The MELCO/MERL System Combination Approach for the Fourth CHiME Challenge

Yuuki Tachioka[1], *Shinji Watanabe*[2], *Takaaki Hori*[2]

[1]Information Technology R&D Center, Mitsubishi Electric Corporation
[2]Mitsubishi Electric Research Laboratories
Tachioka.Yuki@eb.MitsubishiElectric.co.jp, watanabe@merl.com, thori@merl.com

Abstract

This paper describes our approach for all three tracks of the fourth CHiME challenge. Front-end process prepared two speech enhancements. Back-end process extracted three types of different features and after decoding, it used neural network based rescoring. Finally, the hypotheses of the multiple systems were combined and the word error rate of our best system became less than half of that of the state-of-the-art baseline.

1. Background

The 4th CHiME challenge provides three tracks: 1ch, 2ch, and 6ch track [1]. We entered all three tracks. For all tracks, state-of-the-art baseline scripts were prepared. They employed discriminatively trained deep neural network (DNN) acoustic models and recurrent neural network (RNN) based rescoring with advanced speech enhancement. There are four different environments in the tasks and for these kinds of tasks, system combination was effective. To realize more effective combination, we prepared multiple systems with different speech enhancement and different feature extractions. This paper separately confirmed the effectiveness of our approach in terms of the word error rate (WER).

2. Front-end process

For single-channel track, sparse non-negative matrix factorization (NMF) [2] was used to suppress noise. To reduce distortions, enhanced speech was mixed with original noisy speech. For multi-channel track, in addition to the provided beamformer (BeamformIt), minimum variance distortionless response (MVDR) beamformer with precise steering vector estimation [3] was employed.

3. Back-end process

In addition to the provided 13-dimensional MFCC $+\Delta+\Delta\Delta$ with feature-space maximum likelihood linear regression (fM-

Table 1: System description for Table 2. All systems used DNN acoustic model.

{m,p,f}-{s,m}-{n,s,b,m}-{u,a,a2}+{r,l}	
{m,p,f}	MFCC / PLP / fbank
{s,m}	Single / multi-channel data training
{n,s,b,m}	Noisy / sparse NMF / BeamformIt / MVDR
{u,a,a_2}	Unadapted / adapted / adapted-2 DNN
{r,l}	RNN / LSTM-LM rescoring

Table 2: Average WER [%] for the tested systems. For 1ch, "baseline1" was "m-s-n-u" and "baseline2" was "m-s-n-u+r". For 2ch and 6ch, "baseline1" was "m-s-b-u" and "baseline2" was "m-s-b-u+r". "best" combined asterisk-marked systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	baseline1	14.67	15.67	27.69	24.15
	baseline2	11.69	15.43	23.71	20.95
	m-m-n-u	12.67	13.55	22.17	20.29
	m-m-n-u+l*	7.76	8.92	15.66	15.12
	p-m-n-u+l*	7.74	9.23	16.03	15.31
	f-m-n-u+l*	5.60	7.60	11.76	12.75
	f-m-n-a+l*	5.58	7.70	11.85	12.72
	m-m-s-u+l*	7.78	8.86	15.49	15.08
	p-m-s-u+l*	7.60	9.33	15.47	15.61
	f-m-s-u+l*	5.56	7.30	11.64	12.76
	f-m-s-a+l*	5.41	7.48	11.64	12.90
	best	5.15	7.15	11.13	12.15
2ch	baseline1	10.90	12.36	20.44	19.03
	baseline2	9.63	10.72	18.08	16.88
	m-m-b-u	9.90	10.60	16.89	16.27
	m-m-b-u+l*	5.59	6.33	11.43	10.55
	p-m-b-u+l*	5.51	6.48	11.71	10.77
	f-m-b-u+l*	4.19	5.23	8.38	9.10
	f-m-b-a+l*	3.96	5.15	8.23	8.49
	m-m-m-u+l*	5.34	6.09	11.21	11.55
	p-m-m-u+l*	5.03	6.40	11.11	11.61
	f-m-m-u+l*	3.96	5.23	8.45	9.62
	f-m-m-a+l*	3.80	5.06	7.99	9.10
	best	3.50	4.63	7.28	8.03
6ch	baseline1	8.14	9.07	15.04	14.20
	baseline2	5.75	6.77	11.47	10.91
	m-m-b-u	7.69	8.23	12.57	12.66
	m-m-b-u+r	4.99	5.72	9.22	8.96
	m-m-b-u+l*	3.94	4.49	7.77	7.51
	p-m-b-u+l*	3.90	4.62	7.64	7.71
	f-m-b-u+r	4.18	4.95	7.20	7.47
	f-m-b-u+l*	3.10	3.63	5.94	6.28
	f-m-b-a+l*	3.05	3.60	5.71	5.94
	m-m-m-u+r	4.45	4.19	7.45	7.51
	m-m-m-u+l*	3.47	3.06	6.42	6.39
	p-m-m-u+l*	3.43	2.99	6.36	6.23
	f-m-m-u+r	3.72	3.66	6.11	6.67
	f-m-m-u+l*	2.75	2.61	5.19	5.72
	f-m-m-a+l*	2.60	2.53	5.06	5.01
	f-m-m-a_2+l*	2.47	2.45	4.75	4.39
	best	2.30	2.32	4.31	4.18

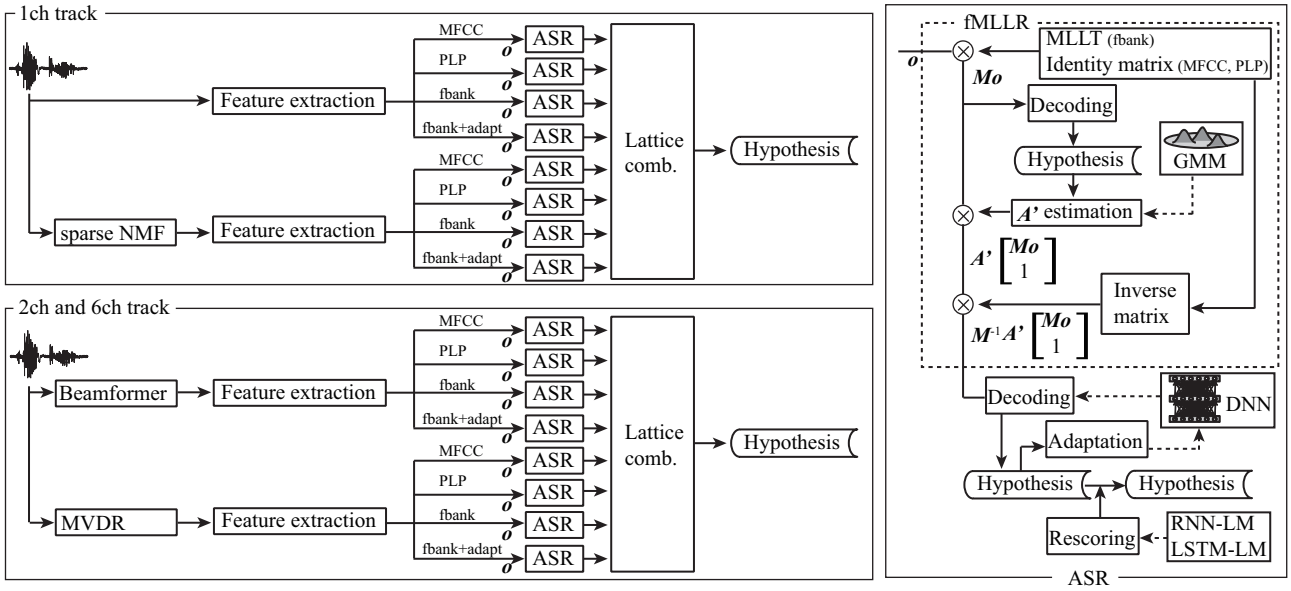

Figure 1: Schematic diagram of the proposed ASR systems.

Table 3: WER [%] per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	7.15	6.24	18.00	8.55
	CAF	5.19	9.81	11.73	13.93
	PED	3.05	4.97	7.81	11.71
	STR	5.19	7.57	6.99	14.40
2ch	BUS	4.54	3.92	11.42	5.08
	CAF	3.63	6.28	7.08	9.41
	PED	2.21	3.38	5.59	8.33
	STR	3.63	4.96	5.04	9.28
6ch	BUS	3.07	2.01	5.16	2.95
	CAF	2.40	2.99	3.90	4.63
	PED	1.64	1.76	4.00	4.18
	STR	2.11	2.51	4.17	4.97

LLR) transformation, we employed 13-dimensional PLP $+\Delta+\Delta\Delta$ with fMLLR transformation and 40-dimensional filterbank (fbank) feature $+\Delta+\Delta\Delta$ with maximum likelihood linear transformation (MLLT) and fMLLR transformation [4]. Features in the consecutive 11 frames were input to the DNN.

In addition to the feature-space adaptation, model-space adaptation of DNN [5] was also used where the second layer of DNN was switched for each speaker. To train DNN acoustic models, multi-channel (6ch) data were all used whereas baseline only used single-channel data. These modification increased the training data size [3]. All training data were noisy without any speech enhancement, i.e., noisy data training.

After decoding, we used long short-term memory (LSTM)-language model (LM) rescoring [6] instead of the baseline recurrent neural network (RNN)-LM. Figure 1 shows the schematics of the proposed method. In each track, there were two types of speech enhancement. For each enhancement, three different features were used; and for fbank feature, model-space speaker adaptation was performed. In total, hypotheses of eight systems are combined by using lattice combination.

4. Experimental evaluation

Table 2 shows the WERs of the challenge. Descriptions of the system ID is shown in Table 1. Comparison of baseline1 and "m-m-n-u" shows the effectiveness of multi-channel data training, which was especially effective for 1ch track and improved the WERs by around 2–5%. Comparison of baseline1 and baseline2 and that of "m-m-n-u" and "m-m-n-u+l" show the effectiveness of LSTM-LM rescoring, which improved WER more than RNN-LM rescoring. The performances of MFCC and PLP features were almost equivalent but fbank feature significantly improved the WERs. DNN model adaptation was also effective. MVDR beamformer shows its effectiveness for the 6ch track more than 2ch track, compared with the baseline beamformer. Combining multiple systems additionally improved WERs by around 0.3–0.6%. WERs of the best system were less than half of those of "baseline2" except one case (Test and simu in the 1ch track).

Table 3 shows the WER of the best system per environment in Table 2. Increasing the number of microphones was effective for all conditions. In real data, "BUS" was the most difficult task.

5. Conclusion

This paper showed our approach for the fourth CHiME challenge. Multi-channel data training, fbank feature, and LSTM-LM based rescoring were the most effective. System combination gave additional improvements for all conditions.

6. 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 and Language*, to appear, 2016.

[2] J. Eggert and E. Komer, "Sparse coding and NMF," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 4. IEEE, 7 2004, pp. 2529–2533.

[3] 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 *Proceedings of ASRU*. IEEE, 12 2015, pp. 436–443.

[4] T. N. Sainath, B. Kingsbury, A. R. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, “Improvements to deep convolutional neural networks for LVCSR,” in *Proceedings of ASRU*. IEEE, 12 2013, pp. 315–320.

[5] T. Ochiai, S. Matsuda, H. Watanabe, X. Lu, C. Hori, and S. Katagiri, “Speaker adaptive training for deep neural networks embedding linear transformation networks,” in *Proceedings of ICASSP*. IEEE, 4 2015, pp. 4605–4609.

[6] T. Hori, C. Hori, S. Watanabe, and J. Hershey, “Minimum word error training of long short-term memory recurrent neural network language models for speech recognition,” in *Proceedings of ICASSP*. IEEE, 3 2016.