Semi-Supervised Sound Source Localization Based on Manifold Regularization

Bracha Laufer-Goldshtein, *Student Member, IEEE*, Ronen Talmon, *Member, IEEE*,
and Sharon Gannot, *Senior Member, IEEE*

Abstract—Conventional speaker localization algorithms, based merely on the received microphone signals, are often sensitive to adverse conditions, such as: high reverberation or low signal-to-noise ratio (SNR). In some scenarios, e.g., in meeting rooms or cars, it can be assumed that the source position is confined to a predefined area, and the acoustic parameters of the environment are approximately fixed. Such scenarios give rise to the assumption that the acoustic samples from the region of interest have a distinct geometrical structure. In this paper, we show that the high-dimensional acoustic samples indeed lie on a low-dimensional manifold and can be embedded into a low-dimensional space. Motivated by this result, we propose a semi-supervised source localization algorithm based on two-microphone measurements, which recovers the inverse mapping between the acoustic samples and their corresponding locations. The idea is to use an optimization framework based on manifold regularization, that involves smoothness constraints of possible solutions with respect to the manifold. The proposed algorithm, termed manifold regularization for localization, is adapted while new unlabelled measurements (from unknown source locations) are accumulated during runtime. Experimental results show superior localization performance when compared with a recently presented algorithm based on a manifold learning approach and with the generalized cross-correlation algorithm as a baseline. The algorithm achieves $2°$ accuracy in typical noisy and reverberant environments (reverberation time between 200 and 800 ms and SNR between 5 and 20 dB).

Index Terms—Diffusion distance, manifold regularization, relative transfer function (RTF), reproducing kernel Hilbert space (RKHS), sound source localization.

I. INTRODUCTION AND MOTIVATION

THE problem of source localization has attracted the attention of many researchers during the last decades. Various applications rely on the recovery of the spatial position of an emitting source, such as: automated camera steering [1], speaker separation [2] and beamformer steering for robust speech recognition [3]. For this reason, considerable amount of efforts have been devoted to investigate this field and a wide range of methods have been proposed over the years. Common to all

localization approaches is the utilization of multiple microphone recordings to infer the spatial information. The fundamental challenge is to attain robust localization in poor conditions, i.e., in the presence of high reverberation and background noises.

Conventional localization approaches can be roughly divided into two main categories: single- and dual-step approaches. In the first class of algorithms, the source location is determined directly from the microphone signals. The most dominant member of this class is the maximum likelihood (ML) algorithm. The algorithm is derived by applying the ML criterion to a chosen statistical model of the received signals. This optimization often involves maximization of the output power of a beamformer, steered to all potential source locations [4]–[6]. Among these methods we mention the well-known steered response power phase transformation (SRP-PHAT) algorithm [7], [8]. Another type of single-stage approaches is high resolution spectral estimation methods, such as the multiple signal classification algorithm [9], and the estimation of signal parameters via rotational invariance techniques [10].

In the dual-step approaches category, the first stage involves time difference of arrival (TDOA) estimation from spatially separated microphone pairs [11]. The classical method for TDOA estimation is the generalized cross-correlation (GCC) algorithm introduced in the landmark paper by Knapp and Carter [12]. The generalized cross-correlation (GCC) method relies on the assumption of a reverberant-free model such that the acoustic transfer function (ATF), which relates the source and each of the microphones, is a pure delay. However, this assumption does not hold in the presence of room reverberation, rendering a performance deterioration [13]. Consequently, improvements of the GCC method for the reverberant case were proposed [14]–[18]. Another type of approaches are subspace methods based on adaptive eigenvalue decomposition [19] and generalized eigenvalue decomposition [20].

In the second algorithmic stage, the noisy TDOA estimates are combined to carry out the actual localization. Each TDOA estimate is associated with an infinite set of source positions, lying on a half of an hyperboloid. The locus of the speaker can be recovered by intersecting the hyperboloid surfaces corresponding to the measurements of different pairs of microphones. However, the computation of a 3-dimensional hyperboloids intersection is a cumbersome task and tends to be sensitive to TDOA estimation errors. In far-field regime the hyperboloid can be approximated by a cone, and linear intersection estimate can be applied [21]. Another simplifying approach is to recast the hyperbolic equations into a spherical form, and apply the nonlinear least squares approach [22].

All the prementioned methods utilize the spatial information conveyed by the received signals, but do not rely on any prior information about the enclosure in which the measurements are obtained. In some scenarios, e.g., in meeting rooms or cars, the source position is confined to a predefined area. It is reasonable to assume that representative samples from the region of interest can be measured in advance. Examining the structures and patterns characterizing the representative samples can be utilized for formulating a data-driven model which relates the measured signals to their corresponding source positions. The additional information may help to better cope with the challenges posed by reverberation and noise. So far, only few attempts were made to involve training information for performing source localization.

Deleforge and Horaud in [23], discussed a two-dimensional sound localization scheme, in the binaural hearing context. Their central assumption is that the binaural observations lie on an intrinsic manifold which is locally linear. Accordingly, they proposed a probabilistic piecewise affine regression model, that learns the localization-to-interaural mapping and its inverse. In [24] and [25], the authors have generalized the algorithm to deal with multiple sources using variational expectation maximization framework. Another approach presented in the field of binaural hearing, is based on a Gaussian mixture model which is used to learn the azimuth-dependent distribution of the binaural feature space [26].

In [27] the task of direction of arrival (DOA) estimation was formulated as a classification problem and a learning-based approach was presented. They proposed to extract features from the GCC vectors and use a multilayer perceptron neural network to learn the nonlinear mapping from such features to the DOA.

Talmon *et al.* [28] introduced a supervised method based on *manifold learning*, using *diffusion kernels*. The main idea is specifying the fundamental controlling parameters of the acoustic impulse response (AIR) using a manifold learning scheme. Assuming that the position of the source is the only varying degree-of-freedom of the system at hand, this process is capable of recovering the unknown source locations. The key point of the algorithm is to use an appropriate diffusion kernel with a specifically-tailored distance measure, that is capable of finding the underlying independent parameters, dominating the system. Talmon *et al.* [29] have applied this method to a single microphone system with a white Gaussian noise (WGN) input.

In [30] we adopted the paradigm of [29] and adapted it to a more realistic setting where the source is a speech signal rather than a WGN signal. The power spectral density (PSD) of the speech signal is non-flat (as well as non-stationary). Hence, the spectral variations may blur the variations attributed to the different possible locations of the source. In order to mitigate this problem, we committed two major changes in the algorithm presented in [29]: 1) a second microphone was added and 2) the feature vector, that was originally based on the correlation function has been replaced by a PSD-based vector. It should be emphasized that in [29] the feature vector was associated with the AIR, whereas in [30] the feature vector relied on the relative transfer function (RTF) which is the Fourier transform of the relative impulse response.

Though localization algorithms based on the diffusion framework were shown to perform well, their fundamental drawback is that they do not provide any guarantee for optimality. In general, the diffusion-based methods are implemented by a dual-stage approach. First, a low-dimensional embedding of the representative samples is recovered in an unsupervised manner. Second, the new representation is used to estimate the unknown locations based on the labelled samples. The separation into two stages where one is entirely unsupervised and the other is entirely supervised is not necessarily optimal. Moreover, the unlabelled data are not exploited for the estimation itself.

The significance of combining both labelled and unlabelled data, in the source localization context, should be further emphasized. Classification and regression algorithms which rely on training data are very popular in various applications, such as: text categorization, handwriting recognition, images classification and speech recognition. Nowadays, there exist a rich database for each of these tasks, with considerable amount of examples with true labellings. Thus, these problems are more usefully solved using fully supervised approaches. On the contrary, in the localization problem the training should fit to the specific acoustic environment in which the measurements are obtained, thus, we cannot create a general database that corresponds to all possible acoustic scenarios. Instead, the training set should be generated individually for each acoustic environment. To obtain labelled data, one needs to generate recordings in a controlled manner and calibrate each of them precisely. Generating a large amount of labelled data is a cumbersome and impractical process. However, unlabelled data is freely available since it can be collected whenever someone is speaking. This greatly motivates the use of semi-supervised approaches, which mostly rely on unlabelled data, for the source localization problem. Another motivation is related to the special characteristics of the acoustic environment. As will be further elaborated in this paper, the unlabelled data can be utilized for forming a data-driven model of the acoustic environment that is very useful for performing robust source localization.

To address the limitations of the previous diffusion-based approaches, and to better utilize the unlabelled data, we propose the manifold regularization for localization (MRL) algorithm based on two- microphone measurements. The method recovers the inverse mapping between the RTFs and their corresponding locations. In this approach we assume that the RTF samples which originally have a high dimensional representation, pertain to a low dimensional manifold. This assumption is supported by the fact that any relative impulse response has a certain structure of a decaying exponential envelope which consists of different reflections. The gist of the algorithm is based on the concepts of manifold regularization on a reproducing kernel Hilbert space (RKHS), introduced by Belkin *et al.* [31]. The idea is to extend the standard supervised estimation framework by adding an extra regularization term which imposes a smoothness constraint on possible solutions with respect to a data-driven model. The model is learned empirically by forming a data adjacency graph over both labelled and unlabelled training samples. In this approach, the estimated location relies not only on the labelled samples, but also on the unlabelled ones. Moreover, in order to

efficiently utilize unlabelled samples received during runtime, we propose an adaptive implementation. The Manifold Regularization for Localization (MRL) algorithm iteratively updates the system, based on the new information which becomes available while accumulating new unlabelled data. We compare the proposed algorithm, with the diffusion distance search (DDS) method, which is a diffusion-based algorithm. The discussion is supported by an experimental study based on simulated data.

The paper is organized as follows. In Section II, we formulate the problem in a general noisy and reverberant environment. We motivate the choice of the RTF for forming a feature vector and describe how it can be estimated based on the microphone measurements. In Section III, we discuss the existence of an acoustic manifold and formulate an optimization problem which relies on a data-driven model computed based on both labelled and unlabelled data. This formulation leads to the MRL algorithm which is sequentially adapted by the unlabelled data accumulated during runtime. We briefly describe our previous localization method based on the diffusion framework [30] in Section IV. Accordingly, we describe the derivation of the Diffusion Distance Search (DDS) algorithm which conducts a neighbors' search using the diffusion distance as an affinity measurement between RTFs. In Section V, we demonstrate the algorithms' performance by an extensive simulation study. A comparison between the MRL and the DDS algorithms is carried out in Section VI. Section VII concludes the paper.

II. PROBLEM FORMULATION

We consider a standard enclosure, e.g., a conference room or a car interior, with a moderate reverberation time. A single source located at $\mathbf{p} = [p_x, p_y, p_z]^T$ generates an unknown speech signal $s(n)$, which is received by a pair of microphones. The received signals, denoted by $x(n)$ and $y(n)$, are contaminated by an additive stationary noise, and are given by:

$$x(n) = a_1(n, \mathbf{p}) * s(n) + u_1(n) \qquad (1)$$

$$y(n) = a_2(n, \mathbf{p}) * s(n) + u_2(n) \qquad (2)$$

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. Linear convolution is denoted by $*$. Each of the AIRs consists of the direct path between the source and the microphone, as well as reflections from the surfaces characterizing the enclosure (both early reflections and diffuse field components). Consequently, even in moderate reverberation conditions, the AIR is typically modelled as a long FIR filter.

The purpose is to localize the speaker based on the current received microphone signals $x(n)$ and $y(n)$. We assume that we are also given a set of prerecorded representative samples from the region of interest. The training set is composed of N samples of measured signals $\{\bar{x}_i(n), \bar{y}_i(n)\}_{i=1}^N$ from various positions within the specified region. Only l samples among the set are labelled, i.e., their originating position $\bar{\mathbf{p}}_i$ is known. The rest $u = N - l$ samples are unlabelled, namely, their corresponding source locations are unknown. To summarize, the training set

is composed of l labelled examples $\{\bar{x}_i(n), \bar{y}_i(n), \bar{p}_i\}_{i=1}^l$ and u unlabelled examples $\{\bar{x}_i(n), \bar{y}_i(n)\}_{i=l+1}^N$.

We are interested in a realistic scenario, where the amount of labelled data is significantly smaller than the amount of unlabelled data which can be collected online. Our goal is to build an on-line system which is initially given a small amount of labelled data, and is gradually adapted as new unlabelled samples are acquired.

The first step is to define an appropriate feature vector that faithfully represents the characteristics of the acoustic path and is invariant to the other factors, i.e., the stationary noise and the varying speech signals. For this purpose, we use an equivalent representation of (1) and (2), defined using the relative impulse response $h(n, \mathbf{p})$, which satisfies $a_2(n, \mathbf{p}) = h(n, \mathbf{p}) * a_1(n, \mathbf{p})$. Following this definition and substituting (1) into (2), we obtain an equivalent formulation, which directly relates the measurements in two microphones [32]

$$y(n) = h(n, \mathbf{p}) * x(n) + v(n)$$

$$v(n) = u_2(n) - h(n, \mathbf{p}) * u_1(n). \qquad (3)$$

In this representation, the relative impulse response $h(n, \mathbf{p})$ represents the system relating the measured signal $x(n)$ as an input and the measured signal $y(n)$ as an output. For convenience, we use the Fourier transform of the relative impulse response, which is termed the RTF, and satisfies $H(k, \mathbf{p}) = A_2(k, \mathbf{p})/A_1(k, \mathbf{p})$, where $A_1(k, \mathbf{p})$ and $A_2(k, \mathbf{p})$ are the ATFs of the respective AIRs, and k denotes a discrete frequency index. Using the PSD and cross power spectral density (CPSD) of the measured signals $y(n)$ and $x(n)$, defined in (1) and (2), the RTF $H(k, \mathbf{p})$ is obtained by:

$$H(k, \mathbf{p}) = \frac{S_{yx}(k, \mathbf{p})}{S_{xx}(k, \mathbf{p}) - S_{u_1 u_1}(k)} \quad k = 0, \ldots, D - 1$$

$$= \frac{S_{ss}(k) A_2(k, \mathbf{p}) A_1^*(k, \mathbf{p})}{S_{ss}(k) |A_1(k, \mathbf{p})|^2} = \frac{A_2(k, \mathbf{p})}{A_1(k, \mathbf{p})} \qquad (4)$$

where $S_{yx}(k, \mathbf{p})$ is the CPSD between $y(n)$ and $x(n)$, $S_{xx}(k, \mathbf{p})$ is the PSD of $x(n)$, $S_{u_1 u_1}(k)$ is the PSD of the noise in the first microphone $u_1(n)$, and $S_{ss}(k)$ is the PSD of the source $s(n)$. The choice of the value of D should balance the tradeoff between the correspondence with the relative impulse response length (large value) and latency considerations (small value).

Since $A_1(k, \mathbf{p})$ and $A_2(k, \mathbf{p})$ are unavailable, we estimate the RTF by

$$\hat{H}(k, \mathbf{p}) \equiv \frac{\hat{S}_{yx}(k, \mathbf{p})}{\hat{S}_{xx}(k, \mathbf{p})}. \qquad (5)$$

Note that this estimator is biased since we neglect the PSD of the noise $S_{u_1 u_1}(k)$. Alternatively, unbiased estimators can be used, such as the RTF estimator based on the non-stationarity of the speech signal [32]. However, we are not concerned with robust estimation of the RTF since we will show that the proposed method performs well using the biased RTF estimator. Accordingly, we define the feature vector $\mathbf{h}(\mathbf{p}) = [\hat{H}(0, \mathbf{p}), \ldots, \hat{H}(D-1, \mathbf{p})]^T$ as the concatenation of estimated

RTF values in the D frequency bins (corresponding to the first half of the Fourier transform due to symmetry for real-valued functions). In practice, we discard low and high frequencies in which the ratio in (5) is meaningless due to weak speech components. Since speech signals are mainly concentrated in 0.2–2.5 kHz we use only the estimated RTF values in this frequency band. For the sake of clarity, we omit the dependency on the position, and denote the RTF feature vector by $\mathbf{h}$.

III. MANIFOLD REGULARIZATION FOR LOCALIZATION

Our goal is to recover the target function which transforms each RTF to its corresponding location, based on the training set comprised of both labelled and unlabelled samples. Finding such an inverse mapping is non-trivial due to the complex nonlinear relation between the high-dimensional RTFs and the originating locations. To mitigate this problem we adopt the concepts of manifold regularization, introduced by Belkin *et al.* [31], [33], and present it in the light of the acoustic environment and, in particular, for the source localization problem at hand. It is important to note that, originally, the concepts of manifold regularization were implemented for classification, whereas, here, it is applied to the problem of source localization which is a regression problem.

Two guiding principles are in the core of the proposed method, that will be termed MRL. First, instead of using complex variational calculus for estimating the target function, we assume that the function resides in an RKHS. Due to the special characteristics of the functions belonging to the RKHS, the problem can be formulated simply as a system of linear equations. Second, we incorporate geometrical considerations, i.e., we use the information implied by the intrinsic patterns observed in the set of RTFs to build a data-driven model. Then, the solution is constrained to behave smoothly with respect to this data-driven model, representing the intrinsic structure of the RTFs.

A. The Acoustic Manifold

As mentioned in Section II, the RTFs have a high-dimensional representation in $\mathbb{C}^D$ that corresponds to the vast amount of reflections from the different surfaces characterizing the enclosure. However, the coefficients of a typical relative impulse response are confined to an exponentially decaying pattern. Furthermore, RTFs drawn from a certain region of interest in the enclosure have an even more specific structure and thus are not spread uniformly in the entire space of $\mathbb{C}^D$. Instead, they are confined to a compact manifold $\mathcal{M}$ of dimension d, which is much smaller compared to the dimension of the ambient space, i.e., $d \ll D$.

This assumption is further justified by the fact that the RTFs are influenced by only a small set of parameters related to the physical characteristics of the environment, such as: the enclosure dimensions and shape, the surfaces' materials and the positions of the microphones and the source. Moreover, we focus on a static configuration, in which the properties of the enclosure and the position of the microphones remain fixed. In such an acoustic environment, the only varying degree of freedom is the source location. Accordingly, we assume that the RTFs

can be intrinsically embedded in a low-dimensional manifold which is governed by the position of the source. The existence of such an acoustic manifold was discussed in detail in [34], and was demonstrated with respect to the DOA of the source. The main results will be briefly described in the experimental part, in Section V-B.

Roughly, we consider a manifold of reduced dimensions which may have a complex nonlinear structure. However, in small neighborhoods the manifold is locally linear, meaning that in the vicinity of each point it is flat and coincides with the tangent plane to the manifold at that point. Here, we assume that a small movement in the physical position has a little effect on the corresponding RTF, yet large movements give rise to very different RTFs. The same principle ensures the ability of acoustic echo cancellation (AEC) methods [35] to adaptively identify and track acoustic systems. Hence, the Euclidean distance can faithfully measure affinities between points that reside close to each other on the manifold. For larger scales, the Euclidean distance is meaningless, and we should rather use the geodesic distance on the manifold. However, the geodesic distance can be evaluated only when the structure of the manifold is known. In order to respect the manifold structure we will only examine local connections between points and disregard larger distances.

B. Background of RKHS

Our goal is to find the inverse-mapping function that receives an RTF sample and returns the corresponding source location. In general, estimating a function that minimizes a cost function, is a cumbersome task that requires complex mathematical tools, such as variational calculus. One simplifying approach is to assume that the target function belongs to a certain class of functions with a specific structure. For example, it can be assumed that the target function belongs to a certain space of functions, spanned by an orthogonal basis. Hence, the target function can be represented by a linear combination of the basis functions, where the weights are determined according to the projections of the function on each of the basis functions. In our case we assume that the target function belongs to an RKHS associated with a unique kernel function that evaluates each function in the space by an inner product. Rather than computing the basis functions spanning the space, we use an analogue representation with linear combinations of the kernel function. According to this representation, the problem can be converted to a simple linear estimation of a finite set of parameters.

We will first represent the kernel function and its properties, and then define the RKHS and discuss its representation by the kernel function that will be used for deriving the optimization problem in Section III-C. In Appendix VII, we show that the eigenfunctions associated with the kernel form an orthogonal basis for the RKHS, and discuss an analogue representation in terms of these basis functions.

As implied by its name, an RKHS is associated with a kernel function $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ that measures a pairwise affinity between RTFs. The kernel function must satisfy the following two conditions:

1) Symmetry: $k(\mathbf{h}_i, \mathbf{h}_j) = k(\mathbf{h}_j, \mathbf{h}_i) \quad \forall \mathbf{h}_i, \mathbf{h}_j \in \mathcal{M}$.

2) Positive semi-definite: the $n \times n$ matrix $\mathbf{K}$ with $K_{ij} = k(\mathbf{h}_i, \mathbf{h}_j)$ is positive semi-definite, for any arbitrary finite set of points $\{\mathbf{h}_i\}_{i=1}^n \in \mathcal{M}$.

Another essential requirement from the kernel is that it defines a notion of locality, determined with accordance to a scaling factor ε_k: for $\|\mathbf{h}_i - \mathbf{h}_j\| \ll \varepsilon_k$, $k(\mathbf{h}_i, \mathbf{h}_j) \to 1$, and for $\|\mathbf{h}_i - \mathbf{h}_j\| \gg \varepsilon_k$, $k(\mathbf{h}_i, \mathbf{h}_j) \to 0$. A common choice is to use a Gaussian kernel with variance ε_k^2:

$$k(\mathbf{h}_i, \mathbf{h}_j) = \exp\left\{-\frac{\|\mathbf{h}_i - \mathbf{h}_j\|^2}{\varepsilon_k^2}\right\}. \qquad (6)$$

Clearly, the Gaussian kernel is a symmetric positive semi-definite function, and satisfies the locality property. In general, other radial basis functions have this locality property and thus can be used instead of the Gaussian function.

The locality property is of major importance in our case, since the kernel receives RTFs, sampled from the manifold $\mathcal{M}$. As discussed above, the manifold is in general nonlinear and is assumed to be locally linear over small patches. Due to its property of locality, the kernel function constitutes an affinity measure that respects the manifold structure.

An RKHS, denoted as $\mathcal{H}_k$, is a Hilbert space of functions, mapping each $\mathbf{h} \in \mathcal{M}$ to $\mathbb{R}$, which is associated with a kernel k. We skip the formal definition of an RKHS (for details see [36] and [37]). Instead, we state the two main properties of an RKHS:

1) For all $\mathbf{h} \in \mathcal{M}$, $k_{\mathbf{h}}(\cdot) \in \mathcal{H}_k$
2) *The reproducing property:* for all $f \in \mathcal{H}_k$ and $\mathbf{h} \in \mathcal{M}$, $\langle f(\cdot), k_{\mathbf{h}}(\cdot) \rangle = f(\mathbf{h})$

where for each $\mathbf{h} \in \mathcal{M}$ we define the real valued function $k_{\mathbf{h}}(\cdot) \equiv k(\mathbf{h}, \cdot)$. The first property simply states that the RKHS consists of all functions defined by the kernel k at some point on the manifold. The second property implies that the kernel k has a special property that it evaluates all the functions in the space by an inner product. For example, in L_2 the delta function has the reproducing property since it evaluates all the functions in L_2: $\langle \delta(y, \cdot), f(\cdot) \rangle_{L_2} = f(y)$. Nevertheless, this does not define an RKHS, since the delta function does not belong to L_2. However, the space of bandlimited functions in the range $-a < \omega < a$, where ω is the radial frequency, is an RKHS and the reproducing kernel is the sinc function: $k_x(y) = \sin(a(y - x))/(\pi(y - x))$.

We have seen that an RKHS is associated with a unique reproducing kernel function. In the opposite direction, known as the Moore-Aronszajn theorem [36], every symmetric, positive definite kernel k defines a unique RKHS $\mathcal{H}_k$ that is given by the *completion* (an expansion that includes the limits of all Cauchy sequences) of the space of functions spanned by the set $\{k_{\mathbf{h}_i}(\cdot)\}$:

$$\left\{ f | f(\cdot) = \sum_i c_i k_{\mathbf{h}_i}(\cdot); i \in \mathbb{N}, c_i \in \mathbb{R}, \mathbf{h}_i \in \mathcal{M} \right\} \qquad (7)$$

with respect to the following inner product:

$$\langle f(\cdot), g(\cdot) \rangle = \left\langle \sum_i c_i k_{\mathbf{h}_i}(\cdot), \sum_j d_j k_{\mathbf{h}_j}(\cdot) \right\rangle$$

$$= \sum_{i,j} c_i d_j k(\mathbf{h}_i, \mathbf{h}_j). \qquad (8)$$

It can be easily verified that the two mentioned properties of an RKHS are satisfied by this definition. Obviously, the reproducing kernel belongs to the space, and the reproducing property holds, since:

$$\langle f(\cdot), k_{\mathbf{h}}(\cdot) \rangle = \left\langle \sum_i c_i k_{\mathbf{h}_i}(\cdot), k_{\mathbf{h}}(\cdot) \right\rangle$$

$$= \sum_i c_i k(\mathbf{h}_i, \mathbf{h}) = f(\mathbf{h}). \qquad (9)$$

An equivalent definition of an RKHS, based on Mercer's theorem [38], is discussed in Appendix VII. According to this view point, any function $f \in \mathcal{H}_k$ can be represented by an orthogonal basis of functions $\{\psi_i(\cdot)\}$ related to the kernel k:

$$\mathcal{H}_k = \left\{ f | f(\cdot) = \sum_i \alpha_i \psi_i(\cdot) \text{ and } \|f\|_{\mathcal{H}_k} < \infty \right\}. \qquad (10)$$

where $\| \cdot \|_{\mathcal{H}_k}^2$ is the RKHS norm that corresponds to the inner product defined in (9). To circumvent the computation of the basis functions, we use the representation of (7), in terms of the kernel function.

C. Optimization and Manifold Regularization

In this section, we present the optimization over the target function assuming that it belongs to an RKHS $\mathcal{H}_k$ with a reproducing kernel k. Formally, we search for a function $f_c : \mathbb{C}^D \to \mathbb{R}$ $c \in \{x, y, z\}$ which is the inverse mapping between an RTF and its corresponding position, i.e., $f_c(\mathbf{h}) = p_c$. In this paper, we focus on estimating one position coordinate, thus, we omit the coordinate subscript. However, the analysis, the results and the algorithm described here can be naturally extended to estimating several coordinates.

The search will be formulated by the following optimization problem:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_k} \frac{1}{l} \sum_{i=1}^l V(f(\bar{\mathbf{h}}_i), \bar{p}_i) + \gamma_k \|f\|_{\mathcal{H}_k}^2 + \gamma_M \|f\|_{\mathcal{M}}^2$$

$$(11)$$

where $\| \cdot \|_{\mathcal{H}_k}^2$ is the RKHS norm, $\| \cdot \|_{\mathcal{M}}^2$ is the *intrinsic* norm defined with respect to the manifold $\mathcal{M}$, and γ_k, γ_M are scalar parameters. The RKHS norm is the general norm defined in $\mathcal{H}_k$, while the intrinsic norm expresses the behaviour of the function with respect to the geometrical structure of the manifold.

The optimization problem consists of three components. The first term is an empirical cost function defined over the labelled samples $\{\bar{\mathbf{h}}_i\}_{i=1}^l$. The function V evaluates the extent of correspondence between the evaluations of the target function $f(\bar{\mathbf{h}}_i)$ and the true labels $\bar{p}_i$. In our case, we set the cost function to be the squared loss function $(\bar{p}_i - f(\bar{\mathbf{h}}_i))^2$. Note that while the L_2 norm is not suitable for comparing between RTFs [34], it is a reasonable choice for evaluating localization quality.

The two last terms in (11) are regularization conditions. Roughly, their role is to prevent the solution from overfitting to the labelled examples. The second term is the Tikhonov regularization which penalizes the RKHS norm of the function to impose smoothness condition in $\mathcal{H}_k$. The additional regularization

term, defined by the last term in (11), was introduced by Belkin *et al.* [31]. This is an intrinsic regularization that represents a smoothness penalty of the function with respect to the manifold $\mathcal{M}$.

One natural choice for the intrinsic norm is to measure the gradient of the function along the manifold, i.e., to measure the variability of the function with respect to small movements on the manifold. Since the manifold structure is unknown, this term should be approximated on the basis of both labelled and unlabelled samples. The training set $\{\bar{\mathbf{h}}\}_{i=1}^N$, which includes different realizations of possible acoustic paths, can be viewed as a discrete sampling of the manifold $\mathcal{M}$. The manifold can be empirically represented by a graph in which the training samples are the graph nodes, and the weights of the edges are defined according to an $N \times N$ adjacency matrix $\mathbf{W}$ between the samples:

$$W_{ij} = \begin{cases} \exp\left\{-\frac{\|\bar{\mathbf{h}}_i - \bar{\mathbf{h}}_j\|^2}{\varepsilon_w^2}\right\} & \text{if } \bar{\mathbf{h}}_j \in \mathcal{N}_i \text{ or } \bar{\mathbf{h}}_i \in \mathcal{N}_j \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

where $\mathcal{N}_j$ is a set consisting of the n nearest-neighbors of $\bar{\mathbf{h}}_j$ among $\{\bar{\mathbf{h}}_i\}_{i=1}^N$, and ε_w is a scaling factor. The set $\mathcal{N}_j$ of nearest neighbors is determined according to the regular Euclidean distance between RTFs which is reliable for small scales, and meaningless only for large distances.

The adjacency matrix $\mathbf{W}$ is used to form the graph Laplacian $\mathbf{L}$, by $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^N W_{ij}$. It can be shown, under certain conditions, that the graph Laplacian $\mathbf{L}$ converges to a differential operator on the manifold $\mathcal{M}$, as was discussed in detail in [39]–[41]. Hence, the gradient of the function along the manifold can be approximated using the graph Laplacian. Accordingly, an intrinsic measure of data-dependent smoothness is given by: $\|f\|_{\mathcal{M}}^2 = \mathbf{f}^T \mathbf{L} \mathbf{f}$, where $\mathbf{f} = [f(\bar{\mathbf{h}}_1), ..., f(\bar{\mathbf{h}}_N)]$. Thus, the optimization problem (11) can be recast as:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^l (\bar{p}_i - f(\bar{\mathbf{h}}_i))^2 + \gamma_k \|f\|_{\mathcal{H}_K}^2 + \gamma_M \mathbf{f}^T \mathbf{L} \mathbf{f}. \quad (13)$$

Further insight can be obtained by the expansion of the intrinsic regularization:

$$\mathbf{f}^T \mathbf{L} \mathbf{f} = \sum_{i,j=1}^N f(\bar{\mathbf{h}}_i) L_{ij} f(\bar{\mathbf{h}}_j)$$

$$= \sum_{i=1}^N \left(\sum_{j=1}^N W_{ij} - W_{ii}\right) f^2(\bar{\mathbf{h}}_i) - \sum_{\substack{i,j=1 \\ i \neq j}}^N W_{ij} f(\bar{\mathbf{h}}_i) f(\bar{\mathbf{h}}_j)$$

$$= \sum_{i,j=1}^N W_{ij} f^2(\bar{\mathbf{h}}_i) - \sum_{i,j=1}^N W_{ij} f(\bar{\mathbf{h}}_i) f(\bar{\mathbf{h}}_j)$$

$$= \frac{1}{2} \sum_{i,j=1}^N W_{ij} \left(f(\bar{\mathbf{h}}_i) - f(\bar{\mathbf{h}}_j)\right)^2. \quad (14)$$

Intuitively, in (14), large W_{ij}, corresponding to strong similarity between $\bar{\mathbf{h}}_i$ and $\bar{\mathbf{h}}_j$, implies a tendency of $f(\bar{\mathbf{h}}_i)$ and $f(\bar{\mathbf{h}}_j)$ to be close to each other. For this reason, a truncated kernel was chosen in (12), since it is reasonable to penalize the function only when the corresponding RTFs reside in the same local neighborhood.

Note that (13) is a semi-supervised formulation, since it involves both labelled and unlabelled samples. While the first term is merely based on the labelled samples, the last two terms are based on both labelled and unlabelled data. The two regularization parameters γ_k and γ_M balance between maximizing the correspondence to the labelled data, and maintaining low complexity of possible solutions. In some respects, both regularization terms try to relate the target function to the manifold $\mathcal{M}$ by the two different kernels defined in (6) and (12). Involving two kernels associated with different scales represents two different measurements of smoothness with respect to the manifold. Since the real structure of the manifold is unknown, the combination of both kernels is essential for obtaining a more accurate modelling of the manifold.

The representer theorem [42] states that the minimizer f^* of (13) is a linear combination of the kernel functions only in the set of labelled and unlabelled points $\{\bar{\mathbf{h}}_i\}_{i=1}^N$, i.e., it is given by:

$$f^*(\mathbf{h}) = \sum_{i=1}^N a_i k(\bar{\mathbf{h}}_i, \mathbf{h}) \quad (15)$$

where $\{a_i\}$ are the interpolation weights. In Appendix VII, we provide the proof of the theorem [31], which is derived by a simple orthogonality argument, and relies on the specific structure of the functions in $\mathcal{H}_k$ implied by (7), together with the reproducing property that uniquely characterizes the RKHS. The representer theorem dramatically simplifies the regularized optimization problem of (13) so it can be formulated as a linear optimization over a finite set of parameters $\{a_i\}$.

D. Derivation of the Localization Algorithm

In the previous section we formulated an optimization problem with manifold regularization for recovering the target function f in (13). Based on the representer theorem stated in (15), the optimization boils down to estimating the interpolation weights $\{a_i\}$. Substituting (15) in (13) yields a second-order polynomial objective function of $\mathbf{a} = [a_1, ..., a_N]^T$:

$$\mathbf{a}^* = \operatorname*{argmin}_{\mathbf{a} \in \mathbb{R}^N} \frac{1}{l} (\mathbf{q} - \mathbf{J}\mathbf{K}\mathbf{a})^T (\mathbf{q} - \mathbf{J}\mathbf{K}\mathbf{a})$$

$$+ \gamma_k \mathbf{a}^T \mathbf{K} \mathbf{a} + \gamma_M \mathbf{a}^T \mathbf{K} \mathbf{L} \mathbf{K} \mathbf{a} \quad (16)$$

where $\mathbf{K}$ is the $N \times N$ Gram matrix of k defined by $K_{ij} = k(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j)$; $\mathbf{J}$ is a $N \times N$ diagonal matrix: $\mathbf{J} = \operatorname{diag}(1, ..., 1, 0, ..., 0)$ with l ones and u zeros on its diagonal (functions as an indicator for the labelled samples in the set); and $\mathbf{q} = [\bar{p}_1, ..., \bar{p}_l, 0, ..., 0]^T$ is a label vector comprising the l known positions of the labelled samples with $q_i = 0$, for all $i > l$. Differentiating with respect to $\mathbf{a}$ and comparing to zero, yields

$$\frac{1}{l} (-\mathbf{J}\mathbf{K})^T (\mathbf{q} - \mathbf{J}\mathbf{K}\mathbf{a}) + (\gamma_k \mathbf{K} + \gamma_M \mathbf{K}\mathbf{L}\mathbf{K}) \mathbf{a} = 0. \quad (17)$$

By rearranging (17), we obtain the following linear system:

$$[\mathbf{JK} + l\gamma_k \mathbf{I}_N + l\gamma_M \mathbf{LK}]\, \mathbf{a} = \mathbf{q} \qquad (18)$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Accordingly, the interpolation weights $\mathbf{a}$ are given by:

$$\mathbf{a}^* = [\mathbf{JK} + l\gamma_k \mathbf{I}_N + l\gamma_M \mathbf{LK}]^{-1}\, \mathbf{q}. \qquad (19)$$

Thus far, the computations were carried out offline based only on the training set, composed of both labelled and unlabelled samples. The input to the algorithm is a new pair of measurements $\{x(n), y(n)\}$, generated by an unknown source from an unknown location on the manifold. The corresponding feature vector $\mathbf{h}$ is estimated according to (5). The kernel between the new sample $\mathbf{h}$ and each of the training samples $\{\bar{\mathbf{h}}_i\}_{i=1}^N$, is evaluated. The position of the new measurement is estimated according to (15) by a weighted sum of these kernel evaluations multiplied by the weights given by (19):

$$\hat{p} = f(\mathbf{h}) = \sum_{i=1}^N a_i^* k(\bar{\mathbf{h}}_i, \mathbf{h}). \qquad (20)$$

E. Adaptive MRL

In this section we summarize the algorithm and formulate it in a dual-stage structure. The algorithm is composed of two main parts: system adaptation and localization. In the adaptation stage, the interpolation weights $\mathbf{a}^*$ are computed according to (19) based on the labelled and unlabelled samples, which were collected up to this point in time. In the localization stage, we receive a new pair of measurements $\{x(n), y(n)\}$ of an unknown source from an unknown location, and estimate the corresponding position based on the weights computed in the previous stage.

We take advantage of the fact that the optimization is derived in a semi-supervised manner, and propose an adaptive version. At first, the system is initialized with a small amount of labelled and unlabelled data, and the corresponding weights are computed. When new test samples become available, their corresponding positions are estimated based on the computed weights. After acquiring q unlabelled samples (that are known to lie on the manifold), they are added to the existing set of unlabelled samples, i.e. $u \leftarrow u + q$. Then, the weights are recomputed for the new extended set, i.e., the system is adapted and the total number of weights increases by q. We continue to adapt the system whenever q additional unlabelled samples are accumulated, until we acquire a sufficient amount of unlabelled samples such that there is no additional benefit in recomputing the weights. Note that the adaptation process can potentially adjust to changes in the environmental conditions if also the labelled samples can be updated. However, this attribute was not examined in the current paper that focuses on static configurations. Examining dynamic scenarios with changing environmental conditions is left for future work.

The proposed MRL algorithm is summarized in Algorithm 1 and is illustrated in a flow diagram in Fig. 1. The flow diagram emphasizes the duality between the two parts of the algorithm and the interaction between them. In the downward direction,

---

Algorithm 1: Manifold Regularization for Localization.

Parameters: $\varepsilon_k, \varepsilon_w, n, \gamma_k, \gamma_M$

System Adaptation:

Input : $N = l + u$ training points:
l labelled samples $\{\bar{x}_i(n), \bar{y}_i(n), \bar{p}_i\}_{i=1}^l$
u unlabelled samples $\{\bar{x}_i(n), \bar{y}_i(n)\}_{i=l+1}^N$
Output: Interpolation weights $\mathbf{a}^*$
   1) For each point estimate the corresponding RTF $\bar{\mathbf{h}}_i$
      according to (5).
   2) Construct the reproducing kernel matrix $\mathbf{K}$ and the
      adjacency matrix $\mathbf{W}$, according to (6) and (12)
      respectively, based on $\{\bar{\mathbf{h}}_i\}_{i=1}^N$.
   3) Compute the expansion weights $\mathbf{a}^*$ according to (19).

Localization:

Input : A new pair of measurements $\{x(n), y(n)\}$
produced by an unknown source from an unknown location
Output: Estimated position $\hat{p}$
   1) Estimate the corresponding RTF $\mathbf{h}$ according to (5).
   2) Compute the affinity between $\mathbf{h}$ and each of $\{\bar{\mathbf{h}}_i\}_{i=1}^N$,
      using the reproducing kernel.
   3) Estimate the new point location using the estimated
      interpolation weights: $\hat{p} = f(\mathbf{h}) = \sum_{i=1}^N a_i^* k(\bar{\mathbf{h}}_i, \mathbf{h})$.
After acquiring q new samples, return to *System Adaptation*
and add the new unlabelled samples, $u \leftarrow u + q$.

---

the model of the system derived in the adaptation part is utilized for localization. In the upward direction, the new unlabelled samples acquired in the localization stage, are propagated and utilized for system adaptation, in addition to the existing unlabelled samples. Moreover, note that the two rightmost (blue) blocks are semi-supervised whereas the rest of the blocks are unsupervised.

It should be emphasized that we do not present an update mechanism, but instead the weights are computed from scratch in each adaptation iteration. The development of a recursive version of the algorithm is left for future work.

The number q of new unlabelled samples that should be accumulated before the system is adapted, is chosen empirically to obtain satisfactory performance. Note that when a too small value is chosen, the computational complexity is increased, without gaining a significant performance improvement. Furthermore, adding only a small amount of unlabelled information does not change the weights significantly, yet it gradually affects the complexity of the solution.

IV. REVIEW OF LOCALIZATION BASED ON DIFFUSION MAPPING

In this section, we briefly review a method for semi-supervised localization that was presented in [30]. This method, that will be termed DDS, is a dual-stage approach based on the concepts of diffusion maps [43], [44]. In the first stage we recover the mapping between the original space $\mathbb{C}^D$ and the embedded space $\mathbb{R}^d$ which is governed by the controlling parameter, i.e., the position of the source. The second step is
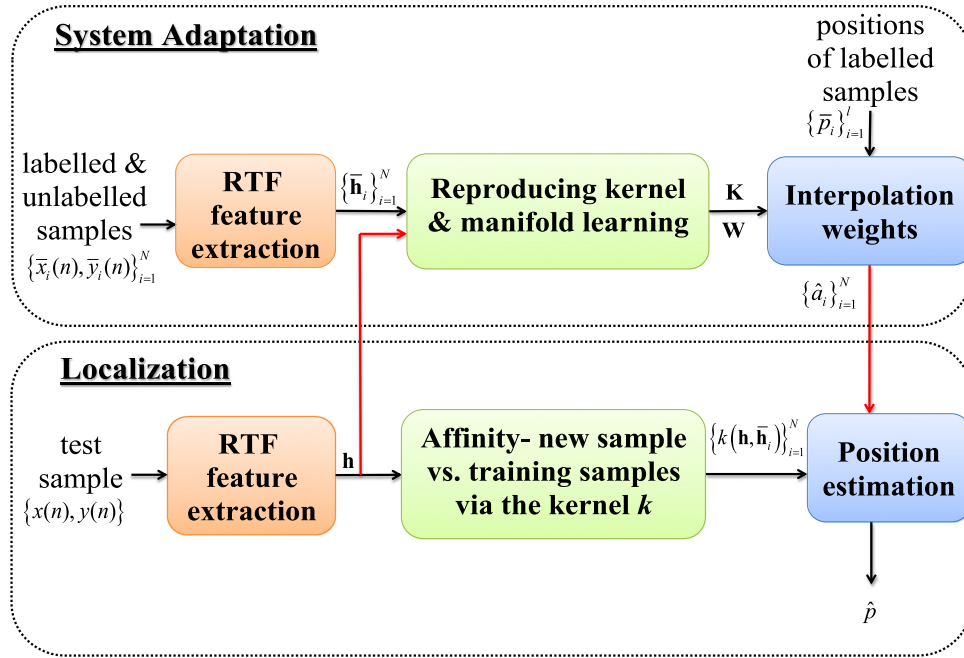
Fig. 1. Flow diagram of the proposed MRL algorithm. The algorithm consists of two parts: system adaptation and localization. In the adaptation part, both labelled and unlabelled samples are utilized to build a data-driven model for the RTFs and relate it to the position of the source. In the localization part, the position of a new pair of measurements is estimated based on the model learnt in the adaptation stage. The newly acquired unlabelled samples in the localization stage, are propagated and utilized for system adaptation.

performing the localization by searching the neighbors of the new point among the training set in the new recovered space. Note that both the MRL and DDS algorithms rely on the information implied by the manifold $\mathcal{M}$. Nevertheless, there are several fundamental aspects that distinguish between the two, as will be elaborated in Section VI.

A. Parametrization of the Manifold

In the previous section we introduced a discrete representation of the manifold by a graph in which the training samples are the graph nodes, and the weights of the edges are defined according to the adjacency matrix $\mathbf{W}$ of (12). The adjacency graph is normalized to obtain the transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, which defines a Markov process on the graph. Accordingly, $p(\mathbf{h}_i, \mathbf{h}_j) \equiv P_{ij}$ represents the probability of transition in a single Markov step from node $\mathbf{h}_i$ to node $\mathbf{h}_j$.

A nonlinear mapping of the samples into a new embedded space is obtained by a spectral decomposition of the transition matrix $\mathbf{P}$. The embedding is based on a parametrization of the manifold $\mathcal{M}$, which forms an intrinsic representation of the data. We apply a singular value decomposition to the transition matrix $\mathbf{P}$, and obtain a set of the N principal right-singular vectors $\{\boldsymbol{\varphi}_j\}_{j=0}^{N-1}$, and N singular values $\{\lambda_j\}_{j=0}^{N-1}$. The d principal right-singular vectors that correspond to the d largest singular values, form the diffusion mapping of the samples into an Euclidean space $\mathbb{R}^d$, defined by

$$\boldsymbol{\Phi}_d : \mathbf{h}_i \mapsto \left[\lambda_1 \varphi_1^{(i)}, \ldots, \lambda_d \varphi_d^{(i)}\right]^T \qquad (21)$$

where $\varphi_k^{(i)}$ denotes the ith entry of the vector $\boldsymbol{\varphi}_k$. Usually, $\boldsymbol{\varphi}_0$ is ignored since it is equal to a column vector of ones.

In the localization stage, the embedding should be extended, given a new RTF sample $\mathbf{h}$, corresponding to a new pair of measurements $\{x(n), y(n)\}$ produced by an unknown source from unknown location. Further spectral decomposition is unnecessary according to Nyström extension [45]. The new spectral coordinates are obtained by:

$$\varphi_j^* = \frac{1}{\lambda_j} \mathbf{b}^T \boldsymbol{\varphi}_j \quad j \in \{1, \ldots, d\} \qquad (22)$$

where $\mathbf{b}$ is an affinity vector between the training set and the new test point:

$$b_i = \exp\left\{ -\frac{\|\bar{\mathbf{h}}_i - \mathbf{h}\|^2)}{\varepsilon_w^2} \right\}. \qquad (23)$$

B. Nearest Neighbor Search on the Manifold

In Section III-A we described the structure of the acoustic manifold $\mathcal{M}$ of the RTFs. We stated that in order to properly measure affinities between RTFs, we should use the geodesic distance, which is the shortest path on the manifold. An approximation of the geodesic distance is given by diffusion distance, defined as:

$$D_{\text{Diff}}^2 (\mathbf{h}_i, \mathbf{h}_j) = \|p(\mathbf{h}_i, \cdot) - p(\mathbf{h}_j, \cdot)\|_{\phi_0}^2$$

$$= \sum_{r=1}^{N} \left(p(\mathbf{h}_i, \mathbf{h}_r) - p(\mathbf{h}_j, \mathbf{h}_r)\right)^2 / \phi_0^{(r)}$$

where ϕ_0 is the most dominant left-singular vector of $\mathbf{P}$.

The diffusion distance incorporates information of the entire set to determine the connectivity between pairs of samples on the graph. Pairs of points that are closely related to the same subset of points in the graph, are considered close to each other and vice versa. It can be shown that the diffusion distance is equal to the Euclidean distance in the diffusion maps space when using all N eigenvectors. This equivalence emphasizes the virtue of the diffusion mapping as it indicates that the mapping preserves the affinity between points with respect to the manifold. The diffusion distance can be well approximated by only the first d principal eigenvectors [43], i.e.,

$$D_{\text{Diff}}(\mathbf{h}_i, \mathbf{h}_j) \cong \|\mathbf{\Phi}_d(\mathbf{h}_i) - \mathbf{\Phi}_d(\mathbf{h}_j)\|. \tag{24}$$

Equipped with the ability to measure distances along the manifold using the diffusion distance, we are able to properly quantify the affinities between RTF samples. Samples which reside next to each other on the manifold, are assumed to be physically adjacent, i.e., they are likely to represent sources from close positions. Thus, the position of a new sample can be estimated by searching for its neighbors on the manifold. Accordingly, the estimate will be formulated as a weighted sum of the positions of the labelled samples, where the weights are proportional to the corresponding diffusion distance between the new sample and each of the labelled samples:

$$\hat{p} = \sum_{i \in \mathcal{N}_i} \gamma\left(\bar{\mathbf{h}}_i\right) \bar{p}_i \tag{25}$$

where the weights $\gamma\left(\bar{\mathbf{h}}_i\right)$ are given by:

$$\gamma\left(\bar{\mathbf{h}}_i\right) = \frac{\exp\left\{-D_{\text{Diff}}^2\left(\mathbf{h}, \bar{\mathbf{h}}_i\right)/\varepsilon_\gamma^2\right\}}{\sum_{j \in \mathcal{N}_i} \exp\left\{-D_{\text{Diff}}^2\left(\mathbf{h}, \bar{\mathbf{h}}_j\right)/\varepsilon_\gamma^2\right\}}. \tag{26}$$

where ε_γ is a scaling factor. The DDS procedure is summarized in Algorithm 2. Here, n_1 represents the number of neighbors in (12), and n_2 represents the number of neighbors in (25).

Note that both labelled and unlabelled samples participate in the first stage, for the construction of the graph Laplacian. However, in the localization stage only the labelled samples are utilized because we rely on the labellings. Though both the MRL and the DDS algorithms have evident similarities, we show in the experimental part that the later is inferior due to its different utilization of unlabelled data.

V. EXPERIMENTAL RESULTS

A. Setup

We describe the simulated setup used for conducting the experimental study. We simulated a $6 \times 6.2 \times 3$ m room, using an efficient implementation [46], of the image method [47]. In the room there are two microphones located at $(3, 3, 1)$ m and $(3.2, 3, 1)$ m, respectively. The source is known to be positioned at 2 m distance with respect to the first microphone, on the same height. The goal is to recover the azimuth angle of the source. The initial analysis and examination of algorithms is carried out assuming that the azimuth angle of the source is ranging between $10° \div 60°$. Then, the algorithm performance is further demonstrated on a wider range of azimuth angles between $0° \div 180°$. Fig. 2 illustrates the simulation setup.

Algorithm 2: DDS

Parameters: ε_w, ε_γ, n_1, n_2, d
Diffusion Mapping:
Input : $N = l + u$ training points:
l labelled samples $\{\bar{x}_i(n), \bar{y}_i(n), \bar{p}_i\}_{i=1}^l$
u unlabelled samples $\{\bar{x}_i(n), \bar{y}_i(n)\}_{i=l+1}^N$
Output: Embedding $\mathbf{\Phi}_d(\cdot)$
  1) For each point estimate the corresponding RTF $\bar{\mathbf{h}}_i$ according to (5).
  2) Construct the graph $\mathbf{W}$ based on $\{\bar{\mathbf{h}}_i\}_{i=1}^N$, and form the transition matrix $\mathbf{P}$.
  3) Employ singular value decomposition of $\mathbf{P}$ and obtain the singular-values $\{\lambda_j\}$ and the right-singular vectors $\{\boldsymbol{\varphi}_j\}$.
  4) Construct the map $\mathbf{\Phi}_d$ according to (21) to obtain an embedding that represents the intrinsic structure of the manifold $\mathcal{M}$.

Localization:
Input : A new pair of measurements $\{x(n), y(n)\}$ produced by an unknown source from an unknown location
Output: Estimated position $\hat{p}$
  1) Estimate the corresponding RTF $\mathbf{h}$ according to (5).
  2) Apply Nyström extension according to (22) to obtain the spectral coordinates of $\mathbf{h}$.
  3) Compute the approximated diffusion distance between $\mathbf{\Phi}_d(\mathbf{h})$ and each of the labelled samples $\{\mathbf{\Phi}_d(\bar{\mathbf{h}}_i)\}_{i=1}^N$, according to (24).
  4) Estimate the new point location by (25) as a linear combination of the positions of the labelled samples according to distances in the diffusion mapped space.
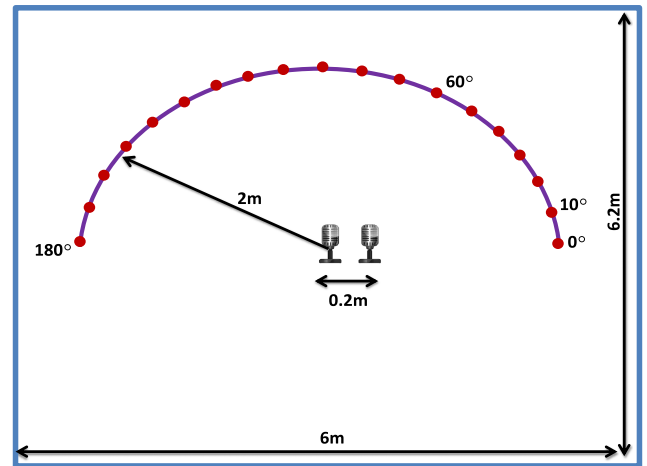

Fig. 2. An illustration of the room setup. The purple arc marks the region where the source is assumed to be positioned. The red dots define the grid of the labelled examples.

For each location, we simulate a unique 3 s speech signal, sampled at 16 kHz. The clean speech is convolved with the corresponding AIR and is contaminated by a WGN. This forms the measured signals in the two microphones. For each source location, the CPSD and the PSD are estimated with Welch's method with 0.128 s windows and 75% overlap and are utilized
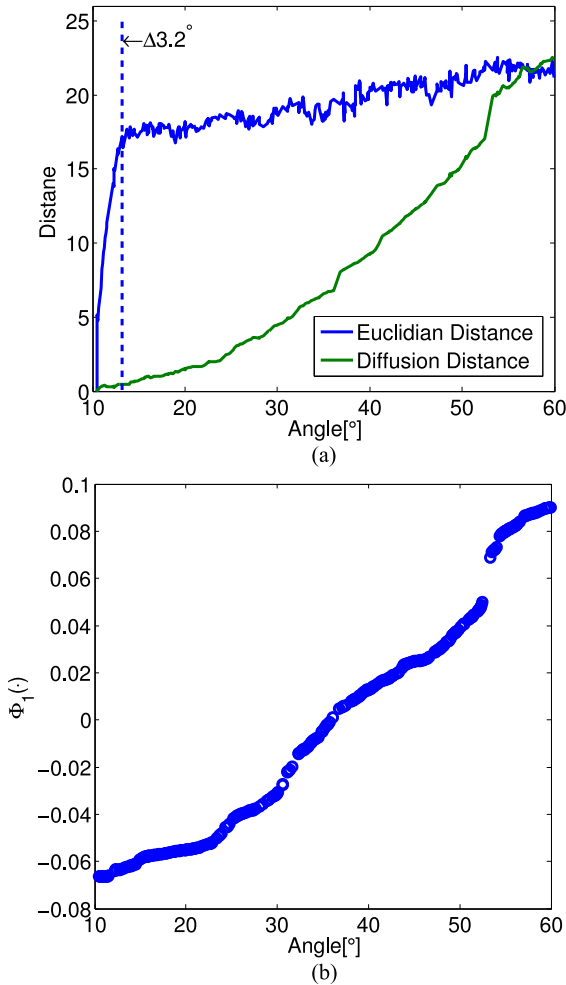
Fig. 3. (a) The Euclidean distance and the diffusion distance between each of the RTFs and the RTF corresponding to $10°$, as a function of the angle. The dashed line shows the boundary angle until which monotonicity is preserved for the Euclidean distance. (b) Single-element diffusion mapping $\Phi_1(\cdot)$.

for estimating the RTF in (5) for 2048 frequency bins. The RTF vector consists of $D = 286$ frequency bins corresponding to 0.2–2.5 kHz, in which most of the speech components are concentrated.

B. Analysis of the Manifold

In this section we review the main results presented in [34]. We investigate the acoustic manifold of the RTFs and examine the proper distance between them that maintains physical adjacency. The analysis is carried out using a set of $N = 400$ RTF samples, corresponding to 400 positions distributed uniformly in the specified range, between $10° \div 60°$ at 2 m distance from the first microphone. Two alternative distance measures for quantifying the affinity between different RTFs, are addressed. We start with the Euclidean distance defined by:

$$D_{\mathrm{Euc}}(\mathbf{h}_i, \mathbf{h}_j) = \|\mathbf{h}_i - \mathbf{h}_j\|. \qquad (27)$$

The Euclidean distance is compared with the diffusion distance presented in Section IV-B.

Fig. 3(a) depicts the Euclidean distance and the diffusion distance between each of the RTFs and a reference RTF cor-

responding to $10°$, as a function of the angle. We used a moderate reverberation time of 300 ms and WGN of 20 dB SNR. We observe that the monotonic behaviour of the Euclidean distance with respect to the angle is confined to approximately $3.2°$ range. Consequently, we conclude that the Euclidean distance is meaningful only for small arcs. Thus, in general the Euclidean distance is not a good distance measure between RTFs. However, it can be properly utilized when inserted into a Gaussian kernel in either the manifold regularization framework or the diffusion framework. According to its scaling parameter, the Gaussian kernel preserves small distances and suppresses large distances which are meaningless. The kernel scale should be adjusted to the distance at which monotonicity is maintained by the Euclidean distance, in order to preserve locality.

For the diffusion distance, only the first element in the mapping ($d = 1$) was considered. This choice will be justified in the sequel. We can see that for almost the entire range, the diffusion distance remains monotonic with respect to the angle, indicating that it is an appropriate metric in terms of the source DOA. Further insight into the mapping itself, is gained by plotting the single-element mapping $\Phi_1(\cdot)$, as depicted in Fig. 3(b). We observe that the mapping corresponds well with the angle up to a monotonic distortion. Thus, the diffusion mapping successfully reveals the latent variable, namely, the position of the source. The almost perfect matching between the first element of the mapping and the corresponding angle, justifies the use of $d = 1$ for estimating the diffusion distance.

To summarize, the presented results strengthen the claim on the existence of a nonlinear acoustic manifold. In small neighborhoods around each point, the manifold is approximately flat, meaning that it resembles an Euclidean (linear) space. For larger scales the affinity between RTFs should be determined according to the geodesic distance on the manifold. The diffusion framework successfully reveals the latent variable controlling the acoustic manifold, and the diffusion distance properly reflects the distances on the manifold. These results motivate the involvement of manifold aspects in the localization process, as introduced by either the MRL or the DDS algorithms.

C. Localization Results

In this section we examine the ability of both DDS and MRL to recover the DOA of the source. The training set consists $N = 400$ representative samples distributed uniformly between $10° \div 60°$. Among the training set, only $l = 6$ samples were labelled, creating a grid with approximately $10°$ distance between adjacent labelled samples, as depicted in Fig. 2. The performance is examined on a set of $T = 120$ additional samples produced by unknown sources from unknown locations, confined to the defined range between $10° \div 60°$. The performance is measured according to the root mean square error (RMSE), defined by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{T} \sum_{i=1}^{T} \|p_i - \hat{p}_i\|^2} \qquad (28)$$

where p stands for the azimuth angle of the source. To prevent the results from being dependent on a specific reflection pattern of a

TABLE I
PARAMETERS VALUES FOR THE MRL AND THE DDS ALGORITHMS

MRL	ε_k	ε_w	n	γ_k	γ_M
	$1.73 \cdot \text{med}$	$0.316 \cdot \text{med}$	12	$1e^{-7}$	$1e^{-3}$
DDS	ε_w	d	n_1	n_2	ε_γ
	$0.316 \cdot \text{med}$	3	400	3	$1e^{-4}$

certain room section, we repeated the simulation with rotations of the constellation described above. The rotation angle was generated uniformly between $0° \div 360°$. The positions of the second microphone, the training points and the test points were rotated by this angle, with respect to the first microphone. The RMSE was averaged over 50 rotations of the constellation.

The values of the parameters for the MRL and DDS algorithms are given in Table I, where "med" stands for the median of the distances between all the training samples. In general, we chose the variances of all the Gaussians to be proportional to the median distance "med," that roughly expresses the limit for which the distances are monotonic with respect to the angle, as discussed in Section V-B. The exact proportion was determined by an exhaustive search, and a similar search was conducted for determining the values of all the other parameters as well.

The results of the MRL and the DDS algorithms are compared with that obtained by the classical GCC algorithm [12] for both noisy and reverberant conditions. The GCC algorithm is based on finding the maximum of the generalized cross-correlation function, defined as:

$$\hat{R}_{yx}^g(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_g(\omega)\hat{S}_{yx}(\omega)e^{j\omega\tau}d\tau \qquad (29)$$

where τ is a time-lag and $\Psi_g(\omega)$ is a weight function. We used the generalized cross-correlation-phase transformation (GCC-PHAT) variant, which is associated with the weight function $\Psi_g(\omega) = 1/|\hat{S}_{yx}(\omega)|$, since it is considered more robust to reverberation. The CPSD between the two measurements was estimated using the same parameters as defined in Section V-A.

For comparing between the three algorithms we used different levels of diffuse noise with speech-like PSD. A spherically diffused noise field was generated according to [48], using the simulator provided in [49]. In the first scenario we examine the algorithms' performance for different reverberation times with fixed SNR of 5 dB and 20 dB. In the second scenario the reverberation time is set to 300 ms and 600 ms, and different noise levels are examined. In all scenarios the training set is generated with a fixed SNR level of 10 dB and with the same reverberation time as that used in the test phase. The RMSE of the three algorithms in both scenarios, are shown in Fig. 4(a) and (b), respectively.

It can be seen in Fig. 4(a) that the GCC-PHAT performs well for low reverberation. However, its performance deteriorates gradually as reverberation increases, and becomes
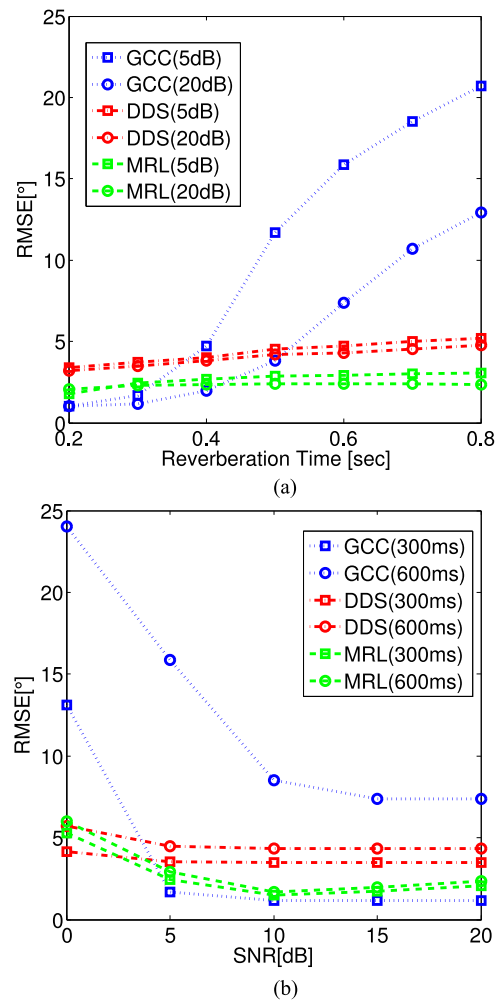

Fig. 4. The RMSE of GCC-PHAT, DDS and MRL (a) as a function of the reverberation time (SNR=5, 20 dB), and (b) as a function of SNR (T_{60} = 300, 600 ms).

inferior compared with the performance of both the DDS and the MRL algorithms. In high reverberation, the GCC-PHAT is incapable of distinguishing between the direct arrival and the reflections. A misidentification of the direct path, results in a large estimation error. The proposed algorithms are more robust to reverberation, since the variations in the entire RTFs are taken in account. In addition, the GCC is significantly affected by the amount of noise, whereas the semi-supervised approaches are almost invariant regarding the noise level.

Similar behaviour is observed in Fig. 4(b) in which different noise levels are examined. Here too, the GCC-PHAT method behaves well in high and moderate SNR conditions, and its performance degrades as noise level increases. When the measurements are contaminated by a significant amount of noise, the correlation between the two measurements is also very noisy, and the GCC-PHAT cannot correctly identify the peak corresponding to the direct path. On the contrary, the semi-supervised algorithms are much more robust with respect to the background noise, and most of the time obtain lower error. These type of algorithms can compensate for the information loss caused by the poor conditions, by capitalizing on the prior information inferred from the training samples.
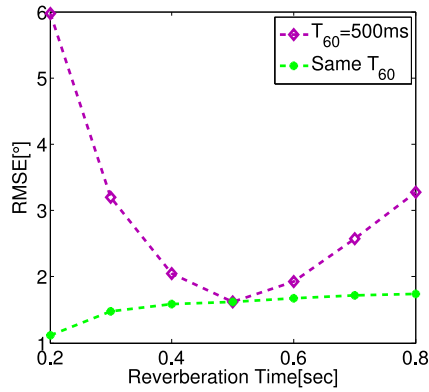
Fig. 5. The RMSE of the MRL algorithm when trained with $T_{60} = 500$ ms, and when trained with the same T_{60} as in the test stage (SNR=10 dB for both the training and the test).

We also observe that most of the time the MRL approach exhibits better results compared with the DDS method. The reason for the visible gap between the RMSEs of the two algorithms is related to the different ways they utilize unlabelled data, and will be further elaborated in Section VI. In addition, in Fig. 4(b) we observe that the minimum error for the MRL method, is obtained for an SNR of 10 dB, when the noise level in the test samples matches exactly the noise level used for training. However, in general the algorithm performs well even when the test and the training do not meet exactly the same noise conditions.

Another aspect to be investigated is the algorithm sensitivity to changes in the environmental conditions between the training and the test stages. In Fig. 5, we see the RMSE of the MRL algorithm for different reverberation times, when trained with a fixed reverberation time of 500 ms. For comparison, we also present the error obtained for each reverberation time when the test samples are measured at the same reverberation level as in the training (as was done in the previous simulations). It can be observed that for small differences in the reverberation time of around 100 ms, the increase in the error is quite small (less then $0.5°$). However, as the difference in the reverberation time between the training and the test increases, the error deviates significantly from that obtained ideally when the same reverberation time is used for both the training and the test. From the manifold perspective, slight variations in the reverberation time have a mild effect on the manifold structure, while significant variations correspond to large differences in the geometry of the associated manifolds.

Finally, we examine the iterative process of the MRL algorithm through the following sequential simulation. We used reverberation time of 300 ms and diffuse noise with 20 dB SNR. This time we examined a wider range of angles between $0° \div 180°$. The initial adaptation was based on only 19 labelled samples, creating a grid of $10°$ distance between adjacent labelled samples, as depicted in Fig. 2. We conducted 10 cycles of the sequential algorithm, each comprised of both stages of system adaptation and localization. In the localization stage, we estimated the angles of 80 new samples from unknown locations. The total RMSE of the entire set was computed. In the following iteration, these 80 new samples were treated as
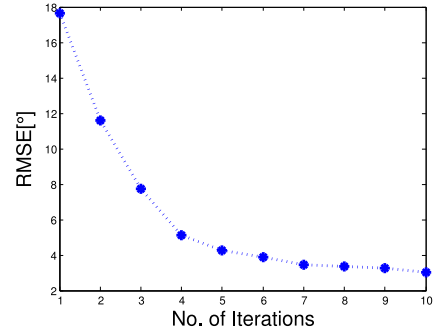

Fig. 6. The RMSE of an iterative simulation of MRL for angles in the range $0° \div 180°$, where 80 unlabelled points are added in each iteration. $T_{60} = 300$ ms and SNR=20 dB
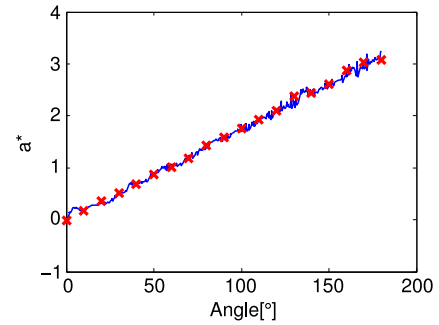

Fig. 7. The estimated expansion weights $\mathbf{a}^*$ with respect to the corresponding angle. The blue line corresponds to the weights of $u = 441$ unlabelled examples, while the red x-marks correspond to the weights of $l = 19$ labelled examples.

additional unlabelled data, utilized for system adaptation. The results are summarized in Fig. 6.

In this figure we observe that the RMSE decreases as a function of the number of iterations, indicating that the unlabelled data has an important role in reducing the estimation error. However, after a considerable amount of unlabelled data is accumulated, the process stabilizes on a certain error, and additional samples are redundant.

VI. DISCUSSION

In the previous section we demonstrated the robustness of the MRL and the DDS algorithms to noisy and reverberant conditions. We have also seen that the performance of the DDS method is inferior with respect to that of the MRL algorithm. In this section we discuss the interfacing points of both algorithms, on the one hand, and highlight the major differences between them, on the other hand.

To investigate the role of the unlabelled data in the MRL method, we inspect the expansion weights $\mathbf{a}^*$ derived by the algorithm when using the kernel in (6) normalized by the sum of its rows. In Fig. 7, the N entries of the vector $\mathbf{a}^*$ in (15), are depicted with respect to the associated angle. The red x-marks correspond to the weights $\{a_i\}_{i=1}^{l}$ of $l = 19$ labelled examples while the blue line corresponds to the weights $\{a_i\}_{i=l+1}^{u}$ of $u = 441$ unlabelled examples. We observe a monotonic, almost linear, behaviour of the coefficients with respect to the angle. The obtained behaviour of the MRL coefficients, resembles the monotonic relation between the single-element diffusion mapping $\mathbf{\Phi}_1(\cdot)$ and the corresponding angle, depicted in Fig. 3(b).

The correspondence between the two algorithms, suggests that they share similar aspects which lead to a parametrization of the manifold and recovery of the DOA of the source.

However, we have seen that the MRL is a better localizer compared with the DDS. The difference between the two, is attributed to their different utilization of the unlabelled data. In the DDS algorithm, the unlabelled data are used only in the learning phase, and the estimation merely comprises the positions of the labelled samples. In contrast, in MRL the unlabelled data do not only take part in the recovery of the manifold, but also participate in the estimation itself, involving both labelled and unlabelled data (15). Another advantage of MRL over DDS is that it is sequentially updated, hence, it is more suitable for on-line implementations.

VII. CONCLUSION

A novel approach for semi-supervised localization, based on state-of-the-art manifold learning techniques, was presented. A set of representative samples in a defined room section is utilized for learning the acoustic manifold of the RTFs and building a data-driven model. Equipped with this knowledge, we find the function relating the samples and the corresponding positions by solving an optimization problem in an RKHS. The optimization is based on manifold regularization and involves smoothness constraints of possible solutions with respect to the manifold. According to the representer theorem the solution is given as an expansion in terms of a kernel function sampled at the training points. The implementation of the proposed algorithm is formulated in an adaptive manner. The initialization is conducted with only few labelled samples, and then the system is gradually adapted as new unlabelled samples are received. Simulation results confirm the algorithm robustness in noisy and reverberant environments. It is important to note that the results of the algorithm were only demonstrated in a simulated room setup, while the performance on real recordings as well as the examination of other complex environments, are left for future work.

Integrating between traditional signal processing techniques and novel machine learning tools may be the key for better addressing adverse conditions, such as high noise levels and reverberations, that are the main causes for performance degradation of classical localization approaches. The current results indicate that the manifold perspective exhibits an interesting insight into the general structure of the acoustic responses and offers better solutions for common signal processing problems.

APPENDIX A

We define the integral operator on functions, associated with the kernel k, by the following integral transform:

$$[T_k f] = \int k(\mathbf{t}, \mathbf{s}) f(\mathbf{s}) d\mathbf{s} = g(\mathbf{t}). \tag{30}$$

The eigenfunctions $\{\psi_i(\cdot)\}$ and eigenvalues $\{\lambda_i\}$ of the integral operator satisfy:

$$[T_k \psi_i] = \int k(\mathbf{t}, \mathbf{s}) \psi_i(\mathbf{s}) d\mathbf{s} = \lambda_i \psi_i(\mathbf{t}). \tag{31}$$

According to Mercer's theorem [38], the kernel k can be expanded by:

$$k(\mathbf{t}, \mathbf{s}) = \sum_i \lambda_i \psi_i(\mathbf{t}) \psi_i(\mathbf{s}) \tag{32}$$

where the convergence is absolute and uniform. The eigenfunctions $\{\psi_i(\cdot)\}$ form an orthogonal set and the RKHS can be defined as the space of functions spanned by this set:

$$\mathcal{H}_k = \left\{ f | f(\cdot) = \sum_i \alpha_i \psi_i(\cdot) \text{ and } ||f||_{\mathcal{H}_k} < \infty \right\} \tag{33}$$

where the RKHS norm is defined by the inner product:

$$\langle f, g \rangle = \left\langle \sum_i \alpha_i \psi_i(\cdot), \sum_j \beta_j \psi_i(\cdot) \right\rangle = \sum_i \frac{\alpha_i \beta_i}{\lambda_j}. \tag{34}$$

The reproducing property holds in this representation, since

$$\langle f(\cdot), k_{\mathbf{h}}(\cdot) \rangle = \left\langle \sum_i \alpha_i \psi_i(\cdot), \sum_j \lambda_j \psi_j(\mathbf{h}) \psi_j(\cdot) \right\rangle$$

$$= \sum_i \sum_j \alpha_i \lambda_j \psi_j(\mathbf{h}) \langle \psi_i(\cdot), \psi_j(\cdot) \rangle$$

$$= \sum_i \alpha_i \psi_j(\mathbf{h}) = f(\mathbf{h}) \tag{35}$$

APPENDIX B

Theorem 1: The minimizer of the optimization problem (13) has an expansion in terms of labelled and unlabelled examples:

$$f^*(\mathbf{h}) = \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \mathbf{h}) \tag{36}$$

Proof: Any function $f \in \mathcal{H}_k$ can be uniquely decomposed into two components, where one is lying in the linear subspace spanned by the kernel functions in the training examples $f_\parallel = \text{span} \{k(\bar{\mathbf{h}}_i, \cdot), \ i = 1, \dots, N\}$ and the other is lying in the orthogonal complement $f_\perp$

$$f = f_\parallel + f_\perp = \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \mathbf{h}) + f_\perp \tag{37}$$

where $\langle f_\perp, k(\bar{\mathbf{h}}_j, \cdot) \rangle = 0$ for all $1 \le j \le N$.

The above orthogonal decomposition and the reproducing property together, show that the evaluation of f on any training point $\bar{\mathbf{h}}_j$, $1 \le j \le N$ is independent of the orthogonal component $f_\perp$:

$$f(\mathbf{h}_j) = \langle f(\cdot), k(\bar{\mathbf{h}}_j, \cdot) \rangle$$

$$= \left\langle \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \cdot) + f_\perp(\cdot), k(\bar{\mathbf{h}}_j, \cdot) \right\rangle$$

$$= \left\langle \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \cdot), k(\bar{\mathbf{h}}_j, \cdot) \right\rangle = \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j). \tag{38}$$

Consequently, the value of the empirical terms involving the loss function and the intrinsic norm in the optimization problem

(the first and the third terms, respectively), are independent of $f_\perp$. For the second term (the norm of f in $\mathcal{H}_k$), since $f_\perp$ is orthogonal to $\sum_{i=1}^{N} a_i k(\mathbf{h}_i, \cdot)$ and only increases the norm of f in $\mathcal{H}_k$, we have

$$
\begin{aligned}
\|f\|_{\mathcal{H}_k}^2 &= \left\| \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \mathbf{h}) + f_\perp(\cdot) \right\|_{\mathcal{H}_k}^2 \\
&= \left\| \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \mathbf{h}) \right\|_{\mathcal{H}_k}^2 + \left\| f_\perp(\cdot) \right\|_{\mathcal{H}_k}^2 \\
&\geq \left\| \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \mathbf{h}) \right\|_{\mathcal{H}_k}^2 .
\end{aligned} \tag{39}
$$

Therefore setting $f_\perp = 0$ does not affect the first and the third terms of (13), while it strictly decreases the second term. It follows that any minimizer f^* of (13) must have $f_\perp = 0$, and therefore can be represented as: $f^*(\mathbf{h}) = \sum_{i=1}^{N} a_i k(\bar{\mathbf{h}}_i, \mathbf{h})$. ∎

REFERENCES

[1] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 2, pp. II909–II912.

[2] K. H. Knuth, "Bayesian source separation and localization," *Proc. SPIE's*, vol. 3459, pp. 147–158, 1998.

[3] T. Yamada, S. Nakamura, and K. Shikano, "Robust speech recognition with speaker localization by a microphone array," in *Proc. Int. Conf. Spoken Lang.*, 1996, vol. 3, pp. 1317–1320.

[4] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 7, pp. 1132–1143, Jul. 1990.

[5] 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 Trans. Signal Process.*, vol. 50, no. 8, pp. 1843–1854, Aug. 2002.

[6] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 1, pp. I-125–I-128.

[7] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, May 1997.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. New York, NY, USA: Springer, 2001, pp. 157–180.

[9] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[10] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[11] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 170–170, Jan. 2006.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[13] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[14] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1997, vol. 1, pp. 375–378.

[15] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Process.*, vol. 59, no. 3, pp. 253–266, 1997.

[16] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 2, pp. ii–133.

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

[18] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.

[19] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.

[20] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1110–1124, 2003.

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

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

[23] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," presented at the IEEE Int. Workshop Machine Learning Signal Processing, Santander, Spain, Sep. 2012.

[24] A. Deleforge, F. Forbes, and R. Horaud, "Variational EM for binaural sound-source separation and localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 76–80.

[25] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *Int. J. Neural Syst.*, vol. 25, no. 1, pp. 1440003-1–1440003-21, 2015.

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

[27] 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 Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 76–80.

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

[29] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 245–248.

[30] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics*, New Paltz, NY, USA, Oct. 2013.

[31] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learning Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

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

[33] M. Belkin, P. Niyogi, and V. Sindhwani, "On manifold regularization," in *Proc. 10th Int. Workshop Artificial Intell. Statist.*, 2005, pp. 17–24.

[34] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Study on manifolds of acoustic responses," *Proc. Int. Conf. Latent Variable Analysis*, Liberec, Czech Republic, Aug. 2015.

[35] C. Breining *et al.*, "Acoustic echo control. an application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, Jul. 1999.

[36] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. no. 68, pp. 337–404, 1950.

[37] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York, NY, USA: Springer Science & Business Media, 2011.

[38] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosoph. Trans. Roy. Soc. London. Series A*, vol. 209, pp. 415–446, 1909.

[39] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learning*, vol. 56, no. 1–3, pp. 209–239, 2004.

[40] R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 21, pp. 7426–7431, May 2005.

[41] M. Hein and J. Y. Audibert, "Intrinsic dimensionality estimation of submanifold in $\mathbb{R}^d$," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 289–296.

[42] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory*. New York, NY, USA: Springer, 2001, pp. 416–426.

[43] R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmonic Anal.*, vol. 21, pp. 5–30, Jul. 2006.

[44] R. Talmon, I. Cohen, S. Gannot, and R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 75–86, Jul. 2013.

[45] E. J. Nyström, *Über die praktische Auflösung von linearen Integralgleichungen mit Anwendungen auf Randwertaufgaben der Potentialtheorie*. Akademische Buchhandlung, 1929.

[46] E. A. P. Habets. (2006, Nov.). Room impulse response (RIR) generator. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, Jul. 2006.

[47] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[48] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multi-sensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[49] E. A. P. Habets. (2008, Nov.). Noise generator. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/noise-generators.

Bracha Laufer-Goldshtein (S'15) received the B.Sc. (*summa cum laude*) and the M.Sc. degrees (*cum laude*) in electrical engineering from Bar-Ilan University, Ramat-Gan, Israel, in 2013 and 2015, respectively. She is currently working toward the Ph.D. degree at the Speech and Signal Processing Laboratory, Faculty of Engineering, Bar-Ilan University. Her research interests include statistical signal processing, speaker localization, array processing, and geometric methods for data analysis.

Ronen Talmon (S'09–M'11) received the B.A. degree (*cum laude*) in mathematics and computer science from the Open University, Ra'anana, Israel, in 2005, and the Ph.D. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2011. He is currently an Assistant Professor of electrical engineering at the Technion—Israel Institute of Technology .

From 2000 to 2005, he was a Software Developer and Researcher at a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant with the Department of Electrical Engineering, Technion. From 2011 to 2013, he was a Gibbs Assistant Professor at the Department of Mathematics, Yale University, New Haven, CT, USA. In 2014, he joined the Department of Electrical Engineering, Technion.

His research interests include statistical signal processing, analysis and modeling of signals, speech enhancement, biomedical signal processing, applied harmonic analysis, and diffusion geometry.

Dr. Talmon received the Irwin and Joan Jacobs Fellowship, the Andrew and Erna Fince Viterbi Fellowship, and the Horev Fellowship.

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (*summa cum laude*) from the Technion Israel Institute of Technology, Haifa, Israel, in 1986, and the M.Sc. (*cum laude*) and Ph.D. degrees from Tel-Aviv University, Tel Aviv-Yafo, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he was a Postdoctoral Researcher at the Department of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology. He is currently a Full Professor at the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel, where he is heading the Speech and Signal Processing Laboratory and the Signal Processing Track. He received the Bar-Ilan University Outstanding Lecturer Award for 2010 and 2014. He was an Associate Editor of the *EURASIP Journal of Advances in Signal Processing* during 2003–2012, and also an Editor of several special issues on multi-microphone speech processing of the same journal. He was also a Guest Editor of *ELSEVIER Speech Communication and Signal Processing* journals. He was an Associate Editor of IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, during 2009–2013, where he is currently a Senior Area Chair. He was also a Reviewer of many IEEE journals and conferences. Since January 2010, he has been a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, where he is currently the Committee Vice-Chair. He has been a Member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005, and was the General Cochair of IWAENC held at Tel-Aviv, Israel, in August 2010. He has served as the General Cochair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), in October 2013. He was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013, and EUSIPCO 2013. His research interests include multimicrophone speech processing and specifically distributed algorithms for *ad hoc* microphone arrays for noise reduction and speaker separation, dereverberation, single microphone speech enhancement, and speaker localization and tracking.