Unsupervised network adaptation and phonetically-oriented system combination for the CHiME-4 challenge

Yusuke Fujita[1]*, Takeshi Homma*[2]*, Masahito Togami*[1]

[1]Hitachi, Ltd. Research and Development Group, Japan
[2]Hitachi America, Ltd., USA

yusuke.fujita.su@hitachi.com

Abstract

In this paper, we describe our submitted systems for the CHiME-4 challenge and report the experimental results.

We first examine unsupervised speaker adaptation method for deep neural network (DNN) based acoustic model. The speaker-dependent DNN is constructed by re-training the speaker-independent DNN using evaluation data per speaker. Experiments show that the method provides up to 29% relative gain on the word error rate (WER).

Second, we describe a phonetically-oriented system combination method. The method utilizes phonetic similarity to construct a word alignment. It gives a better treatment of insertion and deletion errors in the word alignment. Experiments show that the method provides up to 16% relative gain.

Finally, we combine the above methods with our previous approaches for the submitted system. We utilize multi-output signals from local Gaussian modeling (LGM) based source separation as augmented training data. We also used the LGM as a preprocessing of beamforming at frontend. The submitted system achieved 4.68% of WER for the real evaluation set.

1. Background

We participate in the CHiME-4 challenge [1] and we submit all (1, 2, 6ch) tracks. We explain how the speaker-dependent deep neural network (DNN) is constructed and a new development of system combination method for this challenge. The local Gaussian model (LGM) is also emphasized because it is successfully applied to the speech enhancement for the past CHiME-3 challenge [2].

2. Contributions

2.1. Unsupervised network adaptation

Speaker adaptation is successfully applied in a lot of tasks. The CHiME-4 baseline system employs feature-space maximum likelihood linear regression (fMLLR) transform for speaker adaptation. In the CHiME-3 best paper [3] used re-training of convolution neural network (CNN) for speaker adaptation and reported significant gain of word error rate (WER). While unsupervised re-training of DNN has been shown no improvement in [4], we observed an improvement on the CHiME-4 data set.

Figure 1 shows the decoding process with unsupervised network adaptation. In this work, the baseline DNN trained with state-level minimum Bayes risk criterion (DNN+sMBR) is used as an initial acoustic model for speaker adaptation. Labels for re-training are generated from 1-best decoding results of test data. The decoding for re-training is performed using the initial acoustic model and 3-gram language model, followed by rescoring using the 5-gram and recurrent neural network language model (RNNLM). Alignments of the 1-best decoding results are generated using the initial model. Then, re-training is performed using mini-batch stochastic gradient descent (SGD) algorithm with a cross entropy criterion. The parameters of mini-batch SGD are tuned using the development set.

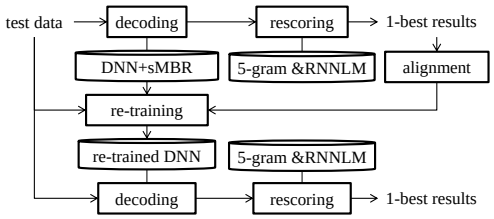

Figure 1: decoding with unsupervised network adaptation

2.2. Phonetically-oriented system combination

The ROVER [5] is a well-known technique to reduce word errors using multiple sentences obtained from multiple systems. In the approach, word alignments among multiple sentences are constructed by word-based DP matching. The word alignment makes a word set which contains words obtained from different systems in the same second. Based on the word alignment, the most trustable word within a word set is chosen. However, the word alignment often generates irrelevant word sets.

The left of Fig. 2 shows such an example: the word "their" from recognizer 1 is put into a word set containing "are" from recognizer 2 and 3. The ideal alignment in this case is that "their" from recognizer 1 is be associated with "there are" from recognizer 2 and "they are" from recognizer 3.

In this study, we employ the phonetically-oriented word alignment (POWA) proposed in [6]. A word alignment example with POWA is shown in the bottom right of Fig. 2.

Based on the POWA-based word set, we perform word selection utilizing machine learning [2].

The feature vector x used for the correct word estimators is formed as:

$$\mathbf{x} = (\mathbf{x}_{oc}^\top, \mathbf{x}_{cf}^\top, \mathbf{x}_{nl}^\top)^\top \in \mathbb{R}^{\binom{N}{2}+2N} \quad (1)$$

$$\mathbf{x}_{oc} = (\delta(w_i, w_j); 1 \le i < j \le N)^\top \in \mathbb{R}^{\binom{N}{2}} \quad (2)$$

$$\mathbf{x}_{cf} = (c_i; 1 \le i \le N)^\top \in \mathbb{R}^N \quad (3)$$

$$\mathbf{x}_{nl} = (\delta(w_i, \mathrm{NULL}); 1 \le i \le N)^\top \in \mathbb{R}^N \quad (4)$$

where $\delta()$ is the Kronecker delta function, N is the number of

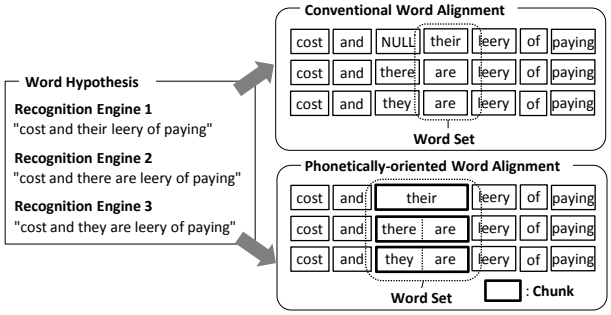

Figure 2: Phonetically-oriented word alignment

recognizers, each element of $\mathbf{x}_{oc}$ is an indicator showing the chunk from a recognizer i is the same with the chunk from another recognizer j , c_i is a confidence of a chunk from a recognizer i, which is calculated as a geometric mean of words' confidences within a chunk, and NULL means the word is empty. The label vector $\mathbf{y}$ is formed as following:

$$\mathbf{y} = (\delta(w_i, w_{true}); 1 \leq i \leq N)^\top \in \mathbb{R}^N \tag{5}$$

where w_{true} means a chunk which consists of correct words. Given feature vector x and label vector y, the correct word estimator is trained by logistic regression model. The correct word estimator was trained from the development set.

2.3. LGM based source separation

2.3.1. Data augmentation using LGM

In this work, we use the data augmentation method using multiple output signals from LGM based source separation [7]. In the LGM based source separation [8], the multi-microphone signal in the time-frequency domain $\mathbf{x}(f,t)$ is expressed as

$$\mathbf{x}(f,t) = \sum_{j=1}^{J} \mathbf{c}_j(f,t), \tag{6}$$

where $\mathbf{c}_j(f,t) = [c_{1j}(f,t), \cdots, c_{Ij}(f,t)]^\top$ is the contribution of the jth source to the mixture signals, J is the number of sources, and I is the number of microphones. The source separation problem is to estimate $\mathbf{c}_j(t)$ from $\mathbf{x}(t)$.

In the LGM approach, the multichannel covariance matrix of each speech source is assumed to be a multiplication of a time-variant scalar $v_j(f,t)$ and a time-invariant multichannel matrix $\mathbf{V}_j(f)$ for jth source.

$$\mathbf{c}_j(f,t) \sim \mathcal{N}_{\mathbb{C}}(0, v_j(f,t)\mathbf{V}_j(f)) \tag{7}$$

The LGM estimates the maximum likelihood value of $v_j(f,t)$ and $\mathbf{V}_j(f)$ by using expectation-maximization algorithm. Then, the separated signal can be obtained by multichannel Wiener filtering:

$$\mathbf{c}_j(f,t) = v_j(f,t)\mathbf{V}_j(f)\mathbf{R}_x^{-1}(f,t)\mathbf{x}(f,t), \tag{8}$$

where $\mathbf{R}_x(f,t)$ is the covariance matrix of the input signal $\mathbf{x}(f,t)$ which is the sum of covariance matrix of every sources.

In this study, the number of sources is set to 3. All channels of the target source signals are used as augmented training data for acoustic modeling.

2.3.2. Semi-stationary noise separation using LGM

In the original LGM framework, all of source signals are assumed to be time-varying signals. However, in the real environments, there are a lot of semi-stationary noises. To deal with these noises, we introduce moving average smoothing of activities for the non-target noise sources in addition to the original LGM. The modification to the original LGM for non-target noise sources ($j > 0$) is following:

$$\mathbf{c}_j(f,t) \sim \mathbf{N}_{\mathbb{C}}(0, \hat{v}_j(f,t)\mathbf{V}_j(f)) \; ; j > 0 \tag{9}$$

, where $\hat{v}_j(f,t)$ is a smoothed activity:

$$\hat{v}_j(f,t) = \sum_{\tau=0}^{T_j} v_n(f, t-\tau) \tag{10}$$

T_j is the number of smoothing frames.

That modification works as a kind of regularization for avoiding over-fitting problem especially in semi-stationary noise environments. Applying the moving average filter in the each EM iteration, the target source, i.e. the most active source is extracted onto $\mathbf{c}_0$. So we no longer select the target source from separated signals using SRP-PHAT.

In this work, we use this modification of LGM for the frontend speech enhancement. For non-target two sources, the numbers of smoothing frames are set to 3 and 6. The test utterance is processed by the modified LGM before the baseline beamforming is applied.

3. Experimental evaluation

3.1. Tuning adaptation parameters

We first evaluated the sensitivity to hyper-parameters for unsupervised network adaptation. The system for this evaluation used the LGM based source separation. The structure of acoustic model and language models are the same as the baseline DNN+RNNLM system except the acoustic feature, which was 40 dimensional log mel filterbank with an energy term, followed by per utterance mean variance normalization and delta and acceleration feature augmentation.

The evaluation results for the development set are shown in Table 1. We observed the results of unsupervised network adaptation with any set of hyper-parameters always better than the non-adapted result. The small learning rate and the small number of iteration gives good result. Through this evaluation, the mini-batch size was set to 12000, the learning rate was set to 0.0004, and the number of iteration was set to 2 for further evaluations.

Table 1: Average WER (%) for various adaptation parameters

iteration	learn rate	mini-batch	WER (dev avg)
No adaptation			4.85
1	0.01	256	4.115
1	0.008	512	4.08
1	0.001	256	3.7
1	0.0004	256	3.745
1	0.0004	512	3.735
1	0.0004	12000	3.7
1	0.0001	256	3.865
2	0.0004	12000	**3.695**
10	0.0004	256	4.305

3.2. Evaluation on submitted system

The evaluation results are shown in Table 2, and the WERs per environment for the submitted systems are shown in Table 3.

The **adapted** system used unsupervised network adaptation. The initial model was from the baseline DNN+sMBR system. The baseline system used fMLLR transformed MFCC feature. The adapted system was tested only on "6ch track".

The **combined** system used phonetically-oriented system combination. The system combined four baseline systems (GMM, DNN+sMBR, DNN+5gram, DNN+RNNLM). The combined system was tested only on "6ch track".

The **LGM** system used the LGM based source separation. For the frontend of 1ch track, we applied no speech enhancement. The structure of acoustic model and language models are the same as the baseline system except the acoustic feature, which was 40 dimensional log mel filterbank with an energy term, followed by per utterance mean variance normalization and delta and acceleration feature augmentation.

The **LGM+adapted** system used unsupervised network adaptation. The initial model was from the LGM system.

The **submitted** system used phonetically-oriented system combination. The system combined 24 recognizers (12 backend models and 2 frontend methods). The backend models are comprised of 4 baselines (GMM, DNN+sMBR, DNN+5gram, DNN-RNNLM), 4 LGM-based data augmented models, 2 adapted DNN models (DNN+5gram, DNN+RNNLM) and 2 LGM-based data augmented and adapted DNN models. The frontend methods are baseline beamforming (beamformit) and LGM based beamforming as described in Section 2.3.2.

The results of the real test set on the 6ch track show the effectiveness of the unsupervised network adaptation. The relative gain was 6% of WER from the baseline and 29% from the LGM system. While the phonetically-oriented system combination was not effective for baseline systems, combination with the LGM and LGM+adapt systems achieved 16% relative gain. The LGM constantly reduced the WER and boosted the effectiveness of unsupervised network adaptation and phonetically-oriented system combination.

Table 2: Average WER (%) for the tested systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	baseline	11.56	12.99	23.59	20.72
	LGM	9.27	11.97	16.88	17.76
	LGM+adapted	7.29	9.56	13.57	13.96
	submitted	5.89	7.36	11.42	9.23
2ch	baseline	8.21	9.50	16.55	15.40
	LGM	6.51	8.37	12.08	10.98
	LGM+adapted	5.13	6.36	9.09	7.79
	submitted	4.22	5.88	8.61	7.32
6ch	baseline	5.76	6.77	11.46	10.91
	adapted	5.37	6.36	10.77	9.18
	combined	5.77	6.80	11.48	10.72
	LGM	4.49	5.20	7.78	6.35
	LGM+adapted	3.58	3.81	5.56	4.47
	submitted	2.68	3.33	4.68	4.15

4. Conclusion

We wrote our development for the CHiME-4 challenge and reported the experimental results. We examined unsupervised speaker adaptation for DNN based acoustic model and shown that the adaptation gives up to 29% relative gain on the 6ch track. Second, we evaluated a phonetically-oriented system combination method. Experiments showed that the system combination results up to 16% relative gain. Finally, we evaluated the combination of the above methods with LGM based source separation. The experimental results of the submitted system show that 4.68% of WER for the real evaluation set.

Table 3: WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	7.85	6.31	15.93	6.69
	CAF	6.02	9.87	11.86	9.86
	PED	4.01	5.78	9.81	9.69
	STR	5.68	7.48	8.09	10.68
2ch	BUS	5.24	4.78	12.26	4.78
	CAF	4.38	7.79	8.98	8.24
	PED	3.05	5.04	7.03	7.56
	STR	4.20	5.91	6.16	8.69
6ch	BUS	3.38	3.01	6.13	3.19
	CAF	2.20	3.92	4.50	4.17
	PED	2.33	2.88	3.87	4.20
	STR	2.80	3.53	4.24	5.02

5. References

[1] E. Vincent, S. Watanabe, 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.

[2] Y. Fujita, R. Takashima, T. Homma, R. Ikeshita, Y. Kawaguchi, T. Sumiyoshi, T. Endo, and M. Togami, "Unified ASR system using LGM-based source separation, noise-robust feature extraction, and word hypothesis selection," in *Proc. IEEE ASRU*, 2015, pp. 416–422.

[3] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE ASRU*, 2015, pp. 436–443.

[4] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. ICASSP*, May 2013, pp. 7947–7951.

[5] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE ASRU*, 1997, pp. 347–354.

[6] N. Ruiz and M. Federico, "Phonetically-oriented word error alignment for speech recognition error analysis in speech translation," in *Proc. IEEE ASRU*, Dec 2015, pp. 296–302.

[7] Y. Fujita, R. Takashima, T. Homma, and M. Togami, "Data augmentation using multi-input multi-output source separation for deep neural network based acoustic modeling," in *Proc. Interspeech*, 2016, pp. 3818–3822.

[8] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Speech Audio Process.*, vol. 18, pp. 1830–1840, Sep. 2010.