Supervised Graph-based Processing for Sequential Transient Interference Suppression

Ronen Talmon, *Member, IEEE*, Israel Cohen, *Senior Member, IEEE*,
Sharon Gannot, *Senior Member, IEEE*, and Ronald R. Coifman

Abstract

In this paper we present a *supervised* graph-based framework for sequential processing and employ it to the problem of transient interference suppression. Transients typically consist of an initial peak followed by decaying short-duration oscillations. Such sounds, e.g. keyboard typing and door knocking, often arise as an interference in everyday applications: hearing aids, hands-free accessories, mobile phones, and conference-room devices. We describe a graph construction using a noisy speech signal and training recordings of typical transients. The main idea is to capture the transient interference structure, which may emerge from the construction of the graph. The graph parametrization is then viewed as a data-driven model of the transients and utilized to define a filter that extracts the transients from noisy speech measurements. Unlike previous transient interference suppression studies, in this work the graph is constructed in advance from training recordings. Then, the graph is extended to newly acquired measurements, providing a sequential filtering framework of noisy speech.

Index Terms

Ronen Talmon and Ronald R. Coifman are with the Department of Mathematics, Yale University, New Haven 06520, CT (e-mail: ronen.talmon@yale.edu, ronald.coifman@math.yale.edu).

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

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

R. Talmon and I. Cohen work was supported by the Israel Science Foundation (grant no. 1130/11). R. Talmon was partially funded by the Viterbi Fellowship, Technion. R. Coifman work was supported by DARPA and the SPAWAR System Center Pacific under Contract No. N66001-11-C-4092.

Speech enhancement, speech processing, acoustic noise, transient noise, graph filtering

I. Introduction

Transients typically consist of an initial peak followed by decaying short-duration oscillations of length ranging from 10ms to 50ms. Such sounds, e.g. keyboard typing and door knocking, often arise as an interference in everyday applications: hearing aids, hands-free accessories, mobile phones, and conference-room devices. Unfortunately, the wide-spread assumption of stationary noise poses a major limitation on traditional speech enhancement algorithms. In particular, it makes them inadequate in transient interference environments, as transients are characterized by a sudden burst of sound.

In [1] and [2] we proposed an algorithm that infers the geometric structure of the transient interference using nonlocal (NL) diffusion filtering [3] [4] [5] [6] [7] [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 implicitly identified, and the transient interference was extracted by averaging over all these instances. In [9] and [10] this work was improved and extended to support a wider variety of transient interferences. A robust approach to distinguish between transients and speech was employed based on the observation that speech components are slowly varying with respect to transient interferences, just as pseudo-stationary noise is slowly varying with respect to speech. Thus, by employing common speech enhancement techniques, configured to track rapid variations, the "abrupt" transients can be enhanced while suppressing the slowly varying speech components. In addition, a manifold learning approach termed *diffusion maps* was utilized to compute a robust intrinsic metric for comparison [11]. It enabled to cluster different transient interference types, and when incorporated into the NL filter, it provided a better affinity metric for averaging over transient instances.

Recently several *supervised* speech enhancement algorithms, which rely on the prior knowledge of the typical interference patterns, have been proposed [12], [13], [14]. In these algorithms, nonnegative matrix factorization (NMF) is employed to compute a basis for the interferences, which is then utilized to enhance the speech and suppress the noise. However, these algorithms suffer from several limitations. They require training recordings of both the interference and the speech, which, as shown in [13], makes the algorithms speaker-dependent. In addition, the application of NMF is required for every new measurement and its computational burden is high. Finally, when applied to enhance speech and suppress noise, as in [14], a temporal smoothing is applied which makes the algorithm inadequate for transient interferences.

In this paper we present a *supervised* graph-based framework for sequential processing and employ it to the problem of transient interference suppression. In [15], Haddad et al. presented a novel filtering framework based on a reference set. They introduce a graph-based method that relies on local models and enables to extract given patterns from images. Based on this work, we describe a graph construction relative to a measured signal and training recordings. The objective of the graph is to capture the underlying structure of the training data, which has to represent all the variations of a certain signal of interest. The graph parametrization is then viewed as a data-driven model of the signal of interest and utilized to define a filter that extracts this signal from the measurement. The construction of the graph is based on an affinity kernel between the measurement and the training recordings. As proposed in [15], we rely on a specially-adapted metric based on local models of the signal of interest obtained from the training data.

We show that the application of the proposed scheme to the task of transient interference suppression provides accurate and efficient speech enhancement. Common speech enhancement algorithms fail to deal with transient interferences since their noise estimation component is not designed to track the rapid variations characterizing transients. Thus, similarly to [1] and [9], the main component of the proposed algorithm is the estimation of the spectral variance of the transient interference. Then, the optimally modified log-spectral amplitude (OM-LSA) estimator [16] [17], which is a single-channel speech enhancement algorithm, is employed to enhance the speech based on the estimate of the transient signal spectral variance. In this setting, the training recordings include typical transient interferences. Based on training recordings of the transient signal, the graph enables to accurately capture the structure of the transients. Then, the graph-based filter extracts it from the noisy speech and provides accurate spectral variance estimate. Previous studies, e.g., [1] and [9], infer the geometric structure of the transients from the noisy signal and employ batch processing. In this work, the graph is constructed in advance from training recordings, and a special focus is given to extending the graph to new measurements and to proposing a sequential filtering framework of the noisy signal.

This paper is organized as follows. In Section II, we formulate the problem. In Section III, we present the graph construction and the corresponding processing framework. In this section, we describe a batch processing of a finite observation interval. In Section IV, we compute local models of the transients from the training data and incorporate them into the construction of the graph. In Section V, we present an efficient sequential implementation which may be adapted to realtime speech communication systems. Finally, in Section VI, experimental results are presented, demonstrating the improved performance of the proposed algorithm.

II. Problem Formulation

Let $x(n)$ denote a clean speech signal picked up with a single microphone. The observed signal $y(n)$ is given by

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

where $t(n)$ and $u(n)$ are additive transient interference and stationary background noise, respectively, and n is the time index. The transient component $t(n)$ may consist of multiple types of interferences.

Let $Y(l, k)$ denote the short-time Fourier transform (STFT) of the microphone signal $y(n)$ in time-frame l and frequency-bin k. Let N denote the number of nonnegative frequency bins corresponding to analysis and synthesis windows of length $2(N - 1)$, and let R denote the time frame shift. Accordingly, (1) is represented in the STFT domain as

$$Y(l, k) = X(l, k) + T(l, k) + U(l, k)$$

where $X(l, k)$, $T(l, k)$ and $U(l, k)$ are the STFTs of $x(n)$, $t(n)$ and $u(n)$, respectively.

Define $\lambda_y(l, k) = \mathbb{E}\left[|Y(l, k)|^2\right]$ to be the short-time spectral variance of the measured signal. Assuming the speech, the transient interference, and the stationary noise are mutually uncorrelated, the spectral variance of the measurement is given by

$$\lambda_y(l, k) = \lambda_x(l, k) + \lambda_t(l, k) + \lambda_u(l, k) \tag{2}$$

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

In this work, our objective is to estimate the clean speech signal $x(n)$ given the noisy measurements $y(n)$. The processing of the measured signal is performed sequentially in the time-frequency domain. In order to exploit the spectral structure of the transients, we collect the spectral features from all the frequency bins of each time frame into vectors. Let $\boldsymbol{\lambda}_y(l)$ be a vector of the spectral variance values of the measured signal corresponding to time frame l, defined by

$$\boldsymbol{\lambda}_y(l) = \left[\lambda_y(l, 0), \ldots, \lambda_y(l, N - 1)\right]^T \tag{3}$$

and let $\boldsymbol{\lambda}_t(l)$ be a vector of spectral variance values of the transient signal, defined similarly as

$$\boldsymbol{\lambda}_t(l) = \left[\lambda_t(l, 0), \ldots, \lambda_t(l, N - 1)\right]^T . \tag{4}$$

As described in the introduction, our focus is on estimating the spectral variance of the transient interference. Given a new time frame of measurements, our objective is to estimate $\boldsymbol{\lambda}_t(l)$ based on $\boldsymbol{\lambda}_y(l)$. Then, the estimated spectrum is used for enhancing the speech.

Suppose a training recording of a typical transient signal $\bar{t}(n)$ is available in advance[1]. The recording comprises a collection of transient instances representing the various possible types which are assumed to be known a-priori. The training recording is processed in the time-frequency domain using the STFT with the same analysis and synthesis windows and the same time shift. Let $\bar{\boldsymbol{\lambda}}_t(\bar{l},k)$ be the spectral variance of the training recording, and let $\bar{M}$ be the number of available training time frames. Similarly to (3) and (4) we define

$$\bar{\boldsymbol{\lambda}}_t(\bar{l}) = \left[\bar{\lambda}_t(\bar{l},0),\ldots,\bar{\lambda}_t(\bar{l},N-1)\right]^T . \tag{5}$$

Each of the vectors can be viewed as an N-dimensional point. Collecting all the vectors yields a set $\left\{\bar{\boldsymbol{\lambda}}_t(\bar{l})\right\}_{\bar{l}}$ of $\bar{M}$ training points in an N-dimensional space.

Let N_t be the number of transient types in the training recording, and let $\overline{\mathcal{T}}_i$ be the set of training time frame indices containing the ith type. We assume no more than a single transient exists in one time frame which implies that $\overline{\mathcal{T}}_i \cap \overline{\mathcal{T}}_j = \emptyset$ for $i \neq j$. In addition, we assume the duration of each transient event is shorter than a single short-time frame. Longer transient interferences are broken into separate sets and considered as few transient types. In [9], the examination of a wide variety of transient interferences led us to the observation that each transient event consists of an abrupt sound followed by decaying oscillations. Thus, in [9], a transient is modeled as a composition of two parts - abrupt and decaying. In this work, each part is treated independently as a different type of transient. Let $\overline{\mathcal{T}} = \overline{\mathcal{T}}_1 \oplus \cdots \oplus \overline{\mathcal{T}}_{N_t}$ denote the set of training time frames indices containing any transient interference. The remaining time frames of the training recording are silent.

III. Graph-Based Processing

A. Graph Construction

Following [18] [19] [20], we define a "one-sided" kernel consisting of an affinity measure between the observed data points and the training points. Let M be the number of available observation time frames. In Section V we extend the following derivation to support sequential processing where the observations are not available in advance. Let $\mathbf{W}$ be an $M \times \bar{M}$ kernel matrix defined using a Gaussian as

$$\mathbf{W}_{l,\bar{l}} = \exp\left\{-\frac{\left\|\log(\boldsymbol{\lambda}_y(l)) - \log(\bar{\boldsymbol{\lambda}}_t(\bar{l})) - \boldsymbol{\eta}\right\|^2}{2\sigma^2}\right\} \tag{6}$$

where σ^2 is the variance and $\boldsymbol{\eta}$ is a constant vector. For simplicity, $\log(\mathbf{x})$ denotes a pointwise logarithm operation on the coordinates of the vector $\mathbf{x}$. We operate in the logarithmic domain because empirical

[1] For simplicity, in the remainder of the paper we denote with a bar all the terms associated with the training recording.

experiments show better results than the linear domain. As in many speech processing applications in the logarithmic domain, small values are clipped. For simplicity, the clipping is omitted from the derivation. The presence of the unusual constant $\boldsymbol{\eta}$ becomes apparent in Section III-B, where we discuss its role and describe how to determine its value.

The one-sided kernel defines a bipartite graph [21], where $\left\{\bar{\boldsymbol{\lambda}}_t(\bar{l})\right\}_{\bar{l}}$ and $\{\boldsymbol{\lambda}_y(l)\}_l$ are the two disjoint sets of nodes, and $\mathbf{W}_{l,\bar{l}}$ determines the weight of the edge connecting $\boldsymbol{\lambda}_y(l)$ and $\bar{\boldsymbol{\lambda}}_t(\bar{l})$. We normalize the one-sided kernel to create a transition matrix of a Markov process on the graph, i.e., $\mathbf{A} = \mathbf{D}^{-1}\mathbf{W}$ with $\mathbf{D}$ a diagonal matrix defined by $\mathbf{D}_{l,l} = \sum_{\bar{l}=1}^{\bar{M}} \mathbf{W}_{l,\bar{l}}$. Accordingly, $\mathbf{A}_{l,\bar{l}}$ is the transition probability in a single step from node $\boldsymbol{\lambda}_y(l)$ to node $\bar{\boldsymbol{\lambda}}_t(\bar{l})$.

Let $\bar{\mathbf{K}}$ be a "two-sided" kernel of size $\bar{M} \times \bar{M}$ defined on the training nodes by $\bar{\mathbf{K}} \triangleq \mathbf{A}^T\mathbf{A}$. According to the definition, each component of the two-sided kernel is given by

$$\bar{\mathbf{K}}_{\bar{l},\bar{l}'} = \sum_{l=1}^{M} \mathbf{A}_{l,\bar{l}}\mathbf{A}_{l,\bar{l}'}.$$

Thus, $\bar{\mathbf{K}}_{\bar{l},\bar{l}'}$ can be interpreted as an affinity metric between a training node $\bar{\boldsymbol{\lambda}}_t(\bar{l})$ and a training node $\bar{\boldsymbol{\lambda}}_t(\bar{l}')$ via any observable node $\boldsymbol{\lambda}_y(l)$.

Similarly, $\mathbf{K}$ is a "two sided" kernel of size $M \times M$ defined on the observed points by $\mathbf{K} \triangleq \mathbf{A}\mathbf{A}^T$, i.e.,

$$\mathbf{K}_{l,l'} = \sum_{\bar{l}=1}^{\bar{M}} \mathbf{A}_{l,\bar{l}}\mathbf{A}_{l',\bar{l}}.$$

Then, $\mathbf{K}_{l,l'}$ can be interpreted as an affinity metric between an observed node $\boldsymbol{\lambda}_y(l)$ and an observed node $\boldsymbol{\lambda}_y(l')$ via any training node $\bar{\boldsymbol{\lambda}}_t(\bar{l})$. It further implies that two observations are similar if they "see" the training points in the same way.

B. Probabilistic Interpretation

Suppose that the transient part in the observation at time frame l equals to one of the training points, i.e., $\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l})$. By (2) we have for every frequency bin k

$$\log\left(\lambda_y(l,k)\right) - \log\left(\bar{\lambda}_t(\bar{l},k)\right) = \log\left(1 + \frac{\lambda_x(l,k) + \lambda_u(l,k)}{\lambda_t(l,k)}\right) > 0.$$

Our experiments show that the empirical probability density function of the right hand term has a single peak. We observe that the peak (mean) is located remotely from zero, and the empirical probability density function is almost symmetric. Thus, we approximate the probability density function by a normal distribution with $\boldsymbol{\eta}$ mean and σ^2 variance, such that the negative tail is negligible. The values of the mean and variance can then be determined according to the empirical mean and variance of the set $\{\log(\boldsymbol{\lambda}_y(l)) - \log(\boldsymbol{\lambda}_t(l))\}$. Accordingly,

$$\begin{aligned}&\Pr\left(\log(\boldsymbol{\lambda}_y(l))|\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\|\log(\boldsymbol{\lambda}_y(l)) - \log(\bar{\boldsymbol{\lambda}}_t(\bar{l})) - \boldsymbol{\eta}\|^2}{2\sigma^2}\right\}. \qquad (7)\end{aligned}$$

We further assume that the transient signal in the observation and the training transient signal have similar distributions. In addition, we assume that the spectral feature vector of the transient signal in each time frame can uniformly take one of a finite set of spectral feature vectors of cardinality γ (as each transient type has a distinct characteristic structure), i.e., $\Pr\left(\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right) = 1/\gamma$. By the law of total probability we obtain

$$\Pr\left(\log(\boldsymbol{\lambda}_y(l))\right) = \frac{1}{\gamma}\sum_{\bar{l}}\Pr\left(\log(\boldsymbol{\lambda}_y(l))|\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right). \qquad (8)$$

We assume statistically independent frames neglecting potential frame overlap. This assumption is not respected in practice, especially since we use 75% overlapping frames. However, it enables us to provide a probabilistic interpretation of the kernel. The conditional joint probability of frames with the same transient component can be expressed similarly

$$\begin{aligned}&\Pr\left(\log(\boldsymbol{\lambda}_y(l)), \log(\boldsymbol{\lambda}_y(l'))\,\middle|\,\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l')\right)\\ &= \frac{1}{\gamma}\sum_{\bar{l}}\Pr\left(\log(\boldsymbol{\lambda}_y(l)), \log(\boldsymbol{\lambda}_y(l'))\,\middle|\,\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l') = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right)\\ &= \frac{1}{\gamma}\sum_{\bar{l}}\Pr\left(\log(\boldsymbol{\lambda}_y(l))\,\middle|\,\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right)\\ &\times\Pr\left(\log(\boldsymbol{\lambda}_y(l'))\,\middle|\,\boldsymbol{\lambda}_t(l') = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right). \qquad (9)\end{aligned}$$

A significant benefit from this particular kernel is expressed by the following proposition.

Proposition 1 (Probabilistic Interpretation)**.** *Under the probabilistic assumptions (7), (8), and (9), the elements of the kernel satisfy*

$$\mathbf{K}_{l,l'} = \Pr\left(\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l')\,\middle|\,\boldsymbol{\lambda}_y(l), \boldsymbol{\lambda}_y(l')\right)$$

Proof: See Appendix I. ■

Proposition 1 implies that the affinity metric defined by the kernel is the probability of comparing a pair of observable vectors with the same transient pattern. Accordingly, this kernel entails a comparison between the underlying spectral features of the transients “neutralizing” the speech and background noise. This way, the constructed graph may convey the desired transient interference spectral structure.

C. Graph-based Filter

Let $\{\mu_j, \boldsymbol{\psi}_j\}_j$ be the eigenvalue decomposition of $\mathbf{K}$, which satisfies

$$\mathbf{K} = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^T \tag{10}$$

with

$$\boldsymbol{\Psi} = \left[\boldsymbol{\psi}_0 \cdots \boldsymbol{\psi}_{M-1}\right]$$

and $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of the eigenvalues in a descending order $\mu_0 \geq \mu_1 \geq \ldots > 0$. Each eigenvector $\boldsymbol{\psi}_j$ is of length M and its lth coordinate parameterizes the lth time frame. By the orthogonality of the eigenvectors, the set $\{\boldsymbol{\psi}_j\}_j$ forms a complete basis for any function $f : \Gamma \to \mathbb{R}$ with $\Gamma = \{\boldsymbol{\lambda}_y(l)\}_l$. In particular, let $i_k : \Gamma \to \mathbb{R}$ be a function that retrieves the kth frequency bin from the spectral vector $\boldsymbol{\lambda}_y(l)$, i.e., $i_k\left(\boldsymbol{\lambda}_y(l)\right) = \lambda_y(l,k)$. It implies that each spectral component can be expanded according to the set of eigenvectors as

$$\lambda_y(l,k) = i_k\left(\boldsymbol{\lambda}_y(l)\right) = \sum_{j=0}^{M-1} \mu_j \langle i_k, \boldsymbol{\psi}_j \rangle \boldsymbol{\psi}_j(l)$$

where the inner product is defined as $\langle i_k, \boldsymbol{\psi}_j \rangle \triangleq \boldsymbol{\lambda}_y^f(k)\boldsymbol{\psi}_j$ with $\boldsymbol{\lambda}_y^f(k) = [\lambda_y(1,k), \ldots, \lambda_y(M,k)]$.

The constructed graph captures the structure of the transients, characterized by a distinct spectral structure, by connecting similar spectral observations. Specifically, as implied by Proposition 1, strong connections represent a high probability that the same transient pattern exits in the connected time frames. Consequently, there exists a subset of ℓ eigenvectors which represents the transient interference. For simplicity, we assume this subset consists of the dominant eigenvectors, i.e., $\{\boldsymbol{\psi}_j\}_{j=0}^{\ell-1}$. In practice, we may determine the appropriate eigenvectors by observing their spectral structure.

We define the following graph-based filter that approximates the transient spectral component by projecting the spectral variance of the observation onto the eigenvectors spanning the transient interference subspace

$$\hat{\lambda}_t(l,k) = \sum_{j=0}^{\ell-1} \mu_j \langle i_k, \boldsymbol{\psi}_j \rangle \boldsymbol{\psi}_j(l). \tag{11}$$

Let $\boldsymbol{\lambda}_y$ be an $M \times N$ matrix where its (l,k)th element is defined as $\lambda_y(l,k)$. Then (11) can be re-written in a matrix form as

$$\hat{\boldsymbol{\lambda}}_t(l) = \boldsymbol{\lambda}_y^T \sum_{j=0}^{\ell-1} \mu_j \boldsymbol{\psi}_j(l) \boldsymbol{\psi}_j. \tag{12}$$

In practice, few speech "leftovers" may appear in the estimated spectral variance. Human speech consists of both harmonic and nonharmonic sounds and it can span across a wide range of frequencies.

Thus, many speech phonemes can be represented (at least partially) by the transients "building blocks". Existence of such residuals in the spectral variance estimate of the transient signal degrades the quality of the speech when incorporated into an enhancement algorithm. Since the leftovers usually exist in periods where the transient signal is absent, we are able to easily distinct them by their low magnitude compared to the magnitude of the transients. Thus, we remove potential leftovers by employing a hard threshold.

D. Speech Enhancement

To enhance the speech, we employ the OM-LSA with a modified noise estimate. 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 obtained by

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

In [16], the optimal spectral gain with respect to the minimum log spectral amplitude (LSA) error criterion is controlled by the speech presence probability. Since it is unknown, the speech presence probability is estimated based on the timefrequency distribution of the a-priori signal-to-noise ratio (SNR), where the noise variance is estimated using the improved minima controlled recursive averaging (IMCRA) [22]. Unfortunately, short and abrupt bursts of transient interferences are falsely detected as speech components. Hence, the transient interference is not a part of the noise PSD estimate obtained by the IMCRA approach, and as a result, is not attenuated. In this work, we set the optimal spectral gain to correspond to the sum of the spectral variance estimate of the transient interference $\hat{\lambda}_t(l,k)$ and the stationary noise $\hat{\lambda}_u(l,k)$. The former estimate is obtained by the graph-based filter (11) following the hard thresholding, and the latter estimate is obtained by the IMCRA. The IMCRA and the OM-LSA parameters used in this stage are similar to the set of parameters used to enhance speech and reduce stationary background noise as described in [16].

Since the optimal spectral gain is controlled by the transient interference spectrum, the suppression of transients is now attainable. For more details regarding the optimal gain derivation and estimation of the speech presence probability and the noise spectrum, we refer the reader to [16] and references therein. A Matlab code of the OM-LSA is available at [23].

IV. Transient Local Models and an Affinity Function

The estimation of the spectral variance of the transient interference is employed by the graph-based filter defined in (11). Thus, the estimation accuracy heavily depends on the ability of the graph to extract the structure of the spectral variance of the transients. As discussed in Section III and implied by Proposition

1, the graph connects nodes with the same transient type. In order to enhance this property, we define a local data-driven model for each transient interference type based on the training recording. We assume the labeling of the transient recording $\left\{\overline{\mathcal{T}}_i\right\}_{i=1}^{N_t}$ is available. Let $\left\{\bar{\boldsymbol{\lambda}}_t(\bar{l})\right\}_{\bar{l}\in\overline{\mathcal{T}}_i}$ be the set of training spectral vectors corresponding to the ith transient type. We assume it consists of several transient events which define the variability of the transient type. Let $\bar{\boldsymbol{\eta}}_i$ be the empirical mean vector of the set, i.e.,

$$\bar{\boldsymbol{\eta}}_i = \frac{1}{\left|\overline{\mathcal{T}}_i\right|} \sum_{\bar{l}\in\overline{\mathcal{T}}_i} \log\left(\bar{\boldsymbol{\lambda}}_t(\bar{l})\right)$$

and let $\bar{\mathbf{C}}_i$ be the empirical covariance matrix of the set

$$\bar{\mathbf{C}}_i = \frac{1}{\left|\overline{\mathcal{T}}_i\right|} \sum_{\bar{l}\in\overline{\mathcal{T}}_i} \left(\log\left(\bar{\boldsymbol{\lambda}}_t(\bar{l})\right) - \bar{\boldsymbol{\eta}}_i\right) \left(\log\left(\bar{\boldsymbol{\lambda}}_t(\bar{l})\right) - \bar{\boldsymbol{\eta}}_i\right)^T$$

where $\left|\overline{\mathcal{T}}_i\right|$ is the cardinality of the set $\overline{\mathcal{T}}_i$. The pair $(\bar{\boldsymbol{\eta}}_i, \bar{\mathbf{C}}_i)$ may be used as the learned model of the ith transient type. This implicit Gaussian representation is set for simplicity and was previously used in [13] and [24]. This assumption is supported by the fact that the logarithm has support for both negative and positive values. By employing principal component analysis (PCA), the large eigenvectors of $\bar{\mathbf{C}}_i$, which correspond to the principal "parameters", capture most of the information disclosed in the data. Hence, the dimensionality is significantly reduced by considering only the subspace spanned by a few principal eigenvectors. Let $\{\bar{\mathbf{v}}_{i,j}\}_{j=1}^{L}$ be the set of L such principal eigenvectors. A well-known limitation of PCA is that it is linear and able to capture only the global structure of the training data. The training set of transient instances admits a complicated global structure (often referred to as a non-linear manifold). Thus, a low-dimensional linear subspace may not faithfully describe the data in our setting. However, a PCA-based approach may perform rather well when applied locally, i.e., on a data set sufficiently condensed in a small neighborhood. In our setting, this corresponds to defining a model for each transient interference type. Then, incorporating these local models in the graph provides integration of all the acquired models together.

We define P_i to be a linear projection operator of each spectral feature vector onto the local model of the ith transient type as

$$P_i(\boldsymbol{\lambda}_y(l)) = \bar{\boldsymbol{\eta}}_i + \sum_{j=1}^{L} \langle \log\left(\boldsymbol{\lambda}_y(l)\right) - \bar{\boldsymbol{\eta}}_i, \bar{\mathbf{v}}_{i,j}\rangle \bar{\mathbf{v}}_{i,j} \tag{13}$$

where the inner product is defined as $\langle \log\left(\boldsymbol{\lambda}_y(l)\right) - \bar{\boldsymbol{\eta}}_i, \bar{\mathbf{v}}_{i,j}\rangle \triangleq \left(\log\left(\boldsymbol{\lambda}_y(l)\right) - \bar{\boldsymbol{\eta}}_i\right)^T \bar{\mathbf{v}}_{i,j}$. The linear projection (13) can be used as a stand-alone estimator for the spectral variance of the transients. In practice it does not yield satisfactory results. However, it provides essential information which may be incorporated into the graph construction. The graph provides integration of all transient instances and their local models together. Capitalizing the connections between the entire set of data, rather than using a single local model, attains significantly improved results.

Based on the projection, we define a pairwise metric between spectral feature vectors for each transient type

$$d_i\left(\boldsymbol{\lambda}_y(l), \boldsymbol{\lambda}_y(l')\right) = \left\| P_i(\boldsymbol{\lambda}_y(l)) - P_i(\boldsymbol{\lambda}_y(l')) \right\|. \tag{14}$$

The definition of the local metric (14) enables to adjust the kernel computation in (6). We now define the affinity kernel as

$$\begin{aligned} \mathbf{W}_{l,\bar{l}} = \exp&\left\{ -\frac{\left\| \log(\boldsymbol{\lambda}_y(l)) - \log(\bar{\boldsymbol{\lambda}}_t(\bar{l})) - \boldsymbol{\eta} \right\|^2}{2\sigma^2} \right\} \\ \times \exp&\left\{ -\frac{d_i^2\left(\boldsymbol{\lambda}_y(l), \bar{\boldsymbol{\lambda}}_t(\bar{l})\right)}{2\tilde{\sigma}^2} \right\}. \end{aligned} \tag{15}$$

for $\bar{l} \in \overline{\mathcal{T}}_i$ with scale $\tilde{\sigma}^2$ corresponding to the values of d_i. The first term ensures that the kernel is defined locally by comparing the spectral features of the vectors. The second term conveys the affinity of the observable vector to the training vector in terms of the ith transient interference type. Consequently, two vectors are similar if their underlying transient is similar and the observable speech component does not distort the transient significantly. The remainder of the graph construction, namely, the computation of the transition matrix $\mathbf{A}$ and the kernels $\bar{\mathbf{K}}$ and $\mathbf{K}$, remains unaltered. Compared to the kernel defined in (6), the new kernel enhances the connection between time frames that consist of transient events. Consequently, the spectral representation of the constructed graph better captures the transient structure, and the estimation of the transient spectral variance in (11) becomes more accurate. Experimental results show improved transient extraction and speech enhancement using the adjusted local kernel (15) compared with (6).

V. IMPLEMENTATION

We start by drawing the algebraic connection between the eigen-decomposition of the kernels $\mathbf{K}$ and $\bar{\mathbf{K}}$.

Proposition 2. *The kernels $\mathbf{K}$ and $\bar{\mathbf{K}}$ share the same eigenvalues μ_j. The eigenvector $\boldsymbol{\psi}_j$ of $\mathbf{K}$ corresponding to nonzero eigenvalues $\mu_j > 0$ satisfies*

$$\boldsymbol{\psi}_j = \frac{1}{\sqrt{\mu_j}} \mathbf{A} \boldsymbol{\varphi}_j$$

where $\boldsymbol{\varphi}_j$ is the eigenvector of $\bar{\mathbf{K}}$ corresponding to eigenvalue μ_j. In addition, the eigenvector sets $\left\{\boldsymbol{\varphi}_j\right\}_j$ and $\left\{\boldsymbol{\psi}_j\right\}_j$ are orthogonal.

Proof: See Appendix II. ■

The main property emerged from Proposition 2 is the natural extension of the eigenvalue decomposition. Given a training recording and an initial observation interval, the matrix $\mathbf{A}$ and the kernel $\bar{\mathbf{K}}$ can be constructed. Next, the singular value decomposition (SVD) of $\mathbf{A}$ is computed, which allows us to define the graph-based filter (11) used to estimate the spectral variance of a transient in the initial observation interval. Proposition 2 can then be applied to extend the spectral representation of the kernel matrix $\mathbf{K}$, which defines the filter, to a new observation. The extension implied by Proposition 2 is efficiently computed and can be implemented in a sequential manner based on the spectral representation of $\bar{\mathbf{K}}$ (which is computed in advance using the training data).

For each spectral feature vector $\boldsymbol{\lambda}_y(l')$ corresponding to a new time frame observation l', we have by Proposition 2 that

$$\boldsymbol{\psi}_j(l') = \frac{1}{\sqrt{\mu_j}}\mathbf{a}_{l'}^T\boldsymbol{\varphi}_j \tag{16}$$

where $\mathbf{a}_{l'}$ is a vector of length $\bar{M}$ with elements given by

$$\begin{aligned}\mathbf{a}_{l'}(\bar{l}) = \frac{1}{d_{l'}} \exp\left\{-\frac{\left\|\log(\boldsymbol{\lambda}_y(l')) - \log(\bar{\boldsymbol{\lambda}}_t(\bar{l})) - \boldsymbol{\eta}\right\|^2}{2\sigma^2}\right\}\\ \times \exp\left\{-\frac{d_i^2\left(\boldsymbol{\lambda}_y(l'), \bar{\boldsymbol{\lambda}}_t(\bar{l})\right)}{2\tilde{\sigma}^2}\right\}\end{aligned} \tag{17}$$

and where

$$\begin{aligned}d_{l'} = \sum_{\bar{l}'=1}^{\bar{M}} \exp\left\{-\frac{\left\|\log(\boldsymbol{\lambda}_y(l')) - \log(\bar{\boldsymbol{\lambda}}_t(\bar{l}')) - \boldsymbol{\eta}\right\|^2}{2\sigma^2}\right\}\\ \times \exp\left\{-\frac{d_i^2\left(\boldsymbol{\lambda}_y(l'), \bar{\boldsymbol{\lambda}}_t(\bar{l}')\right)}{2\tilde{\sigma}^2}\right\}.\end{aligned}$$

Then, the corresponding graph-based estimator based on the extended eigenvector is given by (12), i.e.,

$$\hat{\boldsymbol{\lambda}}_t(l') = \boldsymbol{\lambda}_y^T \sum_{j=0}^{\ell-1} \mu_j \boldsymbol{\psi}_j(l')\boldsymbol{\psi}_j. \tag{18}$$

The sequential estimation of the spectral variance of the transient signal via the graph-based processing framework is summarized in Algorithm 1. A particular attention should be given to the efficiency and low computational complexity of the enhancement stage of each time frame. Following is a description of the naïve computational cost (number of operations) for each step in the enhancement stage. Step 1 involves fast Fourier transform which yields $\mathcal{O}(N\log N)$ operations. Computing the affinity between the new observation and the $\bar{M}$ training vectors in Step 2 yields $\mathcal{O}(N\bar{M})$ operations, treating the number of principal local-model eigenvectors ℓ as a constant. Accordingly, Step 3 costs $\mathcal{O}(\bar{M})$ operations. Finally,

```
Algorithm 1 Graph-based Processing Algorithm
Training stage:
    1) Obtain a training recording of typical transients and compute a training set $\{\bar{\boldsymbol{\lambda}}_t(\bar{l})\}_{\bar{l}=1}^{\bar{M}}$ of $\bar{M}$ spectral variance feature vectors.
    2) Obtain an initial measurement and compute a set $\{\boldsymbol{\lambda}_y(l)\}_{l=1}^{M}$ of $M$ spectral variance feature vectors.
    3) Compute the "one-sided" kernel matrix $\mathbf{W}$ of size $M \times \bar{M}$ according to (15).
    4) Construct the transition matrix $\mathbf{A}$ of size $M \times \bar{M}$.
    5) Obtain the eigenvalue decomposition $\{\mu_j, \boldsymbol{\varphi}_j\}_j$ and $\{\mu_j, \boldsymbol{\psi}_j\}_j$ of $\bar{\mathbf{K}}$ and $\mathbf{K}$, respectively, by computing the SVD of $\mathbf{A}$.
Enhancement stage:
    1) Obtain a new time frame of the observable signal and compute the corresponding new feature vector $\boldsymbol{\lambda}_y(l')$.
    2) Compute the affinity of the new observation vector to the training vectors according to (17).
    3) By Proposition 2, extend the eigenvectors to the new frame according to (16).
    4) Construct the graph-based filter corresponding to the new frame according to (18) using the extended vector. Obtain an estimate of the spectral variance for the transient interference $\hat{\boldsymbol{\lambda}}_t(l')$.
    5) Compute the optimal gain of the OM-LSA based on $\hat{\boldsymbol{\lambda}}_t(l')$ and employ it on the new time frame to enhance the speech.
    6) Return to 1 in the Enhancement stage.
```

employing the graph-based filter in Step 4 requires $\mathcal{O}(M\bar{M})$. By assuming that $M, \bar{M} > N$, we have a total computational burden of $\mathcal{O}(M\bar{M})$. We note that this cost is mainly due to a matrix multiplication which can be implemented very efficiently.

VI. Experimental Results

A. Experimental Setup

We evaluate the performance of the proposed algorithm on recorded speech and transient signals sampled at 16 KHz. Speech signals are taken from the TIMIT database [25], and recorded transient interferences are taken from an online free corpus [26]. The time domain 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 SNR of 20 dB. Each measurement is 20 seconds long and consists of several speech utterances of 5 different speakers and 30 transient events. For the time-frequency representation, we use time frames of 512 samples length which correspond to $N = 257$ positive frequency bins. In addition, we use 75% overlap between successive frames.

We examine the suppression of three transient interference signals. The first transient interference is keyboard typing. We enhance a measurement interval containing 30 key strokes with different amplitudes. The different key strokes are organized into three clusters of similar spectral structures. Based on a training recording of similar keyboard strokes, we train three transient models corresponding to the three key stroke types as described in Section IV. The second interference consists of three types of household knocks. One of the knocks has a relatively long duration which exceeds a single time frame. Consequently, we attach two models to this interference type (one for the first abrupt part and one for the following decaying part) and another two models corresponding to the other two types of knocks, which results in four different models. The measurement signal consists of several different instances of each type with varying amplitudes. Finally, the third interference consists of three types of door knocks. Accordingly, we train three corresponding models based on the training recordings. Similarly to the other transient interferences, the measurement consists of several different instances of these door knocks with varying amplitudes. We note that each training recording consists of 10 instances of transients from each type. In addition, in order to represent the transients and define the graph-based filter (11) we use the principal $\ell = 20$ eigenvectors of the graph. For each transient interference we empirically set the parameters (kernel scale) which yield maximal performance.

B. *Performance Evaluations*

In Fig. 1 we show an example for the transient spectral variance estimation. Figure 1(a) presents the waveform and spectrogram of an instance of a door knock, and Figure 1(b) presents the waveform and spectrogram of the transient instance estimate by the graph-based filter (11). We observe similar waveform and spectral features. A particular attention should be given to the accurate estimate of the spectral "pattern" of the abrupt first part of the transient. Unfortunately, we also detect inaccurate estimation of the high frequencies in the decaying second part of the transient. The decaying part is noise-like and less structured compared to the abrupt part. Thus, it is more difficult to capture its characteristic geometry. On the other hand, it is usually of low energy and thus in practice inaccurate estimation may not have significant influence.

Figure 2 depicts the waveforms and spectrograms of the measurements and enhanced signals. Figures

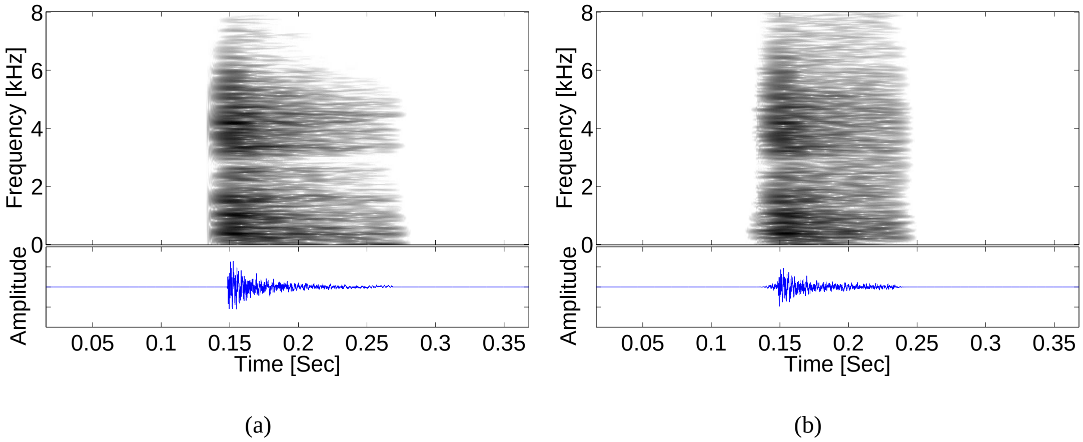
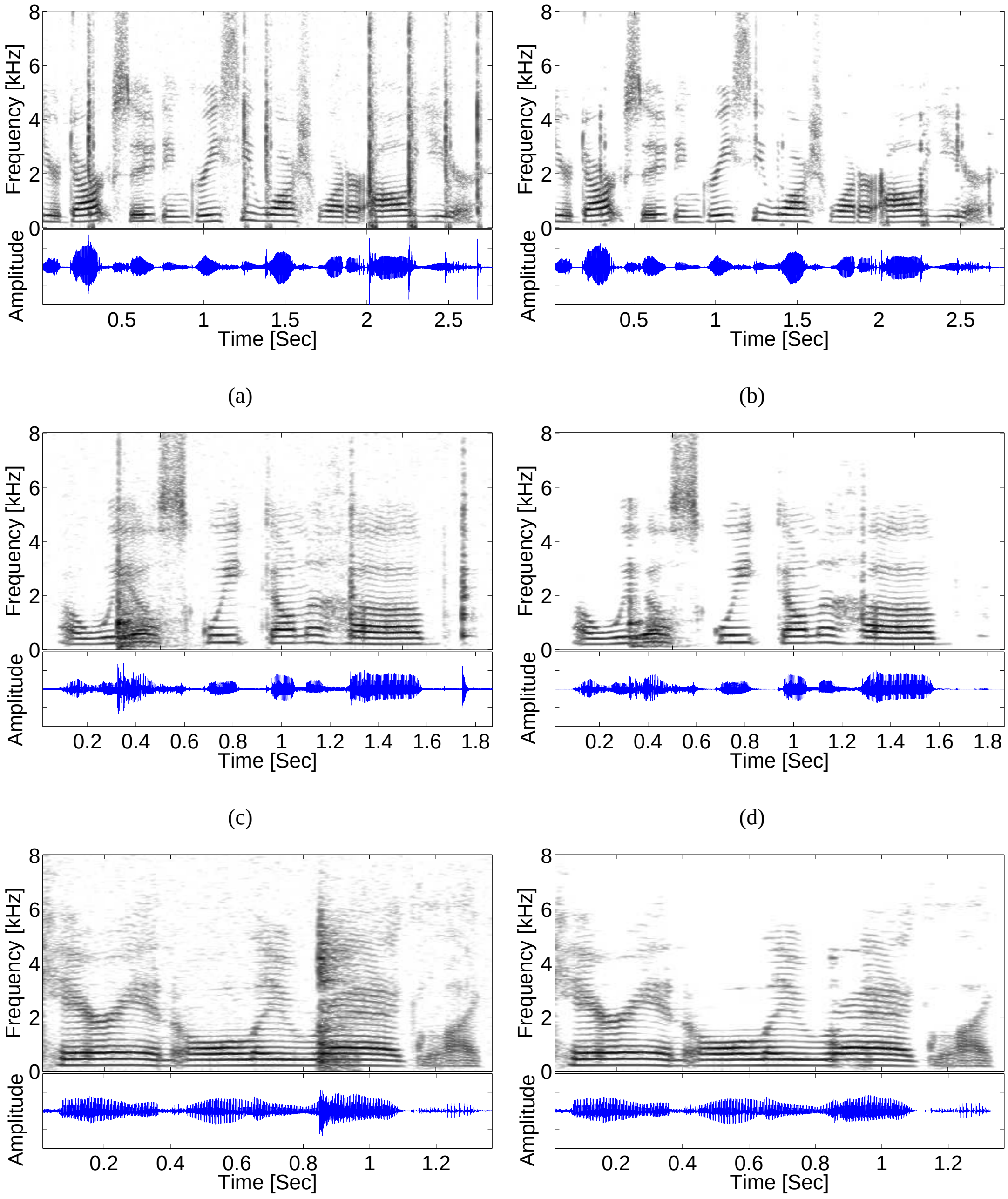


Fig. 1. Transient waveforms and spectrograms. (a) A clean transient (door knock) event. (b) The estimated transient.

2 (a), (c), and (e) show the noisy signals with keyboard typing, household interferences, and door knocks, respectively. Figures 2 (b), (d), and (f) show the corresponding enhanced signals. We observe that the proposed method attains significant transient interference reduction, while imposing very low distortion. Merely few transient residuals (e.g., near 1.3 s in Fig. 2 (b)) appear in the enhanced signal. Furthermore, the waveforms of the enhanced signals suggest that the transient suppression does not leave “holes” in the signal, but rather maintains the speech component.

We compare the performance of the proposed algorithm to the algorithm proposed in [9]. The proposed algorithm introduces two new aspects with respect to the previous work: learning transient models from training recordings and online processing, which are both incorporated into an integrated processing framework. We note that the online processing is obtained naturally given the trained models, since the employment of the models on the entire observation interval is equivalent to the employment of the models frame-by-frame. Thus, the comparison between the algorithms does not reflect the additional training stage of the proposed algorithm nor the advantage that the measurement is processed frame-by-frame. The online processing makes the proposed algorithm more adequate to communication applications. In addition, learning transient models in advance circumvents the requirement of the algorithm proposed in [9] to have several instances of transients in order to properly capture the model from the measurements. In the following experiment we expect better results using the batch algorithm in case the observation interval contains several instances of transients with similar structure and amplitude. On the other hand, the graph-based algorithm is advantageous in case of multiple transient types and in case of high variability in the amplitudes of the transients.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. A segment of the measurements and enhanced signals waveforms and spectrograms. (a) Noisy signal with 7 key strokes. (b) Enhanced speech with suppressed keyboard typing. (c) Noisy signal with 4 events of household interferences. (d) Enhanced speech with suppressed household interferences. (e) Noisy signal with a door knock. (f) Enhanced speech with suppressed door knocks.

TABLE I

SPEECH ENHANCEMENT EVALUATION.

Transient Type	SNR Improvement [dB]		LSD Improvement [dB][2]	
	Batch Algorithm Proposed in [9]	Online Graph-based Filtering	Batch Algorithm Proposed in [9]	Online Graph-based Filtering
Keyboard Typing	9.47	7.78	2.71	2.12
Household Interferences	5.20	6.62	1.83	2.04
Door Knocks	8.17	9.79	2.96	2.39

We evaluate the output of the algorithms using two objective measures [27]. The first is the common SNR, defined as

$$\begin{aligned} \mathrm{SNR}_{in} &= 10\log_{10}\frac{\mathbb{E}\left\{x^2(n)\right\}}{\mathbb{E}\left\{(y(n)-x(n))^2\right\}} \\ \mathrm{SNR}_{out} &= 10\log_{10}\frac{\mathbb{E}\left\{x^2(n)\right\}}{\mathbb{E}\left\{(\hat{x}(n)-x(n))^2\right\}} \end{aligned} \tag{19}$$

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

$$\mathrm{LSD}_{in} \triangleq \mathbb{E}_l\left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(\lambda_x(l,k))-\ell(\lambda_y(l,k))\right|^2\right]^{\frac{1}{2}} \tag{20}$$

$$\mathrm{LSD}_{out} \triangleq \mathbb{E}_l\left[\frac{1}{N}\sum_{k=0}^{N-1}\left|\ell(\lambda_x(l,k))-\ell(\hat{\lambda}_x(l,k))\right|^2\right]^{\frac{1}{2}} \tag{21}$$

where

$$\ell\left(\lambda\right)=\max\left\{10\log_{10}\lambda,\delta\right\} \tag{22}$$

and δ is a small value defined by $\delta = \max\lambda_x(l,k) - 50$, used to confine the dynamic range of the log-spectrum to 50 dB. These measures are computed only 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 I summarizes the objective evaluation of the speech enhancement algorithms. We observe improvement in all tested cases. For keyboard typing the batch algorithm indeed demonstrates better SNR

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

TABLE II

PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ) SCORES.

Transient Type	Noisy PESQ Scores	Batch Algorithm Proposed in [9] PESQ Scores Improvement	Online Graph-based Filtering PESQ Scores Improvement
Keyboard Typing	2.165	0.601	0.749
Household Interferences	2.028	0.663	0.644
Door Knocks	1.933	0.593	0.536

and LSD improvements since it exploits the presence of similar key strokes with similar amplitudes. For door knocks the proposed algorithm yields better SNR improvement whereas the batch algorithm yields better LSD improvement. Thus, no obvious advantage to any of the algorithms is reported; The repeating door knocks in the observation interval have a similar structure which may be better exploited by the batch algorithm, however, the knocks have high amplitude variability which can be better handled by the graph-based algorithm. For household interferences the proposed online algorithm outperforms the batch algorithm. In this case the noisy signal consists of multiple types of interferences with various spectral structures and with both short- and long-durations. Thus, it demonstrates the robustness and flexibility of the proposed algorithm attained by training several interference models.

Table II depicts the improvement of the perceptual evaluation of speech quality (PESQ) scores [28] with respect to the noisy signal. This measure cover a different aspect compared to Table I. We note that even a small increase in the PESQ score suggests noticeable improvement, as any sudden increase of power (e.g., attenuated transients) is audible. We observe that the speech quality is improved in all tested cases in comparison with the noisy signal. In addition, the PESQ score improvement is larger when using the proposed algorithm compared to the algorithm in [9] in case of keyboard typing, whereas it is smaller in household interferences and door knocks. This trend complements the reported results in Table I. In general, we note that milder transient suppression (conveyed by lower SNR and LSD improvements) usually leads to smaller speech distortion (conveyed by higher PESQ values).

It is worthwhile noting that informal hearing tests confirm the objective measures and demonstrate significant reduction of the transient interference. In addition, we employed the proposed algorithm on noisy speech corrupted by keyboard typing recorded in a laptop. The obtained results are comparable to the reported results on the simulated data. Audio samples of the presented results are available online in [29].

The comparison between the algorithms shows similar results where neither of the algorithms

TABLE III

SPEECH ENHANCEMENT EVALUATION IN MULTI-CONDITION CASE.

Transient Type	SNR Improvement [dB]	LSD Improvement [dB]	PESQ Score Improvement
Keyboard Typing	7.46	2.04	0.597
Household Interferences	4.72	1.69	0.418
Door Knocks	8.75	1.81	0.528

demonstrates clear advantage based on the objective measures. Thus, the preferred algorithm mainly depends on the listener preferences. However, the proposed algorithm results are achieved by online processing and demanding lower computational burden. In addition, the proposed algorithm does not introduce lag into the system. In practice, these properties make the proposed algorithm more suitable for real-time communication systems.

In Tables I and II, the reported results correspond to a matched-condition setup, where each testing sample contains a certain type of transient, and the training data that is used for applying the algorithm to the testing sample contains exactly this type of transient. This scenario is suitable for applications in which the typical transients are known in advance, e.g., keyboard typing in phone- and conference-call software. To further illustrate the applicability of the proposed algorithm under real-world conditions, we evaluate the proposed algorithm in a multi-condition training scenario. In this experiment, transients from all types are used for training a single model, which is then used to suppress all the testing samples. For a fair comparison we employed the testing stage on the same noisy recordings as in the matched-condition experiment. Table III presents the SNR and LSD improvements and the PESQ score obtained under the multi-condition case. As expected in this challenging scenario, we observe degradation in the transient suppression and speech quality compared to the matched-condition case in Tables I and II. However, the suppression of the transients and the enhancement of the speech are significant and audible. This illustrates the ability of the proposed algorithm to train a generic single model consisting of a dictionary of a wide variety of transients, which can then be suppressed from real-world recording in various scenarios.

VII. CONCLUSIONS

We have presented a supervised graph-based processing framework for sequential transient interference suppression. Based on training recordings, we propose to construct a graph that captures the intrinsic structure of the transients. Then, by relying on the graph parametrization we define a filter that extracts the transients from noisy speech measurements. The application of the filter is shown to be efficient

$$\mathbf{K}_{l,l'} = \frac{\sum_{\bar{l}} \exp\left\{-\frac{\|\log(\boldsymbol{\lambda}_y(l))-\log(\bar{\boldsymbol{\lambda}}_t(\bar{l}))-\boldsymbol{\eta}\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\|\log(\boldsymbol{\lambda}_y(l'))-\log(\bar{\boldsymbol{\lambda}}_t(\bar{l}))-\boldsymbol{\eta}\|^2}{2\sigma^2}\right\}}{\sum_{\bar{l}'} \exp\left\{-\frac{\|\log(\boldsymbol{\lambda}_y(l))-\log(\bar{\boldsymbol{\lambda}}_t(\bar{l}'))-\boldsymbol{\eta}\|^2}{2\sigma^2}\right\} \sum_{\bar{l}'} \exp\left\{-\frac{\|\log(\boldsymbol{\lambda}_y(l'))-\log(\bar{\boldsymbol{\lambda}}_t(\bar{l}'))-\boldsymbol{\eta}\|^2}{2\sigma^2}\right\}} \tag{24}$$

$$\mathbf{K}_{l,l'} = \frac{\sum_{\bar{l}} \Pr\left(\log(\boldsymbol{\lambda}_y(l))|\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right) \Pr\left(\log(\boldsymbol{\lambda}_y(l'))|\boldsymbol{\lambda}_t(l') = \bar{\boldsymbol{\lambda}}_t(\bar{l})\right)}{\sum_{\bar{l}'} \Pr\left(\log(\boldsymbol{\lambda}_y(l))|\boldsymbol{\lambda}_t(l) = \bar{\boldsymbol{\lambda}}_t(\bar{l}')\right) \sum_{\bar{l}'} \Pr\left(\log(\boldsymbol{\lambda}_y(l'))|\boldsymbol{\lambda}_t(l') = \bar{\boldsymbol{\lambda}}_t(\bar{l}')\right)} \tag{25}$$

---

and adapted to online processing, by sequentially extending the graph parametrization to newly acquired observations. To capture the underlying structure of the transients, a suitable metric is defined based on local models computed from the training recordings. Experimental results show significant transient interference suppression and low speech distortion for various transient interference types.

The ability to capture the underlying structure of training recordings and then sequentially extracting it from noisy measurements provides efficient, generic, and robust processing framework. Given sufficient training recordings, this framework may handle a wider variety of interferences, and might be naturally extended to other problems and applications.

Appendix I
Probabilistic Interpretation

Proof: By definition we have

$$\begin{aligned}\mathbf{K}_{l,l'} &= \left(\mathbf{A}\mathbf{A}^T\right)_{l,l'} = \sum_{\bar{l}} \mathbf{A}_{l,\bar{l}}\mathbf{A}_{l',\bar{l}} = \sum_{\bar{l}} \frac{\mathbf{W}_{l,\bar{l}}}{\sum_{\bar{l}'} \mathbf{W}_{l,\bar{l}'}} \frac{\mathbf{W}_{l',\bar{l}}}{\sum_{\bar{l}'} \mathbf{W}_{l',\bar{l}'}} \\ &= \frac{\sum_{\bar{l}} \mathbf{W}_{l,\bar{l}}\mathbf{W}_{l',\bar{l}}}{\sum_{\bar{l}'} \mathbf{W}_{l,\bar{l}'} \sum_{\bar{l}'} \mathbf{W}_{l',\bar{l}'}}\end{aligned} \tag{23}$$

Substituting the “one-sided” affinity function (6) into (23) yields (24). Then, by the probability assumption (7) we have (25).

Substituting (8) and (9) into (25) yields

$$\begin{aligned}\mathbf{K}_{l,l'} &= \frac{1}{\gamma}\frac{\Pr\left(\log(\boldsymbol{\lambda}_y(l)), \log(\boldsymbol{\lambda}_y(l'))\,|\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l')\,\right)}{\Pr\left(\log(\boldsymbol{\lambda}_y(l))\right)\Pr\left(\log(\boldsymbol{\lambda}_y(l'))\right)} \\ &= \frac{\Pr\left(\log(\boldsymbol{\lambda}_y(l)), \log(\boldsymbol{\lambda}_y(l'))\,|\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l')\,\right)}{\Pr\left(\log(\boldsymbol{\lambda}_y(l)), \log(\boldsymbol{\lambda}_y(l'))\right)} \\ &\quad \times \Pr\left(\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l')\right).\end{aligned}$$

Finally, by Bayes’ theorem we obtain

$$\mathbf{K}_{l,l'} = \Pr\left(\boldsymbol{\lambda}_t(l) = \boldsymbol{\lambda}_t(l')\,\middle|\boldsymbol{\lambda}_y(l), \boldsymbol{\lambda}_y(l')\right) \tag{26}$$

■

Appendix II
Eigen-decomposition Connection

Proof: By the definition of the kernels, namely $\mathbf{K} = \mathbf{A}\mathbf{A}^T$ and $\bar{\mathbf{K}} = \mathbf{A}^T\mathbf{A}$, we obtain (1) the left singular vectors of $\mathbf{A}$ are the eigenvector ψ_j of $\mathbf{K}$; (2) the right singular vectors of $\mathbf{A}$ are the eigenvectors φ_j of $\bar{\mathbf{K}}$; (3) the nonzero singular values of $\mathbf{A}$ are the square roots of the eigenvalues μ_j of either $\mathbf{K}$ or $\bar{\mathbf{K}}$. According to the singular value decomposition, it implies that $\mathbf{K}$ and $\bar{\mathbf{K}}$ share the same eigenvalues and the sets $\left\{\varphi_j\right\}_j$ and $\left\{\psi_j\right\}_j$ are orthogonal. Moreover, we obtain

$$\mathbf{A}\varphi_j = \sqrt{\mu_j}\psi_j$$

which yields

$$\psi_j = \frac{1}{\sqrt{\mu_j}}\mathbf{A}\varphi_j$$

■

Acknowledgment

The authors thank Prof. Amit Singer for helpful discussions. They also 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 Transaction on Audio, Speech and Language Processing*, 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,” *Proc. 35th IEEE Internat. Conf. Acoust. Speech and Signal Process. (ICASSP-2010), Dallas, Texas,* Mar. 2010.

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

[4] D. Barash, “A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 844–847, 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 Processing Letters*, 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 Journal on Imaging Sciences*, vol. 2, no. 1, pp. 118–139, 2009.

[9] R. Talmon, I. Cohen, and S. Gannot, “Single-channel transient interference suppression using diffusion maps,” *to appear in IEEE Transaction on Audio, Speech and Language Processing*, May 2012.

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

[11] R. Coifman and S. Lafon, “Diffusion maps,” *Applied and Computational Harmonic Analysis*, vol. 21, pp. 5–30, Jul. 2006.

[12] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Tran. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 1–12, Jan. 2007.

[13] K.W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” *Proc. 33th IEEE Internat. Conf. Acoust. Speech and Signal Process. (ICASSP-2008), Las Vegas, NV*, vol. 14, pp. 4029 – 4032, Mar. 2008.

[14] N. Mohammadiha, T. Gerkmann, and A. Leijon, “A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization,” *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’11)*, 2011.

[15] A. Haddad, D. Kushnir, and R. R. Coifman, “Filtering via a reference set,” *Technical Report YALEU/DCS/TR-1441*, Feb. 2011.

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

[17] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[18] D. Kushnir, A. Haddad, and R. Coifman, “Anisotropic diffusion on sub-manifolds with application to earth structure classification,” *Applied and Computational Harmonic Analysis*, vol. 32, no. 2, pp. 280–294, 2012.

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

[20] R. Talmon, I. Cohen, and S. Gannot, “Supervised source localization using diffusion kernels,” *Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics, WASPAA-2011, New Paltz, NY*, 2011.

[21] A. Bondy and U.S.R. Murty, *Graph Theory*, Springer, 2008.

[22] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

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

[24] T. Koren, R. Talmon, and I. Cohen, “Supervised system idenfication based on local PCA models,” *Proc. 37rd IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2012, Kyoto, Japan*, Mar. 2012.

[25] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database,” National Inst. of Standards and Technology (NIST), Gaithersburg, MD, Feb 1993.

[26] *[Online]. Available: http://www.freesound.org*.

[27] S. R. Quachenbush, T. P. Barnwell III, and M. A. Clements, *Objective measures of speech quality*, Prentice Hall, 1988.

[28] “Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” *Tech. Rep. ITU-T P.862*, 2001.

[29] *[Online]. Available: http://users.math.yale.edu/rt294/*.

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

From 2000 to 2005, he was a software developer and researcher at a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (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 Irwin and Joan Jacobs Fellowship for 2011, the Viterbi Fellowship for 2011-2012, 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) received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion – Israel Institute of Technology, Haifa, Israel, 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, 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, 20010), 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 is an Associate Professor in the Faculty of Engineering at Bar-Ilan University, Israel. He received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, 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) 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.

Dr. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for the year 2010. He is a coeditor of the Speech Enhancement section of the Springer Handbook of Speech Processing (Springer, 2008), and a coeditor of Speech Processing in Modern Communication: Challenges and Perspectives (Springer, 2010). Dr. Gannot serves as 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 served as an Associate Editor of EURASIP Journal on Advances in signal Processing in 2003-2011, an Editor of two special issues on Multi-microphone Speech Processing of the same journal, and a guest editor of ELSEVIER Speech Communication journal. 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 was the general co-chair of IWAENC 2010 held in Tel-Aviv, Israel. He is a general co-chair of the International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) to be held in Mohonk Mountain House, New Paltz, New York in 2013.

His research interests include parameter estimation, statistical signal processing and speech processing using either single- or multi-microphone arrays and in particular, speaker extraction, noise reduction, dereverberation and speaker localization in adverse conditions.

Ronald R. Coifman is Phillips professor of mathematics at Yale University. He received his Ph.D. from the University of Geneva in 1965. Prior to coming to Yale in 1980, he was a professor at Washington University in St Louis. Prof. Coifman's recent publications have been in the areas of nonlinear Fourier Analysis, wavelet theory, numerical analysis and scattering theory. Professor Coifman is currently leading a research program to develop new mathematical tools for efficient transcription of data, with applications to feature extraction recognition, denoising, and information organization. He was chairman of the Yale mathematics department 1986-89. He is a member of the National Academy of Sciences, American Academy of Arts and Sciences, and the Connecticut Academy of Sciences and Engineering. He received the DARPA Sustained Excellence Award in 1996, the 1996 Connecticut Science Medal, the 1999 Pioneer award from the International Society for Industrial and applied Mathematics, the National Science Medal in 1999, and the Wavelet Pioneer Award in 2007.