A Study on Manifolds of Acoustic Responses

Bracha Laufer-Goldshtein[1], Ronen Talmon[2], and Sharon Gannot[1](✉)

[1] Bar-Ilan University, 5290002 Ramat-gan, Israel
bracha.gold@walla.com, Sharon.Gannot@biu.ac.il
[2] Technion – Israel Institute of Technology, Technion City, 3200003 Haifa, Israel
ronen@ee.technion.ac.il

Abstract. The construction of a meaningful metric between acoustic responses which respects the source locations, is addressed. By comparing three alternative distance measures, we verify the existence of the acoustic manifold and give an insight into its nonlinear structure. From such a geometric view point, we demonstrate the limitations of linear approaches to infer physical adjacencies. Instead, we introduce the diffusion framework, which combines local and global processing in order to find an intrinsic nonlinear embedding of the data on a low-dimensional manifold. We present the diffusion distance which is related to the geodesic distance on the manifold. In particular, simulation results demonstrate the ability of the diffusion distance to organize the samples according to the source direction of arrival (DOA).

1 Introduction

Speech processing in reverberant environments facilitates a very complex relation between the emitted speech and the signal received by the microphones. Many algorithms, such as beamformers and localizers, try to distinguish between signals based on their propagation vector. In some scenarios, e.g. meeting rooms or cars, it can be assumed that the source position is confined to a predefined area. It can be reasonable to assume that representative samples from the region of the interest can be measured in advance. Due to reverberation, common practice is to represent the acoustic responses using a large number of variables, corresponding to the vast amount of reflections from the different surfaces characterizing the enclosure. In fact, the acoustic responses are only influenced by 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. As a result, the high-dimensional acoustic responses are not uniformly scattered in their original space, but are rather concentrated on a manifold of much lower dimension. We therefore investigate the manifold of the acoustic responses and examine the proper distance between them.

In the context of multi-channel echo cancellation, Fozunbal et al. [6] presented a system identification algorithm by learning a low dimensional *linear* model of the room. Talmon and Gannot [10] proposed a different approach for supervised system identification, utilized for generalized sidelobe canceller (GSC) beamformer, based on the diffusion maps concept [2]. A similar approach is discussed in this paper.

E. Vincent et al. (Eds.): LVA/ICA 2015, LNCS 9237, pp. 203–210, 2015.
DOI: 10.1007/978-3-319-22482-4_23

The manifold perspective was also examined in light of the source localization problem. The existence of a binaural manifold was discussed by Deleforge et al. in [3–5] and a localization algorithm was presented. Another approach for supervised source localization based on the diffusion framework was introduced in [8].

In the current study, we show how to construct an informative metric between acoustic responses which respects the position of the source in the enclosure. For simplicity, the demonstration will focus only on the DOA of the source. We are interested in 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. This is the latent variable which distinguishes between different acoustic responses. Accordingly, we will embed the acoustic responses in an intrinsic low-dimensional space representing the manifold and show that this embedding corresponds with the position of the source.

2 Problem Formulation

We consider a single source generating an unknown speech signal $s(n)$, which is received by a pair of microphones. Both the speaker and the microphones are located in an enclosure, e.g., a conference room or a car interior, with moderate reverberation time. The received signals, denoted by $x(n)$ and $y(n)$, are contaminated by additive stationary noise sources and are given by:

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

where n is the time index, $a_i(n),\ i = \{1, 2\}$ are the corresponding acoustic impulse response (AIRs) relating the source and each of the microphones, and $u_i(n),\ i = \{1, 2\}$ are uncorrelated white Gaussian noise (WGN) signals. Each of the AIRs is composed of the direct path between the source and the microphone, as well as reflections from the surfaces characterizing the enclosure. Consequently, even in moderate reverberation, the AIR is typically modelled as a long finite impulse response (FIR) filter.

Common practice 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. An equivalent representation of (1) is given by:

$$\begin{aligned} y(n) &= h(n) * x(n) + v(n) \\ v(n) &= u_2(n) - h(n) * u_1(n) \end{aligned} \tag{2}$$

where $h(n)$ is the relative impulse response between the microphones with respect to the source, satisfying $a_2(n) = h(n) * a_1(n)$. In (2), the relative impulse response represents the system relating the measured signal $x(n)$ as an input and the measured signal $y(n)$ as an output.

For convenience, we represent (2) in the frequency domain. Assuming high signal to noise ratio (SNR) conditions, the Fourier transform of the relative impulse response, termed the relative transfer function (RTF), is obtained by:

$$H(k) = \frac{S_{yx}(k)}{S_{xx}(k)} = \frac{S_{ss}(k)A_2(k)A_1^*(k)}{S_{ss}(k)|A_1(k)|^2} = \frac{A_2(k)}{A_1(k)} \quad k = 0, \ldots, D-1 \tag{3}$$

where $H(k)$ is the RTF, $S_{yx}(k)$ is the cross power spectral density (CPSD) between $y(n)$ and $x(n)$, $S_{xx}(k)$ is the power spectral density (PSD) of $x(n)$ and $S_{ss}(k)$ is the PSD of the source. $A_1(k)$ and $A_2(k)$ are the acoustic transfer functions (ATFs) of the respective AIRs, and k denotes a discrete frequency index. Since $A_1(k)$ and $A_2(k)$ are unavailable, we use the estimated RTF $\hat{H}(k) \equiv \frac{\hat{S}_{yx}(k)}{\hat{S}_{xx}(k)}$, based on the estimated PSD and CPSD. The choice of the value of D should balance the tradeoff between correspondence to the relative impulse response length (large value) and latency considerations (small value). Accordingly, we define the feature vector $\mathbf{h} = [\hat{H}(0), \ldots, \hat{H}(D-1)]^T$ as the concatenation of estimated RTF values in all frequency bins. In practice, we discard high frequencies in which the ratio in (3) is meaningless due to the lack of speech components. When noise influence cannot be neglected, we use, instead, an RTF estimator based on the non-stationarity of the speech signal [7].

3 Manifold-Based Distance Measures

Three alternative distance measures for quantifying the affinity between different RTFs, are addressed. We start with linear distance measures, namely, the Euclidean distance and the distance derived by principal component analysis (PCA) mapping. Next, we describe the concept of diffusion maps and present the diffusion distance [2]. For deriving the PCA-based distance and the diffusion distance, we assume the availability of a considerable amount of representative RTFs from various locations in the region of interest in the enclosure. The geometric interpretation of each of the distance measures is examined, and their hidden assumptions are highlighted and discussed.

3.1 Linear Distance Measures

The Euclidean distance between RTFs is denoted by:

$$D_{\text{Euc}}(\mathbf{h}_i, \mathbf{h}_j) = \|\mathbf{h}_i - \mathbf{h}_j\|. \tag{4}$$

The Euclidean distance does not assume an existence of a manifold and compares two RTFs in their original space. In particular, the Euclidean distance is equal to the geodesic distance when the manifold is flat, thus, inexplicitly, the Euclidean distance respects flat manifolds. Therefore, it is a good affinity measure only when the RTFs are uniformly scattered all over the space, or when they lie on a flat manifold.

The second distance we consider is based on PCA, which is the most common method to date for *linear* dimensionality reduction. First, the empirical mean and covariance matrix of the data are computed by:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i, \quad \mathbf{R} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{h}_i - \boldsymbol{\mu})(\mathbf{h}_i - \boldsymbol{\mu})^T \tag{5}$$

where N is the number of available representative RTFs in the region of interest. Then, by applying eigenvalue decomposition to the covariance matrix $\mathbf{R}$, we obtain a set of D eigenvectors and eigenvalues, denoted by $\{\mathbf{v}_i, \lambda_i\}_{i=0}^{D-1}$. The d eigenvectors, corresponding to the d largest eigenvalues, are viewed as the principal components of the data and form a new low-dimensional coordinate system. Finally, the RTFs are linearly projected onto the new coordinates/principal components:

$$\boldsymbol{\nu}\left(\mathbf{h}_i\right) = \left[\mathbf{v}_1, \ldots \mathbf{v}_d\right]^T\left(\mathbf{h}_i - \boldsymbol{\mu}\right). \tag{6}$$

The corresponding distance is given by the Euclidean distance between the projections:

$$D_{\text{PCA}}(\mathbf{h}_i, \mathbf{h}_j) = \|\boldsymbol{\nu}(\mathbf{h}_i) - \boldsymbol{\nu}(\mathbf{h}_j)\|. \tag{7}$$

PCA is essentially a global approach; the principal directions of the *entire* set of RTFs are extracted from the covariance matrix. Then, the RTFs are *linearly* projected onto these directions, assuming that the manifold is linear/flat. As a result, the algorithm filters undesired samples' perturbations with respect to the manifold, which are caused by artifacts, such as: noise, estimation error and non-uniform sampling. Assuming that the manifold is indeed flat, PCA performs better than the Euclidean distance, due to this element of filtering.

3.2 Diffusion Distance

The concept of diffusion maps was introduced by Coifman and Lafon [2] as a general method for data-driven nonlinear dimensionality reduction. The diffusion framework consists of the following steps [9].

First, the affinity between RTFs is measured based on a pairwise weight function $k(\cdot,\cdot)$. Typically, the affinity is defined by a Gaussian function:

$$k(\mathbf{h}_i, \mathbf{h}_j) = \exp\left\{-\frac{\|\mathbf{h}_i - \mathbf{h}_j\|^2}{\varepsilon}\right\}. \tag{8}$$

Such an affinity preserves locality since it defines local neighbourhoods according to the value of the scale parameter ϵ: for $\|\mathbf{h}_i - \mathbf{h}_j\| \ll \epsilon$, $k(\mathbf{h}_i, \mathbf{h}_j) \rightarrow 1$, and for $\|\mathbf{h}_i - \mathbf{h}_j\| \gg \epsilon$, $k(\mathbf{h}_i, \mathbf{h}_j) \rightarrow 0$.

Second, $\{\mathbf{h}_i\}$ are interpreted as nodes in a Graph. A Markov process can be defined on the graph via a construction of the transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{W}$ is the Gram matrix defined by $W_{ij} = k(\mathbf{h}_i, \mathbf{h}_j)$, and $\mathbf{D}$ is a diagonal matrix whose elements are the row sums of $\mathbf{W}$. 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 $\mathbf{h}_j$.

In the next step, a nonlinear mapping of the RTFs into a new low-dimensional Euclidean space is built according to:

$$\mathbf{\Phi}_d : \mathbf{h}_i \mapsto \left[\boldsymbol{\varphi}_1^{(i)}, \ldots, \boldsymbol{\varphi}_d^{(i)}\right]^T \tag{9}$$

where $\{\boldsymbol{\varphi}_j\}_{j=1}^{d}$ are the d-principal right-singular vectors of the transition matrix $\mathbf{P}$, and $\boldsymbol{\varphi}_k^{(i)}$ denotes the ith entry of the vector $\boldsymbol{\varphi}_k$. Note that $\boldsymbol{\varphi}_0$ is ignored since it is an all-ones column vector.

The diffusion distance, which describes the relationships between pairs of samples in terms of their graph connectivity, is defined by:

$$D_{\text{Diff}}(\mathbf{h}_i, \mathbf{h}_j) = \|p\left(\mathbf{h}_i, \cdot\right) - p\left(\mathbf{h}_j, \cdot\right)\|_{\boldsymbol{\phi}_0} = \sum_{r=1}^{N} \left(p\left(\mathbf{h}_i, \mathbf{h}_r\right) - p\left(\mathbf{h}_j, \mathbf{h}_r\right)\right)^2 / \boldsymbol{\phi}_0^{(r)} \tag{10}$$

where $\boldsymbol{\phi}_0$ is the most dominant left-singular vector of $\mathbf{P}$. The diffusion distance reflects the flow between two RTFs on the manifold, which is related to the geodesic distance on the manifold. It can be shown that the diffusion distance is equal to the Euclidean distance in the diffusion maps space when using all N eigenvectors, and can be well approximated by only the first few d eigenvectors [2], i.e.,

$$D_{\text{Diff}}(\mathbf{h}_i, \mathbf{h}_j) \cong \|\mathbf{\Phi}_d(\mathbf{h}_i) - \mathbf{\Phi}_d(\mathbf{h}_j)\|. \tag{11}$$

Though both diffusion maps and PCA construct a low-dimensional representation of the data, the two algorithms differ by the following fundamental distinctions. First, in PCA the data is globally viewed as of one piece drawn from some probability distribution and only the second order statistics is regarded. In contrast, diffusion maps combines local connections via the kernel construction and global processing via the spectral decomposition. Second, in PCA the hidden assumption is that the manifold is flat, thus linear projections are appropriate. On the other hand, diffusion maps is nonlinear and the data is embedded in new coordinates rather than linearly projected.

4 Analysis of the Manifold

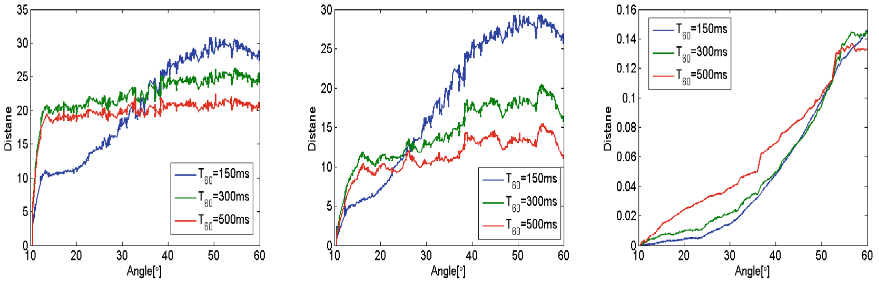
In this section we examine the ability of each of the distance measurements, discussed in this paper, to organize the RTFs according to the corresponding DOA. For this purpose, we used the following setup. A source located in a $6 \times 6.2 \times 3\,\text{m}$ room, is picked up by two microphones located in $(3, 3, 1)\,\text{m}$ and $(3.2, 3, 1)\,\text{m}$, respectively. The position of the source is confined to an arc of $10^\circ \div 60^\circ$ at $2\,\text{m}$ distance with respect to the first microphone. The manifold analysis is carried out using a set of $N = 400$ samples, generated uniformly in the specified range. For each location, we simulate a unique $3\,\text{s}$ speech signal, sampled at 16 kHz. The received signals are obtained by convolving the clean speech signal with the corresponding AIR, simulated based on the image method [1], and contaminated by a WGN with 20 dB SNR. 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 for estimating the RTF in (3) for $D = 2048$ frequency bins.

Both PCA and the diffusion map procedures were applied to the data. For computing the PCA-based distance, we used the projections on the $d = 10$ eigenvectors associated with the 10 largest eigenvalues (d was chosen empirically to obtain maximal range of monotonic behaviour). For the diffusion distance, only the first element in the mapping ($d = 1$) was considered. This choice will be justified in the sequel. All the distance measures were averaged over 50 rotations of the constellation described above with respect to the first microphone.

Figure 1(a) depicts the average Euclidean distance between each of the RTFs and a reference RTF corresponding to 10°, as a function of the angle, for three different reverberation times: 150 ms, 300 ms and 500 ms. We observe a monotonic behavior of the Euclidean distance with respect to the angle, which is confined to a certain region that becomes smaller as the reverberation time increases. Consequently, we conclude that the Euclidean distance is meaningful for small arcs, whose size is determined by the amount of reverberation. This implies that, in general, the Euclidean distance is not a good distance measure between RTFs, however, it can be utilized in the Gaussian kernel for diffusion maps, which takes into account only nearby RTFs through ϵ.

Figure 1(b) depicts the same illustration as Fig. 1(a) for the PCA-based distance. We observe similar trends with respect to the reverberation time compared to that inspected by the Euclidean distance. Here as well, the monotonicity of the distance with respect to the angle is maintained only in a limited region. However, this region is larger than the one exhibited by the Euclidean distance.

It follows that both the Euclidean distance and the PCA-based distance are not appropriate for measuring angles' proximity. The reason is that they both

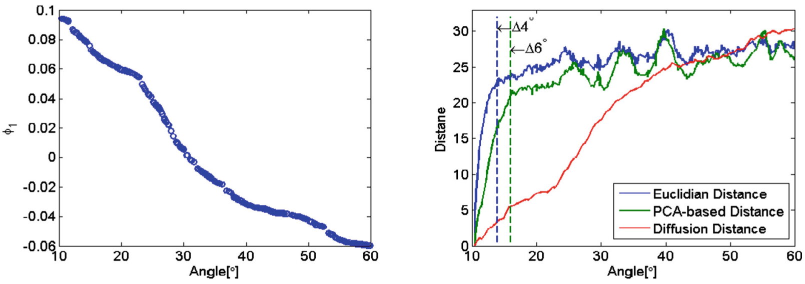


(a) (b) (c)

Fig. 1. Averages of the Euclidean distance (a), the PCA-based distance (b) and the diffusion distance ($\epsilon \approx 50$) (c) between each of the RTFs and the RTF corresponding to 10°, as functions of the angle.

rely on the assumption that the manifold is flat. However, monotonicity is preserved only for local environments indicating that the manifold is only locally linear, capturing its tangent plane, but generally, has a nonlinear structure. From the fact that locality is preserved for smaller regions when reverberation increases, we conclude that the complexity and nonlinearity of the manifold goes hand in hand with the amount of reverberation.

We now turn to the diffusion distance. The kernel scale ϵ should be adjusted to the distance at which monotonicity is maintained by the Euclidean distance, and should ignore longer distances. In Fig. 1(c) we examine the diffusion distance. It can be seen that for almost the entire range, the diffusion distance is monotonic with respect to the angle, indicating that it is an appropriate distance measure in terms of the source DOA. Moreover, by comparing the distance measures in Fig. 1(a)–(c), we observe that the diffusion distance is almost invariant with respect to the reverberation time, whereas the other distance measures significantly vary with the amount of reverberation.



(a) (b)

Fig. 2. (a) Single-element diffusion mapping $\mathbf{\Phi}_1(\cdot)$ for $T_{60} = 300\,\text{ms}$. (b) The three distance measures (normalized to the same scale) between each of the RTFs and the RTF at 10°, as a function of the angle for $T_{60} = 300\,\text{ms}$. The dashed lines show the boundary angles until which monotonicity is preserved.

Further insight into the mapping itself is gained by plotting the single-element mapping $\mathbf{\Phi}_1(\cdot)$, as depicted in Fig. 2(a). 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 position justifies the use of $d = 1$ for estimating the diffusion distance.

In practice we will be interested in a single scenario rather than in the average behaviour. Figure 2(b) compares between the range of monotonicity for each of the distance measures, for a single arbitrary scenario with moderate reverberation time of 300 ms. We observe that the monotonic behaviour is approximately maintained along $\Delta = 4°$ for the Euclidean distance, and along $\Delta = 6°$ for the PCA-based distance. The diffusion distance is monotonic for almost the entire range of $\Delta = 50°$. This confirms our previous conjecture that the diffusion distance is advantageous in measuring the physical position of the source.

5 Conclusions

In this paper we strengthen the claim on the existence of a nonlinear acoustic manifold, whose complexity is influenced by the amount of reverberation. We demonstrate the shortcomings of both the Euclidean distance and the PCA-based distance and their inability to measure the real physical distance. Instead, we propose to use the diffusion distance derived under the diffusion framework, which measures the distance between samples with respect to the manifold. Simulation results show that the diffusion distance properly arranges the RTFs according to the corresponding DOA.

This research lays the foundations for robust source localization algorithms based on a data-driven manifold. Moreover, the existence of an acoustic manifold paves the way for a better understanding of the acoustic environment, and will hopefully lead to simplified and improved acoustic models.

References

1. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
2. Coifman, R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**, 5–30 (2006)
3. Deleforge, A., Forbes, F., Horaud, R.: Variational EM for binaural sound-source separation and localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 76–80 (2013)
4. Deleforge, A., Forbes, F., Horaud, R.: Acoustic space learning for sound-source separation and localization on binaural manifolds. Int. J. Neural Syst. **25**(1), 19 (2015)
5. Deleforge, A., Horaud, R.: 2D sound-source localization on the binaural manifold. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP). Santander, Spain, September 2012
6. Fozunbal, M., Kalker, T., Schafer, R.W.: Multi-channel echo control by model learning. In: The International Workshop on Acoustic Echo and Noise Control (IWAENC). Seattle, Washington, September 2008
7. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
8. Laufer, B., Talmon, R., Gannot, S.: Relative transfer function modeling for supervised source localization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, October 2013
9. Talmon, R., Cohen, I., Gannot, S., Coifman, R.: Diffusion maps for signal processing: a deeper look at manifold-learning techniques based on kernels and graphs. IEEE Signal Process. Mag. **30**(4), 75–86 (2013)
10. Talmon, R., Gannot, S.: Relative transfer function identification on manifolds for supervised GSC beamformers. In: 21st European Signal Processing Conference (EUSIPCO). Marrakech, Morocco, September 2013