LOCALIZATION OF AN UNKNOWN NUMBER OF SPEAKERS IN ADVERSE ACOUSTIC CONDITIONS USING RELIABILITY INFORMATION AND DIARIZATION

Andreas Brendel[1]*, Bracha Laufer-Goldshtein*[2]*, Sharon Gannot*[2]*, Ronen Talmon*[3]*, and Walter Kellermann*[1]

[1]Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 7, D-91058 Erlangen, Germany, Andreas.Brendel@FAU.de,
[2]Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel
[3]Viterbi Faculty of Electrical Engineering, Technion, Haifa 3200003, Israel

ABSTRACT

This paper investigates localization of an arbitrary number of simultaneously active speakers in an acoustic enclosure. We propose an algorithm capable of estimating the number of speakers, using reliability information to obtain robust estimation results in adverse acoustic scenarios and estimating individual probability distributions describing the position of each speaker using convex geometry tools. To this end, we start from an established algorithm for localization of acoustic sources based on the EM algorithm. There, the estimation of the number of sources as well as the handling of reverberation has not been addressed sufficiently. We show improvement in the localization of a higher number of sources and in the robustness in adverse conditions including interference from competing speakers, reverberation and noise.

Index Terms— Acoustic source localization, number of speakers estimation, diarization, EM algorithm

1. INTRODUCTION

Acoustic source localization is an essential task for many signal processing applications, e.g., steering a beamformer or a camera to the position of interest. If multiple distributed sensor nodes, i.e., an Acoustic Sensor Network (ASN), are available, source localization algorithms can benefit from the different perspectives on the acoustic scene provided by the distributed nodes. This additional information fueled the research in the field of acoustic source localization, and a variety of algorithms for localization in ASNs have been proposed. Most of them are based on triangulation of node-wise Direction of Arrival (DOA) estimates, e.g., [1], [2]. If the environment is fixed, e.g., a smart home scenario, learning-based localization approaches [3], [4] can be used, which are trained based on acoustic features labeled by the corresponding positions in the room. Subsequently, the trained algorithm is capable of source localization based on unlabeled data and unseen positions. Examples for this class of algorithms are neural network-based methods [5], [6], manifold learning approaches using Relative Transfer Functions (RTFs) as features [7] or the distributed learning-based approach of [8] relying on the Coherent-to-Diffuse power Ratio (CDR). However, these learning-based localization algorithms rely on labeled training data, gathered in a specific environment. In [9], a method which relies on a model-based EM source separation and localization (MESSL), was presented. In this method, STFT-based features are used to estimate the Time Difference of Arrival (TDOA) of multiple speakers by using Gaussian Mixture Model (GMM) clustering. To this end, a set of candidate source TDOAs is created where each of them is associated with a single Gaussian component. The estimated source TDOA is given by the candidate associated with the highest probability. The MESSL algorithm for TDOA estimation has been extended to a multichannel version in [10] and to support reverberant and noisy environments [11, 12], where [12] directly uses the raw data as features. For localizing and tracking acoustic sources in an ASN, the ideas of the MESSL algorithm have been exploited in [13], opening the door for multiple applications and extensions: a distributed version of the algorithm has been proposed in [14], and the GMM-based model has been replaced by a mixture of von Mises distributions in [15] to take the directionality of the feature into account.

In this paper, we propose an algorithm extending [13] by addressing three of its prominent shortcomings. Firstly, reverberation and additive noise, which invariably impair the observed signals in any enclosure, render many time frequency bins useless for the localization task by introducing a bias to the estimate. This has been addressed by a model of the human hearing system in [16]. However, there are a lot of algorithmic parameters to be chosen for this model. Secondly, the estimation of an unknown number of speakers introduces further challenges, and currently, it does not have a satisfactory solution: the number of sources is either assumed to be known or determined by thresholding the probability map [14]. However, the optimum threshold depends on the number of speakers as well as on the reverberation and noise level, and is hence difficult to set. Thirdly, the estimation of the positions of multiple speakers [14, 15] requires the extraction of local maxima from the estimated probability map. Alternatively, multiple maps, each corresponding to only one speaker, could be estimated [9, 13], but the performance of this algorithm highly depends on an appropriate initialization of the individual maps using coarse prior knowledge about the position of the speakers [13], which is usually not available. To counteract the deteriorating effect of reverberation, we propose to incorporate reliability information from acoustic scene analysis by controlling the influence of each single observation on the estimate into the proposed algorithm. This is accomplished by applying a weighted Expectation Maximization (EM) algorithm [17]. To this end, the CDR [18] is used as a measure of degradation of the bin-wise observation by reverberation. The CDR has been used for STFT bin-selection and weighting for localization algorithms relying on the sparsity of speech signal mixtures in [19, 20]. For localizing an unknown number of speakers, we pre-estimate the number of speakers and their probability of dominance over time using methods of convex geometry [21, 22]. The obtained dominance probabilities are exploited

This work was supported by DFG under contract no <Ke890/10-1> within the Research Unit FOR2457 "Acoustic Sensor Networks"

by the proposed probabilistic model to obtain separate probability maps for each speaker. This especially improves the localization performance for a larger number of sources in the acoustic scene.

2. PROPOSED EM-BASED LOCALIZATION ALGORITHM

We consider here a fully synchronized ASN containing M sensor nodes, each equipped with two microphones observing S sources at unknown positions. In the following, $t \in \{1, \dots, T\}$ and $k \in \{1, \dots, K\}$ denote the time and frequency indices, and $i \in \{1, 2\}$ and $m \in \{1, \dots, M\}$ denote microphone and node indices, respectively. We model the ith microphone signal at the mth sensor node in the STFT domain as

$$x_m^i(t,k) = \sum_{s=1}^{S} h_{sm}^i(t,k) x_s(t,k) + v_m^i(t,k), \tag{1}$$

where h_{sm}^i denotes the relative transfer function corresponding to source $s \in \{1, \dots, S\}$ describing the acoustic path between the ith microphone at the mth sensor node and the first microphone of the first sensor node, which serves as the reference. The image of source s observed at the reference microphone is denoted by $x_s(t,k)$ and the additive noise at the ith microphone of the mth sensor node is denoted by v_m^i. We use the Pair-wise Relative Phase ratios (PRPs) of each sensor node m

$$\phi_m(t,k) = \frac{x_m^2(t,k)}{x_m^1(t,k)} \left| \frac{x_m^1(t,k)}{x_m^2(t,k)} \right| \tag{2}$$

as a feature and stack them in a vector $\boldsymbol{\phi}(t,k) = \text{vec}_m\, \phi_m(t,k)$. We assume W-disjoint orthogonality, i.e., each STFT bin is dominated by a single speaker [23] and hence $\boldsymbol{\phi}(t,k)$ describes a single speaker.

For each position $\mathbf{p}$ from a finite set of candidate target positions $\mathcal{P}$, the candidate PRPs are approximated by the free-field model

$$\tilde{\phi}_m^k(\mathbf{p}) = \exp\left(-j \frac{2\pi f_k}{c} \left(\|\mathbf{p} - \mathbf{p}_m^1\|_2 - \|\mathbf{p} - \mathbf{p}_m^2\|_2 \right) \right), \tag{3}$$

where $\mathbf{p}_m^i$ denotes the position of the ith microphone at the mth sensor node, c is the speed of sound, and f_k is the physical frequency corresponding to STFT frequency index k. The PRPs corresponding to the different sensor nodes are collected in a vector for concise notation $\tilde{\boldsymbol{\phi}}^k(\mathbf{p}) = \text{vec}_m\, \tilde{\phi}_m^k(\mathbf{p})$. By using the complex Gaussian Probability Density Function (PDF)

$$\mathcal{N}^c\left(\phi_m(t,k) | \tilde{\phi}_m^k(\mathbf{p}), \sigma^2 \right) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2}{\sigma^2} \right) \tag{4}$$

and assuming independence of the observations over time, frequency and nodes, the data-likelihood can be modeled as the following GMM with class probabilities $\psi_{\mathbf{p}}$

$$\prod_{t,k=1}^{T,K} \sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} \prod_{m=1}^{M} \mathcal{N}^c\left(\phi_m(t,k) | \tilde{\phi}_m^k(\mathbf{p}), \frac{\sigma^2}{w_m(t,k)} \right). \tag{5}$$

Note that we use the same variance σ^2 for all Gaussian components here for simplicity. Here, compared to [13, 14, 15], we introduced the weighting factor $w_m(t,k)$ which favors reliable observations by assigning a large weight to them [17]. This weighting strategy can be interpreted as increasing the variance of all Gaussians for erroneous estimates corresponding to small weights, i.e., an observation with small weight contributes to all positions in the probability map, and as it does not favor or penalize a certain position, it does not affect the localization results.

However, this model does not distinguish between the contributions of the individual sources. To address this issue, we introduce probabilities of source dominance $P_s(t)$ at frame t and incorporate them into the model as convex weights of GMMs representing the individual sources

$$p(\boldsymbol{\phi}; \boldsymbol{\psi}, \Theta) = \prod_{t,k=1}^{T,K} \sum_{s=1}^{S} P_s(t) \cdots \tag{6}$$
$$\cdots \sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p},s} \prod_{m=1}^{M} \mathcal{N}^c\left(\phi_m(t,k) | \tilde{\phi}_m^k(\mathbf{p}), \frac{\sigma^2}{w_m(t,k)} \right).$$

Here, we introduced the abbreviation $\boldsymbol{\psi} = \text{vec}_{\mathbf{p},s}\, \psi_{\mathbf{p},s}$ and the set of fixed parameters $\Theta = \{\mathbf{w}, \mathbf{P}\}$, with $\mathbf{w} = \text{vec}_{t,k,m}\, w_m(t,k)$ and $\mathbf{P} = \text{vec}_{t,s}\, P_s(t)$. The corresponding maximum likelihood parameter estimation problem is not straightforwardly solvable in closed-form. Hence, the latent random indicator variable

$$z(t,k,\mathbf{p},s) = \begin{cases} 1 & \text{if observation } (t,k) \text{ corresponds to} \\ & \quad \text{source } s \text{ at position } \mathbf{p} \\ 0 & \text{else} \end{cases} \tag{7}$$

is introduced which allows to use an EM algorithm for this task. The auxiliary function, i.e., the posterior mean of the complete data log-likelihood, is calculated with $\mathbf{z} = \text{vec}_{t,k,\mathbf{p},s}\, z(t,k,\mathbf{p},s)$ as

$$\mathcal{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(l-1)}) = \mathcal{E}\left\{ \log p(\mathbf{z}, \boldsymbol{\phi}; \boldsymbol{\psi}, \Theta) | \boldsymbol{\phi}; \boldsymbol{\psi}^{(l-1)}, \Theta \right\} \tag{8}$$
$$= \sum_{t,k,s=1}^{T,K,S} \sum_{\mathbf{p}\in\mathcal{P}} \mathcal{E}\left\{ z(t,k,\mathbf{p},s) | \boldsymbol{\phi}; \boldsymbol{\psi}^{(l-1)}, \Theta \right\} \Bigg[\log \psi_{\mathbf{p},s} + \dots$$
$$\dots \log P_s(t) + \sum_{m=1}^{M} \log \mathcal{N}^c\left(\phi_m(t,k) | \tilde{\phi}_m^k(\mathbf{p}), \frac{\sigma^2}{w_m(t,k)} \right) \Bigg],$$

with l representing the iteration index and $\mathcal{E}\{\cdot\}$ the expectation operator. The E step is obtained by evaluating the posterior expectation of the latent variable

$$\gamma^{(l)}(t,k,\mathbf{p},s) = \mathcal{E}\left\{ z(t,k,\mathbf{p},s) | \boldsymbol{\phi}; \boldsymbol{\psi}^{(l-1)}, \Theta \right\} \tag{9}$$
$$= \frac{P_s(t) \psi_{\mathbf{p},s}^{(l-1)} \prod_{m=1}^{M} \mathcal{N}^c\left(\phi_m(t,k) | \tilde{\phi}_m^k(\mathbf{p}), \frac{\sigma^2}{w_m(t,k)} \right)}{\sum_{\mathbf{p},s} P_s(t) \psi_{\mathbf{p},s}^{(l-1)} \prod_{m=1}^{M} \mathcal{N}^c\left(\phi_m(t,k) | \tilde{\phi}_m^k(\mathbf{p}), \frac{\sigma^2}{w_m(t,k)} \right)}.$$

The M step is derived by optimizing the auxiliary function w.r.t. $\psi_{\mathbf{p},s}$. The estimator for the class probabilities of the Gaussian components corresponding to the existence of source s at position $\mathbf{p}$ is given by

$$\psi_{\mathbf{p},s}^{(l)} = \frac{1}{T \cdot K} \sum_{t,k=1}^{T,K} \gamma^{(l)}(t,k,\mathbf{p},s). \tag{10}$$

Finally, the source positions are estimated by determining the maximum of each probability map, i.e., by evaluating the different class probabilities for each source s

$$\hat{\mathbf{p}}_s = \underset{\mathbf{p}\in\mathcal{P}}{\text{argmax}} \psi_{\mathbf{p},s}^{(L)}, \tag{11}$$

where L is the maximum number of iterations.

3. CALCULATION OF OBSERVATIONS RELIABILITY

In Section 2, weights $w_m(t,k)$ have been introduced in (5) to account for the varying reliability of the observations. This raises the

question of how to quantify the reliability of observations? Here, we assume that the source signal consists of a coherent signal component, corresponding to the direct path and early reflections, and a diffuse signal component, corresponding to the late reverberation and noise. The power ratio of these signal components is called CDR [18] and has been used for dereverberation [18], localization [8], and bin-selection and weighting in DOA estimation [19, 20]. Note that the proposed weighting strategy is general and the weights $w_m(t,k)$ can be chosen differently, depending on the available information of the observed signals.

In order to estimate the CDR, the auto- and cross- Power Spectral Density (PSD) of the microphone signals i, j at node m are estimated by recursive averaging over time

$$\hat{\Phi}_{x_m^i x_m^j}(t,k) = \mu \hat{\Phi}_{x_m^i x_m^j}(t-1,k) + (1-\mu) x_m^i(t,k)(x_m^j)^*(t,k), \tag{12}$$

where $i, j \in \{1,2\}$ and $0 < \mu < 1$ is a forgetting factor. The complex spatial coherence of the observed microphone signals is then estimated by

$$\hat{\Gamma}_x^{(m)}(t,k) = \frac{\hat{\Phi}_{x_m^1 x_m^2}(t,k)}{\sqrt{\hat{\Phi}_{x_m^1 x_m^1}(t,k)\hat{\Phi}_{x_m^2 x_m^2}(t,k)}}. \tag{13}$$

The coherence of a diffuse sound field can be expressed in closed form by $\Gamma_n^{(m)}(k) = \mathrm{sinc}\,(2 f_k d_{\mathrm{mic},m}/c)$, where $\mathrm{sinc}(\cdot)$ denotes the normalized sinc function and $d_{\mathrm{mic},m}$ denotes the spacing of the microphones at node m. Here, we use the DOA-independent CDR estimator [13], which is a function of $\hat{\Gamma}_x^{(m)}(t,k)$ and $\Gamma_n^{(m)}(k)$. Finally, the CDR-based weight is expressed as

$$w_m(t,k) = \max\left(\frac{\mathrm{CDR}_m(t,k)}{\mathrm{CDR}_m(t,k)+1}, \epsilon\right) \tag{14}$$

to obtain values limited to $[0,1]$ for convenience. Hereby, ϵ is small positive number avoiding division by zero in (9).

4. CALCULATION OF SOURCE DOMINANCE PROBABILITIES

To estimate the number of sources and the frame-wise speaker probabilities, we shortly summarize a recently proposed approach for speaker diarization based on spectral analysis of the correlation matrix of consecutive time frames [21, 22]. As we assume W-disjoint orthogonality, the ratio of the STFT bins of the ith microphone at the mth node w.r.t. the corresponding bins of the first microphone at the first node is assumed to correspond to an RTF of one of the speakers $R_m^i(t,k) = \frac{x_m^i(t,k)}{x_1^i(t,k)}$. Based on this observation, the following real-valued feature vector $\mathbf{a}(t)$ of length $D = 2(2M-1)K$ is computed for all time frames

$$\mathbf{a}_m^i(t) = \left[R_m^i(t,1), \ldots, R_m^i(t,K)\right]^{\mathrm{T}},$$

$$\mathbf{a}^c(t) = \left[\mathbf{a}_1^2(t), \mathbf{a}_2^1(t), \mathbf{a}_2^2(t) \ldots, \mathbf{a}_M^1(t), \mathbf{a}_M^2(t)\right]^{\mathrm{T}},$$

$$\mathbf{a}(t) = \left[\mathrm{Re}\{\mathbf{a}^c(t)\}^{\mathrm{T}}, \mathrm{Im}\{\mathbf{a}^c(t)\}^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{15}$$

We assume that there exist a few frames with only one speaker active, which is a valid assumption for typical human speakers who have some natural pauses while talking. The aim of this algorithm is to estimate these frame-wise speaker dominance probabilities $\mathbf{P}(t) = [P_1(t), \ldots, P_S(t)]^{\mathrm{T}}$, which occupy the standard probability simplex as they have to sum up to one. To this end, the $T \times T$ correlation matrix $\mathbf{C}$ with elements $C_{tt'} = \frac{1}{D}\mathcal{E}\{\mathbf{a}^{\mathrm{T}}(t)\mathbf{a}(t')\}$ is approximated as $\mathbf{C} \approx \mathbf{B}\mathbf{B}^{\mathrm{T}}$ [22], where $\mathbf{B}$ is a $T \times S$ matrix containing the source dominance probabilities, i.e., $B_{tj} = P_j(t)$. Therefore, the rank of the correlation matrix $\mathbf{C}$ is equal to the rank of $\mathbf{B}$ and hence to the number of speakers S [21]. The correlation matrix $\mathbf{C}$ is estimated by the approximation $\hat{C}_{tt'} = \frac{1}{D}\mathbf{a}^{\mathrm{T}}(t)\mathbf{a}(t')$ (which is statistically justified in [21]). The eigenvalue decomposition of the estimated correlation matrix $\hat{\mathbf{C}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$ yields an orthonormal matrix $\mathbf{U}$ containing its eigenvectors $\mathbf{u}_j$ and a diagonal matrix $\mathbf{\Lambda}$ containing its eigenvalues λ_j. The number of sources can now be estimated by [21]

$$\hat{S} = \left(\underset{j}{\mathrm{argmin}} \frac{\lambda_j}{\lambda_1} < \alpha\right) - 1, \tag{16}$$

where λ_1 denotes the largest eigenvalue. Each time frame t can be represented by a slice of the extracted and sorted eigenvectors as $\boldsymbol{\nu}(t) = [u_1(t), \ldots, u_{\hat{S}}(t)]^{\mathrm{T}}$. These $\boldsymbol{\nu}(t)$ vectors occupy a rotated and scaled simplex, whose S vertices $\{\boldsymbol{\nu}(t_s)\}_{s=1}^{S}$ can be computed with the successive projection algorithm [24]. To finally obtain the source probabilities per time frame $\mathbf{P}(t)$, we construct the back-transformation matrix from the identified vertices $\hat{\mathbf{Q}} = [\boldsymbol{\nu}(t_1), \ldots, \boldsymbol{\nu}(t_{\hat{S}})]^{\mathrm{T}}$, and rotate and scale $\boldsymbol{\nu}(t)\ \forall t$ back to the standard simplex to obtain the source probabilities

$$\hat{\mathbf{P}}(t) = \hat{\mathbf{Q}}^{-1}\boldsymbol{\nu}(t). \tag{17}$$

The complete localization algorithm is summarized in Alg. 1.

Algorithm 1 Localization of Unknown Number of Speakers

INPUT: Microphone signals $x_m^i(t,k), \forall m,t,k,i$
Compute PRPs $\phi_m(t,k)$ by (2) and candidate PRPs by (3)
Compute weights $w_m(t,k)$ according to (14)
Estimate source number S using (16) and speaker probabilities $P_s(t)\ \forall s,t$ according to (17)
Initialize probability maps $\psi_{\mathbf{p},s}^{(0)} = \frac{1}{\hat{S}|\mathcal{P}|}\ \forall s, \mathbf{p} \in \mathcal{P}$
for $l = 1$ **to** L **do**
  E step: Estimate soft-assignment of observations to sources and positions $\gamma^{(l)}(t,k,\mathbf{p},s)\ \forall t,k,s,\mathbf{p} \in \mathcal{P}$ by (9)
  M step: Estimate probability maps $\psi_{\mathbf{p},s}^{(l)}\ \forall s, \mathbf{p} \in \mathcal{P}$ via (10)
end for
Estimate source positions by (11)
OUTPUT: Position estimates $\hat{\mathbf{p}}_s\ \forall s \in \{1, \ldots, \hat{S}\}$

5. EVALUATION

We carried out experiments in a simulated enclosure [25, 26] of dimensions $6\,\mathrm{m} \times 6\,\mathrm{m} \times 3.1\,\mathrm{m}$ accompanying $M = 12$ distributed microphone pairs (distance between them $\approx 1\,\mathrm{m}$) of spacing $d_{\mathrm{mic}} = 0.2\,\mathrm{m}$. All sources and microphones lie in the same plane of $1\,\mathrm{m}$ height. We created $J = 25$ sets of Room Impulse Responses (RIRs) for each acoustic condition corresponding to random positions in the room with a minimum spatial separation of $0.5\,\mathrm{m}$ and a distance from the walls of $1\,\mathrm{m}$. The corresponding simulated RIRs are convolved with speech signals of $10\,\mathrm{sec}$ duration at a sampling rate of $f_s = 16\,\mathrm{kHz}$. We tested speech mixtures according to the speaker activities depicted in Fig. 1 which represents a situation of a natural conversation (for $S = 1$, the speaker is continuously active). The speech signals were transformed into the STFT domain by a von Hann window of $50\,\mathrm{ms}$ length and frame shift of $10\,\mathrm{ms}$. We only used the frequency interval $[600\,\mathrm{Hz}, 1000\,\mathrm{Hz}]$ as it contains the most relevant part of the speech signal spectrum. The estimation of

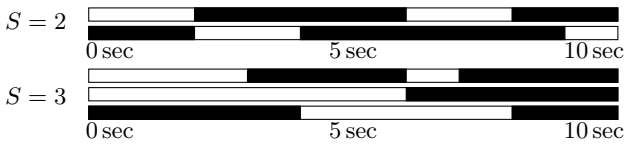

Fig. 1. Temporal activity of the speakers for $S \in \{2, 3\}$. For $S = 1$, the speaker is constantly active.

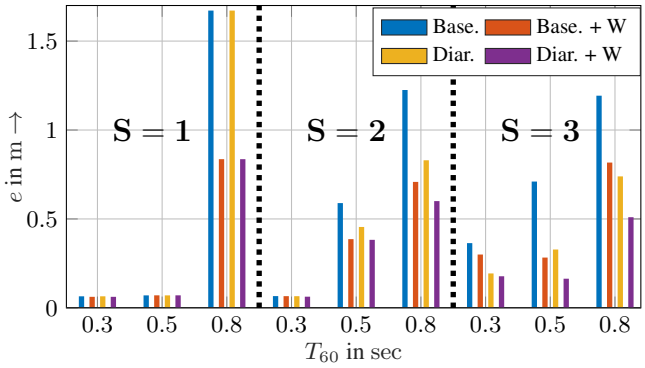

Fig. 2. Average localization error e for varying reverberation times and number of sources S. The underlying scenario is noise-free.

the variance σ^2 of the model (6) yields always the same value for a given grid resolution and is hence set to a fixed value $\sigma^2 = 1$. To enhance the assignment of the speakers to individual maps, we applied a threshold to the speakers probability such that $P_s(t)$ is set to zero if $P_s(t) < 0.5$, i.e., speaker s only contributes to the corresponding probability map if it dominates time frame t. To maintain a valid probabilistic model, we normalized appropriately. All probability maps were initialized with $\psi^{(0)}_{\mathbf{p},s} = \frac{1}{\hat{S}|\mathcal{P}|}\ \forall s, \mathbf{p} \in \mathcal{P}$, where $|\cdot|$ denotes the cardinality of a set, as no prior knowledge about the source positions is assumed. The forgetting factor for the estimation of the PSDs in (12) was set to $\mu = 0.75$. The threshold for estimating the number of speakers was set to $\alpha = 0.4$. The maximum number of iterations for the EM algorithm has been set to $L = 10$. A grid of 41×41 candidate source positions corresponding to a resolution of 10 cm was used. For the algorithm presented in [13], the association of the estimated source position with the ground truth was determined in a greedy fashion by minimizing the overall estimation error and subtracting the corresponding peaks from the probability map to avoid spurious detections. Note that here some algorithmic tuning is necessary, which is unnecessary in the proposed method. For a meaningful comparison and to avoid the problem of quantifying the error when a wrong number of sources S was estimated, we evaluated the estimation of the number of sources and the estimation of the position separately. The estimation of the number of sources is quantified by misdetection (MD) and false alarm (FA) rates, i.e., the relative amount of undetected and wrongly detected sources. To assess the localization performance of the algorithm, we compute the position error averaged over $J = 25$ trials and $S \in \{1, 2, 3\}$ source signals $e = \frac{1}{J \cdot S} \sum_{j=1}^{J} \sum_{s=1}^{S} \|\mathbf{p}_{s,j} - \hat{\mathbf{p}}_{s,j}\|_2$, where $\mathbf{p}_{s,j}$ denotes the position of the sth source in trial j and $\hat{\mathbf{p}}_{s,j}$ is its estimate. To exclude the influence of particular source signals, we picked randomly S out of a set of 7 available source signals in each trial.

To assess the contribution of both the source dominance probabilities and the weights $w_m(t, k)$, we evaluate four different versions of the algorithm: algorithm [14] as a baseline (Base.), the baseline algorithm with incorporated weighting (Base. + W), and the proposed algorithm using diarization with (Diar. + W) and without weighting

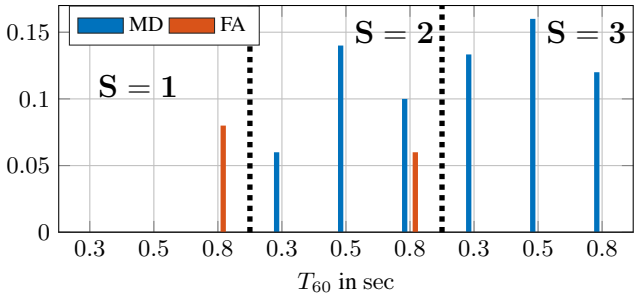

Fig. 3. Misdetection and false alarm rates in a noise-free scenario.

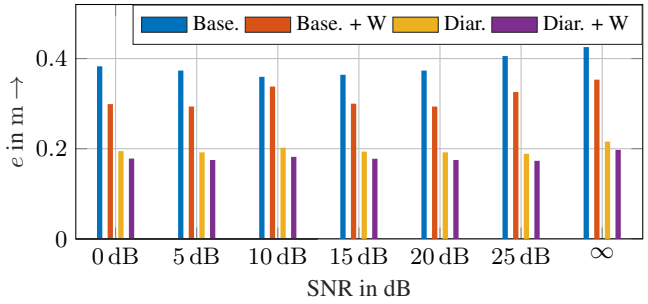

Fig. 4. Average localization error for $S = 3$ and $T_{60} = 0.3$ sec for varying SNR conditions.

(Diar.) included. For the comparison of the baseline algorithm with other state-of-the-art methods, we refer to [13, 14] and do not repeat this comparison here. The average localization errors for noise-free scenarios for varying reverberation time (T_{60}) are depicted in Fig. 2. As the T_{60} and the number of speakers increase, the advantage of the proposed extensions becomes more pronounced, where the combination of diarization and weighting (Diar. + W) always performs best. The corresponding performance in estimating the number of sources is presented in Fig. 3. It can be observed that the task becomes more difficult for increasing reverberation time and number of speakers, yet the source counting error is still very low even in adverse scenarios. Finally, the localization error for varying noise levels at $T_{60} = 0.3$ sec and for $S = 3$ is depicted in Fig. 4. While the localization error is almost not affected by the investigated levels of white additive Gaussian noise, the same trend can be observed for all SNRs: the baseline method performed worst, followed by its weighted version. The diarization approaches performed best, where the weighting yielded further improvement. Note that due to the random selection of source signals the results in the presented experiments vary additionally.

6. CONCLUSIONS

In this paper, we extended an established family of localization algorithms to cope with adverse acoustic scenarios including varying reverberation times, noise levels, and an unknown number of target sources. We incorporated reliability measures to weigh the observations depending on their contribution to the position estimate. A recently proposed technique for speaker diarization based on convex geometry was applied to estimate the number of speakers and to thereby allow for the estimation of individual probability distributions for each source, which circumvents error-prone peak detection strategies. The efficacy of the proposed algorithm has been demonstrated in simulations for various realistic scenarios especially as the number of speakers or the reverberation level increases.

7. REFERENCES

[1] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 54–67, Feb. 2015.

[2] A. Brendel and W. Kellermann, "Localization of multiple simultaneously active sources in acoustic sensor networks using ADP," in *IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Process. (CAMSAP)*, Curacao, Netherlands Antilles, Dec. 2017, pp. 1–5.

[3] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 2814–2818.

[4] A. Saxena and A.Y. Ng, "Learning sound location from a single microphone," in *IEEE Int. Conf. on Robotics and Automation*, Kobe, Japan, May 2009, pp. 1737–1742.

[5] A. Shareef, Y. Zhu, and M. Musavi, "Localization Using Neural Networks in Wireless Sensor Networks," in *Proceedings of the 1st International Conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications*, Innsbruck, Austria, 2008.

[6] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates," *Sensors*, vol. 18, no. 10, Oct. 2018.

[7] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-Supervised Source Localization on Multiple Manifolds With Distributed Microphones," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 25, no. 7, pp. 1477–1491, July 2017.

[8] A. Brendel and W. Kellermann, "Learning-based Acoustic Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," in *European Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sept. 2018.

[9] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[10] M. I. Mandel and J. Barker, "Multichannel Spatial Clustering for Robust Far-Field Automatic Speech Recognition in Mismatched Conditions," in *Proc. of Interspeech*, San Francisco, CA, USA, Sept. 2016, pp. 1991–1995.

[11] O. Schwartz, Y. Dorfan, E. A. P. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *Proc. of the 15th Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sept. 2016, pp. 1–5.

[12] O. Schwartz, Y. Dorfan, M. Taseska, E. A. P. Habets, and S. Gannot, "DOA Estimation in Noisy Environment with Unknown Noise Power using the EM Algorithm," in *Hands-free Speech Commun. and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Mar. 2017.

[13] O. Schwartz and S. Gannot, "Speaker Tracking Using Recursive EM Algorithms," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[14] Y. Dorfan and S. Gannot, "Tree-Based Recursive Expectation-Maximization Algorithm for Localization of Acoustic Sources," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, no. 10, pp. 1692–1703, Oct. 2015.

[15] A. Brendel, S. Gannot, and W. Kellermann, "Localization of Multiple Simultaneously Active Speakers in an Acoustic Sensor Network," in *IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, Sheffield, UK, 2018.

[16] Y. Dorfan, A. Plinge, G. Hazan, and S. Gannot, "Distributed Expectation-Maximization Algorithm for Speaker Localization in Reverberant Environments," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 26, no. 3, pp. 682–695, Mar. 2018.

[17] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 38, no. 12, pp. 2402–2415, Dec. 2016.

[18] A. Schwarz and W. Kellermann, "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, no. 6, June 2015.

[19] A. Brendel, C. Huang, and W. Kellermann, "STFT Bin Selection for Localization Algorithms based on the Sparsity of Speech Signal Spectra," in *EURONOISE*, Crete, Greece, May 2018.

[20] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *IEEE Workshop on Applicat. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[21] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Diarization and Separation Based on a Data-Driven Simplex," in *European Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Oct. 2018.

[22] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *accepted for puclication in IEEE Transactions on Signal Processing*, Feb. 2018.

[23] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. on Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[24] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, July 2001.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustic Soc. of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[26] E. A. P. Habets, "Room Impulse Response Generator," Tech. Rep., Int. Audio Laboratories, Sept. 2010.