Multiple Speaker Localization using Mixture of Gaussian Model with Manifold-based Centroids

Avital Bross, Bracha Laufer-Goldshtein and Sharon Gannot
Faculty of Electrical Engineering, Bar-Ilan University, Ramat-Gan, Israel
{avital.bross,bracha.laufer,sharon.gannot}@biu.ac.il

***Abstract*—A data-driven approach for multiple speakers localization in reverberant enclosures is presented. The approach combines semi-supervised learning on multiple manifolds with unsupervised maximum likelihood estimation. The relative transfer functions (RTFs) are used in both stages of the proposed algorithm as feature vectors, which are known to be related to source positions. The microphone positions are not known. In the training stage, a nonlinear, manifold-based, mapping between RTFs and source locations is inferred using single-speaker utterances. The inference procedure utilizes two RTF datasets: A small set of RTFs with their associated position labels; and a large set of unlabelled RTFs. This mapping is used to generate a dense grid of localized sources that serve as the centroids of a Mixture of Gaussians (MoG) model, used in the test stage of the algorithm to cluster RTFs extracted from multiple-speakers utterances. Clustering is applied by applying the expectation-maximization (EM) procedure that relies on the sparsity and intermittency of the speech signals. A preliminary experimental study, with either two or three overlapping speakers in various reverberation levels, demonstrates that the proposed scheme achieves high localization accuracy compared to a baseline method using a simpler propagation model.**

***Index Terms*—Manifold-learning, semi-supervised inference, mixture of Gaussians**

I. Introduction

Speaker localization is an essential component in various audio applications, e.g. automated camera steering and teleconferencing systems, speaker separation [1] and robot audition [2]. The problem of localizing (and tracking) speakers has therefore attracted the attention of the research community for more than two decades. Localizing speakers "in the wild", namely in scenarios characterized by noise, reverberation and multiple competing speakers, is still a challenge. A recent special issue in IEEE Selected Topics in Signal Processing [3] was dedicated to audio source localization in real-life scenarios. Moreover, a recently introduced community-wide challenge enables fair comparison between various methods using a common dataset [4]. In the current contribution we will focus on the problem of multiple concurrent speaker localization in reverberant environment. We will restrict the discussion to static scenarios.

Recent years have witnessed a change of paradigm in the localization literature, with the introduction of learning-based methods, many of these are based on supervised learning of deep neural networks (DNNs) [5]–[8]. Time-frequency masking is used in [9] to circumvent the need to train the network with spatial information. Weak supervision employing ranking-loss is proposed in [10] for reducing the requirements for labelling the acoustic data.

Each of the above learning-based methods use different spatial features as the input to the network. In this contribution, we will use the RTF, as it is known to provide a meaningful acoustic "fingerprint" uniquely characterizing the source position [11]. We have also demonstrated that the collection of RTFs pertain to a low-dimensional acoustic manifold, which intrinsic degrees of freedom (DoF) are limited to a small number of variables, namely that the acoustic manifold is smooth [12]. In a fixed environment and microphone constellation, these acoustic responses intrinsically differ only by the source position, and can hence be used to infer a nonlinear mapping from an RTF to source position. This observation led to the introduction of a dual microphone semi-supervised localization scheme based on Gaussian process regression [13]. This concept was later extended to multiple microphone pairs [14]. The use of multiple features in the Gaussian process regression was proposed in [15].

While these schemes outperform state-of-the-art methods, they are still limited to the localization of a single speaker. In this paper, we extend the manifold-based approach to the case of multiple speakers with overlapping activity. Our proposed approach utilizes [14] to generate a dense grid of localized RTFs using a small amount of labelled data and a large set of unlabelled acoustic features. Following [16], we show that these localized RTFs can serve as the centroids of a MoG model instead of the original centroids, which are only taking into account the direct sound propagation. Simulation study demonstrates that the proposed method outperforms the baseline method [16], especially in high reverberation levels.

II. Problem formulation

Consider an array of M microphone pairs. The microphone signals in the short-time Fourier transform (STFT) domain are given by $z_{m,j}(t,k) = \sum_{s=1}^{S} a_{m,j}(t,k,\mathbf{q}_s)v_s(t,k)+n_{m,j}(t,k)$ with $j = 1,2$ and $m = 1,\ldots,M$. The speech signals emanating from speaker $s = 1,\ldots,S$ located at position $\mathbf{q}_s$ are denoted $v_s(t,k)$. The number of speakers $S \geq 1$ is assumed to be a priory known. The acoustic transfer functions (ATFs) $a_{m,j}(t,k,\mathbf{q}_s)$ are describing the propagation from $\mathbf{q}_s$

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245; and the Israeli Innovation Authority through KAMIN Project No. 61916. Avital Bross is also funded by grant for advancement of woman in science and technology of the Israeli Ministry of Science and Technology.

to microphone j at pair m, and $n_{m,j}(t,k)$ are the additive noise signals as received by the microphones. The time index is $t = 0, \ldots, T-1$ and the frequency bin index is k. These measurements can be recast in terms of the relative transfer functions (RTFs):

$$z_{m,2}(t,k) = \sum_{s=1}^{S} h_m(t,k,\mathbf{q}_s) z_{m,1}(t,k) + \tilde{n}_{m,2}(t,k) \quad (1)$$

with the first microphone $z_{m,1}(t,k)$ of each node serving as the reference microphone and $\tilde{n}_{m,2}(t,k)$ a noise term related to the noise signals at the microphones. The RTF $h_m(t,k,\mathbf{q}_s)$ is defined as [17]:

$$h_m(t,k,\mathbf{q}_s) = \frac{a_{m,2}(t,k,\mathbf{q}_s)}{a_{m,1}(t,k,\mathbf{q}_s)}. \quad (2)$$

We further assume a static scenario, namely $h_m(t,k,\mathbf{q}_s) = h_m(k,\mathbf{q}_s)$ and define a frequency concatenated vector:

$$\mathbf{h}^m(\mathbf{q}_s) = [h_m(k_1,\mathbf{q}_s), \ldots, h_m(k_F,\mathbf{q}_s)]^\top \quad (3)$$

with $k_1, \ldots, k_F$ a pre-defined frequency range of length $K = k_F - k_1 + 1$, where reliable RTF estimates can be expected. Finally, let

$$\mathbf{h}(\mathbf{q}_s) = \left[\mathbf{h}^{1\top}(\mathbf{q}_s), \ldots, \mathbf{h}^{M\top}(\mathbf{q}_s)\right]^\top \quad (4)$$

denote the aggregated RTF (aRTF), which is a concatenation of the RTF vectors of all pair of microphones. The RTF and the aRTF vectors in (3) and (4), respectively, are known to provide meaningful *fingerprinting* for the acoustic environment and are particularly useful for source localization [11], [12].

III. A COMBINED SEMI-SUPERVISED AND UNSUPERVISED LOCALIZATION SCHEME

In this section we present a method for localizing sources with overlapping activity. The method consists of two stages. In the first, training stage, a sparse grid of RTFs with *known* locations and a denser grid of unlabeled RTFs is generated. These labeled and unlabeled points are jointly used to localize a dense grid of RTFs by applying the multiple-manifold Gaussian process (MMGP) algorithm [14]. The outcome of this training stage is a dense grid of RTFs with associated positions (although the location information obtained by the localizer may not be perfectly accurate). In the second stage, the actual localization of multiple (concurrent sources) is carried out, by employing the MoG model [16], with the previously localized dense grid as the Gaussians' centroids.

A. The Feature Vectors

We have already defined in (3) the RTF-based feature vector (at the mth node), $\mathbf{h}^m(\mathbf{q}_s)$, which is associated with speaker $s \in \{1, \ldots, S\}$, located at position $\mathbf{q}_s$. Since in our scenario multiple speakers can be concurrently active, we do not have access to the separated sources and hence cannot estimate the RTFs using the entire utterance. Instead, we define the instantaneous RTF (iRTF), which was shown to facilitate tracking of a single acoustic source [18], and in our case will facilitate localization of arbitrary activity patterns of multiple overlapping speakers.

The iRTF at node m and time-frequency bin (t,k) is estimated as:

$$\hat{h}_m(t,k) = \frac{\frac{1}{2L+1}\sum_{i=t-L}^{t+L} z_{m,1}(i,k)\cdot z^*_{m,2}(i,k)}{\frac{1}{2L+1}\sum_{i=t-L}^{t+L} |z_{m,1}(i,k)|^2}, \quad (5)$$

where the denominator and numerator are the power spectral density (PSD) and the cross-PSD (cPSD) estimates at node m, respectively, using Bartlett method with $L \geq 0$ to robustify the PSD estimation accuracy. If only a single (static) source is present in the scene, L can be increased to cover the entire speech utterance. In that case, $\hat{h}_m(t,k)$ in (5) will converge to the RTF $h_m(t,k,\mathbf{q}_s)$ in (2) of the active speaker located in $\mathbf{q}_s$. If more than one source is active in the scene, i.e. $S > 1$, then L should be kept small enough to capture the activity of all sources. In the training phase only one source is active, therefore we use $L \gg 1$. In the test phase when concurrent speaker activity is assumed we use a small value for L.

Similarly to the RTFs and aRTF defined in (2) and (4), respectively, their respective instantaneous estimate can now be defined:

$$\hat{\mathbf{h}}^m(t) = \left[\hat{h}_m(t,k_1), \ldots, \hat{h}_m(t,k_F)\right]^\top \quad (6a)$$

$$\hat{\mathbf{h}}(t) = \left[\hat{\mathbf{h}}^{1\top}(t), \ldots, \hat{\mathbf{h}}^{M\top}(t)\right]^\top. \quad (6b)$$

Utilizing the W-disjoint orthogonality (WDO) property of speech signals [19] and assuming high signal-to-noise ratio (SNR), the frequency bins of the iRTF are assumed to be dominated by at most a single speaker RTF (see also [20]). This property will be the basis of the MoG clustering approach, discussed in Sec. III-C. Note that although we focus on static scenarios, i.e. $h_m(t,k,\mathbf{q}_s) = h_m(k,\mathbf{q}_s)$, the iRTF is time-varying to enable capturing the speakers' intermittent activities.

B. Manifold-based Grid Generation

In the training stage we use the MMGP algorithm [14] to generate a dense grid of positions associated with RTFs. The MMGP algorithm starts with a sparse grid of labelled points, namely RTF with associated positions. As measuring positions in a room is a tedious task, only a small number of labeled points is assumed to be available. The role of the MMGP algorithm is therefore to generate a much denser grid of points. For that, it will use many utterances of speech signals from random positions in the region of interest (RoI), without measuring their precise positions. The algorithm in [14] is only capable of localizing a single source and will therefore require a subsequent localization stage that will be applied in the test phase.

For localizing a source, we first define a mapping function, associated with the mth node, $f^m : \mathcal{M}_m \to \mathbb{R}$, which maps the ith RTF sample $\mathbf{h}^m_i \in \mathcal{M}_m$ to the corresponding source position, namely $p^m_i \equiv f^m(\mathbf{h}^m_i)$. This mapping is independently applied for each Cartesian coordinate. The

coordinate index is omitted for brevity. Note that although the position of the source does not depend on the specific node, the notation p^m is used to express that the mapping is obtained from the *point of view* of the mth node. We also assume that the mapping function $f^m(\cdot)$ obeys a Gaussian process.

We further define the scalar function $f : \cup_{m=1}^{M} \mathcal{M}_m \to \mathbb{R}$ which attaches an aRTF sample $\mathbf{h}_i$ with the (coordinate of the) source position, $p_i \equiv f(\mathbf{h}_i)$. To fuse the different perspectives presented by the different nodes, we define the multiple-manifold Gaussian process (MMGP) p_i as the mean of the Gaussian processes of all the nodes, i.e. each position p_i drawn from this process, is given by:

$$p_i = \frac{1}{M}\left(p_i^1 + p_i^2 + \ldots + p_i^M\right). \tag{7}$$

Due to the assumption that the processes are jointly Gaussian, the process p_i is also Gaussian with zero-mean and covariance $\tilde{k}$:

$$p_i \sim \mathcal{GP}(0, \tilde{k}). \tag{8}$$

The covariance between two positions p_r and p_l is given by:

$$\text{cov}(p_r, p_l) \equiv \tilde{k}(\mathbf{h}_r, \mathbf{h}_l) = \frac{1}{M^2}\sum_{i=1}^{n}\sum_{q,w=1}^{M} k_q(\mathbf{h}_r^q, \mathbf{h}_i^q) k_w(\mathbf{h}_l^w, \mathbf{h}_i^w). \tag{9}$$

where $n = n_L + n_U$ is the total number of training points including n_L labelled points and n_U unlabelled points.

In (9), k_m is a standard pairwise function $k_m : \mathcal{M}_m \times \mathcal{M}_m \longrightarrow \mathbb{R}$, often termed "kernel function". A common choice is the Gaussian kernel, with a scaling factor ε_m:

$$k_m(\mathbf{h}_i^m, \mathbf{h}_j^m) = \exp\left\{-\frac{\|\mathbf{h}_i^m - \mathbf{h}_j^m\|^2}{\varepsilon_m}\right\}. \tag{10}$$

The definition of the covariance in (9), induces a new type of manifold-based kernel that takes into account all training points and all points of view of the different nodes.

Define the set of measured positions $\mathcal{P}_L = \{\bar{p}_i\}_{i=1}^{n_L}$ of the labelled set arising from a noisy observation model, given by:

$$\bar{p}_i = p_i + \eta_i;\ i = 1, \ldots, n_L \tag{11}$$

where $\eta_i \sim \mathcal{N}(0, \sigma_p^2)$ $i = 1, \ldots, n_L$ are i.i.d. Gaussian noises, independent of p_i. The noise in (11) reflects uncertainties due to imperfect measurements of the source positions while acquiring the labelled set. Note that since the Gaussian variables p_i and η_i are independent, they are jointly Gaussian. Consequently, p_i and $\bar{p}_i$ are also jointly Gaussian.

To localize the position of a new test RTF $\mathbf{h}_t$ of source from an unknown position, we propose an estimator based on the posterior probability $\Pr(p_t = f(\mathbf{h}_t)|\mathcal{P}_L, \mathcal{H}_L, \mathcal{H}_U)$, where $\mathcal{H}_L, \mathcal{H}_U$ are the set of labelled and unlabelled RTFs, respectively. According to (11) and (8), the function value at the test point p_t and the concatenation of all labelled training positions $\bar{\mathbf{p}}_L = [\bar{p}_1, \ldots, \bar{p}_{n_L}]^\top$ are jointly Gaussian, with probability density function (p.d.f.):

$$\begin{bmatrix}\bar{\mathbf{p}}_L \\ p_t\end{bmatrix}\Bigg|\mathcal{H}_L, \mathcal{H}_U \sim \mathcal{N}\left(\mathbf{0}_{n_L+1}, \begin{bmatrix}\tilde{\mathbf{\Sigma}}_L + \sigma_{\mathbf{p}}^2\mathbf{I}_{n_L} & \tilde{\mathbf{\Sigma}}_{Lt} \\ \tilde{\mathbf{\Sigma}}_{Lt}^\top & \tilde{\Sigma}_t\end{bmatrix}\right) \tag{12}$$

where $\tilde{\mathbf{\Sigma}}_L$ is an $n_L \times n_L$ covariance matrix defined over the function values at the labelled samples in $\mathcal{H}_L$, $\tilde{\mathbf{\Sigma}}_{Lt}$ is an $n_L \times 1$ covariance vector between the function values of the labelled RTFs in $\mathcal{H}_L$ and p_t, $\tilde{\Sigma}_t$ is the variance of p_t, and $\mathbf{I}_{n_L}$ is the $n_L \times n_L$ identity matrix. This implies that the conditional distribution $\Pr(p_t|\bar{\mathbf{p}}_L, \mathcal{H}_L, \mathcal{H}_U)$ is a multivariate Gaussian with the following mean and variance:

$$\hat{p}_t = \tilde{\mathbf{\Sigma}}_{Lt}^\top\left(\tilde{\mathbf{\Sigma}}_L + \sigma_p^2\mathbf{I}_{n_L}\right)^{-1}\bar{\mathbf{p}}_L \tag{13a}$$

$$\text{Var}(\hat{p}_t) = \tilde{\Sigma}_t - \tilde{\mathbf{\Sigma}}_{Lt}^\top\left(\tilde{\mathbf{\Sigma}}_L + \sigma_p^2\mathbf{I}_{n_L}\right)^{-1}\tilde{\mathbf{\Sigma}}_{Lt}. \tag{13b}$$

Although the unlabelled samples do not appear explicitly in (13a),(13b), they do take a role in the computation of the correlation terms, as implied by (9). In fact, the unlabelled samples are an essential component in inferring the manifold structure, hence facilitating a more accurate representation of the high-dimensional RTF samples. The training procedure can be applied to any point in the region of interest. In our implementation it is used to localize all RTFs in the unlabelled set $\mathcal{H}_U$, resulting in a set of associated positions $\mathcal{P}_U$.

The entire MMGP algorithm for a single source localization is summarized in Algorithm 1.

Input : Microphone signals from a single source $z_{m,i}(t,k)$; $m = 1, \ldots, M$, $i = \{1, 2\}$
Data: Set of labelled RTFs in $\mathcal{H}_L$ associated with a set of (noisy) positions $\mathcal{P}_L$; Set of unlabelled RTFs in $\mathcal{H}_U$; Test RTF $\mathbf{h}_t$
Output: Position estimate $\hat{p}_t$ associated with $\mathbf{h}_t$

1 Estimate RTFs using (5),(6a),(6b) with $L \gg 1$
2 Calculate multi-manifold covariance matrix entries $\tilde{\mathbf{\Sigma}}_{lr} = \text{cov}(p_r, p_l)$ using (9)
3 Localize the source using (13a): $\hat{p}_t = \tilde{\mathbf{\Sigma}}_{Lt}^\top\left(\tilde{\mathbf{\Sigma}}_L + \sigma_p^2\mathbf{I}_{n_L}\right)^{-1}\bar{\mathbf{p}}_L$

Algorithm 1: MMGP algorithm for single source localization. The algorithm can be applied to any point in the RoI, including all $\mathbf{h}_t \in \mathcal{H}_U$, resulting in a set of associated positions $\mathcal{P}_U$.

C. Multiple Sources Localization using MoG Clustering

In the test phase, we wish to localize multiple sources with overlapping activity patterns. According to the WDO property of the speech signals [19], even if multiple sources are concurrently active, each of the time-frequency (TF) bins of the received microphone signals in the STFT domain is dominated by only a single source. This property can be utilized to cluster feature vectors using any unsupervised clustering approach, e.g. by maximizing the likelihood of the parameters of a MoG model. Following [1], [16], we do not estimate the centroids of the Gaussians in the MoG model, but rather assume that they are set in advance to a predefined grid points. While originally [16], the grid points were set according to a regular grid of positions, here we use the set of

RTFs that were localized in the training stage. We therefore define the set of grid points as all $h_m(t,k,\mathbf{p})$, $m=1,\ldots,M$ in the labelled and unlabelled sets with $\mathbf{p}\in\mathcal{P}=\mathcal{P}_L\cup\mathcal{P}_U$. As these RTFs are already localized in the training stage, they can be organized according to their position, and thus serve as the centroids of Gaussians. Assuming independence across the TF bins and between nodes, the p.d.f. of the entire utterance can be written as:

$$f(\hat{\mathbf{h}}(t=1,\ldots,T)) = \prod_{t=1}^{T}\prod_{k=k_1}^{k_F}\ldots \sum_{\mathbf{p}\in\mathcal{P}}\psi_{\mathbf{p}}\prod_{m=1}^{M}\mathcal{N}^c(\hat{h}_m(t,k);h_m(t,k,\mathbf{p}),\sigma^2) \quad (14)$$

with

$$\mathcal{N}^c(\hat{h}_m(t,k);h_m(t,k,\mathbf{p}),\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|\hat{h}_m(t,k)-h_m(t,k,\mathbf{p})|^2}{\sigma^2}\right). \quad (15)$$

Note the difference between the labelled and unlabelled points. While the former are accurately localized with small measurement noise σ_p^2, the latter are only known up to the localization accuracy of the training stage, as given in (13b). For simplicity, we neglect this difference and assume fixed variance for all Gaussians and that these localized RTFs can indeed serve as grid points in the clustering scheme. Furhermore, this variance σ^2 is not estimated by the algorithm but rather set empirically.

Now, the EM procedure can be straightforwardly applied to estimate the parameters of MoG model. As all other parameters are set in advance, the only parameters to be determined are $\psi_{\mathbf{p}}$, $\mathbf{p}\in\mathcal{P}$.

For applying the EM procedure we define the hidden data as $x(t,k,\mathbf{p})$, the indicator function associating each TF bin of the iRTF $\hat{h}_m(t,k)$ to one of the grid positions $h_m(t,k,\mathbf{p})$. The E-step at iteration $\ell=1,\ldots,L$ then estimates a soft association of a TF bin to a grid point:

$$\mu^{(\ell)}(t,k,\mathbf{p}) \triangleq E\left\{x(t,k,\mathbf{p})|\hat{\mathbf{h}}(t);\hat{\psi}_{\mathbf{p}}^{(\ell-1)}\right\} = \frac{\hat{\psi}_{\mathbf{p}}^{(\ell-1)}\prod_m\mathcal{N}^c(\hat{h}_m(t,k);h_m(t,k,\mathbf{p}),\sigma^2)}{\sum_{\mathbf{p}}\hat{\psi}_{\mathbf{p}}^{(\ell-1)}\prod_m\mathcal{N}^c(\hat{h}_m(t,k);h_m(t,k,\mathbf{p}),\sigma^2)} \quad (16)$$

and the M-step provides an estimate the MoG weights:

$$\hat{\psi}_{\mathbf{p}}^{(\ell)} = \frac{\sum_{t=1}^{T}\sum_{k=k1}^{k_F}\mu^{(\ell)}(t,k,\mathbf{p})}{T\cdot K}. \quad (17)$$

As the number of active sources in the scene are assumed to be known in advance, their position can be estimated by finding the S highest peaks of the weight map after the last EM iteration:

$$\mathbf{p}_s = \underset{\mathbf{p}}{\operatorname{argmax}}\,\psi_{\mathbf{p}}^{(L)},\ s=1,\ldots,S. \quad (18)$$

The weight map is uniformly initialized in the RoI, namely $\psi_{\mathbf{p}}^{(0)}=\frac{1}{|\mathcal{P}|}$. An alternative method to utilize the intermittent activity of the speakers is presented in [21]. The method incorporates a diarization stage and a data reliability measure to improve the MoG-based clustering.

Note that in the original unsupervised localization scheme in [16], the feature vectors are selected as the pair-wise relative phase ratios (PRPs), namely $\frac{z_{m,1}(t,k)}{z_{m,2}(t,k)}\cdot\frac{|z_{m,2}(t,k)|}{|z_{m,1}(t,k)|}$ and the Gaussian centrods as

$$\exp\left(-j\frac{2\pi k}{K}\frac{\cdot(||\mathbf{p}-\mathbf{p}_{m,2}||-||\mathbf{p}-\mathbf{p}_{m,1}||)}{c\cdot T_s}\right). \quad (19)$$

IV. Experimental Results

A. Setup

1) Signal and Room Parameters: A room with dimensions $5.2\times6.2\times3.5$ m was simulated. Defining the left lower corner of the room as the origin of the coordinate system, the RoI is defined as $[2,4]\times[2,4]$ m in both the x and y axes and 1.5 m in the z-axis. Eight pairs of microphones with microphone inter-distance of 0.2 m were placed in the perimeter of the RoI. The speech measurements were simulated by convolving speech utterances drawn from the TIMIT corpuswith impulse responses simulated by a room impulse response (RIR) generator.[1] A spatially and spectrally white Gaussian noise (WGN) was added to all microphone signals with SNR=20 dB. The signals were analyzed by an STFT with 1024 frequency bins and 75% overlap between frames. Only bins corresponding to the frequency range $150-1500$ Hz were considered.

2) Algorithm Parameters: In the training phase, the algorithm starts with a sparse grid of $n_L=49$ labelled samples, creating a uniform grid with 33 cm resolution in both x-axis and y-axis. The RTFs associated with the labelled positions were estimated using a WGN signal. We then estimated $n_U=400$ unlabelled RTFs, by transmitting speech signals from random positions in the RoI. The kernel variance ϵ_m was set to 5000 for all nodes. The standard deviation of the labelled positions was set to $\sigma_p=0.71$ cm.

In the test stage, two- and three-speakers' scenarios were considered. The source signals were randomly drawn from the TIMIT database. The source positions were randomized in the RoI with minimum distance between sources of 0.3 m. The variance of all Gaussians in the MoG model were set to $\sigma^2=1$. For estimating the iRTFs we used $L=3$.

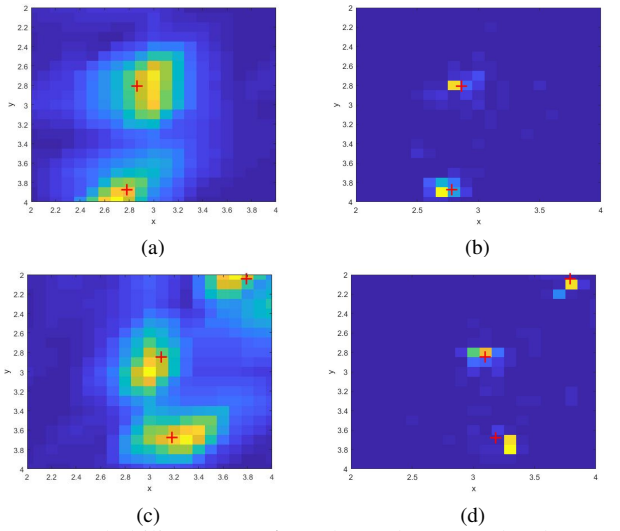
B. Results

As a baseline method we selected the unsupervised method described [16]. In Table I the position estimation error (in meters), averaged over 100 Monte-Carlo trials, is depicted. First, the localization accuracy as a function of the activity overlap between the sources for the two speakers scenario in mild reverberation level $T_{60}=0.3$ s is depicted, demonstrating that the algorithm gracefully degrade as the overlap percentage between the sources' activity increases. We then depict the localization accuracy with full activity overlap (concurrent speakers) for both two and three speakers scenarios and for three reverberation levels: $T_{60}=0.3,0.5,0.7$ s. It is

[1] Available at `https://github.com/ehabets/RIR-Generator`

TABLE I: Comparison between the proposed method and the baseline method [16], averaged over 100 trials. $T_{60} = 0.3, 0.5, 0.7$ s, two- or three-speakers and various overlap percentage. SNR=20 dB, $M = 8$ for all tests.

T_{60} [s]	S [#]	Overlap [%]	Proposed [m]	Baseline [16] [m]
0.3	2	0	0.11	0.33
0.3	2	25	0.11	0.35
0.3	2	50	0.13	0.38
0.3	2	75	0.14	0.41
0.3	2	100	0.15	0.44
0.5	2	100	0.18	0.50
0.7	2	100	0.20	0.69
0.3	3	100	0.20	0.44
0.5	3	100	0.21	0.54
0.7	3	100	0.23	0.69



(a) (b) (c) (d)

Fig. 1: Probability maps $\psi_{\mathbf{p}}$. The red '+' marks denote the ground-truth speakers locations. $T_{60} = 0.3$ s. Both two-speakers (a)+(b) and three-speakers (c)+(d) scenarios are presented. The maps of a clustering algorithm using $n_L = 49$ labelled points as centroids, are presented in (a)+(c) and using $n = 449$ labeled and unlabeled points in (b)+(d).

clearly demonstrated that the proposed method maintains high estimation accuracy even in high reverberation level and that it outperforms the baseline, unsupervised, method in all scenarios. Recall that the method in [16] requires the microphone positions. The probability maps $\psi_{\mathbf{p}}$ after the Lth iteration of two trials are depicted in Fig. 1 for both the two- and three-speakers scenarios for $T_{60} = 0.3$ s. It can be verified that the probability maps peak at the correct source positions. As a comparison we also present the interpolated probability maps using similar EM-based clustering that only uses the labelled RTFs as centroids. Despite the lower resolution, these maps still peak close to the correct positions. Only a small advantage of the higher resolution maps is demonstrated. Future study will investigate the robustness of the proposed two-stage approach to the grid density and the noise level.

REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 2, pp. 382–394, 2010.

[2] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 23, no. 11, pp. 2046–2058, 2015.

[3] S. Gannot, M. Haardt, W. Kellermann, and P. Willett, "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," *IEEE Selec. Topics in Sig. Proc.*, vol. 13, no. 1, pp. 3–7, 2019.

[4] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE Workshop on Sensor Array and Multichannel Sig. Proc. (SAM)*, 2018, pp. 410–414.

[5] 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 *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2015, pp. 2814–2818.

[6] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Selec. Topics in Sig. Proc.*, vol. 13, no. 1, pp. 8–21, 2019.

[7] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE Selec. Topics in Sig. Proc.*, vol. 13, no. 1, pp. 22–33, 2019.

[8] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2018, pp. 1462–1466.

[9] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 27, no. 1, pp. 178–188, 2018.

[10] R. Opochinsky, B. Laufer-Goldshtein, S. Gannot, and G. Chechik, "Deep ranking-based sound source localization," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, 2019, pp. 283–287.

[11] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *Proc. IEEE Workshop on App. of Sig. Proc. to Au. and Acous. (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[12] ——, "Study on manifolds of acoustic responses," in *Proc. Int. Conf. on Latent Variable Anal. and Sig. Sep. (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[13] B. Laufer, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 24, no. 8, pp. 1393–1407, 2016.

[14] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple-manifolds with distributed microphones," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 25, no. 7, pp. 1477–1491, Jul. 2017.

[15] A. Brendel, I. Altmann, and W. Kellermann, "Acoustic source position estimation based on multi-feature Gaussian processes," in *Proc. European Sig. Proc. Conf. (EUSIPCO)*, 2019, pp. 1–5.

[16] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 22, no. 2, pp. 392–402, 2014.

[17] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Sig. Proc.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[18] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A hybrid approach for speaker tracking based on TDOA and data-driven models," *IEEE/ACM Trans. Au., Sp., Lang. Proc.*, vol. 26, no. 4, pp. 725–735, Apr. 2018.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[20] Z. Koldovskỳ, J. Málek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1335–1347, 2015.

[21] A. Brendel, B. Laufer-Goldshtein, S. Gannot, R. Talmon, and W. Kellermann, "Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization," in *Proc. IEEE Intl. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2019, pp. 7898–7902.