Transient Noise Reduction Using Nonlocal Diffusion Filters

Ronen Talmon, *Student Member, IEEE*, Israel Cohen, *Senior Member, IEEE*, and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—Enhancement of speech signals for hands-free communication systems has attracted significant research efforts in the last few decades. Still, many aspects and applications remain open and require further research. One of the important open problems is the single-channel transient noise reduction. In this paper, we present a novel approach for transient noise reduction that relies on non-local (NL) neighborhood filters. In particular, we propose an algorithm for the enhancement of a speech signal contaminated by repeating transient noise events. We assume that the time duration of each reoccurring transient event is relatively short compared to speech phonemes and model the speech source as an auto-regressive (AR) process. The proposed algorithm consists of two stages. In the first stage, we estimate the power spectral density (PSD) of the transient noise by employing a NL neighborhood filter. In the second stage, we utilize the optimally modified log spectral amplitude (OM-LSA) estimator for denoising the speech using the noise PSD estimate from the first stage. Based on a statistical model for the measurements and diffusion interpretation of NL filtering, we obtain further insight into the algorithm behavior. In particular, for given transient noise, we determine whether estimation of the noise PSD is feasible using our approach, how to properly set the algorithm parameters, and what is the expected performance of the algorithm. Experimental study shows good results in enhancing speech signals contaminated by transient noise, such as typical household noises, construction sounds, keyboard typing, and metronome clacks.**

***Index Terms*—Acoustic noise, impulse noise, speech enhancement, speech processing, transient noise.**

I. Introduction

ENHANCEMENT of speech signals is of great interest in many hands-free communication systems. Although this problem has attracted significant research efforts for several decades, many aspects remain open and require further research. Among them is the single-channel transient noise reduction. Traditional speech enhancement approaches usually consist of two components: noise power spectrum estimation and estimation of the desired clean speech signal. In single-channel-based applications, spectral information is usually exploited for the estimation of the noise [1]–[6]. In particular, the noise signal is assumed to remain stationary during the observation interval; hence, its power spectral density (PSD) is time-invariant or slowly varying compared to the speech. Another common and fundamental assumption is that the speech signal is not present during the whole observation interval. A common approach for estimating the noise PSD is to average the noisy measurement over periods where the speech is absent. Using the noise PSD estimate, the speech signal can be estimated based on some statistical model.

The assumption of stationary noise signal poses a major limitation on these traditional algorithms, making them inadequate in many transient noise environments. Transient noises are usually characterized by percussive or impulsive nature, i.e., a sudden burst of sound. Typically, transients consist of an initial peak followed by decaying short-duration oscillations of length ranging from 10 to 50 ms. Among them we mention noise originating from engines, keyboard typing, construction operations, bells, knocking, rings, hammering, etc. Vaseghi and Rayner [7], [8] proposed a method for detection and suppression of such impulsive noise, consisting of relatively short duration noise pulses. After detecting a transient, the corrupted segment is completely removed and the source signal is estimated using interpolation that relies on the assumption that the desired source is auto-regressive (AR). Godsill and Rayner [9] improved the algorithm based on a statistical model and interpolation using a Gibbs sampler. Unfortunately, removing the entire corrupted segment is problematic since acceptable speech completion is obtained only for very short transient occurrences.

Traditional methods typically do not take into account the repetitive nature of many transient noises. Usually a distinct pattern appears a large number of times at different time locations. The fact that the same pattern appears multiple times can be utilized for improved denoising. Specifically, the pattern intervals can be identified, and the transient noise may be extracted by averaging over all of these instances.

This approach naturally leads to nonlocal (NL) denoising methods using an NL neighborhood filter [10]. This method combined with local kernels, enables signal denoising with specially tailored locality metrics adapted to specific tasks at hand [11]–[14]. These methods are also known as bilateral filtering. The main idea in nonlocal filtering is to process the data according to the affinity metric conveyed by a kernel, which enables to capture similar patterns. This results in combining together data samples from different locations in time. Hence, this process is referred to as "nonlocal," whereas "local" filtering is associated with processing of data samples from adjacent locations. Although NL averaging is very simple, it is surprisingly superior to other methods. A diffusion

Manuscript received February 23, 2010; revised October 30, 2010; accepted November 02, 2010. Date of publication November 18, 2010; date of current version June 01, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Seltzer.

R. Talmon and I. Cohen are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel (e-mail: ronenta2@techunix.technion.ac.il; icohen@ee.technion.ac.il).

S. Gannot is with the School of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel (e-mail: gannot@eng.biu.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2093651

interpretation of the NL denoising approaches [15], explains the behavior of NL neighborhood filters and enables improved filtering algorithms. The analysis of Singer *et al.* [15] is mainly based on a probabilistic model and on the relation between the averaging process and the eigenstructure of the denoising filter. Although NL neighborhood filters have recently become common in image processing applications, their potential in audio processing in general, and speech enhancement in particular, has not yet been fully investigated.

In this paper, we present a novel approach for speech enhancement that relies on NL filters. In particular we propose an algorithm for the enhancement of nonstationary AR source contaminated with repeating transient noise events. For simplicity, we assume that the time duration of each reoccurring transient event is relatively short compared to speech phonemes and that all events have the same spectral features up to random amplifications, as presented by Vaseghi and Rayner [8]. It is worthwhile noting that we evaluate the proposed algorithm using real signal recordings, since these restrictive assumptions may seem inadequate in practical scenarios. The proposed algorithm consists of two stages. In the first stage, we estimate the PSD of the transient noise. This is achieved by enhancing the transient noise, relying on the strong auto-correlation of the speech signal in time and the burst-like nature of the transient noise. Then, we employ an NL neighborhood filter to extract the transient noise signal. Unlike the approach proposed in [8], we aim at estimating the transient signal rather than just detecting the locations in time of transient events. The NL filter, employed with a specially tailored similarity function, enables to implicitly capture the underlying structure of the measurements. This structure conveys significant information, which may help to distinguish between the transient noise and the speech source signal. In the second stage, we utilize the optimally modified log spectral amplitude (OM-LSA) estimator [4] for denoising the speech with a modified noise PSD estimator, that relies on the extracted transient signal. We note that the noise estimate from the first stage can also be incorporated into other algorithms. For example, in [16], a transient noise reduction algorithm was proposed relying on a given indicator for transient noise events. Our approach may provide an indicator for transient noise events based on NL filtering.

The proposed algorithm is robust to various transient noise types. We show good results in cleaning a speech signal contaminated with transient noise, such as keyboard typing, typical household noises, construction sounds, and metronome clacks. In addition, we present a probabilistic analysis of the NL filtering by introducing a statistical model for the measurements. Based on the diffusion interpretation indicated by Singer *et al.* [15], we obtain further insight into the algorithm behavior. In particular, for given transient noise, we determine whether estimation of the noise PSD is feasible using our approach, how to properly set the algorithm parameters, and what is the expected performance of the algorithm. Recently, we have presented a transient noise reduction algorithm that relies on a modified NL filter [17]. The modified filter is incorporated to obtain further enhancement and robustness. This work extends [17] and provides a probabilistic analysis.

This paper is organized as follows. In Section II, we describe the geometric approach for data analysis in general, and a diffusion framework in particular. In Section III, we formulate the problem of transient noise reduction. In Section IV, we present the proposed algorithm and analyze it in Section V based on diffusion interpretation of the NL filters. Finally, in Section VI experimental results are presented, demonstrating the performance of the proposed algorithm.

II. Diffusion Framework

In recent years, there has been a growing effort to develop data analysis methods based on the geometry of the acquired raw data. These geometric approaches or manifold learning methods aim at discovering the underlying structure in data sets as a precursor to other types of processing [18]–[24]. In particular, among the geometric approaches, *diffusion maps* [24], [25] is of particular interest, since its derivation and some of its main results pave the way for diffusion interpretation of the NL filtering [15]. The proposed algorithm does not involve the actual mapping of diffusion maps; however, it relies on the derivation and main results of this method. Thus, in this section, we present the general formulation of the diffusion framework.

Let $\Gamma = \{\mathbf{x}_i\}_{i=1}^{M}$ be a given high-dimensional data set of M samples, where $\mathbf{x}_i \in \mathbb{F}^N$ and $\mathbb{F}$ is a field. We note that in the general setting, i is merely an index of a sample in the data set. The diffusion framework consists of the following steps: 1) construction of a weighted graph G on the given data set Γ, based on a pairwise weight function k, that corresponds to a local affinity between samples in Γ; 2) derivation of a random-walk on the graph G via a construction of a transition matrix that is derived from k; and 3) interpretation of the discrete random-walk on the graph as a continuous diffusion process on a manifold.

A. Building a Graph

We construct the graph G on the data set Γ in order to capture the geometry of the set. Let $k_\sigma : \mathbb{F}^N \times \mathbb{F}^N \rightarrow \mathbb{R}$ be a kernel or a weight function representing a notion of pairwise affinity between the data samples, with a scale parameter σ. For all $\mathbf{x}_i, \mathbf{x}_j \in \Gamma$, the weight function has the following properties: 1) symmetry: $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) = k_\sigma(\mathbf{x}_j, \mathbf{x}_i)$; 2) non-negativity: $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) \geq 0$; 3) fast decay: given a positive scale parameter $\sigma > 0$, $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) \rightarrow 1$ for $\|\mathbf{x}_i - \mathbf{x}_j\| \ll \sigma$ and $k(\mathbf{x}_i, \mathbf{x}_j) \rightarrow 0$ for $\|\mathbf{x}_i - \mathbf{x}_j\| \gg \sigma$. For example, a Gaussian kernel $k_\sigma(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right\}$ satisfies these properties. It is worthwhile noting that the Euclidean distance can be replaced by any application-oriented metric. For simplicity, we omit the notation of the scale σ, when referring to the kernel k.

Based on the relation defined by the kernel, we form a weighted graph or a Euclidean manifold, where the data samples are the graph G vertices and the kernel k sets the weights of the edges connecting the data points, i.e., the weight of the edge connecting the node $\mathbf{x}_i$ to the node $\mathbf{x}_j$ is $k(\mathbf{x}_i, \mathbf{x}_j)$. It is worthwhile noting that the kernel conveys the local geometry of the data set $\Gamma = \{\mathbf{x}_i\}$, unlike global methods such as principal component analysis (PCA), which are based on statistical

correlations between data samples. Moreover, a kernel with fast decay [property (3)] intensifies the locality property of this approach, as it defines a neighborhood around each data sample $\mathbf{x}_i$ of radius σ (in other words, samples $\mathbf{x}_j$ subject to $\|\mathbf{x}_i - \mathbf{x}_j\|^2 > \sigma$ 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.

B. Constructing a Random Walk

Following classical construction in spectral graph theory [26], the kernel is normalized to create a non-symmetric pairwise metric, given by

$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{k(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)} \tag{1}$$

where $d(\mathbf{x}_i) = \sum_{j=1}^{M} k(\mathbf{x}_i, \mathbf{x}_j)$ is often referred to as the degree of $\mathbf{x}_i$. Using the non-negativity 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 can be interpreted as a transition probability function of a Markov chain[1] on the data set $\Gamma = \{\mathbf{x}_i\}$. 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 random-walk step from node $\mathbf{x}_i$ to node $\mathbf{x}_j$. We point out that p is not described in a conventional conditional probability notation to emphasize its role as a non-symmetric pairwise metric and to correspond with the common notations from the literature. Let $\mathbf{K}$ denote the matrix corresponding to the kernel function $k(\cdot, \cdot)$, 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 matrix corresponding to the function $p(\cdot, \cdot)$, both on the finite data set $\Gamma = \{\mathbf{x}_i\}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = d(\mathbf{x}_i) = \sum_{j=1}^{M} k(\mathbf{x}_i, \mathbf{x}_j)$. Let $\mathbf{X}$ be a matrix consisting of the data set samples

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_M]^T. \tag{2}$$

Advancing the random-walk on the data set a single step forward can be written as $\mathbf{PX}$. Similarly, propagating the random-walk t steps forward corresponds to raising $\mathbf{P}$ to the power of t and applying it on the data set as $\mathbf{P}^t\mathbf{X}$. We denote the probability function corresponding to $\mathbf{P}^t$ as $p_t(\mathbf{x}_i, \mathbf{x}_j)$, which measures the probability of transition from node $\mathbf{x}_i$ to node $\mathbf{x}_j$ in t steps.

Let $\mathcal{X}_\tau$ denote the Markovian process defined by the transition matrix $\mathbf{P}$, with time index τ. The probabilistic interpretation of a single step is:[2]

$$\begin{aligned}[\mathbf{PX}]_i &= \sum_{j=1}^{M} \mathbf{P}_{ij}\mathbf{x}_j \\ &= \sum_{j=1}^{M} \Pr\{\mathcal{X}_{\tau+1} = \mathbf{x}_j | \mathcal{X}_\tau = \mathbf{x}_i\}\,\mathbf{x}_j \\ &= \mathbb{E}\left[\mathcal{X}_{\tau+1} | \mathcal{X}_\tau = \mathbf{x}_i\right]\end{aligned} \tag{3}$$

which means that running the chain forward gives the expected values of the random-walker starting at the node $\mathbf{x}_i$ after a single step. Consequently, performing t steps corresponds to the expected value after t steps. Hopefully, this process results in revealing the relevant geometric structure of $\Gamma = \{\mathbf{x}_i\}$. As we show in Section IV, (3) may be interpreted as a single iteration of an NL filter. In Appendix I, we present a simple example of denoising a telegraph signal corrupted by additive white Gaussian noise to demonstrate the diffusion framework construction.

C. Diffusion Interpretation

Results from spectral theory can be employed to describe $\mathbf{P}^t$, enabling to study the geometric structure of $\Gamma = \{\mathbf{x}_i\}$ in a compact and efficient way. It can be shown that $\mathbf{P}$ has a complete sequence of left and right eigenvectors $\{\boldsymbol{\varphi}_j, \boldsymbol{\psi}_j\}$ and positive eigenvalues, written in a descending order

$$1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \tag{4}$$

satisfying $\mathbf{P}\boldsymbol{\psi}_j = \lambda_j\boldsymbol{\varphi}_j$ and $\mathbf{P}^T\boldsymbol{\varphi}_j = \lambda_j\boldsymbol{\psi}_j$. The eigenvalues $\{\lambda_j\}$ and the eigenvectors $\{\boldsymbol{\varphi}_j, \boldsymbol{\psi}_j\}$ provide a spectral representation of the geometry of the manifold defined by the data set $\Gamma = \{\mathbf{x}_i\}$ and the kernel function $k(\cdot, \cdot)$.

Let $q(\cdot)$ be the probability density function of the data set samples. When using an exponentially decaying kernel, e.g., a Gaussian kernel, it is shown in [24], [25] that for a large data set $M \longrightarrow \infty$ (corresponds to dense sampling of the manifold) and small-scale $\sigma \longrightarrow 0$ (corresponds to very local kernel), the transition matrix $\mathbf{P}$ of the discrete random-walk on the graph converges to the continuous backward Fokker–Planck operator $\mathcal{L}$, defined for any smooth function $f : \Gamma \longrightarrow \mathbb{R}$ by

$$\mathcal{L}f = \Delta f + 2\frac{\nabla q}{q}\nabla f \tag{5}$$

where Δ is defined as $\Delta f = \nabla \cdot (\nabla f)$, ∇f is the gradient of f, and $\cdot$ is the inner product. Let $U(\mathbf{x}_i) = -2\ln q(\mathbf{x}_i)$ be the potential derived from the probability density function $q(\cdot)$ of the data set samples on the manifold. Then Fokker–Planck operator (5) can be written as

$$\mathcal{L}f = \Delta f - \nabla U \nabla f. \tag{6}$$

For example, the Fokker–Planck operator may describe the motion of a particle in a potential field, where the function f denotes the location of the particle. An analysis of the diffusion process associated with the Fokker–Planck operator, and the characteristics of its eigenfunctions have been extensively studied in the literature [27]. The characteristics of the spectral decomposition of the diffusion operator may be exploited for various tasks and applications [15], [25], [28]–[30]. In Section V, we exploit this convergence for the analysis of transient noise extraction using an NL filter.

III. Problem Formulation

Throughout this paper, we use the following notation convention. Lowercase letters denote scalars, bold letters vectors, and capital bold letters matrices. In addition, signals in the time domain are represented by lowercase letters followed by the time index in brackets, whereas signals in the short-time Fourier transform (STFT) domain are represented by lower-case letters

[1] A Markov chain is a discrete random process subject to the next state depends only on the current state.

[2] $[\mathbf{X}]_i$ extracts the ith row of the matrix $\mathbf{X}$.

(to emphasize time variation) followed by a subscript indicating the time frame and frequency bin indices.

Let $s(n)$ denote a speech signal and let $\xi(n)$ be a contaminating interference, represented by

$$\xi(n) = u(n) + d(n) \tag{7}$$

where $d(n)$ is a dominating transient part, and $u(n)$ is a low variance stationary noise. The signal measured by a microphone is given by

$$y(n) = s(n) + \xi(n). \tag{8}$$

For simplicity, in the remainder of this paper we omit the stationary part. The proposed algorithm is designed for distinction of the transient noise from the rest of the measurement components. Evidently, it is much easier to distinguish the stationary noise from the transients, compared to the nonstationary speech. Consequently, the presence of stationary noise does not change significantly the derivation of the algorithm. In Section VI, we show that the proposed algorithm can handle speech contaminated by both transient and stationary noises. It is worthwhile noting that in the literature numerous methods for enhancement of speech signals contaminated by (quasi) stationary noise can be found. In addition, we can employ one of these methods prior to the proposed algorithm.

We assume that the speech signal is modeled as an AR process in short-time frames [31]. The observation interval is divided into M short-time frames of length N. Accordingly, in each time frame $p = 1, \ldots, M$, the source signal is an AR process, given by

$$s(n) = \sum_{l=1}^{L} a_l^p s(n-l) + w(n) \tag{9}$$

where $w(n)$ is a white noise excitation signal with zero-mean and σ_w^2 variance, and $\{a_l^p\}_{l=1}^L$ are L AR coefficients in frame p. We assume L is large enough to capture the long-term linear prediction coefficients, enabling representation of both voiced and unvoiced phonemes. In practice, we verify that the number of coefficients L is greater than a single period of the pitch to enable its representation. In addition, we do not exploit the whiteness of the excitation signal, but rather rely on the fact that the excitation signal can be distinct from the transient noise.

The transient noise consists of short duration pulses of random amplitudes. It may be modeled as the output of a filter excited by an amplitude-modulated random binary sequence [7], [8], [32], given by

$$d(n) = h(n) * (b(n)v(n)) \tag{10}$$

where $b(n) \in \{0, 1\}$ is a binary valued random sequence of time locations of the transient noise events, $v(n)$ is a continuous valued random process of transient amplitudes, and $h(n)$ is an impulse response of a filter that models the duration and shape of each transient event. In this paper, we use a fixed impulse response $h(n)$, which implies that the transient events have the same spectral features up to random amplitude. Hence, the transient noise can be viewed as a superposition of the impulse response $h(n)$ with random amplitudes. This restrictive assumption is used for simplicity. It is worthwhile noting that in Section VI the proposed algorithm is evaluated in practical scenarios using real transient noise recordings. We use the Gaussian–Poisson statistical model proposed in [8], i.e., the random amplitude $v(n)$ is modeled as a Gaussian process with μ_v mean and σ_v^2 variance, and the transient time locations $b(n)$ are modeled as a Poisson process.[3] The Poisson distribution is assumed for simplicity; however, it does not have a significant role in the algorithm derivations, nor does the Gaussian amplitude and the exact pulse shape. For sufficiently low-rate Poisson process, we assume that no more than one transient event exists in each short time frames. We denote by $\mathcal{H}_0$ the set of time frames free of transient noise occurrences, and by $\mathcal{H}_1$, we denote the set of time frames that include transient occurrences.

IV. Proposed Algorithm

The proposed algorithm consists of two stages. In the first stage, we aim at estimating the PSD of the transient noise. In the second stage, we utilize the OM-LSA estimator [4] for denoising the speech. The OM-LSA that we use is equipped with a modified noise PSD estimator, based on the estimation of the transient noise PSD obtained in the first stage.

A. Transient Noise Spectrum Estimation

Aiming at enhancing the characteristic difference between the transient noise and the AR source signal, we "whiten" (or "decorrelate") the noisy measurement $y(n)$ in each time frame using the AR parameters of the source signal. Let $\tilde{y}_p(n)$ be the whitened measurement in time frame p, which can be written as[4]

$$\tilde{y}_p(n) = y(n) - \sum_{l=1}^{L} a_l^p y(n-l). \tag{11}$$

Substituting (7)–(9) into (11) yields

$$\tilde{y}_p(n) = w(n) + \tilde{d}_p(n) \tag{12}$$

where $\tilde{d}_p(n)$ is a smeared version of the transient noise, given by

$$\tilde{d}_p(n) = d(n) - \sum_{l=1}^{L} a_l^p d(n-l). \tag{13}$$

In (12), we observe that the whitened signal consists of two components—the source excitation signal $w(n)$, and a smeared version of the transient noise $\tilde{d}_p(n)$.

The derivation of (11) and (12) is applicable given the AR coefficients of the source signal $\{a_l^p\}$, which are unknown. Estimation of the coefficients of an AR source from noisy measurements has been a subject of many studies and extends the scope of this paper. In practice, we use the common Levinson–Durbin algorithm to estimate the AR coefficients in time frames free of a transient noise event. By exploiting the short duration of transient impulses and by assuming slow variations of the AR

[3]We use a discrete time version of the continuous time Poisson process, as described in [8].

[4]We note that in our notation, tilde denotes a whitened version of the signal.

coefficients in time, we are able to set the AR coefficients in time frames that contain a transient occurrence according to the estimated AR coefficients of neighboring frames.[5] Moreover, as previously noted, the main role of the whitening is to further enhance the distinction between transient noise events and speech components. Therefore in practice, the proposed algorithm is not sensitive to estimation errors of the AR coefficients.

According to our assumption, typical transient events have unique spectral features. Thus, we apply the STFT to emphasize the difference between the transient noise and the source. We use STFT time-frames of length N. Let $\tilde{y}_{p,k}$ be the whitened measurement in the STFT domain in time frame p and frequency bin k. Using (12) and (13), it can be written as

$$\begin{aligned}\tilde{y}_{p,k} &= w_{p,k} + \tilde{d}_{p,k} \\ &= w_{p,k} + (1 - a_{p,k})\, d_{p,k}\end{aligned} \tag{14}$$

where $w_{p,k}$ is the STFT of the excitation signal, $a_{p,k}$ is the multiplicative transfer function (MTF) approximation of the source AR filter [33], and $d_{p,k}$ is the STFT of the transient noise. Using (10), $d_{p,k}$ is given by

$$d_{p,k} = \begin{cases} h_k v_p \exp\left\{-j\frac{k\tau_p}{N}\right\}, & p \in \mathcal{H}_1 \\ 0, & p \in \mathcal{H}_0 \end{cases} \tag{15}$$

where h_k is the MTF approximation of the transient noise system $h(n)$, $v_p \triangleq v(pN + \tau_p)$ is the random amplitude of the impulse, and τ_p is the random relative location of the impulse from the beginning of the frame, both in frame $p \in \mathcal{H}_1$. In (15) we assume a single impulse per frame. In addition, we circumvent overlaps of the transient occurrences between frames by including in $\mathcal{H}_1$ only frames that contain a significant part of a transient event. Since time frames usually overlap, combined with the fact that typical transient events are shorter than the length of a time frame, we are able to assume that frames in $\mathcal{H}_1$ share the characteristic spectral features of a transient event.

Given the AR coefficients of the source, estimating the PSD of the transient noise $d(n)$ and estimating the PSD of the smeared version of the transient noise $\tilde{d}_p(n)$ are equivalent tasks. Furthermore, (12) and (14) imply that the whitened measurements consist of the smeared version of the transient noise "contaminated" by a white noise $w_{p,k}$. Consequently, we can interpret the estimation of the transient noise PSD as a problem of spectral denoising, where we aim at enhancing the transient events and attenuating the white excitation signal. For that purpose, we employ an NL filter that exploits the divergence between the STFT features of the transient events and the white excitation signal.

Consider the STFT features of each time frame as a single sample in a high-dimensional field. Specifically, let $\tilde{\mathbf{y}}_p$ be a vector of the STFT coefficients from all frequency bins in the pth time-frame of the whitened signal $\tilde{y}_p(n)$, defined as

$$\tilde{\mathbf{y}}_p = [\tilde{y}_{p,0}, \ldots, \tilde{y}_{p,N-1}]^T \tag{16}$$

[5]We note that speech onset or phoneme transition right before or after the transient results in inaccurate estimation of the AR coefficients. We assume such cases occur with very low probability.

and let $\tilde{\mathbf{Y}}$ be an $M \times N$ matrix consisting of all these vectors, given by

$$\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_M]^T. \tag{17}$$

We define an affinity kernel $k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$ between pairs of samples $\tilde{\mathbf{y}}_p$ and $\tilde{\mathbf{y}}_l$ for all p and l. In this paper, we use the following Gaussian kernel based on a Euclidean distance:

$$k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l) = \exp\left\{-\frac{\|\boldsymbol{\phi}_{\tilde{y}}(p) - \boldsymbol{\phi}_{\tilde{y}}(l)\|^2}{2\sigma^2}\right\} \tag{18}$$

where $\boldsymbol{\phi}_{\tilde{y}}(p)$ is a vector of size N, given by

$$\boldsymbol{\phi}_{\tilde{y}}(p) = [\phi_{\tilde{y}}(p,0), \ldots, \phi_{\tilde{y}}(p, N-1)]^T \tag{19}$$

where $\phi_{\tilde{y}}(p,k)$ is the short-time PSD of $\tilde{y}_p(n)$ in time-frame p and the frequency bin k. In practice, we evaluate the short-time PSD by smoothing the periodograms $1/N\tilde{y}_{p,k}\tilde{y}^*_{p,k}$ over time frames p according to the common Welch method. This specific choice of kernel is motivated by the desire to exploit the reoccurring distinct spectral features of time frames containing a transient event, which may be appropriately conveyed by the power spectrum $\phi_{\tilde{y}}(p,k)$. In addition, the phase of the frames that contain transient events should be disregarded in the comparison, since it varies from frame to frame, and depends on the relative location of each event in the frame. Consequently, the phase of the STFT has little role in the estimation of the transient noise PSD derived in this step of the proposed algorithm. However, the phase is taken into consideration in the application of the NL filter, and in the next step, where the speech is estimated according to the OM-LSA gain function calculation.

As described in Section II, we view the STFT features of the time frames $\{\tilde{\mathbf{y}}_p\}$ as nodes of an undirected symmetric graph. Two nodes $\tilde{\mathbf{y}}_p$ and $\tilde{\mathbf{y}}_l$ are connected by an edge with weight $k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$, that corresponds to the affinity between $\tilde{\mathbf{y}}_p$ and $\tilde{\mathbf{y}}_l$. We continue with the construction of a random-walk on the graph nodes by normalizing the kernel $k(\cdot,\cdot)$, similarly to (1). We obtain a non-symmetric metric $p(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$ between two nodes, which represents the probability of transition in a single step from $\tilde{\mathbf{y}}_p$ to $\tilde{\mathbf{y}}_l$. Let $\{\tilde{\mathcal{Y}}_\tau\}$ be the Markovian process associated with this random-walk (where τ represents time index), and let $\mathbf{P}$ be an $M \times M$ matrix consisting of the transition probabilities. Similarly to (3), a single random-walk step is given by

$$\begin{aligned}\left[\mathbf{P}\tilde{\mathbf{Y}}\right]_p &= \sum_{l=1}^{M} \mathbf{P}_{pl}\tilde{\mathbf{y}}_l \\ &= \sum_{l=1}^{M} \Pr\left\{\tilde{\mathcal{Y}}_{\tau+1} = \tilde{\mathbf{y}}_l | \tilde{\mathcal{Y}}_\tau = \tilde{\mathbf{y}}_p\right\} \tilde{\mathbf{y}}_l \\ &= \mathbb{E}\left[\tilde{\mathcal{Y}}_{\tau+1} | \tilde{\mathcal{Y}}_\tau = \tilde{\mathbf{y}}_p\right].\end{aligned} \tag{20}$$

In (20), a single step is interpreted as averaging over similar time frame samples, where the sense of similarity is emerged from the kernel. Thus, the choice of the kernel $k(\cdot,\cdot)$ is of key importance in this method. We rely on the fact that time frames that

contain transient events consist of distinct spectral features compared to time frames free of transient events, as demonstrated in Section VI. Consequently, the kernel (18), which compares between the spectral features of time frames, implicitly leads to separation of frames into two classes. The first class, which was previously denoted by $\mathcal{H}_1$, consists of time frames that contain transient noise occurrences, which are similar to each other (in the kernel sense) since they have similar PSD features that characterize a transient event. The second class, denoted by $\mathcal{H}_0$, consists of time frames free of transient noise, which are similar to each other since they contain only the PSD of the white excitation signal. It is worthwhile noting, that in the latter case, we assume that the whitening using the long-term AR coefficients have captured both the white and pitch excitation characterizing unvoiced and voiced segments, respectively. Thus, the random-walk iteration approximately averages over all the frames from the same class, i.e.,

$$\left[\mathbf{P}\tilde{\mathbf{Y}}\right]_p = \sum_l \mathbf{P}_{pl}\tilde{\mathbf{y}}_l \approx \sum_{l\in H_i} \mathbf{P}_{pl}\tilde{\mathbf{y}}_l,\ p \in H_i. \tag{21}$$

As a result, the smeared transient events are averaged with similar events, whereas the zero-mean random excitation signal $w_{p,k}$ is averaged destructively, and therefore suppressed. Consequently, after a random-walk iteration, the smeared transient noise signal can be extracted from the whitened measurement. We note that unlike the kernel function (18), the application of the NL filter takes into account the phase of the signal. The length of a time frame is chosen to be similar to the lengths of transients. Time frames, which contain similar and aligned transients, are identified as similar frames according to (18), whereas time frames, which contain similar transients but unaligned, are identified as different. Hence, the constructive averaging in (21) is carried out only over time frames with aligned transients. In future work, we intend to include relative alignments before the averaging, that would enable constructive averaging also over time frames consisting of unaligned transients. Let $\hat{\tilde{d}}_{p,k}$ denote the estimate of the smeared transient noise at time frame p and frequency bin k after a single iteration. Let $\hat{\tilde{\mathbf{d}}}_p$ be a vector of length N consisting of the STFT features of the estimated transient noise in time frame p, $\hat{\tilde{\mathbf{d}}}_p = \left[\hat{\tilde{d}}_{p,0},\ldots,\hat{\tilde{d}}_{p,N-1}\right]$, which is given by

$$\hat{\tilde{\mathbf{d}}}_p = \left[\mathbf{P}\tilde{\mathbf{Y}}\right]_p^T. \tag{22}$$

The spectral decomposition of the transition probability function can be written as (using the notations from Section II)

$$\begin{aligned} p(\tilde{\mathbf{y}}_p,\tilde{\mathbf{y}}_l) &= \sum_{j=0}^{M} \lambda_j \boldsymbol{\psi}_j(p)\boldsymbol{\varphi}_j(l) \\ &= \boldsymbol{\varphi}_0(l) + \sum_{j=1}^{M} \lambda_j \boldsymbol{\psi}_j(p)\boldsymbol{\varphi}_j(l). \end{aligned} \tag{23}$$

In the last transition, we used the fact that $\lambda_0 = 1$ and $\boldsymbol{\psi}_0 = \mathbf{1}$, since according to the construction, the sum of each row of $\mathbf{P}$ is 1. Based on (23) we obtain that applying the random-walk step (20) can be expressed as[6]

$$\left[\mathbf{P}\tilde{\mathbf{Y}}\right]_{p,k} = \sum_{j=0}^{M-1} \lambda_j b_{j,k} \boldsymbol{\psi}_j(p) \tag{24}$$

where $b_{j,k}$ is the inner product between the left eigenvector $\boldsymbol{\varphi}_j$ and the whitened measurement at frequency bin k, given by

$$b_{j,k} = \langle \tilde{y}_{(\cdot,k)}, \boldsymbol{\varphi}_j \rangle = [\tilde{\mathbf{Y}}^T]_k \boldsymbol{\varphi}_j. \tag{25}$$

Consequently, t consecutive steps are written as

$$\left[\mathbf{P}^t\tilde{\mathbf{Y}}\right]_{p,k} = \sum_{j=0}^{M-1} \lambda_j^t b_{j,k} \boldsymbol{\psi}_j(p). \tag{26}$$

Using the properties of the eigenvalues of the transition matrix (4), we note that for infinite number of iterations, all the components in the sum (26), except the first, converge to zero, since $\lambda_j^t \xrightarrow{t\longrightarrow\infty} 0,\ \forall \lambda_j < 1$. Consequently, the resulting signal after an infinite number of iterations is "blurred" to a *trivial* steady state $\left[\mathbf{P}^t\tilde{\mathbf{Y}}\right]_{p,k} \xrightarrow{t\longrightarrow\infty} b_{0,k}$. Thus, we conclude that by increasing the number of random-walk steps we do not necessarily obtain a better result, but we might rather degenerate the signal. In order to properly estimate the transient noise signal, a finite number of iterations should be applied. On the one hand, the proper number of steps should be large enough to extract an accurate estimate of the transient noise. On the other hand, we should not use too many steps that would "smear" or "blur" the signal. Setting the proper number of iterations is of key importance and is addressed in Section V. It is worthwhile noting that in our experiments (described in Section VI) we find that the range of the proper number of iterations is between 10 and 200. In addition, we find that beyond 1000 iterations, significant distortions might emerge.

The spectral decomposition enables efficient implementation of such random-walk steps. First, computing a desired iteration does not involve taking powers of the transition matrix $\mathbf{P}$, but rather taking powers of the scalar eigenvalues. In addition, by assuming a fast eigenvalues decay, we may use merely few of the eigenvectors that correspond to the largest eigenvalues for the implementation of (26), and hence, exploit the dimensionality reduction property of this approach [24].

It is worthwhile noting that (20) implies that the transition matrix $\mathbf{P}$ simply constitutes an *NL diffusion filter* as described in [15]. In Section V, we present a statistical model of the PSD estimate of the decorrelated signal $\tilde{y}(n)$ and analyze the behavior and performance of the proposed NL filter based on diffusion interpretation. Using this interpretation we gain further insight into the algorithm. In particular, it enables to address the question of a proper choice of the algorithm parameters, and provides quality measures of the filter capability to extract the transient noise properly.

Finally, we perform inverse filtering using the source signal AR coefficients on the output of the NL filter

$$\hat{d}_{p,k} = \frac{1}{1-a_{p,k}} \hat{\tilde{d}}_{p,k} \tag{27}$$

[6] $[\mathbf{X}]_{p,k}$ returns the element at the pth row and kth column of the matrix $\mathbf{X}$.

where $\hat{d}_{p,k}$ is an estimate of the transient signal in the STFT domain. Since the kernel is based solely on spectral features, an estimate of the short-time PSD of the transient noise $\hat{\phi}_d(p,k)$ is calculated based on smoothing periodograms $1/N\hat{d}_{p,k}\hat{d}^*_{p,k}$ of the outcome signal of the NL filter.

We note that the transient noise PSD estimation presented in this section is an offline algorithm. The entire observable data is processed in two iterations. In the first iteration, the kernel is constructed, which requires the calculation of M^2 pairwise distances between time frames. In the second iteration, the NL filter is applied on each time frame by averaging over M time frames. Both iterations can be efficiently implemented. The kernel function can be calculated only for few nearest neighbors. Then, the corresponding NL averaging is performed only over these neighbors.

B. OM-LSA With a Modified Noise Spectrum Estimator

The optimally modified log spectral amplitude (OM-LSA) speech estimator [4] relies on the optimal spectral gain function, which is controlled by speech presence uncertainty. As proposed in [4], the speech presence probability is estimated based on the time–frequency distribution of the *a priori* signal-to-noise ratio (SNR), where the noise variance is estimated using the minima controlled recursive averaging (MCRA) [5]. Unfortunately, short bursts of transient noise occurrences are falsely detected as speech components. Hence, the transient noise is not estimated by the MCRA approach, and as a result, is not attenuated by the OM-LSA estimator.

In the proposed algorithm, we use an OM-LSA version equipped with a modified noise PSD estimation. From the output of the NL filter we obtain an estimate of the PSD of the transient noise signal $\hat{\phi}_d(p,k)$. We adjust the optimal spectral gain function calculation to rely on the following noise spectral estimation

$$\hat{\phi}_\xi(p,k) = \hat{\phi}_u(p,k) + \hat{\phi}_d(p,k) \tag{28}$$

where $\hat{\phi}_u(p,k)$ is the stationary noise PSD estimate obtained using the MCRA approach. Accordingly, the calculation of the optimal spectral gain function is controlled by both the stationary and transient noise parts, and thus, attenuation of transient occurrences is feasible. It is shown in [1] that the MMSE estimator of the phase of the desired speech signal is simply the phase of the measurement. Consequently, the calculation of the gain function requires estimate of the noise PSD without the phase, which is provided by the first stage of the proposed algorithm. Therefore, the phase of the transient noise signal is not taken into consideration in our work separately, but is processed as part of the noisy measurement. For more details regarding the optimal gain function derivation and estimation of the speech presence probability and the noise PSD, we refer the readers to [4]–[6] and references therein. The outcome of the algorithm is denoted by $\hat{s}(n)$ and $\hat{s}_{p,k}$, corresponding to the enhanced speech in the time and the STFT domains, respectively.

V. Probabilistic Analysis and Diffusion Interpretation

In the limit of a large number of samples (i.e., $M \longrightarrow \infty$) and small kernel scale (i.e., $\sigma \longrightarrow 0$ (18)), the discrete random-walk converges to the continuous diffusion process described by the backward Fokker–Planck operator [24], [25], [34]. As described in Section II, two graph nodes $\tilde{\mathbf{y}}_p$ and $\tilde{\mathbf{y}}_l$ effectively have nonzero affinity $k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$ if their distance is less than σ. Consequently, for each node $\tilde{\mathbf{y}}_p$, $p(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l) > 0$ only for nodes $\tilde{\mathbf{y}}_l$ within a ball of radius σ around $\tilde{\mathbf{y}}_p$. This means that the random-walker has nonzero transition probability from node $\tilde{\mathbf{y}}_p$ only to nodes within radius σ. Thus, in terms of a diffusion process in continuous time, we obtain that a single (discrete) random-walk step corresponds to the evolution defined by the Fokker–Planck operator (6) in a continuous time step of $\Delta\tau = \sigma$. The Fokker–Planck operator (6) implies that the propagation of the diffusion process depends on the distribution $q(\cdot)$ of the data set samples, which is conveyed by the potential $U(\cdot) = -2\ln q(\cdot)$. Moreover, the density of samples $q(\cdot)$ is going to evolve according to the Fokker–Planck equation. Thus, in our case, given the distribution of the PSD estimate of the whitened measurements, we may provide analysis for the behavior of the random-walk [15], [25], [29], [30] . In particular, in this section we evaluate the proper number of iterations that should be used to extract the transient noise signal. In addition, we estimate the probability of misidentifying a transient occurrence and choose the proper kernel scale σ. In Appendix II, we demonstrate the diffusion interpretation on a simple example from [15] of denoising a constant signal corrupted by additive white Gaussian noise.

A. Probabilistic Setup

From (14) and (15), we can write an estimate of the PSD of the whitened measurement as

$$\hat{\phi}_{\tilde{y}}(p,k) = \begin{cases} \sigma_w^2 + e_{p,k} & p \in \mathcal{H}_0 \\ [\Gamma(k) + \varepsilon_{p,k}]\,\phi_d(p,k) + \sigma_w^2 + e_{p,k} & p \in \mathcal{H}_1 \end{cases} \tag{29}$$

where $\phi_d(p,k)$ is the transient noise PSD, which according to (15), is given by[7]

$$\phi_d(p,k) = |h_k|^2 |v_p|^2 \tag{30}$$

$\Gamma(k)$ is the mean power of the AR spectral envelope

$$\Gamma(k) \triangleq \frac{1}{M}\sum_{p=1}^{M} |1 - a_{p,k}|^2 \tag{31}$$

and $\varepsilon_{p,k}$ expresses the diversity of the spectral envelope of time frame p with respect to the mean spectral envelope $\Gamma(k)$, satisfying

$$|1 - a_{p,k}|^2 = \Gamma(k) + \varepsilon_{p,k}. \tag{32}$$

In addition, $e_{p,k}$ is the PSD estimation error. Periodogram, which is used for estimating the PSD, is exponentially distributed. In our work, we improve the PSD estimate by averaging periodograms (as in the Welch method), which gives a single peak distribution, resulting from convolving exponential pdfs. For simplicity, we assume that the PSD estimation error is white and Gaussian $e_{p,k} \sim \mathcal{N}(0, \sigma^2_{e,k})$. It is worthwhile noting that $e_{p,k}$ may also express a model mismatch that can

[7]Notice that the transient amplitude v_p (without the phase) does not depend on the frequency bin.

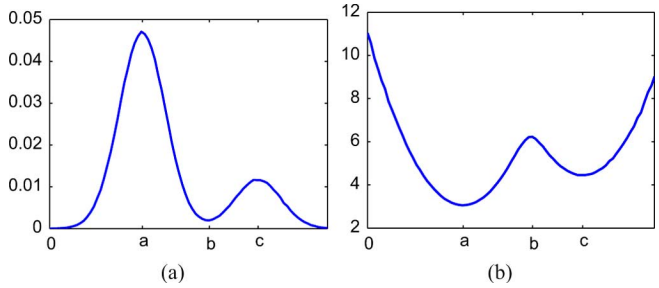

Fig. 1. Illustration of the distribution of the PSD of the decorrelated measurement (a) The probability density function. (b) The potential.

be derived from inaccurate estimation of the AR envelope coefficients.

According to our choice of the kernel (18), the propagation of the Markovian random-walk depends on the potential $U(\cdot) = -2\log q(\cdot)$ corresponding to the distribution $q(\cdot)$ of the PSD estimate $\hat{\phi}_{\tilde{y}}(p,k)$ of the whitened measurements. Consequently, we can view $\hat{\phi}_{\tilde{y}}(p,k)$ presented in (29) as a random variable, and analyze the random-walk propagation accordingly. Specifically, since the density of the random variable $\hat{\phi}_{\tilde{y}}(p,k)$ evolve according to the Fokker–Planck equation, we are able to track the processing enabled by the NL filter on the signal $\tilde{y}_{p,k}$.

B. Setting the Number of Random Walk Steps

First we examine a simple case, where there is no divergence between transient noise events, i.e., $\sigma_v^2 \longrightarrow 0$, and the AR process is stationary, $\varepsilon_{p,k} \longrightarrow 0$. From (29), we have that the PSD of the measurements has a two Gaussian mixture distribution $q(\cdot)$, both with variance σ_e^2, centered at $a \triangleq \sigma_w^2$ and $c \triangleq \sigma_w^2 + \Gamma(k)|h_k|^2\mu_v^2$, respectively,[8] creating a "two wells" potential $U(\cdot)$, as illustrated in Fig. 1. The left well, centered at a, corresponds to time-frames from the class $\mathcal{H}_0$, whereas the right well, centered at c, corresponds to time-frames from the class $\mathcal{H}_1$. Accordingly, our aim in the first stage of the algorithm, i.e., averaging time frames from each class separately, can be interpreted as to bring values from each well to its minimum [or mean according to (29)]. It is worthwhile noting that Fig. 1 illustrates the "two wells" potential corresponds to the two Gaussian mixture distribution of the simplest case. However, as we describe later in this section, the "two wells" shape characterizes the potential of the distribution in the general case, where the number of wells corresponds to the number of hypotheses.

The analysis of the continuous diffusion process in two-wells potential is well studied in the literature, mainly for physical and chemical systems [27], [28] . In particular, two characteristic times enable to analyze the diffusion process [15]. The first is the relaxation (equilibration) time τ_R for each well. It implies that in order to properly bring all samples in a certain well to their mean, we need to propagate the diffusion process for τ_R. Thus, we need to apply at least $t_R = \tau_R/\Delta\tau = \tau_R/\sigma$ random-walk steps (using the fact that each discrete random-walk step corresponds to $\Delta\tau = \sigma$ propagation time). It can be shown that the relaxation time of each well depends on the curvature of the bottom of the potential well. In this simplest case, assuming the two Gaussians are well separated, the relaxation time of both left and right wells can be approximated by the curvature of each Gaussian independently. In one dimension, the curvature is the absolute value of the second derivative. Specifically, for the potential of a Gaussian it is given by

$$\begin{aligned}\tau_R &= \frac{1}{|U''(a)|} = \frac{1}{|U''(c)|} \\ &\cong \left| \frac{\partial^2}{\partial x^2} \left\{ -2\ln\left(\frac{1}{\sqrt{2\pi\sigma_{e,k}^2}} \exp\left\{ -\frac{(x-a)^2}{2\sigma_{e,k}^2} \right\} \right) \right\} \right|_{x=a}^{-1} \\ &= \frac{\sigma_{e,k}^2}{2} \end{aligned} \tag{33}$$

where U is the potential associated with the two Gaussian mixture distribution of $\hat{\phi}_{\tilde{y}}(p,k)$, and U'' is the second derivative of U. However, bringing all the samples to their mean can be obtained only if the samples do not exit their well. Consequently, the second characteristic time is the mean first passage time (MFPT) τ_{exit} to exit a well. Alternatively, it can be described as the time it takes for a particle to surmount the potential barrier on its way to the lowest well. Matkowsky and Schuss [28] showed that the MFPT is exponentially increasing with the height of the barrier between the wells, i.e.,

$$\begin{aligned}\tau_{\text{exit}}\left(a \longrightarrow c\right) &= \frac{2\pi}{\sqrt{U''(a)U''(b)}} \exp\left\{|U(a) - U(b)|\right\} \\ \tau_{\text{exit}}\left(c \longrightarrow a\right) &= \frac{2\pi}{\sqrt{U''(c)U''(b)}} \exp\left\{|U(c) - U(b)|\right\}. \end{aligned} \tag{34}$$

where b is the location of the barrier between the wells as illustrated in Fig. 1. In addition, they showed that the MFPT is closely related to the convergence rate of the diffusion process to the steady state, which is determined by the first nontrivial eigenvalue λ_1 of the transition matrix as implied in (26). Accordingly, to properly bring the samples to their mean, we should not apply more than $t_{\text{exit}} = \tau_{\text{exit}}/\sigma$ random-walk iterations, which can be approximated by using λ_1.

Thus, in order to be able to obtain a good extraction of the transient noise signal, the two potential wells of the PSD of the whitened measurement should be well separated to distinguish the two classes, indicating the presence/absense of a transient occurrence. In particular, the two characteristic times of interest of the potential should satisfy $\tau_{\text{exit}} \gg \tau_R$, and the proper number of iterations t should be

$$t_R < t \ll t_{\text{exit}}. \tag{35}$$

For the simple case, (35) can be written explicitly using (33) and (34) as

$$\frac{\sigma_{e,k}^2}{\sigma} < t \ll \min\left(\tau_{\text{exit}}\left(a \longrightarrow c\right), \tau_{\text{exit}}\left(c \longrightarrow a\right)\right). \tag{36}$$

As the transient noise occurrences become more diverse, i.e., σ_v^2 increases, the Gaussian distribution is smeared. The PSD distribution in this case is a convolution (due to summation of two independent random variables) between Gaussian and χ^2 pdfs. The χ^2 probability of a single degree of freedom corresponds to the random variable $\phi_d(p,k)$ presented in (30) as square of a Gaussian random variable with mean μ_v and variance σ_v^2.

[8] μ_v and σ_v are the mean and variance of the Gaussian random amplitude $v(n)$ of the transient signal (10).

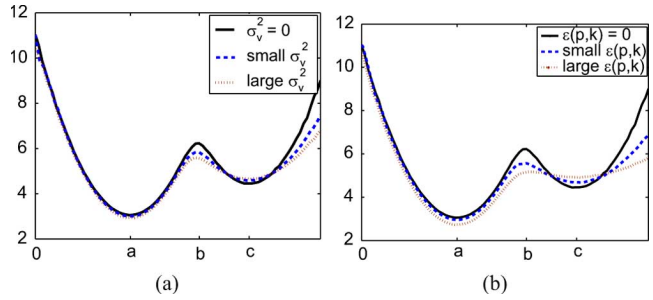

Fig. 2. Illustration of the potential of the PSD of the decorrelated measurement. (a) In case of diverse transient occurrences. (b) In case of nonstationary AR source.

Fig. 2(a) shows the potential corresponding to three values of σ_v^2: zero (solid line), arbitrary small value (dashed line), and a ten times larger value (dotted line). As illustrated in Fig. 2(a), the right well becomes wider and shallower and the barrier between the wells is lower. According to (33) and (34), it results in a longer relaxation time and a shorter MFPT to exit a well. Consequently, it is more difficult to distinguish the presence of a transient occurrence. In addition, since the number of iterations should satisfy (35), more iterations should be applied; however, the maximum number of iterations is more restricted. In addition, the extracted transient noise signal obtained by averaging over all transient events, which are more diverse, varies from each individual occurrence. It implies that degraded extraction is achieved at the cost of a larger computational effort. In Fig. 2(b), similar trends can be observed as the AR process becomes nonstationary and more diverse. We assume that $\varepsilon_{p,k}$ has a Gaussian distribution, and compare in Fig. 2(b) three values of $\varepsilon_{p,k}$: zero (solid line), low value (dashed line), and a ten times larger value (dotted line). According to (29), the distribution of frames in $\mathcal{H}_1$ corresponds to a sum of two independent Gaussian random variables. Consequently, we observe that as $\varepsilon_{p,k}$ increases, the right well becomes wider and shallower, which increases the relaxation time and decreases the MFPT to exit this well.

C. Identification Probability of Transient Events

We exploit the propagation of the diffusion process in a two wells potential to evaluate the probability of misidentifying a transient occurrence. It implies that a misidentification occurs in case the PSD estimate $\hat{\phi}_{\tilde{y}}(p,k)$ of a frame in $p \in \mathcal{H}_1$, falls in the wrong left well (due to large PSD estimation error $e_{p,k}$). As $\hat{\phi}_{\tilde{y}}(p,k)$ is presented in our analysis as a random variable (29), we are able to calculate this probability. Specifically, in the simplest case ($\sigma_v^2 \longrightarrow 0$ and $\varepsilon_{p,k} \longrightarrow 0$), the probability of misidentifying a transient occurrence can be explicitly written as

$$\Pr\left\{\hat{\phi}_{\tilde{y}}(p,k) < b|\mathcal{H}_1\right\} = \Phi\left(\frac{b-c}{\sigma_e}\right) \tag{37}$$

where $\Phi(x)$ is the standard Gaussian cumulative distribution function. Similarly, the probability of falsely identifying a transient occurrence, which occurs when the PSD estimate $\hat{\phi}_{\tilde{y}}(p,k)$ of a frame in $p \in \mathcal{H}_0$ is in the right well (again, due to a large PSD estimation error), is given by

$$\Pr\left\{\hat{\phi}_{\tilde{y}}(p,k) > b|\mathcal{H}_0\right\} = 1 - \Phi\left(\frac{b-a}{\sigma_e}\right). \tag{38}$$

We observe that the misidentification probability mainly depends on the distance between the potential wells minima.

It is worthwhile noting that these probabilities are closely related to the analysis of spectral clustering limitations. The diffusion interpretation implies that the question of whether transient occurrences can be distinguished is analogous to the question of whether spectral clustering can be employed. In this problem setup, in case the condition $\tau_{\text{exit}} \gg \tau_R$ is satisfied, it indicates that time frames free of transient events can be distinguished from time frames that contain transient events using spectral clustering methods. For more details see [29] and [30], where the authors discuss the limitation of spectral clustering extensively using similar diffusion interpretation. Based on this analogy, we can conclude that spectral clustering algorithms [35]–[38] relying on the proposed diffusion operator may enable identification of the locations in time of transient events. In particular, Shi and Malik [36] proposed to use the first nontrivial eigenvector of the diffusion operator $\mathbf{P}$ as an indicator for the clusters. They showed that calculating the first nontrivial eigenvector is equivalent to finding the minimum normalized cut of the graph G we constructed, whose nodes are the data set samples and the edges weights are determined by the affinity kernel.

D. Setting the Kernel Scale

The convergence to the continuous diffusion operator is further utilized for properly choosing the kernel scale σ. It can be shown [39], [40] that the convergence rate of the random-walk to the continuous diffusion process depends on a balance between a bias term and a variance term. The bias term is associated with discretization of the diffusion in time, and hence, calls for small σ (corresponds to the propagation time of a single random-walk step), which results in small random-walk steps. The variance term is associated with discretization of the diffusion in space, and hence, calls for large σ which results in increasing the number of neighbors for each node, and hence integrating over a larger number of samples. In [39] and [41], it was proposed to automatically set the scale by examining a logarithmic scale of the sum of the kernel weights, without computing the spectral decomposition of the transition matrix. The sum of the kernel matrix elements can be approximated by an integral. In particular, for a *proper* scale, the samples are assumed to lie on a manifold $\mathcal{M}$, and this integral is approximated by [39]

$$\sum_{p,l} k(\mathbf{y}_p, \mathbf{y}_l) \approx \frac{M^2}{vol(\mathcal{M})}(2\pi\sigma)^{d/2} \tag{39}$$

where $vol(\mathcal{M})$ is the volume of the manifold, and d is the manifold dimensionality. In a logarithmic scale, we have

$$\log\left(\sum_{p,l} k(\mathbf{y}_p, \mathbf{y}_l)\right) \approx \frac{d}{2}\log(\sigma) + \log\left(\frac{M^2(2\pi)^{d/2}}{vol(\mathcal{M})}\right) \tag{40}$$

which implies that the slope of the logarithmic scale of the sum of the kernel weights as a function of σ is $d/2$. However, in the limit $\sigma \longrightarrow \infty$, we have $k(\mathbf{y}_p, \mathbf{y}_l) \longrightarrow 1$ and $\sum_{p,l} k(\mathbf{y}_p, \mathbf{y}_l) \longrightarrow M^2$. On the other hand, for $\sigma \longrightarrow 0$, we have $k(\mathbf{y}_p, \mathbf{y}_l) \longrightarrow \delta_{p,l}$, and $\sum_{p,l} k(\mathbf{y}_p, \mathbf{y}_l) \longrightarrow M$. These two limits

suggest that the logarithmic plot cannot be linear for all σ. In the linear region, both the bias and the variance errors are relatively small, and therefore σ may be chosen from that region.

E. High-Dimensional Processing

Based on the probabilistic analysis, we obtain additional explanation for the usefulness of comparing a few frequency bins collected into a single vector for each time frame. In Section IV, we stated that by comparing the spectrum of all frequency bins of each time frame (18) rather than the individual sub-bands, we exploit the unique spectral structure of each time frame consisting of a transient occurrence. In terms of diffusion process in a two *high-dimensional* potential wells, the distance between the potential wells minima is increased and the barrier between the wells becomes higher. As a result, the probability of misidentification (37) decreases as the distance between c and b increases. Similarly, the probability of false alarm (38) decreases since the distance between a and b increases. Moreover, according to (34), the MFPT to exit a well τ_{exit} increases, and therefore, additional random-walk steps satisfying (35) can be employed, yielding a more accurate transient noise signal extraction.

VI. Experimental Results

In this section, we evaluate the performance of the proposed method. First, we examine the results on synthetic signals and explore the trends according to our analysis. Second, we test the algorithm on real recorded signals, and compare the results of the proposed algorithm with the results of the OM-LSA estimator [4].

In the first experiment, we generate signals according to the time domain model. The sampling frequency is set to 16 kHz. The source signal is simulated as a stationary fourth order autoregressive source according to (9) with white Gaussian excitation signal $w(n)$ of unit variance $\sigma_w = 1$. The transient noise is generated according to (10). The fixed filter $h(n)$ has distinct spectral features containing three harmonies at frequencies 1600, 3200, and 6400 Hz, as shown in Fig. 3(a), and the transient occurrences are determined according to a Gaussian–Poisson distribution. The mean and variance of the random amplitude $v(n)$ are $\mu_v = 2$ and $\sigma_v = 0.003$, and the Poisson rate is 0.0005. The parameters for the simulated scenario are chosen such that the potential-wells are well separated, which according to (36) enables estimation of the transient noise PSD using diffusion filtering. Empirical testing showed that simulation parameters that did not satisfy (36), indeed resulted in poor performance.

For illustration, we plot in Fig. 3 the measurement along with the first (nontrivial) eigenvector. To correspond with the measurement, we show the magnitude of the eigenvector as a function of time (the eigenvector consists of magnitudes per time frame). We clearly observe that the eigenvector provides an indicator for transient noise occurrences. First, it implies according to the analogy between the transient noise reduction problem and spectral clustering, that proper denoising is feasible. Second, as a by-product of the proposed algorithm, we obtain an identifier for transient events. It is worthwhile noting

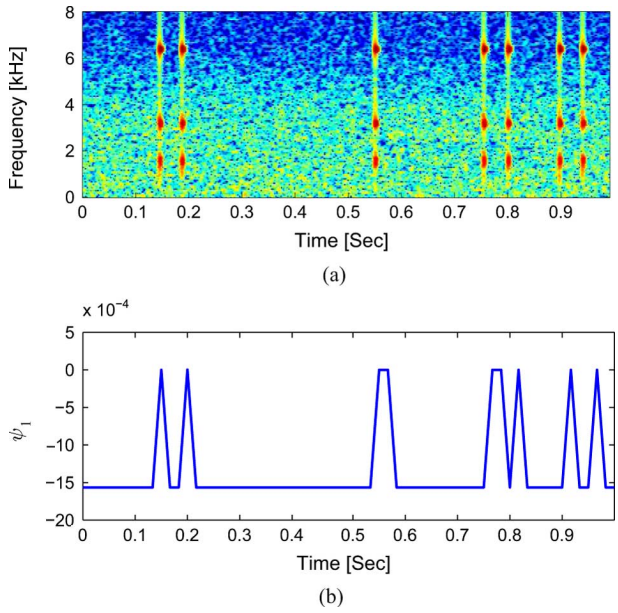

Fig. 3. Synthetic signals experiment. (a) Noisy measurement. (b) The first (nontrivial) eigenvector $\boldsymbol{\psi}_1$.

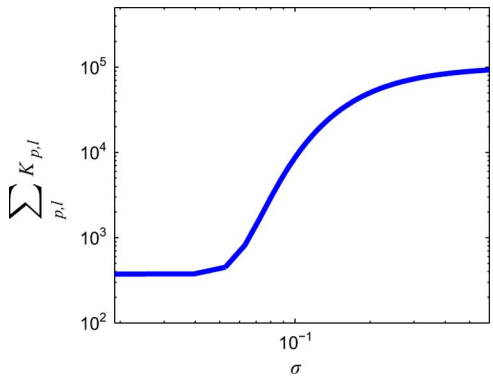

Fig. 4. Logarithmic scale plot of the sum of the kernel weights $\sum_{p,l} k(\mathbf{y}_p, \mathbf{y}_l)$ versus the kernel scale σ.

that our proposed method aims at a more difficult outcome—extracting the transient noise signal from the measurement, rather than just indicating transient occurrences.

Fig. 4 illustrates the proposed automatic choice of kernel scale. We plot a logarithmic scale of the sum of the kernel weights as a function of the scale. We observe that the curve is nonlinear, and we choose the scale σ for this experiment from the linear region (i.e., $\sigma = 10^{-1}$). According to our empirical experiments, choosing the scale from this region indeed yields good results.

We evaluate the transient signal estimation obtained using the NL diffusion filter. For measuring the performance we use the transient to signal ratio (TSR) defined by

$$\begin{aligned} \text{TSR}_{\text{in}} &= 10\log_{10}\frac{\mathbb{E}\left\{d^2(n)\right\}}{\mathbb{E}\left\{(y(n)-d(n))^2\right\}}; \\ \text{TSR}_{\text{out}} &= 10\log_{10}\frac{\mathbb{E}\left\{d^2(n)\right\}}{\mathbb{E}\left\{(\hat{d}(n)-d(n))^2\right\}} \end{aligned} \tag{41}$$

where $\hat{d}(n)$ is obtained by applying inverse STFT (ISTFT) on the estimated transient noise $\hat{d}_{p,k}$.

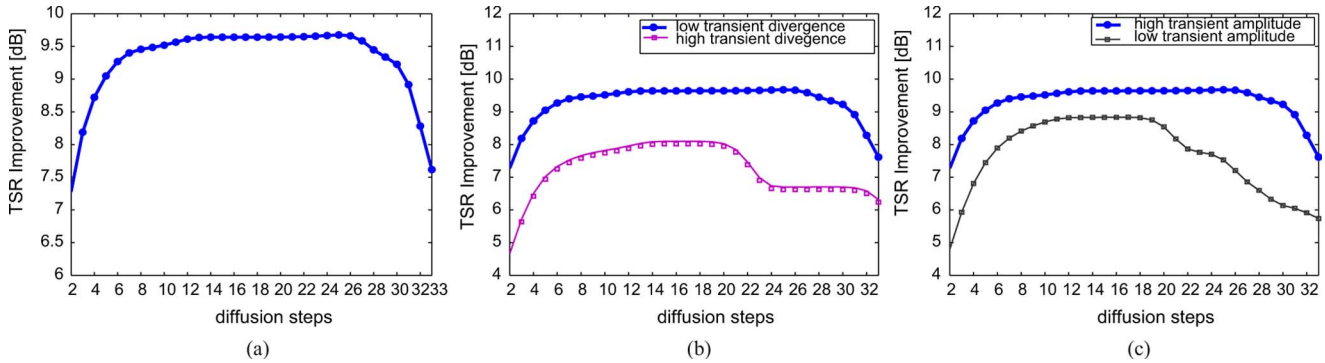

Fig. 5. (a) TSR improvement (in dB) obtained as a function of the number of denoising steps. (b) Comparison of the obtained TSR improvement. (in dB) between low and high transient occurrence divergence. (c) Comparison of the TSR improvement (in dB) between two different mean amplitude values.

Fig. 5(a) shows the TSR improvement obtained as a function of the number of denoising steps. We note that in this experiment we use powers of the transition matrix (the NL filter) $\mathbf{P}$, indicating several random-walk steps in each diffusion iteration. We clearly observe the tradeoff in setting the proper number of steps that emerges from the results. Initially, we obtain an increase in the TSR as we employ more steps. Then, after reaching a certain number of steps (greater than the relaxation time), the TSR remains constant and applying more denoising steps does not improve the results. Finally, when the number of steps reaches the MFPT to exit a well, applying additional denoising steps smears the signal, and the TSR decreases.

In Fig. 5(b), we compare the TSR improvement between low and high transient occurrence divergence. For the low divergence case, we set the variance of the amplitude modulation σ_v^2 in the simulation to be a small value ($\sigma_v = 0.003$), and for the high divergence case, we set σ_v^2 in the simulation to be twice higher. First, the TSR obtained when the transient occurrences divergence is high, is smaller than the TSR obtained when the divergence is low. Since transient occurrences are less similar, the averaging obtained by the diffusion filter is less accurate since the resulting mean value varies from each transient occurrence. Second, we observe that in the high diversity case, we need to apply more steps in order to reach optimal TSR values. However there is a sharp decline starting from smaller number of iterations. By increasing the transient noise amplitude variance, the potential well becomes wider, and the barrier lower. Consequently, the relaxation time of the right-hand well (corresponding to the transient occurrences) increases, and more diffusion steps should be used. In addition, the MFPT decreases, implying that less diffusion steps can be used.

Fig. 5(c) shows the obtained TSR improvement for two different mean amplitude values $\mu_v = 1.5, 2$ (with the same variance σ_v). We observe that the TSR obtained when the mean amplitude is small, is lower than the TSR obtained when the mean amplitude is large. In addition, a decay in the TSR occurs after a smaller number of iterations, in the case of small mean amplitude. By decreasing the mean transient noise amplitude μ_v, the potential wells become closer. As a consequence, the MFPT decreases, and less diffusion steps can be used. In addition, since the separation of the two potential wells is worse, the probability for misidentification increases.

In the second experiment, we use recorded speech and transient noise signals. Speech signals sampled at 16 kHz are taken from TIMIT database [42]. Various recorded transient noises are taken from [43]. The measurements are constructed according to (7) and (8). The additive stationary noise part is a computer generated white Gaussian noise with an SNR of 20 dB. The length of the speech utterance and the recorded transient noise is 10 s. Such transient noise signal consists of 10 to 12 transient events. We use short-time frames of 256 samples length both for the LPC estimation and for the STFT. The corresponding time frame length is 16 ms, which is longer than the duration of the tested transients (approximately 10 ms). In each time frame, we estimate AR envelope consisting of $L = 50$ coefficients, in order to obtain a white excitation signal for both voiced and unvoiced phonemes (we verify that the pitch period is of shorter length). According to our empirical tests, such a relatively short envelope enabled sufficient whitening of the speech. Fig. 6 shows spectrograms of the noisy measurement with metronome noise and the whitened signal $\tilde{y}_p(n)$. We observe that the impulsive nature of transient events is maintained, while the speech phonemes are whitened. Specifically, we notice that $L = 50$ AR coefficients provide satisfactory whitening of both voiced and unvoiced phonemes to enable better distinction of the transient events from the speech components.

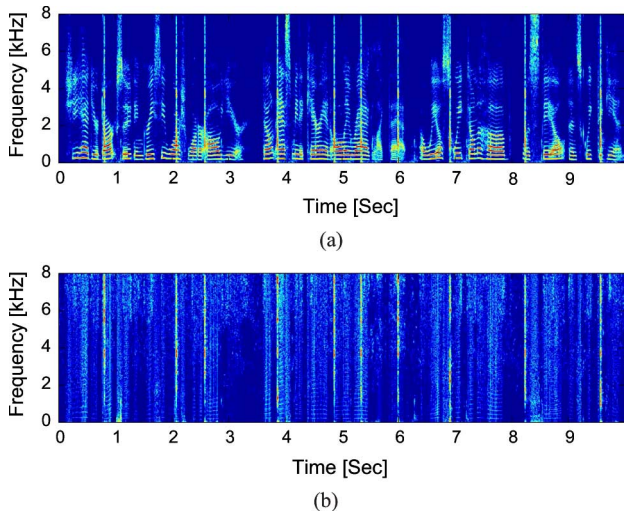

Fig. 6. Signal spectrograms. (a) The noisy signal. (b) The whitened. signal $\tilde{y}_p(n)$.

The transient noise is extracted by the diffusion filter using $t = 128$ iterations. This specific number of iterations was

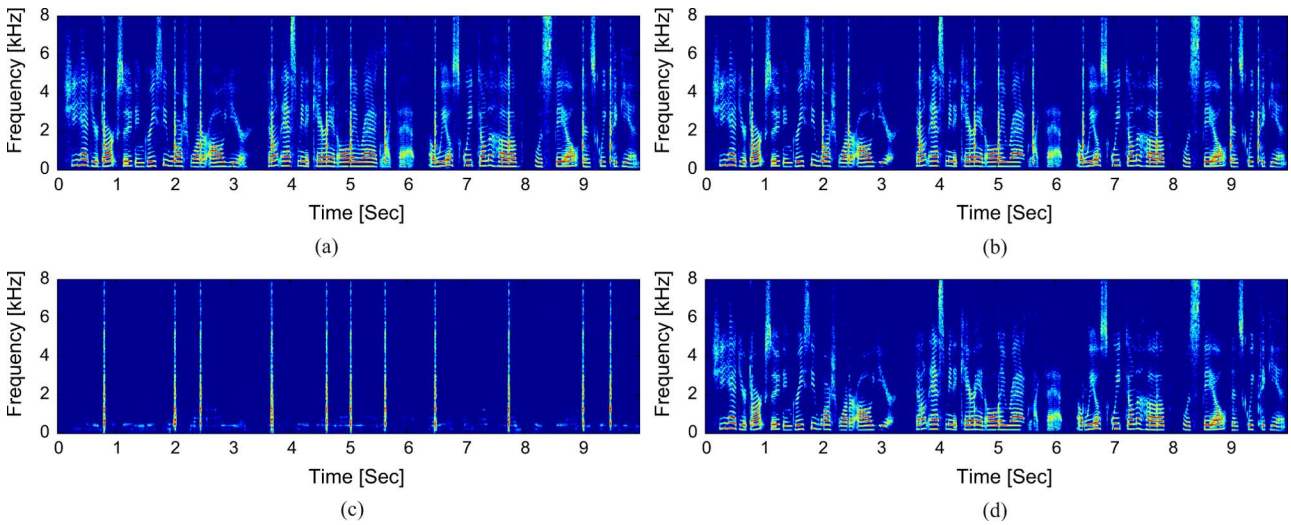

Fig. 7. Signal spectrograms. (a) The noisy signal. (b) The enhanced signal obtained by the OM-LSA. (c) The transient estimate obtained by the NL diffusion filter. (d) The enhanced signal obtained by the proposed method.

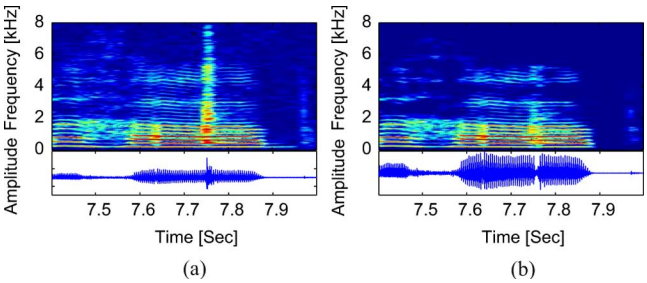

Fig. 8. Signal spectrograms and waveforms in the area near the transient event at 7.7 s. (a) The noisy signal. (b) The enhanced signal obtained by the proposed method.

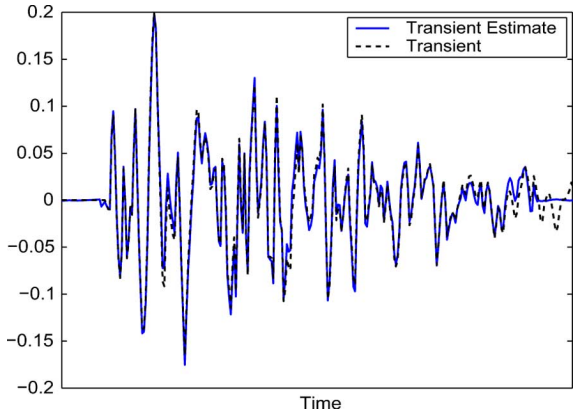

Fig. 9. Estimation of a single metronome "clack."

chosen both according to the analysis from Section V along with empirical testing.

Fig. 7 shows spectrograms of the noisy speech signal corrupted by metronome noise, denoised signal using the OM-LSA, transient PSD estimation from the output of the diffusion filter, and denoised signal using the proposed method. Audio samples of these signals are available online in [44]. The proposed algorithm does not require periodic occurrences of the transient noise signal. Thus, we artificially change the gaps between each metronome "clack" to demonstrate the full potential of the proposed algorithm. We see in Fig. 7(d) that the proposed estimator achieves both stationary and transient noise reduction, while imposing very low distortion. The satisfactory transient noise reduction is enabled due to the accurate noise PSD estimation, as shown in Fig. 7(c), where we obtain a relatively clean estimation of the PSD of the metronome signal. In addition, we observe that the spectral shape of the metronome "clacks" are estimated accurately. Fig. 8 zooms into the area near the transient event at 7.7 s, and further illustrates the removal of the transient component and the preservation of the speech. It is worthwhile noting that a small speech distortion was also detected in informal subjective listening tests. To demonstrate the accuracy of the estimation of the transient signal, we present in Fig. 9 a single metronome "clack" compared to its estimate taken from the output of the diffusion filter.

We evaluate the transient noise estimation using two objective measures. The first is the TSR improvement, defined in (41). The second is the mean spectral distance (SD) defined as

$$\mathrm{SD}_{\mathrm{in}} \triangleq \frac{1}{M}\sum_{p=1}^{M}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left(|d_{p,k}|-|y_{p,k}|\right)^2\right]^{1/2}$$

$$\mathrm{SD}_{\mathrm{out}} \triangleq \frac{1}{M}\sum_{p=1}^{M}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left(|d_{p,k}|-|\hat{d}_{p,k}|\right)^2\right]^{1/2}.$$

The TSR measure provides evaluation of the estimation in terms of power, whereas the SD provides evaluation of the estimation accuracy of the spectral features.

In Table I we examine the estimation of various transient noise types. We compare the TSR improvement and the SD improvement, obtained by the proposed method and the OM-LSA in various transient noise environments. The presented results are averaged over ten realizations of noise and different speech signals (both male and female). We obtain significant improvement in both measures, indicating good estimation of the transient noise. Keyboard typing noise PSD estimation is of particular interest. We obtain marginal improvement in SD, which implies less accurate extraction of the spectral features of the

TABLE I
EVALUATION OF THE TRANSIENT SIGNAL ESTIMATION

Input	TSR Improvement [dB]	Transient SD Improvement [dB]
Metronome	14.2	0.93
Door Knocks	10.7	0.79
Kitchen Pocks	6.3	0.69
Keyboard Typing	5.7	0.56

TABLE II
ENHANCEMENT EVALUATION IN TRANSIENT OCCURRENCE PERIODS

Input	SNR Improvement [dB]		LSD Improvement [dB]	
	OM-LSA	Proposed	OM-LSA	Proposed
Metronome	0.03	9.58	0.07	6.05
Door Knocks	0.23	6.85	0.14	4.23
Kitchen Pocks	0.39	6.91	0.35	2.34
Keyboard Typing	0.68	3.45	1.22	2.27

transient noise. Due to the high divergence between transient occurrences belonging to different key strokes, both in shape and power, the averaging process results in a "mean tapping shape" which varies from the spectral shape of each individual key press. Nevertheless, we observe good improvement in TSR, which indicates good estimation in terms of noise power. Since the potential-well associated with the excitation signal is well separated, the transient signal is properly extracted by the NL filter, which enables accurate identification of transient events time locations, yielding significant reduction in the total noise power at the output of the proposed algorithm. This particular noise type demonstrates the robustness of the proposed algorithm. Even though the transient interference caused by keyboard typing do not correspond to our assumptions, the proposed algorithm still enables improved result.

We evaluate the output of the proposed method using another two objective measures. The first is the common signal to noise ratio, defined as

$$\mathrm{SNR}_{\mathrm{in}} = 10\log_{10}\frac{\mathbb{E}\left\{s^2(n)\right\}}{\mathbb{E}\left\{(y(n)-s(n))^2\right\}};$$
$$\mathrm{SNR}_{\mathrm{out}} = 10\log_{10}\frac{\mathbb{E}\left\{s^2(n)\right\}}{\mathbb{E}\left\{(\hat{s}(n)-s(n))^2\right\}}. \tag{42}$$

The second is the mean log spectral distance (LSD) between the measured signal and the desired source, which is specifically adapted to speech signals and defined as

$$\mathrm{LSD}_{\mathrm{in}} \triangleq \frac{1}{M}\sum_{p=1}^{M}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(s_{p,k})-\ell(y_{p,k})\right|^2\right]^{1/2}$$
$$\mathrm{LSD}_{\mathrm{out}} \triangleq \frac{1}{M}\sum_{p=1}^{M}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(s_{p,k})-\ell(\hat{s}_{p,k})\right|^2\right]^{1/2}$$

where

$$\ell\left(f(t)\right) = \max\left\{10\log_{10}\left|f(t)\right|, \delta\right\}$$

and δ is a small value defined by $\delta = \max_t\left|f(t)\right| - 50$, used to confine the dynamic range of the log-spectrum to 50 dB.

To test the estimation of the speech in the presence of the transient noise events, we present in Table II the results obtained only in time frames in $\mathcal{H}_1$ (instead of the whole observation interval). We compare the speech enhancement results obtained using the proposed method and the OM-LSA estimator. We present the two objective measures (SNR improvement and LSD improvement) in dB. We clearly observe that the proposed method achieves better results compared to the OM-LSA in all noise types. It is worthwhile noting that similar results were obtained for other transient noise types taken from [43]: roof hammering, door slams, household clacks, and other percussive noises. The results emphasize the advantage of the proposed algorithm in obtaining good transient noise reduction, while preserving speech components, even under the adverse conditions created by the presence of transient noise events.

VII. CONCLUSION

We have presented a novel approach for handling speech corrupted with transient interferences. The proposed approach is based on the NL diffusion filter, that exploits the intrinsic geometric structure of the measurements. In particular, it relies on the variation of speech components and sharp impulses of repeating transient noise occurrences. By using diffusion interpretation of the NL filters, we gained insight into the behavior of the proposed method. Using the diffusion framework, we addressed the problem of proper choice of parameters and evaluated the performance and limitations of the proposed method. Experimental results have demonstrated that for repetitive and short transient occurrences, the proposed method obtains improved results, compared to those obtained by the OM-LSA estimator. In addition, the proposed method is robust to various types of transient interferences.

The main component of the proposed algorithm is the estimation of the transient noise PSD using NL diffusion filtering. Here, we have incorporated the PSD estimate into the OM-LSA estimator for speech enhancement. However, the PSD estimate may be exploited for other tasks as well. For example, it can be of major importance when developing a voice activity detector (VAD), adapted to transient noise environments. Future work will address real-time implementation of the algorithm, and developing a model for the spectral variations and durations of the transient events. For example, we aim at developing a more robust algorithm based on two iterations of the NL filter. The first iteration will provide just an estimate of the locations of the transients. In the second iteration, given the transients locations, each transient amplitude, shape and duration will be estimated and handled.

APPENDIX I
GAUSSIAN NOISE EXAMPLE

We present a simple example of denoising a step function corrupted by Gaussian noise using NL filters. Let $\Gamma = \{x_i\}_{i=1}^{M}$ be a data set consisting of M real samples $x_i \in \mathbb{R}$. Each data sample, which consists of a desired constant corrupted by additive white Gaussian noise, is given by

$$x_i = d_i + n_i \tag{43}$$

where d_i are in $\{-1, 0, 1\}$ with equal probability and n_i are independent and identically distributed Gaussian random variables with zero mean and $\sigma_n^2 = 0.2$ variance. For example, x_i

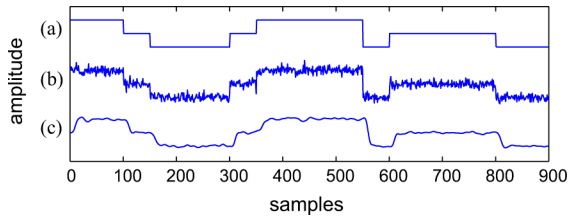

Fig. 10. (a) Source signal. (b) The noisy measurement. (c) The denoised signal using low-pass filter.

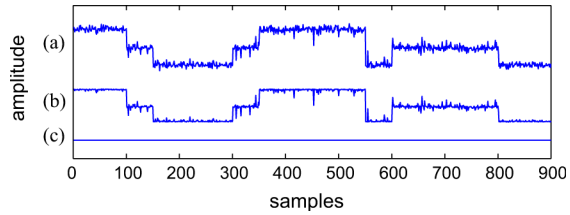

Fig. 11. Denoised signals using 1-D NL filters. (a) The denoised signal after a single step ($t = 1$). (b) The denoised signal after $t = 10$ steps. (c) The denoised signal after $t = 3000$ steps.

may be seen as a time series with time index i, consisting of measurements of a noisy telegraph signal. Fig. 10(a) and (b) shows the source signal and the noisy measurement. In Fig. 10(c), we present a denoised signal using low-pass filter. We use a finite impulse response (FIR) filter of length 20 with cutoff frequency of 0.1π rad to maintain the low frequencies of the source step function. We observe that the noise is suppressed; however, significant distortions are introduced, especially in the source function edges. It is worthwhile noting that other common denoising algorithms would enable similar trends. For example, wavelet denoiser might improve the performance of the low-pass filter since it provides multiscale resolution of the signal. However, the distortion of the edges, which occurs due to the processing of samples from two levels of the step function together, remains. In the remainder of this section, we demonstrate how a nonlocal filter solves this artifact.

We define a Gaussian kernel $k : \Gamma \times \Gamma \rightarrow \mathbb{R}$ as $k(x_i, x_j) = \exp\{-(x_i - x_j)^2/2\sigma^2\}$ which conveys a notion of pairwise affinity between the samples. As described in Section II, we construct a weighted graph G based on the data samples and the kernel in three steps. 1) We set the data samples $\Gamma = \{x_i\}$ to be the graph nodes. 2) The weights of the edges connecting the nodes are set according to the kernel, i.e., the edge connecting x_i and x_j is of weight $k(x_i, x_j)$. 3) By normalizing the kernel according to (1), we create a nonsymmetric affinity metric $p(x_i, x_j)$. This metric can be viewed as a transition probability function of a Markov chain on the graph, i.e., $p(x_i, x_j)$ represents the probability of transition in a single step of the random-walk from node x_i to node x_j. Let $\mathbf{P}$ be a matrix corresponding to the function p, where its (i, j)th element is $p(x_i, x_j)$, and let $\mathbf{x}$ be a vector consisting of all the data samples $\mathbf{x} = [x_1, \ldots, x_M]^T$. Accordingly, advancing the random-walk on the graph a single step forward can be written as $\mathbf{Px}$. Using the eigendecomposition of the matrix $\mathbf{P}$,

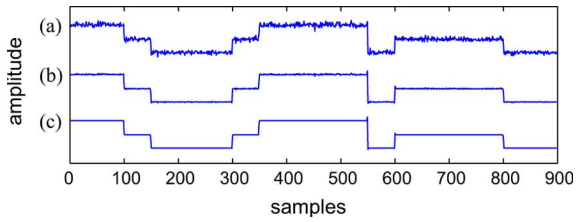

Fig. 12. Denoised signals using 3-D NL filters. (a) The denoised signal after a single step ($t = 1$). (b) The denoised signal after $t = 3$ steps. (c) The denoised signal after $t = 10$ steps.

described in Section II, we can present the expansion of the samples on the eigenbasis as[9]

$$\mathbf{x} = \sum_{j=0}^{M-1} b_j \boldsymbol{\psi}_j \tag{44}$$

where $\boldsymbol{\psi}_j$ are the matrix $\mathbf{P}$ right eigenvectors, and b_j are given by the inner product between the left eigenvectors $\boldsymbol{\varphi}_j$ and the samples $\mathbf{x}$, i.e., $b_j = \boldsymbol{\varphi}_j^T \mathbf{x}$. Applying the random-walk (i.e., the matrix $\mathbf{P}$) on the data set results in

$$[\mathbf{Px}]_i = \sum_{j=0}^{M-1} \lambda_j b_j \boldsymbol{\psi}_j(i) \tag{45}$$

where λ_j are the matrix $\mathbf{P}$ eigenvalues satisfying (4). Now, applying t random-walk steps is given by

$$\left[\mathbf{P}^t \mathbf{x}\right]_i = \sum_{j=1}^{M} \lambda_j^t b_j \boldsymbol{\psi}_j(i). \tag{46}$$

Fig. 11(a)–(c) shows the denoised signal after $t = 1$, $t = 100$, and $t = 3000$ random-walk steps. In Fig. 11(a), we observe that the step function is denoised without the distortions that were introduced by using the low-pass filtering. Fig. 11(b) presents further noise suppression by using ten random-walk steps, still without distorting the step function edges. However, we observe in Fig. 11(c) that the signal is completely degenerated to a constant value when using too many steps ($t = 3000$). We elaborate and discuss this issue in details in Sections IV and V.

In practice, the affinity metric is usually extended to improve the performance of the NL filter. Instead of the 1-D metric between single samples, a high-dimensional metric between the samples entire neighborhoods or patches is used. Consequently, let the pairwise kernel be $k(x_i, x_j) = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\}$, where $\mathbf{x}_i$ is a vector consisting of the neighborhood of the sample x_i. For example, let $\mathbf{x}_i$ be a vector of length 3 given by $\mathbf{x}_i = [x_{i-1}, x_i, x_{i+1}]^T$. Fig. 12(a)–(c) shows the denoising results using the 3-D kernel after $t = 1$, $t = 3$, and $t = 10$ steps. We observe that the noise is completely suppressed, whereas the edges are maintained.

Appendix II
Diffusion Interpretation Example

In order to provide another interpretation of the NL filter, we degenerate the example presented in Appendix I. Now we assume the desired source signal is a constant corrupted by addi-

[9]The eigenvectors are a complete set spanning the space of the samples.

tive Gaussian noise, i.e., $d_i = d$. In this case, from (43), we have that the density of the samples $\{x_i\}$ is Gaussian

$$q(x_i) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left\{-\frac{(x_i - d)^2}{2\sigma_n^2}\right\}. \quad (47)$$

Thus, up to an additive constant, the potential, defined by $U = -2\ln q$, is parabolic

$$U(x_i) = \frac{(x_i - d)^2}{\sigma_n^2}. \quad (48)$$

As shown in [24] and [25], for a large data set $M \longrightarrow \infty$ and small kernel scale $\sigma \longrightarrow 0$, the transition matrix $\mathbf{P}$, which represents the discrete random-walk on the graph, converges to the continuous backward Fokker–Planck operator $\mathcal{L}$ (6). When using scalars, we have that for every smooth function $f : \Gamma \longrightarrow \mathbb{R}$, the resulting Fokker–Planck operator is merely a second-order differential equation, given by

$$\mathcal{L}f = f'' - U'f' \quad (49)$$

where f' and f'' are first- and second-order derivatives of f, and U' is the first derivative of the potential U.

It can also be shown that the eigenvectors of $\mathbf{P}$ are discrete approximation of the eigenfunctions of $\mathcal{L}$. In our case, using (48) and (49), the eigenfunctions $\psi_j(x_i)$ (which can be viewed as a smooth function on the data samples) satisfy the second-order differential equation

$$\mathcal{L}\psi_j(x_i) = \psi_j''(x_i) - \frac{2(x_i - d)}{\sigma_n^2}\psi_j'(x_i) = \mu_j\psi_j(x_i) \quad (50)$$

where μ_j are the corresponding eigenvalues of the continuous Fokker–Planck operator. The eigenfunctions that solve (50) are known as the Hermite polynomials $\psi_j(x_i) = H_j((x_i - d)/\sigma_n)$. The first three are given by $H_0(x) = 1$, $H_1(x) = x$, and $H_2(x) = x^2 - 1$. Thus, from (44) and the special form of the eigenfunction, we obtain that the expansion of the samples on the eigenbasis consists of only the first two terms

$$\begin{aligned}\mathbf{x} &= b_0\boldsymbol{\psi}_0 + b_1\boldsymbol{\psi}_1 \\ &= b_0\mathbf{1} + \frac{b_1}{\sigma_n}(\mathbf{x} - d\mathbf{1}) \\ &= d\mathbf{1} + (\mathbf{x} - d\mathbf{1})\end{aligned} \quad (51)$$

where $\mathbf{1}$ is a vectors of ones of length M. Combining (46) and (51) yields

$$\left[P^t\mathbf{x}\right]_i = d\mathbf{1} + \lambda_1^t(\mathbf{x} - d\mathbf{1}) \quad (52)$$

which means that each step of the random-walk shrinks the noise in $\mathbf{x}$ towards the desired mean value d at rate λ_1. Consequently, we obtained that applying the random-walk on the data set suppresses the additive Gaussian white noise and provides an estimate of the desired constant.

ACKNOWLEDGMENT

The authors would like to thank Prof. Ronald Coifman for helpful discussions. They also thank the anonymous reviewers for their constructive comments and useful suggestions.

REFERENCES

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

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

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

[5] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[7] S. V. Vaseghi and P. J. W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *IEE Proc. I: Commun. Speech Vis.*, pp. 38–46, Feb. 1990.

[8] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 3rd ed. New York: Wiley, 2006.

[9] S. J. Godsill and P. J. W. Rayner, "Statistical reconstruction and analysis of autoregressive signals in impulsive noise using the gibbs sampler," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 352–372, Jul. 1998.

[10] L. P. Yaroslavski, *Digital Picture Processing*. Berlin, Germany: Springer-Verlag, 1985.

[11] D. Barash, "A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 844–847, Jun. 2002.

[12] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Model. Simul.*, vol. 4, pp. 490–530, 2005.

[13] M. Mahmoudi and G. Sapiro, "Fast image and video denoising via nonlocal means of similar neighborhoods," *IEEE Signal Process. Lett.*, vol. 12, no. 12, pp. 839–842, Dec. 2005.

[14] A. D. Szlam, M. Maggioni, and R. Coifmain, "Regularization on graphs with function-adapted diffusion processes," *J. Mach. Learn. Res.*, 2007.

[15] A. Singer, Y. Shkolnisky, and B. Nadler, "Diffusion interpretation of non local neighborhood filters for signal denoising," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 118–139, 2009.

[16] A. Abramson and I. Cohen, "Enhancement of speech signals under multiple hypotheses using an indicator for transient noise presence," in *Proc. 32rd IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-2007*, Honolulu, HI, Apr. 2007, pp. IV-553–556.

[17] R. Talmon, I. Cohen, and S. Gannot, "Speech enhancement in transient noise environment using diffusion filtering," in *Proc. 35th IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP'10)*, Dallas, TX, Mar. 2010, pp. 4782–4785.

[18] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1996.

[19] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 260, pp. 2319–2323, 2000.

[20] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 260, pp. 2323–2326, 2000.

[21] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, 2003.

[22] D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 5591–5596, 2003.

[23] R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Nat. Acad. Sci.*, vol. 102, no. 21, pp. 7426–7431, May 2005.

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

[25] 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, 2006.
[26] F. R. K. Chung, *Spectral Graph Theory*, 1997, CBMS-AMS.
[27] G. W. Gardiner, *Handbook of Stochastic Processes for Physics*, 2nd ed. Berlin, Germany: Springer-Verlag, 2002.
[28] B. J. Matkowsky and Z. Schuss, "Eigenvalues of the Fokker-Planck operator and the approach to equilibrium for diffusions in potential fields," *SIAM J. Appl. Math.*, vol. 40, pp. 242–254, 1981.
[29] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators," *Neural Inf. Process. Syst. (NIPS)*, vol. 18, 2005.
[30] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering," *Neural Inf. Process. Syst. (NIPS)*, vol. 19, 2006.
[31] T. F. Quatieri, *Discrete Time Speech Signal Processing Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall, 2001.
[32] S. Godsill, *Digital Audio Restoration—A Statistical Model Based Approach*. London, U.K.: Springer-Verlag, 1998.
[33] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[34] M. Belkin and P. Niyogi, "Convergence of Laplacian Eigenmaps," in *Advances in Neural Information Processing Systems (NIPS) 19*, B. Schölkopf and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 129–136.
[35] Y. Weiss, "Segmentation using eigenvectors: A unifying view," in *Proc. Int. Conf. Comput. Vis.*, 1999.
[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[37] M. Meila and J. Shi, "A random walks view of spectral segmentation," in *8th Int. Workshop Artif. Intell. Statist.*, 2001.
[38] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 849–856.
[39] M. Hein and J. Y. Audibert, L. De Raedt and S. Wrobel, Eds., "Intrinsic dimensionality estimation of submanifold in r^d," in *Proc. 22nd Int. Conf. Mach. Learn., ACM*, 2005, pp. 289–296.
[40] A. Singer, "From graph to manifold Laplacian: The convergence rate," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 128–134, 2006.
[41] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph Laplacian tomography from unknown random projections," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1891–1899, Oct. 2008.
[42] J. S. Garofolo, "Getting Started With the DARPA TIMIT CD-ROM: An Acoustic-Phonetic Continous Speech Database," Nat. Inst. of Standards and Technology (NIST), Gaithersburg, MD, 1993.
[43] [Online]. Available: http://www.freesound.org
[44] [Online]. Available: http://www.eng.biu.ac.il/~gannot/

Ronen Talmon (S'09) received the B.A degree in mathematics and computer science from the Open University, Ra'anana, Israel, in 2005. He is currently pursuing the Ph.D. degree in electrical engineering at the Technion—Israel Institute of Technology, Haifa.

From 2000 to 2005, he was a Software Developer and Researcher in a technological unit of the Israeli Defense Forces. Since 2005, he has been a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion. His research interests are statistical signal processing, speech enhancement, system identification, harmonic analysis, and geometric methods for data analysis.

Israel Cohen (M'01–SM'03) received the B.Sc. (*summa cum laude*), M.Sc., and Ph.D. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, 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, CT. In 2001, he joined the Electrical Engineering Department, Technion, where he is currently an Associate Professor. 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 cochair of the 2010 International Workshop on Acoustic Echo and Noise Control.

Dr. Cohen is a recipient of the Alexander Goldberg Prize for Excellence in Research and the Muriel and David Jacknow Award for Excellence in Teaching. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as Guest Editor of a special issue of the *EURASIP Journal on Advances in Signal Processing* on Advances in Multimicrophone Speech Processing and a special issue of the *EURASIP Speech Communication Journal* on Speech Enhancement.

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (*summa cum laude*) from the Technion-Israel Institute of Technology, Haifa, in 1986 and the M.Sc. (*cum laude*) and Ph.D. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering.

In 2001, he held a post-doctoral position at the Department of Electrical Engineering (SISTA), K.U.Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Faculty of Electrical Engineering, Technion. Currently, he is an Associate Professor in the School of Engineering, Bar-Ilan University, Ramat-Gan, Israel.

Dr. Gannot is the recipient of Bar-Ilan University Outstanding Lecturer Award for the year 2010. He is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He is also Associate Editor of the *EURASIP Journal on Advances in Signal Processing*, an Editor of two special issues on Multi-Microphone Speech Processing of the same journal, a guest editor of the *ELSEVIER Speech Communication* journal and a reviewer of many IEEE journals and conferences. He has been a member of the Technical and Steering Committee of the International Workshop on Acoustic Echo and Noise Control (IWAENC) since 2005 and the general co-chair of IWAENC 2010 held in Tel-Aviv. His research interests include parameter estimation, statistical signal processing, and speech processing using either single- or multi-microphone arrays.