SPEECH ENHANCEMENT WITH A LOW-COMPLEXITY ONLINE SOURCE NUMBER ESTIMATOR USING DISTRIBUTED ARRAYS

Maja Taseska, Affan Hasan Khan, and Emanuël A. P. Habets

International Audio Laboratories Erlangen[*]
Am Wolfsmantel 33, 91058 Erlangen, Germany
{maja.taseska, emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

Enhancement of a desired speech signal in the presence of background noise and interferers is required in various modern communication systems. Existing multichannel techniques often require that the number of sources and their locations are known in advance, which makes them inapplicable in many practical situations. We propose a framework which uses the microphones of distributed arrays to enhance a desired speech signal by reducing background noise and an initially unknown number of interferers. The desired signal is extracted by a minimum variance distortionless response filter in dynamic scenarios where the number of active interferers is time-varying. An efficient, geometry-based approach that estimates the number of active interferers and their locations online is proposed. The overall performance is compared to the one of a geometry-based probabilistic framework for source extraction, recently proposed by the authors.

Index Terms— Source extraction, PSD matrix estimation, distributed arrays, number of sources

1. INTRODUCTION

Modern communication systems involve hands-free acquisition of desired speech in a variety of applications, such as smart homes, hands-free telephony, etc. The signals captured by the available microphones are often corrupted by background noise and interfering speech. If the second order statistics (SOS) of the desired and undesired signals are known, the desired signal can be estimated by applying linear spatial filters to the microphone signals. The SOS of a particular signal can be estimated from the microphone signals by temporal averaging, during periods where the particular signal is dominant. Recently, spatial cues have been used to detect the dominant signal in each time-frequency (TF) bin and estimate the corresponding SOS [1–3]. If multiple speech interferers are present, the SOS of each interfering signal needs to be estimated.

The detection of the number of speech sources in reverberant environments has been recently adressed in [3–7], where spatial information for each TF-bin extracted using microphone arrays is employed. On the one hand, the authors in [3–5] use probabilistic mixture models, where the number of sources represents a model parameter to be estimated. These algorithms are used in a batch mode [4, 5], for instance, prior to a blind source separation algorithm, or require a training phase to estimate the model parameters [3]. Hence, they are not suited for real-time applications with varying number of sources. On the other hand, the authors in [6, 7] use histograms of bin-wise direction of arrival (DOA) estimates to count and localize the active sources. While the algorithm in [6] is based on histograms of the complete observation, the algorithm in [7] is based on short-time histograms, and hence suited for a real-time tracking of the number of speech sources.

In this work, we propose a non-probabilistic approach for localization and number of source estimation, applied to extract a desired signal in the presence of varying number of interferers. Similarly to [7], our proposed approach estimates the number of sources online, at each time frame. Based on bin-wise positions obtained by triangulating the DOA estimates of multiple distributed arrays, the number of sources is monitored, new sources are detected, or inactive sources are discarded. Given the number and the location of the speech sources and the bin-wise position estimates, the corresponding SOS are estimated from the microphone signals by temporal averaging. Finally, a minimum variance distortionless response (MVDR) is employed that promptly adapts to emerging and disappearing interferers, while maintaining low distortion of the desired speech signal. The framework could potentially be used for source separation, where each source is extracted by a filter that reduces the remaining sources and the background noise.

The paper is organized as follows: in Section 2 we present the signal model. The source extraction problem consists of (i) detecting the number of active sources online, as described in Section 3, and (ii) estimating the SOS of the active sources based on bin-wise position estimates, as described in Section 4. Evaluation results are presented in Section 5 and Section 6 concludes the paper.

2. PROBLEM FORMULATION

Consider a scenario where M microphones from S distributed arrays capture a desired speech signal, unknown and possibly time-varying number of interfering signals that are coherent across the arrays, and a background noise signal. The short-time spectral coefficients of the microphone signals are given by an $M \times 1$ vector

$$\boldsymbol{y}(n,k) = \sum_{i \in \mathcal{Q}_n} \boldsymbol{x}_i(n,k) + \boldsymbol{v}(n,k), \tag{1}$$

where the set $\mathcal{Q}_n$, contains the labels of all active sources, the vectors $\boldsymbol{x}_i$ and $\boldsymbol{v}$ are the speech and the noise signals, respectively, and n and k are the time and frequency indices. Let d denote the label of the desired source signal $\boldsymbol{x}_d$, and let the set $\mathcal{Q}_n^{\mathrm{i}} = \mathcal{Q}_n \setminus \{d\}$ denote the labels of all interfereing signals. The different signals in (1) are considered to be realizations of zero-mean, mutually uncorrelated random processes. The power spectral density (PSD) matrix of the microphone signals is denoted by $\boldsymbol{\Phi}_{\boldsymbol{y}}(n,k) = \mathrm{E}\left[\boldsymbol{y}(n,k)\boldsymbol{y}^{\mathrm{H}}(n,k)\right]$, where $\mathrm{E}\left[\cdot\right]$ is the expectation operator. The PSD matrices of $\boldsymbol{x}_d$, $\boldsymbol{x}_i$,

[*] A joint institution of the University Erlangen-Nuremberg and Fraunhofer IIS, Germany

and $\boldsymbol{v}$ are defined similarly. As the signal components are uncorrelated, it follows that

$$\boldsymbol{\Phi}_{\boldsymbol{y}}(n,k) = \boldsymbol{\Phi}_{\boldsymbol{x}_d}(n,k) + \sum_{i \in \mathcal{Q}_n^{\mathrm{i}}} \boldsymbol{\Phi}_{\boldsymbol{x}_i}(n,k) + \boldsymbol{\Phi}_{\boldsymbol{v}}(n,k). \quad (2)$$

The goal is to estimate the desired signal captured at the m-th microphone, by a linear combination of the microphone signals, i.e.,

$$\widehat{X}_{d,m}(n,k) = \boldsymbol{w}_{d,m}^{\mathrm{H}}(n,k)\, \boldsymbol{y}(n,k). \quad (3)$$

If the PSD matrices of $\boldsymbol{x}_d$, $\boldsymbol{x}_i$, and $\boldsymbol{v}$ are known, an MVDR filter that minimizes the undesired signal power while preserving the desired signal can be computed as follows [8]

$$\boldsymbol{w}_{d,m} = (\boldsymbol{g}_{d,m}^{\mathrm{H}}\, \boldsymbol{\Phi}_{\mathrm{u}}^{-1} \boldsymbol{g}_{d,m})^{-1}\, \boldsymbol{\Phi}_{\mathrm{u}}^{-1}\, \boldsymbol{g}_{d,m}\ , \quad (4)$$

where the time and frequency indices were omitted for brevity. The relative transfer function $\boldsymbol{g}_{d,m}$ between the desired source and the m-th microphone is obtained as the m-th column of $\boldsymbol{\Phi}_{\boldsymbol{x}_d}$ normalized by the desired signal PSD at the m-th microphone, and $\boldsymbol{\Phi}_{\mathrm{u}}$ is the PSD matrix of all undesired signals, i.e.,

$$\boldsymbol{\Phi}_{\mathrm{u}}(n,k) = \sum_{i \in \mathcal{Q}_n^{\mathrm{i}}} \boldsymbol{\Phi}_{\boldsymbol{x}_i}(n,k) + \boldsymbol{\Phi}_{\boldsymbol{v}}(n,k). \quad (5)$$

Note that if an estimate of $\boldsymbol{\Phi}_{\boldsymbol{x}_d}$ is available, the matrix $\boldsymbol{\Phi}_{\mathrm{u}}$ is given by $\boldsymbol{\Phi}_{\boldsymbol{y}} - \boldsymbol{\Phi}_{\boldsymbol{x}_d}$. While this estimate can be useful in applications such as signal detection, the sensitivity to estimation errors leads to unpleasant artifacts and inconsistent quality in spatial filtering applications. Therefore, we aim at estimating the PSD matrix of each interferer and the background noise separately, and compute $\boldsymbol{\Phi}_{\mathrm{u}}$ using (5). Estimating the number of interferers and their PSD matrices $\boldsymbol{\Phi}_{\boldsymbol{x}_i}$, for $i \in \mathcal{Q}_n^{\mathrm{i}}$, using spatial cues extracted by distributed arrays is the main contribution of this paper. A block diagram of the proposed framework is illustrated in Figure 1.

3. NUMBER OF SOURCE ESTIMATION

The total number of active sources is estimated each time frame using bin-wise position estimates from the L most recent frames and is denoted by $|\mathcal{Q}_n|$. Given the position estimates, a time-varying convex quadrilateral in the form of a kite is assigned to each active source, where the kite vertices are completely determined by the set of position estimates. The number of sources at each frame is then obtained by counting the number of kites.

3.1. Bin-wise position estimation

Given bin-wise DOA estimates from two distributed arrays, a position can be estimated by triangulation. The DOAs can be for instance computed using the estimator which was used in the number of source estimation algorithm in [6]. Let $\hat{\boldsymbol{n}}_s = [\cos(\hat{\theta}_{\mathrm{s}}),\, \sin(\hat{\theta}_{\mathrm{s}})]$ denote the unit vector pointing towards the DOA $\hat{\theta}_s$ estimated at array s. While we consider only the azimuthal angle, it is straightforward to extend the algorithm to consider also the elevation angle. For each TF bin, the DOAs of two arrays i and j are used to obtain the position estimate $\hat{\boldsymbol{r}}$ by triangulation, which are chosen based on a simple criterion, described as follows: (i) given the DOA vectors for all arrays $\hat{\boldsymbol{n}}_1, \ldots, \hat{\boldsymbol{n}}_S$, the inner product between each pair of vectors is computed; (ii) the arrays corresponding to the vector pair with angle closest to 90 degrees are used for triangulation. The criterion can be formally expressed as follows

$$(i,j) = \underset{i',j'}{\arg\min}\, |\hat{\boldsymbol{n}}_{i'}^{\mathrm{T}} \hat{\boldsymbol{n}}_{j'}|, \quad i' \neq j'. \quad (6)$$

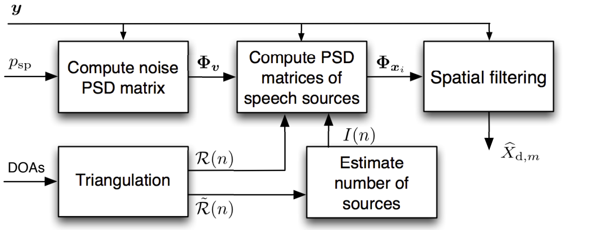

$\boldsymbol{y}$
p_{sp}
Compute noise PSD matrix
$\boldsymbol{\Phi}_{\boldsymbol{v}}$
Compute PSD matrices of speech sources
$\boldsymbol{\Phi}_{\boldsymbol{x}_i}$
Spatial filtering
$\widehat{X}_{\mathrm{d},m}$
$I(n)$
DOAs
Triangulation
$\mathcal{R}(n)$
$\tilde{\mathcal{R}}(n)$
Estimate number of sources

Fig. 1: Block diagram of the proposed source extraction framework, where p_{sp} denotes the speech presence probability.

The reasoning behind this criterion is that for angles which are closer to 90 degrees, the triangulation is less sensitive to DOA errors. As the angle between the DOA vectors becomes small, a small DOA error results in a large position error, and as the angle approaches 180 degrees such that the DOA vectors are almost parallel, triangulation might fail.

3.2. TF bin selection and source kite computation

For robust number of source estimation, a subset of reliable TF bins needs to be selected where the bin-wise position estimates accurately represent the locations of the sources. For instance, in [6, 7], the TF bins where only one source is dominant were selected, requiring an algorithm to detect single-source bins. We propose to select reliable TF bins using a speech presence probability (SPP) and the DOA estimates of all arrays. The set of position estimates corresponding to the reliable TF bins at a time frame n is denoted by $\mathcal{R}(n)$. Given a position estimate $\hat{\boldsymbol{r}}$, and the array positions $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_S$, we denote by $\theta_{r,s}$ the angles between the vectors $\hat{\boldsymbol{r}} - \boldsymbol{p}_s$ for $s = 1, \ldots, S$ and the horizontal axis. Moreover, for each array s, a histogram of all DOA estimates at frame n is computed, and the maximum of the histogram is denoted by $\tilde{\theta}_s$. We used a histogram resolution of 10°. A position $\hat{\boldsymbol{r}}(n,k)$ is considered reliable, and hence belongs to $\mathcal{R}(n)$ if $p_{\mathrm{sp}}(n,k) > p_{\mathrm{thr}}$ and the angles $\theta_{r,s}$ for each array s satisfy $|\theta_{r,s}(n,k) - \tilde{\theta}_s| \leq \Delta_\theta$. The SPP threshold p_{thr} ensures speech presence and Δ_θ is chosen small enough such that, with a high probability, the position estimates in $\mathcal{R}(n)$ correspond to a single source. Hence, the set $\mathcal{R}(n)$ is associated to the source that corresponds to the peak in the DOA histograms of all arrays at frame n. This check is performed for all arrays in order to minimize the size of the clusters that are formed by the reliable position estimates for each source, leading to more robust number of source estimation.

Let the set $\mathcal{R}_L(n)$ contain all reliable position estimates from the past L frames, i.e.,

$$\mathcal{R}_L(n) = \mathcal{R}(n) \cup \mathcal{R}(n-1) \cup \ldots \cup\ \mathcal{R}(n-L+1). \quad (7)$$

A kite $K_i(n)$ for source i is defined using a set $\mathcal{S} \subseteq \mathcal{R}_L(n)$ of position estimates related to that source, where the assignment of position estimates to sources is described in Section 3.3. Given the set $\mathcal{S}$, the kite vertices are computed as follows (see Figure 2):

(i) The vertex $\boldsymbol{p}_1$ is the position of the array s with the smallest Euclidean distance to the mean of $\mathcal{S}$.

(ii) Determine the maximum $\hat{\theta}_{s,\max}$ and minimum $\hat{\theta}_{s,\min}$ among the DOAs related to the position estimates in $\mathcal{S}$.

(iii) Determine the position estimate in $\mathcal{S}$ with the largest Euclidean distance $d_{\max}$ from array s.

(iv) We introduce three parameters related to the kites: angular margin θ_{mar}, distance margin d_{mar}, and maximum angle θ_{res}. The margins θ_{mar} and d_{mar} allow for deviations of the position estimates of

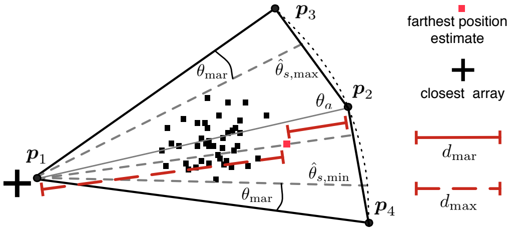


Fig. 2: Illustration of the kite computation

Algorithm 1 *Inputs:* $\mathcal{Q}_{n-1}$, $K_{i,n-1}$, $\forall i \in \mathcal{Q}_{n-1}$, $\boldsymbol{t}$(n-1), $\mathcal{R}(n)$, $\mathcal{R}_L(n)$

```
IF  $\mathcal{Q}_{n-1} \neq \emptyset$ (there are sources present)
  1. $\mathcal{Q}_n = \mathcal{Q}_{n-1}$, $K_{i,n} = K_{i,n-1}$,  for $i \in \mathcal{Q}_n$ (propagate sources)
  2. $\boldsymbol{t}(n) = \boldsymbol{t}(n-1) + \mathbf{1}_{|\mathcal{Q}_n|\times 1}$
  3. Compute the sets $\Pi_i = K_{i,n} \cap \mathcal{R}(n)$  for $i \in \mathcal{Q}_n$
  IF $|\Pi_{i'}| > 0$, where $i' = \arg\max_i(|\Pi_i|)$
    4. Set the $i'$-th entry of $\boldsymbol{t}(n)$ to zero
    5. $\mathcal{S}_{i'} = [\mathcal{R}_L(n) \cap K_{i',n}] \setminus [\mathcal{R}_L(n) \cap [\cup_{i \neq i'} K_{i,n}]]$
    6. The vertices of $K_{i',n}$ are recomputed using $\mathcal{S}$ (see Section 3.2)
  ELSE (adding a new source)
    7. Create an auxiliary kite $K_{\rm aux}$ using the points in $\mathcal{R}(n)$
    8. $\mathcal{S}_{\rm aux} = \mathcal{R}_L(n) \cap K_{\rm aux}$
    IF $|\mathcal{S}| > \eta_{\rm det}$ (minimum number of points to detect a source)
      9. Add a kite $K_{i_{\rm new},n}$ using $\mathcal{S}_{\rm aux}$, as described in Section 3.2
      10. Add a source: add $i_{\rm new}$ to $\mathcal{Q}_n$, add a token $\boldsymbol{t}(n) = [\ \boldsymbol{t}(n), 0\ ]$
    END
  END
ELSE (no sources present)
  11. Execute lines 7-10.
END
13. Count the number $l$ of entries in $\boldsymbol{t}(n)$ that exceed $t_{\max}$
14. Remove the $l$ sources from $\mathcal{Q}_n$ and their respective kites
Outputs: $\mathcal{Q}_n$,  $K_{i,n}$ for $i \in \mathcal{Q}_n$, $\boldsymbol{t}(n)$
```

a source, hence accounting for estimation errors. The maximum angle $\theta_{\rm res}$ limits the size of the kites.

(v) We define $\theta_a = {}^{(\hat{\theta}_{s,\min}+\hat{\theta}_{s,\max})}/_2$, $\theta_b = \min(\hat{\theta}_{\rm s,max} + \theta_{\rm mar}, \theta_{\rm a} + \theta_{\rm res})$, and $\theta_c = \max(\hat{\theta}_{\rm s,min} - \theta_{\rm mar}, \theta_{\rm a} - \theta_{\rm res})$. The remaining three vertices are then given as follows

$$\boldsymbol{p}_2 = \boldsymbol{p}_1 + (d_{\max} + d_{\rm mar}) \cdot [\cos(\theta_{\rm a}),\ \sin(\theta_{\rm a})] \tag{8a}$$

$$\boldsymbol{p}_3 = \boldsymbol{p}_1 + (d_{\max} + d_{\rm mar}) \cdot [\cos(\theta_{\rm b}),\ \sin(\theta_{\rm b})] \tag{8b}$$

$$\boldsymbol{p}_4 = \boldsymbol{p}_1 + (d_{\max} + d_{\rm mar}) \cdot [\cos(\theta_{\rm c}),\ \sin(\theta_{\rm c})]. \tag{8c}$$

3.3. Counting the number of kites

At each time frame, the sets $\mathcal{R}(n)$, $\mathcal{R}_L(n)$, the set of source labels from the previous frame $\mathcal{Q}_{n-1}$, and the existing source kites are inputs to the source counting algorithm, given by Algorithm 1. The first step is to find the kite $K_{i'}$ which contains the maximum number of points from $\mathcal{R}(n)$. If the maximum is nonzero, the kite is updated using the set $\mathcal{S}_{i'}$ that consists of all points in $R_L(n)$ which are in $K_{i'}$, but not in any other kite (line 5). If the set $\mathcal{R}(n)$ has no intersection with the existing kites, or if there are no existing kites so far (lines 7-11), an auxiliary kite $K_{\rm aux}$ is formed using $\mathcal{S}_{\rm aux} = \mathcal{R}(n)$. A new source is added if a sufficient number of points $\eta_{\rm det}$ from $\mathcal{R}_L(n)$ belong to $K_{\rm aux}$. The value of $\eta_{\rm det}$ should be chosen such that new sources are promptly detected, while minimizing false alarms.

To track the number of active sources, we use a so-called token for each source. The tokens for all active sources are stored in a vector $\boldsymbol{t}(n)$. Once a new source is detected, it is assigned a token with value 0 (line 10). Once $\mathcal{R}(n)$ is associated with an existing source, the token of that source is reset to 0 (line 4), while the tokens of all other sources are incremented by one (line 2). When the token of a source reaches $t_{\max}$, the source is declared inactive, its kite is removed and the source label is deleted from $\mathcal{Q}_n$ (lines 13-14). The parameter $t_{\max}$ should provide a good trade-off between (i) prompt inactivity detection, so that the spatial filter aims at reducing only the actually active interferers, and (ii) accounting for short speech pauses without removing a particular interferer. In our implementation we used $t_{\max} = 90$ (corresponding to approximately 3 seconds).

4. POSITION-BASED PSD MATRIX ESTIMATION

To compute a spatial filter that extracts one of the sources in $\mathcal{Q}_n$, while reducing the remaining sources, the PSD matrix of each source is required. Recently proposed approaches for estimating PSD matrices of speech sources from a mixture are based on recursive temporal averaging. The computation of the the averaging parameter is a crucial factor for the estimation accuracy. Different probabilistic methods have been recently proposed [1–3], where given a probability $p_{x_i}(n,k)$ that the i-th source is dominant at TF bin (n,k), the averaging parameter is computed as

$$\alpha_{x_i}(n,k) = 1 - p_{x_i}(n,k)(1 - \tilde{\alpha}), \quad \tilde{\alpha} \in [0,1), \tag{9}$$

and the PSD matrix is estimated according to

$$\boldsymbol{\Phi}_{\boldsymbol{x}_i}(n) = \alpha_{x_i}(n)\,\boldsymbol{\Phi}_{\boldsymbol{x}_i}(n-1) + [1 - \alpha_{x_i}(n)]\,\boldsymbol{y}(n)\boldsymbol{y}^{\rm H}(n). \tag{10}$$

The frequency index in (10) was omitted for brevity. To compute an averaging parameter $\alpha_v(n,k)$ for noise PSD matrix estimation SPP can be used, as done in [9]. In this work, we propose a position-based non-probabilistic indicator function $\mathcal{I}_{x_i}(n,k)$ to compute the averaging parameter $\alpha_{x_i}(n,k)$, for each source i as follows

$$\alpha_{x_i}(n,k) = 1 - \mathcal{I}_{x_i}(n,k)(1 - \tilde{\alpha}) \ \ \tilde{\alpha} \in [0,1). \tag{11}$$

In this way, $\alpha_{x_i} = \tilde{\alpha}$ when $\mathcal{I}_{x_i} = 1$ such that $\boldsymbol{\Phi}_{\boldsymbol{x}_i}$ is recursively updated, and $\alpha_{x_i} = 1$ when $\mathcal{I}_{x_i} = 0$, such that $\boldsymbol{\Phi}_{\boldsymbol{x}_i}$ takes the value from the previous frame.

As described in Section 3, each source $i \in \mathcal{Q}_n$ is associated with a kite $K_{i,n}$. We can estimate the location of the i-th source at frame n as the mean $\boldsymbol{\mu}_i$ of the position estimates in $\mathcal{S}_i$, where $\mathcal{S}_i$ was defined in line 5 of Algorithm 1. Given a position estimate $\hat{\boldsymbol{r}}(n,k)$, and an SPP $p_{\rm sp}(n,k)$, the indicator functions at TF bin (n,k) are computed as follows

$$\mathcal{I}_{x_i} = \begin{cases} 1, & \text{if } \ \|\hat{\boldsymbol{r}} - \boldsymbol{\mu}_i\| < \|\hat{\boldsymbol{r}} - \boldsymbol{\mu}_j\| \ \ \forall j \neq i, \ \text{ and } \ p_{\rm sp} > p_{\rm thr} \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where the SPP was computed as in [9], and $p_{\rm thr}$ is a threshold to ensure speech presence. In this manner, at a TF bin (n,k) only the PSD matrix of the source whose estimated location $\boldsymbol{\mu}_i$ has the shortest Euclidean distance to the bin-wise position $\hat{\boldsymbol{r}}(n,k)$ is updated. Note that the indicator function is computed separately at each frequency k, meaning that PSD matrices of different sources can be estimated during periods when multiple sources are active as well.

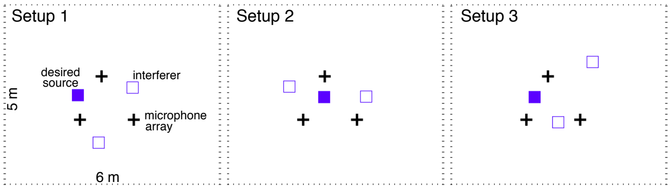


Fig. 3: Scenarios considered for the evaluation.

α_{x_i}	p_{thr}	Δ_θ	η_{det}	$t_{\max}$	d_{mar}	θ_{mar}	θ_{res}	L
0.8	0.9	15°	20	90	25 cm	10°	20°	90

Table 1: Parameters used in the performance evaluation

5. PERFORMANCE EVALUATION

5.1. Performance measures and experimental setup

To evaluate the proposed system, we focused on two main aspects: (i) estimation of number of sources in different acoustic conditions and (ii) extracted signal quality in terms of segmental speech distortion index ν_{sd} [8], segmental interference reduction (segIR) Δ_{i}, and segmental noise reduction (segNR) Δ_v, where the noise consists of diffuse background noise and sensor noise. The quantities Δ_v and Δ_{i} were obtained by averaging segment-wise values of the segNR and segIR computed over non-overlapping 30 ms segments. The averaging was done in the linear domain for ν_{sd}, and in the logarithmic domain for Δ_v and Δ_{i}. Given the residual noise $\widehat{v}_m$, the segment-wise NR for the i-th segment at microphone m was computed as

$$\mathrm{SegNR}_m(i) = \frac{\sum_t v_m^2(t) \cdot w_i(t)}{\sum_t \widehat{v}_m^2(t) \cdot w_i(t)}, \tag{13}$$

where w_i is a rectangular window equal to one in segment i and zero elsewhere. The segIR is computed similarly. For the performance evaluation, one of the sources illustrated in Figure 3 was selected as desired, and the remaining sources were considered as interferers. In practice, the desired source can be for instance a source that is located in certain region of the room, or if source separation is required, each source can be extracted separately while reducing the remaining sources. Note that as mentioned in the problem formulation in Section 2, the proposed system assumes that the active sources are coherent across the arrays. This is satisfied in scenarios where the reverberation time is low to moderate or where the sources are in the vicinity of the arrays. Only in such cases the sources can be robustly detected using bin-wise position estimates. This is a reasonable assumption for instance in meeting/teleconferencing scenarios or smart home applications where multiple sources may be in the vicinity of the array but only the signal of one source is desired.

Simulations were done in a 6×5×3 m^3 shoe-box room, for the scenarios in Figure 3. The microphone signals were obtained by convolving clean speech signals with simulated room impulse responses [10], adding diffuse babble noise as background noise [11], and adding uncorrelated sensor noise with desired signal-to-sensor noise ratio of approximately 40 dB. While only the results for reverberation times T_{60} of 0.2 s and 0.35 s are evaluated here, the number of source estimation algorithm has shown to work robustly up to $T_{600} = 600$ ms, provided that the sources are in the vicinity of the arrays. The sampling frequency was 16 kHz and the STFT length was 64 ms with 50% overlap. Three uniform circular arrays were used with three omnidirectional microphones each and a diameter 2.5 cm. The remaining parameters are given in Table 1.

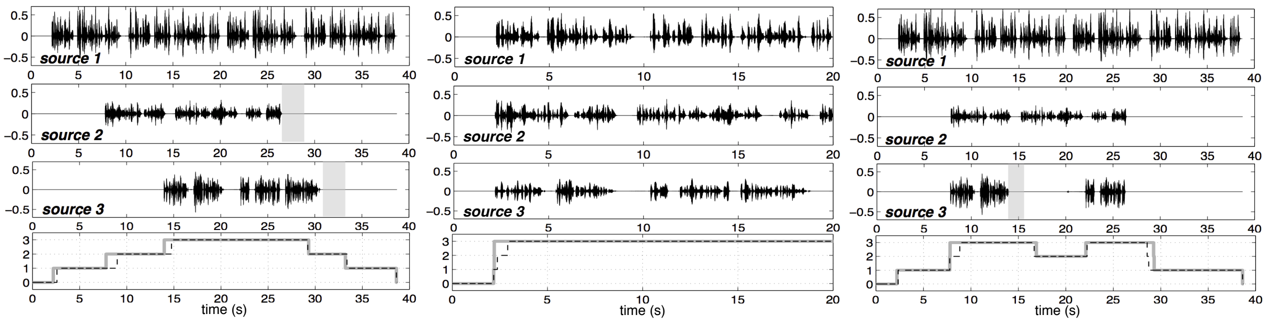
5.2. Results: estimating number of sources

In Figure 4, the results of the number of source estimation algorithm are presented for three activity patterns of the sources, and compared to the ground truth value of the number of sources. Based on the discussion in Section 3.3, where the source activity is monitored using the so-called tokens, the ground truth is computed as follows: a source is considered active in time frame n if it has been active in at least one of the most recent L frames. The results in Figure 4 correspond to $T_{60} = 350$ ms, where the diffuse noise level was set such that the input signal-to-noise ratio (iSNR) was approximately 5 dB. This was the worst case scenario among the evaluated acoustic conditions. Whenever a new source appears, the algorithm promptly detects it, even if multiple new sources appear simultaneously, as in Fig. 4(c). The effect of the tokens stored in the vector $\boldsymbol{t}(n)$ is illustrated in Fig. 4(a) and 4(c), where the shaded regions indicate the time it takes for an inactive source to be discarded. Moreover, the fact that the source number remains unchanged during speech pauses in Figure 4(b) is also due to the token-based activity monitoring. In some applications, removing a source during short pauses might have an adverse effect on the interference reduction if the interferer becomes active again. The latency in detecting a new source, which can be controlled by the parameter η_{det}, is visible in Figure 4, where the estimated number of sources deviates from the ground truth at the onsets of a new source. The time it takes for a source to be detected depends on the reverberation time, the noise level, and the value of η_{det}. For $T_{60} = 200$ ms and iSNR of 15 dB, the latency was smaller than the one illustrated in Figure 4.

5.3. Results: evaluation of extracted signals

We evaluated the performance of the proposed framework and added as a reference the results obtained with oracle source detection for the PSD matrix updates, and with the probabilistic approach proposed by the authors in [1]. The latter is denoted by ”EM-based” (EM for Expectation Maximization) and requires the number of sources in advance. Moreover, this approach can not be applied if the number of sources varies online. The three approaches were evaluated in a constant triple talk, for all scenarios in Fig. 3. As no significant dependence on source positions was observed, the results were averaged over the three scenarios. Diffuse noise was added to obtain an iSNR of approx. 15 dB and 5 dB (upper and lower part of Table 2, respectively). The results indicate that the ability to handle dynamic scenarios comes at the cost of a slightly worse performance compared to the EM-based approach. The most significant performance drop is observed in terms of speech distortion, especially at higher reverberation times. The segIR is only reduced by 1 dB on average, and the segNR is slightly better for the proposed approach. The results for the remaining activity patterns at iSNR $\approx$ 15dB are given in Table 3. For the EM-based approach, the number of sources was fixed to three, regardless of the activity pattern.

Online detection of active sources increases the response time of the spatial filter to adapt to reduce new interferers. The segment-wise IR during an onset of an interferer is shown in Fig. 5. Nevertheless, position-based PSD matrix computation results in prompt adaptation of the MVDR such that in 0.5 s the performance converges to the one of the EM-based approach where the number of interferers and their locations are known in advance. At the cost of artifacts, such as musical noise and false alarms for new interferers, the response time can be reduced by adjusting the parameters α_{x_i} and η_{det}. The audio files used in the evaluation are available at http://www.audiolabs-erlangen.de/resources/2014-EUSIPCO-SE.



(a) Pattern1

(b) Pattern2

(c) Pattern3

Fig. 4: Estimated number of sources. Simulations with the illustrated activity patterns were done for all scenarios in Fig. 3 and for T_{60}= 0.2 s and 0.35 s. The estimated number of sources in all cases was as illustrated above. The solid gray line denotes the ground truth value, whereas the dashed line denotes the estimated number of sources. The iSNR was approximately 5 dB.

	Oracle		EM-based		Proposed	
T_{60}[s]	0.20	0.35	0.20	0.35	0.20	0.35
iSNR [dB]	14.0	14.9	14.0	14.9	14.0	14.9
iSIR [dB]	-1.0	-1.6	1.0	-1.6	1.0	-1.6
segNR [dB]	6.9	6.8	4.7	4.9	5.1	5.2
segIR [dB]	18.8	16.5	15.5	12.8	15.1	11.2
ν_{sd}	0.023	0.066	0.025	0.058	0.034	0.120
iSNR [dB]	5.0	5.4	5.0	5.4	5.0	5.4
iSIR [dB]	-1.0	-1.6	-1.0	-1.6	-1.0	-1.6
segNR [dB]	9.6	8.8	6.9	6.3	7.2	6.6
segIR [dB]	18.7	16.4	16.0	12.9	15.9	11.2
ν_{sd}	0.026	0.075	0.020	0.060	0.026	0.120

Table 2: Results for `pattern2` (see Fig. 4).

	Oracle		EM-based		Proposed	
T_{60}[s]	0.20	0.35	0.20	0.35	0.20	0.35
iSNR [dB]	14.0	14.9	14.0	14.9	14.0	14.9
iSIR [dB]	1.5	0.7	1.5	0.7	1.5	0.7
segIR [dB]	20.1	17.6	16.2	13.7	16.5	11.8
ν_{sd}	0.018	0.063	0.022	0.063	0.030	0.127

Table 3: Average results for `pattern1` and `pattern3`.

6. CONCLUSIONS

A non-probablistic source extraction framework was proposed with a geometry-based online number of source estimator. A desired source was extracted using the microphone signals from distributed arrays. The proposed approach does not require the number of sources in advance and can be applied to extract a desired signal in dynamic scenarios with time-varying number of interferers. Scenarios with different source positions and source activities over time were simulated and satisfactory results were obtained for different noise conditions and mild to moderate reverberation times. Future work includes evaluation of the proposed framework using measured data in different acoustic conditions, as well as extension of the number of sources estimator to handle moving sources.

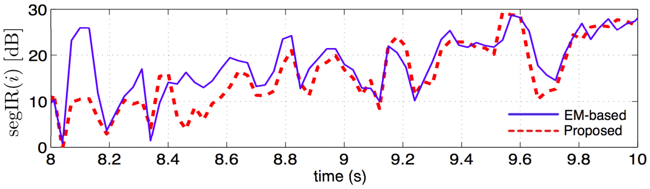


Fig. 5: Segment-wise IR during an onset of an interferer.

REFERENCES

[1] M. Taseska and E.A.P Habets, "MMSE-based source extraction using position-based posterior probabilities," in *ICASSP*, 2013.

[2] D. H. Tran Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *IWAENC*, 2010.

[3] M. Taseska and E. A. P. Habets, "Informed spatial filtering with distributed arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1195–1207, 2014.

[4] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *ICASSP*, 2012.

[5] I. Jafari, N. Ito, M. Souden, S. Araki, and T. Nakatani, "Source number estimation based on clustering of speech activity sequences for microphone array processing," in *IEEE Intl Workshop on MLSP*, Sept 2013, pp. 1–6.

[6] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *IWAENC*, 2008.

[7] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct 2013.

[8] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[9] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *IWAENC*, Sept. 2012.

[10] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2006.

[11] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.