Informed Spatial Filtering for Sound Extraction Using Distributed Microphone Arrays

Maja Taseska, Student Member, IEEE, and Emanuël A. P. Habets, Senior Member, IEEE

Abstract—Hands-free acquisition of speech is required in many human-machine interfaces and communication systems. The signals received by integrated microphones contain a desired speech signal, spatially coherent interfering signals, and background noise. In order to enhance the desired speech signal, state-of-the-art techniques apply data-dependent spatial filters which require the second order statistics (SOS) of the desired signal, the interfering signals and the background noise. As the number of sources and the reverberation time increase, the estimation accuracy of the SOS deteriorates, often resulting in insufficient noise and interference reduction. In this paper, a signal extraction framework with distributed microphone arrays is developed. An expectation maximization (EM)-based algorithm detects the number of coherent speech sources and estimates source clusters using time-frequency (TF) bin-wise position estimates. Subsequently, the second order statistics (SOS) are estimated using bin-wise speech presence probability (SPP) and a source probability for each source. Finally, a desired source is extracted using a minimum variance distortionless response (MVDR) filter, a multichannel Wiener filter (MWF) and a parametric multichannel Wiener filter (PMWF). The same framework can be employed for source separation, where a spatial filter is computed for each source considering the remaining sources as interferers. Evaluation using simulated and measured data demonstrates the effectiveness of the framework in estimating the number of sources, clustering, signal enhancement, and source separation.

Index Terms—Distributed arrays, EM algorithm, PSD matrix estimation, source extraction, spatial filtering.

I. INTRODUCTION

The extraction of a desired speech signal from a mixture of signals from multiple simultaneously active talkers and background noise is of interest in many hands-free communication systems, including modern mobile devices, smart homes, and teleconferencing systems. In some applications, e.g., where automatic speech recognition is required, the goal is to obtain an estimate of the signal from a desired talker, while reducing noise and signals from interfering talkers. In other applications, an estimate of each talker’s signal is required. In practice, information about the number and location of the different talkers, or the presence and type of background noise is unavailable and the estimation is based solely on the microphone signals. Traditionally, multichannel noise and interference reduction is achieved by linearly combining the signals from closely spaced microphones, known as beamforming, initially developed for radar and sonar applications in the mid twentieth century [1]–[3]. In case of wideband signals such as speech, spatial filtering (beamforming) is often performed in the TF domain [4], where the signal at each frequency satisfies the narrowband assumption, allowing for standard beamforming techniques. Moreover, TF domain processing offers the flexibility to tune the spatial filter performance at each TF bin separately. To compute the coefficients of a spatial filter that is optimum with respect to a certain statistical criterion, e.g., minimum mean squared error (MMSE), the SOS of the noise and the interfering signals need to be estimated from the microphone signals [5]. The estimation accuracy is a crucial performance factor, since overestimation could lead to a cancellation of the desired signal, while underestimation could result in high levels of residual noise and interference.

State-of-the-art methods for single- and multichannel noise power spectral density (PSD) estimation are based on recursive temporal averaging, controlled by a single- or multichannel SPP [6]–[8]. In contrast to traditional voice activity detectors [9], SPP allows for updates of the noise PSD during speech presence as well, leading to a better performance, especially in scenarios with non-stationary noise and low signal-to-noise ratios (SNRs). Similarly, to estimate the source PSD matrices by recursive averaging, it is crucial to accurately associate each TF bin with the active sources at that TF bin. In recent research [10], [11], the source activity in each TF bin is described by a set of disjoint states, such that each state indicates that a particular source is dominant at a given TF bin. Such description relies on the assumption that speech signals are sparse in the TF domain [12], which usually holds in mildly reverberant environments with few simultaneously active talkers.

In order to determine the dominant source at each TF bin, spatial cues extracted using multiple microphones are commonly used. The microphone signal vectors contain spatial information that can be used for this task, as it has been done in [10], [11], [13]–[15]. Alternatively, parametric information such as binaural cues [16], direction of arrival (DOA) [17], and bin-wise position estimates [18] can be extracted from the microphone signals. Spatial filters that are obtained by employing parametric information in the SOS estimation are referred to as informed spatial filters. The signal vectors and the parametric information can also be used jointly, as it has been recently done in
[19]. Additionally, spectral cues [14, 20] and temporal correlations [15] can be exploited to improve the detection of dominant sources, and hence the estimation of their SOS. In the majority of these contributions, probabilistic frameworks prevail, where the spatial cues extracted when a particular source is dominant are modeled by an underlying probability distribution. Hence, the distribution of all observed spatial cues is modeled by a mixture probability distribution. To detect the dominant source, it is then required to estimate the mixture parameters and compute probabilities related to the activity of each source.

The EM algorithm has been often employed to estimate the mixture parameters and the source probabilities [10, 13, 15, 16, 18, 21]. In some of the first related contributions [21–23], the source probabilities were used as TF masks for source separation. Although TF masking can achieve significant interference reduction and improvement in desired speech intelligibility, violation of the sparsity assumption rapidly results in high distortion of the desired signal, especially in reverberant multi-talk scenarios. Moreover, TF masking does not fully utilize the spatial diversity offered by the microphone arrays, as the desired signal estimate is obtained by applying a spectral weighting to a single reference microphone. For further applications of TF masking and its relation to computational auditory scene analysis (CASA), the reader is referred to [24] and references therein. On the other hand, using the source probabilities to estimate SOS for spatial filtering as done in [10], [11], [18], has been shown to provide a very good interference reduction, while maintaining low distortion of the desired signal.

In standard EM-based algorithms, the number of mixture components, i.e., the number of sources, needs to be known in advance. This represents a significant drawback, as the number of sources is often unknown in practice and needs to be estimated from the microphone signals. To overcome this limitation, the authors in [25] use a maximum a-posteriori version of the EM algorithm, where the number of sources is modeled as a random variable with a Dirichlet prior. However, according to the reported results, the algorithm requires a significant number of iterations to converge, even in mildly reverberant scenarios. Recently, the authors in [26] used a variational EM algorithm which can estimate the number of sources and the mixture parameters, at the cost of an increased computational complexity compared to the standard maximum likelihood (ML) EM algorithm. Further sparsity-based methods to detect the number of sources have for instance been considered in [27, 28], where instead of the EM algorithm, different clustering and classification methods are employed. In this work, we build upon our previous work in [18] and propose an efficient EM-based algorithm which uses bin-wise position estimates. The main components of the algorithm are (i) a standard ML EM iteration, (ii) a bin-wise position-based estimation of the number of sources, and (iii) pruning of the noisy and reverberant training data. The number of sources and the mixture parameters are accurately estimated in a few iterations, even in challenging multi-talk scenarios.

Besides the variety of source extraction algorithms which employ microphone arrays with co-located microphones, distributed microphone arrays (DMAs) have been often considered in related research over the last decade (see [29] and references therein). Researchers propose different methods to exploit the additional spatial diversity offered by DMAs. For instance, the authors in [20] extract source posterior probabilities for each array and merge them into a single probability before updating the mixture parameters and before using the probability for SOS estimation. For different distributed EM algorithms in the context of source separation the reader is referred to [30–32]. Several methods to compute optimal spatial filter coefficients using DMA have been proposed in [33–35] and references therein. DMAs can also be used to extract additional spatial cues such as the level difference between the signals at different arrays [20]. In our contribution, the motivation for using DMAs is twofold: firstly, we use the position estimate as a spatial cue, which is obtained by triangulation of the DOA estimates from at least two DMAs. Secondly, we compute an estimate of the desired speech signal by combining all available microphones, which in most cases results in superior interference reduction compared to a single microphone array with co-located microphones.

To summarize, in this work we develop a spatial filtering framework which estimates the number of sources and the SOS using parametric information extracted from DMAs. We use the direct to-diffuse ratio (DDR) to estimate the SPP, and bin-wise position to estimate the source probabilities in each TF bin. The DDR-based SPP and the position-based source probability estimation were recently proposed by the present authors in [18], [36]. The novel contributions of this work include an extension of the framework in [18] to handle unknown number of sources. We propose an efficient EM-based algorithm that simultaneously estimates the number of sources and the associated mixture parameters. Moreover, we compare the source extraction performance of the MVDR, MWF and PMWF, and propose a method to control the PMWF trade-off parameter using the source probabilities. We consider scenarios where the number of detected sources does not change, however we do not impose restrictions on the source activity, i.e., speech pauses, simultaneously active talkers, or inactivity of some of the talkers. The source clustering and the source extraction performance were extensively evaluated with simulated and measured data.

The rest of the paper is organized as follows: in Section II, we define the signal model in the TF domain and formulate the source extraction problem. In Section III, the spatial filters used in this contribution are briefly derived and explained. The estimation of the SOS of the noise and the different source signals is discussed in Section IV. The SPP and the source probabilities required for the PSD matrix estimation are detailed in Section V. The main contributions of this work are presented in Section VI, where the proposed EM-based algorithm which detects the number of sources is described, and in Section VII, where the proposed PMWF trade-off parameter is described. A comprehensive performance evaluation of the two main blocks of the framework, namely, (i) the number of source estimation and clustering, and (ii) using the cluster information in a probabilistic framework for source extraction is provided in Section VIII. Section IX concludes the paper.

II. PROBLEM FORMULATION

A. Linear Signal Model

The spatial filtering framework developed in this contribution is defined in the frequency domain. A short-time Fourier transform (STFT) is applied to the time domain microphone signals
and each TF bin is processed independently. If the total number of microphones is denoted by $M$ and the total number of talkers by $J$, the microphone signals in the STFT domain are given as follows

$$y(n, k) = \sum_{j=1}^{J} x_j(n, k) + v(n, k),$$  

(1)

where the $M \times 1$ vectors $y(n, k)$, $x_j(n, k)$ and $v(n, k)$ contain the complex spectral coefficients of the microphone signals, the $j$-th talker’s signals and the noise signals respectively, and $n$ and $k$ are the time and frequency indices respectively. The speech signals $x_j(n, k)$, for $j = 1, \ldots, J$, and the noise signal $v(n, k)$ represent realizations of zero-mean, mutually uncorrelated random processes.

The signal of a desired talker, denoted by an index $d$, at a reference microphone $m$, can be estimated by linearly combining the $M$ microphone signals as follows

$$\hat{x}_{d,m}(n, k) = h_{d,m}^H(n, k)y(n, k),$$  

(2)

where $h_{d,m}(n, k)$ contains the complex filter coefficients at a TF bin $(n, k)$. The goal in this paper is to compute a filter $h$, which reduces the signals of the interfering talkers and the noise, while preserving the signal of the desired talker $d$. Moreover, the number of talkers $J$ needs to be estimated from the microphone signals.

### B. Second Order Statistics

The SOS required to compute the spatial filters consist of the PSD matrices of the interfering talker signals and the noise, and the relative array propagation vector of the desired talker signal. The PSD matrix of the microphone signals is defined as

$$\Phi_y(n, k) = E[y(n, k)y^H(n, k)],$$

(3)

where $E[\cdot]$ represents the expectation of a random variable. The PSD matrices of $x_j$ and $v$ are defined similarly. Due to the assumption that the different speech signals and the noise signal are mutually uncorrelated, the following relation holds

$$\Phi_y(n, k) = \sum_{j=1}^{J} \Phi_{x_j}(n, k) + \Phi_v(n, k).$$

The array propagation vector for a given source can be obtained from the respective PSD matrix, according to

$$g_{j,m}(n, k) = \frac{\Phi_{x_j}(n, k) e_m}{\phi_{j,m}(n, k)}, \text{ with } e_m = [1 \cdots 0 1 \cdots 0]^T.$$  

(7)

Note that if the source positions are time invariant, the array propagation vectors $g_{j,m}$ do not depend on the time index $n$.

### III. Optimum Linear Filtering

In this section, a brief overview of the MVDR filter, the MWF and the PMWF is provided. Although the three filters arise by optimizing different statistical criteria, they are inherently related to each other [5], [37]. The MWF and the PMWF can be written as an MVDR filter multiplied by a single-channel post filter [38], which uses the temporal variations of the desired and undesired signal PSDs to achieve better noise and interference reduction. For brevity, in the following we omit the microphone, time, and frequency indices wherever possible.

#### A. Minimum Variance Distortionless Response (MVDR) Filter

An MVDR filter is obtained by minimizing the residual undesired signal power, while requiring distortionless response for the signal of the desired talker. To extract a desired talker $d$, the MVDR filter is obtained as the solution of the following optimization problem

$$h_{\text{mwdr},d} = \arg \min_h \Phi_y h, \quad \text{subject to}$$

$$h^H d = 1,$$  

(8)

where $\Phi_y$ denotes the undesired signal PSD matrix obtained as the sum of the PSD matrices of the interfering talker signals and the background noise. Solving the optimization problem leads to the well-known MVDR or Capon beamformer [39] given by

$$h_{\text{mwdr},d} = (\Phi_d \Phi_u^{-1} g_d)^{-1} \Phi_u^{-1} g_d.$$  

(9)

#### B. Multichannel Wiener Filter

The MWF provides an MMSE estimate of the desired signal, by minimizing the following cost function

$$J(h) = E[(X_d - h^H y)^2],$$

(10)

where $X_d$ is the signal of the desired talker at the reference microphone. Setting the derivative with respect to $h$ to zero and solving for $h$, the following expression is obtained

$$h_{\text{mwf},d} = \phi_d \Phi_y^{-1} g_d,$$  

(11)

where $\phi_d = E[X_d^2]$ denotes the PSD of the desired signal. Applying the matrix inversion lemma [40] to $\Phi_y^{-1}$ and rearranging, the MWF can be written as

$$h_{\text{mwf},d} = (\phi_d^{-1} + g_d^H \Phi_u^{-1} g_d)^{-1} \Phi_u^{-1} g_d.$$  

(12)
or as the product of an MVDR and a single-channel post filter as [37]

\[ h_{\text{mvdr,d}} = h_{\text{mvdr,d}} \cdot \frac{g_d H^H \Phi_u g_d}{\beta^{-1} + g_d H^H \Phi_u^{-1} g_d}. \]  

(13)

C. Parametric Multichannel Wiener Filter

The concept of a PMWF has been used for time domain filtering in earlier works including [41], [42], whereas in [5], [37] several equivalent expressions of the frequency domain PMWF are derived. The PMWF is obtained by minimizing the residual noise power, while imposing a constraint on the maximum allowable distortion of the desired signal, as follows [5]

\[ \text{subject to} \quad \mathbb{E} \left[ \left( y - h^H x_d \right)^2 \right] \leq a^2. \]  

(14)

Solving the optimization problem results in the following expression for the PMWF

\[ h_{\text{pmwf,d}} = \arg \min_h h^H \Phi_u h, \quad \text{subject to} \quad \mathbb{E} \left[ \left( y - h^H x_d \right)^2 \right] \leq a^2. \]  

(15)

where \( \beta \) is a parameter that controls the trade-off between distortion of the desired speech signal and reduction of noise and interfering signals. By utilizing the matrix inversion lemma and rearranging, the PMWF can be rewritten as

\[ h_{\text{pmwf,d}} = \left( \beta \Phi_u^{-1} + g_d H^H \Phi_u^{-1} g_d \right)^{-1} \Phi_u^{-1} g_d. \]  

(16)

or as the product of an MVDR and a single channel post-filter as

\[ h_{\text{pmwf,d}} = h_{\text{mvdr,d}} \cdot \frac{g_d H^H \Phi_u^{-1} g_d}{\beta \Phi_u^{-1} + g_d H^H \Phi_u^{-1} g_d}. \]  

(17)

As a part of this contribution, we propose a method to control the trade-off parameter \( \beta \), which will be described in Section VII.

IV. ESTIMATION OF THE SECOND ORDER STATISTICS (SOS)

A crucial factor that determines the quality of the extracted source signals at the spatial filter output is the estimation accuracy of the SOS. The SOS need to be estimated from the microphone signals, without prior information about the source positions, the number of sources, and their activity over time. State-of-the-art approaches for estimation of the SOS of multiple signals from their mixtures involve recursive updates based on which signal is dominant at a particular TF bin. For this purpose, we introduce the following hypotheses

\[ \mathcal{H}_o : y(n,k) = v(n,k), \]  

indicating speech absence, \( 18a \)

\[ \mathcal{H}_j : \text{indicating that the} j\text{-th talker is dominant, i.e} \]  

\[ y(n,k) \approx x_j(n,k) + v(n,k). \]  

(18b)

Consequently, speech presence is indicated by

\[ \mathcal{H}_x = \mathcal{H}_1 \sqcup \mathcal{H}_2 \sqcup \ldots \sqcup \mathcal{H}_d. \]  

(19)

Let \( p(\mathcal{H}_x | y), p(\mathcal{H}_j | y), p(\mathcal{H}_j | y), p(\mathcal{H}_d | y) \) denote the posterior probabilities of the hypotheses after observing the microphone signals. These probabilities can be used to estimate the SOS, as done in several recently proposed source extraction frameworks [10], [11], [18].

A. Computation of the Noise PSD Matrix

The noise PSD matrix at the TF bin \( (n, k) \) is recursively estimated as a weighted sum of the instantaneous noise PSD matrix at TF bin \( (n, k) \) and the noise PSD matrix estimate from TF bin \( (n - 1, k) \). In [7], [8], the weights are computed using the SPP \( p(\mathcal{H}_d | y) \), such that for \( \alpha \leq \alpha_s < 1 \) an averaging parameter is computed as

\[ \alpha_s(n) = p(\mathcal{H}_s | y(n)) + \alpha_s(1 - p(\mathcal{H}_s | y(n))). \]  

(20)

and the noise PSD matrix is recursively estimated according to

\[ \hat{\Phi}(n) = \alpha_s(n) \hat{\Phi}(n-1) + (1 - \alpha_s(n))|y(n)|^2 \]  

(21)

B. Computation of the PSD Matrices of Speech Sources

Similarly, the PSD matrix of each source is recursively estimated using the source posterior probabilities \( p(\mathcal{H}_s | y) \). As the background noise is always present, we introduce the following PSD matrix for each source \( j \)

\[ \Phi_x = \hat{\Phi} - \hat{\Phi}_v. \]  

(22)

which can be recursively estimated as follows

\[ \hat{\Phi}_x = \alpha_s(n) \hat{\Phi}_x(n-1) + (1 - \alpha_s(n))|y(n)|^2 \]  

(23)

where for a chosen constant \( 0 < \alpha_s < 1 \), the averaging parameter \( \alpha_s \) is computed as

\[ \alpha_s(n) = 1 - p(\mathcal{H}_s | y(n)) + \alpha_s p(\mathcal{H}_s | y(n)). \]  

(24)

An important difference between the source PSD matrix estimation and the noise PSD matrix estimation is the fact that prior to performing the recursive update (23) for source \( j \), a classification step takes place, such that at TF bin \( (n, k) \), a PSD matrix update is performed only for the source \( j_{n,k} \) that satisfies

\[ j_{n,k} = \arg \max_j p(\mathcal{H}_s | y(n,k)). \]  

(25)

Finally, the PSD matrix for source \( j \) is computed as

\[ \hat{\Phi}_x(n, k) = \hat{\Phi}_x(n, k) - \hat{\Phi}_v(n, k). \]  

(26)

The remaining task, which contains a part of the main contribution of this work, is to estimate the posterior probabilities in (20) and (24).

V. ESTIMATING POSTERIOR PROBABILITIES

The posterior probability that the \( j \)-th source is dominant, given the current microphone signals \( y \) at a particular TF bin can be decomposed as follows

\[ p(\mathcal{H}_s | y) = p(\mathcal{H}_x | y) \cdot p(\mathcal{H}_s | y, \mathcal{H}_x). \]  

(27)
where we made use of the fact that \( p[H_x | y] = p[H_x, \hat{H}_x | y] \).

Clearly, the first factor in (27) represents the SPP, and the second factor is the probability that the \( j \)-th source is dominant, conditioned on speech presence. In the following, we describe the computation of these two probabilities by using the microphone signals and extracted parametric information for each TF bin.

A. Speech Presence Probability

If the spectral coefficients of the speech and the noise signals are modeled as complex Gaussian vectors, the multichannel SPP was derived in [43] as follows

\[
p[H_x | y(n)] = \left\{ 1 + \frac{q(n)}{1-q(n)} [1 + \xi(n)] e^{-\frac{x(n)}{1+\tau(n)}} \right\}^{-1}, \tag{28}
\]

where \( q(n) = p[H_x(n)] \) denotes the a priori speech absence probability (SAP) and

\[
\xi(n) = \text{tr}\{\Phi^{-1}(n)\Phi_x(n)\}, \tag{29}
\]

\[
x(n) = y(n)^T \Phi^{-1}(n) \Phi_x(n) \Phi^{-1}(n) y(n), \tag{30}
\]

where \( \text{tr}\{\cdot\} \) denotes the trace operator. The speech signal PSD matrix \( \Phi_x(n) \) can be computed as \( \Phi_y(n) - \Phi_B(n) \).

In this paper, we use the DDR-based a priori speech absence probability \( q \) proposed by the present authors in [36]. In this manner, onsets of coherent speech signals are accurately detected and do not leak into the noise PSD matrix estimate. The DDR was computed using the complex coherence between two microphones from an array, as proposed in [44]. Due to the small inter-microphone distances in one array, the DDR is overestimated at low frequencies even in noise-only frames. To detect noise-only frames accurately, each frame is subdivided into two frequency bands and the average DDR for the two bands is computed. Subsequently, a binary mask is computed which is equal to zero if the ratio of DDRs is larger than a threshold, and one otherwise. Eventually, the a priori SAP as computed in [36] is multiplied by the binary mask. As the SPP in this work is computed using distributed arrays, the DDR is computed for each microphone array separately, and the maximum DDR is chosen for the a-priori SAP estimation.

B. Source Posterior Probabilities Conditioned on Speech Presence

In order to estimate the conditional posterior probabilities \( p[H_x | y] \) for \( j = 1, \ldots, J \), position estimates are computed for each TF bin. Using the DMAs, multiple DOAs can be computed per TF bin and triangulated to obtain a position estimate \( \hat{r}(n, k) \). In [18], the present authors proposed the following position-based approximation of the posterior probabilities

\[
p[H_x | y(n, k)] \approx p[H_x | \hat{r}(n, k)], \tag{31}
\]

where it is assumed that the conditional source probabilities \( p[H_x | y(n, k)] \) are completely determined by the source position estimate at TF bin \( (n, k) \). The fullband distribution of \( \hat{r} \) given that speech is present, was modeled by a Gaussian mixture (GM) with \( J \) components as follows

\[
p[\hat{r} | H_x] = \sum_{j=1}^{J} \pi_j \mathcal{N}(\hat{r}; \mu_j, \Sigma_j), \tag{32}
\]

where \( \pi_j \) denote the mixing coefficients and \( \mathcal{N}(\hat{r}; \mu_j, \Sigma_j) \) denotes a Gaussian distribution with mean \( \mu_j \) and covariance matrix \( \Sigma_j \). If the mixture parameters are known, the required conditional source probabilities can be computed as

\[
p[H_x | \hat{r}, H_x] = \sum_{j' = 1}^{J} \pi_j \mathcal{N}(\hat{r}; \mu_{j'}, \Sigma_{j'}). \tag{33}
\]

A ML estimation of mixture parameters using unlabeled data is often done by the EM algorithm. Recently, the EM algorithm has been used in several source extraction frameworks, to cluster spatial cues extracted from the microphone signals [13], [16], [18], [21].

VI. ESTIMATION OF NUMBER OF SOURCES AND GAUSSIAN MIXTURE PARAMETERS

A limitation of the standard ML EM algorithm is that the number of sources needs to be known in advance [10], [11], [13], [14], [16], [18]. In this paper, we propose an ML-based variant of the EM algorithm that jointly detects the number of GM components (sources) and estimates the GM parameters. The algorithm requires a training phase of a short duration, where each talker is active for at least 1-2 seconds, without constraints on the number of simultaneously active talkers. One iteration of the algorithm consists of (i) a standard ML EM iteration, (ii) position-based number of sources estimation, and (iii) training data pruning based on the estimated number of sources. Steps (ii) and (iii) can also be interpreted as a re-initialization of the ML EM algorithm that is based on position-based criteria. By using the SPP explicitly in the M-step, the algorithm is able to cluster the sources even in the presence of background noise. In the rest of this section, we briefly review the concept of tolerance regions of a Gaussian distribution, which are used in deriving position-based re-initialization criteria, and describe the steps of the proposed EM-based algorithm in detail.

A. Tolerance Region of A Gaussian Distribution

A tolerance region of a distribution can be interpreted as a region of minimum volume, centered at \( \mu \), that contains a certain probability mass \( p_{\text{mass}} \). Let us consider a multivariate Gaussian distribution with a mean vector \( \mu \) and covariance matrix \( \Sigma \). A point \( \hat{r} \) belongs to a tolerance region of probability \( \pi \) if the following holds

\[
(\hat{r} - \mu)^T \Psi^{-1}(\hat{r} - \mu) \leq \Psi_{\text{mass}} \tag{34}
\]

where \( \Psi_{\text{mass}} \) depends on \( \psi_{\text{mass}} \) as detailed next. It can be shown [46] that for an \( N \)-dimensional Gaussian distribution, the quadratic form \( (\hat{r} - \mu)^T \Psi^{-1}(\hat{r} - \mu) \) follows a Chi-squared distribution with \( N \) degrees of freedom. For the 2-dimensional (2D) case, the cumulative distribution function of a Chi-squared
Algorithm 1 Number of sources detection and clustering

**Initialization**
1. Select a number of initial Gaussian components in the mixture $j^{(0)}$, corresponding to a maximum number of sources.
2. Initialize the GMM by a K-means clustering [45].

**repeat**
2. Estimate number of components.
   2(a). Removing a Gaussian component.
   2(b). Merging Gaussian components.
3. Prune training data.

**until** the GM parameters difference between two iterations is sufficiently small.

Run a final Mahalanobis distance-based merger.

distribution reduces to an exponential distribution, leading to the following relation between $p_{\text{max}}$ and $\Psi_{\text{max}}$

$$p_{\text{max}} = 1 - e^{-\Psi_{\text{max}}/2}.$$ 

(35)

For a 2D Gaussian distribution, the locus of points defined by (34) represents the interior of an ellipse with center $\mu$ and axes aligned with the eigenvectors of $\Sigma$.

**B. The Steps of the Proposed EM Algorithm**

Algorithm 1 presents a brief outline of the proposed algorithm for number of sources detection and clustering. After an initialization step where the maximum number of Gaussian components $j^{(0)}$ is selected and the means of the clusters are initialized with the K-means algorithm [45], the following steps are repeated until convergence:

1. **Standard EM iteration.** Given a training set of unlabeled position estimates $D = \{\hat{x}_1, \ldots, \hat{x}_N\}$, the GM parameters $P = \{\pi_j, \mu_j, \Sigma_j, \ldots\}$ are found by maximizing the log likelihood

$$\ln p(D|P) = \sum_{n=1}^{N} \ln \sum_{j=1}^{J} \pi_j N(\hat{r}_n; \mu_j, \Sigma_j).$$

(36)

which can be done iteratively, alternating between the E-step and the M-step of the EM algorithm. In the E-step, the posterior probabilities conditioned on speech presence are computed using the current model parameters according to (33), whereas in the M-step the mixture parameters are updated as follows

$$\mu_j \leftarrow \frac{\sum_{n=1}^{N} p(\hat{r}_n | \hat{y}_n, \mu_j, \Sigma_j)}{\sum_{n=1}^{N} p(\hat{r}_n | \hat{y}_n)} \hat{r}_n,$$

(37)

$$\Sigma_j \leftarrow \sum_{n=1}^{N} p(\hat{r}_n | \hat{y}_n, \mu_j, \Sigma_j) (\hat{r}_n - \mu_j)(\hat{r}_n - \mu_j)^T$$

(38)

$$\pi_j \leftarrow \frac{1}{N} \sum_{n=1}^{N} p(\hat{r}_n | \hat{y}_n),$$

(39)

where the posterior probabilities $p(\hat{r}_n | \hat{y}_n, \mu_j, \Sigma_j)$ are estimated as

$$p(\hat{r}_n | \hat{y}_n, \mu_j, \Sigma_j) = \frac{p(\hat{r}_n | \hat{y}_n, \mu_j, \Sigma_j)}{\sum_{j=1}^{J} p(\hat{r}_n | \hat{y}_n, \mu_j, \Sigma_j)},$$

(40)

and $\hat{y}_n$ contains the microphone signals from the TF bin corresponding to the position estimate $\hat{r}_n$.

2. **Estimate the number of sources.** In this step, the position estimates are used to update the number of Gaussian components, by removing components that do not model a source, and by merging components that model the same source.

2(a). Removing Gaussian components. Three empirical criteria $C_1$, $C_2$, and $C_3$ are used to determine if the $j$-th Gaussian component in the mixture of the current iteration models a source. The first criterion is based on the fact that components which do not model a source exhibit a significantly larger variance compared to the ones that model a source. Moreover, due to the initialization of the algorithm with an overestimated number of sources, some of the Gaussian components might model more than one source simultaneously, leading to a large variance. Formally, the variance criterion is given by

$$C_1 : \text{tr} (\Sigma_j) > c_{\text{var}} \cdot \text{tr} (\hat{\Sigma}_j^*),$$

(41)

where $c_{\text{var}}$ is a pre-defined constant which determines the maximum variance that is allowed along the principal axes of a Gaussian component that models a source, and $\hat{\Sigma}_j^*$ is the covariance matrix of the Gaussian component with minimum principal axes variance among all Gaussian components in the current iteration.

The second criterion $C_2$ relates to the condition number of the covariance matrix $\Sigma_j$, and can be computed as the ratio of the largest eigenvalue to the smallest eigenvalue of $\Sigma_j$, where the eigenvalues determine the variances along the two principal axes. Assuming that noise and reverberation are localized randomly in the room, noisy and reverberant position estimates can be modeled by a distribution with a balanced variance along all principal axes. This criterion can be quantified by the condition number of the corresponding covariance matrix. If $\text{cond}\{\cdot\}$ denotes the condition number, components that do not satisfy

$$C_2 : \text{cond}\{\Sigma_j\} < c_{\text{cond}}$$

(42)

are likely to model a speech source. The pre-defined constant $c_{\text{cond}}$ denotes the maximum condition number that is characteristic for a Gaussian component that models noisy or reverberant position estimates.

The third criterion seeks to remove the $j$-th Gaussian component if the component contains the means of at least two other components within a tolerance region defined by a probability $p_u$. Formally, this can be written as follows

$$C_3 : (\mu_j - \mu_j')^T \Sigma_j^{-1} (\mu_j - \mu_j') \leq \Psi_{p_u}$$

(43)

for at least two values of $j'$, where $j' \neq j$.

(44)

where $\Psi_{p_u}$ is computed using $p_u$ and (35).

Finally, the $j$-th Gaussian component is removed if the following statement is true

$$(C_1 \land C_2) \lor C_3,$$

(45)
where \( \land \) and \( \lor \) denote logical conjunction and disjunction. The expression (45) is crucial for robust number of sources estimation: (i) the conjunction \( C_1 \land C_2 \) eliminates sources with high variance (criterion \( C_1 \)), only if the variance is balanced along the principal axes (criterion \( C_2 \)); (ii) the disjunction with \( C_3 \) ensures that a Gaussian component that models more than one source simultaneously is always discarded, provided that each source is already modeled by a separate Gaussian component.

When removing the \( j \)-th Gaussian component, the remaining mixture coefficients need to be re-normalized so that their sum is equal to one. Alternatively, the Mahalanobis distances between the mean of the removed Gaussian component and the means of the remaining Gaussian components can be taken into account, such that the new coefficient \( \pi_{j'} \) of the \( j' \)-th component is computed as

\[
\pi_{j'} \leftarrow \pi_{j'} + \pi_j \frac{(\mu_j - \mu_{j'})^T \Sigma_j^{-1} (\mu_j - \mu_{j'})}{\sum_{k \in \mathcal{G}} (\mu_j - \mu_k)^T \Sigma_k^{-1} (\mu_j - \mu_k)},
\]

where \( \mathcal{G} \) denotes the set of remaining Gaussian components.

2(b). Merging Gaussian components. Components with closely located means are likely to be modeling a single source. Two components \( j \) and \( j' \) are merged if the following holds

\[
||\mu_{j'} - \mu_j|| < c_\mu,
\]

where \( || \cdot || \) denotes the Euclidean norm, and \( c_\mu \) is a pre-defined constant. The \( j \)-th and the \( j' \)-th component are merged to form a single component with the following parameters

\[
\mu_{\text{merge}} = \frac{\mu_j + \mu_{j'}}{2}, \quad \Sigma_{\text{merge}} = \frac{\Sigma_j + \Sigma_{j'}}{2} - \pi_{j'} + \pi_j.
\]

3. Pruning training data. After removing one or more Gaussian components, or merging multiple Gaussian components, certain position estimates from the training set \( \mathcal{D} \) are no longer accurately modeled by the remaining mixture components. If \( p_{\text{miim}} \) denotes a chosen probability mass and \( \Psi_{p_{\text{max}}} \) the associated Mahalanobis distance computed by (35), a position estimate \( \hat{r} \) is removed from the training set \( \mathcal{D} \) if for all components \( j' \in \mathcal{G} \) the following holds

\[
(\hat{r} - \mu_{j'})^T \Sigma_{j'}^{-1} (\hat{r} - \mu_{j'}) \geq \Psi_{p_{\text{max}}},
\]

This means that if a position estimate does not belong to a tolerance interval \( p_{\text{max}} \) of any Gaussian component in the current iteration, it is removed from the training data for the next iteration.

Steps 1-4 are repeated until convergence. The algorithm has converged if the difference of the means and covariance matrices of the GM between two iterations is smaller than a threshold. After convergence, a final Mahalanobis distance-based merging is performed in order to assure that each source is modeled by a single Gaussian component. In particular, two Gaussian components \( j \) and \( j' \) are merged if at least one of the following inequalities is satisfied

\[
(\mu_{j'} - \mu_j)^T \Sigma_j^{-1} (\mu_{j'} - \mu_j) \leq \Psi_{p_{\text{merge}}}, \quad \text{or}
\]

\[
(\mu_j - \mu_{j'})^T \Sigma_{j'}^{-1} (\mu_j - \mu_{j'}) \leq \Psi_{p_{\text{merge}}}.
\]

The proposed algorithm exhibited extremely fast convergence. For the tested scenarios with different reverberation and noise levels, we found that no more than 7 iterations were required.

VII. PROPOSED PMWF TRADEOFF COMPUTATION

In many practical situations involving multiple talkers, the activity of the different talkers changes over time, with periods where certain (or all) talkers are inactive, such as in typical meeting scenarios. Information about the activity of the talkers can be utilized to achieve stronger interference reduction during inactivity of the desired talker. On the other hand, when the desired talker is active, strong interference reduction might result in undesired distortions. As the PMWF offers the possibility of a tradeoff between the noise and interference reduction and the distortion of the desired speech, our goal is to use source posterior probabilities to control the PMWF. We propose a frequency-independent tradeoff parameter, where the source posterior probabilities \( p(T_n | \hat{r}, \hat{s}) \) are used to track the activity of the different talkers. For the \( j \)-th talker, the posterior probabilities from a sliding window of \( L \) frames are used to compute the following activity indicator

\[
p_j(n) = \frac{\sum_{i=n-L}^{n} p(T_i | \hat{r}, \hat{s})}{\sum_{i=n-L}^{n} p(T_i | \hat{r}, \hat{s})},
\]

which attains values between 0 and 1. Finally, \( p_j(n) \) is then mapped to a tradeoff parameter using a sigmoid-like function

\[
\beta_j(n) = \frac{10^\frac{p_j(n) \cdot (l_{\text{u,ax}} - l_{\text{miim}})}{10} + 1}{1 + 10^\frac{p_j(n) \cdot (l_{\text{u,ax}} - l_{\text{miim}})}{10}},
\]

where \( l_{\text{miim}} \) and \( l_{\text{max}} \) are the minimum and maximum values for \( \beta_j \), \( c \) determines a shift of the function along the \( p_j \) (horizontal) axis, and \( \rho \) determines the steepness of transition region of the function. In this work, the parameters were set to \( l_{\text{miim}} = 0, l_{\text{max}} = 7, c = -12 \) and \( \rho = 3 \). In this manner we obtained a function which for \( p_j \geq 0.2 \) results in \( \beta_j = 0 \) (MVD filter), for \( p_j = 0.1 \) results in \( \beta_j = 1 \) (standard MWF) and for \( p_j \in [0, 0.1] \) the tradeoff parameter \( \beta \) rapidly increases to its maximum value \( l_{\text{max}} \), leading to strong noise and interference reduction. Since \( p_j \) is computed using a temporal window over frames, it is important to use \( \beta \approx 0 \) even for low values of the activity indicator \( (p_j \geq 0.2) \). This avoids undesired distortions of the desired signal at the onsets, where the desired signal is only present during a portion of the \( L \) considered frames.

VIII. EXPERIMENTS AND PERFORMANCE EVALUATION

The proposed source extraction framework was evaluated with both simulated and measured data. In the following, the performance measures, the experimental setup, and the evaluation results are presented. Simulated data is used to demonstrate the performance of the proposed EM-based algorithm in environments with different reverberation levels. The extracted signal quality for different spatial filters, different background
noise levels, and different number of sources was evaluated using measured data.

A. Performance Measures

The signal quality at the output of the different spatial filters was evaluated in terms of the following measures:

1) Segmental speech distortion index $\nu_{sd}$, as defined in [4, Eq. 4.44].

2) PESQ score improvement [47], denoted by $\Delta$-PESQ. The $\Delta$-PESQ is computed as the difference of the PESQ score of the inverse STFT of $\hat{X}_{d,m}(n, k)$ and the PESQ score of the mixture $y_m(t)$ received at the reference microphone $m$.

3) Segmental interference reduction (segIR), where the segIR for the $i$-th frame of length $T$ is computed according to

$$\text{segIR}(i) = 10 \log_{10} \left( \frac{\sum_{j \neq d} |x_{j,m}(t)|^2}{\sum_{j \neq d} |\hat{x}_{j,m}(t)|^2}, (i-1)T < t < iT, \right.$$  

where $\{\cdot\}$ denotes the average over all frames, $\{\cdot\}_{d}$ denotes a signal filtered by a filter designed to extract a desired source $d$. The final segIR value is obtained by averaging the segment-wise values $\text{segIR}(i)$.

4) Segmental noise reduction factor

$$\text{segNR}(i) = 10 \log_{10} \left( \frac{|y_m(t)|^2}{|\hat{y}_{m}(t)|^2}, (i-1)T < t < iT. \right.$$  

In the following, we denote the input desired-speech-to-noise ratio (segDSNR) and desired-speech-to-interference (segDSIR) by $S_{d}$ and $S_{x}$, respectively. The respective segment-wise values are given by

$$S_{d}(i) = 10 \log_{10} \left( \frac{|x_{d,m}(t)|^2}{|y_{m}(t)|^2}, (i-1)T < t < iT. \right.$$  

$$S_{x}(i) = 10 \log_{10} \left( \sum_{j \neq d} |x_{j,m}(t)|^2, (i-1)T < t < iT. \right.$$  

All segmental performance measures were computed using non-overlapping frames of 30 ms. For the spatial filters and the performance measures for a given source, an arbitrary microphone from the array nearest to the source was chosen as the reference $m$.

B. Experimental Setup

The sampling frequency for all experiments was 16 kHz and the frame length of the STFT was $L = 512$ samples, with 50% overlap. The smoothing parameter $\alpha_r$ used in the background noise PSD matrix estimation in Section IV-A was set to 0.9. The PSD matrix $\Phi_\alpha$ of the microphone signals was also obtained by recursive averaging with a smoothing parameter equal to 0.9. However, in order to learn the noise PSD matrix more accurately, it was assumed that first second contains only background noise and the update parameter $\alpha_r$ during these frames was set to 0.75. The averaging constant $\alpha_\varphi$ was set to 0.8 for all talkers. Note that due to estimation errors, the PSD matrix estimates $\hat{\Phi}_\alpha(n, k)$ given by (26) might not be positive semi-definite. Positive semi-definiteness can be ensured, for instance, by applying a singular value decomposition (SVD), setting the negative singular values to zero, and applying inverse SVD with the new singular values. The different parameters related to the proposed clustering algorithm in Section VI are summarized in Table I. The given values offered a stable performance in all tested scenarios with mild to moderate reverberation and noise levels.

The simulated microphone signals were obtained by convolving simulated room impulse responses (RIRs) with four different speech signals of approximately equal power. The RIRs in a $3.2 \times 3.3 \times 2.8$ m shoebox room and reverberation $T_{90} = 200$ ms and $T_{90} = 400$ ms were simulated using an efficient implementation of the image source model [48]. To obtain the noisy microphone signals, an ideal diffuse noise component [49] and a spatially uncorrelated noise component with segDSNR of 30 dB were added to the convolved speech signals. Two circular DMAs were used with four microphones each, diameter 3 cm and inter-array distance 1.16 m. Note that the proposed framework does not impose any restriction on the geometry and position of the arrays. However, it is required that the arrays are designed to cover the full angular range of 360 degrees.

The measurements were carried out in a room with $T_{90} \approx 250$ ms and dimensions $3.2 \times 3.3 \times 2.8$ m. Two circular arrays with four DPA miniature microphones each, diameter 3 cm and inter-array distance 1.16 m were used. In order to avoid erroneous DOA estimates in the high frequency range due to spatial aliasing, all signals for the evaluation were bandlimited to 7 kHz. In principle, frequencies above the aliasing frequency can be used and processed if, for instance, the phase wrapping of the DOAs is correctly compensated before the triangulation. An approach to map the DOA estimates above the aliasing frequency.
frequency to the true DOA was proposed in [50], in the context of source separation. The RIRs for each source-microphone pair were measured, where the signals were emitted by GENERELEC loudspeakers arranged in two different setups as shown in Fig. 1(a) and Fig. 1(b). The RIRs for the setup illustrated in Fig. 1(c) were used to generate a diffuse sound, such that a different babble speech signal for each loudspeaker was convolved with the measured RIRs. To ensure that the generated signal is sufficiently diffuse, the first 30 ms of the measured RIRs were set to zero. Finally, the resulting microphone signals were obtained by adding the convolved speech signals, the diffuse signal with a given segDSNR, and the measured sensor noise scaled appropriately to achieve a segDSNR of 30 dB. To evaluate the effect of background noise on the signal extraction performance, segDSNRs of approx. 11.6 dB, 21 dB and 30 dB were considered, where the background noise consists of the diffuse babble speech and the measured sensor noise. In the case with 30 dB, the background noise contains only the sensor noise, without diffuse babble speech.

C. Results

The evaluation results emphasize two main aspects of the proposed framework: 1) number of source detection and clustering in different simulated and measured scenarios, and 2) evaluation of the extracted source signals in terms of objective performance measures. The objective performance evaluation is carried out for each source present in a given scenario and for different spatial filters, i.e., the standard MVDR and MWF filters, as well as the PMWF with the proposed tradeoff parameter.

1) Number of Source Detection and Clustering: The performance was evaluated for different reverberation times, different diffuse noise levels and different number of sources. In all cases, the length of the signal used for training was 2J seconds where J denotes the number of sources. In multi-talk scenarios all sources are active during the training period, whereas in single-talk scenarios each source is active for two seconds. In Fig. 3, the resulting clusters of four sources in simulated environment with two different $T_{60}$ values are illustrated. The clusters in Fig. 3(a) correspond to a training done during single-talk, whereas the clusters in Fig. 3(b) correspond to a training done during constant multi-talk of all sources. Although this is a challenging scenario, where the sources are less likely to be sparse in the TF domain, the algorithm successfully detects the number of sources and the respective clusters for both $T_{60}$ values.

The clustering results with the measured RIRs are shown in Fig. 4 and Fig. 5. We considered input segDSNRs of $S_n = 30$ dB and $S_n = 11.5$ dB. For all scenarios, the clustering was performed during single-talk (Fig. 4) and during multi-talk (Fig. 5). The algorithm was tested with four sources, corresponding to setup 1 (see Fig. 1), with three sources corresponding to setup 2, and with two sources corresponding to setup 1 with only two of the four sources active. The results demonstrate that the clustering algorithm is robust to low to moderate background noise levels in moderately reverberant environments. As expected, the performance deteriorates when training is done during multitalk, where the errors in cluster orientation are more significant as the number of sources increases. Nevertheless, the number of sources and the source locations are estimated in all cases with a good accuracy, with maximum error in the estimated source position of 35 cm for source 3 from setup 2 [see Fig 1(b)] at $T_{60} = 400$ ms and multi-talk training. It can be observed that the sensitivity of the cluster orientation and cluster center estimation depends on the relative position of the source with respect to the DMAs.

2) Objective Performance Evaluation of Extracted Signals: In order to evaluate the objective quality of the extracted source signals in different scenarios, the following experiments were performed.
TABLE II
INPUT SEGDSNR IN dB FOR THE FOUR SOURCES IN A PARTIAL MULTI-TALK SCENARIO (LEFT) AND A CONSTANT MULTI-TALK SCENARIO (RIGHT)

<table>
<thead>
<tr>
<th></th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
<th></th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.7</td>
<td>0.6</td>
<td>-5.3</td>
<td>2.2</td>
<td></td>
<td>-1.3</td>
<td>-3.9</td>
<td>-10.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

1) Compare the performance when using all the available microphone signals from the DMAs to the performance when using the microphones from only one array. In the latter case, for each source the array closer to the estimated source location (the mean of the respective Gaussian component) is chosen.

2) Examine how the accuracy of the estimated clustering algorithm affects the objective quality of the extracted source signals at the output of a spatial filter. For this purpose different scenarios were evaluated with training done during multi-talk, and training during single-talk. The corresponding clusters for the evaluated scenarios were shown in Fig. 4 and Fig. 5.

3) Compare the performance of the MVDR, the MWF, and the proposed PMWF.

The experiments were done with signals that contain background noise, periods of multi-talk, and periods of single-talk. In the first two experiments, different number of sources were considered, and in the third experiment, only a four-source scenario was considered. The activity of the sources over time and the segDSNRs \( S_s \) for all sources in the different scenarios are given in Fig. 2 and Table II.

In Table III, comparison of spatial filtering with one versus spatial filtering with two arrays is presented for different segDSNRs. The results are obtained by averaging over all talkers in the respective scenario. The average segDSNR (averaged over the talkers) \( S_s = 0.3 \) dB. As expected, spatial filtering with two arrays achieves superior interference reduction, for the different noise levels, and on average 4 dB, 2.2 dB, and 3.2 dB for the four, three and two sources scenarios, respectively. Two arrays score better in terms of PESQ as well and achieve better diffuse and sensor noise reduction in all cases. However, as the number of sources increases, the segmental SD index \( \nu_{sd} \) is lower when spatial filtering is done with one array. The performance gain when using two arrays instead of one depends on the geometry and the relative source-array positions.

In Table IV, the signal quality at the output of the MVDR filter is compared when training is done in single-talk versus multi-talk. We consider only the segDSIR and the segmental SD index \( \nu_{sd} \), which are averaged over all talkers in the respective scenario. The segDSNR and the PESQ scores followed the trend of the segDSIR. The difference in extracted signal quality for single-talk versus multi-talk training becomes more significant as the number of simultaneously active sources increases and the sparsity assumption is more likely to be violated. An advantage of the algorithm is that the training setup does not have a significant effect on the segDSIR even for the multi-talk training of four sources. The SD index \( \nu_{sd} \) on the other hand, is more sensitive to errors in the cluster estimation. Interestingly, for the scenario with two sources, the MVDR filter applied after a multi-talk training achieves lower speech distortion (SD) index than the same filter applied after a single-talk training. This observation can be explained by the fact that two simultaneously active sources are sparse in the TF domain and the cluster means and orientations are accurately estimated. In addition, the source clusters estimated during multi-talk have higher variance and therefore the source PSD matrices computed using the related source posterior probabilities will capture more of the desired signal energy, resulting in lower SD index than in the single-talk case.

Finally, the different filters were compared for all performance measures and a scenario with four sources, as illustrated in Fig. 1(a). The results averaged over the four different sources, with single-talk training and source activity as shown in Fig. 2, are given in Fig. 6. In terms of the SD index, the MVDR filter...
Fig. 6. Objective performance evaluation of different spatial filters for the four source scenario in Fig. 1(a), for three different noise levels. The results are averaged over the four sources (a) Speech distortion index $r_{s, a}$ (b) PESQ score improvement (c) Interference reduction (segIR) (d) Noise reduction (segNR).

Fig. 7. Objective performance evaluation of the spatial filters for the four source scenario in Fig. 1(a). Input segDSNR $\delta_s = 30$ dB (a) Speech distortion index $r_{s, a}$ (b) PESQ score improvement (c) Interference reduction (segIR) (d) Noise reduction (segNR).

Fig. 8. Objective performance evaluation of the spatial filters for a four source scenario in Fig. 1(a). The four sources are simultaneously active at all times (also during training). Input segDSNR $\delta_s = 30$ dB (a) Speech distortion index $r_{s, a}$ (b) PESQ score improvement (c) Interference reduction (segIR) (d) Noise reduction (segNR).

TABLE IV

| Interference Reduction and Desired Speech Distortion Comparison for MVDR Filtering Based on a Single-Talk Training Versus MVDR Filtering Based on Multi-Talk Training, for Three Different Diffuse Noise Levels: 30 dB (Top), 21 dB (Middle), and 11.5 dB (Bottom). Spatial Filtering is Performed Using All Available Microphones. Average Input SegDSIR Ratio $\delta_s = 0.3$ dB |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 2 talkers       | 3 talkers       | 4 talkers       |                 |                 |                 |
|                 | $\mu_{\text{segIR}}$ | $\nu_{\text{segIR}}$ | $\mu_{\text{segIR}}$ | $\nu_{\text{segIR}}$ | $\mu_{\text{segIR}}$ | $\nu_{\text{segIR}}$ |
| single          | 15.0            | 0.055           | 10.9            | 0.076           | 12.5            | 0.069           |
| multi           | 14.4            | 0.046           | 9.9             | 0.082           | 11.4            | 0.085           |
| single          | 15.0            | 0.053           | 11.1            | 0.078           | 12.1            | 0.059           |
| multi           | 14.4            | 0.047           | 10.1            | 0.084           | 11.2            | 0.080           |
| single          | 14.9            | 0.083           | 9.9             | 0.087           | 11.6            | 0.084           |
| multi           | 14.9            | 0.058           | 9.5             | 0.076           | 10.7            | 0.096           |

achieves the best performance with SD index lower than 0.1 in all cases, whereas the MWF results in SD index between 0.1 and 0.2. Due to the PMWF trade-off parameter that does not distort the signal even at quite low probability of desired source activity, as proposed in Section VII, the PMWF approaches the low SD index of the MVDR. In terms of PESQ, all filters show similar performance. The segDSIR and segDSNR, were computed separately over segments where the desired source is present and segments where the desired source is absent. As expected, in most cases the filters achieve better noise and interference reduction during periods where the desired source is silent, where the performance difference is clearly most significant for the proposed PMWF filter. During periods when the desired talker is not active, the PMWF reduces up to 12 dB more interference and up to 7 dB more background noise as compared to periods where the desired talker is active. The results also demonstrate that the interference reduction performance of the algorithm is not affected by the background noise level, at least for the considered low to moderate segDSNRs. Furthermore, to demonstrate that all sources are successfully extracted, the performance measures for each source separately
at $S_n = 30$ dB are shown in Fig. 7. Comparing the results for the different sources, it is confirmed that in contrast to the segDSIR and the segDSNR which are robust to errors in the cluster estimation, the SD index is more sensitive. To clarify, we can make the following observations based on the clustering results for this particular scenario: in Fig. 4, the clusters for $\text{source}_1$ and $\text{source}_2$ [see Fig. 1(a) for the source labeling] are accurately estimated and have comparable variances in both dimensions; the cluster of $\text{source}_3$ exhibits significant variance in only one direction, making it more sensitive to errors and hence underestimation of the source probability; the mean of the cluster associated with $\text{source}_4$ is estimated with error of 35 cm. This explains why the SD index is higher for $\text{source}_3$ and $\text{source}_4$ than for $\text{source}_1$ and $\text{source}_2$. Note that the sensitivity of the clustering algorithm and its effect on the spatial filtering performance can be significantly reduced by incorporating more than two DMAs for triangulation and position estimation.

Finally, to demonstrate a worst case scenario, we considered a constant multi-talk with four sources simultaneously active both during training and during the whole evaluated segment. The results are shown in Fig. 8. Note that in this case, the input $S_m$ differs from the previous scenario and is much lower for all sources (see Table II). The results demonstrate that even in this adverse case where the sparsity assumption is likely to be violated, the spatial filters are able to extract the source signals with a good quality. The largest performance drop is observed for the SD index, which reaches 0.6 for $\text{source}_4$. The PESQ improvement of 0.7 points on average is similar to the previous scenario where the improvement was 0.8 points on average. Note that as all sources are active at all times, the segDSIR and segDSNR need to be compared to the respective values in Fig. 7 where the target source is active. Notably, even in the challenging multi-talk scenario, there is not a significant performance deterioration in terms of segDSIR and segDSNR.

IX. CONCLUSIONS

We developed an informed spatial filtering framework for source extraction in the presence of interfering coherent sources and background noise. The work was based on a recently proposed probabilistic approach for SOS estimation, followed by a spatial filtering using the MVDR, the MWF, and the PMWF. Bin-wise position information extracted from distributed microphone arrays was used to cluster the sources in the TF domain, and estimate the SPP and the source probabilities. An efficient EM-based clustering algorithm was proposed that simultaneously detects the number of sources and clusters them using a very small number of iterations. Moreover, we proposed a PMWF with a fullband probabilistic source activity detection-based tradeoff parameter. Comprehensive performance evaluation with both simulated data and measured data demonstrated the applicability of the framework for source clustering and source extraction for different number of sources, different background noise levels, different training conditions and different spatial filters. It was shown that the framework extracts the signals with good quality even in adverse multi-talk environments.

REFERENCES


Maja Taseska (‘S13) was born in 1988 in Ohrid, Macedonia. She received her B.Sc. degree in electrical engineering at the Jacobs University, Bremen, Germany, in 2010, and her M.Sc. degree at the Friedrich-Alexander-Universität, Erlangen, Germany in 2012. She then joined the International Audio Laboratories Erlangen, where she is currently pursuing a Ph.D. in the field of informed spatial filtering.

Her current research interests include informed spatial filtering, source localization and tracking, and noise reduction.

Emanuël A. P. Habets (‘S02–M’07–SM’11) received his B.Sc degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and his M.Sc and Ph.D. degrees in electrical engineering from the Technische Universität Eindhoven, The Netherlands, in 2002 and 2007, respectively. From March 2007 until February 2009, he was a Postdoctoral Fellow at the Technion—Israel Institute of Technology and at the Bar-Ilan University in Ramat-Gan, Israel. From February 2009 until November 2010, he was a Research Fellow in the Communication and Signal Processing group at Imperial College London, United Kingdom. Since November 2010, he is an Associate Professor at the International Audio Laboratories Erlangen (a joint institution of the University of Erlangen and Fraunhofer IIS) and Head of the Spatial Audio Research Group at Fraunhofer IIS, Germany.

His research interests center around audio and acoustic signal processing, and he has worked in particular on dereverberation, noise estimation and reduction, echo reduction, system identification and equalization, source localization and tracking, and crosstalk cancellation.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, a general co-chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York, and general co-chair of the 2014 International Conference on Spatial Audio (ICSA) in Erlangen, Germany. He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2011–2013) and a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013–2015). Since 2013 he is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.