REAL-TIME STEREO SPEECH ENHANCEMENT WITH SPATIAL-CUE PRESERVATION BASED ON DUAL-PATH STRUCTURE

Masahito Togami, Jean-Marc Valin, Karim Helwani, Ritwik Giri, Umut Isik, and Michael M. Goodwin

Amazon Web Services, Palo Alto, CA, USA

ABSTRACT

We propose a real-time stereo speech enhancement algorithm which preserves spatial cues under the condition that there are multiple speech sources. Instead of the conventional common-gain based method with a single-path structure that enhances a single source with spatial-cue preservation, we propose a novel common-gain based method with a dual-path structure for enhancement of multiple speech sources with spatial-cue preservation. The proposed method enhances the dominant source and the other sources separately with a monaural speech enhancement algorithm to increase speech enhancement performance. A final stereo signal is obtained by remixing spatial images of the enhanced speech sources. The approach provides perfect reconstruction of multiple speech sources with spatial-cue preservation under the condition that the monaural speech enhancement is ideal. Separation of the dominant source and the other sources is carried out by spatial beamforming with estimated steering vectors. The steering vectors are estimated based on a time-frequency mask calculated using the output of the speech enhancement as feedback. We evaluate the proposed method with two datasets, one with fully overlapped mixtures and one with sparsely overlapped mixtures. Objective and subjective evaluation results show that the proposed method can provide improved speech enhancement and spatial-cue preservation with respect to discrete-channel processing and the conventional single-path common-gain method for both types of mixtures.

Index Terms— speech enhancement, multichannel processing, spatial-cue preservation, common gain

1. INTRODUCTION

Teleconferencing systems are often used in noisy and reverberant environments, so speech enhancement techniques are needed to ensure clear communication [1, 2]. In teleconferencing applications, enabling two-way communication imposes constraints of real-time processing and low input/output latency. Additionally, stereo input is becoming increasingly common in current teleconferencing situations, so speech enhancement methods for two-channel input are needed. In such solutions, it is important not only to achieve high speech enhancement performance but also to preserve the spatial cues of speech sources, because spatial-cue information is an important clue to know who is speaking.

Many approaches for monaural speech enhancement have been studied [1, 3, 4], e.g. spectral subtraction [4] and Bayesian methods such as minimum mean-square error (MMSE) methods [3]. However, the speech enhancement performance of MMSE methods is not sufficient for nonstationary noise. Recently, deep neural network (DNN) based methods have been applied for monaural speech enhancement or speech source separation [5, 6, 7, 8, 9, 10]. Due to the strong expression capability of DNNs for speech characteristics, the performance of monaural speech enhancement algorithms for non-stationary noise conditions has improved dramatically.

While research on DNN-based speech enhancement has focused on monaural algorithms, multichannel speech enhancement using DNNs has also been recently explored in an effort to improve on established classical methods [11, 12]. Spatial beamforming with time-frequency masking [11] is a popular approach which combines DNN and multichannel spatial beamforming. However, the speech enhancement performance of DNN-based multichannel spatial beamforming is constrained by the upper performance limit of the traditional spatial beamformer in the system. Recently, methods which incorporate DNNs more directly into multichannel processing have been studied, e.g., [13, 14, 15, 16]. However, it is necessary to train a specific DNN model for multichannel signals.

In parallel with the study of DNN-based multichannel speech enhancement, multichannel speech enhancement techniques with spatial-cue preservation based on traditional signal processing frameworks have been also actively studied [17, 18, 19, 20, 21, 22, 23]. The common-gain method [23] is a straightforward speech enhancement approach with spatial-cue preservation for single-source cases. It estimates a common time-frequency gain for all microphone input signals. The common-gain based method ensures preservation of the interaural phase difference (IPD) and interaural level difference (ILD) between channels.

In this paper, we propose a real-time stereo speech enhancement method with spatial-cue preservation under the condition that there are multiple speech sources. The proposed method adopts a novel common-gain based method with a dual-path structure in which the main speaker and additional speakers are enhanced separately to increase speech enhancement performance, after which the sources are remixed with spatial-cue preservation. The proposed dual-path structure ensures that multiple speech sources are perfectly reconstructed with spatial-cue preservation under the condition where the monaural speech enhancement works ideally. In the proposed system, separation is carried out by a spatial beamformer which is adapted using DNN-based time-frequency masking. Speech enhancement is done by using both spatial beamforming and a pretrained DNN-based monaural speech enhancement. The output signal of the DNN-based monaural speech enhancement is also utilized for estimation of the time-frequency masks. Thus, it is not needed to train a stereo-specific DNN model. The state-of-the art PercepNet algorithm [24] is utilized as the monaural speech enhancer. We evaluate the proposed method using two datasets with different speech overlap characteristics. Objective and subjective evaluation results are given to demonstrate the effectiveness of the proposed method.

2. PROBLEM STATEMENT

2.1. Signal model

It is assumed that there are two microphone input signals. The m-th microphone input signal $x_{m,t}$ (t is the time-index) is modeled in the time domain as follows:

$$x_{m,t} = \sum_{i=1}^{N_s} s_{i,t} * h_{i,m} + n_t, \tag{1}$$

where $s_{i,t}$ is the i-th speech source signal, N_s is the number of speech sources, $h_{i,m}$ is the impulse response between the i-th speech source location and the m-th microphone location, and n_t is the background noise signal. Speech enhancement is carried out in the time-frequency domain. The time-frequency representation of $x_{m,t}$ can be written as follows:

$$\boldsymbol{x}_{l,k} = \sum_{i=1}^{N_s} s_{i,l,k}\boldsymbol{a}_{i,k} + \boldsymbol{n}_{l,k}, \tag{2}$$

where l is the frame-index, k is the frequency index, $\boldsymbol{x}_{l,k} = [\ x_{1,l,k}^T \quad x_{2,l,k}^T\ ]^T$, T is the transpose operator of a matrix or a vector, and $x_{m,l,k}$ is the time-frequency representation of the time-domain signal $x_{m,t}$. $s_{i,l,k}$ and $\boldsymbol{n}_{l,k}$ are defined similarly. $\boldsymbol{a}_{i,k} = [\ a_{i,1,k} \quad a_{i,2,k}\ ]^T$ is a steering vector and $a_{i,m,k}$ is the time-frequency representation of $h_{i,m}$.

In this paper, we focus on stereo speech enhancement with spatial-cue preservation. Thus, the objective is defined as extraction of $\sum_{i=1}^{N_s} s_{i,l,k}\boldsymbol{a}_{i,k}$ from the microphone input signal $\boldsymbol{x}_{l,k}$. We assume that a pretrained DNN-based monaural speech enhancement method is incorporated in the system. For the experiments in this paper, the state-of-the art PercepNet method [24] is used for monaural speech enhancement. PercepNet satisfies our design requirements in for real-time, low-latency, high-quality enhancement; it operates on 10-ms frames with 30 ms of lookahead, and ranked second in the real-time track of the recent DNS challenge despite operating at much lower than the allowed complexity [25].

2.2. Discrete channel processing

One approach for enhancing a stereo speech signal is to perform monaural speech enhancement independently for each channel (Fig. 1). We use this discrete-channel processing approach as a baseline for assessing our algorithm. In discrete-channel processing, it is not guaranteed that each speech source is enhanced the same way in each output channel, and as a result the output spatial image can be unstable.

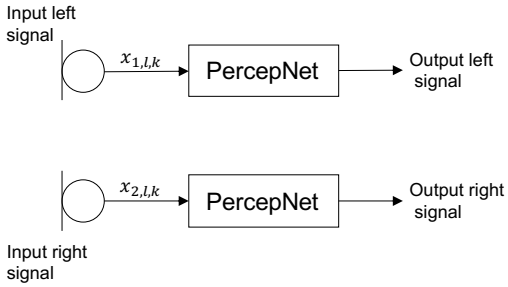

Fig. 1. Block diagram of discrete-channel processing method.

2.3. Common gain method

An alternate method is proposed in [23] where a common time-frequency gain for both microphone signals is estimated (Fig. 2). The output signal of each channel is obtained by multiplying the common time-frequency gain by the input signal of the corresponding channel. The common-gain method ensures that the ILD and IPD of the output stereo signal are the same as those of the microphone input stereo signal. In our implementation of this approach depicted in Fig. 2, the common band gain is calculated by PercepNet operating on a downmix of the left-channel input signal and the right-channel input signal. PercepNet estimates N band gains, where N is set to be smaller than the number of frequency bins in the input time-frequency representation. The band gains derived from the downmix are then shared between channels. Time-frequency gains are calculated by interpolating the shared band gains along the frequency axis with the envelop postfiltering [24]. Although each band gain is shared between channels, the time-frequency gain $G_{l,m,k}$ varies on each channel due to the influence of global gain compensation in the envelope postfiltering. We use this common-gain method as a baseline method in evaluation.

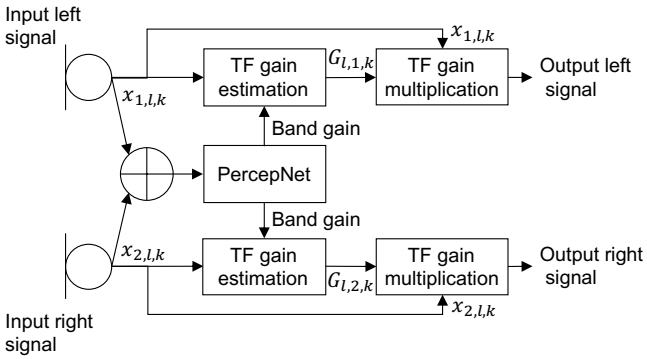

Fig. 2. Block diagram of common-gain method.

3. PROPOSED FRAMEWORK

3.1. Overview

An overview of the proposed framework is shown in Fig. 3. The approach is designed to achieve high-quality speech enhancement using spatial selectivity as well as monaural speech enhancement. Considering the top half of the diagram, a delay-and-sum beamformer (DSBF) is first used to derive enhanced speech signals to serve as inputs to common-gain estimation. The DSBF steering vector is determined from the outputs of the monaural speech enhancement model and fed back to the beamformer, as will be explained later. In the common-gain estimation stage, a spatial image of the enhanced speech signal for each channel is used instead of the microphone input signals; this provides improved speech enhancement performance. To ensure robustness in the presence of multiple speech sources, we use a dual-path structure with two instances of the DSBF and common-gain method. In the first path (the top half), a dominant speech source is enhanced. In the second path (the bottom half), any other speech sources are enhanced. Separating the speech sources initially via spatial beamforming improves the signal-to-noise ratio of the input signal for the monaural speech enhancement and correspondingly improves the quality of the output signal. After the two paths are processed, the output signal for each

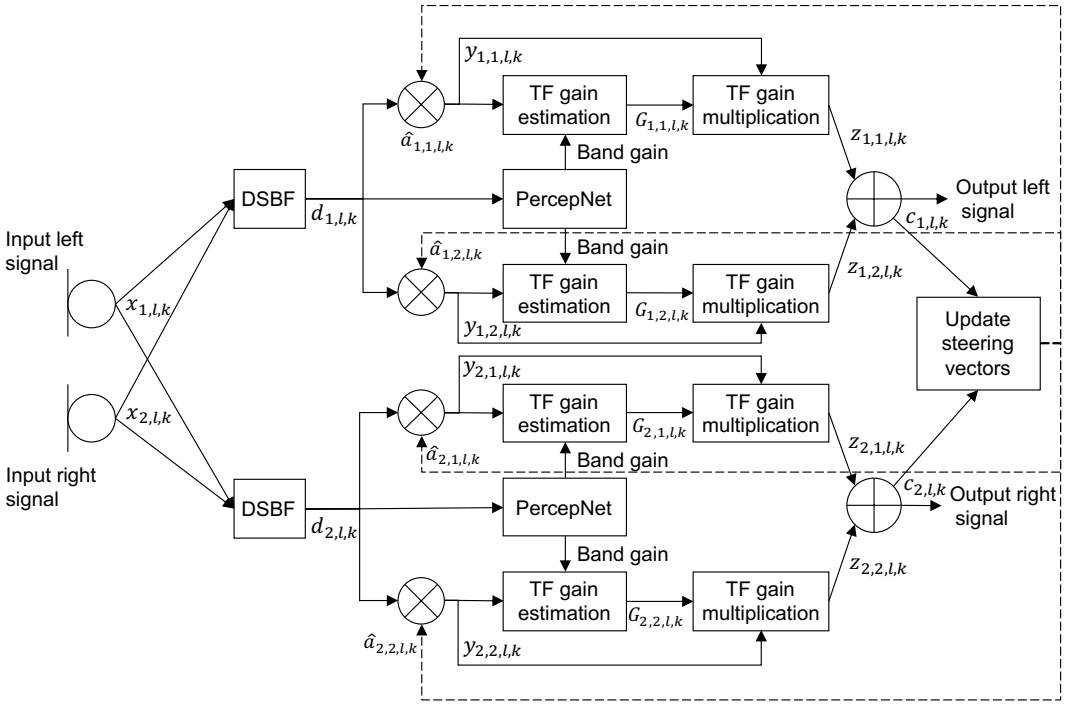

Fig. 3. Block diagram of proposed stereo speech enhancement algorithm.

channel is generated by remixing the respective output channel signals of both paths. Note that under the assumption that the monaural speech enhancement algorithm provides a distortion-free speech output, the dual-path system ensures that multiple speech sources are perfectly reconstructed.

3.2. Spatial beamforming

Let $\hat{\boldsymbol{a}}_{1,l,k}$, $\hat{\boldsymbol{a}}_{2,l,k}$ be the estimated steering vectors of the first path and the second path, respectively. The input signal for the monaural speech enhancement for each path $d_{i,l,k}$ is obtained by using the DSBF as follows:

$$d_{i,l,k} = \hat{\boldsymbol{a}}_{i,l,k}^{H} \boldsymbol{x}_{lk}, \tag{3}$$

where H denotes the Hermitian transpose operator. The m-th input signal for the i-th path, $y_{i,m,l,k}$, is estimated as a spatial image of the enhanced speech signal of the i-th path as follows:

$$y_{i,m,l,k} = d_{i,l,k} \hat{a}_{i,m,l,k}. \tag{4}$$

3.3. Monaural speech enhancement

For each path, PercepNet estimates a common band gain from $d_{i,l,k}$. Each band gain is shared between channels and a time-frequency gain $G_{i,m,l,k}$ is calculated by interpolating the shared band gains along the frequency axis with the envelop postfiltering. The output signal after monaural speech enhancement for the m-th channel of the i-th path $z_{i,m,l,k}$ is obtained as follows:

$$z_{i,m,l,k} = y_{i,m,l,k} G_{i,m,l,k}, \tag{5}$$

where $G_{i,m,l,k}$ is the time-frequency gain. The final output signal for each channel $c_{m,l,k}$ is obtained as follows:

$$c_{m,l,k} = \sum_{i=1}^{2} z_{i,m,l,k}. \tag{6}$$

We also evaluate a method with a single-path structure and spatial beamforming with the estimated steering vector, in which the output signal is obtained as $c_{m,l,k} = z_{1,m,l,k}$.

3.4. Estimation of steering vectors

To perform DSBF effectively, it is necessary to estimate the steering vector $\hat{\boldsymbol{a}}_{i,l,k}$ accurately. The proposed method estimates $\hat{\boldsymbol{a}}_{1,l,k}$ as the steering vector of the dominant speech source by using principal component analysis (PCA) with an estimated spatial covariance matrix (SCM). The steering vector for the second path $\hat{\boldsymbol{a}}_{2,l,k}$ is estimated such that $\hat{\boldsymbol{a}}_{2,l,k}$ is orthogonal to $\hat{\boldsymbol{a}}_{1,l,k}$. The l^2-norm of each steering vector is set to 1. The SCM $\boldsymbol{R}_{l,k}$ of the dominant speech source is updated in an online manner as follows:

$$\boldsymbol{R}_{l,k} = \gamma_{l,k} \boldsymbol{R}_{l-1,k} + (1 - \gamma_{l,k}) \boldsymbol{x}_{l,k} \boldsymbol{x}_{l,k}^{H}, \tag{7}$$

$$\gamma_{l,k} = 1 - \mathcal{M}_{l,k}(1 - \alpha), \tag{8}$$

where α is a forgetting factor. α is set to 0.99. $\mathcal{M}_{l,k}$ is a time-frequency mask which controls updating of the SCM depending on the speech presence at each time-frequency bin. To avoid performance degradation in SCM estimation, it is important to accurately estimate this time-frequency mask, which essentially selects the time-frequency bins where speech sources are dominant. This mask estimation is done by reusing the monaural speech enhancement results:

$$\mathcal{M}_{l,k} = \min\left(\frac{\|\boldsymbol{c}_{l,k}\|}{\|\boldsymbol{x}_{l,k}\|}, 1 \right), \tag{9}$$

where $\boldsymbol{c}_{l,k} = [\ c_{1,l,k} \quad c_{2,l,k}\ ]^T$ is a stereo signal which contains enhanced speech signals. Additional DNN training or inference is not needed for steering vector estimation.

Table 1. Evaluation results

	Dataset	WSJ1			Stereo Sparse Librimix		
Single / Dual		**IPD error**	**ILD error**	**MOS**	**IPD error**	**ILD error**	**MOS**
	Discrete-channel processing	0.234	3.53	3.27 ± 0.06	0.192	3.68	3.47 ± 0.06
	Common gain (Baseline)	0.232	3.03	3.27 ± 0.06	0.194	3.22	3.48 ± 0.06
Single	Common gain (Proposed)	0.207	**1.93**	3.29 ± 0.07	0.158	**1.63**	3.48 ± 0.06
	Non-adaptive steering vectors	0.239	2.64	3.26 ± 0.06	0.187	2.72	3.51 ± 0.06
Dual	Common gain (Proposed)	**0.195**	2.65	$\mathbf{3.30} \pm 0.06$	**0.147**	2.62	$\mathbf{3.52} \pm 0.06$

3.5. Perfect reconstruction

Under the assumption that the band gain $G_{i,m,l,k}$ outputs speech sources with no degradation and completely removes noise, $\boldsymbol{c}_{l,k}$ can be written as follows:

$$\begin{aligned}\boldsymbol{c}_{l,k} &= \left(\sum_{i=1}^{2} \hat{\boldsymbol{a}}_{i,l,k}\hat{\boldsymbol{a}}_{i,l,k}^{H}\right)\sum_{n=1}^{N_s} s_{n,l,k}\boldsymbol{a}_{n,k} \\ &= \sum_{n=1}^{N_s} s_{n,l,k}\boldsymbol{a}_{n,k}.\end{aligned} \tag{10}$$

Thus, in this ideal case, stereo speech sources are perfectly reconstructed with spatial-cue preservation in the output signal. Although the conventional common-gain based method also reconstructs a stereo speech signal perfectly under the condition that the common gain outputs a speech signal without degradation, it does not ensure that multiple speech sources are enhanced without degradation. On the other hand, in the proposed method, even when there are multiple sound sources, it is possible to focus on enhancement of a relatively small number of speech sources in each common-gain based monaural speech enhancement path. Thus, it can be expected that noise can be suppressed more effectively with relatively less distortion of speech sources in the proposed dual-path structure.

4. EVALUATION

4.1. Setup

We evaluated the proposed stereo speech enhancement framework with objective and subjective experiments. The sampling rate was 16000 Hz. The number of the band gains N was set to 32. We developed our evaluation datasets by using wsj1_2345_db [1], in which room dimension, source location, microphone location, SNR, and reverberation time are simulated similarly to spatialized wsj0-2mix [26]. We added noise signals which were extracted from CHiME3 dataset [27]. For speech source signals, we used two datasets, WSJ1 [28] and LibriSpeech ASR corpus [29]. For the WSJ1 corpus, fully overlapped mixtures were generated. For the LibriSpeech ASR corpus, the overlap of multiple speech sources was determined by Sparse LibriMix [2], and the overlap ratio was set to 0.2. We call this dataset Stereo Sparse LibriMix. The number of speech sources was set to 2. We compared several methods with a single-path structure and a dual-path structure. We also evaluated "Non-adaptive steering vectors" in which the steering vector is fixed to $\hat{\boldsymbol{a}}_{1,l,k} = [\ 1 \quad 1\ ]^T$ and $\hat{\boldsymbol{a}}_{2,l,k} = [\ 1 \quad -1\ ]^T$ so as to confirm the effectiveness of adaptive steering vector estimation in the proposed method.

[1] https://github.com/fakufaku/create_wsj1_2345_db

[2] https://github.com/popcornell/SparseLibriMix

4.2. Experimental results

For our evaluations, we use IPD error and ILD error [dB] as objective measures, and mean opinion score (MOS) as a subjective measure. IPD error is defined as a distance between the IPD of an output stereo signal $\phi_{\hat{\boldsymbol{c}}}$ and the IPD of a non-reverberant stereo signal $\phi_{\boldsymbol{c}}$ as follows:

$$\text{IPD error} = \frac{|\phi_{\boldsymbol{c}} - \phi_{\hat{\boldsymbol{c}}}|}{\pi}, \tag{11}$$

where phase compensation is incorporated in the calculation. The ILD error is defined as the distance between the ILD of an output stereo signal $L_{\hat{\boldsymbol{c}}}$ and the ILD of a non-reverberant stereo signal $L_{\boldsymbol{c}}$ as follows:

$$\text{ILD error} = |20\log_{10} L_{\hat{\boldsymbol{c}}} - 20\log_{10} L_{\boldsymbol{c}}|. \tag{12}$$

For subjective testing, MOS testing [30] was carried using the crowd sourcing methodology described in P.808 [31, 32]. The number of listeners was 10. The evaluation results are shown in Table 1. The proposed common-gain method with the dual-path structure outperformed the discrete-channel processing method and the conventional common-gain method. The proposed common-gain method with the dual-path structure achieved the best performance in terms of IPD error and MOS for both the WSJ1 and Stereo Sparse Librimix datasets. Also, by comparing the proposed common-gain method with the dual-path structure with "Non-adaptive steering vectors", it is shown that the adaptive steering vector estimation in the proposed method is also effective.

5. CONCLUSIONS

We proposed a stereo speech enhancement technique which combines DNN-based monaural speech enhancement and spatial beamforming. By using a dual-path structure with a common band gain between channels in each path, the approach preserves the spatial cues of multiple input speech sources. The approach also avoids the complex training requirements of dedicated stereo DNN speech enhancement models by relying on a pretrained monaural model. Experimental results show that the method to be effective under the condition that there are multiple speech sources.

6. REFERENCES

[1] Philipos C. Loizou, *Speech enhancement : theory and practice / Philipos C. Loizou.*, CRC Press, Boca Raton, Fla, 2nd ed. edition, 2013.

[2] Jacob Benesty, Shoji Makino, and Jingdong Chen, *Speech Enhancement*, Springer, 2005.

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

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[5] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[6] Hakan Erdogan, John R. Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 708–712.

[7] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[8] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[9] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[10] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, 2016, pp. 196–200.

[12] Takuya Higuchi, Keisuke Kinoshita, Nobutaka Ito, Shigeki Karita, and Tomohiro Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *Proc. IEEE ICASSP*, 2018, pp. 531–535.

[13] Masahito Togami, "Multi-channel itakura saito distance minimization with deep neural network," in *Proc. IEEE ICASSP*, 2019, pp. 536–540.

[14] Yoshiki Masuyama, Masahito Togami, and Tatsuya Komatsu, "Multichannel Loss Function for Supervised Speech Source Separation by Mask-Based Beamforming," in *Proc. Interspeech*, 2019, pp. 2708–2712.

[15] Cong Han, Yi Luo, and Nima Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE ICASSP*, 2020, pp. 6404–6408.

[16] Bahareh Tolooshams and Kazuhito Koishida, "A training framework for stereo-aware speech enhancement using deep neural networks," *arXiv preprint arXiv:2112.04939*, 2020.

[17] Tim Van den Bogaert, Jan Wouters, Simon Doclo, and Marc Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel wiener filter," in *Proc. IEEE ICASSP*, 2007, vol. 4, pp. IV–565–IV–568.

[18] Joseph Szurley, Alexander Bertrand, Bas Van Dijk, and Marc Moonen, "Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 952–966, 2016.

[19] Andreas I. Koutrouvelis, Jesper Jensen, Meng Guo, Richard C. Hendriks, and Richard Heusdens, "Binaural speech enhancement with spatial cue preservation utilising simultaneous masking," in *Proc. EUSIPCO*, 2017, pp. 598–602.

[20] T.J. Klasen, M. Moonen, T. Van den Bogaert, and J. Wouters, "Preservation of interaural time delay for binaural hearing aids through multichannel wiener filtering based noise reduction," in *Proc. IEEE ICASSP*, 2005, vol. 3, pp. iii/29–iii/32 Vol. 3.

[21] Bram Cornelis, Simon Doclo, Tim Van dan Bogaert, Marc Moonen, and Jan Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[22] Márcio H. Costa and Patrick A. Naylor, "ILD preservation in the multichannel wiener filter for binaural hearing aid applications," in *Proc. EUSIPCO*, 2014, pp. 636–640.

[23] Stefan Thaleiser and Gerald Enzner, "Cue-preserving mmse filter with bayesian snr marginalization for binaural speech enhancement," in *Proc. IEEE ICASSP*, 2021, pp. 6124–6128.

[24] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," in *Proc. Interspeech*, 2020, pp. 2482–2486.

[25] Chandan K. A. Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," *arXiv preprint arXiv:2001.08662*, 2020.

[26] Zhong-Qiu Wang, Jonathan Le Roux, and John R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE ICASSP*, 2018, pp. 1–5.

[27] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, 2015, pp. 504–511.

[28] "Linguistic data consortium, and nist multimodal information group, csr-ii (wsj1) complete ldc94s13a, linguisticdata consortium, philadelphia, 1994, web download.," .

[29] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[30] ITU-T, "Recommendation p.800: Methods for subjective determination of transmission quality," 1996.

[31] ITU-T, "Recommendation p.808: Subjective evaluation of speech quality with a crowdsourcing approach," 2018.

[32] B. Nadari and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *Proc. Interspeech*, 2020.