MANIFOLD-BASED BAYESIAN INFERENCE FOR SEMI-SUPERVISED SOURCE LOCALIZATION

Bracha Laufer-Goldshtein[1]*, Ronen Talmon*[2] *and Sharon Gannot*[1]

[1] Faculty of Engineering
Bar-Ilan University
Ramat-Gan, 5290002, Israel
bracha.laufer@biu.ac.il, Sharon.Gannot@biu.ac.il

[2] Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa, 3200003, Israel
ronen@ee.technion.ac.il

ABSTRACT

Sound source localization is addressed by a novel Bayesian approach using a data-driven geometric model. The goal is to recover the target function that attaches each acoustic sample, formed by the measured signals, with its corresponding position. The estimation is derived by maximizing the posterior probability of the target function, computed on the basis of acoustic samples from known locations (labelled data) as well as acoustic samples from unknown locations (unlabelled data). To form the posterior probability we use a manifold-based prior, which relies on the geometric structure of the manifold from which the acoustic samples are drawn. The proposed method is shown to be analogous to a recently presented semi-supervised localization approach based on manifold regularization. Simulation results demonstrate the robustness of the method in noisy and reverberant environments.

Index Terms— relative transfer function (RTF), kernel function, manifold-based prior, manifold regularization.

1. INTRODUCTION

Sound source localization is of great merit in a large variety of applications, including: video conferencing, automatic camera steering and speaker separation. For this reason, it has attracted the attention of many researchers along the years, and a wide variety of localization methods have been proposed. However, classical localization algorithms are highly sensitive to adverse conditions, namely, to the presence of high reverberation and background noise. Thus, the challenge is to form a robust localizer that successfully circumvents these limiting factors.

Conventional localization approaches can be roughly divided into two main categories: single- and dual-step approaches. In the first class of algorithms, the location of the source is estimated directly from the measured signals. Numerous methods fall under this category, most of which derived by applying the maximum likelihood (ML) criterion [1–4], or spectral methods such as the well-known multiple signal classification (MUSIC) algorithm [5]. In the dual-stage class, the first step is to estimate the time difference of arrival (TDOA) for each pair of microphones [6–10]. Next, the TDOA readings are combined to attain the actual localization [11, 12].

Common to most conventional localization methods is that they solely depend on the measured signals, and do not utilize any prior information regarding the acoustic environment in which the source is located. However, in some scenarios, e.g. in meeting rooms or cars, the source is expected to be positioned in a specified region in the enclosure. Thus, representative samples from the region of interest can be measured in advance. Such representative samples provide an additional information about the acoustic environment that may be utilized to develop robust localization methods. Thus far, only few attempts were made to involve training information for performing source localization [13–18].

In this paper we present a semi-supervised approach on the basis of labelled (attached with corresponding locations) and unlabelled (from unknown locations in the predefined region of interest) representative samples. It is important to emphasize that these representative samples should be generated uniquely for each specific acoustic environment. Generating labelled data is a cumbersome task, and hence the amount of labelled data is assumed to be very limited. However, the availability of unlabelled data is much greater, since it can be collected whenever someone is speaking in the enclosure of interest. This observation motivates the development of a semi-supervised localization approach.

Our goal is to estimate the target function which receives an acoustic sample and returns its corresponding location. The target function is estimated in this work using a Bayesian inference framework which involves a likelihood function and a prior probability. While the likelihood function measures the correspondence of the target function to the labelled examples, the prior probability reflects our prior belief regarding the distribution of the target function. In particular, following Sinhwani et al. [19], we propose to use a manifold-based prior which relies on the geometric structure of the RTF samples, implied by unlabelled samples. We discuss the analogy of the Bayesian approach to a recently presented semi-supervised source localization method based on manifold regularization [20]. The paper is supported by simulation results that demonstrate the robustness of the proposed method to noise and reverberation.

2. PROBLEM FORMULATION

We consider the following acoustic environment. A source is located at position $\mathbf{p} = [p_x, p_y, p_z]$ in a reverberant enclosure. The source is emitting an unknown signal $s(n)$ which is measured by a pair of microphones, also located in the enclosure. The noisy measurements, denoted by $x(n)$ and $y(n)$, are given by a convolution between the clean source signal and the corresponding acoustic impulse response (AIR), contaminated by stationary noise signals:

$$\begin{aligned} x(n) &= a_1(n, \mathbf{p}) * s(n) + u_1(n) \\ y(n) &= a_2(n, \mathbf{p}) * s(n) + u_2(n) \end{aligned} \tag{1}$$

where n is the time index, $a_i(n, \mathbf{p}),\ i = 1, 2$ are the corresponding AIRs relating the source at position $\mathbf{p}$ and each of the microphones, and $u_i(n),\ i = 1, 2$ are the noise signals.

A feature vector that represents the characteristics of the acoustic environment and is independent of the source signal, is constructed

based on the two measured signals. We propose to use the relative transfer function (RTF) [21, 22], defined by: $H_{yx}(k,\mathbf{p}) = \frac{A_2(k,\mathbf{p})}{A_1(k,\mathbf{p})}$, where $A_1(k,\mathbf{p})$ and $A_2(k,\mathbf{p})$ are the acoustic transfer functions (ATFs) of the respective AIRs, and k denotes a discrete frequency index. Since the ATFs are unavailable, the RTF is estimated instead based on the measured signals:

$$\hat{H}_{yx}(k,\mathbf{p}) \equiv \frac{\hat{S}_{yx}(k,\mathbf{p})}{\hat{S}_{xx}(k,\mathbf{p})} \cong \frac{A_2(k,\mathbf{p})}{A_1(k,\mathbf{p})} \tag{2}$$

where $S_{yx}(k,\mathbf{p})$ and $S_{xx}(k,\mathbf{p})$ are the cross power spectral density (CPSD) between $y(n)$ and $x(n)$ and the power spectral density (PSD) of $x(n)$, respectively. Note that the estimator of (2) is biased due to the additive noise [23]. However, we will show that the proposed method is insensitive to this type of estimation errors. Accordingly, we define the feature vector $\mathbf{h}(\mathbf{p}) = [\hat{H}_{yx}(0,\mathbf{p}),\ldots,\hat{H}_{yx}(D-1,\mathbf{p})]^T$ as the concatenation of estimated RTF values in D frequency bins. In practice, we discard high frequencies in which the ratio in (2) is meaningless due to weak speech components. For the sake of brevity, we omit the dependency on the position from the notation, and denote the RTF feature vector by $\mathbf{h}$.

From a probabilistic view point, $\mathbf{h}$ is a random vector, drawn from some probability distribution p_H. Though originally the RTFs have a high dimensional representation due to reverberation, we have shown in [22] that they pertain to a nonlinear manifold of much lower dimensions. The support of p_H, representing the manifold from which the RTFs are drawn, will be denoted by $\mathcal{M}$.

3. BAYESIAN INFERENCE FOR SEMI-SUPERVISED LOCALIZATION

We assume that we have a training set consisting of l labelled RTF samples, attached with their respective locations, and u unlabelled RTF samples from unknown locations. Let $H_L = \{\mathbf{h}_i\}_{i=1}^{l}$ be the set of l labelled samples, and $P_L = \{p(\mathbf{h}_i)\}_{i=1}^{l}$ their associated labels. The set of unlabelled samples is denoted by $H_U = \{\mathbf{h}_i\}_{i=l+1}^{n}$, where $n = l+u$. The training set, consisting both labelled and unlabelled samples is denoted by $H_D = H_L \cup H_U = \{\mathbf{h}_i\}_{i=1}^{n}$.

The aim of this work is to estimate the locations corresponding to a test set of q pairs of measurements $\{x_i(n), y_i(n)\}_{i=n+1}^{m}$ of unknown sources from unknown locations, where $m = n+q$. The corresponding set of RTF samples is denoted by $H_T = \{\mathbf{h}_i\}_{i=n+1}^{m}$. The entire set, comprised of both the training and the test samples, is denoted by $H = H_D \cup H_T = \{\mathbf{h}_i\}_{i=1}^{m}$.

It is important to emphasize that both the RTF samples and the measured positions are treated as random vectors/variables. We assume that they are generated by the following stochastic model: First, an RTF sample residing in the manifold $\mathcal{M}$ is drawn according to p_H (not used explicitly in the following computations). The position of the source is a random variable obtained as an output of the target function that receives the RTF sample as an input. The target function is assumed to follow a stochastic process. Finally, the measured position is a noisy version of the actual position due to uncertainty or imperfection of the measurements. Following this probabilistic model, a Bayesian inference framework can be formulated for the problem of estimating the position attached with an observed RTF sample, given the sets of labelled and unlabelled samples. A flow diagram of the statistical model is illustrated in Fig. 1.

3.1. Bayesian Formulation

Let $f_c : \mathbb{C}^D \rightarrow \mathbb{R}$ $c \in \{x,y,z\}$ be the function that attaches each RTF sample to one of the coordinates of the corresponding source position, i.e. $f(\mathbf{h}) = p_c$. In this paper we focus on estimating one

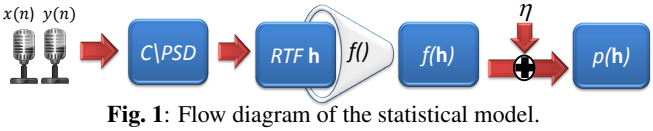

Fig. 1: Flow diagram of the statistical model.

position coordinate, hence, we omit the coordinate subscript. However, the analysis, the results and the algorithm described henceforth can be naturally extended to the estimation of multiple coordinates.

We treat the target function as a stochastic process, i.e., as a collection of random variables of the form $f(\mathbf{h})$, $\mathbf{h} \in \mathcal{M}$. Accordingly, the target function can be estimated based on the following posterior probability, given by Bayes' rule:

$$p(f|P_L,H_L,H_U) \propto p(P_L|f,H_L)\cdot p(f|H_L,H_U) \tag{3}$$

The posterior function is composed of two parts: the likelihood function $p(P_L|f,H_L)$ and the prior probability $p(f|H_L,H_U)$ of f. The likelihood function measures the correspondence of the values of the function f, at the labelled samples H_L, to the measured positions P_L. The prior probability specifies our a priori belief about the properties of f.

It is assumed that the measured positions $P_L = \{p(\mathbf{h}_i)\}_{i=1}^{l}$ follow a noisy observation model, given by:

$$p(\mathbf{h}_i) = f(\mathbf{h}_i) + \eta_i \; i = 1,\ldots,l \tag{4}$$

where $\eta_i \sim \mathcal{N}(0,\sigma^2)$ $i = 1,\ldots,l$ are i.i.d. Gaussian noises, independent of f. This model reflects the uncertainty due to imprecise microphones' calibration or imperfect measurement of the source position while acquiring the labelled set. Under this model, the likelihood function is given by:

$$p(P_L|f,H_L) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{l}(p(\mathbf{h}_i)-f(\mathbf{h}_i))^2\right\} \tag{5}$$

As for the prior of the function f, we assume that it follows a Gaussian process [24–26]:

$$f \sim \mathcal{GP}(\nu,k) \tag{6}$$

where ν is the mean function and k is the covariance function, that specify the Gaussian process. The choice of a Gaussian process as a prior was shown to give good results in regression problems [26]. Moreover, this choice is justified by its analogy to the optimization framework discussed in Section 4.

In the following, we assume that the mean function ν is constant and equals zero to maintain simplicity of the equations. However, all the results apply also to any general mean function, with only small changes. The function $k : \mathcal{M}\times\mathcal{M} \longrightarrow \mathbb{R}$, often referred to as a *kernel* function, is a pairwise function that evaluates the covariance of each pair of samples drawn from the process f. In order to serve as an admissible covariance function, k should be a symmetric and positive-definite kernel.

According to (6), the random vector $\mathbf{f}_H = [f(\mathbf{h}_1),\ldots,f(\mathbf{h}_m)]$, has a joint Gaussian distribution, i.e.,

$$\mathbf{f}_H \sim \mathcal{N}(\mathbf{0}_m, \mathbf{\Sigma}_{HH}) \tag{7}$$

where $\mathbf{0}_m$ is an $m\times 1$ vector of all zeros and $\mathbf{\Sigma}_{HH}$ is the covariance matrix with elements $k(\mathbf{h}_i,\mathbf{h}_j)$, $\mathbf{h}_i,\mathbf{h}_j \in H$. Usually, the covariance between $f(\mathbf{h}_i)$ and $f(\mathbf{h}_j)$, given by $k(\mathbf{h}_i,\mathbf{h}_j)$, depends

only on the corresponding coordinates $\mathbf{h}_i$ and $\mathbf{h}_j$, and is independent of the rest of the set. For example, the covariance can be represented by a Gaussian kernel with variance ε_k: $k(\mathbf{h}_i, \mathbf{h}_j) = \exp\{-\|\mathbf{h}_i - \mathbf{h}_j\|^2/\varepsilon_k\}$, where $\|\cdot\|$ denotes the l_2 norm. In the following, this type of kernel and the corresponding Gaussian process, will be referred to as a *standard kernel* and a *standard Gaussian process*, respectively. In the remainder of this section we use a standard kernel function which forms a prior probability that does not exploit the available set of unlabelled data H_U. In the next section, we define a *modified* kernel function $\tilde{k}$ that depends also on the unlabelled data, thereby better exploiting the geometric properties of the manifold $\mathcal{M}$.

We use a transductive view point [26], i.e. rather than deriving a general estimator of the function f, which is a cumbersome task, we estimate the function value at some specific test point $\mathbf{h}_t \in \mathcal{M}$ from an unknown position. The corresponding posterior probability is $p(f(\mathbf{h}_t)|P_L, H_L)$. According to (5) and (7), the function at the test point $f(\mathbf{h}_t)$ and the concatenation of all labelled training positions $\mathbf{p}_L = \text{vec}\{P_L\} \equiv [p(\mathbf{h}_1), \ldots, p(\mathbf{h}_l)]^T$ are jointly Gaussian, with:

$$\begin{bmatrix} \mathbf{p}_L \\ f(\mathbf{h}_t) \end{bmatrix} \Bigg| H_L \sim \mathcal{N}\left(\mathbf{0}_{l+1}, \begin{bmatrix} \mathbf{\Sigma}_{LL} + \sigma^2\mathbf{I}_l & \mathbf{\Sigma}_{Lt} \\ \mathbf{\Sigma}_{Lt}^T & \Sigma_{tt} \end{bmatrix}\right) \tag{8}$$

where $\mathbf{\Sigma}_{LL}$ is an $l \times l$ covariance matrix defined over the function values at the labelled samples H_L, $\mathbf{\Sigma}_{Lt}$ is an $l \times 1$ covariance vector between the function values at H_L and $f(\mathbf{h}_t)$, Σ_{tt} is the variance of $f(\mathbf{h}_t)$, and $\mathbf{I}_l$ is the $l \times l$ identity matrix. This implies that the conditional distribution $p(f(\mathbf{h}_t)|P_L, H_L)$ is a multivariate Gaussian with mean μ_{cond} and variance σ^2_{cond} given by:

$$\begin{aligned} \mu_{\text{cond}} &= \mathbf{\Sigma}_{Lt}^T \left(\mathbf{\Sigma}_{LL} + \sigma^2\mathbf{I}_l\right)^{-1} \mathbf{p}_L \\ \sigma^2_{\text{cond}} &= \Sigma_{tt} - \mathbf{\Sigma}_{Lt}^T \left(\mathbf{\Sigma}_{LL} + \sigma^2\mathbf{I}_l\right)^{-1} \mathbf{\Sigma}_{Lt}. \end{aligned} \tag{9}$$

Hence, the maximum a posteriori probability (MAP) estimator of $f(\mathbf{h}_t)$ (which coincides with the minimum mean squared error (MMSE) estimator in the Gaussian case) is given by:

$$\widehat{f}(\mathbf{h}_t) = \mu_{\text{cond}} = \mathbf{\Sigma}_{Lt}^T \left(\mathbf{\Sigma}_{LL} + \sigma^2\mathbf{I}_l\right)^{-1} \mathbf{p}_L \tag{10}$$

3.2. Data-Driven Prior

In this section we follow Sinhwani et al. [19] and introduce an alternative prior for the function f, which is based on the manifold $\mathcal{M}$ from which the RTF samples are drawn. The new prior, computed based on both labelled and unlabelled data, is a Gaussian process with a modified kernel function that reflects the intrinsic patterns in the data. We form a discrete representation of the manifold by a graph defined over the entire training set H_D. The graph nodes are the training samples and the weights of the edges, constituting an $n \times n$ affinity matrix $\mathbf{W}$, are computed using a kernel function. We denote by $\mathcal{G}$ an abstract collection of random variables that represents the *geometric structure* of the manifold. Accordingly, the likelihood of the geometry variables $\mathcal{G}$ can be defined by:

$$P(\mathcal{G}|\mathbf{f}_D) \propto \exp\left\{-\frac{\gamma_M}{2}\left(\mathbf{f}_D^T\mathbf{M}\mathbf{f}_D\right)\right\} \tag{11}$$

where γ_M is a scaling factor, $\mathbf{f}_D = [f(\mathbf{h}_1), \ldots, f(\mathbf{h}_n)]^T$ and $\mathbf{M}$ is the graph Laplacian given by $\mathbf{M} = \mathbf{S} - \mathbf{W}$. Here, the diagonal matrix $\mathbf{S}$ is given by $S_{ii} = \sum_{j=1}^n W_{ij}$. The probability in (11) reflects the tendency of $f(\mathbf{h}_i)$ and $f(\mathbf{h}_j)$ to have similar values when the corresponding RTF samples have strong similarity, namely they are closely connected in the graph $\mathbf{W}$. In this sense, the likelihood function is a measure of correspondence between the values of the target function f and the structure of the manifold, implied by the geometry variables.

In order for the model to be extendible to additional test data H_T, we make the assumption that given $\mathbf{f}_D$, the geometry variables are independent of the function values in other points, i.e. $p(\mathcal{G}|\mathbf{f}_H) = p(\mathcal{G}|\mathbf{f}_D)$. By this assumption we avoid the recomputation of the graph Laplacian $\mathbf{M}$ for the new dataset.

Accordingly, the posterior of $\mathbf{f}_H$, given the geometry variables, constitutes a manifold-based prior for $\mathbf{f}_H$, which can be written as:

$$\begin{aligned} p(\mathbf{f}_H|\mathcal{G}) &= p(\mathcal{G}|\mathbf{f}_H) \cdot p(\mathbf{f}_H)/p(\mathcal{G}) = p(\mathcal{G}|\mathbf{f}_D) \cdot p(\mathbf{f}_H)/p(\mathcal{G}) \\ &\propto \exp\left\{-\frac{\gamma_M}{2}\left(\mathbf{f}_D^T\mathbf{M}\mathbf{f}_D\right)\right\} p(\mathbf{f}_H) \end{aligned} \tag{12}$$

where $p(\mathbf{f}_H) \sim \mathcal{N}(\mathbf{0}_m, \mathbf{\Sigma}_{HH})$ is the prior probability of samples drawn from the standard Gaussian process defined in (7). Hence, the posterior distribution can be written as the Gaussian distribution $p(\mathbf{f}_H|\mathcal{G}) \propto \exp\left\{-\frac{1}{2}\mathbf{f}_H^T\tilde{\mathbf{\Sigma}}_{HH}^{-1}\mathbf{f}_H\right\}$, where:

$$\tilde{\mathbf{\Sigma}}_{HH}^{-1} = \begin{bmatrix} \mathbf{\Sigma}_{DD} & \mathbf{\Sigma}_{DT} \\ \mathbf{\Sigma}_{DT}^T & \Sigma_{TT} \end{bmatrix}^{-1} + \gamma_M \begin{bmatrix} \mathbf{M} & \mathbf{0}_n \\ \mathbf{0}_n^T & 0 \end{bmatrix}. \tag{13}$$

Based on the matrix inversion lemma it can be shown [19] that the elements of $\tilde{\mathbf{\Sigma}}_{HH}$ are given by:

$$\tilde{k}(\mathbf{h}_i, \mathbf{h}_j) = k(\mathbf{h}_i, \mathbf{h}_j) - \gamma_M \mathbf{\Sigma}_{Di}^T(\mathbf{I}_n + \gamma_M\mathbf{M}\mathbf{\Sigma}_{DD})^{-1}\mathbf{M}\mathbf{\Sigma}_{Dj} \tag{14}$$

for $\mathbf{h}_i, \mathbf{h}_j \in H$, where $\mathbf{\Sigma}_{Di}$ denotes the column vector $[k(\mathbf{h}_1, \mathbf{h}_i), \ldots, k(\mathbf{h}_n, \mathbf{h}_i)]^T$.

To conclude, the Gaussian process conditioned on the geometry variables $\mathcal{G}$ is associated with a modified covariance function $\tilde{k}$, that will be termed *manifold-based kernel*. Based on this new data-driven prior with the manifold-based kernel $\tilde{k}$, an alternative estimator for $f(\mathbf{h}_t)$ is then given by:

$$\widehat{f}(\mathbf{h}_t) = \tilde{\mathbf{\Sigma}}_{Lt}^T \left(\tilde{\mathbf{\Sigma}}_{LL} + \sigma^2\mathbf{I}_l\right)^{-1} \mathbf{p}_L. \tag{15}$$

which is the similar to (10), where the covariance terms $\tilde{\mathbf{\Sigma}}$ are computed using $\tilde{k}$ rather than k.

4. ANALOGY TO MANIFOLD REGULARIZATION FOR LOCALIZATION

In [20] we have presented a semi-supervised source localization algorithm, based on manifold regularization. We briefly describe the concepts of the proposed approach, and its equivalence to the Bayesian framework introduced here. The idea is to formulate the estimation of the target function f as a regularized optimization problem. We assume that the target function belongs to a reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{H}_k$. An RKHS is a Hilbert space of functions $\mathcal{M} \to \mathbb{R}$, associated with a unique kernel function k. The kernel function has the reproducing property, which means that it evaluates each function in the space by inner product, i.e.: $\langle f(\cdot), k(\mathbf{h}, \cdot)\rangle = f(\mathbf{h})$, for all $f \in \mathcal{H}_k$ and $\mathbf{h} \in \mathcal{M}$. Accordingly, every function in the space can be represented as a linear combination of the kernel functions.

There is a close relation between the RKHS and the Gaussian process, when they are associated with the same kernel function [19]. In this section we show that these two different view points lead to the same estimators.

As introduced by Belkin et. al. [27], a regularized optimization problem in an RKHS, can be defined by:

$$\widehat{f} = \underset{f \in \mathcal{H}_k}{\text{argmin}} \frac{1}{\sigma^2}\sum_{i=1}^{l}(p(\mathbf{h}_i) - f(\mathbf{h}_i))^2 + \|f\|^2_{\mathcal{H}_k} + \gamma_M\mathbf{f}_D^T\mathbf{M}\mathbf{f}_D \tag{16}$$

where $\|\cdot\|^2_{\mathcal{H}_k}$ is the RKHS norm. The optimization consists of three parts: a cost function defined over the labelled examples (the first term), a smoothness penalty in $\mathcal{H}_k$ (the second term) and a smoothness penalty with respect to the manifold $\mathcal{M}$ (the third term). The role of the squared cost function is to measure how f fits the data, which is analogous to the role of the Gaussian likelihood function in (5). In the same manner, the second and the third terms in (16) are analogous to the data-driven prior presented in Section. 3.2. It was shown in [27], that the two regularization terms in (16) can be merged into a single regularization term, by defining a new RKHS $\tilde{\mathcal{H}}_k$, i.e.: $\|f\|^2_{\tilde{\mathcal{H}}_k} = \|f\|^2_{\mathcal{H}_k} + \gamma_M \mathbf{f}_D^T \mathbf{M} \mathbf{f}_D$. The new RKHS $\tilde{\mathcal{H}}_k$ is associated with a modified kernel $\tilde{k}$ defined similarly to (14). Therefore, the optimization problem (16) can be recast as:

$$\widehat{f} = \underset{f \in \tilde{\mathcal{H}}_k}{\operatorname{argmin}} \frac{1}{\sigma^2} \sum_{i=1}^{l} (p(\mathbf{h}_i) - f(\mathbf{h}_i))^2 + \|f\|^2_{\tilde{\mathcal{H}}_k} \tag{17}$$

According to the Representer theorem, the target function minimizing (17), can be written as a linear combination of the kernel functions, only in the set of the labelled samples, i.e.: $\widehat{f}(\mathbf{h}_t) = \sum_{i=1}^{l} a_i \tilde{k}(\mathbf{h}_i, \mathbf{h}_t)$. Thus, the optimization (17) is reduced to estimating the interpolation weights $\{a_i\}$, by substituting this form in (17), and differentiating with respect to $\mathbf{a} = [a_1, ..., a_l]^T$. Following this computation, we receive that the interpolation weights are given by: $\mathbf{a}^* = \left(\tilde{\mathbf{K}}_{LL} + \sigma^2 \mathbf{I}_l\right)^{-1} \mathbf{p}_L$, where $(\tilde{K}_{LL})_{ij} = \tilde{k}(\mathbf{h}_i, \mathbf{h}_j)$. Thus, the estimated value of the target function f at any point $\mathbf{h}_t \in \mathcal{M}$, is given by (equivalent to (15)):

$$\widehat{f}(\mathbf{h}_t) = \tilde{\mathbf{K}}_{Lt}^T \left(\tilde{\mathbf{K}}_{LL} + \sigma^2 \mathbf{I}_l\right)^{-1} \mathbf{p}_L \tag{18}$$

where $\tilde{\mathbf{K}}_{Lt} = [\tilde{k}(\mathbf{h}_1, \mathbf{h}_t), \ldots, \tilde{k}(\mathbf{h}_n, \mathbf{h}_t)]^T$.

To conclude, both the Bayesian approach and the regularized optimization problem defined in an RKHS give rise to the same estimators, when the same kernel function serves as the covariance function of the Gaussian process and as the reproducing kernel of $\tilde{\mathcal{H}}_k$, receptively. In both view points, the underlying structure of the manifold is taken into consideration by using a manifold-based kernel function.

5. EXPERIMENTAL STUDY

In this section we examine the performance of the proposed estimator (15) in recovering the azimuth angle of a source uttering speech signals (consisting of both female and male speech). To simulate a reverberant room of size $6 \times 6.2 \times 3$, we use an efficient implementation [28] of the image method [29]. There are two microphones located in the room, at $[3, 3, 1]$m and $[3.2, 3, 1]$m, respectively. The source is at 2 m distance from the first microphone, on the same latitude. Our goal is to estimate the azimuth angle of the source in the range between $10^\circ \div 60^\circ$.

Each of the measured signals is generated by convolving the clean speech signal with the AIR relating the source and the corresponding microphone and contaminating the filtered signal by a white Gaussian noise (WGN). The training set comprises $n = 400$ pairs of measurements, among which only $l = 6$ are associated with their corresponding positions, forming a grid of approximately 10° distance between adjacent labelled samples. For each location, we use a unique speech signal, 3 s long sampled at $f_s = 16$ kHz. The CPSD and the PSD are estimated with Welch's method with 0.128 s windows and 75% overlap and are utilized for estimating the RTF in (2) for 2048 frequency bins. The RTF vector consists of $D = 400$ frequency bins corresponding to the frequency range 0-3kHz, in which most of the speech components are concentrated. A Gaussian kernel is set to the covariance function k of the Gaussian process. For constructing the graph Laplacian $\mathbf{M}$ we use a truncated Gaussian kernel, i.e., with non-zeros entries for the 12 nearest-neighbours of each sample. The performance is examined over a set of $q = 120$ test samples, of unknown sources from unknown locations. To prevent the results from being dependent on a specific reflection pattern of a certain room section, we repeat the simulation and present the average root mean squared error (RMSE) over 50 rotations of the entire constellation, with respect to the first microphone.

We compare the performance of the two estimators of (10) and (15) with standard and manifold-based kernels, respectively. For comparison, we also apply the classical generalized cross-correlation phase transformation (GCC-PHAT) algorithm [6]. Two scenarios are investigated: different reverberation levels (signal to noise ratio (SNR) fixed to 20 dB) and different noise levels (T_{60} fixed to 300 ms). Note that in each scenario, we choose moderate fixed values for T_{60} and SNR, in order to isolate the two disturbing factors, namely, noise and reverberation. The RMSE of all three methods in both scenarios, is depicted in Fig. 2.

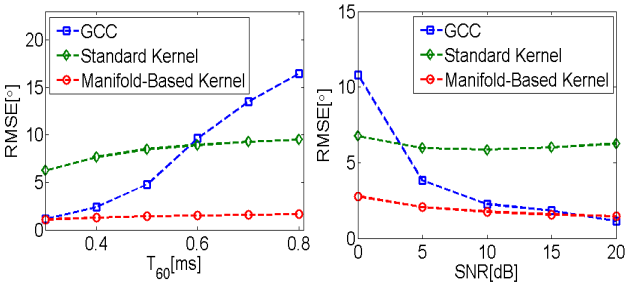

Fig. 2: The RMSE (a) for differnet reverberation times, and (b) for differnet noise levels. ($\sigma^2 = 0.005, \gamma_M = 20$)

We observe that the GCC algorithm performs well in moderate conditions, but exhibits a significant performance deterioration as reverberation or noise level increases. In adverse conditions the correlation between the measured signals is distorted, hence, the peak corresponding to the direct path is usually misidentified. In contrast, the training based algorithms are shown to be much more robust to the presence of noise and reverberation. This type of approaches takes advantage of the prior information implied by the training set to compensate for the information loss in adverse conditions. It can also be observed that the error is significantly reduced by using the manifold-based kernel compared to the standard kernel. The manifold-based kernel is tailored to the underlying structure of the RTF samples and is hence more appropriate for estimating the target function which maps each RTF to its corresponding position.

6. CONCLUSIONS

A novel semi-supervised Bayesian approach was derived for the source localization problem. Both labelled and unlabelled samples are utilized for representing the geometric structure of the RTF samples. We proposed a manifold-based prior, associated with a unique manifold-based kernel, which reflects the correspondence of the target function to the structure of the manifold. The resulting Bayesian estimator was shown to be robust to noise and reverberation. In addition, we have shown the equivalence between the proposed method and a manifold-regularized optimization in an RKHS, when the reproducing kernel coincident with the covariance function of the Gaussian process. The new Bayesian formulation motivates further examination of the appropriate statistical assumptions, as well as of possible extensions, such as tracking moving sources.

7. REFERENCES

[1] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 38, no. 7, pp. 1132–1143, 1990.

[2] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843–1854, 2002.

[3] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2004, pp. ii–133.

[4] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2007, pp. I–125.

[5] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[7] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1997, pp. 375–378.

[8] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[9] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177 – 204, 2005.

[10] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on applied signal processing*, vol. 2006, pp. 170–170, 2006.

[11] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, 1997.

[12] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[13] R. Talmon, D. Kushnir, R. Coifman, I. Cohen, and S. Gannot, "Parametrization of linear systems using diffusion kernels," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1159–1173, Mar. 2012.

[14] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 245–248.

[15] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[16] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 1, 2015.

[17] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 76–80.

[18] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[19] V. Sindhwani, W. Chu, and S. S. Keerthi, "Semi-supervised gaussian process classifiers." in *IJCAI*, 2007, pp. 1059–1064.

[20] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *pre-print*, Aug 2015, arXiv:1508.03148v1.

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

[22] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Study on manifolds of acoustic responses," in *Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[23] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 451–459, 2004.

[24] D. J. MacKay, "Introduction to gaussian processes," *NATO ASI Series F Computer and Systems Sciences*, vol. 168, pp. 133–166, 1998.

[25] C. K. Williams and D. Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[26] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[27] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 824–831.

[28] E. A. P. Habets, "Room impulse response (RIR) generator," http://home.tiscali.nl/ehabets/rir_generator.html, Jul. 2006.

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