SIMULTANEOUS OPTIMIZATION OF FORGETTING FACTOR AND TIME-FREQUENCY MASK FOR BLOCK ONLINE MULTI-CHANNEL SPEECH ENHANCEMENT

Masahito Togami

LINE Corporation

ABSTRACT

In this paper, we propose a block-online multi-channel speech enhancement technique which simultaneously optimizes time-frequency masks and forgetting factors for estimation of multi-channel covariance matrices of the desired speech signal and the noise signal so as to maximize speech enhancement performance under the condition that environmental changes occur. The proposed method reduces the noise signal by using a multi-channel Wiener filter (MWF) which is generated by the covariance matrices with the estimated forgetting factors and the estimated time-frequency masks which are outputs of the proposed neural network. The proposed method learns all the parameters of the proposed neural network so as to maximize the speech enhancement performance. Three types of the input features for the forgetting factors adaptation are proposed. The first one is the magnitude spectral of the microphone input signal. The second one is the MWF output with the previous-block filter that is adapted in the previous block. The third one is the inner product between the microphone input signal and the estimated covariance matrices in the previous block. Experimental results show that the proposed method can reduce noise signal more accurately than the conventional equally weight sample averaging.

Index Terms— Deep Learning, block online speech enhancement, multi-channel Wiener filtering, forgetting factor adaptation, time-frequency mask estimation

1. INTRODUCTION

Microphone input signal is typically contaminated by background noise signal, which is a cause for degradation of speech quality in human listening devices and degradation of automatic speech recognition accuracy. Therefore, previously, many noise reduction techniques have been studied for extracting only desired speech signal from microphone input signal [1]. Noise reduction techniques can be categorized into two categories, i.e., 1) offline techniques which outputs noise reduced signal after that whole speech signal is recorded and 2) online techniques which outputs noise reduced signal time-by-time. For human-listening purpose or automatic speech recognition systems which is needed to output recognition results on the fly, it is highly needed to output the noise reduced signal with minimum time-delay, so online noise reduction techniques are highly needed.

Single-channel noise reduction techniques [2] have been widely studied as online stationary noise reduction techniques. Time-frequency power spectral of a speech source is estimated by multiplying estimated Signal-to-Noise Ratio (SNR) with microphone input signal power. One SNR estimation method is the decision directed (DD) approach [3]. A posteriori SNR and a priori SNR are alternately updated with the estimated noise power in the DD approach. The noise power is adaptively estimated under the assumption that the noise signal is stationary, e.g., minima controlled recursive averaging (MCRA) approaches [4]. Therefore, when the noise signal is non-stationary, it is highly difficult to reduce the noise signal by the conventional single-channel noise reduction techniques.

Multi-channel noise reduction techniques [1], e.g., minimum variance distortion-less response (MVDR) beamformer [5] can reduce non-stationary noise signal by controlling spatial directivity. MVDR beamformers do not utilize temporal characteristics of the desired speech signal and the noise signal. Recently, local Gaussian modeling (LGM) based multi-channel Wiener filtering (MWF) have been also actively studied [6], which can be regarded as a combination of a MVDR beamformer and a single-channel noise reduction. The MWF techniques calculate SNR of the microphone input signal by estimating a speech source activity at each time-frequency point so as to perform single-channel noise reduction. Multi-channel noise reduction techniques require for the second order statistics of the desired speech signal and the noise signal. There are online extensions of time-varying MWF techniques [7, 8], which estimates the second order statistics with forgetting factors in an online manner. Typically, the forgetting factors have been defined empirically. However, it is difficult to configure the forgetting factors so as to fit any acoustic conditions, and the predefined forgetting factors cannot track sudden acoustic enviromental changes.

Recently, neural network based mask estimation (NNME) approaches have been applied for MVDR adaptation [9, 10, 11, 12, 13]. Similar to conventional time-frequency masking based MVDR techniques [14, 15, 16, 17, 18], NNME approaches estimate the second order statistics with time-frequency masking. An online estimation technique of the second order statistics with time-frequency masking is proposed in [11]. However, only mask estimation is done in this method. The forgetting factors do not utilize in this method. Old samples and new samples equally affect the estimated covariance matrices. This leads to slow tracking speed for acoustic environmental changes. Especially, when a speech enhancement technique is applied for human listening devices, it is inevitable to forget the past acoustic information, because microphone input signal is recorded continuously. Therefore, how to configure the forgetting factors is an important issue in speech enhancement techniques.

In this paper, we propose a block-online multi-channel speech enhancement technique, which estimates forgetting factors adaptively with a neural network depending on the microphone input signal. The proposed method also estimates a time-frequency mask which is utilized for multi-channel covariance matrices estimation in the same neural network. Unlike the conventional methods that estimates the parameters of the neural network for a time-frequency mask estimation so as to minimize the estimation error of the time-frequency mask, the proposed method learns all the parameters of the neural network so as to maximize the speech enhancement performance. Experimental results show that the proposed method can reduce noise signal more accurately than the conventional equally

weight sample averaging.

2. PROBLEM STATEMENT

2.1. Input signal model

Multi-channel microphone input signal in the time-frequency domain, $\boldsymbol{x}_{l,k} \in \boldsymbol{C}^{N_m}$ (l is the frame index, k is the frequency index, and N_m is the number of the microphones), is defined as follows:

$$\boldsymbol{x}_{l,k} = \boldsymbol{s}_{l,k} + \boldsymbol{n}_{l,k}, \tag{1}$$

where $\boldsymbol{s}_{l,k}$ is the desired speech signal and $\boldsymbol{n}_{l,k}$ is the noise signal. The purpose of speech enhancement is to extract the desired signal $\boldsymbol{s}_{l,k}$ from the observed microphone input signal $\boldsymbol{x}_{l,k}$. In the block-online speech enhancement, $X_b = \{\boldsymbol{x}\}_{l=bL_b \cdots bL_b+L_b-1,k}$ is assumed to be given, where b is the block index and L_b is the frame length of each block. From X_b, parameters which is needed for speech enhancement are adapted.

2.2. Multi-channel spatial filtering

There are several multi-channel spatial filtering techniques which extracts $\boldsymbol{s}_{l,k}$ from the microphone input signal $\boldsymbol{x}_{l,k}$. In the multi-channel spatial filtering framework, the desired speech signal is estimated as follows:

$$\boldsymbol{y}_{\text{out},l,k} = \boldsymbol{W}_{l,k}\boldsymbol{x}_{l,k}, \tag{2}$$

where $\boldsymbol{W}_{l,k}$ is a $N_m \times N_m$ separation matrix. As an adaptation technique of the separation matrix $\boldsymbol{W}_{l,k}$, the multi-channel Wiener filtering (MWF) is one of commonly utilized techniques. The MWF technique adapts $\boldsymbol{W}_{l,k}$ so as to minimize a mean square error between $\boldsymbol{y}_{\text{out},l,k}$ and $\boldsymbol{s}_{l,k}$ under the assumption that $\boldsymbol{s}_{l,k}$ and $\boldsymbol{n}_{l,k}$ are uncorrelated as follows:

$$\boldsymbol{W}_{l,k} = \boldsymbol{R}_{s,l,k}(\boldsymbol{R}_{s,l,k} + \boldsymbol{R}_{n,l,k})^{-1}, \tag{3}$$

where $\boldsymbol{R}_{s,l,k}$ is the covariance matrix of the desired speech signal and $\boldsymbol{R}_{n,l,k}$ is the covariance matrix of the noise signal. Therefore, after the covariance matrices of the desired speech signal and the noise signal are estimated, we can obtain the multi-channel Wiener filter based on Eq. 3. The two covariance matrices are defined as follows:

$$\boldsymbol{R}_{s,l,k} = E[\boldsymbol{s}_{l,k}\boldsymbol{s}_{l,k}^H], \tag{4}$$

$$\boldsymbol{R}_{n,l,k} = E[\boldsymbol{n}_{l,k}\boldsymbol{n}_{l,k}^H], \tag{5}$$

where H is the Hermite transpose operator of a matrix/vector and E is the operator of mathematical expectation. An easy way to estimate the covariance matrices is to approximate the covariance matrices by the averaged sample covariance matrices as follows:

$$\boldsymbol{R}_{s,l,k} \approx \frac{1}{L_l}\sum_{\tau} \boldsymbol{s}_{l-\tau,k}\boldsymbol{s}_{l-\tau,k}^H, \tag{6}$$

$$\boldsymbol{R}_{n,l,k} \approx \frac{1}{L_l}\sum_{\tau} \boldsymbol{n}_{l-\tau,k}\boldsymbol{n}_{l-\tau,k}^H, \tag{7}$$

where L_l is the length of the averaged frames. The sample covariance matrices estimated by Eq. 6 and Eq. 7 equally weighs each time sample. However, equally weight sample averaging is not appropriate for tracking environmental change, because old samples before the environmental change will affect estimation of the covariance matrices to the same extent as new samples after the environmental change. Another problem is how to estimate the desired speech signal, $\boldsymbol{s}_{l,k}$ and the noise signal, $\boldsymbol{n}_{l,k}$, because we can observe only noisy microphone input signal $\boldsymbol{x}_{l,k}$.

2.3. Covariance matrix estimation based on time-frequency mask

In time-frequency mask based methods [14, 15, 16, 9, 11], prior to estimation of the desired covariance matrix and the noise covariance matrix, $\boldsymbol{s}_{l,k}$ and $\boldsymbol{n}_{l,k}$ are estimated by using time-frequency masks as follows:

$$\hat{\boldsymbol{s}}_{l,k} = M_{s,l,k}\boldsymbol{x}_{l,k}, \tag{8}$$

$$\hat{\boldsymbol{n}}_{l,k} = M_{n,l,k}\boldsymbol{x}_{l,k}, \tag{9}$$

where $M_{s,l,k}$ and $M_{n,l,k}$ are the time frequency masks. $\hat{\boldsymbol{s}}_{l,k}$ and $\hat{\boldsymbol{n}}_{l,k}$ are roughly estimated under the assumption that there is only one speech source at each time-frequency point. By using estimated $\hat{\boldsymbol{s}}_{l,k}$ and $\hat{\boldsymbol{n}}_{l,k}$ as alternatives of $\boldsymbol{s}_{l,k}$ and $\boldsymbol{n}_{l,k}$ in Eq. 6 and Eq. 7, the covariance matrices, $\boldsymbol{R}_{s,l,k}$ and $\boldsymbol{R}_{n,l,k}$, can be estimated.

One way to increase the tracking speed for environmental changes is to estimate the sample covariance matrices with exponentially decay weights. In the conventional methods [15, 16], the sample covariance matrices with exponentially decay weights are estimated with forgetting factors and time-frequency masks in a block-online way as follows:

$$\boldsymbol{R}_{s,b,k} = \alpha_s\boldsymbol{R}_{s,b-1,k} + \frac{1-\alpha_s}{L_b}\sum_{l=bL_b\cdots bL_b+L_b-1} M_{s,l,k}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H, \tag{10}$$

$$\boldsymbol{R}_{n,b,k} = \alpha_n\boldsymbol{R}_{n,b-1,k} + \frac{1-\alpha_n}{L_b}\sum_{l=bL_b\cdots bL_b+L_b-1} M_{n,l,k}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H, \tag{11}$$

where $\boldsymbol{R}_{s,b,k}$ and $\boldsymbol{R}_{s,n,k}$ are the covariance matrices estimated in the bth block and α_s and α_n are the forgetting factors. The forgetting factors control tracking speed for environmental changes. When the environmental changes occur suddenly, α should be a small value. On the other hand, when acoustic environments are close to be stationary, α should be a big value. Therefore, fixed forgetting factors are not appropriate for the environments in which environmental changes occur, and adaptive forgetting factors are required so as to track acoustic environmental changes.

3. PROPOSED METHOD

3.1. Overview

Instead of fixed-valued forgetting factors, the forgetting factors are estimated adaptively in the proposed method. The block diagram of the proposed method is shown in Fig. 1. The adaptive forgetting factors are estimated via a neural network at each time-frequency point. The time-frequency mask are also estimated via a neural network at each time-frequency point. Instead of the conventional time-frequency mask estimation method that minimizes the mask estimation error [9, 11], the proposed method estimates both the time frequency masks and the forgetting factors jointly under the same cost function so as to maximize speech enhancement performance.

3.2. Time-frequency mask and forgetting factors estimation

The proposed method estimates the time-frequency mask of the desired speech source and the noise signal, $M_{s,l,k}$ and $M_{n,l,k}$ by using multiple dense layers with batch normalization and dropout (training phase only). The magnitude spectral of the microphone input signal is utilized as the input feature. The forgetting factors are also adapted based on multiple dense layers with batch normalization and dropout (training phase only). The proposed method utilizes three types of input features, f_b. The first one is the magnitude spectral of

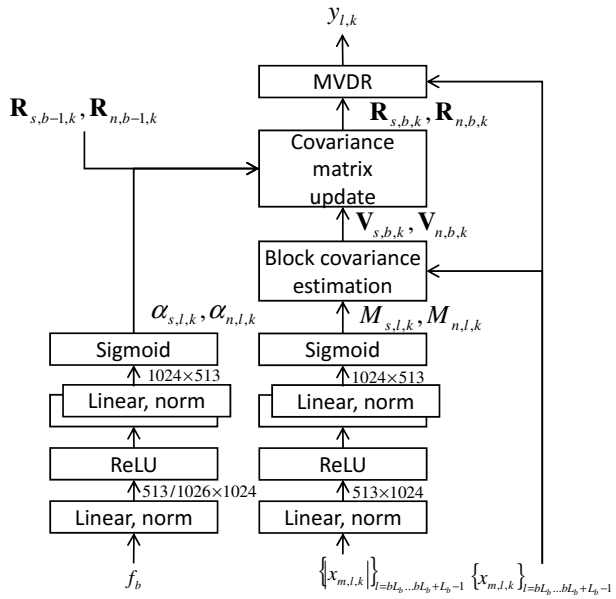

Fig. 1. Block diagram of proposed method

the microphone input signal which is the same input feature as the time-frequency mask. The second one is the magnitude spectral of the output signal of speech enhancement, $\boldsymbol{y}_{\text{pre},l,k}$, which is obtained by the MWF adapted in the previous block as follows:

$$\boldsymbol{y}_{\text{pre},l,k} = \boldsymbol{W}_{b-1,k}\boldsymbol{x}_{l,k}, \tag{12}$$

where $\boldsymbol{W}_{b-1,k}$ is the MWF that is estimated from the covariance matrices $\boldsymbol{R}_{s,b-1,k}$ and $\boldsymbol{R}_{n,b-1,k}$ that is estimated in the previous block. The second feature reflects the amount of the acoustical environmental change between the $b-1$th block and the bth block, because when the environmental changes occur, the amount of the noise or the amount of speech distortion in $\boldsymbol{y}_{\text{pre},l,k}$ will increase.

The third one is the inner product between the previous covariance matrices and the microphone input signal, $P_{s,b,k}$ and $P_{n,b,k}$, which is defined as follows:

$$P_{s,l,k} = \boldsymbol{x}_{l,k}\boldsymbol{R}_{s,b-1,k}\boldsymbol{x}_{l,k}, \tag{13}$$

$$P_{n,l,k} = \boldsymbol{x}_{l,k}\boldsymbol{R}_{n,b-1,k}\boldsymbol{x}_{l,k}. \tag{14}$$

A covariance matrix contains the acoustical spatial information. Therefore, $P_{s,b,k}$ and $P_{n,b,k}$ can be interpreted as the amount of the environmental change such as the change of the desired source location, the change of the noise source location.

The forgetting factors $\alpha_{s,l,k}$ and $\alpha_{n,l,k}$ are estimated via a neural network with one of the three input features. After estimating the forgetting factors for each frame, $\alpha_{s,l,k}$ and $\alpha_{n,l,k}$ are averaged in the bth block, and the averaged forgetting factors $\alpha_{s,b,k}$ and $\alpha_{n,b,k}$ are used for the covariance matrices adaptation.

3.3. Parameter learning in an end-to-end manner

All the parameters in the proposed method are learned so as to minimize the distance between the estimated speech source signal and the target speech source signal. The loss function based on the weighting average of the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) is defined as follows:

$$\begin{aligned} Q_\theta &= -\frac{\beta}{K}\sum_k \log \text{SDR}_\theta(\boldsymbol{y}_{\text{out},l,k}, \boldsymbol{s}_{l,k}) \\ &- \frac{1-\beta}{K}\sum_k \log \text{SIR}_\theta(\boldsymbol{y}_{\text{out},l,k}, \boldsymbol{n}_{\text{out},l,k}), \end{aligned} \tag{15}$$

where $\boldsymbol{n}_{\text{out},l,k}$ is the noise signal contained in the output signal , θ is the parameter of the proposed neural network, SDR is the SDR of the output signal , and SIR is the SIR of the output signal. SDR and SIR at each frequency bin is defined as follows:

$$\text{SDR}_\theta(\boldsymbol{y}_{\text{out},l,k}, \boldsymbol{s}_{l,k}) = \frac{\sum_l \|\boldsymbol{s}_{l,k}\|^2}{\sum_l \|\boldsymbol{s}_{l,k} - \boldsymbol{y}_{\text{out},l,k}\|^2}, \tag{16}$$

$$\text{SIR}_\theta(\boldsymbol{y}_{\text{out},l,k}, \boldsymbol{s}_{l,k}) = \frac{\sum_l \|\boldsymbol{y}_{\text{out},l,k}\|^2}{\sum_l \|\boldsymbol{n}_{\text{out},l,k}\|^2}. \tag{17}$$

The proposed method utilizes $\beta = 0.5$.

4. EVALUATION

Table 1. Evaluation results with CHiME-3 dataset: SDR (dB) and SIR (dB)

	Training		Development	
Approaches	SDR (dB)	SIR (dB)	SDR (dB)	SIR (dB)
(a)	13.65	16.20	12.97	14.02
(b)	**18.15**	**18.73**	15.78	15.71
(c)	17.93	18.69	**16.07**	**15.88**
(d)	16.71	17.67	14.77	15.06

4.1. Experimental setup

The proposed method evaluated by using CHiME-3 dataset [19]. CHiME-3 dataset consists of recording speech signal in four noisy areas, i.e., a bus, cafe, pedestrian, and street junction area. In the training phase, 7138 simulated noisy data is utilized. The best parameter of the neural network is selected based on the averaged loss value of 1640 simulated development data. Adam optimizer (learning rate was 0.001) was utilized. The number of the epochs was set to 25. Early stopping based on the averaged loss of the development data was also utilized, and patience was set to 5. The sampling rate was 16000Hz. The number of the microphones N_m was 6. The microphone input signal was converted into time-frequency domain via a short-term Fourier transform. Frame size was set to 1024 pt. Frame shift was set to 512 pt. Hanning window was utilized. The SIR and SDR (dB) were evaluated for the training dataset and the development dataset. The number of the frame length in each block L_b was set to 5. Back propagation was performed for each block separately. The mini-batch size was set to 100.

4.2. Neural network architecture

4.2.1. Time-frequency mask estimation

The neural network for time-frequency mask estimation consists of two layers. The first dense layer has 513 input units and 1024 output

units with a ReLU function. The second dense layer has 1024 input units and 1024 output units with a Sigmoid function. In each dense layer, batch normalization is performed for the input feature. The first dense layer performs dropout (Probability was 0.5).

4.2.2. Forgetting factor estimation

The neural network for forgetting factor estimation consists of two layers. The number of the input units in the first dense layer is depending on the input feature. The first dense layer has 513 or 1026 units and 1024 output units with a ReLU function. The second dense layer has 1024 input units and 1024 output units with a Sigmoid function. In each dense layer, batch normalization is performed for the input feature. The first dense layer performs dropout (Probability was 0.5).

4.3. Comparative methods

We compared the following four methods:

- (a) **Equally weight sample averaging**: Similar to the conventional method [11], the covariance matrices at each block are estimated by using the equally weight sample averaging, Eq. 6 and Eq. 7.
- (b) **Forgetting factors adaptation (Input)**: Forgetting factors are adapted by the proposed method. The magnitude spectral of the microphone input signal is utilized as the input feature of the neural network for forgetting factors adaptation.
- (c) **Forgetting factors adaptation (Output)**: Forgetting factors are adapted by the proposed method. The magnitude spectral of the output enhanced speech signal by the MWF filter adapted in the previous block (Eq. 12) is utilized as the input feature.
- (d) **Forgetting factors adaptation (Inner product)**: Forgetting factors are adapted by the proposed method. The inner products between the previous covariance matrices and the current microphone input signal (Eq. 13 and Eq. 14) are utilized as the input feature.

4.4. Experimental results

The experimental result is shown in Table 1. It is shown that the proposed method with the adaptive forgetting factors outperformed the Equally weight sample averaging. The input feature for the adaptive forgetting factors based on the magnitude spectral of the microphone input signal was the best for the training dataset. However, for the development dataset, the magnitude spectral of the MWF output with the previous MWF filter was the best. This means that there was a little bit over-fitting for the training dataset in the "Forgetting factors adaptation (Input)", and the magnitude spectral of the MWF output with the previous MWF filter that reflects the amount of the environmental change from the previous block to the current block is effective so as to control the tracking speed. Samples of output spectrograms are shown in Fig. 4.4. The output signal of the BeamformIt toolkit [20] is also shown. The area was set to "bus". It is shown that the proposed method can reduce more noise than BeamformIt and equally weight sample averaging.

5. CONCLUSION

In this paper, we proposed a block-online speech enhancement technique, which can estimate forgetting factors and time-frequency masks based on a neural network. The proposed method estimates forgetting factors and time-frequency masks jointly based on the same loss function so as to maximize the speech enhancement performance. Three types of the input features for adaptation of the forgetting factors have been proposed. Experimental results show that the proposed method outperformed the conventional equally weight sample averaging based method from the speech enhancement perspective. Furthermore, it was shown that the input feature for adaptation of the forgetting factors based on the spatial filtering results by using the filter obtained at the previous block achieved the best performance.

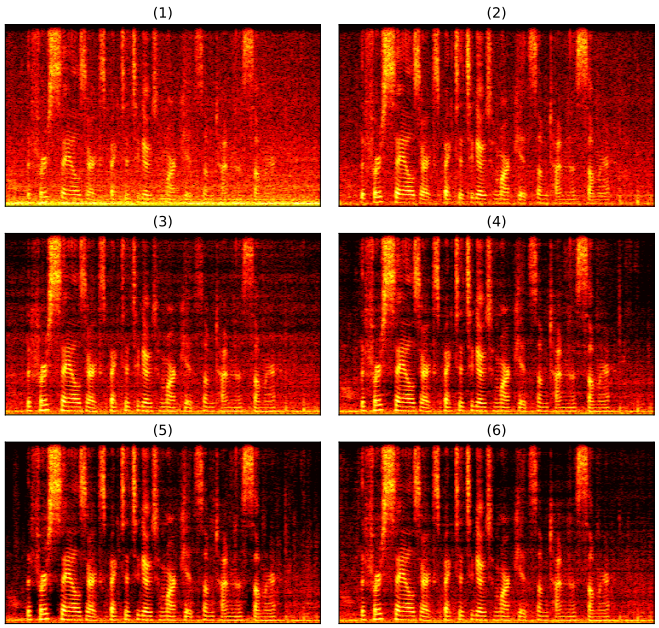

Fig. 2. Examples of spectrogram: (1) Microphone input signal, (2) BeamFormIT, (3) Equally weight sample averaging, (4) Forgetting factors adaptation (Input), (5) Forgetting factors adaptation (Output), (6) Forgetting factors adaptation (Inner product)

6. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Springer Publishing Company, Incorporated, 1st edition, 2010.

[2] P.C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, CRC Press, 1st edition, 2007.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan 2002.

[5] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug 1972.

[6] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] M. Togami, "Online speech source separation based on maximum likelihood of local gaussian modeling," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 213–216.

[8] L.S. Simon and E. Vincent, "A general framework for online audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 397–404.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[10] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 536–540.

[11] H. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 531–535.

[12] Z. Wang and D. Wang, "Mask weighted stft ratios for relative transfer function estimation and its application to robust asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5619–5623.

[13] Y. Liu, A. Ganguly, K. Kamath, and T. Kristijansson, "Neural network based time-frequency masking and steering vector estimation for two-channel mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 6717–6721.

[14] M. Togami and A. Amano, "Adaptation methodology for minimum variance beam-former based on frequency segregation," in *Proc. of the 2005 Autumn Meeting of the Acoustical Society of Japan (in Japanese)*, Sept. 2005.

[15] M. Togami, Y. Obuchi, and A. Amano, *Automatic Speech Recognition of Human-Symbiotic Robot EMIEW*, chapter 22, pp. 395–404, I-tech Education and Publishing, 2007.

[16] Y. Kawaguchi and M. Togami, "Soft masking based adaptation for time-frequency beamformers under reverberant and background noise environments," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aug. 2010, pp. 736–740.

[17] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5210–5214.

[18] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based mvdr beamformer," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5694–5698.

[19] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 504–511.

[20] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.