask MVDR	6.66	5.55	11.39	6.39
	BLSTM-Mask GEV	7.19	7.50	10.32	9.62

The second group of experiments compares acoustic model techniques with fixed front end based on BeamformIt [19]. Table 2 shows that using all 6 channels for training is particularly effective for generalization to the test set, presumably due to the increase in speech signal variety in the training data. Although an individual BLSTM acoustic model does not outperform the DNN sMBR, the state posterior patterns of both models seem to be complementary, and the hybrid BLSTM-DNN acoustic model achieves significant improvement. Based on the result, we adopt BLSTM-DNN acoustic model as the main system, but still use DNN sMBR as a complementary system to investigate several features and training variations due to its lower computational cost.

Table 3 reports the results on combined front-end techniques and BLSTM-DNN acoustic modeling. Here, we also report the speaker adaptation and language model re-scoring on top of the systems for both BLSTM-Mask MVDR and GEV beamformers. Note that the speaker adaptation is only performed for the DNN part of the BLSTM-DNN. The table clearly shows the improvement of the combination of beamforming and BLSTM-DNN from Tables 1 and 2, and the effectiveness of the LM re-scoring and speaker adaptation is also confirmed.

Table 2: Average WER (%) for the back-end systems with fixed BeamformIt front-end.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	Baseline: DNN sMBR 3gram	8.14	9.07	15.00	14.23
	6ch Training	7.71	8.21	12.79	12.67
	BLSTM 6ch Training	8.50	8.96	13.59	13.28
	BLSTM-DNN	7.44	7.48	11.51	11.51

Table 3: Average WER (%) for combined single systems with BLSM-mask beamformers, BLSTM-DNN, LM re-scoring using LSTM, and speaker adaptation

Track	System	Dev		Test	
		real	simu	real	simu
6ch	BLSTM-Mask MVDR	5.80	4.68	8.57	5.23
	+ LM re-scoring	2.92	2.27	4.83	2.51
	+ Adaptation	2.54	1.95	4.18	1.84
	BLSTM-Mask GEV	6.26	6.34	8.13	7.83
	+ LM re-scoring	3.12	3.11	4.23	4.06
	+ Adaptation	2.77	2.63	3.81	2.94

Finally we have combined our main BLSTM-DNN systems with DNN sub systems. In addition to the two beamformer results (MVDR and GEV) in Table 3, we have additionally prepared comparable systems by changing features with PNCC [20] (pncc), ETSI AFE [21] (afe), and PLP [22] with pitch features (plp+p) using DNN sMBR acoustic models (DNN), retrained DNN with beamformed features (DNN, ret), and using alternative implementation of the BLSTM-Mask GEV beamformer by [3] (GEV [3]). After that, we have combined all the lattices obtained by these systems and performed system combination using minimum Bayes risk decoding.

Table 4: Average WER (%) with final system combination.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	GEV [3], DNN	2.62	2.58	3.74	3.26
	GEV, DNN, ret	2.63	2.48	3.63	2.87
	MVDR, DNN, ret	2.47	1.79	4.13	1.67
	MVDR, afe, DNN	3.20	2.89	4.99	2.57
	GEV, pncc, DNN	3.62	3.68	5.49	4.66
	GEV [3], plp+p, DNN	3.31	3.26	4.85	4.43
	Combination	2.11	1.95	2.98	1.97

Using these complementary systems, the system combination achieved 2.98%, representing an improvement of 0.6% absolute over our best single system.

Table 5: WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
6ch	BUS	2.73	1.53	4.30	1.59
	CAF	2.08	2.14	2.71	1.83
	PED	1.78	1.67	2.37	1.81
	STR	1.81	1.92	3.10	1.72

4. Summary

This paper describes the MERL/Sabanci submission system for the CHiME-4 speech separation and recognition challenge. Our main single system consists of BLSTM-mask-estimation based beamforming, DNN-BLSTM hybrid acoustic model, and rescoring based on LSTMLM, leading to a system that employs LSTMs all the way through. With acoustic model adaptation and system combination, we finally obtained 2.98% WER, placing third among 15 submissions. Future work will consider how to integrate these complicated modules within a deep learning framework, including beamforming network [23, 24] and end-to-end ASR [25, 26, 27].

5. References

[1] 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*, to appear.

[2] 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 *Proc. ASRU*, 2015, pp. 475–481.

[3] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016.

[4] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. ASLP*, vol. 18, no. 2, pp. 260–276, 2010.

[5] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2014, pp. 5532–5536.

[6] Z. Chen, S. Watanabe, H. Erdoğan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. Interspeech*, 2015.

[7] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU*, 2015, pp. 436–443.

[8] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013, pp. 2345–2349.

[9] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*, 2013, pp. 7893–7897.

[10] S. Tokui, K. Oono, S. Hido, and J. 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.

[11] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, , and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010, pp. 1045–1048.

[12] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling." in *Proc. Interspeech*, 2012.

[13] T. Hori, C. Hori, S. Watanabe, and J. R. Hershey, "Minimum word error training of long short-term memory recurrent neural network language models for speech recognition," in *Proc. ICASSP*, 2016, pp. 5990–5994.

[14] A. Stolcke, "Entropy-based pruning of backoff language models," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 270–274.

[15] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," in *Proc. ICASSP*, 1995, pp. 181–184.

[16] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. ACL*, 1996, pp. 310–318.

[17] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. NIST Speech Transcription Workshop*, 2000.

[18] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[19] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[20] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*. IEEE, 2012, pp. 4101–4104.

[21] "ETSI - speech recognition: Advanced front-end feature extraction algorithm," ttp://www.etsi.org/technologies-clusters/technologies/past-work/speech-recognition.

[22] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[23] 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 *Proc. ICASSP*, 2016, pp. 5745–5749.

[24] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. Interspeech*, 2016.

[25] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks." in *ICML*, vol. 14, 2014, pp. 1764–1772.

[26] D. Bahdanau, J. Chorowski, D. Serdyuk, Y. Bengio *et al.*, "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP*, 2016, pp. 4945–4949.

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