A Real-Life Experimental Study on Semi-Supervised Source Localization Based on Manifold Regularization

Bracha Laufer-Goldshtein
Faculty of Engineering
Bar-Ilan University
Ramat-Gan, 5290002, Israel
bracha.laufer@biu.ac.il

Ronen Talmon
Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Haifa, 3200003, Israel
ronen@ee.technion.ac.il

Sharon Gannot
Faculty of Engineering
Bar-Ilan University
Ramat-Gan, 5290002, Israel
sharon.gannot@biu.ac.il

***Abstract*—Recently, we have presented a semi-supervised approach for sound source localization based on manifold regularization. The idea is to estimate the function that maps each relative transfer function (RTF) to its corresponding position. The estimation is based on an optimization problem which takes into consideration the geometric structure of the RTF samples, which is empirically deduced from prerecorded training measurements. The solution is appropriately constrained to be smooth, meaning that similar RTFs are mapped to close positions. In this paper, we conduct a comprehensive experimental study with real-life recordings to examine the algorithm performance in actual noisy and reverberant conditions. The influence of the amount of training data as well as changes in the environmental conditions are also being examined. We show that the algorithm attains accurate localization in such challenging conditions.**

I. Introduction

A large variety of applications rely on the ability to locate a sound source; examples for such applications include: automatic camera steering, video conferencing and robotics. Therefore, the problem of source localization has attracted much research attention, and various localization methods were presented along the years. Traditional localization methods can be roughly divided into three main categories. In the first category, the methods are based on maximization of the steered response power (SRP) of a beamformer output [1], [2]. Another type of approaches are based on high-resolution spectral estimation techniques, such as multiple signal classification (MUSIC) [3] and estimation of signal parameters via rotational invariance (ESPRIT) [4] algorithms. Other methods are dual-stage approaches consisting of time difference of arrival (TDOA) estimation for the different microphone pairs, which are then geometrically combined to perform the actual localization. Various methods are used for TDOA estimation, such as: the celebrated generalized cross-correlation (GCC) method [5] and its variants [6], [7] and adaptive eigenvalue decomposition [8], [9].

Recently, there is a growing interest in learning-based localization. In these approaches, it is assumed that we have prerecorded measurements in the enclosure of interest. From these representative measurements, we can learn the characteristics of the acoustic environment and infer the relations between the acoustic paths and the corresponding positions. Thus far, several attempts were made to involve training information for performing source localization [10]–[15].

In [16], we have presented a semi-supervised source localization algorithm, termed Manifold Regularization for Localization (MRL), which utilizes both labelled (attached with corresponding positions) and unlabelled measurements (from unknown locations). The prerecorded measurements are used to identify the geometrical structure of the acoustic paths, which are assumed to lie on a manifold, and to build a data-driven model. The idea is to solve a regularized optimization problem in which the solution has to fit to the labelled examples, and also to respect the data-driven model extracted from unlabelled data.

It is often claimed that in training-based approaches, the setting is unrealistic and that it requires many pre-conditions. In this paper, we perform a comprehensive experimental study based on real recordings, in order to address these claims. We examine the performance by several practical aspects: robustness to noise and reverberation, utilization of labelled and unlabelled samples and the influence of changes in the environmental conditions. The goal is to demonstrate the ability of the proposed method to locate the source in an actual noisy and reverberant environment in which the physical conditions are not completely fixed. The localization is performed using only a limited number of labelled measurements and several unlabelled measurements.

II. Problem Formulation

Consider a source located at position $\mathbf{p}$ in a reverberant enclosure, producing a speech signal $s(n)$. The signal is contaminated by noise signals and picked up by a pair of microphones located in the enclosure. The measured signals in the two microphones are:

$$y_i(n) = a_i(n, \mathbf{p}) * s(n) + u_i(n), \quad i = 1, 2 \tag{1}$$

where n is the time index, i is the microphone index, $a_i(n, \mathbf{p})$ are the corresponding acoustic impulse responses

978-1-5090-2152-9/16/$31.00

(AIRs) relating the source at position $\mathbf{p}$ and each of the microphones, and $u_i(n)$ are the noise signals. From these two measurements, we extract a power spectral density (PSD)-based feature vector $\mathbf{h}$, which represents the characteristic of the acoustic paths and is independent of the source signal. For this purpose, we estimate the RTF $H(k)$, where k is the frequency index, which is defined as the ratio between the two transfer functions relating the source and the two microphones. The feature vector $\mathbf{h} = [H(k_1), \ldots, H(k_D)]$ consists of RTF estimates in a certain frequency range. In general, the RTF is a high-dimensional vector since it represents the acoustic paths which have a long and complex nature in typical reverberant enclosures. In [17], we have shown that the distinct structure of the RTFs and the fact that they are influenced by only a small set of parameters, related to the physical properties of the environment, imply that they have a compact representation. The RTFs pertain to a low-dimensional manifold $\mathcal{M}$ which has a (possibly) nonlinear structure, but is locally linear in small patches.

To localize the source, we estimate the function f which receives an RTF sample $\mathbf{h}$, and returns the corresponding position $f(\mathbf{h})$. We concentrate on 1-dimensional localization, hence the function's output is scalar. However, the algorithm and the results can be extended to full 3-dimensional localization as well. We assume to have a training set of prerecorded measurements, which consists of a limited number of labelled samples which are attached with corresponding positions, and a larger amount of unlabelled samples from unknown locations. The labelled set consists of n_L samples $\{\mathbf{h}_i, p_i\}_{i=1}^{n_L}$, and the unlabelled set consists of n_U samples $\{\mathbf{h}_i\}_{i=n_L+1}^{n}$, where $n = n_L + n_U$. The unlabelled samples are utilized for recovering the manifold structure, while the labelled samples serve as anchor points to form the mapping between RTFs and positions. Based on the model learned from the training set we estimate the position of a new RTF sample $\mathbf{h}_t$, which is extracted from the measurements of an unknown source from an unknown location.

III. Manifold Regularization for Localization

We briefly describe the MRL algorithm presented in [16]. Let $\mathbf{W}$ be an $n \times n$ matrix in which the (i, j) entry expresses the local similarity between the RTF samples $\mathbf{h}_i$ and $\mathbf{h}_j$. The graph-Laplacian is computed by $\mathbf{L} \equiv \mathbf{S} - \mathbf{W}$, where $\mathbf{S}$ is a diagonal matrix comprising the rows of $\mathbf{W}$ on its main diagonal. Let $\mathbf{f}$ be a concatenation of the function values for all the training samples, i.e. $\mathbf{f} = [f(\mathbf{h}_1), ..., f(\mathbf{h}_n)]$. The estimation of the function f relating RTFs and positions, is formulated by the following optimization problem, presented in [18]:

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

where the search space is a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$, which is a Hilbert space in which all the functions can be represented as a linear combination of a reproducing kernel $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. The reproducing kernel has to be symmetric and positive semi-definite. The norm of a function in the RKHS $\mathcal{H}_k$ is denoted by $\| \cdot \|_{\mathcal{H}_k}$. The optimization (2) consists of three terms: a cost function that measures the correspondence between the function values and the given labelled positions and two regularization terms which are weighted by the parameters γ_k and γ_M. The role of the two regularization terms is to control the complexity of the solution of (2). The first regularization term constrains the norm of the function f, and expresses a general smoothness penalty in $\mathcal{H}_k$. The second regularization term is *manifold-regularization* that controls the smoothness of the function with respect to the manifold structure. The idea in manifold regularization is to ensure that two RTF samples which lie close to each other on the manifold, will be mapped to adjacent physical positions, i.e. their corresponding function values will be similar.

The representer theorem [19] states that the solution of (2) is given as a finite linear combination of the function values of the training samples: $f^*(\mathbf{h}_t) = \sum_{i=1}^{n} a_i k(\mathbf{h}_i, \mathbf{h}_t)$. Thus, the optimization over all the functions in the RKHS is simply translated to a quadratic optimization over the weights a_i. The optimal weight vector $\mathbf{a} = [a_1, \ldots, a_n]$, computed by setting the derivative of the optimization with respect to the vector $\mathbf{a}$ to zero, is given by: $\mathbf{a}^* = \left[\mathbf{JK} + l\gamma_k \mathbf{I}_N + l\gamma_M \mathbf{LK}\right]^{-1} \mathbf{q}$. Here, $\mathbf{K}$ is the $n \times n$ Gram matrix of k defined by $K_{ij} = k(\mathbf{h}_i, \mathbf{h}_j)$; $\mathbf{J}$ is a $n \times n$ diagonal matrix serving as indicator: $\mathbf{J} = \text{diag}(1, ..., 1, 0, ..., 0)$ with n_L ones and n_U zeros on its diagonal; and $\mathbf{q} = [p_1, ..., p_{n_L}, 0, ..., 0]^T$ is a label vector comprising the n_L known positions of the labelled samples with $q_i = 0$, for all $i > n_L$.

IV. Experimental Study

The algorithm performance was tested using real recordings carried out in the speech and acoustic lab of Bar-Ilan University. This is a $6 \times 6 \times 2.4$m room that has a reverberation time controlled by 60 interchangeable panels covering the room facets. The measurement equipment consists of a RME Hammerfall HDSPe MADI sound-card and an Andiamo.mc (for Microphone pre-amplification and digitization (A/D)). As source signals we used Fostex 6301BX loudspeakers which have a rather flat response in the frequency range 80Hz-13kHz. The signals were measured by 2 AKG type CK-32 omnidirectional microphones, which were placed 0.2m apart. All the measurements were carried out with a sampling frequency of 48kHz and a resolution of 24-bits. The measured signals were then downsampled to 16kHz. The reverberation level was determined by the room panels configuration and was set to $T_{60} = 620$ms. An example of the room layout is depicted in Fig. 1.[1]

The source position is confined to a 4m long line at approximately 2.5m distance from the two microphones. Along this line we generated $n_L = 9$ equally-spaced labelled samples

[1]We would like to thank Mr. Pini Tandeitnik for his useful assistance in the lab recordings

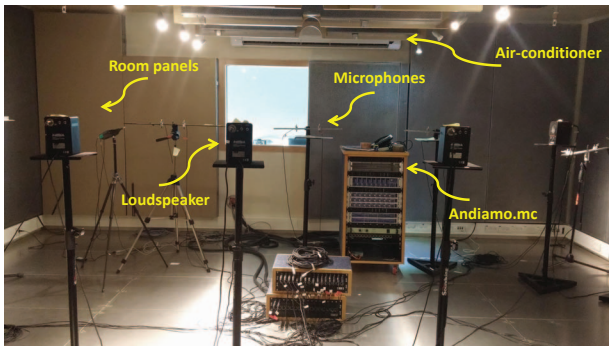


Fig. 1: The room configuration.

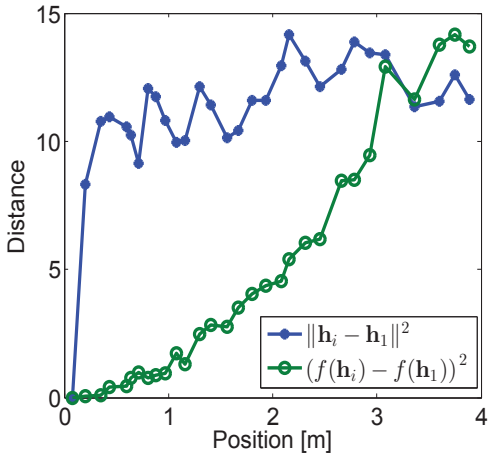


Fig. 2: The Euclidean distance between the first RTF and each of the other RTFs along the line, and the distances between the corresponding mappings.

with resolution of 0.5m. Additional $n_U = 75$ unlabelled measurements were generated along the line in random positions. The algorithm performance was examined on 30 test samples also generated in random positions, unknown to the algorithm, along the same line. For generating the labelled samples a chirp signal, 30s long, was used, while for generating both the unlabelled samples and the test samples we used 105 different speech signals of both males and females, 10s long, drawn from the TIMIT database. The RTF values were estimated in 2048 frequency bins, using Welch's method with 0.128 s long windows and 75% overlap. For the MRL algorithm, the RTF vectors consist of 115 frequency bins, which corresponds to the frequency range between $0.5 - 1.5$kHz, where most of the speech components are concentrated. We used a Gaussian kernel for both the reproducing kernel k associated with the RKHS, and for computing the graph Laplacian $\mathbf{L}$ which empirically represent the manifold structure, with scaling parameters 100 and 7, respectively. For the graph Laplacian, we used a truncated kernel, i.e., with non-zeros entries for only the 5% nearest-neighbours of each sample among the training set, where the remaining entries were set to zero. The regularization parameters were set to $\gamma_k = 10^{-7}$ and $\gamma_M = 10^{-3}$.

The first experiment addresses the geometrical structure of the RTF samples. For this purpose, we examine the 30 noiseless RTF samples of the test set (resolution of approximately 0.13m). Fig. 2 illustrates the Euclidean distance between the RTF of the left-most measurement along the line and all the other samples on that line: $\|\mathbf{h}_i - \mathbf{h}_1\|$, as a function of the corresponding position. In addition, we compare to the distances obtained after mapping each RTF to the output of the function f estimated by the MRL algorithm, i.e.: $(f(\mathbf{h}_i) - f(\mathbf{h}_1))^2$. It can be seen that the Euclidean distance is monotonic with respect to the physical position only in a limited region (of about 0.3m). For larger scales there is no evident correspondence between the distances in the RTF domain and distances in the physical domain. However, the mapping of the MRL algorithm which relates the RTFs to the corresponding positions, is monotonic for almost the entire range. This analysis, strengthen the assumption that the RTFs lie on a nonlinear manifold, in which linearity can be assumed only for small patches.

The performance of the proposed MRL algorithm is evaluated by estimating the positions of the test set, and computing the corresponding root mean squared error (RMSE). For comparison, we also applied the nearest-neighbour (NN)-search algorithm, which searches among the labelled set for the RTF that is the closest to the RTF of the new test point. The position attached to the closest RTF in the labelled set, is the estimated position for the new point. For the NN-search algorithm, we used RTF vectors consisting of 280 frequency bins, which corresponds to the frequency range between $0.5 - 2.5$kHz, since it yields the best results. We also compare to the generalized cross-correlation phase transformation (GCC-PHAT) algorithm as a baseline, since it is considered robust to reverberation. For the GCC algorithm we used the original sampling frequency of 48kHz, without downsampling. For the MRL and the NN algorithms, we use $n_L = 5$ labelled samples, creating a grid with 1m distance between adjacent samples. We examine two different types of noise sources: air-conditioner noise and babble noise, which was simultaneously played from 3 loudspeakers located in the room. The RMSEs obtained for different signal to noise ratio (SNR) levels, are presented in Fig. 3.

It can be observed that the proposed algorithm outperforms the two other methods for both noise types and for all SNR levels. The GCC algorithm suffers from the high reverberation level, but seems not to be significantly affected by the amount of noise.

In addition, we examined the effect of the amount of labelled samples on the MRL and the NN algorithms (the GCC algorithm does not use training data, hence it is irrelevant in this comparison). We started with 9 labelled samples (grid of 0.5m), reduced the number of points by half to 5 (grid of 1m as in Fig. 3), then reduced to 3 samples (grid of 2m), and finally used only 2 samples (grid of 4m). The results for both algorithms in these 4 scenarios, are summarized in Fig. 4(a). It can be seen that as the size of the labelled

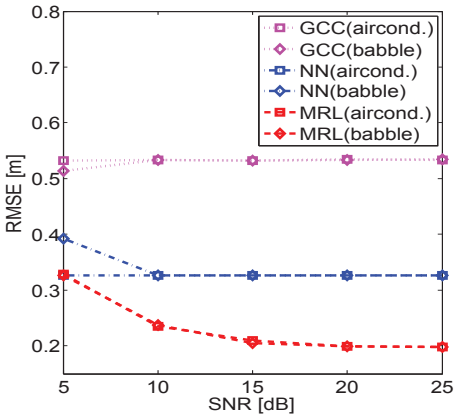
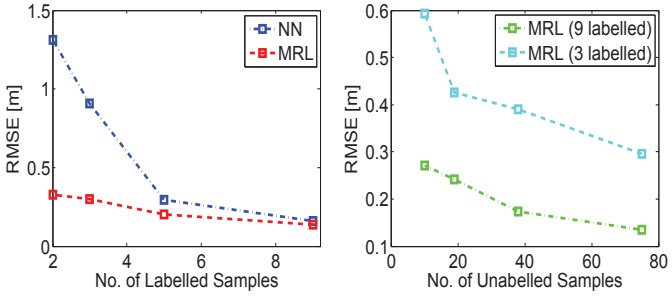


Fig. 3: The RMSE for different noise levels.



(a) (b)

Fig. 4: The RMSE (a) for different labelled set sizes and (b) for different unlabelled set sizes.

set is reduced, the difference in the performance between the two algorithms, gradually increases. This comparison also emphasizes the importance of the semi-supervised approach, i.e., the contribution of the unlabelled samples. Moreover, the later case (with merely two labelled samples) shows that the unlabelled samples carry sufficient information to accurately identify the true degrees of freedom (i.e., the location), and the labelled samples are only used for aligning the obtained representation to the physical domain of the positions.

Another aspect that we wish to examine is the influence of the amount of unlabelled data on the performance of the MRL algorithm. Fig. 4(b), presents the error of the MRL as a function of the number of unlabelled samples. It can be observed that as the amount of unlabelled data increases, the performance of the MRL algorithm gradually improves. We conclude, that the unlabelled data have a significant impact on the accuracy of the solution. The unlabelled samples provide useful information for mapping RTFs to positions, since they recover the manifold structure of the RTFs.

Finally, we investigated the effect of changes in the environmental conditions between the training and the test stages. We examined two types of changes: the door of the room changed from closed (during training) to open (during test) and slight changes in the panel configuration (decreasing the room reverberation time by about 5%). We repeated the measurements of 20 test samples in both scenarios (the training samples are left unchanged), and compared the results obtained under these conditions, to the nominal results where there is no change in the environmental conditions between the training set and the test set. Here as well, the labelled set consists of $n_L = 5$ labelled samples. This comparison is summarized in Table I, which presents the RMSEs in all the defined scenarios. It can be seen that either opening the door or changing the panel configuration does not have a significant impact on the localization results of the MRL method, which indicates that the algorithm is robust to slight changes that are likely to occur in practical scenarios. Note that the results of the GCC algorithm are slightly improved under these changes due to the reduction in the reverberation level.

	Nominal	**Door**	**Panel**
MRL	0.204	0.204	0.25
NN	0.368	0.317	0.363
GCC	0.52	0.477	0.493

TABLE I: Comparison between the RMSE obtained in the case where the training and the test sets are generated exactly with the same conditions (first column) and when the test is generated under some environmental changes: open door (second column) or changes in the panel configuration (third column).

V. Conclusions

A comprehensive experimental study with real recordings is conducted to examine the performance of the proposed semi-supervised localization approach. The algorithm uses a regularized optimization to estimate the function that relates the RTF samples to their corresponding positions. The main idea is to constrain the estimated function to change smoothly with respect to an underlying manifold of the RTF samples, whose structure is empirically learned from unlabelled samples. We have examined the performance of the proposed method using real-life measurements. The experimental results, demonstrate the robustness of the proposed method in noisy and reverberant conditions. The method utilizes both labelled and unlabelled data, hence reducing the amount of either labelled or unlabelled samples has a direct influence on the algorithm performance. Even when there are only few labelled samples the proposed method is able to locate the source using the information implied by the unlabelled measurements. Moreover, typical changes in the environmental conditions are shown to have a little effect on the obtained results.

REFERENCES

[1] 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.

[2] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *Multimedia, IEEE Transactions on*, vol. 10, no. 3, pp. 538–548, 2008.

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

[4] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[5] 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.

[6] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[7] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using csp analysis," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 921–924.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] B. Laufer-Goldshtein, 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.

[16] ——, "Semi-supervised sound source localization based on manifold regularization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[17] ——, "Study on manifolds of acoustic responses," in *Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[18] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, Nov. 2006.

[19] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational learning theory*. Springer, 2001, pp. 416–426.