Single-Channel Transient Interference Suppression With Diffusion Maps

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

***Abstract*—A transient is an abrupt or impulsive sound followed by decaying oscillations, e.g., keyboard typing and door knocking. Such sounds often arise as interference in everyday applications, e.g., hearing aids, hands-free accessories, mobile phones, and conference-room devices. In this paper, we present an algorithm for single-channel transient interference suppression. The main component of the proposed algorithm is the estimation of the spectral variance of the interference. We propose a statistical model of the transient interference and combine it with non-local filtering. We exploit the unique spectral structure of the transients along with their impulsive temporal nature to distinct them from speech. A particular attention is given to handling both short- and long-duration transients. Experimental results show that the proposed algorithm enables significant transient suppression for a variety of transient types.**

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

I. Introduction

TRANSIENTS, which are characterized by sudden bursts of sound, often arise as interference in everyday applications, such as hearing aids, hands-free accessories, mobile phones, and conference-room devices. In this work a special focus is given to the suppression of repeating transient appearances, e.g., keyboard typing and construction operations. While a single transient may be ignored, repeating events make the interference especially annoying. In addition, such persistent reoccurrences may significantly hamper automatic speech recognition systems. Although transient interferences are very common their suppression in speech signals is still considered an open problem. To date, most of the existing single-channel noise reduction algorithms are based on estimation of stationary noise from segments in which the desired signal is absent. Clearly, this approach does not suit the abrupt nature of transient interference; hence, such algorithms are inadequate in this scenario.

Manuscript received August 29, 2011; revised February 26, 2012, May 28, 2012; accepted August 19, 2012. Date of publication August 27, 2012; date of current version October 18, 2012. This work was supported by the Israel Science Foundation under Grant 1130/11. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Man-Wai Mak.

R. Talmon is with the Department of Mathematics, Yale University, New Haven, CT 06520 USA (e-mail: ronen.talmon@yale.edu).

I. Cohen is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il).

S. Gannot is with the Faculty 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.2012.2215593

In [1] and [2], we circumvented this assumption by proposing an algorithm that learns the geometric structure of the transient interference using nonlocal (NL) diffusion filtering [3]–[8]. The key idea was to exploit the intrinsic transient structure instead of relying on estimates of noise statistics. We utilized the fact that a distinct pattern appears multiple times. Specifically, the locations of the repeating pattern were identified, and the transient interference was extracted by averaging over all these instances. Experimental studies showed significant enhancement of speech signals and attenuation of transient interferences, which are characterized by a short duration. Unfortunately, the algorithm could not well handle transients with a slowly decaying oscillatory part and varying amplitudes.

In this paper, we improve and extend [1] to support a wider variety of transient interferences. We utilize a manifold learning approach termed *diffusion maps* [9] to compute a robust intrinsic metric for comparison. In particular, it enables to cluster different transient interference types. We show that when the diffusion distance is incorporated into the NL filter, it provides a better affinity metric for averaging over transient instances. In addition, as was presented first in [10], we propose an intermediate step to distinguish between transients and speech. The approach is based on the observation that speech components are slowly varying with respect to transient interferences, just as stationary noise is slowly varying with respect to speech. Thus, by employing common speech enhancement techniques, configured to track faster variations, the "abrupt" transients can be enhanced while suppressing the slowly varying speech components. The enhanced transients are then utilized to improve the estimation of the spectrum of the interference. We note that exploiting the rate of change of the signal was previously introduced in RASTA [11], where bandpass filtering of the short-term power spectrum is employed to suppress both slowly and rapidly varying interferences.

In order to handle transients with slowly decaying oscillatory part, we adapt a statistical model used to describe room reverberation [12]. We split each transient instance into an abrupt part and a decaying part. The abrupt part has distinct spectral features, which can be captured by the NL filter. The decaying part is random and may resemble a speech component, and therefore it is estimated based on the statistical model, similarly to the variance estimator proposed in [12] for reverberant speech.

This paper is organized as follows. In Section II, we formulate the problem. In Section III, a statistical model for the transient interference is given. In Section IV, we present the proposed algorithm. A diffusion maps approach is described in Section V. Finally, in Section VI, experimental results are presented, demonstrating the improved performance of the proposed algorithm.

II. Problem Formulation

Let $x(n)$ denote a speech signal and let $t(n)$ and $u(n)$ be contaminating transient interference and stationary noise. The signal measured by a microphone is given by

$$y(n) = x(n) + t(n) + u(n). \tag{1}$$

It is worthwhile noting that numerous methods for enhancement of speech signals contaminated by (quasi) stationary noise can be found. Thus, we can employ one of these methods prior to the proposed algorithm.

The transient interference is represented as in [13], [14]

$$t(n) = h(n) * w(n) \tag{2}$$

where $w(n)$ is a sequence of impulses of varying amplitudes indicating the time locations of the transients, and $h(n)$ is an impulse response that models the duration and shape of the transient interference type.

In this work we use a fixed impulse response $h(n)$, which implies that all transient instances of a single interference type in the measurement have the same spectral features up to random amplitudes. Hence, the transient interference can be viewed as a superposition of the impulse response $h(n)$ with random amplitudes. It is worthwhile noting that in Section VI, the proposed algorithm is evaluated in practical scenarios using real transient interference recordings, with arbitrary shapes.

Let $Y(l,k)$ denote the short-time Fourier transform (STFT) of the measured signal $y(n)$ in time-frame l and frequency-bin k. We use analysis and synthesis windows of length N with time shift R. Accordingly, (1) can be represented in the STFT domain as

$$Y(l,k) = X(l,k) + T(l,k) + U(l,k) \tag{3}$$

where $X(l,k)$, $T(l,k)$ and $U(l,k)$ are the STFT of $x(n)$, $t(n)$ and $u(n)$, respectively. In (2), the transient interference is described as a linear convolution between a sequence of impulses $w(n)$ and a fixed filter $h(n)$. In order to properly represent the convolution in the STFT domain we cannot use the common multiplicative transfer function (MTF) approximation, since for long-duration transients, $h(n)$ may be longer than the time frame. Thus, according to the analysis presented in [15], we approximate the convolution using band-to-band filters as

$$T(l,k) \cong \sum_{l'=0}^{\infty} H(l',k) W(l-l',k) \tag{4}$$

where $W(l,k)$ is the STFT of $w(n)$ and $H(l,k)$ is a band-to-band filter of frequency-bin k corresponding to the impulse response $h(n)$. This representation enables to represent a linear convolution of filters longer than the time frame in the STFT domain, as often required in our case. For further details on the representation of the linear convolution in the STFT domain and the dependency on the transform parameters we refer the readers to [15] and the references therein.

We assume that no more than one transient event exists in each short time frame. We denote by $\mathcal{T}$ the set of time frames that contain a transient, and by $\bar{\mathcal{T}}$, we denote the set of time frames free of transient occurrences.

In this work, we aim at estimating the clean speech signal $x(n)$ given the noisy measurements $y(n)$.

III. Preliminaries

Following [12], we propose a statistical model for the band-to-band filters. We have

$$H(l,k) = \begin{cases} B_a(k), & l = 0 \\ B_d(l,k) e^{-\alpha(k) l R}, & l \geq 1 \end{cases} \tag{5}$$

where $\alpha(k)$ denotes the decay rate of the filter. $B_d(l,k)$ are zero-mean, independent and identically distributed (i.i.d.) Gaussian random variables, representing the decaying part of the transient. $B_a(k)$ is a zero-mean Gaussian random variable independent of $B_d(l,k)$, representing the abrupt part of the transient. We note that $B_a(k)$ determines the spectral features which characterize the transient type. On the other hand, $B_d(l,k)$ represent the random unstructured decaying part in frequency bin k along the time frames $l \geq 1$. We assume the abrupt and decaying parts have different statistical characteristics and that the abrupt part entails most of the energy. In addition, we assume $B_d(l,k)$ are i.i.d., which implies that for a given transient type the random oscillations across time frames have the same statistical characteristics.

Let $\beta_a(k) = \mathbb{E}[|B_a(k)|^2]$ and $\beta_d(k) = \mathbb{E}[|B_d(l,k)|^2]$ be the spectral variances of the abrupt and decaying parts, respectively. We can now compute the spectral variance of the filter

$$\begin{aligned} \lambda_h(l,k) &= \mathbb{E}\left[|H(l,k)|^2\right] \\ &= \begin{cases} \beta_a(k), & l = 0 \\ \beta_d(k) e^{-2\alpha(k) l R}, & l \geq 1 \end{cases}. \end{aligned} \tag{6}$$

In (2), $w(n)$ is a sequence of impulses of varying amplitudes. Thus, we model its spectral variance $\lambda_w(l,k)$ as a fixed value across the frequency bins, which is determined by the impulse amplitude. We have

$$\lambda_w(l) \triangleq \lambda_w(l,k) = \begin{cases} B_w(l), & l \in \mathcal{T} \\ 0, & l \notin \mathcal{T} \end{cases} \tag{7}$$

where $B_w(l)$ denote the amplitudes of the impulses in the short-time spectrum domain.

Let $\lambda_y(l,k) = \mathbb{E}[|Y(l,k)|^2]$ be the spectral variance of the measured signal. We assume that the speech, the transient interference, and the stationary noise are uncorrelated. Thus, the spectral variance of the measurement is given by

$$\lambda_y(l,k) = \lambda_x(l,k) + \lambda_t(l,k) + \lambda_u(k) \tag{8}$$

where $\lambda_x(l,k) = \mathbb{E}[|X(l,k)|^2]$, $\lambda_t(l,k) = \mathbb{E}[|T(l,k)|^2]$, and $\lambda_u(k) = \mathbb{E}[|U(l,k)|^2]$.

IV. Proposed Algorithm

The proposed algorithm consists of four components in a cascade: 1) An optimally modified log spectral amplitude (OM-LSA) algorithm for enhancing the transients; 2) a spectral variance estimator for separating the abrupt and decaying parts of the transients; 3) a non-local filter for estimating the power spectral density (PSD) of the abrupt parts, and 4) an additional OM-LSA for suppressing the transients and the stationary

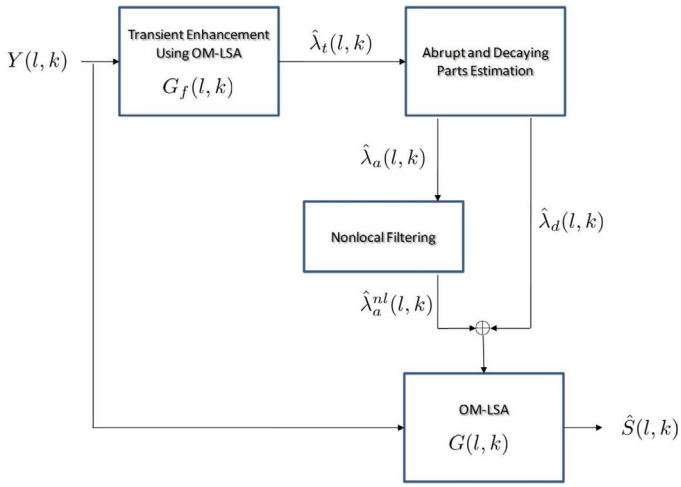

Fig. 1. A block diagram of the proposed algorithm.

noise, and enhancing the speech. Fig. 1 depicts a diagram of the proposed algorithm.

A. Transient Enhancement Using OM-LSA

Our goal in this stage is to enhance transients and attenuate mainly structured speech phonemes in order to obtain a better distinction in the succeeding nonlocal filtering stage. We observe that as stationary noise is slowly varying with respect to speech, speech is slowly varying with respect to transient interference. Thus, transient enhancement is attained by *short-term* averaging over past spectral power values, such that the fluctuations of speech segments are reduced. We use the OM-LSA method [16] and modify it to enhance the transient interference and suppress the speech. The applied modifications are described in this Section.

The log-spectral amplitude (LSA) estimator [17] is useful for reducing background noise in speech signals. In [16], new estimators were introduced for the a priori signal to noise ratio (SNR) and for the a priori speech absence probability (SAP). The spectral gain function of the algorithm is then obtained as a weighted geometric mean of the hypothetical gains associated with signal presence and absence. The algorithm components are based on the noise spectrum, estimated by the minima controlled recursive averaging (MCRA) [18]. The MCRA recursively averages past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in sub-bands.

We configure the MCRA algorithm to track rapid variations. This way, employing the MCRA enables to estimate the PSD of the slower speech $\lambda_x(l,k)$ and stationary noise $\lambda_u(k)$ given the measurements $Y(l,k)$. We use very short time frames of length 16 ms in order to reduce the variations of the speech between subsequent frames. In addition, the following temporal smoothing is carried out

$$S(l,k) = \gamma S(l-1,k) + (1-\gamma)|Y(l,k)|^2 \tag{9}$$

where $S(l,k)$ is the smoothed PSD of the measurements, and γ is a recursion parameter. We choose a relatively small recursion parameter $\gamma = 0.5$ to enable quick tracking of speech components. However, the recursion parameter should not be too small to discard abrupt changes attributed to transients.

The described modification enables to capture most of the speech parts, but sudden changes characterizing voiced phoneme onsets are overlooked. Beginnings of transients can be distinguished from beginnings of phonemes by exploiting the fact that transients are typically shorter and decay faster than voiced phonemes. Thus, by introducing a short lag, we are able to observe future samples and make this distinction. Phoneme onset identification is obtained by using two sliding windows. One window is causal, and used to detect the minimum power in previous frames, as described in [19], [18], [20]. The other window is anti-causal, and used to detect the minimum power in future frames. We note that the window should be shorter than a typical speech phoneme, but longer than a typical transient. The PSD estimate is taken as the maximum of the two minima detected in the two windows. Formally, let L_1 and L_2 be the length of the causal and anti-causal windows, and l be the frame under consideration. Thus, the minima values are picked in the causal window

$$S^c_{\min}(l,k) = \min_{l'} \{S(l',k) \,|\, l' = l - L_1 + 1, \ldots, l\} \tag{10}$$

and in the anti-causal window as

$$S^{ac}_{\min}(l,k) = \min_{l'} \{S(l',k) \,|\, l' = l + 1, \ldots, l + L_2\}\,. \tag{11}$$

Then, the PSD estimate of the speech and stationary noise is obtained by

$$\max\{S^c_{\min}(l,k), S^{ac}_{\min}(l,k)\}\,. \tag{12}$$

Now, we demonstrate the behavior of the described method. At the beginning of a speech phoneme, the minimum in the causal window is low, conveying the power level of the background noise before the phoneme. On the other hand, the minimum in the anti-causal window is high, representing the power level of the phoneme (assuming the window is shorter than the phoneme). Consequently, taking the maximum of the two minima yields the desired estimate of the power level of the phoneme. It is worthwhile noting that a transient instance is not captured in this process. Since both windows are longer than a transient, the minima in such windows must be the power level of the background signal (either speech or background noise) before or after the transient.

The typical lengths of the causal and anti-causal windows range between 40 and 160 ms, which determines the lag introduced into the system. For a typical sampling rate of 16000 Hz and short time frames of length 16 ms (with 75% overlap), it corresponds to 10 to 40 frames (L_1 and L_2).

A particular attention is given to unvoiced phonemes, whose durations might be shorter than 40 ms. We assume that unvoiced phonemes usually appear adjacent to voiced phonemes with higher power and longer duration due to speech harmonies. Thus, unvoiced phoneme onsets are less likely to be local maxima with respect to the causal and anti-causal windows. As a result, we are usually able to distinguish both voiced and unvoiced speech from transients using the temporal averaging.

However, short duration plosive phonemes, especially in high frequency bins, might be wrongly detected and might not be attenuated by this filter. Since such phonemes are usually less structured compared to transients, they are well handled in a succeeding stage of the algorithm as described in Section IV-C. In Section VI, we demonstrate the processing of unvoiced speech.

Let $G_f(l,k)$ denote the spectral gain of the OM-LSA estimator, based on the described MCRA component with two windows configured to track fast variations. Thus, the enhanced transient interference can be written as

$$\begin{aligned}\hat{\lambda}_t(l,k) &= \exp\left\{\mathbb{E}\left[\log|T(l,k)|^2 \mid Y(l,k)\right]\right\} \\ &= |G_f(l,k)Y(l,k)|^2. \end{aligned} \tag{13}$$

In practice we employ a mild recursive temporal smoothing on $\hat{\lambda}_t(l,k)$ to circumvent alignment and offset problems, and to enable extra robustness. We employ on each frequency bin a first order auto-regressive filter with a recursive coefficient of 0.95.

B. Abrupt and Oscillatory Decaying Parts Estimation

The idea in this work is to exploit the transient geometric structure by utilizing the fact that a distinct pattern appears multiple times [1]. Examination of a wide variety of transient interferences led us to the observation that each transient event is characterized by an abrupt sound followed by decaying oscillations. Unfortunately, only the abrupt part demonstrates a characteristic structure, whereas the decaying part has a more random nature. Thus, in order to exploit the repetitive nature of the abrupt part of the transient, we first need to decompose each transient instance into its abrupt and decaying parts. In this section, we propose to estimate the decaying part based on a statistical model adapted from room reverberations modeling [12]. We emphasize that we merely exploit the fact that both applications deal with suppression of decaying interferences. However, there is no additional similarity between dereverberation applications and the proposed work.

The independence of $B_a(k)$ and $B_d(l,k)$ from (5) yields that $\mathbb{E}[H(l,k)H^*(l',k)] = 0$ for $l \neq l'$. By further assuming that $H(l,k)$ and $W(l,k)$ are mutually independent, we have from (4)

$$\begin{aligned}\lambda_t(l,k) &= \mathbb{E}\left[\left|\sum_{l'=0}^{\infty} H(l',k)W(l-l',k)\right|^2\right] \\ &= \sum_{l'=0}^{\infty} \lambda_h(l',k)\lambda_w(l-l'). \end{aligned} \tag{14}$$

Using (6), we can reformulate (14) as

$$\lambda_t(l,k) = \lambda_a(l,k) + \lambda_d(l,k) \tag{15}$$

with

$$\lambda_a(l,k) = \beta_a(k)\lambda_w(l) \tag{16}$$

and

$$\begin{aligned}\lambda_d(l,k) &= \sum_{l'=1}^{\infty} \beta_d(k)e^{-2\alpha(k)l'R}\lambda_w(l-l') \\ &= e^{-2\alpha(k)R}\left[\lambda_d(l-1,k) + \beta_d(k)\lambda_w(l-1)\right]. \end{aligned} \tag{17}$$

Using (16), $\lambda_d(l,k)$ can be rewritten as

$$\lambda_d(l,k) = e^{-2\alpha(k)R}\left[\lambda_d(l-1,k) + \kappa(k)\lambda_a(l-1,k)\right] \tag{18}$$

where $\kappa(k) = \beta_d(k)/\beta_a(k)$. From (15), we obtain

$$\begin{aligned}\lambda_d(l,k) = [1-\kappa(k)]\, e^{-2\alpha(k)R}\lambda_d(l-1,k) \\ +\kappa(k)e^{-2\alpha(k)R}\lambda_t(l-1,k). \end{aligned} \tag{19}$$

As seen from (19), we require that $0 < \kappa(k) \leq 1$, which implies that the energy of the abrupt part is larger than the energy of the decaying part. From (6) and by the derivation in [12], $\kappa(k)$ can be estimated by solving the following equation

$$\frac{1-e^{-2\alpha(k)R}}{e^{-2\alpha(k)R}}\frac{\beta_a(k)}{\beta_d(k)} = \frac{\sum_{n=0}^{n_a-1}[h(n)]^2}{\sum_{n=n_a}^{\infty}[h(n)]^2}$$

where n_a is the length of the abrupt part and is set empirically for every transient type. Now, since $\lambda_t(l-1,k)$ is unavailable, we use in (19) its estimate $\hat{\lambda}_t(l-1,k)$ from Section IV-A to obtain an estimate for the transient abrupt and decaying parts. We note that we may also use the more accurate estimate of the transient obtained in the following section. Let $\hat{\lambda}_d(l,k)$ be the PSD estimate of the decaying part obtained from (19). Accordingly, from (15), we have an estimate of the abrupt part $\hat{\lambda}_a(l,k)$

$$\hat{\lambda}_a(l,k) = \hat{\lambda}_t(l,k) - \hat{\lambda}_d(l,k). \tag{20}$$

C. Nonlocal Filtering

We improve the estimation of the abrupt transient part by exploiting its repetitive nature, i.e., 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 interference may be extracted from the measurement by averaging over all of these instances. This naturally leads to NL filtering [4]–[8].

Let $\bar{p}(\cdot\,,\cdot\,)$ be a non-negative kernel defined between any pair of time frames, such that for any time frame index l, we have $\sum_{l'}\bar{p}(l,l') = 1$. A single step of the NL filter is given by

$$\boldsymbol{\lambda}_a^{nl}(l) = \sum_{l'=1}^{M}\bar{p}(l,l')\hat{\boldsymbol{\lambda}}_a(l') \tag{21}$$

where M is the number of time frames of the measured signal, $\hat{\boldsymbol{\lambda}}_a(l)$ is a vector of length N, obtained by collecting all frequency bins of a single time frame, i.e.,

$$\hat{\boldsymbol{\lambda}}_a(l) = \left[\hat{\lambda}_a(l,0),\ldots,\hat{\lambda}_a(l,N-1)\right]^T, \tag{22}$$

and $\boldsymbol{\lambda}_a^{nl}(l)$ is similarly defined as

$$\boldsymbol{\lambda}_a^{nl}(l) = \left[\lambda_a^{nl}(l,0),\ldots,\lambda_a^{nl}(l,N-1)\right]^T. \tag{23}$$

The NL filter step can be repeated a few times, where each step of the NL filter (21) can be interpreted as averaging over similar time frames according to $\bar{p}$. Let tt denote the number of filter iterations. As we describe in Section V, $\bar{p}$ is associate with the probability of a Markov process to go from one frame to another in a single step. Applying tt NL filter steps is associated with the probability to go from one frame to another in tt steps, and as a result more remote frames may be averaged. For simplicity, the derivation is presented based on a single filter step, however, in Section VI, we use several iterations to enable better performance.

We assume that the kernel implicitly separates between time frames according to the transient events presence. This assumption provides trackable approximated analysis of the NL filtering. We note that this assumption circumvents the risk of applying several steps of the NL filter, which is analyzed in [1]. In particular, it satisfies $\bar{p}(l,l') = 0$ for either $l \in \mathcal{T}$ and $l' \in \bar{\mathcal{T}}$, or $l' \in \mathcal{T}$ and $l \in \bar{\mathcal{T}}$. The values $\bar{p}(l,l')$ in each set convey the uncertainty of the division. The construction of such kernel function is described in Section V.

According to (21), applying a single iteration of the NL filter based on such kernel yields

$$\begin{aligned}\lambda_a^{nl}(l,k) &= \sum_{l'=1}^{M} \bar{p}(l,l')\hat{\lambda}_a(l',k) \\ &= \begin{cases} \sum\limits_{l'\in\mathcal{T}} \bar{p}(l,l')\hat{\lambda}_a(l',k), & l \in \mathcal{T} \\ \sum\limits_{l'\in\bar{\mathcal{T}}} \bar{p}(l,l')\hat{\lambda}_a(l',k), & l \in \bar{\mathcal{T}} \end{cases}.\end{aligned} \tag{24}$$

As a result, since each transient interference has the same abrupt spectral pattern (16), the transient instances are averaged together and enhanced. On the other hand, the "random" speech (after pre-filtering) is averaged incoherently, and therefore suppressed. After a few iterations of the NL filter, the instances of the transient interference may be extracted. In [1] we discuss the diffusion interpretation of the NL filter and present a probabilistic analysis which enables to determine the expected performance and the proper choice of parameters, e.g., the proper number of filter steps tt.

Next, we formulate the employment of the NL filter. We express the estimate of the abrupt transient part obtained in Section IV-B as

$$\hat{\lambda}_a(l,k) = \lambda_a(l,k) + \varepsilon(l,k) \tag{25}$$

where $\varepsilon(l,k)$ is the positive estimation error with mean $\mu_\varepsilon(k) > 0$ and variance $\sigma_\varepsilon^2(k)$, which consists of the spectral variance of the residual speech components and stationary noise. The parameters of the OM-LSA in Section IV-A are set to enable small transient distortion at the expense of significant speech leftovers. This includes a restriction of the maximum attenuation of spectral components to -15 dB. In addition, the windows lengths are set to be longer than the typical transients at the expense of including short-duration speech phonemes.

Substituting (7), (16) and (25) into (24) yields

$$\lambda_a^{nl}(l,k) = \lambda_a(l,k)\pi(l) + \hat{\mu}_\varepsilon(l,k) \tag{26}$$

with the scaling variable

$$\pi(l) = \begin{cases} \hat{\mu}_w(l)/B_w(l), & l \in \mathcal{T} \\ 1, & l \in \bar{\mathcal{T}} \end{cases} \tag{27}$$

and the weighted sums,

$$\hat{\mu}_\varepsilon(l,k) = \begin{cases} \sum\limits_{l'\in\mathcal{T}} \bar{p}(l,l')\varepsilon(l',k), & l \in \mathcal{T} \\ \sum\limits_{l'\in\bar{\mathcal{T}}} \bar{p}(l,l')\varepsilon(l',k), & l \in \bar{\mathcal{T}} \end{cases} \tag{28}$$

and

$$\hat{\mu}_w(l) = \sum_{l'\in\mathcal{T}} \bar{p}(l,l')B_w(l'). \tag{29}$$

In (26), we observe that the output of the NL filter consists of a clean estimate of the abrupt part of the transient, up to scaling, and an additive error term. Unlike the additive error $\varepsilon(l,k)$ in (25), $\hat{\mu}_\varepsilon(l,k)$ can be considered time independent, as it is an average of error terms over time. Thus, we employ spectral subtraction to reduce the additive error. We subtract from (26) the average of $\hat{\mu}_\varepsilon$ obtained from time frames that do not contain transients, yielding

$$\bar{\lambda}_a^{nl}(l,k) \triangleq \lambda_a^{nl}(l,k) - \frac{1}{|\bar{\mathcal{T}}|}\sum_{l'\in\bar{\mathcal{T}}} \lambda_a^{nl}(l',k) \tag{30}$$

where $|\mathcal{A}|$ denote the cardinality of the set $\mathcal{A}$. Using (26), (30) can be written as

$$\bar{\lambda}_a^{nl}(l,k) = \lambda_a(l,k)\pi(l) + \varepsilon^{nl}(l,k) \tag{31}$$

where the residual additive error is given by

$$\varepsilon^{nl}(l,k) = \hat{\mu}_\varepsilon(l,k) - \sum_{l'\in\bar{\mathcal{T}}} \hat{\mu}_\varepsilon(l',k)/|\bar{\mathcal{T}}|. \tag{32}$$

We note that while $\mathbb{E}[\varepsilon(l,k)] = \mu_\varepsilon(k) > 0$ implying that the initial estimate $\hat{\lambda}_a(l,k)$ is biased, the residual error in $\bar{\lambda}_a^{nl}(l,k)$ satisfies $\mathbb{E}[\varepsilon^{nl}(l,k)] = 0$. To further analyze the behavior of the NL filter, we assume perfect kernel clustering, i.e.,

$$\bar{p}(l,l') = \begin{cases} 1/|\mathcal{T}|, & l,l' \in \mathcal{T} \\ 1/|\bar{\mathcal{T}}|, & l,l' \in \bar{\mathcal{T}} \\ 0, & \text{o.w.} \end{cases}. \tag{33}$$

Thus, we obtain from (28) and (32) that

$$\begin{aligned}\mathrm{Var}\left[\varepsilon^{nl}(l,k)\,|\,l\in\mathcal{T}\right] &= \frac{1}{|\mathcal{T}|}\sigma_\varepsilon^2(k) + \frac{1}{|\bar{\mathcal{T}}|^2}\sigma_\varepsilon^2(k) \\ &= \frac{1}{M}\frac{1}{\xi}\sigma_\varepsilon^2(k) + \frac{1}{M^2}\frac{1}{(1-\xi)^2}\sigma_\varepsilon^2(k)\end{aligned} \tag{34}$$

with $M = |\mathcal{T}|+|\bar{\mathcal{T}}|$ the number of time frames and $\xi = |\mathcal{T}|/M$. In case the number of transients grows linearly with the number of time frames, ξ can be approximated by a constant. In addition the number of frames containing transients is relatively small, and hence, the variance of $\varepsilon^{nl}(l,k)$ is attenuated with rate $1/M$. Moreover, sufficient number of time frames ($M \to \infty$) result in

$$\bar{\lambda}_a^{nl}(l,k) = \lambda_a(l,k)\pi(l). \tag{35}$$

In (35), we obtain a consistent estimate of the abrupt part of the transient, up to scaling.

Since the scaling $\pi(l)$ depends only on the time frame and independent of the frequency bin (it represents the amplitude of the entire transient event), it can be estimated based on the variability of spectral speech components. Dividing the NL filter output after spectral subtraction in (35) by the initial estimate in (25) yields the following ratio

$$\frac{\bar{\lambda}_a^{nl}(l,k)}{\hat{\lambda}_a(l,k)} = \pi(l)\left[1+\frac{\varepsilon(l,k)}{\lambda_a(l,k)}\right]^{-1}. \tag{36}$$

We assume that for each time frame $l \in \mathcal{T}$, there exists a frequency bin with negligible residual $\varepsilon(l,k) \approx 0$. Such a frequency bin exists since the speech does not typically span the entire spectrum, and the stationary noise may successfully be suppressed by the OM-LSA in Section IV-A. Thus, since $\varepsilon(l,k)$ and $\lambda_a(l,k)$ are positive, from (36) we have

$$\hat{\pi}(l) = \min_k \left\{ \frac{\bar{\lambda}_a^{nl}(l,k)}{\hat{\lambda}_a(l,k)} \middle| \, k = 0,1,\ldots,N-1 \right\} \tag{37}$$

for each $l \in \mathcal{T}$. In practice, before the minima search, we employ local smoothing between adjacent frequency bins to increase the robustness to small variations of the abrupt part at the expense of smearing the minima. We use a window function b whose length is $2k_b+1$, and for each frame l we smooth the ratio along the frequency bins k

$$\sum_{i=-k_b}^{k_b} b(i)\frac{\hat{\lambda}_a(l,k-i)}{\bar{\lambda}_a^{nl}(l,k-i)} \tag{38}$$

where the natural choice for b is any smooth window, e.g., Hamming.

Finally, from (35) and (37) we compute

$$\hat{\lambda}_a^{nl}(l,k) = \frac{1}{\hat{\pi}(l)}\bar{\lambda}_a^{nl}(l,k) = \frac{\pi(l)}{\hat{\pi}(l)}\lambda_a(l,k) \tag{39}$$

which is an estimate of the abrupt part of the transient, free of the speech leftovers $\varepsilon(l,k)$.

D. Speech Enhancement Using OM-LSA

To enhance the speech, we apply OM-LSA again (at a second time), now with a modified noise PSD estimate, similarly to [1]. We set the optimal spectral gain to the sum of the transient interference and the stationary noise PSD estimates. Let $G(l,k)$ denote the spectral gain of the OM-LSA estimator given the noisy measurement $Y(l,k)$. Thus, the speech estimate is given by

$$\hat{X}(l,k) = G(l,k)Y(l,k). \tag{40}$$

The spectral gain relies on the noise PSD estimate $\hat{\lambda}_u(l,k) + \hat{\lambda}_t^{nl}(l,k)$ where $\hat{\lambda}_u(l,k)$ is the PSD estimate of the stationary noise obtained by the MCRA [18], and $\hat{\lambda}_t^{nl}(l,k) = \hat{\lambda}_a^{nl}(l,k) + \hat{\lambda}_d(l,k)$ is the estimate of the PSD of the transient interference obtained in Sections IV-B and IV-C.

Since the calculation of the optimal spectral gain function is now controlled by both the stationary noise and transient interference, additional suppression of the transients is attainable. For more details regarding the optimal gain function derivation and estimation of the speech presence probability and the noise spectrum, we refer the reader to [16] and references therein. In addition, the OM-LSA parameters used in this stage are similar to the parameters of an OM-LSA set to enhance speech and reduce background noise, as proposed in [16]. We note that as in [16], the phase of the noisy signal is used for reconstructing the enhanced speech. A Matlab code of the OM-LSA is available at [21].

V. Distance Measure of Transients Based on Diffusion Maps

In this section we present the intrinsic distance measure $\bar{p}(\cdot,\cdot)$ defined between any pair of time frames. This distance measure, used in the NL filtering (21), is based on diffusion maps, and is described in the remainder of the section.

We define an affinity metric $k(\cdot,\cdot)$ between pairs of the vectors $\{\hat{\boldsymbol{\lambda}}_a(l)\}$ estimated in Section IV-B using the following Gaussian kernel

$$k(l,l') = \exp\left\{-\|\hat{\boldsymbol{\lambda}}_a(l) - \hat{\boldsymbol{\lambda}}_a(l')\|^2/2\sigma^2\right\} \tag{41}$$

where σ^2 is the variance of the Gaussian kernel which determines the scale of the affinity metric. For more details regarding this specific choice of a kernel see [1]. It is worthwhile noting that different frame length than the length used in Section IV may be used here by concatenating the power spectrum of consecutive time frames into $\hat{\boldsymbol{\lambda}}_a(l)$.

We view the vectors $\{\hat{\boldsymbol{\lambda}}_a(l)\}_{l=1}^M$ as nodes of an undirected symmetric graph, where M denotes the number of available time frames of the measurement. Two nodes $\hat{\boldsymbol{\lambda}}_a(l)$ and $\hat{\boldsymbol{\lambda}}_a(l')$ are connected by an edge with weight $k(l,l')$, that corresponds to the affinity between $\hat{\boldsymbol{\lambda}}_a(l)$ and $\hat{\boldsymbol{\lambda}}_a(l')$. We continue with the construction of a Markov process on the graph nodes by normalizing the kernel k as

$$p(l,l') = k(l,l')/d(l) \tag{42}$$

where $d(l) = \sum_{l'=1}^M k(l,l')$. Consequently, $p(l,l')$ represents the probability of transition in a single step from node $\hat{\boldsymbol{\lambda}}_a(l)$ to node $\hat{\boldsymbol{\lambda}}_a(l')$. Similarly, let $p_t(l,l')$ be the probability of transition in t steps from node $\hat{\boldsymbol{\lambda}}_a(l)$ to node $\hat{\boldsymbol{\lambda}}_a(l')$. Let $\mathbf{K}$ denote the matrix corresponding to the kernel function k, and let $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}$ be the matrix corresponding to the function p, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ll} = d(l)$. Accordingly, $\mathbf{P}^t$ is the matrix corresponding to the function p_t.

Results from spectral theory [22] can be employed to describe $\mathbf{P}$, enabling to study the geometric structure of $\{\hat{\boldsymbol{\lambda}}_a(l)\}$ 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 $1 = \rho_0 > \rho_1 \geq$

$\rho_2 \geq \ldots$, satisfying $\mathbf{P}\boldsymbol{\psi}_j = \rho_j \boldsymbol{\varphi}_j$. Each singular vector, $\boldsymbol{\varphi}_j$ or $\boldsymbol{\psi}_j$, is of length M, and $\varphi_j(l)$ and $\psi_j(l)$ denote access to the lth elements of the singular vectors. We note that since the sum of each row of $\mathbf{P}$ equals to one, the 0th component is trivial, i.e., $\rho_0 = 1$ and $\boldsymbol{\psi}_0$ elements equal to one. The left eigenvector $\boldsymbol{\varphi}_0$, associated with the eigenvalue $\rho_0 = 1$, is the stationary probability, i.e., $\boldsymbol{\varphi}_0 \mathbf{P} = \boldsymbol{\varphi}_0$.

The construction of the Markov process leads to a definition of a new affinity metric between any two vectors [9]

$$D_t^2(l,l') = \|p_t(l,\cdot) - p_t(l',\cdot)\|_{\boldsymbol{\varphi}_0}^2 = \sum_{l''=1}^{M} \left(p_t(l,l'') - p_t(l',l'')\right)^2 / \varphi_0(l'') \quad (43)$$

for any integer t. This metric is termed *diffusion distance* as it relates to the evolution of the transition probability distribution $p_t(l,l')$. Moreover, it enables to describe the relationship between pairs of vectors in terms of their graph connectivity. Consequently, the main advantage of the diffusion distance is that local structures and rules of transitions are integrated together into a global metric. In recent years, this metric was shown to be very useful in various applications from different fields [23]–[27], [7].

We use the right eigenvectors of the transition matrix $\mathbf{P}$ to obtain a new data-driven description of the M vectors $\{\hat{\boldsymbol{\lambda}}_a(l)\}$ via a family of mappings that are termed *diffusion maps* [9]. Let $\Psi_t : \mathbb{R}^N \mapsto \mathbb{R}^\ell$ be the diffusion mappings of the M vectors $\{\hat{\boldsymbol{\lambda}}_a(l)\}$ into a Euclidean space $\mathbb{R}^\ell$ for any integer t, defined as

$$\Psi_t(\hat{\boldsymbol{\lambda}}_a(l)) = \left[\rho_1^t \psi_1(l), \ldots, \rho_\ell^t \psi_\ell(l)\right]^T \quad (44)$$

where ℓ is the new space dimensionality ranging between 1 and $M-1$. We note that a fast decay of the eigenvalues $\{\rho_j\}$ may enable dimensionality reduction, as coordinates in (44) become negligible for large ℓ.

It can be shown that the diffusion distance (43) is equal to the Euclidean distance in the diffusion maps space when using all $\ell = M-1$ eigenvectors [9], i.e.,

$$D_t(l,l') = \|\Psi_t(\hat{\boldsymbol{\lambda}}_a(l)) - \Psi_t(\hat{\boldsymbol{\lambda}}_a(l'))\|. \quad (45)$$

This result provides a justification for using the Euclidean distance in the new space for comparison and clustering purposes. Instead of aggregating the transition probabilities over all possible trajectories as implied by (43), we may simply compute the Euclidean distance between the embedded samples. In particular, since the eigenvalues typically decay fast for a large enough t, the diffusion distance can be well approximated by only the first few ℓ eigenvectors, yielding efficient calculations of the diffusion distance. In Section VI, we show that embedding the vectors into the diffusion maps space naturally organizes the measurements into separate clusters of speech and transient interference.

Similarly to (41), we now define a *new* Gaussian kernel $\bar{k}$ based on the diffusion distance

$$\bar{k}(l,l') = \exp\{-\|\Psi_t(\hat{\boldsymbol{\lambda}}_a(l)) - \Psi_t(\hat{\boldsymbol{\lambda}}_a(l'))\|^2 / 2\bar{\sigma}^2\} \quad (46)$$

and, similarly to (42), construct a corresponding Markovian process to obtain a new transition probability function $\bar{p}(l,l')$. According to the diffusion analysis in [9], the application of the Gaussian kernel in (41) enables parametrization of the lower-dimensional structure of transients. The second application of the Gaussian kernel in (46) intensifies the locality property by implicitly defining a neighborhood around each embedded sample $\Psi_t(\hat{\boldsymbol{\lambda}}_a(l))$ of radius $\bar{\sigma}$. Embedded samples $\Psi_t(\hat{\boldsymbol{\lambda}}_a(l'))$ such that $\|\Psi_t(\hat{\boldsymbol{\lambda}}_a(l)) - \Psi_t(\hat{\boldsymbol{\lambda}}_a(l'))\| > \bar{\sigma}$ are weakly connected to $\Psi_t(\hat{\boldsymbol{\lambda}}_a(l))$.

We emphasize that unlike the kernel (41) used in [1] for the NL filtering which relies on the Euclidean distance between the vectors, in this work we use for the NL filtering a kernel that relies on diffusion distance. The use of diffusion distance conveys the capability to distinguish between different types of transients, and hence, the proposed algorithm enables handling few transient types simultaneously. In Section VI we demonstrate the performance of the new kernel based on the diffusion distance and compare it with the kernel based on the Euclidean distance.

The computational burden of the diffusion maps approach may be significant. The computation of the kernel requires M^2 calculations of the distance between each pair of samples. In addition, a spectral decomposition of an $M \times M$ matrix is employed. In practice, we relax these two steps. First, since we use a Gaussian kernel, we clip small values of the kernel, corresponding to remote samples (not in time but in distance), to zero, and obtain a sparse kernel matrix. Consequently, we are able to use efficient spectral decomposition algorithms adapted for sparse matrices. Second, we employ an efficient version of an approximated k-nearest-neighbors search. Thus, instead of calculating the kernel between each sample and all the rest of the samples, we calculate the kernel only between each sample and its nearest neighbors. In our implementation, for the spectral decomposition of sparse matrices we use the standard MATLAB implementation, and for the k-nearest-neighbors search we use TSTOOL for MATLAB available online in [28].

The derivations of the diffusion maps and distance presented in this section require the availability of all the data. As a result, the presented algorithm entails batch processing of the entire measurement interval. There exists an efficient online version of diffusion maps computation which requires training and calibration [29]. However, this issue is beyond the scope of this paper.

VI. Experimental Results

In this section, we evaluate the performance of the proposed algorithm. We use recorded speech and transient signals sampled at 16 KHz. Speech signals are taken from the TIMIT database [30], and recorded transient interferences are taken from an online free corpus [31]. The measurements are constructed according to (1). We re-scale the speech and transient interference to have equal maximal amplitude in the measured interval. The additive stationary noise part is a computer generated white Gaussian noise with signal to noise ratio (SNR) of 20 dB. The length of each speech utterance and the corresponding transient interference is 20 s. Such transient interference signal typically consists of 25 to 30 transient events.

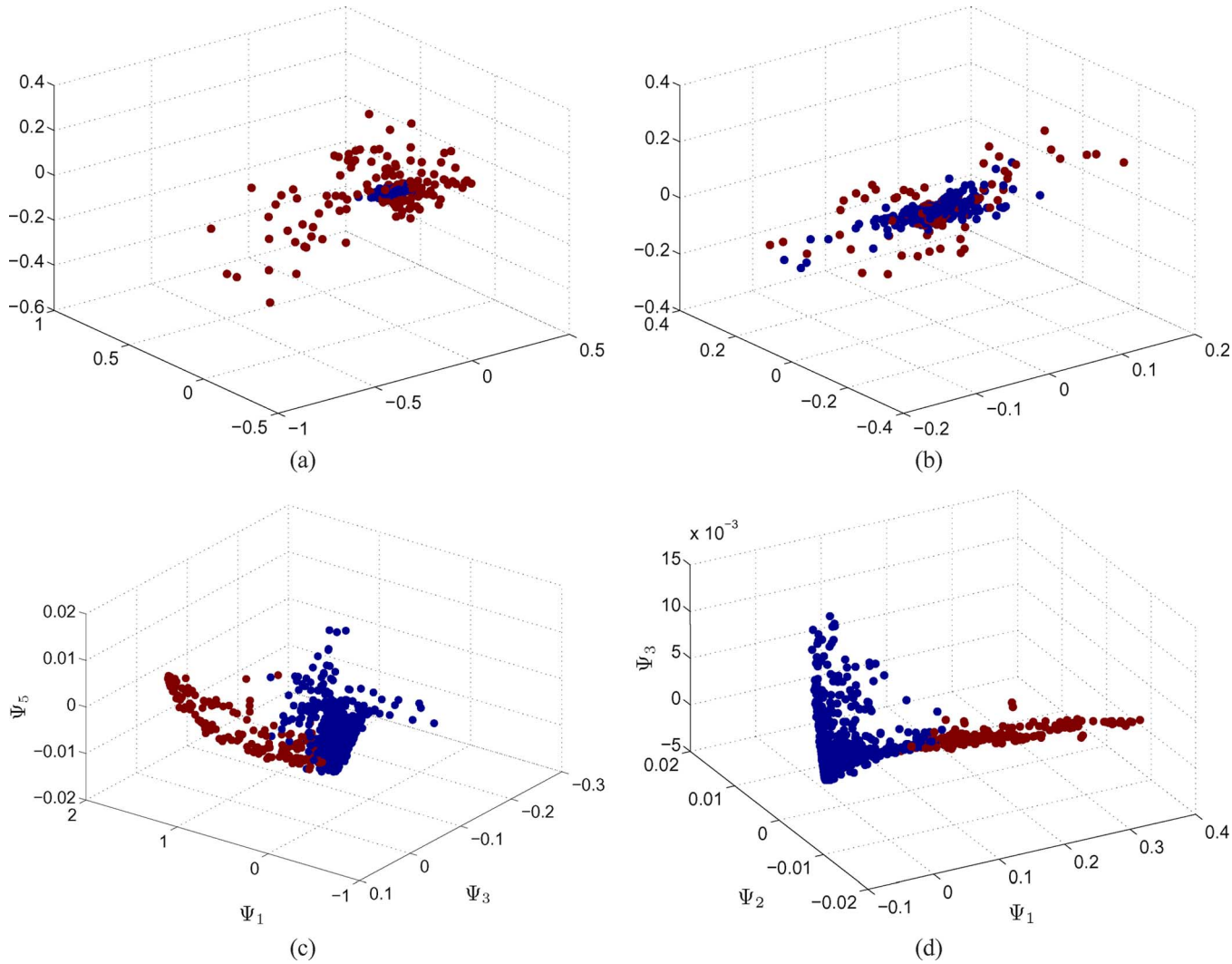

Fig. 2. (a) A scatter plot of the PSD of 3 arbitrary frequency bins of speech contaminated by kitchen knocks. (b) A scatter plot of the PSD of 3 arbitrary frequency bins of speech contaminated by door knocks. (c) A scatter plot of the 1st, 3rd, and 5th coordinates of the diffusion map of speech contaminated by kitchen knocks. (d) A scatter plot of the 1st, 2nd, and 3rd coordinates of the diffusion map of speech contaminated by door knocks.

TABLE I
PARAMETERS OF THE PROPOSED ALGORITHM

Transient type	Anti-causal window length [s]	Oscillation decay (α)	Diffusion Maps frame length [s]
Metronome	0.04	0.2	0.032
Door knocks	0.16	0.7	0.064
Kitchen knocks	0.12	0.7	0.128
Keyboard typing	0.04	0.4	0.128
Household clicks	0.04	0.2	0.032

The algorithm configurable parameters used in the experiments are summarized in Table I. The length of the anti-causal window approximately corresponds to the typical length of the transient. The frame length was chosen empirically to fit each transient type and to produce maximal results. We note that the frame length used for the abrupt and decaying parts estimation, the nonlocal filtering, and the speech enhancement stage equals the specified length used for diffusion maps. Only in the transient enhancement stage we use a fixed length of 16 ms for all transient types. In addition, we used 75% overlap between successive frames. The proper oscillation decay parameter α for each transient type is picked by examining the decay rate of a clean representative transient event. Our experimental results show that it suffices to set a single value of the decay rate corresponding to all frequency bins. In addition, the results show that the algorithm performance is insensitive to different choices of α in a wide range of values. We note that in general it is preferable to choose smaller values, which enable under-estimation of the decaying part. This choice results in decaying parts leftovers in the output signal, whereas larger values may result in speech distortion. As depicted in Table I, we demonstrate the suppression of transient types with both short- and long-duration decaying parts. The enhancement of all transient types is attained using $tt = 128$ iterations of the NL filter. The parameters of the second application of the OM-LSA are set (to enhance the speech) according to [16].

Fig. 2 illustrates the diffusion maps embedding. In Fig. 2(a) and (b) we present a scatter plot, where each point represents the PSD estimate of a time frame of the abrupt part in three arbitrary frequency bins. We note that similar visualization is obtained by different choices of the three bins. Fig. 2(a) depicts points corresponding to the samples of speech contaminated by kitchen knocks, and Fig. 2(b) depicts points corresponding to speech

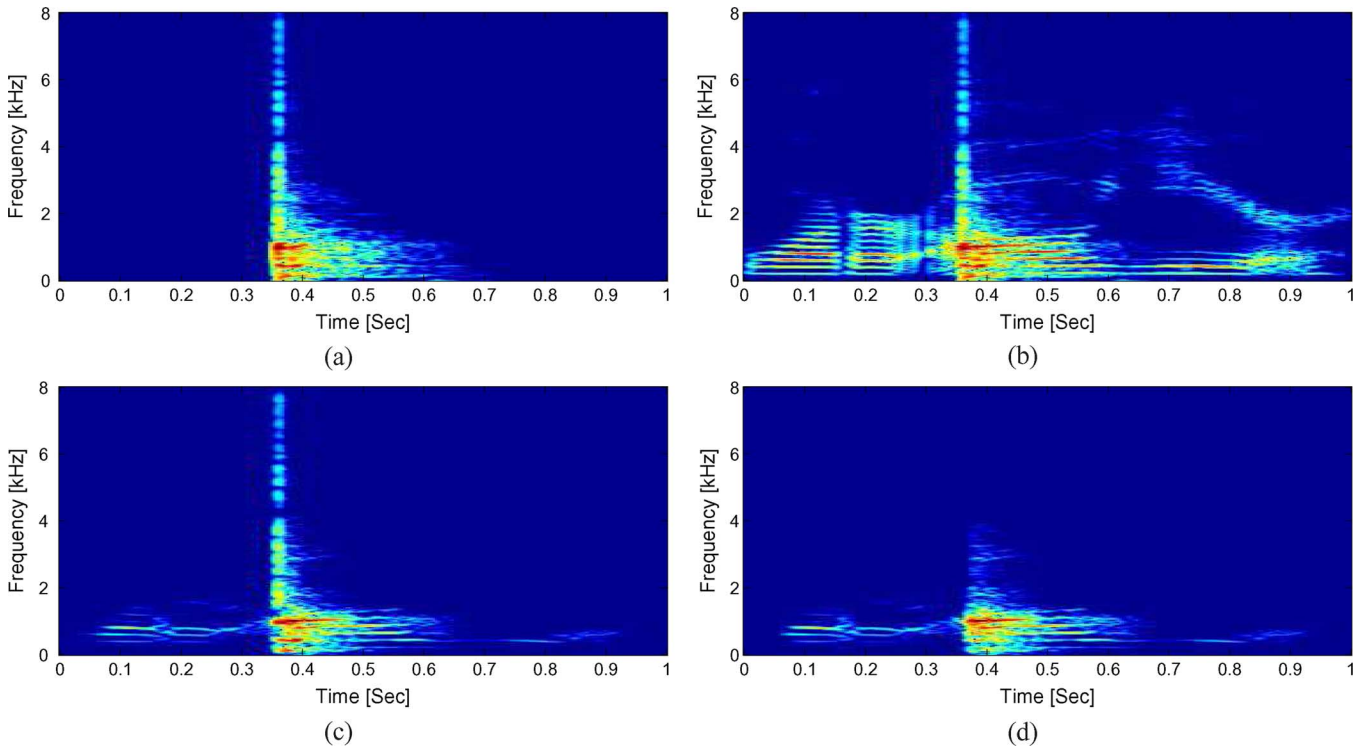

Fig. 3. Signal spectrograms. (a) A clean transient event. (b) The noisy signal. (c) The enhanced transient. (d) The estimated decaying part.

contaminated by door knocks. In Fig. 2(c) and (d) we present a scatter plot of the diffusion maps, where each point represents the *embedding* according to (44) of the samples from Fig. 2(a) and (b), respectively. In Fig. 2(c) we show the 1st, 3rd, and 5th coordinates of the embedding, and in Fig. 2(d) we show the 1st, 2nd, and 3rd coordinates of the embedding. We present merely three coordinates for the 3D illustration. The color of the points represents the frame content: frames containing transients appear in brown and frames without transients appear in blue. In Fig. 2(a) and (b) we do not detect any separation between the time frames according to their content. On the other hand, in Fig. 2(c) and (d) we observe a clear clustering according to transient presence. However, when using merely three coordinates, we see some points with different content overlap. Our empirical testing show that by using $\ell = 10$ dimensions, the diffusion maps embedding provides an adequate separation of the points and minimal overlaps.

To illustrate the performance of the transient enhancement and estimation of the abrupt and decaying parts we present in Fig. 3 spectrograms of a single transient event of a kitchen knock. Fig. 3 shows the clean transient event, the noisy signal, the enhanced transient PSD $\hat{\lambda}_t$, and the estimated decaying part PSD $\hat{\lambda}_d$. In Fig. 3(a) we can clearly see both the abrupt part of transient, characterized by a vertical PSD shape, followed by the decaying part. We note that such a knock has a relatively long decaying part, and hence, it is more difficult to estimate and suppress. In Fig. 3(d) we observe the estimated decaying part, and notice that the abrupt part is successfully excluded from the estimation. We note that the PSD estimate of the abrupt transient part is estimated by subtracting (d) from (c). Fig. 3 also illustrates the important role of the NL filtering. As observed, the enhanced transient signal contains residual speech components. If not attenuated, such residual speech in the transient PSD estimate may result in significant speech distortion, as these speech components would be suppressed along with the transient interference and the background noise by the second application of the OM-LSA. We note that most of the speech leftovers remain in the abrupt part and as a consequence are suppressed by the NL filter. However, as observed in Fig. 3(d), few speech components with a low signal power, located typically around transients, are estimated as parts of the decaying components. Thus, we remove them by applying a threshold on the signal power, where the threshold value is set empirically and was shown to be suitable for all tested transient types.

Fig. 4 shows spectrograms of the noisy speech signal corrupted by metronome interference, the transient estimate, and the enhanced signal. We observe that the proposed method yields an accurate estimation of the spectrum of the transient interference and attains significant transient interference reduction while imposing very low distortion.

In Fig. 5 we demonstrate the enhancement of plosive phonemes attained by the algorithm by presenting the spectrograms corresponding to the utterance “Oh, the time of death” corrupted by household clicks. In Fig. 5(a) we present the noisy speech and in Fig. 5(b) the enhanced signal. We observe that the plosive “t” at 0.6 s is undistorted. Furthermore, the plosive “d” at 1.1 s is undistorted, whereas the adjacent transient event is suppressed.

We test the algorithm on several speech utterances (of both males and females) and transient interference types. We compare the performance of the proposed algorithm using two different kernels: the kernel proposed in [1] based on the Euclidean distance, and the kernel proposed in Section V, based on the diffusion distance.

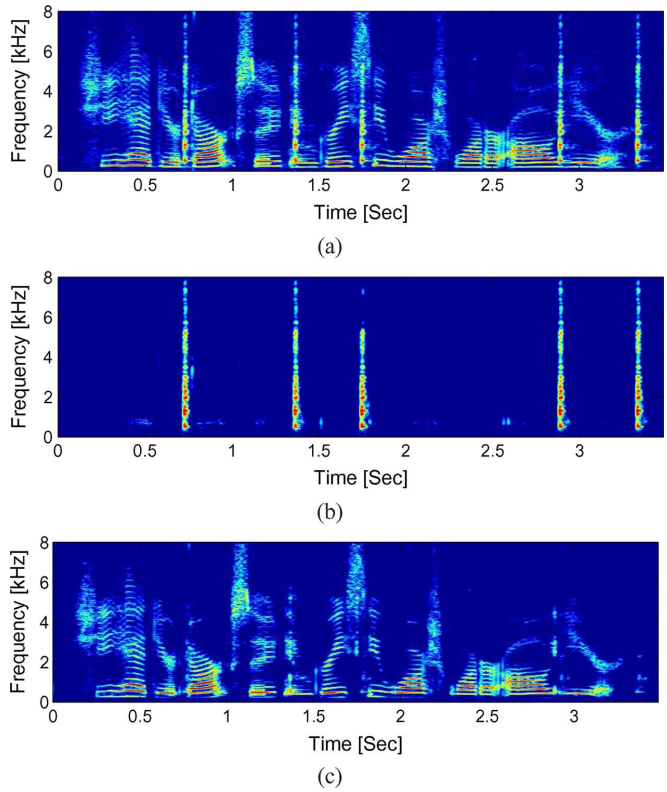

Fig. 4. (a) Spectrogram of the noisy speech signal corrupted by metronome interference. (b) Spectrogram of the transient estimate. (c) Spectrogram of the enhanced signal obtained using the proposed algorithm.

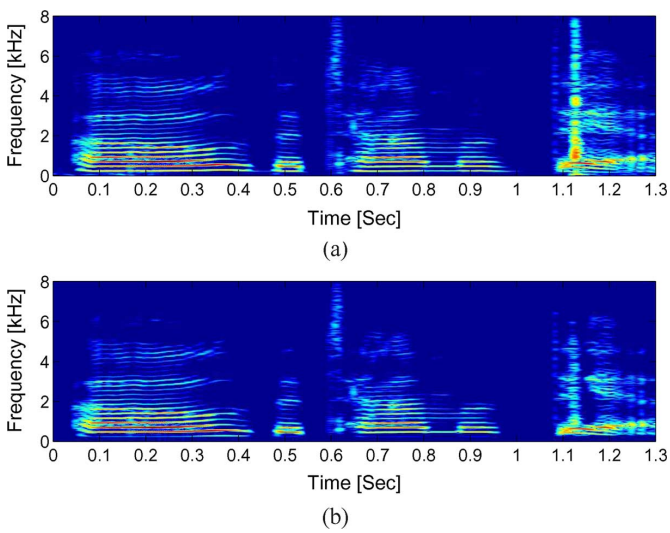

Fig. 5. Spectrograms corresponding to the utterance "Oh, the time of death" corrupted by "household clicks". (a) The noisy speech. (b) The enhanced signal.

We evaluate the transient interference estimation using two objective measures. The first is the transient to signal ratio (TSR), defined by

$$\begin{aligned}\mathrm{TSR}_{\mathrm{in}} &= 10\log_{10}\frac{\mathbb{E}\left\{t^2(n)\right\}}{\mathbb{E}\left\{(y(n)-t(n))^2\right\}}\\ \mathrm{TSR}_{\mathrm{out}} &= 10\log_{10}\frac{\mathbb{E}\left\{t^2(n)\right\}}{\mathbb{E}\left\{(\hat{t}(n)-t(n))^2\right\}}\end{aligned} \tag{47}$$

where $\hat{t}(n)$ is the estimated transient interference corresponding to the PSD $\hat{\lambda}_t^{nl}(l,k)$. We note that in order to obtain the signals in the time domain, we use the phase of the noisy signal $y(n)$. The second measure is the mean spectral distance (SD) defined as

$$\begin{aligned}\mathrm{SD}_{\mathrm{in}} &\triangleq \mathbb{E}_l\left[\frac{1}{N}\sum_{k=0}^{N-1}\left(\lambda_t(l,k)-\lambda_y(l,k)\right)\right]^{\frac{1}{2}}\\ \mathrm{SD}_{\mathrm{out}} &\triangleq \mathbb{E}_l\left[\frac{1}{N}\sum_{k=0}^{N-1}\left(\lambda_t(l,k)-\hat{\lambda}_t^{nl}(l,k)\right)\right]^{\frac{1}{2}}\end{aligned} \tag{48}$$

where $\mathbb{E}_l[\cdot]$ is the ensemble average with respect to the time frame l. 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. We note that both measures are computed only in time periods where the *estimation* contains transient interference.

Table II summarizes the evaluation of the transient interference estimation using Euclidean and diffusion distances, respectively. We report the average measures obtained over several speech utterances for different transient interferences. The results are compared to the transient estimation obtained by the algorithm proposed in [1]. We observe significant improvement of the measures implying good transient estimation. The usage of diffusion distance outperforms the usage of the Euclidean distance in most of the tested cases, except keyboard typing. Compared to the algorithm in [1], the proposed algorithm is advantageous in TSR improvement. SD improvement results are inconclusive and the differences between the algorithms are relatively small.

We evaluate the output of the proposed algorithm using two objective measures as well. The first is the common SNR, defined as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{in}} &= 10\log_{10}\frac{\mathbb{E}\{x^2(n)\}}{\mathbb{E}\{(y(n)-x(n))^2\}}\\ \mathrm{SNR}_{\mathrm{out}} &= 10\log_{10}\frac{\mathbb{E}\{x^2(n)\}}{\mathbb{E}\{(\hat{x}(n)-x(n))^2\}}.\end{aligned} \tag{49}$$

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

$$\begin{aligned}\mathrm{LSD}_{\mathrm{in}} &\triangleq \mathbb{E}_l\left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(\lambda_x(l,k))-\ell(\lambda_y(l,k))\right|^2\right]^{\frac{1}{2}}\\ \mathrm{LSD}_{\mathrm{out}} &\triangleq \mathbb{E}_l\left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(\lambda_x(l,k))-\ell(\hat{\lambda}_x(l,k))\right|^2\right]^{\frac{1}{2}}\end{aligned} \tag{50}$$

where

$$\ell(\lambda) = \max\{10\log_{10}\lambda, \delta\} \tag{51}$$

and δ is a small value defined by $\delta = \max_{l,k}\lambda(l,k) - 50$, used to confine the dynamic range of the log-spectrum to 50 dB.

TABLE II
EVALUATION OF THE TRANSIENT INTERFERENCE ESTIMATION

Transient type	TSR improvement [dB]			SD improvement [dB][1]		
	Algorithm proposed in [1]	Euclidean distance	Diffusion distance	Algorithm proposed in [1]	Euclidean distance	Diffusion distance
Metronome	13.91	15.02	**17.98**	20.3	**20.5**	**20.5**
Door knocks	11.27	8.45	**12.93**	19.6	19.0	**19.7**
Kitchen knocks	4.40	4.02	**5.38**	**18.9**	18.8	**18.9**
Keyboard typing	8.39	**8.83**	**8.83**	**19.7**	19.6	19.5
Household clicks	14.71	15.13	**15.76**	19.7	20.3	**20.5**

[1]Since lower SD is better, SD improvement is defined as $\mathrm{SD}_{in} - \mathrm{SD}_{out}$.

TABLE III
SPEECH ENHANCEMENT EVALUATION

Transient type	SNR improvement [dB]			LSD improvement[2]		
	Algorithm proposed in [1]	Euclidean distance	Diffusion distance	Algorithm proposed in [1]	Euclidean distance	Diffusion distance
Metronome	4.93	4.10	**5.45**	1.37	1.32	**1.45**
Door knocks	4.22	3.89	**4.83**	1.26	1.14	**1.54**
Kitchen knocks	4.53	8.05	**9.25**	0.93	1.35	**2.05**
Keyboard typing	7.27	7.49	**9.70**	2.23	2.34	**3.61**
Household clicks	4.06	4.53	**5.06**	1.03	1.25	**1.42**

[2]Since lower LSD is better, LSD improvement is defined as $\mathrm{LSD}_{in} - \mathrm{LSD}_{out}$.

These measures are computed in time periods where the estimate of the PSD of transients exists. This way we are able to focus on the performance of the proposed algorithm and evaluate the speech enhancement and the artifacts introduced by the algorithm simultaneously. In periods where the transient estimate does not exit, only stationary noise suppression is attained, and the performance of the algorithm equals to the performance of the OM-LSA.

Table III depicts the evaluation of the speech enhancement. We compare the proposed algorithm configured with Euclidean and diffusion based kernel, with the algorithm proposed in [1]. We observe improvement in all tested cases. In addition, the use of a diffusion-based kernel outperforms a Euclidean-based kernel and the algorithm proposed in [1]. It is worthwhile noting that informal hearing tests demonstrate significant reduction of the transient interference. In case of metronome or household clicks the interference is hardly notable in the result. However, kitchen knocks and keyboard typing are attenuated the least, although the SNR and the LSD measures report the most significant improvement. These transient types are particularly difficult to handle. The knocks have a relatively long decaying part, causing the random oscillations to "hide" the distinct spectral pattern of the abrupt part. However as presented in Table III, the improvement obtained by the proposed algorithm (using either distances) is significantly better than the improvement obtained by [1]. This result demonstrates the contribution of the separate handling of the decaying part. In case of keyboard typing, different key strokes have different spectral features, and hence the typing instances are less similar, and the nonlocal averaging is less effective. In this case, the use of diffusion distance which enables better handling to few transient types simultaneously shows superior improvement compared to the other algorithms.

Table IV presents the perceptual evaluation of speech quality (PESQ) score. The PESQ score is computed over the entire utterance and not only in transient periods. We observe improvement of the speech quality in all tested cases, and note that the improvement for short-duration transients, such as metronome and household clicks, is larger than for long-duration transients, e.g., kitchen knocks. Both versions of the proposed algorithm outperform the previous algorithm. It implies the benefit of using the transient enhancement stage and the separation of the abrupt and decaying parts in the proposed algorithm. In addition, as presented in the previous tables, the diffusion based kernel is advantageous compared to the Euclidean based kernel. It is worthwhile noting that even small increase in the PESQ score suggests noticeable improvement, as any sudden increase of power (e.g., attenuated transients) is audible. Audio samples of the presented results are available online in [32].

TABLE IV
PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ) SCORE

Transient type	Noisy	Algorithm proposed in [1]	Euclidean distance	Diffusion distance
Metronome	2.028	2.849	2.890	**2.966**
Door knocks	1.859	2.105	2.418	**2.604**
Kitchen knocks	1.755	1.998	2.071	**2.233**
Keyboard typing	1.456	1.762	2.217	**2.240**
Household clicks	2.096	2.722	2.830	**2.959**

In Table V we demonstrate that the proposed approach is capable of suppression of a few transient interferences simultaneously. We test the suppression of a different keyboard typing recording containing strokes on a wider variety of keys. We note that since the gaps between consecutive strokes is larger compared to the previous sample the results in Tables III and V are not comparable. In addition, we test a new door knocks sample by incorporating more types of door knocks (some with shorter duration), and a new household interferences sample by combining the household clicks and knocks. The reported experimental results show the same trends except for the LSD improvement of household interferences. The difference in the improvement is due to the different distribution of transients and the varying durations. It is worthwhile noting the same decay rate of the oscillatory part is assumed for all the transient interferences. However, as the algorithm is not sensitive to the

TABLE V
EVALUATION OF SUPPRESSION OF MULTIPLE INTERFERENCES

Transient type	SNR improvement [dB]		LSD improvement		PESQ scores		
	Algorithm proposed in [1]	Diffusion distance	Algorithm proposed in [1]	Diffusion distance	Noisy	Algorithm proposed in [1]	Diffusion distance
Keyboard typing	6.46	**9.47**	1.12	**2.71**	2.165	2.519	**2.766**
Household interferences	4.56	**5.20**	**2.07**	1.83	2.028	2.555	**2.691**
Door knocks	7.84	**8.17**	1.33	**2.96**	1.933	2.116	**2.526**

particular choice of the decay rate, this limitation is not very restricting.

VII. CONCLUSION

We have presented an algorithm for transient interference suppression in speech signals. The main component of the proposed algorithm is the estimation of the spectral variance of the transient interference, which is carried out in three steps. First, the transient interference is enhanced by applying a common algorithm for noise PSD estimation, equipped with two sliding windows and configured to track rapid variations characterizing transients. Second, the decaying part of the transient interference is estimated using a variance estimator based on a statistical model adapted from modeling room reverberations. Third, the abrupt part of the transient is estimated. The estimation is based on NL filtering, that exploits the intrinsic geometric structure of the transients. In particular, it relies on the variation between speech components and sharp impulses of repeating transient interference events. The distinction between transient and speech is obtained by incorporating a manifold learning approach termed diffusion maps, which naturally enables to compute an intrinsic metric for the signals at hand. Experimental results have demonstrated that significant interference suppression is attainable for a variety of transient types, which include both short- and relatively long-duration events. In addition, we demonstrated suppression of few transient types simultaneously. In future work, we intend to develop an online version of the diffusion maps and nonlocal filter parts.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] M. Mahmoudi and G. Sapiro, "Fast image and video denoising via nonlocal means of similar neighborhoods," *IEEE Signal Process. Lett.*, vol. 12, pp. 839–842, 2005.

[7] A. D. Szlam, M. Maggioni, and R. R. Coifman, "Regularization on graphs with function-adapted diffusion processes," *J. Mach. Learn. Res.*, vol. 9, pp. 1711–1739, 2008.

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

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

[10] R. Talmon, I. Cohen, and S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," in *Proc. 36th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP-11)*, Prague, Czech Republic, May 2011, pp. 5084–5087.

[11] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[12] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

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

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

[15] Y. Avargel and I. Cohen, "System identification in the short time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] [Online]. Available: http://webee.technion.ac.il/Sites/People/IsraelCohen/

[22] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI: Amer. Math. Soc., 1997.

[23] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using Nystrom method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[24] 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," in *Proc. Nat. Acad. Sci. U.S.A.*, May 2005, vol. 102, no. 21, pp. 7426–7431.

[25] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point could: From transductive to semi supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005.

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

[27] S. Lafon and A. B. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1393–1403, Sep. 2006.
[28] [Online]. Available: http://www.physik3.gwdg.de/tstool/
[29] R. Coifman and S. Lafon, "Geometic harmonics: A novel tool for multiscale out-of-sample extension of empirical functions," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 31–52, Jul. 2006.
[30] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continuous speech database," National Inst. of Standards and Technology (NIST). Gaithersburg, MD, Feb. 1993.
[31] [Online]. Available: http://www.freesound.org
[32] [Online]. Available: http://users.math.yale.edu/rt294/

Ronen Talmon (M'11) received the B.A. degree (cum laude) in mathematics and computer science from the Open University, Ra'anana, Israel, in 2005 and the Ph.D. degree in electrical engineering from the Technion—Institute of Technology, Haifa, in 2011.

From 2000 to 2005, he was a software developer and researcher at a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion. In 2011, he joined the Mathematics Department at Yale University, where he is currently a Gibbs Assistant Professor. His research interests are statistical signal processing, analysis and modeling of signals, speech enhancement, applied harmonic analysis, and diffusion geometry.

Dr. Talmon is the recipient of the Viterbi Fellowship for 2011–2012, the Irwin and Joan Jacobs Fellowship for 2011, the Excellent Project Supervisor Award for 2010, and the Excellence in Teaching Award for outstanding teaching assistants for 2008 and 2011.

Israel Cohen (M'01–SM'03) is an Associate Professor of electrical engineering at the Technion—Israel Institute of Technology, Haifa, Israel. He received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001 he joined the Electrical Engineering Department of the Technion. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2008), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), a coeditor of *Speech Processing in Modern Communication: Challenges and Perspectives* (Springer, 2010), and a general co-chair of the 2010 International Workshop on Acoustic Echo and Noise Control (IWAENC).

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 *Elsevier Speech Communication Journal* on Speech Enhancement.

Sharon Gannot (S'92–M'01–SM'06) received his B.Sc. degree (summa cum laude) from the Technion—Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in electrical engineering. In the year 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U. Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is an Associate Professor at the School of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory. Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010.

Prof. Gannot is currently an Associate Editor of IEEE TRANSACTIONS ON SPEECH, AUDIO AND LANGUAGE PROCESSING. He served as an Associate Editor of the *EURASIP Journal of Advances in signal Processing* between 2003–2001, and as an Editor of two special issues on Multi-microphone Speech Processing of the same journal. He also served as a guest editor of *ELSEVIER Speech Communication* journal and a reviewer of many IEEE journals and conferences. Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASSP) technical committee of the IEEE since Jan., 2010. He is also 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 held at Tel-Aviv, Israel in August 2010. Prof. Gannot will serve as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013. His research interests include parameter estimation, statistical signal processing, especially speech processing using either single- or multi-microphone arrays.