2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC) | 979-8-3503-6185-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IWAENC61483.2024.10694238

AN EFFECTIVE MVDR POST-PROCESSING METHOD FOR LOW-LATENCY CONVOLUTIVE BLIND SOURCE SEPARATION

Jiawen Chua[1], *Longfei Felix Yan*[2], *W. Bastiaan Kleijn*[2]

[1]Eigenspace GmbH, Germany
[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

In this paper, we introduce a post-processing method to minimize the algorithmic latency in traditional blind source separation (BSS) techniques. Our proposed approach involves the incorporation of a minimum variance distortion response method with the spatial covariance matrix, which is derived from conventional BSS methods, to effectively compute short demixing filters. The performance of source separation can be improved either by increasing the number of microphones or by integrating a dereverberation technique as a pre-processing step, or even both. The experimental results confirm the effectiveness and consistency of the proposed approaches on diverse speech databases.

Index Terms— Blind source separation, low-latency, independent vector analysis, beamforming, dereverberation

1. INTRODUCTION

Over the last few decades, blind source separation (BSS) techniques have evolved from separating instantaneous mixtures to more challenging reverberant mixtures [1–4]. Convolutional BSS aims to extract the original sources from recorded reverberant mixtures without prior knowledge of the mixing channels and sources themselves. This process poses a significant challenge, particularly when aiming for low latency. Traditional BSS methods utilize the short-time Fourier transform (STFT) technique to separate sources in the frequency domain. These methods assume that the convolutive process in the time domain can be approximated as an instantaneous multiplicative process in each frequency bin [5]. However, for this assumption to hold, the window length employed in the STFT should significantly exceed the length of the mixing filters [6], resulting in elevated algorithmic latency. One popular approach that follows this assumption is auxiliary independent vector analysis (AuxIVA) [7].

Pre-processing and post-processing approaches can be employed to reduce the algorithmic latency associated with traditional BSS methods. In the pre-processing, [8] proposed the utilization of a speech dereverberation technique known as weighted prediction error (WPE) [9]. This technique alleviates the impact of late reverberations while preserving spatial information, allowing shorter windows during the AuxIVA process. On the other hand, one post-processing approach utilizes the partitioned convolution method [10]. This method involves splitting the long estimated demixing filters into multiple shorter blocks, performing convolution in the frequency domain, and subsequently reconstructing the recovered signals using the overlap-add technique. An alternative post-processing approach [11, 12] involves incorporating additional microphones to calculate shorter demixing filters based on the long estimated ones. This helps minimize the crossband effect when using a shorter window. However, it is important to note that both of these approaches necessitate a high computational load.

Deep neural network methods have emerged as solutions to BSS problems [13–17]. These approaches typically rely on a training process to achieve successful separation, and the quality of the training dataset heavily influences the separation performance. Among these approaches, Conv-TasNet is a prominent approach in this field due to its low delay implementation, for instance, 2 or 4 ms. Recently, Beam-TasNet has been proposed, which combines the DNN approach with the minimum variance distortionless response (MVDR) method, a frequency-domain technique. This combination aims to further improve the separation performance of Conv-TasNet, particularly for reverberant mixtures. However, it comes with the drawback of requiring a 512 ms window length, which significantly increases the algorithmic latency.

In this paper, we present a novel method to reduce the algorithmic latency of the BSS system. Similarly to our previous work [11, 12], we first employ a traditional BSS technique like AuxIVA to estimate the demixing filters using a long window during the STFT processing. Instead of calculating the crossband filters for obtaining the short demixing filters, we compute the spatial covariance matrix (SCM) of the recovered sources. Next, we utilize the MVDR approach on the SCM to obtain short demixing filters for each source. These short demixing filters can then be directly utilized in the short window STFT domain, which is more computationally efficient compared to the partitioned convolution approach. Unlike Conv-TasNet or Beam-TasNet, the proposed technique does not require any pre-training or fine-tuning before separation. Additionally, the separation performance of our method can further be improved by employing the WPE approach as a pre-processing step to perform speech dereverberation.

The structure of the paper is as follows: In Section 2, we introduce our formulation of the BSS problem, which operates in the short STFT domain. In Section 3, we discuss the proposed method aimed at reducing the algorithmic delay of the BSS system. Moving on to Section 4, we present the experimental results of our proposed approach and compare them with other methods. Finally, in Section 5, we draw our conclusions.

2. PROBLEM FORMULATION

Let L denotes the number of sources and M denotes the number of microphones. When disregarding the noise, the observed signals of the mixture in a linear time-invariant (LTI) system can be expressed as:

$$x_m[n] = \sum_{l=0}^{L-1} \sum_{n'=-\infty}^{\infty} h_{ml}[n] s_l[n-n'], \tag{1}$$

where x_m represents the signal captured by the m^{th} microphone, s_l represents the signal emitted by the l^{th} source and h_{ml} represents the room impulse response (RIR) between the m^{th} microphone and the l^{th} source, acting as a mixing filter.

By performing the STFT with a short window length, we obtain the following expression [18]:

$$\mathbf{x}_{p,k} = \sum_{p'} \sum_{k'} \mathbf{A}_{p-p',k,k'} \mathbf{s}_{p',k'}, \tag{2}$$

where $\mathbf{x}_{p,k} \in \mathbb{C}^M$ and $\mathbf{s}_{p,k} \in \mathbb{C}^L$ represent the vector of microphones signals and source signals, respectively, at the p^{th} block index and the k^{th} frequency bin index in the STFT domain. The coefficient $\mathbf{A}_{p-p',k,k'} \in \mathbb{C}^{M\times L}$ indicates the crossband filter coefficients between frequency bands k and k' at the $(p-p')^{\text{th}}$ block index [18].

For the case where the window length is much longer than the length of the RIR, we can simplify the convolutive transfer function to a multiplicative transfer function [5, 6]:

$$\mathbf{x}_{p,k} \approx \mathbf{A}_k \mathbf{s}_{p,k}. \tag{3}$$

It is worth noting that the contribution of the crossband filter coefficients is significantly lower than that of the band-to-band filter coefficients.

To recover the separated signal, we can estimate a demixing matrix $\mathbf{W}_k \in \mathbb{C}^{M\times L}$ in each frequency bin using traditional BSS approaches, such as the AuxIVA approach [7]. By applying the estimated demixing matrix, the recovered sources $\mathbf{y}_{p,k} \in \mathbb{C}^L$ can be obtained as:

$$\mathbf{y}_{p,k} = \mathbf{W}_k^{\text{H}} \mathbf{x}_{p,k}, \tag{4}$$

where $\{\cdot\}^{\text{H}}$ denotes the Hermitian transpose operator.

In the following section, we discuss the process of calculating the demixing operator by utilizing the demixing matrix $\mathbf{W}_k$ in (4). This operator enables separation to be carried out directly within the short window STFT domain to reduce the algorithmic latency.

3. PROPOSED METHOD

Our primary focus is to decrease the algorithmic delay of the BSS methods, which is caused by the length of the window. If we directly use a short window for the STFT, the assumption of an instantaneous multiplicative process in the frequency domain becomes invalid. Consequently, the performance of the traditional BSS methods decreases. To address this issue, we first exploit the advantage of the longer window to estimate the demixing filters $\mathbf{W}_k$ in each frequency bin. This allows us to achieve satisfactory separation performance using the traditional BSS approaches. The long demixing filters $\mathbf{W}_k$ are then used to estimate parameters of an MVDR filter corresponding to the short filters. For real-time separation, only the short demixing filters are required. The estimation process can be performed in the background and updated regularly with the latest one.

Let $\tilde{\mathbf{x}}_{\tilde{p},\tilde{k}}$ denote the mixture in the short window STFT domain. The l^{th} estimated source can be obtained by using the short demixing filters of the l^{th} source, i.e.,

$$\tilde{y}^l_{\tilde{p},\tilde{k}} = \tilde{\mathbf{w}}^{l\,\text{H}}_{\tilde{k}} \tilde{\mathbf{x}}_{\tilde{p},\tilde{k}}. \tag{5}$$

The short demixing filters $\tilde{\mathbf{w}}^l_{\tilde{k}}$ can be computed using the MVDR schema [19, 20] if the steering vector and the SCM of the interfering sources are known. In our case, both of them can be estimated using the long demixing filters that were previously computed.

In this section, we first introduce the problem formulation using the MVDR approach. Next, we explore the estimation of the steering vector. Lastly, we discuss the computation of the SCM of the interfering sources.

3.1. MVDR formulation

From knowledge of the steering vector $\mathbf{u}^l_{\tilde{k}} \in \mathbb{C}^M$ and the SCM of the interfering sources $\mathbf{R}^{\backslash l}_{x,\tilde{k}} \in \mathbb{C}^{M\times M}$, the short demixing filters $\tilde{\mathbf{w}}^l_{\tilde{k}} \in \mathbb{C}^M$ of the desired source in (5) can be computed by minimizing the following expression:

$$\begin{aligned} \min_{\tilde{\mathbf{w}}^l_{\tilde{k}}} \quad & \tilde{\mathbf{w}}^{l\,\text{H}}_{\tilde{k}} \mathbf{R}^{\backslash l}_{x,\tilde{k}} \tilde{\mathbf{w}}^l_{\tilde{k}} \\ \text{s.t.} \quad & \tilde{\mathbf{w}}^{l\,\text{H}}_{\tilde{k}} \mathbf{u}^l_{\tilde{k}} = 1. \end{aligned} \tag{6}$$

The optimization solution can then be computed as

$$\tilde{\mathbf{w}}^l_{\tilde{k}} = \frac{\left(\mathbf{R}^{\backslash l}_{x,\tilde{k}}\right)^{-1} \mathbf{u}^l_{\tilde{k}}}{\mathbf{u}^{l\,\text{H}}_{\tilde{k}} \left(\mathbf{R}^{\backslash l}_{x,\tilde{k}}\right)^{-1} \mathbf{u}^l_{\tilde{k}}}, \tag{7}$$

which is the closed form solution for (6).

3.2. Estimation of the steering vector

Although (6) can be solved for any vector $\mathbf{u}^l_{\tilde{k}}$, it is important to select an appropriate one. In conventional beamforming, the steering vector is determined by the geometry of the microphone array and the direction of sources [21, 22]. For a BSS problem, the source direction information and the array geometry is unknown. However, we can obtain the signal subspace that is spanned by the images of the desired sources in the short window STFT domain by computing the SCM of the reconstructed observation signals $\mathbf{R}^l_{x,\tilde{k}} \in \mathbb{C}^{M\times M}$ that contains only the l^{th} source using the following expression:

$$\mathbf{R}^l_{x,\tilde{k}} = \sum_{\tilde{p}} \tilde{\mathbf{x}}^l_{\tilde{p},\tilde{k}} \tilde{\mathbf{x}}^{l\,\text{H}}_{\tilde{p},\tilde{k}}, \tag{8}$$

where

$$\tilde{\mathbf{x}}^l_{\tilde{p},\tilde{k}} = \underset{\text{short}}{\text{STFT}} \left\{ \underset{\text{long}}{\text{ISTFT}} \left\{ \mathbf{x}^l_{p,k} \right\} \right\}, \tag{9}$$

and $\tilde{\mathbf{x}}^l_{\tilde{p},\tilde{k}} \in \mathbb{C}^M$ represents the image of the desired source captured by the microphone array in the short window STFT domain, while $\mathbf{x}^l_{p,k} \in \mathbb{C}^M$ represents the image in the long window STFT domain. $\text{STFT}\{\cdot\}$ and $\text{ISTFT}\{\cdot\}$ denote the forward STFT and inverse STFT operations, respectively. The terms "long" and "short" refer to the use of a long window and a short window for these operations, respectively.

The image of the desired source $\mathbf{x}^l_{p,k}$ in the long window STFT domain can be estimated as:

$$\mathbf{x}^l_{p,k} = \hat{\mathbf{a}}^l_k y^l_{p,k}, \tag{10}$$

where $y^l_{p,k}$ is the l^{th} estimated source, which is also the l^{th} entry of $\mathbf{y}_{p,k}$ in (4), and $\hat{\mathbf{a}}^l_k \in \mathbb{C}^M$ is the l^{th} column vector of the estimated mixing matrix $\hat{\mathbf{A}}_k$ in the long window STFT domain. The estimated mixing matrix $\hat{\mathbf{A}}_k \in \mathbb{C}^{M\times L}$ can be computed by taking the pseudo-inverse of the demixing matrix, which is $\hat{\mathbf{A}}_k = \mathbf{W}^{\dagger}_k$.

Note that the signal subspace in the short window STFT domain is not solely spanned by a single mixing vector of the desired source. It is also spanned by other vectors that describe the crossband filters of the desired source.

To obtain the vector that describes most of the signal energy defined by the desired source, similarly to the method described in [20], we perform an eigenvalue decomposition on (8) and select the eigenvector associated with the maximum eigenvalue as our estimation of the steering vector $\mathbf{u}^l_{\tilde{k}}$. Instead of computing the time-frequency masks as done in [20], we employ the AuxIVA approach to obtain the images of the desired sources. This ensures that the spatial infor-

mation of the covariance matrix in (8) is well-preserved, as the same linear operation is applied for each time block once the computation of $\mathbf{W}_k$ has converged. We assume the sources remain stationary.

3.3. Estimation of the spatial covariance matrix

The SCM of the interfering sources can be computed as:

$$\mathbf{R}^{\setminus l}_{x,\tilde{k}} = \sum_{\tilde{p}} \tilde{\mathbf{x}}^{\setminus l}_{\tilde{p},\tilde{k}} \tilde{\mathbf{x}}^{\setminus l\ \mathrm{H}}_{\tilde{p},\tilde{k}}, \tag{11}$$

where $\tilde{\mathbf{x}^{\setminus l}}_{\tilde{p},\tilde{k}}$ is the expression of the reconstructed observation signals excluding the l^{th} source in the short window STFT domain. It can be obtained by

$$\tilde{\mathbf{x}^{\setminus l}}_{\tilde{p},\tilde{k}} = \underset{\text{short}}{\mathrm{STFT}} \left\{ \underset{\text{long}}{\mathrm{ISTFT}} \left\{ \mathbf{x}^{\setminus l}_{p,k} \right\} \right\}, \tag{12}$$

where $\mathbf{x}^{\setminus l}_{p,k}$ is the microphone signal in the long window STFT domain excluding the contribution of the l^{th} source:

$$\mathbf{x}^{\setminus l}_{p,k} = \mathbf{x}_{p,k} - \mathbf{x}^{l}_{p,k}. \tag{13}$$

The same estimation process is repeated L times using Eqs. (6) to (13) to obtain the demixing filters for all sources in the short window STFT domain. Finally, the short demixing filters in the separation process are updated with the latest computed $\tilde{\mathbf{w}}^{l}_{\tilde{k}}$, achieving separation performance similar to that of the long filters but with low algorithmic latency.

4. EXPERIMENTAL RESULTS

This section first describes the speech database and the algorithms that we compared. Next, we report the configurations of our proposed approach and the reference algorithms. Then we discuss the metrics used to measure the separation performance. Lastly, we present the experimental results obtained from various approaches in different setup.

4.1. Speech Database

Our experiment utilized two speech databases: spatialized WSJ0-2MIX [23] and LibriSpeech [24]. All speech signals were sampled at a rate of 8 kHz.

The spatialized WSJ0-2MIX dataset provided reverberant sources and convolutive mixtures using a script made available by [23]. For each trial, eight microphones and the sources were randomly located. The aperture size of the microphone array was randomly sampled from 15 to 25 cm. The distance between sources and the center of the microphone array was on average 1.3 m with 0.4 m standard deviation. The room dimensions were not fixed, and the reverberation time, T_{60}, was randomly chosen between 0.2 and 0.6 seconds. The signal-to-noise ratio (SNR) of the source varied between -5 and 5 dB.

As for the LibriSpeech database, reverberant sources and convolutive mixtures were generated during the evaluation stage. Pyroomacoustics [25] was employed as the room impulse response (RIR) generator. For each trial, the T_{60} was randomly selected between 0.1 and 0.3 seconds. Additionally, the microphones and sources were randomly positioned within a room of fixed dimensions (5 x 3 x 3 m). The array aperture size and the source-to-array distance were not constrained.

4.2. Configurations and Reference Algorithms

We conducted a comparison of various algorithms, including a multi-channel version of Conv-TasNet [15], Beam-TasNet [13, 14], two variations of AuxIVA [7], denoted as $\mathrm{AuxIVA_L}$ and $\mathrm{AuxIVA_S}$. These variations indicate the utilization of a long window and a short window, respectively. Additionally, we examined the partitioned convolution method [10], denoted as PConv, and finally our proposed approach.

Both MC-Conv-TasNet and Beam-TasNet were implemented using the publicly available code from the open-source repository [1] provided by [14]. Detailed network configurations can be found in [14] under the label "Baseline". The causal variant of Beam-TasNet was employed, and a window length of 512 ms was used for the beamformer. The refinement technique based on the voice activity detection described in [13] was not utilized in this implementation. For MC-Conv-TasNet, the first output channel of the parallel decoder was selected as the estimated source.

The implementation of AuxIVA utilized the open-source library "libss[2]," which was provided by [7]. A principal component analysis was performed if the microphone number is larger than the number of source. In the case of $\mathrm{AuxIVA_L}$, a Hamming window with a length of 64 ms and a hop size of 16 ms was applied. For $\mathrm{AuxIVA_S}$, the window length was 16 ms, and the hop size was set to 4 ms. As described in [10], only the causal components of the estimated demixing filters of $\mathrm{AuxIVA_L}$ were used for PConv to avoid the artifacts caused by circular convolution.

For the proposed approach, the estimated demixing filters of $\mathrm{AuxIVA_L}$ were employed in the long window STFT domain. In the case of the short window, a Hamming window with a length of 16 ms and a hop size of 4 ms was used.

The algorithmic latencies for MC-Conv-TasNet, Beam-TasNet, $\mathrm{AuxIVA_L}$, $\mathrm{AuxIVA_S}$, PConv and our proposed approach were 2 ms, 512 ms, 64 ms, 16 ms, 16 ms, 16 ms, respectively. Moreover, we also evaluated a variation that utilized the WPE method [26] as a pre-processing stage to reduce the reverberation of the mixture for all these methods. The WPE algorithm was configured with a delay of 3, 5 iterations, 10 taps, and alpha set to 0.9999. To ensure that no additional latency was introduced by the pre-processing stage, the same short window setup was employed for the WPE algorithm.

4.3. Metrics

To assess the separation performance, we computed the scale-invariant signal-to-distortion ratio improvement (SI-SDRi) [27] between the estimated source and the original source. This metric was widely utilized in the speech separation challenge involving the WSJ0 database [15–17]. As our focus is on the separation performance, the signal-to-interference ratio (SIR) [28] is primarily used. Besides, we also reported the perceptual evaluation of subjective quality (PESQ) [29] to indicate the quality of the separated speech. All these metrics were computed using the Asteroid toolkit [30]

4.4. Separation Performance for Fixed Microphone Number

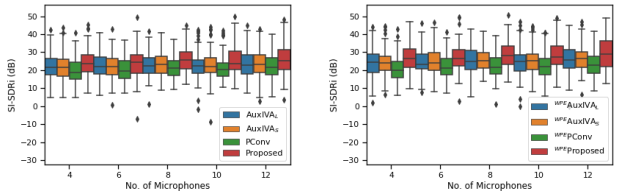
In this experiment, we conducted 100 trials for both spatialized WSJ0-2MIX and LibriSpeech databases and we report the mean values. For Beam-TasNet, the original setup designed to handle mixtures with four channels was utilized, using four microphones. Both MC-Conv-TasNet and Beam-TasNet were trained with the spatialized WSJ0-2MIX dataset and the final models have achieved the separation performance as described in [14].

[1] Available at https://github.com/hangtingchen/Beam-Guided-TasNet

[2] Available at https://github.com/onolab-tmu/libss/blob/main/libss

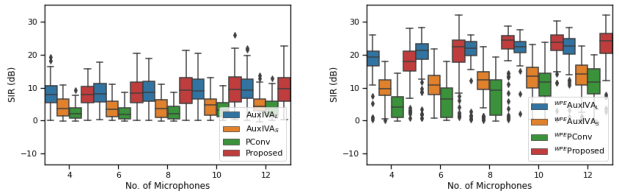
Table 1: Separation performance between various approaches with spatialized WSJ0-2MIX and LibriSpeech

Method	Spatialized WSJ0-2MIX			LibriSpeech		
	SI-SDRi (dB)	SIR (dB)	PESQ	SI-SDRi (dB)	SIR (dB)	PESQ
Beam-TasNet	23.22	16.81	2.05	20.99	4.35	1.55
(with WPE)	**24.32**	**19.12**	2.32	22.55	6.03	1.65
MC-Conv-TasNet	22.91	14.48	1.97	21.03	2.97	1.48
(with WPE)	24.12	17.17	2.27	22.66	4.64	1.53
AuxIVA_L	18.91	6.72	1.88	22.10	8.21	1.97
(with WPE)	19.84	14.64	2.35	24.13	**18.37**	2.53
AuxIVA_S	19.78	5.0	1.86	22.11	3.99	1.75
(with WPE)	20.55	11.33	2.31	24.42	9.59	2.07
PConv	14.45	1.48	1.53	19.76	2.47	1.55
(with WPE)	15.68	3.65	1.66	20.77	4.50	1.64
Proposed	21.80	7.64	1.97	23.85	7.75	2.00
(with WPE)	23.41	15.36	**2.51**	**26.69**	17.24	**2.62**



(a) Without WPE (b) With WPE

Fig. 1: Comparison of the SI-SDRi metric between various approaches using different number of microphone.



(a) Without WPE (b) With WPE

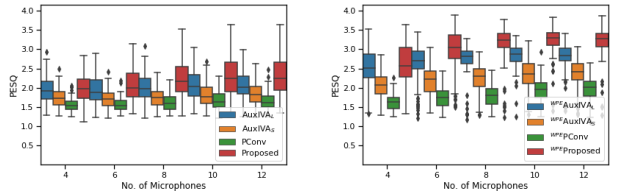
Fig. 2: Comparison of the SIR metric between various approaches using different number of microphone.

The performance of the various approaches is shown in Table 1. Beam-TasNet and MC-Conv-TasNet performed well when evaluated using spatialized WSJ0-2MIX. However, the separation performance was significantly influenced by altering the speech database, as evidenced by the SIR metric during evaluation with LibriSpeech.

The techniques based on AuxIVA demonstrate consistent performance in both speech databases. As anticipated, PConv performs the poorest due to its omission of the filter coefficients, which adversely affects the separation performance. The separation performance of AuxIVA_S is inferior to that of AuxIVA_L because the assumption of a multiplicative process is not valid when using a short window. The proposed approach outperforms the other methods across almost all metrics. Furthermore, the findings show that incorporating the WPE method as a pre-processing step enhances the separation performance of all the methods.

4.5. Separation Performance for Different Microphone Number

In this particular experimental setup, a varying number of microphones were used. Beam-TasNet and MC-Conv-TasNet were omitted as the original configuration allows the models to work exclusively with four microphones. LibriSpeech was chosen due to its flexibility in accommodating more than eight microphones, unlike spatialized WSJ0-2MIX, which is restricted to only eight microphones. We conducted 100 trials for each microphone number setup.



(a) Without WPE (b) With WPE

Fig. 3: Comparison of the PESQ metric between various approaches using different number of microphone.

Fig. 1 illustrates the SI-SDRi metric which indicates the improvement in performance achieved by applying separation to the mixtures. In contrast, Fig. 2 demonstrates the separation performance solely based on the SIR metric. Lastly, Fig 3 depicts the PESQ metric, which assesses the quality of the separated speech.

The experimental results show that the proposed method surpasses all others in all metrics, even under varying circumstances. The performance of the proposed method is further improved with a greater number of microphones. This is due to the increased degree-of-freedom to eliminate the disturbance caused by interfering sources. Furthermore, utilizing the WPE method leads to a substantial enhancement in the separation performance.

5. CONCLUSION

This paper introduces an innovative technique that combines traditional BSS methods with an MVDR architecture. Our approach solves the BSS problem effectively with minimal algorithmic delay. The proposed method computes short demixing filters based on a demixing matrix estimated in the long-window STFT domain, where the assumption of a multiplicative process holds true. These short demixing filters can be directly implemented in the short-window STFT domain. The approach attains excellent separation performance for both the spatialized WSJ0-2MIX and LibriSpeech databases. The performance can be further enhanced by increasing the number of microphones and integrating the WPE method.

6. REFERENCES

[1] S.-I. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," *Advances in neural information processing systems*, vol. 8, 1995.

[2] L. Yan, W. B. Kleijn, and T. Abhayapala, "A linear-time independence criterion based on a finite basis approximation," in *International Conference on Artificial Intelligence and Statistics*, pp. 202–212, PMLR, 2020.

[3] H. Sun, P. Samarasinghe, and T. Abhayapala, "Blind source counting and separation with relative harmonic coefficients," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

[4] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE transactions on speech and audio processing*, vol. 13, no. 1, pp. 120–134, 2004.

[5] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 114–126, 2007.

[6] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 55–69, 1980.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 189–192, IEEE, 2011.

[8] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 506–510, IEEE, 2021.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 85–88, IEEE, 2008.

[10] Y. Kuriki, T. Nakashima, K. Yamaoka, N. Ueno, Y. Wakabayashi, N. Ono, and R. Sato, "Efficient low-latency convolution with uniform filter partition and its evaluation on real-time blind source separation," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 765–769, IEEE, 2022.

[11] J. Chua, G. Wang, and W. B. Kleijn, "Convolutive blind source separation with low latency," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2016.

[12] J. Chua and W. B. Kleijn, "A low latency approach for blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1280–1294, 2019.

[13] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6384–6388, IEEE, 2020.

[14] H. Chen, Y. Yi, D. Feng, and P. Zhang, "Beam-guided TasNet: An iterative speech separation framework with multi-channel output," *arXiv preprint arXiv:2102.02998*, 2021.

[15] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, IEEE, 2021.

[17] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

[18] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE transactions on Audio, Speech, and Language processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[19] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[20] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[21] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2002.

[22] L. Yan, W. Huang, W. B. Kleijn, and T. D. Abhayapala, "Neural optimization of geometry and fixed beamformer for linear microphone arrays," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[23] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2018.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[25] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355, IEEE, 2018.

[26] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, pp. 1–5, VDE, 2018.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, IEEE, 2019.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752, IEEE, 2001.

[30] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, *et al.*, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.