SUPERVISED SYSTEM IDENTIFICATION BASED ON LOCAL PCA MODELS

Tomer Koren[1]*,Ronen Talmon*[2]*, and Israel Cohen*[2]

[1] Department of Computer Science
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
tomerk@cs.technion.ac.il

[2] Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{ronenta2@tx, icohen@ee}.technion.ac.il

ABSTRACT

We propose a supervised system identification method for recovering an acoustic impulse response in a reverberant room. Unlike most existing methods, our algorithm is based on prior information given in the form of a training set of known impulse responses acquired in a controlled environment. By relying on the prior information, we train local Principal Component Analysis (PCA) models of impulse responses corresponding to several different regions in the room. We propose to crudely localize the respective source position, and subsequently, based on the appropriate local model, recover the impulse response. In order to approximate the source location, we introduce a specially-tailored distance measure which is based on an affinity between the trained local models. Experimental results in simulated noisy and reverberant environments demonstrate significant improvements over existing methods.

Index Terms— System identification, acoustic source localization, principal component analysis, local PCA

1. INTRODUCTION

System identification is a fundamental problem in acoustic signal processing applications, including echo cancellation [1], dereveberation [2], noise suppression [3], and beamforming. For example, in echo cancellation, an acoustic impulse response needs to be identified in order to reduce the coupling between the loudspeaker and the microphone. System identification is usually carried out directly by standard least-squares (LS) analysis and adaptive filtering. Unfortunately, in low SNR, this approach enables poor results. The noisy measurements are not reliable, and the estimation of the impulse response is often severely under-determined. This difficulty stems from the large number of independent parameters needed to represent the acoustic system in reverberant environments.

Recently, Fozunbal *et al.* proposed to employ Principal Component Analysis (PCA) for the task of system identification [4]. To overcome the identification challenge, acoustic impulse responses from several known locations in the room are acquired in advance and are used for training and calibration. A supervised algorithm for system identification is then proposed by forming a model based on a training set. Unfortunately, the global model trained by PCA may not be sufficient to capture the local properties of acoustic impulse responses, such as the variability of an impulse response at a certain small region of the room.

In this paper, we propose a method that is capable of preserving and exploiting the local features. By relying on the prior information, we train *local* PCA models of impulse responses corresponding to different regions in the room. We show that when applied locally, the PCA-based approach accurately captures the structure of acoustic impulse responses. Consequently, a two-stage identification algorithm is proposed. In the first stage, we crudely localize the respective source position. For estimating the location, we introduce a specially-tailored distance measure which is based on an affinity between the trained local models. We note that the focus of the work is system identification, and the crude (single-channel) localization is merely a by-product of the algorithm. However, the introduced measure may be beneficial, for example, in supervised source localization methods, e.g. [5]. The coarse localization determines the proper local model of the desired acoustic impulse response. Then, in the second stage, we recover the impulse response by relying on the appropriate local model. The estimated response should correspond to the measured source and microphone signals. In addition, it should fit the local model, which characterizes acoustic impulse responses in this part of the room. Accordingly, the proposed identification is carried out by solving a constrained minimization problem, representing both requirements. Experimental study of the algorithm show significant improvement compared to direct LS and global PCA approaches.

The remainder of the paper is organized as follows. In Section 2, we describe the setup and formulate the problem. In Section 3, we present our proposed algorithm for system identification. Finally, in Section 4, experimental results demonstrate the performance of the algorithm in simulated noisy and reverberant environments.

2. PROBLEM FORMULATION

The acoustic impulse response between a speaker and a microphone is dictated by numerous parameters: the ambient room size and geometry; the presence of objects in the room; the locations of the speaker and the microphone; and the reflective properties of the walls. Although in practice any of these factors is subject to change, we assume they all remain static, except the speaker location. In our setting, it implies that the data acquired in the training (i.e., calibration) phase is indeed applicable in the test phase. Furthermore, the acoustic impulse response between the microphone and the speaker is completely determined by the location of the speaker. We let $h_\theta(n)$ denote the acoustic impulse response between the microphone and a speaker, at position $\theta = [\phi, \varphi, \rho]$, where ϕ and φ are the azimuth and elevation angles relative to the microphone and ρ is the distance between the speaker and the microphone.

We assume the availability of a training set, i.e., a set of known impulse responses $\mathcal{H}$ corresponding to a set of speaker locations $\mathbf{\Theta}$. The set $\mathcal{H}$ may be acquired, for example, by employing clas-

This research was supported by the Israel Science Foundation (grant no. 1130/11).

sic non-blind system identification techniques in a controlled, noiseless environment. We require the set of locations to be of the form $\mathbf{\Theta} = \bigcup_{i=1}^{m} \mathbf{\Theta}_i$, where each set $\mathbf{\Theta}_i = \{\theta_{i_j}\}_{j=1}^{L}$ consists of L small perturbations of a typical, predefined speaker location θ_i. Consequently, our training set takes the form $\mathcal{H} = \bigcup_{i=1}^{m} \mathcal{H}_i$, where each $\mathcal{H}_i = \{h_{i_j}(n)\}_{j=1}^{L}$ is a "cluster" of L impulse responses acquired at the proximity of the location θ_i. While this requirement may seem unrealistic, we note that in practice it is often the case. Typical sources usually move slightly with time due to natural small perturbations. Thus, by dividing a measurement interval into L subintervals, we obtain measurements corresponding to the "cluster" of impulse responses. For more details, see [6], where this assumption was verified on real room recordings.

The input of the algorithm, at the test phase, consists of a target pair $\{x(n),\, y(n)\}$ of source and microphone signals corresponding to an unknown speaker location θ'. The source and microphone signals acquired during the entire observation interval are divided into L' subintervals $\{x_i(n),\, y_i(n)\}_{i=1}^{L'}$ corresponding to locations $\{\theta'_i\}_{i=1}^{L'}$. As with the training data, we assume that these locations are small perturbations of the location θ'. Provided there is *no double-talk*, the signals $y_i(n)$ are expressed as

$$y_i(n) = h_{\theta'_i}(n) * x_i(n) + u_i(n)\,, \tag{1}$$

where $h_{\theta'_i}(n)$ is the unknown acoustic impulse response corresponding to the location θ'_i, and $u_i(n)$ is a local noise. Our goal in this work is to recover the impulse response $h_{\theta'}$, by exploiting the prior information conveyed by the training data $\mathcal{H}$, in conjunction with the test data.

We consider an algebraic formulation of the problem. An acoustic impulse response is denoted by a vector $\mathbf{h}_{\theta'} \in \mathbb{R}^D$, and the source and microphone signals are vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$ respectively, with $M = N + D - 1$. Accordingly, (1) can then be rewritten as

$$\mathbf{y} = \mathbf{X}^T \mathbf{h}_{\theta'} + \mathbf{u} \tag{2}$$

where $\mathbf{u} \in \mathbb{R}^M$ is a noise vector, and the matrix $\mathbf{X}$ is a $D \times M$ convolution matrix of the vector $\mathbf{x}$. In practice, the value of D is dictated by the sampling frequency and the reverberation time of the room, and is usually in the order of a few thousands.

In view of (2), our algorithm seeks to minimize the MSE of the estimation of $\mathbf{h}_{\theta'}$, i.e.,

$$J(\mathbf{h}) = \frac{1}{M}\|\mathbf{X}^T\mathbf{h} - \mathbf{y}\|^2\,. \tag{3}$$

However, solving this problem directly by standard LS analysis often leads to a poor estimate of $\mathbf{h}_{\theta'}$, especially in low SNR and high reverberation conditions. This is due to the high dimensionality of the problem, requiring an extremely large number of samples (i.e., a large value of N) to compensate for the large number of degrees of freedom. Thus, additional prior information on the structure of $\mathbf{h}_{\theta'}$ is necessary to overcome this challenge. In our setting, we form a model for $\mathbf{h}_{\theta'}$ by relying on the training data.

3. PROPOSED ALGORITHM

In [4], the authors proposed to compute the empirical mean vector and covariance matrix of the training set,

$$\bar{\mathbf{h}} = \frac{1}{|\mathbf{\Theta}|}\sum_{\theta\in\mathbf{\Theta}} \mathbf{h}_\theta\,,\; \Sigma = \frac{1}{|\mathbf{\Theta}|}\sum_{\theta\in\mathbf{\Theta}} (\mathbf{h}_\theta - \bar{\mathbf{h}})(\mathbf{h}_\theta - \bar{\mathbf{h}})^T\,. \tag{4}$$

The pair $(\bar{\mathbf{h}}, \Sigma)$ is then used as the learned model. By employing PCA, the large eigenvectors of Σ, which correspond to the principal "parameters", capture most of the information disclosed in the data. Hence, the dimensionality of the problem is significantly reduced by considering only a subspace of $\mathbb{R}^D$, spanned by a few principal eigenvectors.

A well known limitation of PCA is that it is linear and able to capture only the *global* structure (in $\mathbb{R}^D$) of the training data. Unfortunately, a typical set of acoustic impulse responses, acquired at different positions in a room, admits an extremely complex global structure (often referred to as a non-linear *manifold*). As a result, a low-dimensional linear subspace of $\mathbb{R}^D$ may not faithfully describe the data in our setting.

On the other hand, a PCA-based approach may perform rather well when applied *locally*, i.e., on a data set sufficiently condensed in a small neighborhood. The resulting principal eigenvectors may then be thought of as a representation of a "tangent space" of the manifold at that location. In our application, this corresponds to making use of a dataset of acoustic impulse responses measured at the immediate vicinity of a certain position in the room. A crucial observation at this point is that such local model, conveying the local variability of the impulse response, can also serve as a *"signature"* (i.e., a feature set) of the respective source position. This gives rise to a novel metric between acoustic impulse responses, that captures the proximity in terms of the respective source locations. A similar concept was previously presented and analyzed in [6, 7].

Based on the above observations, our algorithm exploits locally-computed covariance matrices to approximate (i.e., localize) the source position θ', and subsequently to recover (i.e., identify) the impulse response $\mathbf{h}_{\theta'}$. The remainder of this section is dedicated to a detailed description of the algorithm.

3.1. Training phase

The training stage involves the computation of local PCA models, at each of the "clusters" given the training set. Specifically, for each $i = 1, \ldots, m$ we define

$$\bar{\mathbf{h}}_{\theta_i} = \frac{1}{|\mathbf{\Theta}_i|}\sum_{\theta\in\mathbf{\Theta}_i} \mathbf{h}_\theta\,,\; \Sigma_{\theta_i} = \frac{1}{|\mathbf{\Theta}_i|}\sum_{\theta\in\mathbf{\Theta}_i} (\mathbf{h}_\theta - \bar{\mathbf{h}}_{\theta_i})(\mathbf{h}_\theta - \bar{\mathbf{h}}_{\theta_i})^T\,,$$

so that our model is formed by the pairs $\{\bar{\mathbf{h}}_{\theta_i}, \Sigma_{\theta_i}\}_{i=1}^{m}$. For storage efficiency, and for reasons that will become apparent shortly, we may keep only a low-rank approximation of each covariance matrix Σ_{θ_i}, determined by its k largest eigenvectors (where k is a predefined parameter).

3.2. Test phase

Given the pairs $\{\mathbf{x}_i,\, \mathbf{y}_i\}_{i=1}^{L'}$ at test time, we first compute a crude, noisy estimate of each of the individual acoustic impulse responses, by solving the problem

$$\hat{\mathbf{h}}_{\theta'_i} = \arg\min_{\mathbf{h}\in\mathbb{R}^D} \|\mathbf{X}_i^T\mathbf{h} - \mathbf{y}_i\|^2\,,$$

where $\mathbf{X}_i$ is $D \times M$ convolution matrix of the vector $\mathbf{x}_i$. The solution to this LS problem is given by

$$\hat{\mathbf{h}}_{\theta'_i} = (\mathbf{X}_i\mathbf{X}_i^T)^{-1}\mathbf{X}\mathbf{y}_i\,.$$

This enables the estimation of a target pair $\{\bar{\mathbf{h}}_{\theta'},\, \Sigma_{\theta'}\}$, defined as the empirical mean and covariance of the vectors $\{\hat{\mathbf{h}}_{\theta'_i}\}_{i=1}^{L'}$. As be-

fore, it is sufficient to store a rank-k approximation of the covariance $\Sigma_{\theta'}$, instead of a full matrix.

One of the key points of this work is to establish an affinity metric that faithfully indicates the distance between two acoustic impulse responses, with respect to the spatial source location. Clearly, the Euclidean distance between the responses is very limited. For example, in anechoic rooms, Euclidean distances merely indicate whether the corresponding sources are located within the same distance from the microphone. In this work, we capitalize the local covariance matrices, which convey the second order statistics of the local spatial variability of the responses, in order to compute an affinity metric.

For any two locations θ_1, θ_2 in the room, for which estimations of the local covariance matrices Σ_{θ_1}, Σ_{θ_2} are at hand, we define an affinity metric via

$$d(\theta_1, \theta_2) = \|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|_F^2 \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenious norm. It is straightforward to show that a good approximation of this metric can be obtained by relying on the rank-k approximations of the matrices Σ_{θ_1}, Σ_{θ_2}, rather than on the full matrices.

For localizing the source of the signal $y(n)$, we naturally choose the training location θ_i whose distance to the unknown target position θ', as measured by (5), is minimized. That is, we let $\tilde{\theta} = \theta_{i^*}$ where

$$i^* = \arg\min_i d(\theta_i, \theta') = \arg\min_i \|\Sigma_{\theta_i} - \Sigma_{\theta'}\|_F^2$$

and $\tilde{\theta}$ is a crude approximation of the source position. Note that, though the location θ' is unknown, we are able to compute the distances $d(\theta_i, \theta')$ since we obtain an estimate of the local covariance at θ'. We also note that the known impulse response $\mathbf{h}_{\tilde{\theta}}$ does not yield sufficiently accurate identification.

Once an approximate coarse location $\tilde{\theta}$ is determined, we utilize the local PCA model $(\bar{\mathbf{h}}_{\tilde{\theta}}, \Sigma_{\tilde{\theta}})$ to identify the acoustic impulse response $\mathbf{h}_{\theta'}$. Note that this model is available beforehand, since the position $\tilde{\theta} = \theta_{i^*}$ is amidst the training locations. It is worthwhile noting that empirical tests showed that the identification based on the PCA model $(\bar{\mathbf{h}}_{\theta'}, \Sigma_{\theta'})$, computed from the noisy measurements, enables poor results. Thus, the prior information and crude localization step are essential to accurate system identification. Let

$$\Sigma_{\tilde{\theta}} = [\mathbf{\Psi}\ \mathbf{\Phi}] \begin{bmatrix} \Lambda & 0 \\ 0 & \bar{\Lambda} \end{bmatrix} \begin{bmatrix} \mathbf{\Psi}^T \\ \mathbf{\Phi}^T \end{bmatrix}$$

be the singular value decomposition (SVD) of $\Sigma_{\tilde{\theta}}$, where Λ is a $k \times k$ diagonal matrix consisting of the k largest singular values, and $\mathbf{\Psi} = [\psi_1, \ldots, \psi_k]$ with $\psi_1, \ldots, \psi_k$ denote the k largest eigenvectors of $\Sigma_{\tilde{\theta}}$. We assume that the columns of $\mathbf{\Psi}$ define a basis that well approximates all possible acoustic paths of this room region. Thus, with high probability, the desired impulse response $\mathbf{h}_{\theta'}$ satisfies

$$\mathbf{h}_{\theta'} - \bar{\mathbf{h}}_{\tilde{\theta}} \in \text{Span}\{\psi_1, \ldots, \psi_k\}. \tag{6}$$

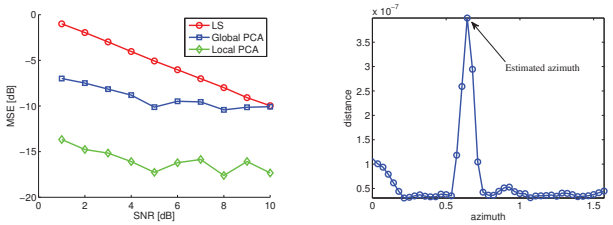
Using the local model, we minimize the estimation error (3) subject to the constraint (6), i.e.,

$$\begin{aligned} &\min_{\mathbf{h} \in \mathbb{R}^D} \|\mathbf{X}^T\mathbf{h} - \mathbf{y}\|^2 \\ &\text{subject to } \mathbf{h} - \bar{\mathbf{h}}_{\tilde{\theta}} \in \text{Span}\{\psi_1, \ldots, \psi_k\}. \end{aligned} \tag{7}$$

We note that the constraint becomes more accurate as more training data from this part of the room is available. On the other hand, in case the source is located remotely from the trained positions, the constraint may impair the LS criterion. This optimization problem boils down to a simple LS problem, which is solved over a k-dimensional subspace of $\mathbb{R}^D$ spanned by the vectors $\psi_1, \ldots, \psi_k$. The solution is given as

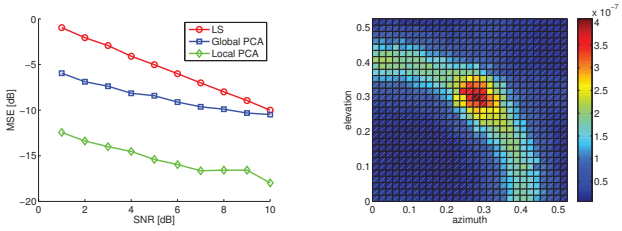
$$\hat{\mathbf{h}}_{\theta'} = \bar{\mathbf{h}}_{\tilde{\theta}} + (\mathbf{\Psi}^T\mathbf{X}\mathbf{X}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{X}(\mathbf{y} - \mathbf{X}^T\bar{\mathbf{h}}_{\tilde{\theta}}) \tag{8}$$

where $\hat{\mathbf{h}}_{\theta'}$ is the estimate of the desired acoustic impulse response. It is therefore apparent that for the identification task, the full matrix $\Sigma_{\tilde{\theta}}$ is redundant and only its k largest eigenvectors should be at our disposal.



(a) (b)

Fig. 1. First experiment results. (a) The obtained MSE of the proposed algorithm (Local PCA) under different noise levels, compared with a standard LS approach (LS) and a global PCA approach (Global PCA). (b) A typical localization result. In this example, the estimated azimuth angle is 0.7053 while the true angle is 0.7017.



(a) (b)

Fig. 2. Second experiment results. (a) The obtained MSE of the algorithms under different noise levels. (b) A typical 2D localization result. In this example, the estimated azimuth and elevation angles are 0.2765 and 0.2965, while the actual angles are 0.2708 and 0.2889.

4. EXPERIMENTAL RESULTS

In this section, we demonstrate the ability of the algorithm to recover the location of a speaker and to identify the corresponding acoustic impulse response. Using the RIR generator [8] we simulate acoustic impulse responses in a room of dimensions $4 \times 4 \times 3.5$m with a moderate reverberation time of $T_{60} = 0.2$s. Due to computational considerations, we set the sampling rate to 16 kHz and generate filters consisting of $D = 1024$ taps. We position a microphone at a fixed central location inside the room, and a speaker in various locations around it.

In the first experiment, we fix the distance between the speaker and the microphone to $\rho = 1.5$m, and the direction of arrival (DOA)

azimuth angle is set in fixed increments of 2° along the interval $0^\circ - 90^\circ$. The elevation angle with respect to the microphone φ is kept at zero throughout the experiment. At each of the 45 different speaker locations, we simulate the corresponding acoustic impulse response at that exact location. Near each location, we simulation additional $L = 50$ impulse responses corresponding to small angular perturbations of the location. As described in Section 2, these perturbations may correspond to the natural movement of the speaker. The perturbations were drawn independently from a Gaussian distribution of zero mean and a standard deviation of 0.2°. We then formed our training set by aggregating the acoustic impulse responses.

At the test phase, we pick a random target azimuth θ' from the interval $0^\circ - 90^\circ$ and simulated an impulse response $\mathbf{h}_{\theta'}$ and a pair of source and microphone signal vectors $\{\mathbf{x}, \mathbf{y}\}$ corresponding to that location. This was accomplished by generating a white Gaussian source signal $\mathbf{x}$ of length 5000, convolving it with $\mathbf{h}_{\theta'}$ and contaminating with additive white Gaussian noise (of a specified SNR) to obtain $\mathbf{y}$. We repeat this exact procedure at another $L' = 50$ positions obtained by perturbing θ' to obtain the test data vectors $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{L'}$.

For evaluating the performance of the algorithm, we compute the following leakage signal

$$\ell(n) = x(n) * \mathbf{h}_{\theta'}(n) - x(n) * \hat{\mathbf{h}}_{\theta'}(n)$$

and measure the normalized leakage variance, corresponding to the normalized mean square error (MSE) between the underlying clean and reconstructed source signals at the microphone. We compare the results of our algorithm to that of an unconstrained LS approach, and to that of the global PCA approach of [4] (trained over the same training set).

In Fig. 1(a) we present the results of the identification task, under various noise levels. The results are averaged over 50 repetitions of the experiment. The curve depicts the MSE of the reconstructed microphone signal obtained by the three approaches, as a function of the SNR. For the proposed algorithm, we used $k = 25$; for the global PCA algorithm, we tuned the parameters (essentially the dimension of the target subspace) to obtain the best possible MSE results. It is evident that the proposed approach yields a significant advantage over both methods. In addition, we observe that the relative improvement obtained by the PCA-based identification methods becomes more significant as the SNR decreases and the measurements become less reliable.

In Fig. 1(b) we present a typical result of the localization stage in SNR level of 10dB. The graph details the distance of the target location, as measured by (5), from each of the possible training locations. It is worthwhile noting that in the neighborhood of the target location, the proposed metric is highly correlated to the true spatial distance from this location.

In order to further demonstrate the robustness of the algorithm, we conducted a second experiment, where both the azimuth and elevation angles ϕ, ψ are varied. In this experiment, we have 45^2 training impulse responses, corresponding to 45 different azimuth and elevation angles in the range $0^\circ - 45^\circ$. In each training location, additional $L = 50$ responses corresponding to perturbations in both angular directions are simulated. The test position of the speaker is randomly picked within the same range of angles as the training positions. The rest of the experiment parameters are taken from the first experiment.

Figure 2(a) presents the MSE obtained by the identification algorithm. We observe similar trends as in Fig. 1(a). The proposed algorithm outperforms the competing methods. Furthermore, the improvement becomes more significant as the SNR decreases. Figure 2(b) depicts the localization results in SNR level of 10dB, where the color coding is set according to the the distance of the target location, as measure by (5), from each of the possible training locations. We clearly observe a peak in the actual position of the source, enabling accurate localization. The results demonstrate the ability of the proposed measure (5) to properly compare impulse responses corresponding to speakers in different angular positions.

5. CONCLUSIONS

We have presented a two-stage supervised algorithm for system identification based on local PCA. The proposed algorithm invokes a data-driven approach that exploits prior measurements for training and calibration. The identification is carried out as an optimization problem that combines the acquired local model along with the current measurements. Experimental results conducted in simulated reverberant environments showed encouraging performance.

For future work, we plan to evaluate the performance of the algorithm on recorded data in real environments. It would be interesting to investigate the effect of environmental changes following the training stage, e.g. when people are moving in the room. In addition, we intend to extend this approach to the problem of relative transfer function identification [9, 10]. Improving the estimation of the relative transfer function between two microphones based on training data, may be highly beneficial for array processing and beamforming.

6. REFERENCES

[1] E. Hänsler and G. Schmidt, "Acoustic echo and noise control: A practical approach," *New York: Wiley*, 2004.

[2] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," *Springer*, 2010.

[3] I. Cohen, J. Benesty, and S. Gannot, "Speech processing in modern communication: Challenges and perspectives," *Springer*, 2010.

[4] M. Fozunbal, T. Kalker, and R.W. Schafer, "Multi-channel echo control by model learning," *Proc. 11th IEEE International Workshop on Acoustic Echo and Noise Control*, 2008.

[5] D. M. Malioutov, M. Cetin, J. W. Fisher III, and A. S. Willsky, "Superresolution source localization through data-adaptive regularization," *Proc. IEEE Sensor Array and Multichannel Signal Process. Workshop*, 2002.

[6] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust.*, 2011.

[7] R. Talmon, D. Kushnir, R. R. Coifman, I. Cohen, and S. Gannot, "Parametrization of linear systems using diffusion kernels," *to appear in IEEE Trans. Signal Process.*, 2012.

[8] E.A.P. Habets, "Room impulse response (RIR) generator," *[Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html*.

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

[10] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 546–555, 2009.