Robust Relative Transfer Function Identification on Manifolds for Speech Enhancement

Amit Sofer*, Tomáš Kounovský†, Jaroslav Čmejla†, Zbyněk Koldovský† and Sharon Gannot*
* The Alexander Kofkin Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel.
Email: {Amit.Sofer, Sharon.Gannot}@biu.ac.il
† Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic.

***Abstract*—Accurate and reliable identification of the relative transfer function (RTF) between microphones with respect to a desired source is an essential component in the design of microphone array beamformers. In this paper, we present a robust RTF identification method on manifolds, tested and trained with real recordings. This method relies on a manifold learning (ML) approach to infer a representation of typical RTFs in a confined area within an acoustic enclosure. We propose a robust supervised identification method that combines the a priori learned geometric structure and the measured signals. A series of experiments using a recently established database of acoustic responses taken at the Bar-Ilan university acoustic lab, demonstrate the effectiveness of the proposed approach over a standard, non-robust, beamforming design method.**

***Index Terms*—Multi-channel speech enhancement, RTF identification, Manifold learning, Robust beamforming**

I. Introduction

Modern acoustic beamformers, which take into account the entire acoustic propagation path, require the acoustic impulse responses (AIRs) relating the source and the microphones (or their respective acoustic transfer functions (ATFs)). Since ATF estimation is a blind problem, it was proposed in [1] to substitute the ATFs with the RTFs in the beamformer design. It was further shown in this work that the RTFs can be non-blindly estimated from the received microphone signals. Accurate identification of the RTFs leads to a significant improvement of the beamformer performance as compared with beamformers that only utilize the direct-path propagation. A plethora of RTF estimation procedures can be found in the literature [1]–[7]. However, the performance of these methods often deteriorates in presence of high-level noise and severe reverberation.

There is a rich literature on robustifying beamforming, usually by some type of beam widening [8]–[15]. In this work we choose to improve the estimated steering vector, the RTF in our case, by utilizing a pre-learned set of RTFs. For this, we harness the manifold leaning paradigm to infer a low-dimensional representation of RTFs.

This work was partly supported by The Czech Science Foundation through Project No. 20-17720S and by the Erasmus+ KA 107 project No. 2017-1-CZ01-KA107-034883; the Israeli Innovation Authority through KAMIN Project No. 61916, "Environment-Aware Data-Driven Acoustic Signal Processing" and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Despite their intricate structure, RTFs are governed by a few parameters, e.g., the size and the geometry of the room, the positions of the source and the microphones, and the reflective properties of the walls, as was analyzed in [16]. Consequently, the acoustic paths exhibit geometric structures of low-dimensionality, which are often referred to as manifolds, and may be accurately parameterized using manifold learning (ML) methods.

In this work, we focus on scenarios where the source position is confined to a region within a known acoustic environment and assume the availability of a training set, specifically, a set of RTFs from the region of possible source positions in this environment.

We follow the footsteps of [17], with some modifications, and learn the manifold of the training RTFs using an extendable kernel method [18]. The method relies on *Laplacian eigenmaps* and *diffusion maps* and applies *spectral graph* theory to infer a low-dimensional embedding of the learned RTFs. Then, we utilize the extendable learned model and construct a robust supervised RTF identification method. The inferred RTFs is then used as the steering vector of a minimum variance distortionless response (MVDR) beamformer. The performance of the proposed method, in terms of noise reduction and speech intelligibility, is validated in various signal-to-noise ratios (SNRs) and reverberation levels using a recently established database recorded at Bar-Ilan acoustic lab [19].

Our method differs from [17] in the following aspects: 1) we use a different RTF estimator based on the generalized eigenvalue decomposition (GEVD), 2) we use a different kernel normalization in the manifold inference, and, most importantly, 3) we use real recorded data rather than simulations in [17].

II. Problem Formulation

We consider a room with an array of M microphones in a fixed location and assume that the possible positions of the desired source are confined to a specific known region. Let $x_m(t), m = 1, \ldots, M$, denote the measured signal at the mth microphone, $s(t)$ denotes the desired speech signal, and $v_m(t), m = 1, \ldots, M$ the contribution of all noise sources, as captured by the mth microphone. Each microphone input is

given by

$$x_m(t) = \{s * a_m\}(t) + v_m(t) \tag{1}$$

where $a_m(t)$ is the AIR from the source to the mth microphone at time t, and $*$ is the convolution operator. In the static case, where the speaker is not moving, the AIR becomes time-invariant. The time-domain convolution in (1) can be approximated by a multiplication in the short-time Fourier transform (STFT) domain and can be written in a vector form as

$$\begin{aligned}\mathbf{x}(n,f) &= s(n,f)\mathbf{a}(f) + \mathbf{v}(n,f), \\ &= \tilde{s}(n,f)\mathbf{h}(f) + \mathbf{v}(n,f),\end{aligned} \tag{2}$$

where n, f are the time-frame and frequency-bin indexes, respectively, $n = 1,\ldots,N$, $f = 1,\ldots,F$, and $\mathbf{a}(f) = [a_1(f),\ldots,a_M(f)]^\top$. We also define $\tilde{s}(n,f) = s(n,f)a_{\text{ref}}(f)$, the source signal as captured by the reference microphone, and $\mathbf{h}(f)$ the vector of RTFs defined as

$$\mathbf{h}(f) \triangleq \frac{\mathbf{a}(f)}{a_{\text{ref}}(f)}, \tag{3}$$

where $a_{\text{ref}}(f)$ is the component of vector $\mathbf{a}(f)$ corresponding to the reference microphone, arbitrarily set as microphone #1.

Given a training set of RTFs from a confined volume in the room, obtained in noiseless conditions, we are interested in improving noisy RTF estimates from the same volume, in order to robustify a beamformer design.

III. RTF IDENTIFICATION ON MANIFOLDS

In this section, we first summarize the GEVD-based RTF estimation procedure, which yields an unbiased estimator using the available input signals and an estimate of the noise-only spatial correlation matrix. Then, we explore the *diffusion maps* framework, which enables us to infer the RTF manifold. Finally, we show that the *geometric harmonics* framework, which facilitates the extension of the learned model to new noisy RTFs, yields a more accurate RTF estimate that may improve the beamformer performance, both in noise reduction and speech distortion measures.

A. GEVD RTF identification

In [2], [4], it was shown that the RTF can be estimated from the GEVD of the spatial correlation matrices of the noisy signals $\mathbf{\Phi}_{xx}(f)$[1] and of the noise-only signals $\mathbf{\Phi}_{vv}(f)$. The latter is estimated from noise-only segments, assumed to be available. The RTF is estimated by solving

$$\mathbf{\Phi}_{xx}(f)\boldsymbol{\varphi}(f) = \mu(f)\mathbf{\Phi}_{vv}(f)\boldsymbol{\varphi}(f). \tag{4}$$

Using $\boldsymbol{\varphi}(f)$, the generalized eigenvector corresponding to $\mu(f)$, the largest generalized eigenvalue, the vector of RTFs

$$\hat{\mathbf{h}}^{\text{GEVD}}(f) \triangleq [\hat{h}_1^{\text{GEVD}}(f),\ldots,\hat{h}_M^{\text{GEVD}}(f)]^\top \tag{5}$$

can be computed as:

$$\hat{\mathbf{h}}^{\text{GEVD}}(f) = \frac{\mathbf{\Phi}_{vv}(f)\boldsymbol{\varphi}(f)}{(\mathbf{\Phi}_{vv}(f)\boldsymbol{\varphi}(f))_{\text{ref}}}. \tag{6}$$

Define $h_m^k(f)$, the RTF associated with the kth training location, $k = 1,\ldots,N_t$, and the mth microphone at the fth frequency bin and $\mathbf{h}_m^k$ the respective vector constructed by concatenating all frequencies. The training sets $\mathcal{R}_m = \{\mathbf{h}_m^k\}_{k=1}^{N_t}$ of all RTFs associated with the mth microphone, are obtained by applying the GEVD procedure to the noiseless training recordings. In the absence of noise Φ_{vv} in (6) is substituted by an identity matrix, such that (4) simplifies to the eigenvalue decomposition (EVD) problem.

B. The diffusion framework

The geometry of $\mathcal{R}_m$ can be studied using the Laplacian operator which is defined by the divergence of the gradient of a manifold in an Euclidean space. The Laplacian contains all information regarding the manifold geometry and describes the time-evolution of a diffusion process over the manifold. It is an infinite-dimension operator defined on continuous spaces. However, in our case, only a finite set of samples on the manifold is available. Based on this discrete set of samples, we represent the manifold by a graph where the observations are the graph nodes, and the weights of the edges are defined using the heat kernel function. We then define a Markov process on the graph by constructing a transition matrix, which is a discretization of the diffusion process on the manifold.

We construct $M-1$ graphs, a graph $G_m = (\mathcal{R}_m, \mathbf{W}_m), m = 2,\ldots,M$ for each microphone except the reference microphone. The graph nodes are the RTFs of the training set and the edges are weighted according to the Gaussian heat kernel $\mathbf{W}_m$, whose (i,k)th element $i,k \in 1,\ldots,N_t$ is given by

$$w_m(i,k) = e^{-\frac{\|\mathbf{h}_m^i - \mathbf{h}_m^k\|^2}{2\epsilon}}, \tag{7}$$

with $\epsilon > 0$ is a kernel scale parameter. The kernel measures the pairwise similarity between the points in $\mathcal{R}_m$.

We normalize the kernel matrix and obtain a Markov transition matrix $\mathbf{P}_m$, where its (i,k)th element is given by

$$p_m(i,k) = \frac{w_m(i,k)}{d_m(i)} \tag{8}$$

with $d_m(i) = \sum_{k=1}^{N_t} w_m(i,k)$ a normalization factor. Since $\mathbf{P}_m$ comprises nonnegative values, and since each row is summed to 1, $p_m(i,k)$ can be viewed as the probability of transition from node i to node k in one random step. These probabilities measure the connectivity of the graph nodes. Since the RTFs are governed by the position of the source, the kernel enables capturing the actual variability in terms of the source position based on the measurements. The transition matrix $\mathbf{P}_m$ is *similar* to a symmetric matrix $\mathbf{A}_m$, and is therefore denoted as its *conjugate*, given by

$$\mathbf{A}_m = \mathbf{D}_m^{\frac{1}{2}}\mathbf{P}_m\mathbf{D}_m^{-\frac{1}{2}} \tag{9}$$

[1] In the more general form it can be time-varying, but here we assume that the RTF is time-invariant, so we use an average over all time segments.

where $\mathbf{D}_m$ is the diagonal matrix with the elements $d_m(i)$ on its diagonal. The symmetric matrix $\mathbf{A}_m$ has N_t real eigenvalues $\{\lambda_j^m\}_{j=0}^{N_t-1}$ and a set of orthogonal eigenvectors $\{\mathbf{v}_j^m\}_{j=0}^{N_t-1}$ in $\mathbb{R}^{N_t}$. Therefore, it can be stated in terms of the following spectral decomposition:

$$\mathbf{A}_m = \sum_{j=0}^{N_t-1} \lambda_j^m \mathbf{v}_j^m (\mathbf{v}_j^m)^\top. \tag{10}$$

The EVD of the kernel captures its significant components and provides a compact parameterization of the manifold of RTFs. Moreover, the eigenvectors can be viewed as functions of the training RTFs, where the ith coordinate of each eigenvector is associated with the ith training RTF. The eigenvectors form a complete basis for any function of the data, and each coordinate of the training RTFs can be expressed as a linear combination of this basis:

$$h_m^i(f) = \sum_{j=0}^{N_t-1} c_j^m(f) v_j^m(i) \tag{11}$$

where $c_j^m(f) = \langle \mathbf{v}_j^m, [h_m^1(f), \ldots, h_m^{N_t}(f)]^\top \rangle$ are the projection coefficients on the basis vectors. Since $\mathbf{P}_m$ is conjugate to $\mathbf{A}_m$, the eigenvalues of both matrices are identical [18]. In addition, $\{\boldsymbol{\phi}_j^m\}_{j=0}^{N_t-1}$ and $\{\boldsymbol{\psi}_j^m\}_{j=0}^{N_t-1}$ are the corresponding left and right eigenvectors of non-symmetric $\mathbf{P}_m$, and the following relations hold $\boldsymbol{\phi}_j^m = \mathbf{D}_m^{\frac{1}{2}} \mathbf{v}_j^m$ and $\boldsymbol{\psi}_j^m = \mathbf{D}_m^{-\frac{1}{2}} \mathbf{v}_j^m$. The eigendecomposition of the transition matrix $\mathbf{P}_m$ is hence given by:

$$\mathbf{P}_m = \sum_{j=0}^{N_t-1} \lambda_j^m \boldsymbol{\psi}_j^m (\boldsymbol{\phi}_j^m)^\top. \tag{12}$$

As was shown in [20] the spectrum (eigenvalues), written in a descending order $1 = \lambda_0^m \geq \lambda_1^m \geq \ldots \lambda_{N_t-1}^m$, decays rapidly such that only a few terms are required to obtain a sufficient accuracy in the sum (12). Hence, we look for the spectral gap to determine ℓ, the number of dominant eigenvalues, which defines the *intrinsic dimension* of the manifold.

The diffusion map $\{\boldsymbol{\Psi}^m\}$, which is defined for each point i in the dataset as

$$\boldsymbol{\Psi}^m(i,:) = [\lambda_1^m \psi_1^m(i), \lambda_2^m \psi_2^m(i), \ldots, , \lambda_\ell^m \psi_\ell^m(i)],$$

embeds each point i in the dataset into an Euclidean space. This embedding is a new parametrization of the data in a lower dimension space, which captures the manifold underlying parameters and respects the manifold geometric structure. It was also shown that, in these new coordinates, the Euclidean distance between two points in the embedded space represents the distance between the two high-dimensional points, as defined by a random walk on the manifold surface, namely, the *diffusion distance*.

C. The geometric harmonics

After obtaining a low dimension embedding of the RTFs in the training set, we would like to extend it to new measured data at test time. By doing so, we wish to improve noisy estimates of RTFs using the information extracted from the learned manifold. Geometric harmonics [21] is a method that extends a low-dimensional embedding to new data points.

Let $\mathbf{B}_m$ be a non-symmetric kernel defined between any RTF in our test set $\tilde{\mathbf{h}}_m^q, q = 1, \ldots, N_{\text{Test}}, N_{\text{Test}} > 1$ and each of the RTFs in the training set, whose (q,i)th element is given by

$$b_m(q,i) = \frac{\tilde{w}_m(q,i)}{\tilde{d}_m(q) z_m(i)} \tag{13}$$

where $\tilde{w}_m(q,i) = e^{-\frac{\|\tilde{\mathbf{h}}_m^q - \mathbf{h}_m^i\|^2}{\epsilon}}$ with $\tilde{d}_m(q) = \sum_i \tilde{w}_m(q,i)$ and $z_m(i) = \sum_q \frac{\tilde{w}_m(q,i)}{\tilde{d}_m(q)}$ normalization factors. It was shown in [22] that the construction of the original training kernel satisfies $\mathbf{A}_m = \mathbf{B}_m^\top \mathbf{B}_m$. Moreover, $\mathbf{C}_m = \mathbf{B}_m \mathbf{B}_m^\top$ can be seen as an extended kernel, whose (q,q')th element measures the probability that any two RTFs $\tilde{\mathbf{h}}_m^q, \tilde{\mathbf{h}}_m^{q'}$ are associated with the same training RTF, and its eigenvectors $\{\boldsymbol{\xi}_j^m\}$ provide an extended parameterization for any RTF. $\mathbf{A}_m$ and $\mathbf{C}_m$ share the same eigenvalues $\{\lambda_j^m\}$, which are the square of the singular values of $\mathbf{B}_m$. The eigenvectors $\{\mathbf{v}_j^m\}$ of $\mathbf{A}_m$ are the right singular vectors of $\mathbf{B}_m$, and the eigenvectors $\{\boldsymbol{\xi}_j^m\}$ of $\mathbf{C}_m$ are the left singular vectors of $\mathbf{B}_m$. The singular value decomposition (SVD) of $\mathbf{B}_m$ describes the algebraic relation between the eigenvectors

$$\boldsymbol{\xi}_j^m = \frac{1}{\sqrt{\lambda_j^m}} \mathbf{B}_m \mathbf{v}_j^m. \tag{14}$$

Based on the Nyström extension and using the extended eigenvectors obtained in (14), the relation in (11) can be expanded into any RTF from the learned region of the room, provided that $\lambda_j^m \neq 0$:

$$\begin{aligned} \tilde{h}_m^q(f) &= \sum_{j=1}^{\ell} c_j^m(f) \xi_j^m(q) + \eta^m(f), \\ &= (\mathbf{B}_m \mathbf{O}_m)(q,f) + \eta^m(f), \end{aligned} \tag{15}$$

where $\eta^m(f)$ is a modelling error term that depends on ϵ and stems from the use of the coefficients $c_j^m(f)$, inferred from the training set, rather than recalculating them with the additional RTFs in the test set. The error term becomes smaller if either the number of training RTFs or ℓ increases. The matrix $\mathbf{O}_m$ can be computed in advance and its (i,f)th element is given by:

$$o_m(i,f) = \sum_{j=1}^{\ell} \frac{c_j^m(f)}{\sqrt{\lambda_j^m}} v_j^m(i). \tag{16}$$

If the parameterization of a single test RTF $\tilde{\mathbf{h}}_m$ is required, the matrix $\mathbf{B}_m$ is reduced to a vector, and (15) should be rewritten by concatenating all frequency bins as $\tilde{\mathbf{h}}_m = \mathbf{O}_m^\top \mathbf{b}_m + \boldsymbol{\eta}^m$, where $\mathbf{b}_m$ is a vector of length N_t whose ith element is defined similarly to (13) as $b_m(i) = e^{-\frac{\|\tilde{\mathbf{h}}_m - \mathbf{h}_m^i\|^2}{\epsilon}} / \tilde{d}_m$, with $\tilde{d}_m = \sum_i e^{-\frac{\|\tilde{\mathbf{h}}_m - \mathbf{h}_m^i\|^2}{\epsilon}}$.

Ideally, we would like to combine (6) and (15) to get an

estimate of the test RTFs $\hat{\mathbf{h}}_m^q, q = 1, \ldots, N_{\text{Test}};\ m = 2, \ldots, M$

$$\hat{\mathbf{h}}^q(f) = \frac{\mathbf{\Phi}_{vv}(f)\boldsymbol{\varphi}(f)}{(\mathbf{\Phi}_{vv}(f)\boldsymbol{\varphi}(f))_{\text{ref}}}, f \in [1, \ldots, F]$$

subject to

$$(\mathbf{B}_m\mathbf{O}_m)(q, f) - \hat{h}_m^q(f) \leq \delta \tag{17}$$

where δ is a small constant (assuming the modelling error is negligible). However, since $\tilde{\mathbf{h}}_m^q$ are required for calculating $\mathbf{B}_m$, the equation becomes highly nonlinear and difficult to solve. Following [17], we relax the problem and apply a two-stage and sub-optimal solution. In the first stage, we obtain a solution $\hat{\mathbf{h}}^{q,\text{GEVD}}(f)$ by solving the GEVD problem (4) for each location (indexed by q) in our test set and for each frequency bin. In the second stage, we utilize the prior geometric information and project the GEVD solution onto the building blocks of the learned manifold, explicitly:

$$\hat{\mathbf{h}}_m^q(f) = (\mathbf{B}_m(\{\hat{\mathbf{h}}_m^{q,\text{GEVD}}\}_{q=1}^{N_{\text{Test}}})\mathbf{O}_m)(q, f) \tag{18}$$

where the notation $\mathbf{B}_m(\{\hat{\mathbf{h}}_m^{q,\text{GEVD}}\}_{q=1}^{N_{\text{Test}}})$ implies substitution of $\tilde{\mathbf{h}}_m^q$ in (13) with the RTFs $\hat{\mathbf{h}}_m^{q,\text{GEVD}}$ that are directly estimated from the measured data. Algorithm 1 summarizes the manifold-based RTF estimation procedure.

Algorithm 1: RTF identification on manifolds

Learning the Manifold of RTFs (Training Stage):

1) Obtain training recordings from the region of interest in the room in noiseless conditions
2) Compute a training set $\{\mathcal{R}_m\}_{m=2}^M$ of typical RTFs (6)
3) Construct the normalized kernels $\mathbf{A}_m$ (7)-(9)
4) Compute the eigenvalue decomposition $\{\lambda_j^m, \mathbf{v}_j^m\}_j$ of the kernels $\mathbf{A}_m$ and construct $\mathbf{O}_m$ the projection coefficients on the basis (16)

Supervised RTF Identification (Test Stage):

1) Obtain a new segment of measurements
2) Estimate the RTF using the GEVD (6)
3) Confine the RTF to the learned manifold (18)

IV. Performance Evaluation

The proposed method was evaluated with several, objective and subjective, performance measures using the MIRaGe dataset [19] comprising multichannel recordings acquired at Bar-Ilan acoustic lab.

Experimental Setup: In the database, a loudspeaker is located on a grid of points in a cube-shaped volume of dimensions $46 \times 36 \times 32$ cm. The possible positions of the loudspeaker form a grid sampled every 2 cm across the x-axis and y-axis and every 4 cm across the z-axis. Overall, there are $24 \times 19 \times 9 = 4104$ possible source positions (grid vertices). Besides, 25 other positions, denoted out of grid (OOG), were used as possible positions of the noise sources. For each position (both in grid and OOG) a chirp signal was played. The entire setup was recorded by six static linear microphone arrays, each of which consisting of $M = 5$ microphones with inter-microphone spacing of $-13, -5, 0, +5$ and $+13$ cm relative to the central microphone (the reference microphone). The entire setup was recorded in three reverberation levels of 100, 300, and 600 ms. For our experiments we used microphone array #2, which is placed directly in front of the grid at the distance of 2 m from the center of the grid. The recordings were randomly divided to 80% training ($N_t = 3283$ positions) and 20% test ($N_{\text{Test}} = 821$ positions).

For the estimation of the RTF at the training stage, the following procedure was applied: 1) using the chirp signals recorded in the MIRaGe database the AIRs from the source position to the microphone arrays were estimated, then 2) speech signals (not used during the subsequent test phase) were convolved with the AIRs to generate clean microphone signals, and finally 3) RTFs were estimated using (6).

The kernel scale was chosen via numerical optimization following a cross-validation procedure and was set to $\epsilon = 3$ for $T_{60} = 100, 300$ ms and to $\epsilon = 0.5$ for the $T_{60} = 600$ ms. The manifold dimensions were similarly set, while respecting the spectral gap and ensuring that $\lambda_k \neq 0, k = 1, \ldots, \ell$. For our database we chose $\ell = 48$ for $T_{60} = 100, 300$ ms and $\ell = 7$ for $T_{60} = 600$ ms. The RTFs length was set to be $F = 2048$, and the sampling rate was set to 16 kHz.

For each point in the test set, three different speech signals were convolved with the AIRs, similarly to the training phase. In addition, additive noise was generated by convolving two pink noise signals with two AIRs corresponding to two fixed OOG positions, with SNR in the range of $[-10 : 10]$ dB.

An MVDR beamformer [1] was constructed for every position in the test set,

$$\mathbf{w}_{\text{MVDR}}^q(f) = \frac{\mathbf{\Phi}_{vv,q}^{-1}(f)\hat{\mathbf{h}}^q(f)}{(\hat{\mathbf{h}}^q(f))^{\mathsf{H}}\mathbf{\Phi}_{vv,q}^{-1}(f)\hat{\mathbf{h}}^q(f)} \tag{19}$$

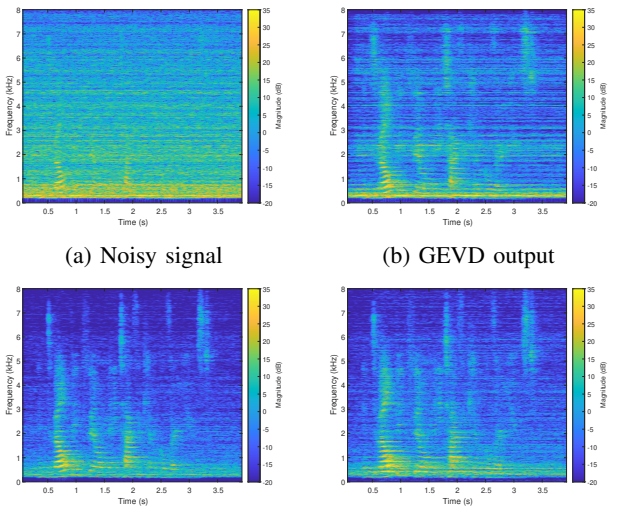
where $\mathbf{\Phi}_{vv,q}(f)$ is the $M \times M$ power spectral density (PSD) matrix of the received noise signals at the fth frequency bin associated with the qth testing location.

The results are analysed using two quality measures, the SNR at the beamformer output and the short-time objective intelligibility (STOI) quality measure [23]. The results are averaged across the 3 different speakers and the 821 test positions. The performance measures are compared with other MVDR beamformers using either the vanilla GEVD estimate of the RTFs or the RTF corresponding to a source at the center of grid (CoG).

Results: Tables I, II and III depict the SNR_{out} and STOI quality measures, for $T_{60} = 100, 300, 600$ ms, respectively. The analyses of the results will be split into two SNR ranges, above and below 5 dB. For SNR < 5 dB, the ML algorithm outperforms the competing algorithms, as it was able to reliably estimate the RTFs even in high reverberation levels.

Comparing the ML-based RTF estimation and the arbitrary CoG RTF selection, demonstrates the usefulness of the proposed approach. For SNR ≥ 5 dB, the robustness of the ML approach becomes a disadvantage and the GEVD demonstrates improved speech intelligibility figures (but still lower SNR_{out},

reflecting slightly higher distortion level). Similar trends were observed for diffused noise. The advantages of the proposed ML scheme for $\text{SNR}_{\text{in}} = -10$ dB and $T_{60} = 600$ ms, are also subjectively demonstrated by assessing sonograms (Fig. 1), and sound samples.[2]



(a) Noisy signal (b) GEVD output

(c) ML output (d) CoG output

Fig. 1: Sonograms: $\text{SNR}_{\text{in}} = -10$ dB and $T_{60} = 600$ ms.

TABLE I: SNR_{out} [dB] and STOI [%] scores for $T_{60} = 100$ ms

	SNR_{out} dB						STOI [%]					
Alg.\SNR_{in}[dB]	−10	−6	−2	2	6	10	−10	−6	−2	2	6	10
Unprocessed	-10	-6	-2	2	6	10	37.8	47.3	58.6	70.3	80.7	88.4
GEVD	1.5	6	11.1	16.5	22.1	27.3	67.1	75.9	83.6	89	**92.5**	**94.7**
CoG	9	13	17	21	25	29	79.7	83.4	85.9	87.4	88.2	88.6
ML	**9.2**	**13.2**	**17.2**	**21.2**	**25.2**	**29.1**	**81.4**	**85.3**	**87.7**	**89.2**	90	90.4

TABLE II: SNR_{out} [dB] and STOI [%] scores for $T_{60} = 300$ ms

	SNR_{out} dB						STOI [%]					
Alg.\SNR_{in}[dB]	−10	−6	−2	2	6	10	−10	−6	−2	2	6	10
Unprocessed	-10	-6	-2	2	6	10	33.7	43.5	55.2	67.1	77.8	86.1
GEVD	-1	2.4	6.1	10.6	15.9	21.5	60.4	69.3	77.3	84	**89.2**	**92.6**
CoG	**5.3**	**9.3**	**13.3**	**17.3**	**21.3**	**25.3**	71.2	75.2	75.8	80.7	82.1	82.9
ML	3.5	7.5	11.6	15.3	19.6	23.5	**73.6**	**78.6**	**82.7**	**85.8**	87.9	89.2

TABLE III: SNR_{out} dB and STOI [%] scores for $T_{60} = 600$ ms

	SNR_{out} [dB]						STOI [%]					
Alg.\SNR_{in}[dB]	−10	−6	−2	2	6	10	−10	−6	−2	2	6	10
Unprocessed	-10	-6	-2	2	6	10	38.9	48.9	60.4	71.9	**81.7**	**88.9**
GEVD	-1.4	2.9	7.6	12.6	17.3	21.9	52.5	62.2	70.6	76.8	80.5	82.5
CoG	6.6	10.7	14.7	18.7	22.7	26.7	64.4	68.4	71	72.7	73.8	74.3
ML	**7.2**	**11.2**	**15.2**	**19.3**	**23.3**	**27.3**	**67.5**	**72.3**	**75.5**	**77.7**	79.1	79.9

V. Conclusions

We have presented a robust supervised RTF identification method in which the manifold of typical RTFs in a particular room is learned in advance, and then, exploited to improve the identification of unknown RTFs based on noisy measurements. The method was tested and trained with real recordings in a wide range of SNRs and reverberation levels, and has shown to provide a robust RTF estimation and consequently to improve the beamformer performances, especially in noisy conditions.

[2]Available `www.eng.biu.ac.il/gannot/speech-enhancement/`

References

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

[2] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *IEEE Int. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2015, pp. 544–548.

[3] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE Int. Conf. on Acous., Sp. and Sig. Proc. (ICASSP)*, 2015, pp. 320–324.

[4] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Tran. on Au., Sp., and Lang. Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[5] Z. Koldovskỳ, J. Málek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Tran. on Au., Sp., and Lang. Proc.*, vol. 23, no. 8, pp. 1335–1347, 2015.

[6] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Tran. on Sp. and Au. Proc.*, vol. 12, no. 5, pp. 451–459, 2004.

[7] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Tran. on Sig. Proc.*, vol. 44, no. 8, pp. 2055–2063, 1996.

[8] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Tran. on Sp. and Au. Proc.*, vol. 12, no. 5, pp. 478–488, 2004.

[9] A. Barnov, V. B. Bracha, S. Markovich-Golan, and S. Gannot, "Spatially robust GSC beamforming with controlled white noise gain," in *Int. Workshop on Acous. Sig. Enhancement (IWAENC)*, 2018, pp. 231–235.

[10] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, and K. R. Liu, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2010.

[11] J. Li and P. Stoica, *Robust adaptive beamforming*. Wiley Online Library, 2006.

[12] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Tran. on Sig. Proc.*, vol. 51, no. 7, pp. 1702–1715, 2003.

[13] ——, "Doubly constrained robust Capon beamformer," *IEEE Tran. on Sig. Proc.*, vol. 52, no. 9, pp. 2407–2423, 2004.

[14] S. A. Vorobyov, A. B. Gershman, Z.-Q. Luo, and N. Ma, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Sig. Proc. Let.*, vol. 11, no. 2, pp. 108–111, 2004.

[15] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Tran. on Sig. Proc.*, vol. 53, no. 5, pp. 1684–1696, 2005.

[16] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *Int. Conf. on Latent Variable Analysis and Sig. Separation*. Springer, 2015, pp. 203–210.

[17] R. Talmon and S. Gannot, "Relative transfer function identification on manifolds for supervised GSC beamformers," in *21st Euro. Sig. Proc. Conf. (EUSIPCO)*, 2013.

[18] N. Rabin, "Data mining in dynamically evolving systems via diffusion methodologies," Ph.D. dissertation, Tel Aviv University, 2010.

[19] J. Čmejla, T. Kounovský, S. Gannot, Z. Koldovský, and P. Tandeitnik, "MIRaGe: multichannel database of room impulse responses measured on high-resolution cube-shaped grid," in *28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 56–60.

[20] R. 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. of the Nat. Ac. of Sci.*, vol. 102, no. 21, pp. 7426–7431, 2005.

[21] R. R. Coifman and S. Lafon, "Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 31–52, 2006.

[22] D. Kushnir, A. Haddad, and R. R. Coifman, "Anisotropic diffusion on sub-manifolds with application to earth structure classification," *Applied and Computational Harmonic Analysis*, vol. 32, no. 2, pp. 280–294, 2012.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Au., Sp., and Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.