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

Optimization of Speech Enhancement Front-end with Speech Recognition-level Criterion

Takuya Higuchi, Takuya Yoshioka, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
{higuchi.takuya, yoshioka.takuya, nakatani.tomohiro}@ntt.lab.co.jp

Abstract

This paper concerns the use of speech enhancement to improve automatic speech recognition (ASR) performance in noisy environments. Speech enhancement systems are usually designed separately from a back-end recognizer by optimizing the front-end parameters with signal-level criteria. Such a disjoint processing approach is not always useful for ASR. Indeed, time-frequency masking, which is widely used in the speech enhancement community, sometimes degrades the ASR performance because of the artifacts created by masking. This paper proposes a speech recognition-oriented front-end approach that optimizes the front-end parameters with an ASR-level criterion, where we use a complex Gaussian mixture model (CGMM) for mask estimation. First, the process of CGMM-based time-frequency masking is reformulated as a computation network. By connecting this CGMM network to the input layer of the acoustic model, the CGMM parameters can be optimized for each test utterance by back propagation using an unsupervised acoustic model adaptation scheme. Experimental results show that the proposed method achieves a relative improvement of 7.7 % on the CHiME-3 evaluation set in terms of word error rate.

Index Terms: speech enhancement, speech recognition, time-frequency masking

1. Introduction

Despite the recent rapid progress made on automatic speech recognition (ASR) technology, ASR systems perform poorly when they are deployed in noisy environments that are not present during training. Performing speech enhancement prior to recognition can alleviate this problem by adaptively removing background noise from test speech signals [1, 2]. Speech enhancement is usually conducted as front-end processing, i.e., in a way that is independent of a back-end recognizer. The front-end parameters are optimized with a signal-level criterion such as a log-likelihood score derived from a generative model and a posterior signal-to-noise ratio (SNR).

The goal of this work is to perform speech enhancement adaptively by using an ASR-oriented front-end optimization criterion. Some classes of speech enhancement approaches, including time-frequency masking, can sometimes hurt the ASR performance because of processing artifacts whereas they effectively improve the SNR [2, 3]. Using the ASR-level criterion allows a speech enhancer to take account of the way in which the enhanced speech signal is processed by the back-end.

An existing approach to achieving such ASR-friendly speech enhancement is to jointly train a denoising auto-encoder (DAE) and a deep neural network (DNN) acoustic model [4, 5, 6]. While this approach allows us to perform speech enhancement based on an ASR-level criterion, the front-end DAE does not adapt to unseen test environments. In addition, the DAE often needs to be initialized by using a parallel corpus comprising pairs of clean and noisy speech samples.

This paper proposes a novel approach for optimizing time-frequency masking-based speech enhancement parameters for each test utterance based on an ASR-level criterion. The basic idea behind the proposed approach is to reinterpret a time-frequency masking process based on a complex Gaussian mixture model (CGMM) as a computation network as in [7, 8, 9]. This CGMM-based front-end network is connected to the input layer of a trained DNN acoustic model. The two networks constitute a large neural network that estimates hidden Markov model (HMM) posteriors from multichannel observations. To make the front-end adaptive to test environments, the front-end network parameters are optimized for each test utterance with an unsupervised acoustic model adaptation scheme [10], i.e., by minimizing the cross entropy between the network outputs and supervision labels created by an initial recognition pass. This enables us to obtain CGMM parameter values that better discriminate the HMM states. An advantage of the proposed approach is that the front-end network is derived from a multichannel signal generation model and thus can be efficiently described by a small number of parameters. This means that the network can be adapted for every single utterance even with an unsupervised approach. An experimental evaluation showed that the proposed method improved the word error rate (WER) by 7.7 % relatively for the CHiME-3 evaluation set compared with a time-frequency masking front-end that optimized the CGMM parameters with a signal-level log-likelihood criterion.

The rest of this paper is organized as follows. Section 2 reviews relevant work. Section 3 overviews our front-end optimization approach using an ASR-level criterion. Section 4 describes a CGMM-based generative model of multichannel observations, which we use for time-frequency mask estimation. Section 5 reformulates the masking process as a computation network. Section 6 describes the proposed ASR-oriented optimization method for the front-end network parameters. Section 7 describes experiments we performed to evaluate the effectiveness of our method. Section 8 concludes the paper.

2. Related Work

Our approach for estimating front-end speech enhancement parameters centers around unsupervised acoustic model adaptation. When we perform unsupervised acoustic model adaptation with a limited quantity of data, the number of parameters to be optimized must be kept small to make the adapted model robust against errors in the supervision labels that are created by an initial decoding pass (see, e.g., [10] for DNN adaptation). Our speech enhancement front-end network is derived from a

http://dx.doi.org/10.21437/Interspeech.2016-681

CGMM-based generative model and thus efficiently models the speech enhancement process with a small number of parameters. Therefore, the parameters can be reliably estimated even from a single utterance.

Some recent studies investigated a novel approach that reformulates generative model-based speech enhancement processes as computation networks [7, 8, 9], one of which dealt with multichannel observations [9]. They modeled multichannel observations as a weighted sum of multi-variate Gaussian sources. A variational parameter inference algorithm for this generative model was "unfolded" to yield a deep network where the network parameters corresponded to the generative model parameters. The parameters were trained by using training data.

Although our proposed approach is inspired by these studies, our generative model and its parameter training scheme are different from those of the previous work. By leveraging the sparsity of speech signals, we model a generative process of a multichannel observation with a two-component CGMM, where one Gaussian component corresponds to speech and one to noise. This generative model can be optimized by a simple parameter estimation algorithm based on the Expectation-Maximization (EM) algorithm [3], thus allowing the speech enhancement process to be reformulated as a shallow network with a small number of parameters. This makes it possible to rapidly adapt the speech enhancement front-end to test environments.

3. Overview of our front-end optimization

Figure 1 shows an overview of our proposed front-end optimization process, which is based on a standard unsupervised acoustic model adaptation approach. First, we perform initial decoding (with or without speech enhancement) to obtain supervision labels for adaptation, i.e., front-end parameter optimization. Second, we reformulate the front-end speech enhancement process as a computation network, which is connected to a feature extractor, followed by a DNN acoustic model. The CGMM-based front-end network parameters are optimized by back propagation, which minimizes the cross entropy between the supervision labels and acoustic model outputs. This allows a speech enhancement front-end to take account of the way in which the back-end recognizer processes the enhanced speech signal. The optimized front-end network is used in the second decoding pass.

4. Complex Gaussian mixture model

This section briefly reviews a speech enhancement method based on a complex Gaussian mixture model (CGMM), which is reformulated as a computation network in the next section.

Let $y_{f,t,m}$ denote the m-th microphone signal at frequency f and time t. The signals from all M microphones can be represented using vector notation as

$$\boldsymbol{y}_{f,t} = [y_{f,t,1}, \ldots, y_{f,t,M}]^{\mathrm{T}}, \tag{1}$$

where superscript T denotes non-conjugate transposition.

By considering the sparseness of the speech energy distribution in the time-frequency domain [3, 11, 12, 13, 14], we assume that observed signals can be clustered into two classes each corresponding to either speech or noise. With this assumption, the observed signal can be described as

$$\boldsymbol{y}_{f,t} = \boldsymbol{r}_f^{(\nu)} s_{f,t}^{(\nu)} \quad (\text{where } d_{f,t} = \nu), \tag{2}$$

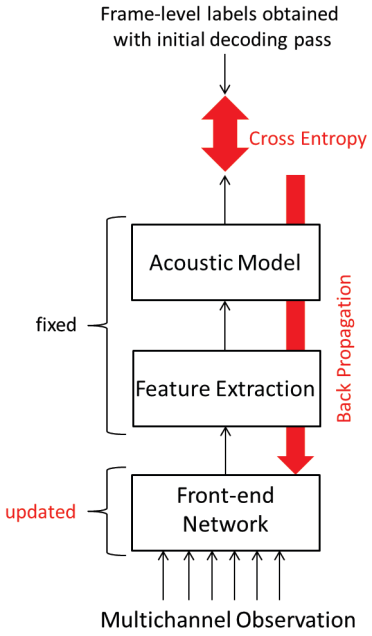

Figure 1: Overview of our front-end optimization process.

where $d_{f,t}$ denotes the category index at the time-frequency point (f, t). ν may take x or n, and $s_{f,t}^{(x)}$ denotes a speech signal at frequency f and time t, while $s_{f,t}^{(n)}$ denotes a noise signal at frequency f and time t. We further assume that $s_{f,t}^{(\nu)}$ locally follows a complex Gaussian distribution as

$$s_{f,t}^{(\nu)} \sim \mathcal{N}_c(0, \phi_{f,t}^{(\nu)}), \tag{3}$$

where $\phi_{f,t}^{(\nu)}$ corresponds to the variance of the signal at the time-frequency point, and $\mathcal{N}_c(x; \mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp - \frac{|x-\mu|^2}{\sigma^2}$.

From Eqs. (2) and (3), the multichannel observed signal follows a complex Gaussian distribution

$$\boldsymbol{y}_{f,t} | d_{f,t} = \nu \sim \mathcal{N}_c(0, \phi_{f,t}^{(\nu)} \boldsymbol{R}_f^{(\nu)}) \tag{4}$$

conditioned on $d_{f,t}$, where $\boldsymbol{R}_f^{(\nu)}$ denotes a frequency-dependent spatial correlation matrix for category ν. The generative model for the observed signal $\boldsymbol{y}_{f,t}$ can be derived by marginalizing the speech/noise indicator $d_{f,t}$, which results in a two-component complex Gaussian mixture model.

The CGMM parameters are usually trained by using the EM algorithm to maximize the log-likelihood criterion, which enable us to obtain CGMM parameters that approximate multichannel observations. The CGMM parameters can be iteratively updated based on the EM algorithm by

$$\phi_{f,t}^{(\nu)} \leftarrow \frac{1}{M} \mathrm{tr}(\boldsymbol{y}_{f,t} \boldsymbol{y}_{f,t}^{\mathrm{H}} \boldsymbol{R}_f^{(\nu)^{-1}}), \tag{5}$$

$$\boldsymbol{R}_f^{(\nu)} \leftarrow \frac{1}{\sum_t \lambda_{f,t}^{(\nu)}} \sum_t \lambda_{f,t}^{(\nu)} \frac{1}{\phi_{f,t}^{(\nu)}} \boldsymbol{y}_{f,t} \boldsymbol{y}_{f,t}^{\mathrm{H}}, \tag{6}$$

where $\lambda_{f,t}^{(\nu)}$ denotes the time-frequency mask for class ν at the time-frequency point (f, t), and superscript H denotes conjugate transposition. The masks, namely the CGMM posteriors,

of the target speech can be obtained at the E-step as follows:

$$\lambda_{f,t}^{(x)} = \frac{p_{f,t}^{(x)}}{p_{f,t}^{(x)} + p_{f,t}^{(n)}}, \quad (7)$$

where

$$p_{f,t}^{(x)} = \frac{det\boldsymbol{R}_f^{(x)^{-1}}}{\pi^M |\phi_{f,t}^{(x)}|^M} \exp\Big\{-\frac{1}{\phi_{f,t}^{(x)}} tr(\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathrm{H}}\boldsymbol{R}_f^{(x)^{-1}})\Big\}, \quad (8)$$

$$p_{f,t}^{(n)} = \frac{det\boldsymbol{R}_f^{(n)^{-1}}}{\pi^M |\phi_{f,t}^{(n)}|^M} \exp\Big\{-\frac{1}{\phi_{f,t}^{(n)}} tr(\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathrm{H}}\boldsymbol{R}_f^{(n)^{-1}})\Big\}. \quad (9)$$

5. Complex Gaussian mixture network

In this section, we reformulate the CGMM-based speech enhancement process as a computation network so that the CGMM parameters can be optimized with a back-end-oriented criterion.

The parameter update rule of $\phi_{f,t}^{(\nu)}$ is regarded as a computation network, where $\boldsymbol{R}_f^{(\nu)}$ is regarded as a free parameter that needs to be optimized. By using Eq. (5), obtained from the log-likelihood criterion, Eqs. (8) and (9) can be described with just spatial correlation matrices and a multichannel observation as

$$p_{f,t}^{(\nu)} = \frac{det\boldsymbol{R}_f^{(\nu)^{-1}}}{\pi^M |\frac{1}{M}\mathrm{tr}(\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathrm{H}}\boldsymbol{R}_f^{(\nu)^{-1}})|^M} \exp\Big\{-M\Big\}. \quad (10)$$

The computation steps given by Eqs. (7) and (10) can be viewed as constituting a network, which can be denoted as

$$\lambda_{f,t}^{(x)} = CGMN(\boldsymbol{R}_f^{(x)^{-1}}, \boldsymbol{R}_f^{(n)^{-1}}; \boldsymbol{y}_{f,t}). \quad (11)$$

The complex Gaussian mixture network (CGMN) has the spatial correlation matrices as "internal" parameters and outputs time-frequency masks from input multichannel observations. Instead of optimizing the CGMM network parameters during training, we estimate the CGMM parameters for each test utterance to realize rapid adaptation to test environments.

6. Parameter training with speech recognition-level criterion

We train the CGMN parameters with a speech recognition-level criterion by using a two-pass-based unsupervised acoustic model adaptation approach. Note that the parameters are optimized independently for each test utterance.

First, by concatenating the CGMN, feature extractor and an acoustic model, we can describe the process for obtaining HMM posteriors as a unified network as follows:

$$\lambda^{(x)} = CGMN(\boldsymbol{R}^{(x)^{-1}}, \boldsymbol{R}^{(n)^{-1}}; \boldsymbol{y}), \quad (12)$$

$$\hat{s} = Masking(\lambda^{(x)}, \boldsymbol{y}), \quad (13)$$

$$\hat{x} = FeatureExtract(\hat{s}), \quad (14)$$

$$\hat{l} = AcousticModel(\hat{x}), \quad (15)$$

where $\hat{s}$, $\hat{x}$ and $\hat{l}$ denote enhanced signals, extracted features and estimated HMM posteriors respectively. $Masking(\lambda^{(x)}, \boldsymbol{y})$ denotes a function that multiplies spectral magnitudes observed by a reference microphone by the time-frequency masks created by the CGMN. $FeatureExtract(\hat{s})$ denotes a function for extracting features from the enhanced signals, which is usually parameterized with fixed parameters (e.g. log-mel feature extraction, calculation of Δ / Δ^2 features and feature normalization by affine transform). $AcousticModel(\hat{x})$ denotes an acoustic model for computing HMM posteriors from the acoustic features. The acoustic model is trained by using training data.

The objective function $\mathcal{L}$ is defined as the cross entropy between supervision labels l^{best} and HMM posteriors $\hat{l}$

$$\begin{aligned}\mathcal{L}(l^{best}, \hat{l}) &= \sum_t \mathcal{L}(l_t^{best}, \hat{l}_t) \\ &= \sum_t CrossEntropy(l_t^{best}, \hat{l}_t),\end{aligned} \quad (16)$$

where l^{best} is generated by an initial decoding pass. Based on the gradient descent algorithm, the CGMN parameters, i.e., the spatial correlation matrices, can be updated by

$$\boldsymbol{R}_f^{(\nu)^{-1}} \leftarrow \boldsymbol{R}_f^{(\nu)^{-1}} - \alpha \frac{1}{T} \sum_t \frac{\partial \mathcal{L}(l_t^{best}, \hat{l}_t)}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}}, \quad (17)$$

where α is the learning rate. By applying the chain rule based on Eqs. (12)-(15), the gradient with respect to (the inverse of) the spatial correlation matrix can be described as

$$\begin{aligned}&\frac{\partial \mathcal{L}(l_t^{best}, \hat{l}_t)}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}} \\ =&\frac{\partial \mathcal{L}(l_t^{best}, \hat{l}_t)}{\partial \hat{x}_t} \cdot \frac{\partial \hat{x}_t}{\partial \hat{s}_{f,t}} \cdot \frac{\partial \hat{s}_{f,t}}{\partial \lambda_{f,t}^{(x)}} \cdot \frac{\partial \lambda_{f,t}^{(x)}}{\partial p_{f,t}^{(\nu)}} \cdot \frac{\partial p_{f,t}^{(\nu)}}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}}.\end{aligned} \quad (18)$$

The first term can be computed because the gradient is calculated when the acoustic model is being trained. The second term is easily computed because the feature extraction procedure is often a combination of linear and log functions (See [4]). The third term is easily computed because $\hat{s}$ is a product of the mask $\lambda^{(x)}$ and the magnitude of the observation at a reference microphone. From Eq. (7), the fourth term is computed as

$$\frac{\partial \lambda_{f,t}^{(x)}}{\partial p_{f,t}^{(x)}} = \frac{p_{f,t}^{(n)}}{(p_{f,t}^{(x)} + p_{f,t}^{(n)})^2}, \quad (19)$$

$$\frac{\partial \lambda_{f,t}^{(x)}}{\partial p_{f,t}^{(n)}} = \frac{-p_{f,t}^{(x)}}{(p_{f,t}^{(x)} + p_{f,t}^{(n)})^2}. \quad (20)$$

From Eq. (10), the final term can be computed as follows:

$$\begin{aligned}\frac{\partial p_{f,t}^{(\nu)}}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}} &= \frac{\partial}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}} \exp\{\log p_{f,t}^{(\nu)}\} \\ &= p_{f,t}^{(\nu)} \cdot \frac{\partial}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}} \log p_{f,t}^{(\nu)},\end{aligned} \quad (21)$$

where

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}} \log p_{f,t}^{(\nu)} \\ =&\frac{\partial}{\partial \boldsymbol{R}_f^{(\nu)^{-1}}} \{-M \log \mathrm{tr}(\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathrm{H}}\boldsymbol{R}_f^{(\nu)^{-1}}) + \log det\boldsymbol{R}_f^{(\nu)^{-1}}\} \\ =&-M \frac{1}{\mathrm{tr}(\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathrm{H}}\boldsymbol{R}_f^{(\nu)^{-1}})} \cdot (\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathrm{H}})^{\mathrm{T}} + (\boldsymbol{R}_f^{(\nu)^{-1}})^{\mathrm{T}^{-1}}\end{aligned} \quad (22)$$

Table 1: WERs for the real data in the development and evaluation sets (real recording only).

Systems	# retrained params	dev					eval				
		avg	bus	caf	ped	str	avg	bus	caf	ped	str
w/o speech enhancement	-	13.16	19.22	11.92	9.59	11.92	24.66	37.82	25.27	19.97	15.56
CGMM	-	13.57	17.33	11.14	12.57	13.23	19.88	25.39	16.62	20.10	17.41
Retrained CNN w/ CGMM	9900	12.77	18.10	10.68	9.91	12.39	24.10	37.45	22.43	19.97	16.57
	42300	13.50	19.69	11.93	10.27	12.09	25.05	38.51	24.82	20.57	16.31
Proposed CGMN w/ CGMM init.	14472	**12.01**	16.36	9.34	10.36	11.99	**18.35**	24.90	15.20	18.16	15.15

We fix the parameters of the back-end acoustic model, and update only the spatial correlation matrices of the CGMN. Note that the CGMN has a structure with a limited number of parameters, which reflects the signal generation process. This prevents the parameters from being excessively affected by supervision errors even with the utterance-wise parameter adaptation.

7. Experimental evaluation

7.1. Settings

We conducted experiments to evaluate the effectiveness of our proposed method in terms of WER by using the CHiME-3 corpus. The corpus consists of read utterances that were recorded with six microphones attached to a tablet device in four different environments: public tranportation (bus), café (caf), pedestrian area (ped), and street junction (str). The sentences were taken from the WSJ0 corpus. The training set comprises 1600 real and 7138 simulated utterances. The audio data from all six microphones were used for training, which amounts to about 108 hours. The development and evaluation sets consist of 1640 and 1320 real utterances, respectively (we did not use the CHiME-3 simulated test sets).

In our experiments, we performed speaker independent decoding by using a deep convolutional neural network (CNN) acoustic model [15, 16] and a tri-gram language model. Inputs to the acoustic model comprised 40-dimensional log mel-filter bank channel outputs. For simplicity, we neither used delta and double-delta features nor performed feature normalization. With these two exceptions, the CNN we used in this experiment was identical to the one we used in our CHiME-3 system [2]. Our CNN consisted of five convolution layers and two max-pooling layers, where all the layers contained 180 feature maps. The last convolution layer was followed by three fully connected layers with 2048 units and a softmax layer. The softmax layer contained 5976 units, i.e., context-dependent HMM states.

The spatial correlation matrices of the CGMN were initialized with generatively optimized CGMM parameters, i.e., those obtained by maximizing the log likelihood for the CGMM generative model. We updated only the spatial correlation matrices for speech to reduce the number of parameters to be optimized. The parameters were updated 30 times with an utterance-batch processing approach. The learning rate α was set at 10^5. Other experimental conditions were set as shown in Table 2.

We compare the performance of the proposed method with those of two existing methods. One method was based on disjoint front-end and back-end processing, where speech enhancement was performed by using CGMM-based spectral masks that maximized the log-likelihood criterion [3]. The other method was based on utterance-level unsupervised acoustic model adaptation. Specifically, the first or the first and second convolution layer(s) of the CNN acoustic model was (were) adapted for each utterance by using supervision labels generated by an initial decoding pass. To take advantage of the CGMM-based spectral masking, we used enhanced speech signals as CNN inputs. For all the two-pass approaches (including the proposed method), the first decoding pass was performed without speech enhancement, which tended to result in a better recognition performance.

Table 2: Experimental conditions.

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Window function	Hanning

7.2. Results

Table 1 shows the WERs we obtained with our proposed method and with the competitors. While CGMM-based spectral masking with the signal-level criterion, i.e., the log likelihood, reduced the WERs for the bus and café environments, it degraded the performance for the pedestrian area and street junction environments. In contrast, the proposed CGMN improved the WERs for almost all the environments. The relative WER improvement was 7.7 % compared with the CGMM optimized by the signal-level criterion. This improvement could not be obtained by simply retraining the bottom layer(s) of the CNN. Table 1 shows that CNN retraining produced no performance gain. It is noteworthy that CNN retraining did not improve the WERs even when we used CGMM speech enhancement to generate adaptation supervisions. These results show the superiority of our proposed front-end optimization.

8. Conclusion

We proposed a novel approach for optimizing a speech enhancement front-end with an ASR-oriented criterion. First, we reformulated the speech enhancement process based on the CGMM as a computation network, which we call the CGMN. The CGMN was connected to the input layer of an acoustic model, and its parameters were optimized by back propagation to minimize the cross entropy between the outputs of the acoustic model and the supervision labels generated by an initial decoding pass. The experimental results showed that the proposed method outperformed spectral masking using a signal-level criterion by 7.7 % relative in terms of WER for the CHiME-3 evaluation set. Our future work will include applying the proposed ASR-level front-end optimization approach to other speech enhancement methods such as beamforming.

9. References

[1] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, 2015, article ID 2015:60, doi:10.1186/s13634-015-0245-7.

[2] 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. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 436–443.

[3] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. Int. Worksh. Acoust. Echo, Noise Contr.*, 2014.

[4] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2504–2508.

[5] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, "Speaker localization and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *Proc. Worksh. Automat. Speech Recognition, Understanding.*, 2015, pp. 30–36.

[6] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5075–5079.

[7] J. R. Hershey, J. L. Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint*, 2014, arXiv:1409.2574v4.

[8] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 66–70.

[9] S. Wisdom, J. Hershey, J. L. Roux, and S. Watanabe, "Deep unfolding for multichannel source separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 121–125.

[10] D. Yu and L. Deng, *Automatic speech recognition: a deep learning approach*. Springer, 2015.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[12] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," *Adv. Neural Inform. Process. Syst.*, vol. 13, pp. 953–960, 2007.

[13] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 33–36.

[14] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[15] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[16] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G. Dahl, and B. Ramabhadrana, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.