INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Data augmentation using multi-input multi-output source separation for deep neural network based acoustic modeling

Yusuke Fujita, Ryoich Takashima, Takeshi Homma, and Masahito Togami

Hitachi, Ltd. Research and Development Group
yusuke.fujita.su@hitachi.com

Abstract

We investigate the use of local Gaussian modeling (LGM) based source separation to improve speech recognition accuracy. Previous studies have shown that the LGM based source separation technique has been successfully applied to the runtime speech enhancement and the speech enhancement of training data for deep neural network (DNN) based acoustic modeling. In this paper, we propose a data augmentation method utilizing the multi-input multi-output (MIMO) characteristic of LGM based source separation. We first investigate the difference between unprocessed multi-microphone signals and multi-channel output signals from LGM based source separation as augmented training data for DNN based acoustic modeling. Experimental results using the third CHiME challenge dataset show that the proposed data augmentation outperforms the conventional data augmentation. In addition, we experiment the beamforming applied to the source separated signals as runtime speech enhancement. The results show that the proposed runtime beamforming further improves the speech recognition accuracy.

Index Terms: speech recognition, acoustic modeling, data augmentation, blind source separation

1. Introduction

Automatic speech recognition (ASR) is essential for robot interaction systems and call center analytics, and it is one of the popular features in smart devices such as tablets and smartphones. Since deep neural network (DNN) has become the dominant acoustic model for ASR in recent years, the accuracy of ASR is sufficiently high when an input signal is recorded in a quiet room or using a close-talking microphone. However, the accuracy is poor when a microphone input signal is contaminated by several noise sources, e.g., background noise sources, the speech of other people, and reverberation in most practical situations for smart devices. Therefore, robustness against these noises is crucial for ASR in smart devices.

To achieve robustness against noises, several speech enhancement techniques have been proposed. Single-channel noise reduction techniques such as spectral subtraction [1], minimum mean-square error short-term spectral amplitude (MMSE-STSA) [2], and optimally-modified log-spectral amplitude (OM-LSA) [3] have been studied. These techniques are mainly utilized for stationary noise reduction.

When the noise sources are highly non-stationary such as speech sources, the performance of single-channel noise reduction is severely degraded. For non-stationary noise reduction, multi-channel noise reduction techniques have been studied. Beamforming techniques [4] are one of the major techniques in this field. When the direction-of-arrival (DOA) of the desired source signal is known in advance, beamforming techniques can accurately reduce non-stationary noise sources. However, when the actual DOA of the desired source signal is different from the pre-given DOA, the output signal is distorted because of the signal cancellation problem [5]. To avoid the signal cancellation problem, blind source separation techniques have been studied, e.g., independent component analysis and local Gaussian modeling (LGM) [6].

Previously, we proposed the ASR system incorporating the LGM based blind source separation method and DNN based acoustic modeling under daily non-stationary noise environment using multi-microphone tablet devices [7]. The LGM based source separation has been successfully applied to the runtime speech enhancement and the speech enhancement of training data for DNN based acoustic modeling. While the speech enhancement applied to both training data set and test data set was effective, the total amount of training data set which we used is thought to be small for DNN training. In this situation, DNNs cannot learn the expected feature transformation such that speaker and noise variations are absorbed and irrelevant features are ignored.

Data augmentation has shown to be an effective approach for DNN training in such limited-resource situations. Elastic spectral distortion method was investigated in [8], and showed the effectiveness of vocal tract length distortion, speech rate distortion, and frequency-axis random distortion. However, such distortions are mainly focused on speaker variations, not on noise variations.

In this paper, we show that the LGM based source separation can be used for data augmentation to improve the noise-robustness of DNN based acoustic models. The multi-input multi-output (MIMO) characteristic of the LGM based source separation can be used for data augmentation. In addition, we show that the effect of runtime beamforming applied after the source separation.

The remainder of the paper is organized as follows. In Section 2 we describe the overview of our ASR system. We then describe the proposed data augmentation method in Section 3. The experimental results of the proposed method are described in Section 4, and finally the conclusion is presented in Section 5.

2. ASR system

2.1. Training speech enhancement and acoustic modeling

Figure 1 shows an overview of the acoustic modeling system incorporated with speech enhancement. At first for reduction of directional noises, we utilize the LGM based blind source separation method [6] against six-channel input signals.

The LGM based method separates input signals into predefined number of sources (three in this paper). Each separated

http://dx.doi.org/10.21437/Interspeech.2016-733

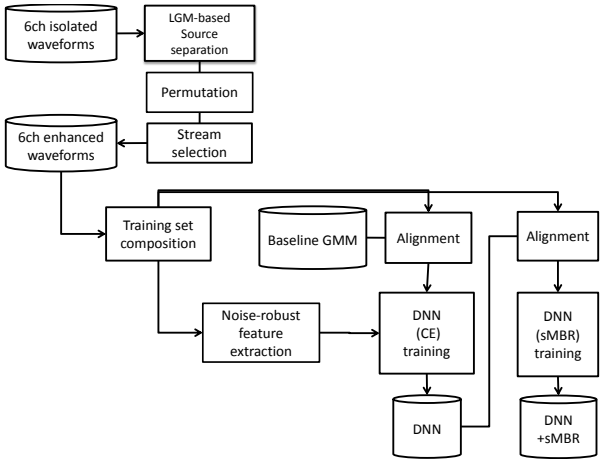

Figure 1: Training system

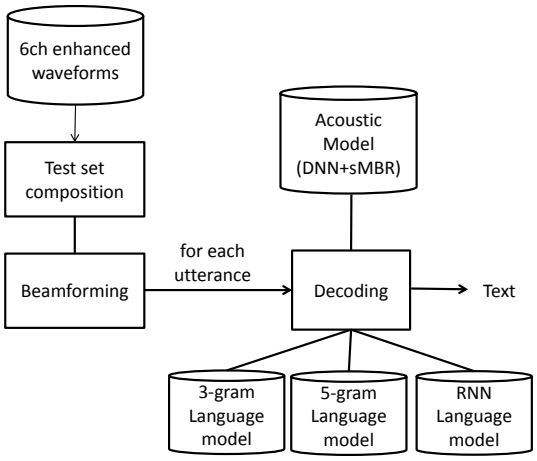

Figure 2: Runtime system

source have six-channel outputs. Since the LGM-based separation is performed for each frequency bin independently, the permutation alignment is solved using power-spectrum correlation based method and DOA based method. Finally, the target source selection is performed to select the signal that most closely related to the target signal amoung the separated signals. To select the target signal, the DOA histogram is calculated for each separated signal by using steered response power with the phase transform (SRP-PHAT) algorithm [9].

Then, the proposed data augmentation is applied for composing the training set. Acoustic modeling is performed using the augmented training set. In this study, we use the Context-dependent deep neural networks with hidden Markov models (CD-DNN-HMM) developed as CHiME-3 baseline system [10]. However, we modified the input features to achieve noise-robustness. The input features are based on 40-dimensional log melfilterbank coefficients with an energy term. The mean and variance normalization is performed per utterance. Then, the delta and delta-delta features are appended.

The system constructs three types of acoustic models: baseline Gaussian mixture models (GMMs), a DNN trained with the cross-entropy criterion, and a DNN trained by the sequence discriminative training method with the state-level minimum Bayes risk (sMBR) criterion.

2.2. Runtime speech enhancement and decoding

Figure 2 shows an overview of the runtime speech enhancement and decoding. At runtime, the LGM based source separation is performed in the same manner as the training speech enhancement. speech enhancement. In this paper, we experiment the beamforming applied to the separated signal.

Then, the signal is decoded through trained acoustic model and language models. In this paper, we use the combination of 3-gram language model, 5-gram language model, and recurrent neural network language model as shown in [11].

3. Proposed method

3.1. Data augmentation using MIMO source separation

A simple approach to data augmentation focused on noise variations is "multi-microphone training". To train an acoustic model using signals from multiple microphones has shown to be effective [12]. However, the signals from multiple microphones are unprocessed noisy data. While more variations on noises are captured than using single microphone, it leads mismatched condition between noisy training signals and enhanced test signals.

We propose the data augmentation method using multiple output signals from LGM based source separation. Because LGM has MIMO characteristic, each source has multiple output signals.

In the MIMO source separation, the multi-microphone signal in the time-frequency domain $\boldsymbol{x}(f,t)$ is expressed as

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

where $\boldsymbol{c}_j(f,t) = [c_{1j}(f,t), \cdots, c_{Ij}(f,t)]^\top$ 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 $\boldsymbol{c}_j(t)$ from $\boldsymbol{x}(t)$.

In the LGM approach, the multichannel covariance matrix of each speech source is assumed to be a multiplication of a time-variant scalar coefficient $v_j(f,t)$ and a time-invariant multichannel matrix $\boldsymbol{R}_j(f)$ for jth source. The LGM estimates the maximum likelihood value of $v_j(f,t)$ and $\boldsymbol{R}_j(f)$ by using expectation-maximization (EM) algorithm. Then, the separated signal can be obtained by multichannel Wiener filtering

$$\boldsymbol{c}_j(f,t) = v_j(f,t)\boldsymbol{R}_j(f)\boldsymbol{R}_x^{-1}(f,t)\boldsymbol{x}(f,t), \tag{2}$$

where $\boldsymbol{R}_x(f,t)$ is the convariance matrix of the input signal $\boldsymbol{x}(f,t)$ which is the sum of covariance matrix of every sources. This separated signal for each source has the same number of channels as the input signal.

While the LGM has MIMO characteristic, the beamforming method has multi-input single-output (MISO) characteristic. In the beamforming method, the separated signal $y(t)$ is obtained as follows:

$$y(t) = \boldsymbol{w}(t)\boldsymbol{x}(t), \tag{3}$$

where $\boldsymbol{w}(t)$ is the multichannel filter of the desired source, which is obtained by using the estimated steering vector and the multichannel covariance matrix of the noise source. The beamforming result cannot be used for the proposed data augmentation because its output is a single-channel signal.

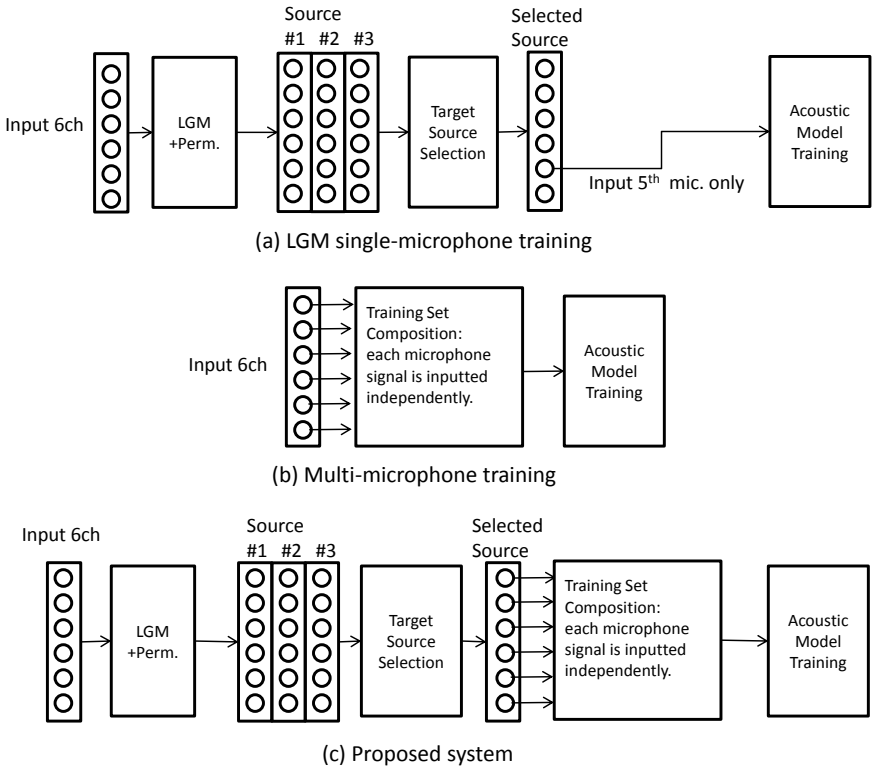

Figure 3: Data augumentation

Figure 3 shows the difference among the conventional method, multi-microphone training, and the proposed method. The proposed method can capture the residual noise observed in the separated signals both training and runtime speech enhancement processes.

3.2. Runtime beamforming after LGM based source separation

In the previous system in [7], LGM based source separation was performed as runtime speech enhancement. However, we only used the single-channel signal for decoding. The fifth microphone was selected because it was closest to a user.

In this paper, we experiment the beamforming after LGM based source separation. Since LGM based source separation has MIMO characteristic and separated multi-channel signals can preserve the spatial information, the beamforming technique can be applied to the separated signals. This approach can be interpreted as MIMO source separation as pre-filter for the beamforming as shown in [13]. We used weighted delay and sum beamforming proposed in [14].

4. Experiment

4.1. Data set

We use the CHiME-3 corpus, which targets the ASR for a multi-microphone tablet device used in everyday noisy environments [10]. A six-channel microphone array embedded in a tablet device was used for recording audio data. The recordings have been made in four varied environments: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED).

The training data consists of 1600 real noisy utterances, and 7138 simulated utterances constructed by mixing clean speeches with noise backgrounds. The total amount of training data set is 18 hours.

To evaluated the proposed system, we calculated the average word error rate (WER) in accordance with the CHiME-3 task specification described in [10].

In the tables shown below, "dev-real" denotes the real utterances in the development set, and "test-real" denotes the real utterances in the test set. In the CHiME-3 corpus, the development set consists of 1640 real utterances, and the test set consists of 1320 real utterances.

4.2. Effect of data augmentation

The evaluation results on the proposed data augmentation method is shown in table 1. The result shows that proposed data augmentation reduces the word error rate (WER) from 12.2% to 10.0% for the real test set. The proposed method outperforms the multi-microphone training with unprocessed signals. For the development set, the proposed data augmentation does not reduce WER. This is because the noise characteristic in the development set is relatively similar to the noise in the training set, while the test set is not very similar. The results show that the proposed data augmentation can achieve the robustness against the noise variations. When the proposed data augmentation and unprocessed multi-microphone training are both applied, the results are better than the proposed data augmentation only. This is because

Table 1: WER(%) on the data augmentation

Method	dev-real	test-real
single-microphone	7.4	13.2
multi-microphone	7.2	11.0
LGM single-microphone	6.0	12.2
proposed	6.6	10.0
proposed + multi-microphone	5.3	9.2

4.3. Effect of the runtime beamforming

The evaluation results on the proposed beamforming method is shown in table 2. The result shows that the beamforming reduces the WER from 10.0% to 9.6% for the real test set. The proposed beamforming method used together with the proposed data augmentation and the multi-microphone training further reduces the WER from 9.2% to 8.8%. There is a mismatch between training set signal and test set signal because the beamforming is not applied to the training data set. However, the degradation introduced by the beamforming is negligible compared with the positive effect that improves the signal to noise ratio.

Table 2: WER(%) on the beamforming (BF)

data augmentation	BF	dev-real	test-real
LGM single-microphone	off	6.0	12.2
	on	6.1	11.7
proposed	off	6.6	10.0
	on	6.1	9.6
proposed + multi-microphone	off	5.3	9.2
	on	5.4	8.8

4.4. Evaluation detail on various noise environment

The detailed results on every environmental conditions compared with the coventional results are shown in Table 3.

The proposed method reduces the WER for every environmental conditions, especially at the CAF environment. The relative improvement is 28.0% in the CAF environment for the real test set. The primary noise source of CAF environment is speech of other people, which is highly non-stationary. The results show that the proposed method works well for such highly non-stationary noise sources.

Table 3: Detailed WER(%) on various noise emvironments

Method	Env.	dev-real	test-real
previous system [7]	Total	7.0	11.8
	BUS	9.8	16.6
	CAF	6.2	11.8
	PED	5.2	10.0
	STR	6.8	8.8
proposed system	Total	5.4	8.8
	BUS	8.6	12.6
	CAF	3.8	7.5
	PED	3.9	8.1
	STR	5.6	7.0

5. Conclusions

In this paper, we investigated the use of local Gaussian modeling (LGM) based source separation for the data augmentation. The proposed data augmentation method is based on the multi-input multi-output (MIMO) characteristic of LGM based source separation. The proposed method was evaluated using the third CHiME challenge dataset. The proposed data augmentation using LGM based source separation achieved promising noise-robustness and shown to be superior to the conventional unprocessed multi-microphone based method. Additionally, we examined the beamforming applied after the source separation as runtime speech enhancement. The experimental results showed that the beamforming at runtime further improves the accuracy of ASR. It is also shown that the proposed method is effective against every environmental condition, especially highly non-stationary noises such as speech of other people are contaminated.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustic., Speech, Signal Process.*, vol. 27, pp. 113–120, Feb. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustic., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.

[3] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[4] D. Johnson and D. Dudgeon, *Array signal processing-concepts and Techniques*. PTR Prentice Hall, 1993.

[5] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer, 2005.

[6] 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.

[7] 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," in *Proc. IEEE ASRU*, 2015, pp. 416–422.

[8] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. IEEE ASRU*, Dec 2013, pp. 309–314.

[9] M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 288–292, 1993.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recgnition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, 2015, pp. 504–511.

[11] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. L. Roux, and V. Mitra, "The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proc. IEEE ASRU*, 2015, pp. 475–481.

[12] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. F. 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. IEEE ASRU*, 2015, pp. 436–443.

[13] M. Togami, Y. Kawaguchi, N. Nukaga, and Y. Obuchi, "Online mvbf adaptation under diffuse noise environments with mimo based noise pre-filtering," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, July 2012, pp. 292–297.

[14] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2023, Sep. 2007.