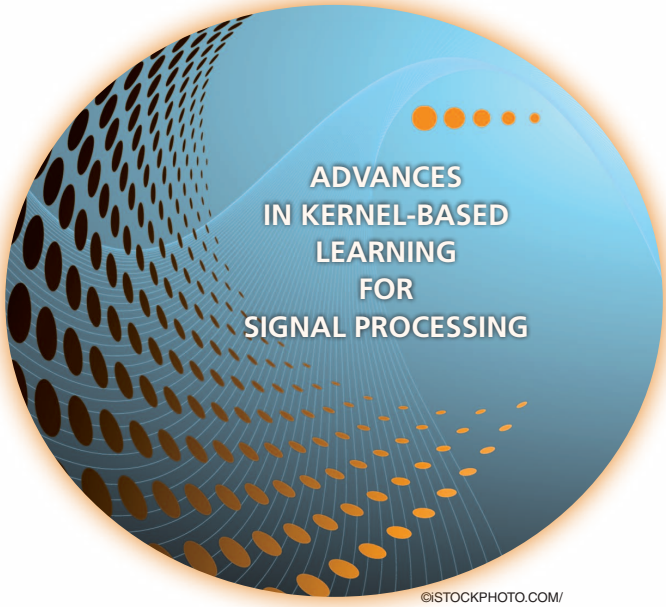
Ronen Talmon, Israel Cohen, Sharon Gannot, and Ronald R. Coifman

Diffusion Maps for Signal Processing



©iSTOCKPHOTO.COM/
ALEKSANDAR VELASEVIC

A deeper look at manifold-learning techniques based on kernels and graphs

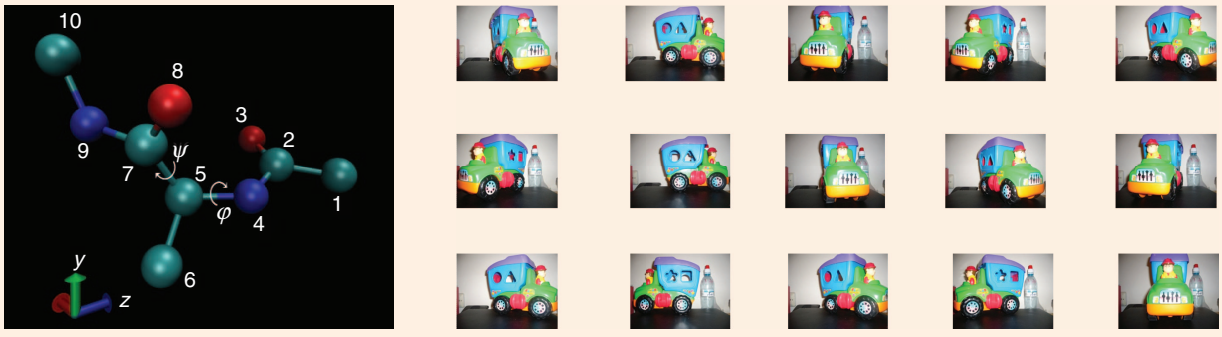
Signal processing methods have significantly changed over the last several decades. Traditional methods were usually based on parametric statistical inference and linear filters. These frameworks have helped to develop efficient algorithms that have often been suitable for implementation on digital signal processing (DSP) systems. Over the years, DSP systems have advanced rapidly, and their computational capabilities have been substantially increased. This development has enabled contemporary signal processing algorithms to incorporate more computations. Consequently, we have recently experienced a growing interaction between signal processing and machine-learning approaches, e.g., Bayesian networks, graphical models, and kernel-based methods, whose computational burden is usually high.

In this article, we review manifold-learning techniques based on kernels and graphs. Our survey covers recent developments and trends and presents ways to incorporate them into signal processing. We integrate theoretical viewpoints, such as compact representations of signals and intrinsic metrics and models, together with practical aspects and concrete signal processing algorithms tackling challenging problems, e.g., transient interference suppression and acoustic source localization. The prime focus is nonlinear signal processing using diffusion maps, which is a recent manifold-learning method. Our hope is that this article will provide an insightful view of these novel methods and will encourage further research into this attractive and exciting new direction.

Digital Object Identifier 10.1109/MSP.2013.2250353
Date of publication: 12 June 2013

MOTIVATION AND BACKGROUND

In a wide range of real-world applications, the observable data sets are controlled by underlying processes or drivers. As a result, these sets exhibit highly redundant representations, and a compact representation of lower-dimensionality exists and may enable a more efficient processing. For instance, molecular dynamics simulations of biologically significant macromolecules (e.g., proteins) provide a unique valuable tool for exploring and developing new drugs and treatments [1]; such simulations describe the motion of a large number of atoms and often occur on time scales well beyond the computational reach of current solvers. Yet, by exploiting the (unknown) coherent structure of molecular motion, such processes can in principle be efficiently represented by only a few, well-chosen reaction coordinates—for example, by a small subset of critical dihedral



[FIG1] Parts (a) and (b) illustrate the degrees of freedom of high-dimensional signals. (a) An alanine dipeptide molecule consisting of ten atoms. The two rotation angles φ and ψ are the main degrees of freedom of the molecule dynamics. Part (b) shows 15 images of a toy, each $127 \times 169 \times 3$ pixels. The toy rotation angle is the sole degree of freedom in the set. (Figure courtesy of Carmeline Dsilva and Neta Rabin.)

angles, rather than by each individual atom trajectory, as illustrated in Figure 1. Similar considerations arise in many signal processing applications. Hence, a fundamental problem is to reveal such latent structures and the associated coarse-grained variables, given their opaque high-dimensional manifestations (such as full atom trajectories). The uncovered latent structures can then be utilized to develop efficient processing algorithms.

The notion of low-dimensional representations is inherent to contemporary data analysis schemes and manifested via the proliferation of seminal computational approaches, such as compressed sensing, sparse representations, and dictionary learning [2], [3]. In particular, there has been a growing effort in recent years to derive analysis schemes based on low-dimensional intrinsic geometry driven by measurements. These are called *manifold-learning techniques* and pave the way for a novel perspective to data analysis, where instead of relying on predefined models and representations, as in the other approaches, the geometry of the data set at hand is captured, and its parametrization is viewed as a data-driven model. The general formulation consists of a low-dimensional latent process in a parametric domain, which is transformed into a set of probability density functions (pdfs) on a statistical manifold [4], [5]. Although statistical representations may be highly suitable for signal processing problems since they may successfully describe random measurements and noise, in this article we focus on a special case in which the set of measurements, and hence the manifold, is deterministic. Such an approach enables to process the measured signal directly without prior knowledge on the distributions nor the need to estimate them.

OVER THE LAST FEW YEARS, WE HAVE PURSUED THE DEVELOPMENT OF GEOMETRIC MODELING AND KERNEL METHODS AND WERE ABLE TO IDENTIFY TWO CRUCIAL BOTTLENECKS THAT LIMIT THEIR INTEGRATION INTO SIGNAL PROCESSING APPLICATIONS.

Manifold learning was first introduced in two pioneer papers published in the same issue of *Science* magazine in 2000: a paper by Tenenbaum et al. presenting the isometric feature mapping (ISOMAP) [6] and a paper by Roweis and Saul presenting locally linear embedding [7]. Following these contributions, many manifold-learning techniques have been proposed in the last decade. Among them we mention Laplacian eigenmaps [8], Hessian eigenmaps [9], and diffusion maps [10]. The core of manifold learning resides in the construction of a kernel based on the connections between the samples of a signal. The main idea is that eigenvectors of the kernel can be viewed as a compact representation of the signal samples. Consequently, the data can be represented (embedded) in a Euclidean space. In addition, the hope is that this new representation will capture the main structures of the data in few dimensions, thereby achieving dimensionality reduction. Diffusion maps, which were introduced by Coifman and Lafon [10], are of particular interest as they provide a general framework that combines many of the other algorithms. Figure 1 presents a toy example demonstrating this goal. Consider a data set of 15 images of a toy. As seen from the images, the sole degree of freedom, which should be revealed to "organize" the images, is the different rotation angle of the toy. This problem is nonlinear and challenging due to the large number of dimensions and small number of samples (images). Later in this article, we demonstrate how this angle can be recovered without any prior knowledge, treating the images merely as high-dimensional vectors.

Over the last few years, we have pursued the development of geometric modeling and kernel methods and were able to identify two crucial bottlenecks that limit their integration

into signal processing applications. First, the measurements are highly dependent on the measuring instrument and the ambient noise conditions. Thus, the analysis of the geometry of the measured signal at hand might be highly influenced by the measurement and instrumental modality, and to successfully apply this approach to signal processing, new robust methods, which capture the intrinsic geometry, should be developed. Second, the ability to sequentially handle streaming data is another important aspect of signal processing. Currently, the common application of geometric analysis methods lies mainly in the areas of machine learning and data mining. Most of the research has been focused thus far on batch processing tasks such as recognition, organization, clustering, and classification of (high-dimensional) data sets. Coping with streaming data raises challenging tasks. For example, data-driven models must be extended accurately and efficiently to newly acquired observations, the acquired geometric information might need to be combined with a statistical model, and the empirical representation should be used for processing. This article attempts to provide answers to these questions from both theoretical and practical perspectives. We provide the readers with a comprehensive review of a particular manifold-learning approach, mainly diffusion maps, and demonstrate its usefulness in signal processing and its relationship to other kernel methods. In addition, we present recent developments that expand the field of kernel methods and specifically designed to handle the aforementioned challenges of signal processing problems.

We begin by describing the geometric approach for data analysis and processing. In particular, we review diffusion maps in detail. We introduce the fundamental concept and ideas and describe the formulation and analysis. Following the introduction, we survey recent developments in geometric methods. Specifically, we focus on affinity metrics based on local models that are incorporated into a diffusion kernel that provides a global model for the signal. This emphasizes the essence of the proposed approach—exploiting both the local and the global structure of the data in the form of distance measures and a kernel, respectively. Then we attempt to provide answers to the paramount questions: how to incorporate the obtained geometric characterization into signal processing tasks, and how to combine the data-driven model with predefined statistical models. We introduce two filtering schemes exploiting the special properties of diffusion maps. The first is based on nonlocal (NL) filters equipped with an intrinsic metric driven by kernels, and the second relies on linear projections on learned dictionaries. Finally, we present the application of the reviewed approaches to two challenging tasks in signal processing: single-channel source localization and transient interference suppression. In particular, we show that locating a sound source based on the geometry of recordings from a single-channel is possible. It further implies that acoustic impulse responses can be indeed parameterized, a fact that deviates from the common belief in this field. In addition, we describe a solution to transient interference suppression. Transient interferences, e.g., keyboard typing, are examples to signals, which are often encountered in real-life speech processing applications and whose representations using temporal statistical models are poor. Thus, existing methods based on statistical signal processing are inefficient in suppressing transients. Instead, we show that techniques based on the geometric structure of transients provide natural solutions and yield state-of-the-art results.

GEOMETRIC APPROACH FOR DATA ANALYSIS AND PROCESSING

Consider a set $\{\mathbf{x}_i\}_{i=1}^{M}$ of M (possibly high-dimensional) signal samples. In the general setting, i is merely an index of a sample in the data set and often denotes the time index of time series. For example, in the case of the images in Figure 1(b), $M = 15$ and $\mathbf{x}_i$ is a column stack representation of a single image. The diffusion framework consists of the following steps:

1) construction of a weighted graph G on the signal at hand based on a pairwise weight function k that measures the affinity between samples.
2) definition of a Markov process on the graph G via a construction of a transition matrix that is derived from k.
3) nonlinear mapping of the samples into a new embedded space based on a parametrization of the graph, which forms an intrinsic representation of the signal.

We note that Steps 1 and 3 are usually common steps in typical manifold-learning techniques. We will elaborate now on the various steps.

BUILDING A GRAPH

We construct the graph G on the samples of the signal. Let $k_\sigma(\mathbf{x}_i, \mathbf{x}_j)$ be a kernel or a weight function representing a notion of pairwise affinity between the samples, with a scale parameter σ. The kernel has the following properties: 1) symmetry: $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) = k_\sigma(\mathbf{x}_j, \mathbf{x}_i)$; 2) nonnegativity: $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) \geq 0$; and 3) locality: given a positive scale parameter $\sigma > 0$, $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) \to 1$ for $\|\mathbf{x}_i - \mathbf{x}_j\| \ll \sigma$ and $k(\mathbf{x}_i, \mathbf{x}_j) \to 0$ for $\|\mathbf{x}_i - \mathbf{x}_j\| \gg \sigma$. For example, a Gaussian kernel $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\}$ satisfies these properties. For simplicity, we omit the notation of the scale σ when referring to the kernel k.

Setting the scale has a great importance and has been a subject of many studies; e.g., see [11]. A small scale intensifies the notion of locality, however, it may result in a poorly connected graph. On the other hand, a large scale guarantees graph connectivity but makes the kernel insensitive to variations in the distances. In practice, the scale is often determined as the empirical standard deviation of the data set, however, there exist analytic methods for setting the scale, which we will not address in this article.

Based on the kernel, we form a weighted graph G, where the samples are the graph nodes and the weight of the edge connecting node $\mathbf{x}_i$ to node $\mathbf{x}_j$ is $k(\mathbf{x}_i, \mathbf{x}_j)$. A kernel with a notion of locality (Property 3) defines a neighborhood around each sample $\mathbf{x}_i$ of radius σ (in other words, samples $\mathbf{x}_j$ s.t. $\|\mathbf{x}_i - \mathbf{x}_j\|^2 > \sigma$ are weakly connected to $\mathbf{x}_i$). Thus, the

choice of the specific kernel function should be application oriented to yield meaningful connections and represent perceptual affinity. In practical applications, the Euclidean distance in the kernel can be replaced by any application-oriented metric. Alternatively, the kernel may be based on the Euclidean distance between features of the signal. It is worthwhile to note that the construction of the graph based on a kernel with a notion of locality is different than global methods. For example, principal component analysis (PCA) does not rely on a graph and is solely based on global statistical correlations between samples in the entire set. Kernel PCA proposed by Scholkopf et al. [12], on the other hand, is similar to diffusion maps since both methods rely on local connections conveyed by a kernel. However, the following steps will distinguish between diffusion maps and kernel PCA.

CONSTRUCTING A MARKOV CHAIN

A classical construction in spectral graph theory, as described by Chung [13], is employed and the kernel is normalized to create a nonsymmetric pairwise metric, given by $p(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)/d(\mathbf{x}_i)$, where $d(\mathbf{x}_i) = \sum_{j=1}^{M} = k(\mathbf{x}_i, \mathbf{x}_j)$ and is often referred to as the degree or the local density of $\mathbf{x}_i$. Using the nonnegativity property of the kernel, which yields that $p(\mathbf{x}_i, \mathbf{x}_j) > 0$, and since $\sum_{j=1}^{M} p(\mathbf{x}_i, \mathbf{x}_j) = 1$, the function p may be interpreted as a transition probability function of a Markov chain on the graph nodes. Specifically, the states of the Markov chain are the graph nodes $\{\mathbf{x}_i\}$ and $p(\mathbf{x}_i, \mathbf{x}_j)$ represents the probability of transition in a single Markov step from node $\mathbf{x}_i$ to node $\mathbf{x}_j$. We note that p is not described in a conventional conditional probability notation to emphasize its role as a nonsymmetric pairwise metric and also maintaining the common notation from the literature.

The underlying assumption of this approach is that the observations are samples of a continuous-time propagation model, which implies a particular organization of the samples. Thus, the objective of the graph is to quantify the pairwise connections between the samples and to uncover the global propagation model. For more details, see [10] and [14].

LAPLACIAN EIGENMAPS, DIFFUSION MAPS, AND DIFFUSION DISTANCE

Let $\mathbf{K}$ denote the kernel matrix, where its (i, j)th element is $k(\mathbf{x}_i, \mathbf{x}_j)$, and let $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}$ be the transition matrix corresponding to p, where $\mathbf{D}$ is a diagonal matrix with the diagonal elements $d(\mathbf{x}_i)$. Propagating the Markov chain t steps forward corresponds to raising $\mathbf{P}$ to the power of t. We also denote the probability function from node $\mathbf{x}_i$ to node $\mathbf{x}_j$ in t steps, which corresponds to $\mathbf{P}^t$, by $p_t(\mathbf{x}_i, \mathbf{x}_j)$.

Applying a singular value decomposition (SVD) to $\mathbf{P}$ yields a complete sequence of left and right singular vectors $\{\boldsymbol{\varphi}_j, \boldsymbol{\psi}_j\}$ and singular values $\{\lambda_j\}$, which are nonnegative and bounded by 1 due to the normalization. The singular values are sorted in a descending order such that $\lambda_0 = 1$. The corresponding right singular vector is trivial and equals to a column vector of ones $\boldsymbol{\psi}_0 = \mathbf{1}$.

The right singular vectors of the transition matrix $\mathbf{P}$ are used to obtain a new data-driven description of the M vectors $\{\mathbf{x}_i\}$ via a family of mappings that are termed *diffusion maps* [10]. Let $\boldsymbol{\Psi}_t(\mathbf{x}_i)$ be the diffusion mappings of the samples into a Euclidean space $\mathbb{R}^{\ell}$, defined as

$$\boldsymbol{\Psi}_t(\mathbf{x}_i) = [\lambda_1^t \boldsymbol{\psi}_1(i), \ldots, \lambda_{\ell}^t \boldsymbol{\psi}_{\ell}(i)]^T, \tag{1}$$

where ℓ is the new space dimensionality ranging between 1 and $M - 1$. Diffusion maps has therefore two parameters, which have to be set by the user: t and the dimension ℓ. The former corresponds to the number of steps of the Markov process on the graph, since the transition matrices $\mathbf{P}$ and $\mathbf{P}^t$ share the same singular vectors, and the singular values of $\mathbf{P}^t$ are the singular values of $\mathbf{P}$ raised to the power of t. The latter indicates the intrinsic dimensionality of the data. Many studies in the literature investigate the problem of intrinsic dimensionality from different perspectives, e.g., a geometric point of view was applied by Kégl [15]. Here, we mention a heuristic approach, exploiting the fact that the eigenvalues are a part of the diffusion maps embedding. The dimension may be set according to the decay rate of the eigenvalues, as the coordinates in (1) become negligible for a large ℓ. In practice when the underlying representation is dominant and the ambient noise is relatively low, we expect to see a distinct "spectral gap" in the decay of the eigenvalues. Such a gap is often a good indicator of the intrinsic dimensionality of the data and its use is a common practice in spectral clustering methods. However, in many (noisy) cases the spectral gap might not be distinct, and hence, the tendency to overestimate the dimensionality may increase the chance of diffusion maps to include the true intrinsic representation. We note that as t increases, the decay rate of the singular values also increases (they are confined in the interval [0, 1]). As a result, we may set ℓ to be smaller, enabling the capture of the underlying structure of the samples in fewer dimensions. Thus, we may claim that a larger number of steps usually brings the samples closer in the sense of the affinity implied by $\mathbf{P}^t$, and therefore, a more "global" structure of the signal is revealed. An example to the ability of diffusion maps to capture the degrees of freedom in a signal is demonstrated by the images of the toy in Figure 2.

IT IS WORTHWHILE NOTING THAT DIFFUSION MAPS ARE REDUCED TO LAPLACIAN EIGENMAPS WHEN THE EIGENVALUES ARE DISCARDED FROM THE MAPPING (USING MERELY THE EIGENVECTORS).

It is worthwhile noting that diffusion maps are reduced to Laplacian eigenmaps [8] when the eigenvalues are discarded from the mapping (using merely the eigenvectors). It implies that Laplacian eigenmaps carry the information on the underlying model and the temporal dynamics without the information

revealed by the Markov process. In addition, we emphasize that manifold-learning methods, e.g., Laplacian eigenmaps and diffusion maps, typically use the eigenvectors themselves as a nonlinear representation of the data. This is a major difference from kernel PCA, in which the data is linearly projected on the obtained eigenvectors of the kernel.

As we described before, the Markov process enables the integration of information from the entire set into the affinity metric $p_t(\mathbf{x}_i, \mathbf{x}_j)$ between two individual samples. This advantage is further expressed in the following derivation of a new affinity metric between any two samples [10], which is defined as

$$D_t^2(\mathbf{x}_i, \mathbf{x}_j) = \| p_t(\mathbf{x}_i, \cdot) - p_t(\mathbf{x}_j, \cdot) \|_{\varphi_0}^2 = \sum_{l=1}^{M} (p_t(\mathbf{x}_i, \mathbf{x}_l) - p_t(\mathbf{x}_j, \mathbf{x}_l))^2 / \boldsymbol{\varphi}_0(l), \quad (2)$$

for any t. This metric is termed *diffusion distance* as it relates to the evolution of the transition probability distribution $p_t(\mathbf{x}_i, \mathbf{x}_j)$. It enables the description of the relationship between pairs of samples in terms of their graph connectivity (see Figure 3). Consequently, the main advantage of the diffusion distance is that local structures and rules of transitions are integrated into a global metric.

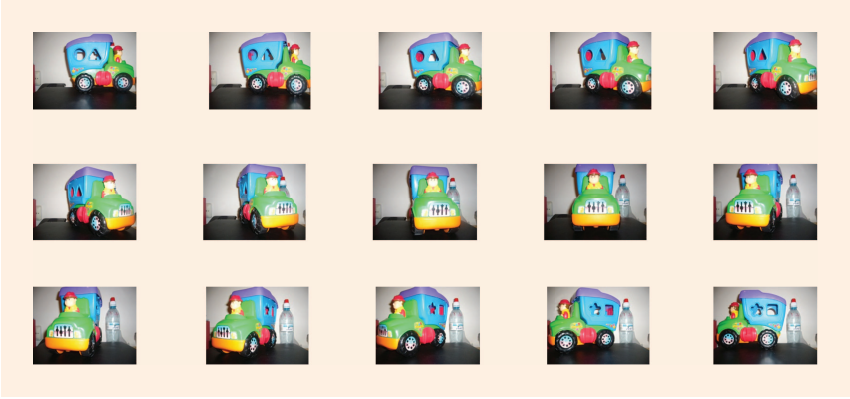
It can be shown that the diffusion distance (2) is equal to the Euclidean distance in the diffusion maps space when using all $\ell = M - 1$ eigenvectors [10], i.e.,

$$D_t^2(\mathbf{x}_i, \mathbf{x}_j) = \| \boldsymbol{\Psi}_t(\mathbf{x}_i) - \boldsymbol{\Psi}_t(\mathbf{x}_j) \|^2. \quad (3)$$

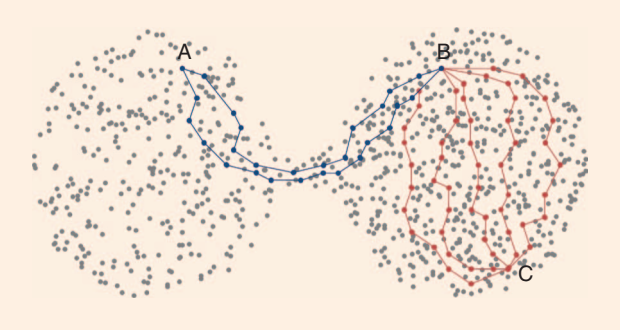
Thus, comparing between embedded samples using the Euclidean distance conveys the advantages of the diffusion distance stated above. In addition, since the spectrum is fast decaying for large enough t, the diffusion distance can be well approximated by only the first few ℓ eigenvectors. Thus, the diffusion distance can be efficiently approximated by the Euclidean distance between embedded samples in low dimensions (setting ℓ to a small value). Enabling meaningful and efficient comparisons between samples makes the Euclidean distance in the new embedded space highly useful. In recent years, this metric was shown to be very effective in various applications from different fields [17], and in this article, we intend to apply it to transient interference suppression in speech signals.

AFFINITY METRICS BASED ON LOCAL MODELS

One of the main advantages of kernel methods compared to classical PCA is the ability to "think globally, fit locally" [6], [7]. The kernel is usually based on an affinity metric between two samples representing local connections, and globally processed via, for example, the eigenvalue decomposition. Recent developments in this field have taken this approach another step forward. The fact that the same metric is used to measure affinity everywhere poses limitations. In addition, the common Gaussian kernel based on the Euclidean distance between the samples is often inadequate to signal processing tasks. In practice, measuring the same phenomena several times usually yields different measurement realizations. Also, the same phenomena can be measured using different instruments or sensors. As a result, each set of related measurements of the same phenomenon will have a different geometric structure, depending on the instrument and the specific realization, and as a result, different pairwise Euclidean distances. Thus, classical manifold-learning methods (including Laplacian eigenmaps and diffusion maps) provide parametrization of merely the observable manifold rather than the desired underlying manifold; this is a significant shortcoming of a kernel based on a Gaussian.



[FIG2] The 15 images from Figure 1(b) sorted according to the diffusion maps embedding. The diffusion maps are computed using the standard Gaussian kernel between the pixels of the images. The dimensionality of the embedding is set to 1, and thus, diffusion maps associate each image with a scalar. (Images courtesy of Neta Rabin.)



[FIG3] The Euclidean distance and the geodesic distance (shortest path) between points A and B are roughly the same as those between points B and C. However, the diffusion distance between A and B is much larger since it has to go through the bottleneck. Thus, the diffusion distance between two individual points incorporates information of the entire set, implying that A and B reside in different clusters, whereas B and C are in the same cluster. This example is closely related to spectral clustering. For details, see [16] and references therein.

A possible way to circumvent the problem is to use application-oriented features and then to compute the Euclidean distance between the features. Recent developments describe alternative ways by relying on universal distance measures based on local structures in the data, which are in turn integrated in the graph to form a global model. In this section, we review two such methods.

INTRINSIC DISTANCE

In signal processing, the cases described above are often represented in a state-space. Let $\boldsymbol{\theta}_i$ be the underlying desired samples of the "state," and let $\mathbf{x}_i$ be the corresponding samples of the measurement, given by $\mathbf{x}_i = f(\boldsymbol{\theta}_i)$, where f is a map (possible nonlinear and probabilistic) which relates the underlying samples to the measurement samples.

Singer and Coifman proposed to compute an intrinsic affinity metric in the underlying (state) space from the measurements [18]. They proposed to divide the measured samples of the signal into N small neighborhoods or "clouds" representing N typical states. Let $\{\mathbf{x}_{i,j}\}_j$ be the set of samples from the ith cloud and let $\mathbf{C}_i$ be their empirical covariance matrix. In practice, these clouds can be computed in several way: 1) to be picked by hand according to prior information; 2) by a nearest neighbors search based on application-oriented features; and 3) in the case of time series, the clouds may be defined according to short windows in time around each sample $\mathbf{x}_i$. Either way, the clouds should consist of measurement samples that correspond to similar underlying samples ("states"), which represent the local variability of the signal. Based on the partition to clouds, a pairwise distance is computed according to

$$d^2_{i,i'}(\mathbf{x}_{i,j}, \mathbf{x}_{i',j'}) = \frac{1}{2}(\mathbf{x}_{i,j} - \mathbf{x}_{i',j'})^T(\mathbf{C}_i^{-1} + \mathbf{C}_{i'}^{-1})(\mathbf{x}_{i,j} - \mathbf{x}_{i',j'}). \quad (4)$$

The distance in (4) is known as the Mahalanobis distance. We note that the rank of the local covariance matrices is equal to the intrinsic dimension and is often smaller than the ambient dimension of the measured data. Thus, we should use the pseudoinverse of the matrices in (4). In addition, the empirical rank of the matrices may serve as additional indicators of the intrinsic dimensionality. Based on a local linearization of the measurement function f, it was shown by Singer and Coifman [18] that the Mahalanobis distance approximates the Euclidean distance between samples of the underlying process to the second order, i.e.,

$$\|\boldsymbol{\theta}_{i,j} - \boldsymbol{\theta}_{i',j'}\|^2 = d^2_{i,i'}(\mathbf{x}_{i,j}, \mathbf{x}_{i',j'}) + O(\|\mathbf{x}_{i,j} - \mathbf{x}_{i',j'}\|^4), \quad (5)$$

where $\boldsymbol{\theta}_{i,j}$ and $\boldsymbol{\theta}_{i',j'}$ are the underlying samples corresponding to $\mathbf{x}_{i,j}$ and $\mathbf{x}_{i',j'}$, respectively. The Mahalanobis distance is a metric between the measurements and can be empirically computed based on the available data; it inverts the measurement function, and therefore, intrinsic. Thus, it may be a very useful distance measure in signal processing applications that replaces the Euclidean distance in a Gaussian kernel.

We applied this approach to the problem of modeling convolution systems [19]. This problem has a key role in signal processing and has long been a task that attracted a considerable research effort. A predefined model is traditionally developed for every type of system, and then the parameters of the model are estimated from observations. We take a different approach. A given system is viewed as a black box controlled by several controlling parameters. By recovering these parameters, we reveal the actual degrees of freedom of the system and obtain its intrinsic modeling. These attractive features are extremely useful for system design, control, and calibration.

Musical instruments are examples of such systems, as each musical instrument is controlled by several parameters. For example, a flute is controlled by covering its holes. Formally, the parameter space can be written as a binary space, assuming the flute has holes and each hole can be either open or covered. An important observation is that the output signal of the flute depends on the blow of air (the input signal) and the covering of the holes. However, the audible music (the notes) depends only on the covering of the holes. In other words, the played music depends solely on a finite set of the instrument's controlling parameters. Another example worth mentioning is a violin. Violin music is determined by the length of the strings. In both examples, by recovering the controlling parameters of the musical instrument, we may naturally characterize the music and identify the played notes. Furthermore, the intrinsic local metric enables to compare signals according to the values of their underlying parameters, thereby providing perceptual comparisons. In these examples, the clouds consist of different measurements of the same notes, which are represented by similar underlying states. It is worthwhile noting that the state is naturally determined by the problem, e.g., as the holes of a flute or length of strings of a violin, unlike traditional formulations in which the state is arbitrarily set. In addition, the use of such an intrinsic metric enables to recover the natural degrees of freedom in the system (the true number of holes or strings) and circumvents the need to use an over representation of the parameters, as is often required by common methods.

LOCAL PCA MODELS

The second metric is based on trained local principal components models [20], which are designed to filter the desired information from the observations [21], [22]. In diffusion maps, for example, the graph connects nodes with similar features. To enhance this property, we define a local data-driven model in local neighborhoods or clouds. A well-known limitation of PCA is that it is linear and able to capture only the global structure of the signal, whereas given signals admit complicated structures. Thus, a low-dimensional linear subspace may not faithfully describe the information. However, a PCA-based approach may perform rather well when applied locally, i.e., on a data set sufficiently condensed in a small neighborhood. In this setting, it corresponds to defining a model for each local region separately. Incorporating these local models in the graph then provides integration of

all the acquired models. Capitalizing the connections between the entire set of data, rather than using a single local model, attains significantly improved results. We assume each cloud of samples represents a certain class of the signal and consists of several instances representing the variability of the class.

As described above, in music, each class can represent a note and the cloud can consist of several different instances of that note. Let $\bar{x}_i$ be the empirical mean of the ith cloud and let C_i be the corresponding empirical covariance matrix of its samples. The pair $(\bar{x}_i, C_i)$ may be used as the learned model of the ith cloud. This implicit Gaussian representation is set for simplicity. By employing PCA, the largest eigenvectors of C_i, which correspond to the principal "parameters," capture most of the information disclosed in each cloud. Hence, the dimensionality is significantly reduced by considering only the subspace spanned by a few principal eigenvectors. Let $\{\mathbf{v}_{i,j}\}_{j=1}^{L}$ be the set of L such principal eigenvectors. Let P_i be a linear projection operator of each sample onto the ith local model, defined as

$$P_i(\mathbf{x}_k) = \bar{\mathbf{x}}_i + \sum_{j=1}^{L} \langle \mathbf{x}_k - \bar{\mathbf{x}}_i, \mathbf{v}_{i,j} \rangle \mathbf{v}_{i,j}, \tag{6}$$

using the standard inner product. Based on the projection, we define a pairwise metric associated with the local model of the ith cloud as

$$d_i(\mathbf{x}_k, \mathbf{x}_l) = \| P_i(\mathbf{x}_k) - P_i(\mathbf{x}_l) \|. \tag{7}$$

The linear projection in (6) extracts essential information and the local metric in (7) enables to adjust the kernel computation. We note that the cloud should be given to be able to use the correct metric.

Both local metrics described in this article exploit the information stored in the local covariance of the samples. In the former approach, the covariance matrices are used to define an intrinsic metric between the samples by locally inverting the measurement function. In the latter, the covariance matrices are viewed as features or a local dictionary of the signal.

INCORPORATING THE GEOMETRIC INFORMATION INTO SEQUENTIAL FILTERING SCHEMES

Thus far, we have presented data-driven methods to obtain models to measured signals. In this section, we present ways to incorporate these models into sequential filtering. The ability to sequentially handle streaming data is an inherent aspect of signal processing, in which the data-driven geometric models must be extended accurately and efficiently to newly acquired observations. However, the common derivation of geometric analysis methods takes the form of batch processing. In this section, we begin with a presentation of an extension scheme that enables efficient sequential processing. Then we present two filters that incorporate the extracted geometric representation by exploiting the diffusion distance and that the eigenvectors of the Markov matrix form a complete orthonormal basis of the signals, respectively. We note that the latter is not a unique property of diffusion maps and therefore it can be replaced by other manifold-learning methods.

SEQUENTIAL PROCESSING

Recently, we have proposed an extension method that relies on a probabilistic interpretation of the kernel [21]. Let $\{\tilde{\mathbf{x}}_i\}$ be a set of samples available in advance. We refer to these samples as a training set that is used to learn the model. We define a nonsymmetric kernel $\mathbf{A}$ between any new unseen set of measurements $\{\mathbf{x}_j\}$ and the training set. The (j,i)th element of the kernel is defined as the probability of the unseen sample $\mathbf{x}_j$ given it is "associated" with the ith local probability class, i.e.,

$$A^{ji} = \frac{1}{d_j} \Pr\{\mathbf{x}_j \,|\, \mathbf{x}_j \in \mathcal{X}_i\}, \tag{8}$$

where $\mathcal{X}_i$ is the local probability class defined by the training sample $\tilde{\mathbf{x}}_i$, and d_j is a normalization factor such that $\sum_i A^{ji} = 1$. An example for a local probability class is a Gaussian $\mathcal{N}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{C}}_i)$ defined in local clouds, as described in the previous section. It was shown by Kushnir et al. [23] that in such a case, $\mathbf{A}^T\mathbf{A}$ is a Gaussian kernel for the training samples, which can be computed directly as described in the previous sections. In addition, $\mathbf{A}\mathbf{A}^T$ serves as an extended kernel defined on the unseen samples, whose (j, j')th element is given by the probability of two unseen samples $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ to be associated with the same local probability class, i.e., $\Pr\{\mathbf{x}_j, \mathbf{x}_{j'} \in \mathcal{X}_i \,|\, \mathbf{x}_j, \mathbf{x}_{j'}\}$. Thus, $\mathbf{A}$ enables to extend the kernel defined on the training set to new samples, and the SVD of $\mathbf{A}$ connects the right and left singular vectors of $\mathbf{A}$ with the eigenvectors of the kernels defined on the training and unseen sample, respectively. It is worthwhile noting that in the nonsymmetric kernel definition in (8) the local model is assumed to be known. As a result, we may use one of the local metrics described in the previous section.

The aforementioned analysis leads to a sequential processing algorithm: 1) in the training stage, the kernel is constructed based on training samples, and its eigen-decomposition is computed; 2) in the test stage, given new samples, the kernel $\mathbf{A}$ between the training samples and the new samples is calculated (in case of a single new point, $\mathbf{A}$ is reduced to a vector); and 3) via the relationship implied by the SVD, the eigenvectors of the extended kernel are efficiently computed indirectly through the algebraic relationship

$$\boldsymbol{\varphi}_j = \frac{1}{\lambda_j} \mathbf{A} \boldsymbol{\psi}_j, \tag{9}$$

circumventing the computation of the entire kernel and the eigenvalue decomposition. A particular attention should be given to the efficiency and low computational complexity of the extension in the test stage. Following is a description of the naïve computational cost (number of operations) for extending the representation to any new sample. Let $\overline{M}$ be the number of training samples. Step 2 involves the computation of the affinity between the new sample and the training samples that yields $\mathcal{O}(\overline{M})$ operations, and by (9), Step 3 yields $\mathcal{O}(\overline{M})$ operations as well.

NONLOCAL FILTERING

NL filters enable signal enhancement and have been proven to be a simple and powerful method, in particular for image denoising [24]. The main idea in NL filtering is to explicitly use the fact that repetitive patterns appear in most natural signals. Specifically, an NL filter is given by

$$\hat{\mathbf{x}}_i = \sum_j \bar{p}(\mathbf{x}_i, \mathbf{x}_j)\mathbf{x}_j, \tag{10}$$

where $\hat{\mathbf{x}}_i$ is the enhanced version of $\mathbf{x}_i$. It implies that each sample of the signal $\mathbf{x}_i$ is enhanced by weighted averaging over other "similar" samples $\mathbf{x}_j$ according to an affinity metric (or kernel) $\bar{p}$. This results in combining together samples from different locations in time or space. Hence, this process is referred to as "nonlocal," whereas "local" filtering is associated with processing of samples from adjacent locations. The affinity kernel, which determines the weights of the NL filter, is of key importance and has a direct impact on the attainable performance. Thus, we should search for a metric that compares the underlying parameters of the sample and is robust to measurement noise.

We presented the diffusion distance (2) as a global measure between samples that incorporates information on the underlying parameters and conveys perceptual connections as captured by diffusion maps. In addition, we showed that it can be empirically computed based on the diffusion maps embedding (3). Thus, it is natural to define a new affinity kernel based on the diffusion distance, e.g., as

$$\bar{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-D_t^2(\mathbf{x}_i, \mathbf{x}_j)/\varepsilon^2\} \tag{11}$$

$$\bar{p}(\mathbf{x}_i, \mathbf{x}_j) = \bar{k}(\mathbf{x}_i, \mathbf{x}_j) \Big/ \sum_l \bar{k}(\mathbf{x}_i, \mathbf{x}_l) \tag{12}$$

and use it in the NL filter (10). We note that the Gaussian kernel implies that remote samples, whose diffusion distances are greater than ε, have negligible weights, and therefore can be discarded. Thus, in practice, the NL filter may be computed based merely on the nearest neighbors.

NONLOCAL FILTERS ENABLE SIGNAL ENHANCEMENT AND HAVE BEEN PROVEN TO BE A SIMPLE AND POWERFUL METHOD, IN PARTICULAR FOR IMAGE DENOISING.

GRAPH-BASED PROCESSING

Next, we present a graph-based framework for sequential processing, which relies on the eigenvectors acquired by diffusion maps. This approach offers sequential filtering compared to the batch NL filtering. By orthogonality, the set of the eigenvectors $\{\boldsymbol{\psi}_j\}_j$ forms a complete basis for any real function defined on the samples. In particular, let $i_k(\mathbf{x}_i)$ be a function that extracts the kth coordinate from the sample $\mathbf{x}_i$. It implies that the kth component of $\mathbf{x}_i$ can be expanded according to the set of eigenvectors as

$$\mathbf{x}_i^k = i_k(\mathbf{x}_i) = \sum_{j=0}^{M-1} \lambda_j \langle i_k, \boldsymbol{\psi}_j \rangle \boldsymbol{\psi}_j(i),$$

where the inner product is defined as $\langle i_k, \boldsymbol{\psi}_j \rangle \triangleq [i_k(\mathbf{x}_1), \ldots, i_k(\mathbf{x}_M)]\boldsymbol{\psi}_j$.

The constructed graph captures the underlying structure of the signal. Consequently, there exists a subset of ℓ eigenvectors which represents the desired information and form a dictionary, whereas the rest of the eigenvectors represent noise. For simplicity, we assume that this subset consists of the leading eigenvectors, i.e., $\{\boldsymbol{\psi}_j\}_{j=0}^{\ell-1}$. In practice, the appropriate eigenvectors might have to be identified manually. A corresponding filter to extract the underlying structure from measurements is defined by a linear projection onto the eigenvectors spanning the underlying parameters subspace, i.e.,

$$\hat{s}_i^k = \sum_{j=0}^{\ell-1} \lambda_j \langle i_k, \boldsymbol{\psi}_j \rangle \boldsymbol{\psi}_j(i), \tag{13}$$

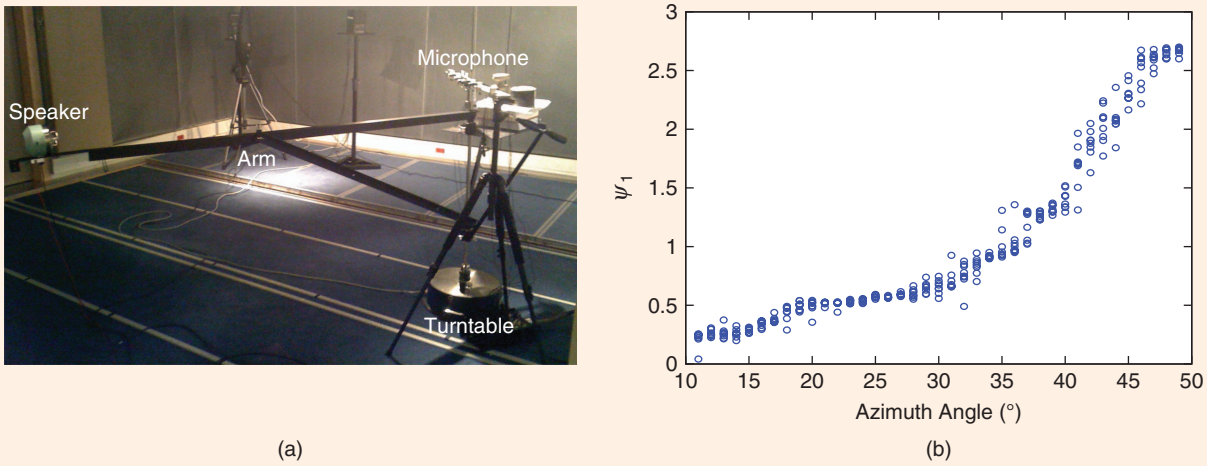
where $\mathbf{s}_i$ is the desired component in the noisy $\mathbf{x}_i$. We note that compared to the previous nonlinear approach (NL filters based on diffusion maps), this processing scheme is based on *linear* projections. In this sense, it is similar to the popular kernel PCA [12]. The new concept lies in the construction of the kernel, which can be based on one of the intrinsic affinity metrics, and will be demonstrated in the next section. In addition to sequential processing, the graph-based filter is characterized by low computational complexity. By (13), obtaining each enhanced coordinate costs $\mathcal{O}(\ell M)$ operations, and hence, the total computational burden to filter each sample is $\mathcal{O}(n\ell M)$, where n is the dimension of $\mathbf{x}_i$.

APPLICATIONS TO SOURCE LOCALIZATION AND SPEECH ENHANCEMENT

The key idea of this novel framework is to combine classical statistical methods with data-driven geometric analysis methods when dealing with real-world applications. We show that capturing the geometric structure of the signals enriches the a priori assumed statistical model and enables good performance. The results presented here indicate that such a combination may be more powerful than strictly model-driven approaches.

SINGLE-CHANNEL SOURCE LOCALIZATION

Assume that a set of single-channel recordings of an unknown source from unknown locations is available, and we aim to find the different locations of the source. At a first glance, this task seems impossible without the spatial information provided by multichannel recordings. To find the position of an acoustic source in a room, the source signal is usually picked up with a microphone array, and the relative delays between pairs of microphone signals are determined. In reverberant rooms, the problem becomes especially challenging since relative room impulse responses between the microphones have to be estimated. Various models of room impulse responses exist in the literature, in which the acoustic path between a source and a



[FIG4] (a) The setup of the recording room at Bar-Ilan University. (b) A scatter plot of the obtained embedding of each measurement as a function of the position azimuth angle.

microphone in an arbitrary enclosure is usually modeled as a long impulse response. Unfortunately, these responses are often difficult to estimate since they consist of a large number of variables.

We observe that a single-channel recording depends on merely few underlying acoustic parameters, e.g., the room dimensions, the positions of the source and microphone, as well as the reflection coefficients of the walls, floor, and ceiling. Thus, a single-channel recording stores the spatial information on the source location, and if we were able to characterize the measurements by the underlying room parameters, we would be able to recover the position of the source. In addition, it implies that the room reverberations may help locating the source, unlike in traditional methods, where reverberations usually reduce the localization performance.

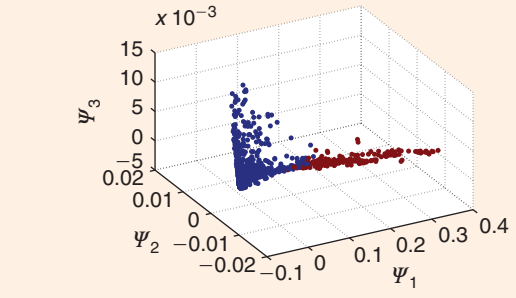
THE ENHANCEMENT OF SPEECH SIGNALS IS OF GREAT INTEREST IN MANY APPLICATIONS RANGING FROM SPEECH RECOGNITION TO HEARING AIDS AND HANDS-FREE MOBILE COMMUNICATION.

In a recent paper [25], we presented a method for locating a source based on diffusion maps and the Mahalanobis distance (4). This method is a specification of a more general algorithm for recovering the parameters of convolution systems [19]. A similar approach has been applied to single-channel source separation (see, for example, Roweis [26]). The following is a description of the experiment that tests the localization on real recordings [25]. Inside a reverberant room, an omnimicrophone was positioned in a fixed location. A 2-m long "arm" was connected to the base of the microphone and attached to a turntable that controlled the horizontal angle of the arm. A speaker was located on the far end of the arm. Thus, the turntable controlled the azimuth angle of the sound played by the speaker with respect to the microphone. Figure 4 depicts the recording room setup.

Several recordings from different angles were performed. From each angle, a zero-mean and unit-variance white Gaussian noise was played from the speaker. The movement of the arm along the entire range of angles was repeated several times. Due to small perturbations of the long arm, we assume that the exact location is not maintained. Thus, the different measurements from the same angle are viewed as a "cloud" of samples. Those clouds are used for the computation of the Mahalanobis distance. The azimuth angle constitutes the sole degree of freedom in this experiment, as the rest of the room parameters, including the location of the microphone, are fixed. Thus, the dimension of the diffusion maps embedding is set to 1. For more details regarding the experimental setup and the algorithm, we refer the readers to [25] and references therein.

We note that the setting of the problem is equivalent to the example presented in Figures 1 and 2. The measured acoustic signals correspond to the different images of the toy. The position of the source plays a similar role as the angle of the toy in the images. Therefore, diffusion maps may reveal the underlying parameters and thereby characterizing the training measurements according to the position of the source.

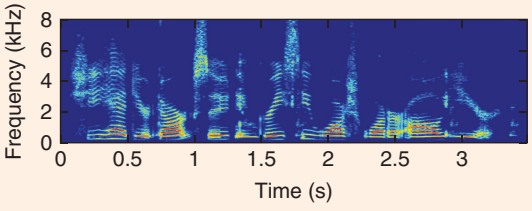
In Figure 4, we show a scatter plot of the values of the obtained leading eigenvector $\boldsymbol{\psi}_1$ as a function of the azimuth angle. It is worthwhile noting that the leading eigenvector is equivalent to the diffusion maps embedding when the dimension is set to 1. We observe that the embedding organizes the measurements according to the azimuth angle. In addition, the relationship between the eigenvector and the angle is close to linear. Thus, this geometric approach was able to reveal and recover the position of the source using single-channel



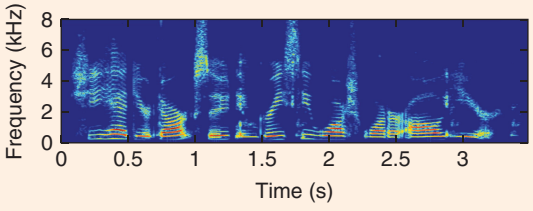
(a)

Transient Type	SNR Improvement (dB)	
	Batch NL Filtering [30]	Online Graph-Based Filtering [21]
Keyboard Typing	9.47	7.78
Household Interferences	5.20	6.62
Door Knocks	8.17	9.79

(b)



(c)



(d)

[FIG5] Transient interference suppression experimental results. (a) Visualization of the diffusion maps embedding in three dimensions. Each point corresponds to a single time frame of the PSD of the noisy speech. This figure depicts points corresponding to speech contaminated by door knocks. The color of the points represents the frame content: frames containing transients appear in brown and frames without transients appear in blue. (b) Speech enhancement evaluation. Parts (c) and (d) show spectrograms of a noisy speech signal containing transient interferences (at .75 s, 1.4 s, 1.75 s, 2.9 s, and 3.3 s) and the enhanced speech signal.

measurements. Based on this embedding, the source position may be easily estimated using training "pivot" measurements from known locations [25].

TRANSIENT INTERFERENCE SUPPRESSION

The enhancement of speech signals is of great interest in many applications ranging from speech recognition to hearing aids and hands-free mobile communication. Although this problem has been studied for several decades, many aspects remain open and require further research. Here, we address the open problem of transient interference suppression and, in particular, suppression of repeating transient appearances, e.g., keyboard typing and construction operations. Transient interference is an example of a signal that does not have a good statistical description, however, exhibits a very distinct geometric structure. Indeed, common speech enhancement algorithms fail to deal with transient interferences since their statistical noise estimation components assume (quasi)stationarity and are not designed to track rapid variations which characterize transients [27], [28].

THE "GEOMETRIC INFORMATION" EXTRACTED BY DIFFUSION MAPS CAN ALSO BE INCORPORATED FOR TRANSIENT INTERFERENCE SUPPRESSION BY GRAPH-BASED FILTERS.

Recently, we devised an algorithm that infers the geometric structure of the transient interference based on the fact that a distinct pattern appears multiple times [29], [30]. The main component of the algorithm is the estimation of the power spectral density (PSD) of the transient interference. Diffusion maps are utilized to compute a robust metric for comparison via the diffusion distance (11), (12). In particular, it enables the clustering of different transient interference types. In fact, the assumption is that the samples of the noisy speech are coarsely distributed similarly to the points in Figure 3, i.e., samples containing the repeated transient events exhibit a distinct structure. Hence, they appear in two clusters according to the presence of transient events, which can be implicitly identified by the embedding. The diffusion distance is then combined with NL filters (10) and enables to enhance the repeating pattern of the transients by averaging over similar instances. On the other hand, the speech samples are very different from each other. Hence, such a NL process yields a destructive averaging of speech samples that reduces their amplitudes. This estimate of the PSD of the transient interference is the missing ingredient in traditional speech enhancement algorithms. Thus, it is added to the estimate of the PSD of the stationary noise and processed using the widely used minimum log spectral amplitude filter [28].

Figure 5 depicts the diffusion maps embedding. Although we may detect the transients by observing the spectrogram, their PSD content does not exhibit any distinct structure that enables to separate them according to the transient presence. However, we observe a clear clustering in the embedding implying that diffusion maps is able to capture the transient presence in few

dimensions. Furthermore, it demonstrates that the Euclidean distance in this domain, which is equal to the diffusion distance, may successfully be used in an NL filter.

The "geometric information" extracted by diffusion maps can also be incorporated for transient interference suppression by graph-based filters. Haddad et al. [22] combined the sequential processing scheme with the graph-based filters and used it to extract textures from images and to identify outliers and anomalies. We used the same technique to suppress transients [21]. In this application, the algorithm infers the transient structure in advance from training recordings which include typical transient interferences. For example, to suppress keyboard typing, representative training key strokes are divided into subsets—organizing strokes with similar sounds in the same subset. Then a local PCA model is computed for each subset and a new intrinsic metric is defined based on linear projections on the local models (7). This enables to construct a graph based on a kernel using the local metric. It is shown that the graph accurately captures the structure of typical transients, and its eigenvectors form a learned dictionary. In the test stage, the learned dictionary is sequentially extended to new incoming noisy samples by building a kernel between the training samples and the new incoming samples (8), (9). Then the graph-based filter (13) is employed to extract the transients (by a linear projection on the "transient dictionary") from the noisy speech and to provide accurate spectrum estimate. We note that the graph filter enables speech enhancement in a sequential manner unlike the batch algorithm based on the NL filter. In addition, it is more efficient and shown to be more robust to handling several transient interference types simultaneously [21] (audio samples are available at http://gauss.math.yale.edu/rt294). In Figure 5, we also demonstrate the performance of the two algorithms. As shown, the enhanced speech exhibits significant transient interference suppression while imposing very low distortion. We note that these algorithms present a solution to this problem for the first time.

SUMMARY

To date, the common application of manifold learning lies mainly in the areas of machine learning and data mining. These methods are usually applied to tasks such as recognition, organization, clustering, and classification of (high-dimensional) data sets. In this article, we intended to show its potential usefulness in signal processing by describing challenging applications that could not be handled using traditional concepts and methods. We began this review by introducing the motivation and background of manifold learning with a focus on a particular method, namely diffusion maps. Then we surveyed recent developments, which included local affinity metrics and filtering schemes. These developments play a key role in bringing manifold learning to signal processing, in particular, by enabling sequential and efficient processing of time series. Finally, we described two algorithms for acoustic source localization and speech enhancement that integrate manifold learning with traditional statistical signal processing. We believe that this is just the tip of the iceberg. Recently, we have been witnessing a growing effort to incorporate kernel methods into signal processing. We are hopeful that this article has contributed insightful views to this effort and will stimulate further research in this emerging direction.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their constructive comments and useful suggestions.

AUTHORS

Ronen Talmon (ronen.talmon@yale.edu) received the B.A. degree (cum laude) in mathematics and computer science from the Open University, Ra'anana, Israel, in 2005 and the Ph.D. degree in electrical engineering from the Technion–Israel Institute of Technology, Haifa, Israel, in 2011. From 2000 to 2005, he was a software developer and researcher at a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a teaching assistant and a project supervisor with the Signal and Image Processing Lab, Electrical Engineering Department, Technion. In 2011, he joined the Mathematics Department at Yale University, where he is currently a Gibbs assistant professor. His research interests are statistical signal processing, analysis and modeling of signals, speech enhancement, applied harmonic analysis, and diffusion geometry. He is the recipient of the 2011 Irwin and Joan Jacobs Fellowship, 2011–2013 Viterbi Fellowship, 2010 Excellent Project Supervisor Award, and the Excellence in Teaching Award for outstanding teaching assistants for 2008 and 2011.

Israel Cohen (icohen@ee.technion.ac.il) is a professor of electrical engineering at the Technion–Israel Institute of Technology, Haifa, Israel. He received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical engineering from the Technion–Israel Institute of Technology in 1990, 1993, and 1998, respectively. From 1990 to 1998, he was a research scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a postdoctoral research associate with the Computer Science Department, Yale University, New Haven, Connecticut. In 2001, he joined the Electrical Engineering Department, Technion. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification, and adaptive filtering. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2008), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), a coeditor of *Speech Processing in Modern Communication: Challenges and Perspectives* (Springer, 2010), and a general cochair of the 2010 International Workshop on Acoustic Echo and Noise Control. He received the Alexander Goldberg Prize for Excellence in Research and the Muriel and David Jacknow Award for Excellence in Teaching. He is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and the IEEE

Speech and Language Processing Technical Committee. He was an associate editor of *IEEE Transactions on Audio, Speech, and Language Processing* and *IEEE Signal Processing Letters*. He was a guest editor of a special issue of the European Association for Signal Processing's (EURASIP's) *Journal on Advances in Signal Processing* on "Advances in Multimicrophone Speech Processing" and a special issue of Elsevier's *Speech Communication* journal on "Speech Enhancement." He is a Senior Member of the IEEE.

Sharon Gannot (Sharon.Gannot@biu.ac.il) received his B.Sc. degree (summa cum laude) from the Technion–Israel Institute of Technology, Haifa, Israel, in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel, in 1995 and 2000, respectively, all in electrical engineering. Currently, he is an associate professor with the Faculty of Engineering, Bar-Ilan University, Israel. He is an associate editor of *IEEE Transactions on Speech, Audio, and Language Processing*. In 2010, he received the Bar-Ilan University Outstanding Lecturer Award. From 2003 to 2012, he was an associate editor of EURASIP's *Journal of Advances in Signal Processing* and an editor of two special issues on multimicrophone speech processing of the same journal. He has also been a guest editor of Elsevier's *Speech Communication* journal and is a reviewer of many IEEE journals and conferences. His research interests include multimicrophone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation, dereverberation, single microphone speech enhancement, and speaker localization and tracking. He is a Senior Member of the IEEE.

Ronald R. Coifman (ronald.coifman@yale.edu) is the Phillips professor of mathematics at Yale University. He received his Ph.D. degree from the University of Geneva in 1965. Prior to coming to Yale in 1980, he was a professor at Washington University in St. Louis, Missouri. His recent publications have been in the areas of nonlinear Fourier analysis, wavelet theory, numerical analysis, and scattering theory. He is currently leading a research program to develop new mathematical tools for efficient transcription of data, with applications to feature extraction recognition, denoising, and information organization. He was chair of the Yale Mathematics Department from 1986 to 1989. He is a member of the National Academy of Sciences, American Academy of Arts and Sciences, and the Connecticut Academy of Sciences and Engineering. He received the 1996 DARPA Sustained Excellence Award, 1996 Connecticut Science Medal, 1999 Pioneer Award from the International Society for Industrial and Applied Mathematics, 1999 National Science Medal, and 2007 Wavelet Pioneer Award.

REFERENCES

[1] G. Hummer and I. G. Kevrekidis, "Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations," *J. Chem. Phys.*, vol. 118, no. 23, pp. 10762–10773, June 2003.

[2] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[4] J. D. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *J. Mach. Learn. Res.*, vol. 6, pp. 129–163, Jan. 2005.

[5] S. M. Lee, A. L. Abbott, and P. A. Araman, "Dimensionality reduction and clustering on statistical manifolds," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–7.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[9] D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. Natl. Acad. Sci. U. S. A*, vol. 100, no. 10, pp. 5591–5596, 2003.

[10] R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 5–30, July 2006.

[11] M. Hein and J. Y. Audibert, "Intrinsic dimensionality estimation of submanifold in r^d," in *Proc. Int. Conf. Machine Learning (ICML)*, 2005, pp. 289–296.

[12] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1996.

[13] F. R. K. Chung, "Spectral Graph Theory," *CBMS Regional Conference Series in Mathematics*, American Mathematical Society, no. 92, 1997.

[14] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *Appl. Comput. Harmon. Anal.*, pp. 113–127, July 2006.

[15] B. Kégl, "Intrinsic dimension estimation using packing numbers," in *Advances in Neural. Information Processing Systems (NIPS)*, 2002, vol. 15, pp. 681–688.

[16] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[17] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1784–1797, Nov. 2006.

[18] A. Singer and R. R. Coifman, "Non-linear independent component analysis with diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 25, no. 2, pp. 226–239, 2008.

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

[20] C. Bregler and S. M. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *Proc. Int. Conf. Computer Vision (ICCV)*, 1995, pp. 494–499.

[21] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 20, no. 9, pp. 2528–2538, Nov. 2012.

[22] A. Haddad, D. Kushnir, and R. R. Coifman, "Filtering via a reference set," Yale University, New Haven, CT, Tech. Rep.YALEU/DCS/TR-1441, Feb. 2011.

[23] D. Kushnir, A. Haddad, and R. Coifman, "Anisotropic diffusion on submanifolds with application to earth structure classification," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 2, pp. 280–294, 2012.

[24] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 490–530, 2005.

[25] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *Proc. IEEE Workshop Applications Signal Processing Audio and Acoustics (WASPAA)*, 2011, pp. 245–248.

[26] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13 (NIPS)*. Cambridge, MA: MIT Press, 2001, pp. 793–799.

[27] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process. (1975–1990)*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[28] I. Cohen and B. Berdugo, "Speech enhancement for non stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[29] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 6, pp. 1584–1599, Aug. 2011.

[30] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression using diffusion maps," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 1, pp. 132–144, Jan. 2013.

[SP]