NCSP'25 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2025
February 27 - March 2, 2025, Pulau Pinang, Malaysia

Moving Blind Source Extraction Based on Constant Separating Vector and Switching Mechanism

Yuehao Zhao, Tetsuya Ueda, and Shoji Makino

Waseda University
E-mail: {zhaoyuehao@akane, t.ueda@akane, s.makino@}waseda.jp

Abstract

This paper proposes a novel method for moving source extraction using a limited number of microphones. In recent years, the constant separating vector (CSV) mixing model has been devised for auxiliary function-based independent vector extraction (AuxIVE) to extract a moving signal stably. However, the extraction performance of AuxIVE based on the CSV mixing model (CSV-AuxIVE) performance significantly decreases with fewer microphones, limiting practical applications. To this end, a switching mechanism is adopted to create multiple extraction filters in each batch for varying spatial positions, even with a few microphones. Experimental results demonstrate that the proposed switching CSV-AuxIVE improves iSDR and iSIR compared to CSV-AuxIVE, particularly in microphone-limited settings.

1. Introduction

Blind source separation (BSS) is an algorithm for separating independent source signals from mixed signals without any prior information. Blind signal extraction (BSE) is a special form of BSS that extracts a specific source of interest (SOI) from mixed signals. With the development of speech applications, BSE has been widely used in automatic speech recognition, teleconferencing and hearing aid devices, etc.

As a traditional BSS method, independent vector analysis (IVA) [1] [2] separates signals by maximizing statistical independence and it also uses a joint statistical source model assuming dependence between different frequency components. Building on this, auxiliary function-based IVA (AuxIVA) [3] has been proposed, which stabilizes and accelerates the convergence of parameter optimization for BSS. As a special from of AuxIVA, auxiliary function-based independent vector extraction (AuxIVE) [4] skips most of the calculations for parameter optimization when dealing with BSE. The methods above are limited to time-invariant scenarios where all sounds including SOI are fixed.

To address time-varying scenarios, the constant separating vector (CSV) mixing model [5] enables moving source extraction by imposing spatial constraints to cover the entire area of the SOI movement during the recording, as shown in Fig. 1 (a). Recently, an algorithm has been proposed based on AuxIVE and CSV mixing model, which is named CSV-AuxIVE [6]. It can maintain extraction performance for dynamic sources. However, CSV-AuxIVE requires a large number of microphones to impose spatial constraints to cover the moving area; it cannot impose a lot of constraints if we have a small number of microphones, as shown in Fig. 1 (b).

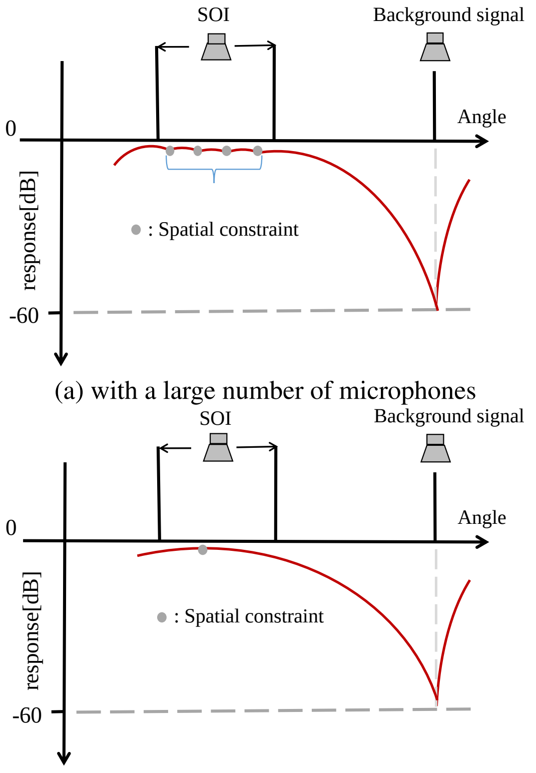
This paper proposes a novel method to extract the moving SOI even with a small number of microphones. Specifically, we incorporate a switching mechanism [7] into CSV-AuxIVE. Switching mechanism has been proposed to improve speech extraction performance when the number of sources exceeds the number of microphones. Using the switching mechanism, our proposed method generates and dynamically applies multiple filters to specific SOI movement ranges. We refer to the proposed method as Switching CSV-AuxIVE (Sw-CSV-AuxIVE). Unlike [7] which has used this mechanism to suppress interference and noise signals, Sw-CSV-AuxIVE focuses on expanding the coverage of SOI movement by the switching mechanism, ensuring robust performance with limited microphones as shown in Fig. 2.

2. Problem formulation

We assume there are M microphones to extract an SOI. After short-time Fourier transform (STFT), we obtain observed microphone signals $\boldsymbol{x}_{f,l} = [x_{1,f,l}, \ldots, x_{M,f,l}]^{\mathsf{T}} \in \mathbb{C}^M$, the SOI $s_{f,l} \in \mathbb{C}$, and the background signals $\boldsymbol{z}_{f,l} \in \mathbb{C}^{M-1}$ at each time frame $l = 1, \ldots, L$ and frequency bin $f = 1, \ldots, F$. In addition, let all frames be divided into $T \geq 1$ time intervals called blocks, and each block includes L_{b} frames for simplicity, hence time frames $L = TL_{\mathrm{b}}$. Hereafter, we treat the frame index $\{(t-1)L_{\mathrm{b}} + 1, ..., tL_{\mathrm{b}}\}$ as the same block index t for $t = 1, \ldots, T$. For example, we denote $\boldsymbol{x}_{f,(t-1)L_{\mathrm{b}}+l'}$ as $\boldsymbol{x}_{f,t,l'}$ for $t = 1, \ldots, T$ and $l' = 1, \ldots, L_{\mathrm{b}}$.

As a main contribution of this paper, we combine the CSV mixing model [8] and the switching mechanism [7]. Figure 3 shows the processing flow that yields SOI $s_{f,t,l'}$. We write the relation between $s_{f,t,l'}$, $\boldsymbol{z}_{f,t,l'}$, and $\boldsymbol{x}_{f,t,l'}$ in a semi-time-varying model as:

$$\boldsymbol{x}_{f,t,l'} = \sum_{j=1}^{J} \delta_{j,f,t} \boldsymbol{A}_{j,f,t} \underbrace{\begin{bmatrix} s_{f,t,l'} \\ \boldsymbol{z}_{f,t,l'} \end{bmatrix}}_{\boldsymbol{s}_{f,t,l'}}, \tag{1}$$



(a) with a large number of microphones

(b) with a small number of microphones (e.g., 2)

Figure 1: Example of directional response in conventional CSV-AuxIVE [6]

where $\boldsymbol{A}_{j,f,t}$ is a mixing matrix parameterized as:

$$\begin{aligned}\boldsymbol{A}_{j,f,t} &= [\boldsymbol{a}_{f,t}\ \boldsymbol{Q}_{j,f,t}]\\ &= \begin{bmatrix}\gamma_{f,t} & \boldsymbol{h}^{\mathsf{H}}_{j,f}\\ \boldsymbol{g}_{f,t} & \frac{1}{\gamma_{f,t}}(\boldsymbol{g}_{f,t}\boldsymbol{h}^{\mathsf{H}}_{j,f} - (\beta^*_{j,f}\gamma_{f,t} + \boldsymbol{h}^{\mathsf{H}}_{j,f}\boldsymbol{g}_{f,t})\boldsymbol{I}_{M-1})\end{bmatrix}.\end{aligned} \tag{2}$$

$(\cdot)^{\mathsf{H}}$ represents the Hermitian transpose. $\delta_{j,f,t}$ is a switching weight which selects block-invariant parameters $\boldsymbol{h}_{j,f}$ and $\beta_{j,f}$ in different time block t. Weight $\delta_{j,f,t}$ takes a binary value so that $\sum_{j=1}^{J}\delta_{j,f,t} = 1$ and $\delta_{j,f,t} \in \{0,1\}$. Hereafter, j and J are referred to as an index of a switching state and the total number of states, respectively. Similarly, the separation model can be written as:

$$\boldsymbol{s}_{f,t,l'} = \sum_{j=1}^{J}\delta_{j,f,t}\boldsymbol{W}^{\mathsf{H}}_{j,f,t}\boldsymbol{x}_{f,t,l'}, \tag{3}$$

where $\boldsymbol{W}_{j,f,t} = \boldsymbol{A}^{-\mathsf{H}}_{j,f,t}$ is a separation matrix parameterized as:

$$\boldsymbol{W}_{j,f,t} = [\boldsymbol{w}_{j,f}\ \boldsymbol{B}_{f,t}] = \begin{bmatrix}\beta_{j,f} & \boldsymbol{g}^{\mathsf{H}}_{f,t}\\ \boldsymbol{h}_{j,f} & -\gamma^*_{f,t}\boldsymbol{I}_{M-1}\end{bmatrix}. \tag{4}$$

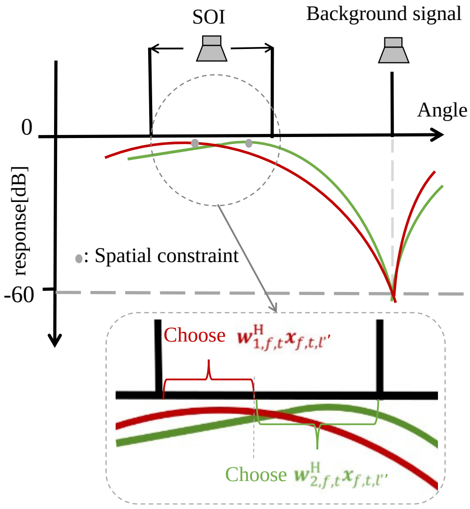


Figure 2: Example of directional response in Sw-CSV-AuxIVE with a small number of microphones (e.g., 2).

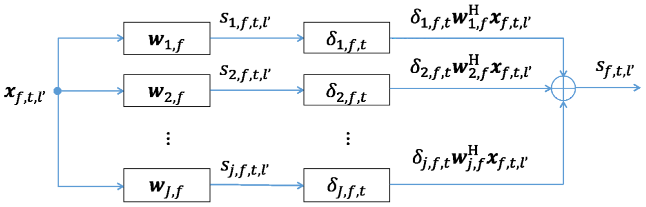

$\boldsymbol{x}_{f,t,l'}$
$\boldsymbol{w}_{1,f}$
$\boldsymbol{w}_{2,f}$
⋮
$\boldsymbol{w}_{J,f}$
$s_{1,f,t,l'}$
$s_{2,f,t,l'}$
$s_{J,f,t,l'}$
$\delta_{1,f,t}$
$\delta_{2,f,t}$
⋮
$\delta_{J,f,t}$
$\delta_{1,f,t}\boldsymbol{w}^{\mathsf{H}}_{1,f}\boldsymbol{x}_{f,t,l'}$
$\delta_{2,f,t}\boldsymbol{w}^{\mathsf{H}}_{2,f}\boldsymbol{x}_{f,t,l'}$
$\delta_{J,f,t}\boldsymbol{w}^{\mathsf{H}}_{J,f}\boldsymbol{x}_{f,t,l'}$
$s_{f,t,l'}$

Figure 3: Flows of obtaining SOI $s_{f,t,l'}$

The time-invariant separation filter $\boldsymbol{w}_{j,f}$ in (4) enables us to extract one source stably. It is possible to parametrize $\boldsymbol{W}_{j,f,t}$ and $\boldsymbol{A}_{j,f,t}$ by assuming a distortion-less constraint $\sum_{j=1}^{J}\delta_{j,f,t}\boldsymbol{w}^{\mathsf{H}}_{j,f}\boldsymbol{a}_{f,t} = 1$. The mixing model in (1) where $J = 1$ corresponds to the CSV mixing model [6], and the separation model in (3) where $T = 1$ corresponds to that with the switching mechanism [7].

Next, we assume a probabilistic model. According to (3), we can write negative log probability of $\boldsymbol{x}_{f,t,l'}$:

$$\begin{aligned}&p(\{\boldsymbol{x}_{f,t,l'}\}_{f,t,l'})\\ &= \prod_{f,t,l'} p(\boldsymbol{s}_{f,t,l'})\left|\det\sum_{j=1}^{J}\delta_{j,f,t}\boldsymbol{W}_{j,f,t}\right|^2.\end{aligned} \tag{5}$$

Then, we assume each component of $\boldsymbol{s}_{f,t,l'}$ (i.e., $s_{f,t,l'}$ and $\boldsymbol{z}_{f,t,l'}$) are mutually independent over all times and frequencies. So their joint probability distribution function (pdf) is equal to the product of marginal pdfs as $p(\boldsymbol{s}_{f,t,l'}) = p(s_{f,t,l'})p(\boldsymbol{z}_{f,t,l'})$, where $p(\boldsymbol{z}_{f,t,l'})$ denote the pdf of $\boldsymbol{z}_{f,t,l'}$, respectively. We set $p(s_{f,t,l'})$ as the following pdf to reflect the block-dependent variance:

$$p(s_{f,t,l'}) = g\left(\frac{s_{f,t,l'}}{\hat{\sigma}_{f,t}}\right)\hat{\sigma}^{-2}_{f,t}, \tag{6}$$

where $\hat{\sigma}_{f,t} = \sum_{j=1}^{J} \delta_{j,f,t}\sqrt{\boldsymbol{w}_{j,f}^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{w}_{j,f}}$ is a frame-based variance of $s_{f,t,l'}$ and $\boldsymbol{C}_{f,t} = \mathbb{E}\left[\boldsymbol{x}_{f,t,l'}\boldsymbol{x}_{f,t,l'}^{\mathsf{H}}\right] = \frac{1}{L_{\mathrm{b}}}\sum_{l'=1}^{L_{\mathrm{b}}}\boldsymbol{x}_{f,t,l'}\boldsymbol{x}_{f,t,l'}^{\mathsf{H}}$ is a frame-based covariance matrix of $\boldsymbol{x}_{f,t,l'}$. We use a time-varying Gaussian distribution to normalized non-Gaussian random variable $g(\cdot)$:

$$g\left(\frac{s_{f,t,l'}}{\hat{\sigma}_{f,t}}\right) \propto \frac{1}{r_{f,t,l'}} \exp\left(-\frac{|s_{f,t,l'}|^2}{r_{f,t,l'}\hat{\sigma}_{f,t}^2}\right), \tag{7}$$

where $\propto$ denotes the proportionality symbol. The pdf of the background is assumed to be circular Gaussian with zero mean and covariance matrix $\boldsymbol{\Omega}_{f,t} = \mathbb{E}[\boldsymbol{z}_{f,t,l'}\boldsymbol{z}_{f,t,l'}^{\mathsf{H}}]$:

$$p(\boldsymbol{z}_{f,t,l'}) \propto \frac{1}{\det \boldsymbol{\Omega}_{f,t}} \exp(-\boldsymbol{z}_{f,t,l'}^{\mathsf{H}}\boldsymbol{\Omega}_{f,t}^{-1}\boldsymbol{z}_{f,t,l'}). \tag{8}$$

Because $\delta_{j,f,t}$ takes 1 for a state j,

$$\begin{aligned}\left|\det \sum_{j=1}^{J} \delta_{j,f,t}\boldsymbol{W}_{j,f,t}\right| &= \sum_{j=1}^{J} \delta_{j,f,t} \log|\det \boldsymbol{W}_{j,f,t}| \\ &= \sum_{j=1}^{J} \delta_{j,f,t}|\gamma_{f,t}|^{(M-2)}|\boldsymbol{w}_{j,f}^{\mathsf{H}}\boldsymbol{a}_{f,t}|.\end{aligned} \tag{9}$$

From the above assumptions, we can obtain the negative log-likelihood of the given signal $\mathcal{X} = \{\boldsymbol{x}_{f,t,l'}\}_{f,t,l'}$:

$$-\log p(\mathcal{X}) \propto \sum_{j=1}^{J}\sum_{f=1}^{F}\sum_{t=1}^{T} \frac{\delta_{j,f,t}}{T_{j,f}} \mathcal{L}(\theta_{j,f,t}), \tag{10}$$

$$\begin{aligned}\mathcal{L}(\theta_{j,f,t}) = \mathbb{E}&\left[\frac{|\boldsymbol{w}_{f,t}^{\mathsf{H}}\boldsymbol{x}_{f,t,l'}|^2}{r_{f,t,l'}\hat{\sigma}_{f,t}^2}\right] + \log r_{f,t,l'} + \log \hat{\sigma}_{f,t}^2 \\ &+ \mathbb{E}\left[\boldsymbol{z}_{f,t,l'}^{\mathsf{H}}\boldsymbol{\Omega}_{f,t}^{-1}\boldsymbol{z}_{f,t,l'}\right] - \log|\gamma_{f,t}|^{2(M-2)} \\ &- \log|\boldsymbol{w}_{j,f}^{\mathsf{H}}\boldsymbol{a}_{f,t}|,\end{aligned} \tag{11}$$

where $\Theta = \{\mathcal{W}, \mathcal{R}, \mathcal{D}\}$, $\mathcal{W} = \{\{\boldsymbol{w}_{j,f}\}_{j,f}, \{\boldsymbol{a}_{f,t}\}_{f,t}\}$, $\mathcal{R} = \{r_{f,t,l'}\}_{f,t,l'}$, $\mathcal{D} = \{\delta_{j,f,t}\}_{j,f,t}$, and $\theta_{j,f,t} = (\boldsymbol{w}_{j,f}, \boldsymbol{a}_{f,t}, r_{f,t,l'}, \delta_{j,f,t})$. $\overset{c}{=}$ denotes equality up to the constant terms. $T_{j,f} = \sum_{t=1}^{T} \delta_{j,f,t}$.

3. Optimization Process

We use a coordinate descent method to reduce the cost function in (11) by repeatedly updating each $\mathcal{W}$, $\mathcal{D}$, and $\mathcal{R}$ one by one.

We use orthogonal constraint [8] that SOI $s_{f,t,l'}$ has zero sample correlation with the noise signal $\boldsymbol{z}_{f,t,l'}$, i.e., $\sum_{j=1}^{J} \delta_{j,f,t}\boldsymbol{w}_{j,f}^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{B}_{f,t} = \boldsymbol{0}_{(M-1)}^{\mathsf{T}}$, where $\boldsymbol{0}_M \in \mathbb{R}^M$ is zero vector. Under the distortionless constraint and the orthogonal constraint, we can estimate t-th mixing vector $\boldsymbol{a}_{f,t}$:

$$\boldsymbol{a}_{f,t} = \sum_{j=1}^{J} \delta_{j,f,t} \frac{\boldsymbol{C}_{f,t}\boldsymbol{w}_{j,f}}{\boldsymbol{w}_{j,f}^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{w}_{j,f}}. \tag{12}$$

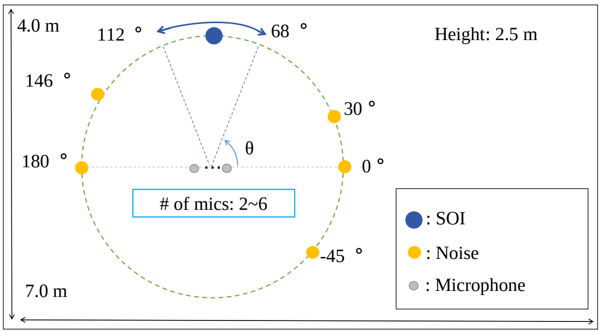


Figure 4: Configurations of sources and microphones

After updating $\boldsymbol{a}_{f,t}$, we update $\boldsymbol{w}_{j,f}$ in each switching state j based on conventional technique [8]:

$$\boldsymbol{w}_{j,f} \leftarrow \left(\sum_{t=1}^{T} \delta_{j,f,t}\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1} \sum_{t=1}^{T} \delta_{j,f,t}\frac{\boldsymbol{w}_{j,f}^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_{j,f}}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}, \tag{13}$$

and $\boldsymbol{V}_{f,t}$ is a covariance matrix:

$$\boldsymbol{V}_{f,t} = \mathbb{E}\left[\frac{\boldsymbol{x}_{f,t,l'}\boldsymbol{x}_{f,t,l'}}{r_{t,l'}}\right], \tag{14}$$

where $r_{t,l'} \leftarrow \frac{1}{F}\sum_{f} r_{f,t,l'}$. Note that this paper adopts a frequency-independent source model only when updating separation matrices $\mathcal{W}$, following a previously proposed practical technique [9].

We update switching weights $\mathcal{D}$ by setting $\delta_{j,f,t} = 1$ for a state j that gives the minimum cost function in (11) among all states at each time frequency:

$$\delta_{j,f,t} \leftarrow \begin{cases} 1 & \text{if } j = \underset{j'}{\operatorname{argmin}}\ \mathcal{L}(\theta_{j',f,t}) \\ 0 & \text{otherwise} \end{cases}. \tag{15}$$

After updating $s_{f,t,l'}$ using (3) and (4), we update the variance $r_{f,t,l'}$ by weighting the power of $s_{f,t,l'}$ by the inverse of block-dependent variance $\hat{\sigma}_{f,t}^2$:

$$r_{f,t,l'} \leftarrow \left|\frac{s_{f,t,l'}}{\hat{\sigma}_{f,t}}\right|^2. \tag{16}$$

4. Experimental evaluation

Table 1: Results of comparison of methods with different numbers of microphones

# of mics.	Method	iSDR [dB]	iSIR [dB]
6	CSV-AuxIVE [6]	2.72	7.06
	Sw-CSV-AuxIVE	2.71	7.93
5	CSV-AuxIVE [6]	2.61	6.82
	Sw-CSV-AuxIVE	2.94	9.24
4	CSV-AuxIVE [6]	2.39	6.27
	Sw-CSV-AuxIVE	2.69	7.17
3	CSV-AuxIVE [6]	2.15	5.65
	Sw-CSV-AuxIVE	2.65	6.47
2	CSV-AuxIVE [6]	0.70	0.85
	Sw-CSV-AuxIVE	1.89	0.89

4.1 Experimental condition

We compared the results of the proposed Sw-CSV-AuxIVE with the existing CSV-Aux-IVE [6]. We obtained the observed signals by mixing one moving SOI and 5 point noises using *signal generator*[1], and generated 10 groups of source signals. We randomly utilized one point-source speech signal from the TIMIT corpus test set [10] to obtain SOI. The number of microphones was reduced to 6, 5, 4, 3, and 2 in that order. $L_b = 100$ frames and we set the number of block $T = 16$. The layout of the experimental environment is shown in Fig. 4.

The speech extraction performance was measured by the average value of improvement of signal-to-distortion ratio (iSDR) and signal-to-interferences ratio (iSIR) [11].

4.2 Result

Our results compare the iSDR between CSV-AuxIVE and the proposed Sw-CSV-AuxIVE using Table 1. The proposed Sw-CSV-AuxIVE consistently outperforms the baseline CSV-AuxIVE in terms of iSIR across all configurations. Although the iSDR of the baseline methods and our proposed Sw-CSV-AuxIVE method shows little difference when the number of microphones is 6, as the number of microphones decreases, the iSDR of Sw-CSV-AuxIVE becomes increasingly significant. Notably, under the condition of having only 2 microphones, Sw-CSV-AuxIVE has about 1 dB higher iSDR than CSV-AuxIVE.

5. Conclusions

This paper proposed Sw-CSV-AuxIVE to enhance BSE of moving sound sources, especially in situations with fewer microphones. Our proposed method was invented by incorporating the switching mechanism to generate multiple separation filters for different spatial positions. Experimental results demonstrated that Sw-CSV-AuxIVE consistently outperforms CSV-AuxIVE.

References

[1] T. Kim, T. Eltoft, and T.-W. Lee, “Independent vector analysis: An extension of ica to multivariate components,” in *Proc. ICA*, Springer, 2006, pp. 165–172.

[2] A. Hiroe, “Solution of permutation problem in frequency domain ICA, using multivariate probability density functions,” in *Proc. ICA*, 2006, pp. 601–608.

[3] N. Ono and S. Miyabe, “Auxiliary-function-based independent component analysis for super-gaussian sources,” in *Proc. LVA/ICA*, 2010, pp. 165–172.

[4] R. Ikeshita, T. Nakatani, and S. Araki, “Block coordinate descent algorithms for auxiliary-function-based independent vector extraction,” *IEEE Trans. SP*, vol. 69, pp. 3252–3267, 2021.

[5] Z. Koldovskỳ, V. Kautskỳ, P. Tichavskỳ, J. Čmejla, and J. Málek, “Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers,” *IEEE Trans. SP*, vol. 69, pp. 2158–2173, 2021.

[6] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, “Auxiliary function-based algorithm for blind extraction of a moving speaker,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–16, 2022.

[7] K. Yamaoka, N. Ono, and S. Makino, “Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement,” *IEEE/ACM Trans. ASLP*, vol. 29, pp. 3461–3475, 2021.

[8] Z. Koldovsky and P. Tichavsky, “Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence,” *IEEE Trans. SP*, vol. 67, no. 4, pp. 1050–1064, 2018.

[9] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, “Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms,” *IEEE/ACM Trans. ASLP*, vol. 30, pp. 1032–1047, 2022.

[10] J. Garofolo, L. Lamel, W. Fisher, *et al.*, “Timit acoustic-phonetic continuous speech corpus LDC93S1. web download. philadelphia: Linguistic data consortium,” in 1993.

[11] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE/ACM Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[1] https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator