This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Kernel Method for Voice Activity Detection in the Presence of Transients

David Dov, Ronen Talmon, *Member, IEEE* and Israel Cohen, *Fellow, IEEE*

***Abstract*—Voice activity detection in the presence of transient interferences is a challenging problem since transients are often detected incorrectly as speech by existing detectors. In this paper, we deviate from traditional approaches and take a geometric standpoint, in which the key element in obtaining an accurate voice activity detection is finding a metric that appropriately distinguishes between speech and transients. For example, speech and transients may often appear similar through the Euclidean distance when represented, e.g., by the Mel-Frequency Cepstral Coefficients, thereby resulting in incorrect speech detection. To address this challenge, we propose to use a metric based on the statistics of the signal in short temporal windows and justify its use by modeling speech and transients by their latent generating variables. These latent variables may be related to physical constraints controlling the generation of the signal, and, as such, they accurately represent the content of the signal – speech or transient. We show that the Euclidean distance between the latent variables is approximated by the proposed metric. Then, by incorporating this metric into a kernel-based manifold learning method, we devise a measure of voice activity and show it leads to improved detection scores compared with competing detectors.**

***Index Terms*—Speech processing, voice activity detection, transient noise, impulse noise, kernel**

I. Introduction

Signals measured in microphones are often contaminated with various environmental noises and interferences. The environmental conditions pose great challenges in a variety of speech processing tasks, e.g., in speech enhancement [1]–[4], voice activity detection [5]–[12] and dominant speaker identification [13]. Here, we focus on the task of voice activity detection in signals measured in a single microphone, i.e., dividing segments of the signal into speech and non-speech clusters.

To appropriately handle noisy environments, a common approach in the literature is to track the statistics of the signal by recursive averaging in short time intervals [1]–[4]. It relies on the assumption that the spectrum of the noise slowly varies in time, whereas the spectrum of speech changes quickly. Hence, sudden variations of the spectrum indicate the presence of speech. Although methods based on this statistical approach successfully distinguish speech from quasi-stationary noises, they fail in distinguishing speech from transients, which are abrupt interferences, such as, knocks, keyboard taps and office noise [14]–[17]. Since the spectrum of such transients varies in time even quicker than the spectrum of speech, transients are wrongly detected as speech using approaches based on recursive averaging.

To overcome the limitations of existing approaches, recent studies have proposed to model transients according to their geometry [16], [18]–[22]. The main assumption in these studies is that transients contain an underlying geometric structure which can be inferred from the signal measurements using manifold learning tools, e.g., those presented in [23]–[27]. In the studies presented in [16], [18], [19], the geometric structure of transients is captured and is exploited to construct an estimator of their spectrum. In turn, the estimated spectrum is incorporated into a denoising filter and is used for speech enhancement. We emphasize that while these studies deal with the estimation of the spectrum of transients, the present study focuses on the problem of distinguishing them from speech.

In [20], an improved distinction between speech and non-speech frames is obtained by a method based on clustering the noisy signal in a specifically designed low-dimensional domain. More precisely, the method is based on representing time frames of the noisy signal using the Mel-Frequency Cepstral Coefficients (MFCCs), and then building a low-dimensional representation of the signal based on local similarities between them. However, the similarities between frames are defined based on the Euclidean distance, which often induces high similarities between speech and transients in standard domains such as the MFCCs and the Short-Time Fourier Transform (STFT) [16], [20]. This results in an incorrect identification of speech and transients, as we demonstrate in this paper.

To deal with this problem, we use a modified version of the Mahalanobis distance [28], which is constructed from the signal measurements and exploits the statistics of the signal in short temporal windows. We analyze the modified Mahalanobis distance using a model of latent variables; we assume that speech and transients are driven by two independent sets of latent variables controlling their generation and refer to them as *the generating variables*. For example, the generation of the complex speech signal is controlled by the few parameters of the vocal tract [29]. The main idea underlying our approach is that comparing signal frames according to the generating variables gives rise to an accurate detection of the content of the frame, particularly, in terms of speech and transients. The challenge is that these variables are unknown and need to be inferred from the noisy signal. We show that the modified Mahalanobis distance locally approximates the Euclidean distance in a domain related to the generating variables.

A particular challenge in the problem of voice activity detection is that speech needs to be detected in frames containing both speech and transients. We found in our experiments that

The authors are with the Department of Electrical Engineering, The Technion-Israel Institute of Technology, Haifa 32000, Israel (e-mail: davidd@tx.technion.ac.il; ronen@ee.technion.ac.il; icohen@ee.technion.ac.il).

This Research was supported by Qualcomm Research Fund and MAFAAT-Israel Ministry of Defense.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

transients are often more dominant than speech, i.e., speech frames containing transients tend to be more similar to frames containing only transients, a fact that hampers the detection of speech presence/absence. The dominance of the transients is related to the high variation of their spectrum in time, to their high amplitudes, and to their typical broad bandwidth. We further show that the modified Mahalanobis distance mitigates the problem and reduces the dominance of the transients by implicitly exploiting the respective difference in the rate of variations of speech and transients [15], [16], [30].

By incorporating the modified Mahalanobis distance in a kernel based manifold learning method, we propose an algorithm for voice activity detection. Since this metric approximates the Euclidean distance between the generating variables, the eigenvectors of the kernel provide a parameterization of the signal in terms of the generating variables, which is viewed as the canonical representation of the signal. We show that this canonical representation improves the distinction between speech and transients compared with the representation obtained using the Euclidean distance. In addition, the canonical representation enables us to define a simple measure of voice activity, which outperforms competing detectors.

It is worthwhile noting that classical voice activity detectors (VAD), such as those presented in [5]–[12], are originally designed to detect speech in the presence of (quasi-) stationary background noise. Based on the assumption that the spectrum of speech rapidly varies compared to the spectrum of (quasi-) stationary background noise, such algorithms detect speech by tracking rapid variations in the spectrum of the noisy signal. In the presence of transients, whose spectrum also rapidly varies over time, such algorithms successfully distinguish the background noise from both speech and transients, but they cannot distinguish between speech and transients. Consequently, in this paper, we focus on distinguishing between speech and transients; in practice, a classical VAD may be applied as a pre-processing stage to distinguish time intervals containing only background noise from both speech and transients.

The remainder of the paper is organized as follows. In Section II, we formulate the problem of voice activity detection. In addition, we present a metric based on the statistics of the signal in short temporal windows, and to justify its use, we propose a model of latent generating variables. Based on this model, we show in Section III that the metric reduces the effect of transients. Using this metric, an algorithm for voice activity detection is introduced in Section IV, and experimental results demonstrating the superior performance of the proposed algorithm are presented in Section V.

II. Problem Formulation

A. The Problem of Voice Activity Detection

Consider a speech signal obtained in a single microphone in the presence of transients and processed in frames. Let $\mathbf{y}_n \in \mathbb{R}^L$ be a feature representation of frame n; in particular, we use the MFCCs such that L is the number of coefficients. The MFCCs are widely used features for speech representation based on the perceptually meaningful Mel-frequency scale [31]. They represent the spectrum of the signal in a compact form and they were previously exploited both for speech recognition [32], [33] and for voice activity detection [34]. We consider a setup where a sequence of N frames $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N$ is available in advance. Assume that the sequence comprises frames where speech is present and frames where it is absent, and let $\mathcal{H}_1^x$ and $\mathcal{H}_0^x$ be two hypotheses representing the presence and the absence of speech, respectively, where x denotes speech. Based on the hypotheses, we define a speech indicator for frame n, which is denoted by $\mathbb{1}_n^x$ and is given by:

$$\mathbb{1}_n^x = \left\{ \begin{array}{ll} 1; & n \in \mathcal{H}_1^x \\ 0; & n \in \mathcal{H}_o^x \end{array} \right\}. \tag{1}$$

The objective in this study is to estimate the speech indicator, i.e., to cluster the sequence according to the two hypotheses.

Similarly to the hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_0^x$, let $\mathcal{H}_1^t$ and $\mathcal{H}_0^t$ be hypotheses of the presence and the absence of transients, respectively, where t denotes transients. We note that frames containing both speech and transients, for which both hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold, are considered as speech frames for the purpose of voice activity detection. Nevertheless, transients are typically more dominant than speech, e.g., due to higher amplitudes and broader bandwidth. Accordingly, the MFCCs of frames containing both speech and transients often appear similar to the MFCCs of frames containing only transients, and, as a result, they are often wrongly identified as we show in Section V. One approach for improving the clustering is to design features, for which the Euclidean distance better distinguishes between the content of the frames of the signal. In this study, we take a different approach and propose to use a different metric instead of the Euclidean distance. Specifically, we propose to measure distances between frames of the signal using a modified Mahalanobis distance, proposed in [28], which is given by:

$$\|\mathbf{y}_n - \mathbf{y}_m\|_M^2 \triangleq \frac{1}{2} \left(\mathbf{y}_n - \mathbf{y}_m\right)^T \left(\mathbf{C}_n^{-1} + \mathbf{C}_m^{-1}\right) \left(\mathbf{y}_n - \mathbf{y}_m\right), \tag{2}$$

where $\mathbf{C}_n \in \mathbb{R}^{L \times L}$ and $\mathbf{C}_m \in \mathbb{R}^{L \times L}$ are the covariance matrices of $\mathbf{y}_n$ and $\mathbf{y}_m$, respectively. The covariance matrices are assumed to be known throughout this section and throughout Section III; in Section V, we describe their estimation from a short time window of samples. The modified Mahalanobis distance was previously presented in [35] for the purpose of solving the problem of non-linear independent component analysis, in which the assumption is that the observable signal is generated by independent latent stochastic dynamical processes. However, these processes are assumed to smoothly evolve in time, i.e., the current state of the process is correlated with previous states. Therefore, such processes cannot properly model transitions between speech presence and absence. Hence, to justify the use of the modified Mahalanobis distance in (2) for voice activity detection, we propose in Section II-B to model the noisy signal using latent variables controlling its generation. By assuming a simplifying statistical model for the generating variables, we show in Section III that the modified Mahalanobis distance approximates a weighted Euclidean distance between the variables, which properly respects the content of the noisy signal.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

B. The Model of The Generating Variables

The generation of many signals can be associated with a small set of physical constraints controlling their production. For example, the generation of speech is controlled by the position of the vocal tract and by the movement of lips, jaw, and tongue [29]. Here, we assume that the measured signal is modeled by two sets of unknown latent variables associated with the generation of speech and the transients. Let $\boldsymbol{\theta}_n^x \in \mathbb{R}^{d^x}$ and $\boldsymbol{\theta}_n^t \in \mathbb{R}^{d^t}$ be two vectors of generating variables underlying the speech signal and the transients in frame n, where d^x and d^t are the number of the variables, respectively. The vector of all generating variables at time frame n is denoted by $\boldsymbol{\theta}_n \in \mathbb{R}^d$, where $d \triangleq d^x + d^t$, and is given by:

$$\boldsymbol{\theta}_n = \left[\left(\boldsymbol{\theta}_n^x\right)^T, \left(\boldsymbol{\theta}_n^t\right)^T \right]^T, \tag{3}$$

where T denotes transpose. The generating variables are assumed hidden, i.e., $\boldsymbol{\theta}_n$ in (3) is not directly measured by the microphone. For example, a variable that is related to the movement of the lips during the production of speech cannot be directly captured in the microphone.

We assume that the relationship between the observable signal $\mathbf{y}_n$ and the vector of the generating variables $\boldsymbol{\theta}_n$ is given by an unknown non-linear transformation $f : \mathbb{R}^d \mapsto \mathbb{R}^L$, such that:

$$\mathbf{y}_n = f\left(\boldsymbol{\theta}_n\right). \tag{4}$$

If we had access to the generating variables, then voice activity detection would become trivial since one may ignore the variables of the transients, $\boldsymbol{\theta}_n^t$, and detect speech merely from the variables of speech, $\boldsymbol{\theta}_n^x$. However, the generating variables are not directly accessible and revealing them is challenging due to their unknown non-linear mapping f in (4) to the observable domain. Still, in the sequel, we assume a simplified model for the generating variables and the non-linear transformation f in (4), and based on this model, we show in Section III that the modified Mahalanobis distance in (2) approximates weighted Euclidean distances between frames in the domain of the generating variables. Specifically, we will show that the proposed metric reduces the effect of transients, thereby allowing improved distinction of frames containing both speech and transients from frames containing merely transients. We emphasize that in practice the generating variables are not directly estimated, but used for the analysis of the modified Mahalanobis distance in (2).

We first assume that the generating variables are statistically independent such that $\boldsymbol{\theta}_n$ has a diagonal covariance matrix. The variables of speech $\boldsymbol{\theta}_n^x$ and the variables of transients $\boldsymbol{\theta}_n^t$ are assumed independent since they are related to two independent phenomena - speech and transients. The independence between each of the variables of (say) speech, i.e., between the entries of $\boldsymbol{\theta}_n^x$, may be associated with a lack of correlation between the corresponding physical constraints. For example, the pronunciation of different parts of speech, e.g., different phonemes, is based on different combinations of the position of the vocal tract and the movement of lips, jaw, and tongue. We note that the independence between variables is a common assumption found in the literature for the analysis of latent models [35]–[37]. For example, in [36], the authors suggest a model of latent independent and identically distributed (IID) variables to provide a probabilistic interpretation of the classical Principal Component Analysis (PCA).

To encode the dominance of the transients, we assume that the generating variables of the transients have larger variances than the variables of speech. Specifically, to keep the statistical model simple, we assume that under hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$, the entries of $\boldsymbol{\theta}_n^x$ and $\boldsymbol{\theta}_n^t$ are IID, with zero mean, and $\sigma_x^2 > 0$ and $\sigma_t^2 > 0$ variances, respectively. We assume that:

$$\sigma_t^2 = r^2 \sigma_x^2, \tag{5}$$

where $r^2 > 1$ is a constant factor encoding the dominance of the transients, such that a larger r implies more dominant transients. The parameter r may be seen as related to the ratio between transients and speech. Typically, even when the transients and speech are normalized to the same maximal value, transients, due to their short duration in time, are more dominant than speech. We note that in order to show in Section III the link between the modified Mahalanobis distance and the generating variables, we do not assume specific distributions of the generating variables and they do not have to be identically distributed. In particular, the variances of the generating variables of (say) speech, i.e., the entries of $\boldsymbol{\theta}_n^x$ do not necessarily equal to the same value σ_x^2, but they are only assumed to have larger variances than the variables of speech. In Section III, we show that the modified Mahalanobis distance approximates weighted distances between the generating variables such that the weights reduce the effect of the more dominant variables, which are the transients, by assumption. Namely, we link the variances of the variables of speech and transients by a single parameter r only for the sake of simplicity. In addition, the mean value of the generating variables is set to zero merely for simplicity and it is not used explicitly in this study. Under the hypotheses $\mathcal{H}_0^x$ and $\mathcal{H}_0^t$, we simply assume that the generating variables of speech and transients equal zero, respectively. Thus, in the presence of speech only, for example, the observable signal $\mathbf{y}_n$ is related only to the generating variables of speech and not to those of the transients.

For the approximation in Section III showing the relation between the modified Mahalanobis distance and the generating variables, we consider the inverse of the function f in (4). However, we consider only frames located within a local neighborhood such that the (Euclidean) distance between them is smaller than a certain value. In such neighborhoods, we assume that the function f in (4) is smooth and *locally* invertible. Note that this assumption is significantly less restrictive than assuming a globally invertible function. In this context, we further note that in Section IV we take a data-driven approach to obtain a representation of the noisy signal based on the generating variables, by exploiting the Mahalanobis distances between the frames of the measured signal. Accordingly, the assumption that the function f in (4) is locally invertible is not strictly imposed, i.e., if it does not hold in practice, the obtained representation may be seen as the best fit of the model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

of the generating variables to the measured signal.

To facilitate the model of the generating variables, the presence of a (quasi-) stationary noise is not considered in this paper. In practice, a classical speech enhancement algorithm, e.g., the one presented in [2], may be used as a preprocessing stage to attenuate stationary noise. Such an algorithm is based on the assumption that the spectrum of the speech signal rapidly varies over time compared to the spectrum of a (quasi-) stationary noise. Hence, the stationary noise is estimated (and then attenuated) by tracking the small variations of the spectrum of the noisy signal. Since the spectrum of transients also rapidly varies over time, it is "seen" by such a speech enhancement algorithm as speech. As a result, the speech enhancement algorithm attenuates only the stationary noise while preserving speech and the transients. Accordingly, we assume that frames which do not contain speech nor transients, i.e., silent frames, are known in advance focusing on the more challenging problem of distinguishing speech from transients. Silent frames are successfully identified even in the presence of stationary noise by classical voice activity detectors, e.g., those presented in [5], [8].

III. Modified Mahalanobis Distance

In this section we show that the modified Mahalanobis distance (2) approximates the following distance:

$$\|\mathbf{y}_n-\mathbf{y}_m\|_M^2=\tfrac{1}{\sigma_x^2}\left(\|\boldsymbol{\theta}_n^x-\boldsymbol{\theta}_m^x\|^2+\tfrac{1}{r^2}\left\|\boldsymbol{\theta}_n^t-\boldsymbol{\theta}_m^t\right\|^2\right)+O\left(\|\mathbf{y}_n-\mathbf{y}_m\|^4\right), \tag{6}$$

which consists of a weighted sum of the Euclidean distances between the generating variables.

As we observed in our experiments, the main challenge in obtaining a successful clustering arises from the fact that speech components may be similar to transient components. Consider, for example, two frames, $\mathbf{y}_n$ and $\mathbf{y}_m$, one consists of only speech and the other consists of only transients. Often, small Euclidean distances are obtained between the MFCCs of such pairs of frames; as a result, these frames are not properly associated with different clusters as we demonstrate in Section V [16], [20]. However, speech and transients are assumed to have different generating variables. As a result, the Mahalanobis distance (6) between these frames is given by

$$\|\mathbf{y}_n-\mathbf{y}_m\|_M^2\approx\frac{1}{\sigma_x^2}\left(\|\boldsymbol{\theta}_n^x\|^2+\frac{1}{r^2}\left\|\boldsymbol{\theta}_m^t\right\|^2\right). \tag{7}$$

Hence, the distance between these two frames is given according to the squared norms of $\boldsymbol{\theta}_n^x$ and $\boldsymbol{\theta}_m^t$ conveying the different content of the frames, in contrast to the Euclidean distance. Assuming for simplicity that $\sigma_x=1$, this example demonstrates that the content of a frame is better represented by the Mahalanobis distance, which approximates the Euclidean distance between the generating variables, i.e., a small Mahalanobis distance between frames truly indicates that they comprise a similar content.

Another property of the Mahalanobis distance (6) stems from the re-scaling of the Euclidean distance between the generating variables of the transients by a factor of r^2. Since transient components are often more dominant than speech components due to their typical abrupt nature and large amplitudes, frames containing both speech and transients tend to be labeled as "transient" frames, i.e., $\mathcal{H}_1^t$, by typical clustering algorithms. This poses a problem for voice activity detection, where the speech presence is required to dominate the clustering. The Mahalanobis distance (7) mitigates the dominance of transients by reducing the weight of the Euclidean distance between their generating variables by a factor of $r^2>1$, thereby allowing for the design of a voice activity detector in which transients are less dominant, and frames tend more to be labeled according to their speech presence and absence, as demonstrated in Section V.

We note that the approximation in (6) holds only for short distances, where the error term $\|\mathbf{y}_n-\mathbf{y}_m\|^4$ is negligible. In Section IV we show how to obtain a global representation of the generating variables by incorporating this metric in a kernel-based manifold learning method.

To derive (6), we follow [30]. Consider the re-scaled vectors $\boldsymbol{\psi}_n^x\in\mathbb{R}^{d^x}$ and $\boldsymbol{\psi}_n^t\in\mathbb{R}^{d^t}$ defined by:

$$\boldsymbol{\psi}_n^x=\frac{\boldsymbol{\theta}_n^x}{\sigma_x} \tag{8}$$

$$\boldsymbol{\psi}_n^t=\frac{\boldsymbol{\theta}_n^t}{\sigma_t} \tag{9}$$

such that the entries of the vectors have unit variances. In addition, let $\boldsymbol{\psi}_n\in\mathbb{R}^d$ denote a vector consisting of all the re-scaled variables in the nth frame:

$$\boldsymbol{\psi}_n=\left[\left(\boldsymbol{\psi}_n^x\right)^T,\left(\boldsymbol{\psi}_n^t\right)^T\right]^T, \tag{10}$$

and let $h:\mathbb{R}^d\mapsto\mathbb{R}^L$ denote the corresponding nonlinear function that maps the re-scaled variables to the observable signal:

$$\mathbf{y}_n=h\left(\boldsymbol{\psi}_n\right). \tag{11}$$

The function h is locally invertible since we assume that the function f in (4) is locally invertible; consequently, let $g:\mathbb{R}^L\mapsto\mathbb{R}^d$ be an inverse map of h, i.e., $\boldsymbol{\psi}_n=g\left(\mathbf{y}_n\right)$. Note that for simplicity we follow [35], and, for all points $\mathbf{y}$ considered throughout the paper, we denote by $g\left(\mathbf{y}\right)$ the local inverse map of the function h even though the function h is assumed invertible only locally.

Singer et al. derived (6) in [35], [38] by using the Taylor expansions of $\boldsymbol{\psi}_n=g\left(\mathbf{y}_n\right)$ and $\boldsymbol{\psi}_m=g\left(\mathbf{y}_m\right)$ at $\mathbf{y}_m$ and $\mathbf{y}_n$, respectively, relying on the symmetry of the expansions. However, in our case, two frames $\mathbf{y}_n$ and $\mathbf{y}_m$ may consist of different signals, e.g, $\mathbf{y}_n$ may consist of only speech and $\mathbf{y}_m$ may consist of only transients, thereby breaking the symmetry between the Taylor expansions of $\boldsymbol{\psi}_n$ and $\boldsymbol{\psi}_m$.

To overcome this problem, we consider the middle point $\mathbf{y}_p$ between $\mathbf{y}_n$ and $\mathbf{y}_m$:

$$\mathbf{y}_p=\frac{\mathbf{y}_n+\mathbf{y}_m}{2}, \tag{12}$$

which does not necessarily exist in practice, but is used here merely as a reference point for the derivation. The mid-point relaxes the symmetry assumption since it contains speech or transients if they are present in one of the frames $\mathbf{y}_n$ or $\mathbf{y}_m$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

First, we focus on the hypothesis that both speech and transients are present, and then extend the derivation to all other possible hypotheses. Specifically, $\mathbf{y}_n$ and $\mathbf{y}_m$ are assumed to contain both speech and transients, and hence, so is the mid-point $\mathbf{y}_p$.

Kushnir et al. have shown in [39] that using a second order Taylor expansions of $\boldsymbol{\psi}_n$ and $\boldsymbol{\psi}_m$ at the mid-point, the Euclidean distance between the two points is given by:

$$\left\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\right\|^2=\left(\mathbf{y}_n-\mathbf{y}_m\right)^T\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n-\mathbf{y}_m\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right), \quad (13)$$

where $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_p\right)$ is a pseudo-inverse of $\boldsymbol{\Lambda}\left(\boldsymbol{\psi}_p\right) \triangleq \mathbf{J}\mathbf{J}^T\left(\boldsymbol{\psi}_p\right) \in \mathbb{R}^{L\times L}$, and $\mathbf{J}\left(\boldsymbol{\psi}_p\right) \in \mathbb{R}^{L\times d}$ is the Jacobian of the function g at the mid-point. The approximation to the fourth order in (13) holds due to the symmetry of the Taylor expansions of $\boldsymbol{\psi}_n$ and $\boldsymbol{\psi}_m$ at the mid-point under our hypothesis; for the sake of completeness, the derivation of (13) is given in Appendix I. In Appendix II, we further show that the term $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_p\right)$ in (13) can be replaced by the term $\frac{1}{2}\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)+\frac{1}{2}\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)$; this result is obtained by the Taylor expansion of $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)$ and $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)$ to the first order at the mid-point. Consequently, we have:

$$\left\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\right\|^2=\frac{1}{2}\left(\mathbf{y}_n-\mathbf{y}_m\right)^T\left(\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)+\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)\right)\left(\mathbf{y}_n-\mathbf{y}_m\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right). \quad (14)$$

where the mid-point, which may not exist in practice, does not appear. Yet, the Jacobian matrices at $\mathbf{y}_n$ and $\mathbf{y}_m$ in (14) are unknown. In Appendix III, we show that these Jacobian matrices can be estimated from the signal at hand based on local temporal statistics. Specifically, we show that the terms $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)$ and $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)$ in (14) are equivalent to the inverse of the local covariance matrices $\mathbf{C}_n^{-1}$ and $\mathbf{C}_m^{-1}$, respectively. Thus, using the definition of the modified Mahalanobis distance (2), we obtain:

$$\left\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\right\|^2=\left\|\mathbf{y}_n-\mathbf{y}_m\right\|_M^2+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right). \quad (15)$$

Reordering (15) yields:

$$\left\|\mathbf{y}_n-\mathbf{y}_m\right\|_M^2=\left\|\boldsymbol{\psi}_n^x-\boldsymbol{\psi}_m^x\right\|^2+\left\|\boldsymbol{\psi}_n^t-\boldsymbol{\psi}_m^t\right\|^2+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right), \quad (16)$$

and substituting the re-scaled variables, $\boldsymbol{\psi}_n^x$ and $\boldsymbol{\psi}_n^t$, by the generating variables, $\boldsymbol{\theta}_n^x$ and $\boldsymbol{\theta}_n^t$, respectively, leads to (6).

Thus far, the derivation of (6) was made under the assumption that hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold for both frames $\mathbf{y}_n$ and $\mathbf{y}_m$; in Appendix IV, we derive (6) under the other hypotheses. We note that the limitation of the result in (6) lies in the assumption that the covariance matrix $\mathbf{C}_n$ is invertible [35], [39], [40]. In practice, when the dimension of the generating variables, d, is smaller than the dimension of the observable signal, L, the covariance matrix is not invertible, and, in this case, a pseudo-inverse should be used; we further discuss the estimation of the covariance matrices in Section V.

IV. Canonical Representation Through Diffusion Maps for Voice Activity Detection

The metric we present in (2) approximates the Euclidean distance between the (re-scaled) generating variables; however, the approximation holds only for short distances, where the factor $O\left(||\mathbf{y}(n)-\mathbf{y}(m)||^4\right)$ in (6) is negligible. Therefore, the proposed metric cannot be directly incorporated in typical clustering or classification methods such as support vector machines (SVM). To overcome this limitation, we use a kernel-based geometric method, termed *diffusion maps*, with a Gaussian kernel which "sees" only local distances between frames [27]. Diffusion maps integrates all local distances into a global parameterization respecting the local distances; since the local distances are based on the generating variables, this global parameterization represents the generating variables and can be viewed as the canonical representation of the signal.

Let $k\left(\mathbf{y}_n,\mathbf{y}_m\right)$ be a similarity kernel between frames $\mathbf{y}_n$ and $\mathbf{y}_m$, given by:

$$k\left(\mathbf{y}_n,\mathbf{y}_m\right)=e^{-\frac{\left\|\mathbf{y}_n-\mathbf{y}_m\right\|_M^2}{\varepsilon}}, \quad (17)$$

where ε is a scaling parameter. Short distances between frame $\mathbf{y}_n$ and frame $\mathbf{y}_m$ provide high values of the kernel, whereas distances much greater than the scaling parameter ε are negligible. In practice, we set the parameter ε according to [41]; since for distances smaller than ε the approximation in (6) holds, the proposed kernel measures local similarities between frames according to the (re-scaled) generating variables. Using the kernel in (17), we construct an affinity matrix $\mathbf{K}\in\mathbb{R}^{N\times N}$ such that its (n,m)th entry, denoted by $K_{n,m}$, represents the similarity between frame $\mathbf{y}_n$ and frame $\mathbf{y}_m$:

$$K_{n,m}=k\left(\mathbf{y}_n,\mathbf{y}_m\right). \quad (18)$$

The affinity matrix $\mathbf{K}$ defines a weighted symmetric graph such that the frames $\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_N$ are the nodes of the graph and the edge between frame $\mathbf{y}_n$ and frame $\mathbf{y}_m$ is given by $K_{n,m}$. We define a Markov chain on the graph by normalizing the kernel [27]:

$$\mathbf{M}=\mathbf{D}^{-1}\mathbf{K}, \quad (19)$$

where $\mathbf{D}\in\mathbb{R}^{N\times N}$ is a diagonal matrix with $D_{m,m}=\sum_n K_{m,n}$. Namely, $\mathbf{M}\in\mathbb{R}^{N\times N}$ is a row stochastic Markov matrix whose rows sum to one. Then, we apply the eigenvalue decomposition to $\mathbf{M}$ yielding eigenvalues $1=\lambda_0>\lambda_1>...>\lambda_{N-1}\in\mathbb{R}$ corresponding to eigenvectors $\boldsymbol{\phi}_0,\boldsymbol{\phi}_1,...,\boldsymbol{\phi}_{N-1}\in\mathbb{R}^N$ [27]. Due to the row normalization, the leading eigenvalue λ_0 equals one, and the leading eigenvector $\boldsymbol{\phi}_0$ is an all ones vector that we ignore since it does not contain information. We use the eigenvectors to form a global parameterization of the signal. Specifically, we construct a matrix $\boldsymbol{\Phi}\in\mathbb{R}^{N\times J}$ using $J<N$ eigenvectors corresponding to the J largest eigenvalues:

$$\boldsymbol{\Phi}\equiv\left[\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,...,\boldsymbol{\phi}_J\right], \quad (20)$$

where the nth row of the matrix is the parameterization of frame $\mathbf{y}_n$. This parameterization respects the local affinities between the generating variables and is independent of the mapping function f. Therefore we view it as the canonical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

representation of the signal. In [20], similarly to the present study, the eigenvectors of a kernel function are used to construct a low-dimensional representation of the signal. The representation is exploited for voice activity detection in a supervised learning framework. Specifically, a measure of voice activity is constructed in the low-dimensional domain using a training set comprising marked speech and transients segments. In this study, we obtain improved clustering between speech and transients using the proposed kernel as we demonstrate in Section V. Hence, we take an *unsupervised* approach and propose to use only the leading (non-trivial) eigenvector, $\boldsymbol{\phi}_1$, as a measure of voice activity, i.e., we set $J = 1$ in (20). We emphasize that the eigenvector $\boldsymbol{\phi}_1$ is of length N, as the number of frames in the sequence, and each of its coordinates describe a frame. Specifically, we estimate the speech indicator of frame n in (1) by comparing the nth entry of $\boldsymbol{\phi}_1$, which we denote by $\phi_1(n)$, to a threshold, such that values above the threshold indicate voice activity:

$$\hat{\mathbb{1}}_n^x = \left\{ \begin{array}{ll} 1; & \phi_1(n) > \tau \\ 0; & \text{otherwise} \end{array} \right\}, \tag{21}$$

where τ is the threshold value. The threshold value may control the trade-off between correct detection and false alarm rates, and, in particular, setting the threshold value to zero may provide a good distinction between speech and non-speech frames as we will show in Section V. We note here that the leading eigenvector, $\boldsymbol{\phi}_1$, solves the well-known normalized cut problem presented in [42] and is widely used for clustering. The main difference with respect to previous studies is that in this study, the use of the modified Mahalanobis distance gives rise to the clustering of the signal according to the generating variables. In addition, we use the leading eigenvector as a *continuous measure* of voice activity in contrast to binary labeling. We will show in Section V that the leading eigenvector successfully distinguishes between speech and transients and provides improved detection scores compared to competing detectors. The proposed algorithm for voice activity detection is summarized in Algorithm 1.

Algorithm 1 Voice activity detection

1: Calculate the MFCCs of the noisy signal $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N$ and estimate the corresponding covariance matrices $\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_N$
2: Calculate the affinity kernel $\mathbf{K}$ based on the modified Mahalanobis distance according to (2), (17) and (18)
3: Calculate $\mathbf{M}$ according to (19)
4: Obtain the leading eigenvector $\boldsymbol{\phi}_1$
5: **for** $n = 1 : N$ **do**
6: **if** $\phi_1(n) > \tau$ **then**
7: $\hat{\mathbb{1}}_n^x = 1$
8: **else**
9: $\hat{\mathbb{1}}_n^x = 0$
10: **end if**
11: **end for**

V. Experimental Results

A. Implementation

To evaluate the performance of the proposed approach, we use speech and transient signals taken from the TIMIT database [43] and an online free corpus [44], respectively. The signals are sampled at 16 kHz, and are processed in frames of 512 samples with 50 percent overlap. We use 40 speech utterances of different speakers and construct 20 sequences, $20 - 30$ s long, by raffling 5 random utterances for each sequence. The transients are synthetically added to the speech sequences, and they are normalized to have the same maximal values. This type of normalization was previously used in [16], [19], [20], and we find it more convenient than for example, normalizing the transients according to their energy, which often has small values due to the short duration of the transients.

The proposed metric in (2) requires the estimation of local covariance matrices for each frame of the signal; one approach for their estimation is to use the sample covariance, as was suggested in [45]:

$$\hat{\mathbf{C}}_n = \frac{1}{2R+1} \sum_{i=-R}^{R} \left(\mathbf{y}_{n+i} - \hat{\boldsymbol{\mu}}_n\right) \left(\mathbf{y}_{n+i} - \hat{\boldsymbol{\mu}}_n\right)^T,$$

where $\mathbf{y}_{n-R}, \mathbf{y}_{n-R+1}, ..., \mathbf{y}_{n+R}$ are consecutive frames at a small temporal neighborhood of frame $\mathbf{y}_n$, and $\hat{\boldsymbol{\mu}}_n = \frac{1}{2R+1} \sum_{i=-R}^{R} \mathbf{y}_{n+i}$ is the sample mean. In our experiments, we set R to 15 and similarly to the finding in [45], we empirically found that a good distinction between speech and transients is obtained using a very small temporal neighborhood with a high overlap between the consecutive frames. However, the use of highly overlapping frames significantly increases the computational cost of the algorithm. Hence, in this study we assume that entries of the observable signal are uncorrelated such that the covariance matrix is diagonal. Accordingly, we estimate the variance of each entry of the observable signal using recursive averaging of the spectrum of the signal, similarly to the method presented in [4]. In this approach, we exploit the entire signal including the silent frames since the variances of speech and the transients are estimated according to variations of the spectrum of the noisy signal with respect to the spectrum estimated in the silent frames. Recall that when the dimension of the generating variables, d, is smaller than the dimension of the observable signal L, a pseudo-inverse is used for the estimation of the inverse of the covariance matrix in (2). We empirically found that applying a pseudo-inverse using three entries of the observable signal with the highest variances provide improved distinction between speech and transients. This finding heuristically implies that the signal is controlled by three generating variables. In our experiments, the estimation of the covariance matrices based on recursive averaging of the spectrum of the signal provides better detection scores compared to the use of the sample covariance, and it is the one used in the simulations in this section. The estimation of the covariance matrix will be further addressed in a future study.

The proposed representation is obtained according to (17)-(20) in a batch manner since all N frames of the sequence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

are required in advance to calculate the affinity matrix in (18). Still, the proposed algorithm may be implemented in an online manner, e.g., by constructing the canonical representation of the signal using a calibration set, given in advance without labels. Then, the eigenvectors of the kernel may be extended to new incoming frames, e.g., using the Nyström method [46]. In this context we note that to reduce the computational cost of the affinity matrix calculation in (18), we exploit a non-symmetric kernel in (17), and address the reader to [39] for more implementation details. We empirically found that it provides better detection scores compared to calculating the symmetric kernel.

B. Voice Activity Detection

The proposed representation of a speech signal, contaminated with a door-knocks transient, is illustrated in Fig. 1 (bottom), and is compared to the representation obtained using the Euclidean distance instead of the Mahalanobis distance in Fig. 1 (top). In both figures, we present a scatter plot of the first two eigenvectors of the affinity kernel such that each point represents a time frame. The points are marked according to the hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ using the labels of the ground truth: frames for which only one of the hypotheses, $\mathcal{H}_1^x$ or $\mathcal{H}_1^t$, holds are marked with red squares and green stars, respectively, and those for which both hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold are marked with blue circles. It can be seen in Fig. 1 (top) that the representation obtained based on the Euclidean distance only partially distinguishes between speech and non-speech frames. In particular, since transients are often more dominant than speech, many frames containing both speech and transients are represented as similar to frames containing only transients. In contrast, the representation obtained based on the proposed metric, illustrated in Fig. 1 (bottom), provides improved clustering between speech and transients. In particular, frames containing both speech and transients tend to be more similar to speech frames than to transients.

The representation obtained from the noisy signal using the proposed metric allows us to devise a measure of voice activity in an unsupervised manner based on the first eigenvector. Specifically, we can estimate the speech indicator for voice activity in (1) by comparing the first eigenvector to a threshold such that values above the threshold indicate voice activity. Specifically, setting the threshold value to zero may provide a good distinction between speech and non-speech frames. At this point we note that the eigenvectors are obtained by the eigenvalue decomposition with arbitrary signs. Therefore, the sign of the first eigenvector has to be set such that the speech cluster corresponds to its high values. In this study, we assume that the correct sign of the eigenvector is known. In practice, the sign of the eigenvector may be set according to the temporal variability of the signal, such that the cluster of transients is assumed to comprise segments of the signal with higher variability rates over time [15]. In addition, we note that although in this study we only use a single eigenvector, more eigenvectors may be used for voice activity detection. For example, in the studies presented in [20], [21], several eigenvectors are used as a low dimensional representation and

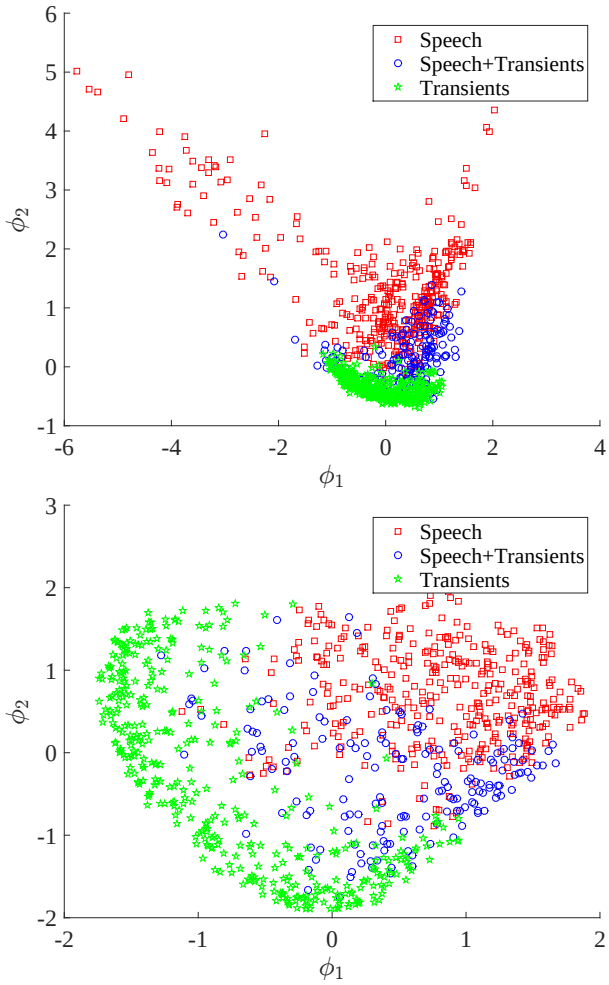


Fig. 1: Scatter plot of the first two non-trivial eigenvectors, for which the speech signal is contaminated by a door-knocks transient. (Top) Kernel based on the Euclidean distance and (bottom) kernel based on the modified Mahalanobis distance.

they are incorporated in a supervised learning framework. However, these studies consider a different problem setup, where the type of transients is known in advance and that they are available in a training set. A different heuristic approach, which does not require a training set, is using a deterministic combination between the eigenvectors, e.g., the sum of the first two eigenvectors; yet, we did not find in our simulations a combination, which consistently provided improved performance. The incorporation of several eigenvectors for voice activity detection will be addressed in a future study.

An example of the obtained voice activity detection of a speech signal contaminated with keyboard taps transients is presented in Fig. 2 and Fig. 3. In Fig. 2, we qualitatively compare the performance of the proposed detector to the one presented in [20], which we term "Mousazadeh" in the plots. For both detectors, we set the threshold value to provide 90 percent correct detection rate and compare between their false alarms. Figure 2 (top) demonstrates that the false alarm rate of Mousadazeh is significantly higher than the false alarm rate of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the proposed detector, especially in the non-speech region after the 15th second. We note that the method presented in [20] is based on representing the noisy signal using MFCCs, and then inferring a low-dimensional representation based on the Euclidean distance in which transient frames tend to be similar to speech frames as demonstrated in Fig. 1 (top). Therefore, it only partially distinguishes speech from transients, whereas the proposed method, based on the improved metric, provides a better distinction between them.

In addition, the method presented in [20] is based on a supervised learning procedure in which the low-dimensional representation is obtained using a training set, and the transients are assumed known in advance. To make a fair comparison, in our simulations, we train the algorithm presented in [20] using several types of transients. In particular, for the evaluation of the algorithm, we use the same types of transients as in the training procedure, but the transients are taken from different recordings than those used for training. In contrast to Mousadazeh, the proposed method performs in an unsupervised manner, and the voice activity measure is learned from the sequence without any prior information.

To further gain insight into the voice activity detection obtained using the leading eigenvector ϕ_1 in (20), we plot the trajectory of ϕ_1 over time in Fig. 3. For the clarity of the presentation, we normalize the eigenvector in the plot to the range of 0 to 1. In addition, we recall that the eigenvector is used as a voice activity measure only for frames containing speech, transients or both of them; silent frames are assumed known in advance and they are assigned with the value zero in the plot. Figure 3 demonstrates that entries of the eigenvector with large values correspond to frames containing speech. Indeed, by setting the threshold to a value that yields 90 percent correct detection rate, the entries of the eigenvector, $\{\phi_1(n)\}$, that correspond to non-speech frames containing transients, receive values below the threshold. As a result, they correctly indicate absence of speech.

In addition to the method presented in [20], the performance of the proposed method is compared to the performance of the methods presented in [5], [8], and [47], which we term "Sohn", "Ramirez" and "Ishizuka" in the plots, respectively. The proposed method is termed "Proposed (MK)", where (MK) is the Mahalanobis kernel, and it is also compared to a similar method based on the Euclidean distance termed "Proposed (EK)", where (EK) is the Euclidean kernel. To better appreciate the results, we report on the delays induced by each method. The methods Sohn and Ishizuka operate in an online manner without a delay; Mousazadeh and Ramirez operate with a delay of two and four frames, respectively; and in the presented experiments, the proposed method operate in a batch manner. Yet, as already noted, the proposed method may be implemented in an online manner without a delay using a calibration set, given in advance without labels, similarly to the method we presented in [21].

The performance of the methods is evaluated in Figs. 4–6 and in Table I. In Figs. 4–6 the methods are evaluated in the form of receiver operating characteristic (ROC) curves, which are curves of probability of detection versus probability of false alarm. The ROC curves are generated by sweeping the threshold value in (21) from the minimal to the maximal entry of the leading eigenvector such that the higher the threshold is, the lower the correct detection and the false alarm rates are. The larger the area under the curve (AUC), the better the performance of the method; the AUC of each method is given in the legend box of each plot. Each of the Figs. 4–6 illustrates the performance of the methods for different types of transients: keyboard taps, hammering and door-knocks, respectively.

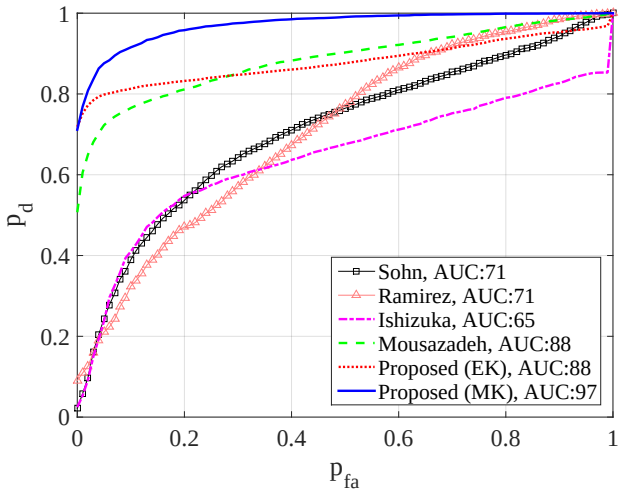


Fig. 4: Probability of detection vs probability of false alarm. Test for a keyboard taps transient.

It can be seen in Figs. 4–6 that the competing methods Sohn, Ramirez and Ishizuka provide poor performance in distinguishing speech from transient frames since they are not designed for this particular task. In Figs. 4 and 5, the proposed method with the Euclidean distance and the method Mousadazeh, which both exploit the Euclidean distance to obtain a low-dimensional representation are comparable and perform significantly better than the methods presented in [5], [8] and [47]. Moreover, the proposed method based on the modified Mahalanobis distance provides the best performance. In Fig. 6, the proposed method provides comparable results to Mousadazeh and significantly outperforms all other methods.

We evaluate the proposed method for different ratios between the transients and speech, and report the AUC obtained for each method for different types of transients in Table I. We define the transient to speech ratio as the ratio between the maximal amplitudes of the transients and speech such that for equal maximal amplitudes, as considered in the previous experiments, the ratio is one. To provide a fair comparison, the Mousadazeh method, which is the only method in our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

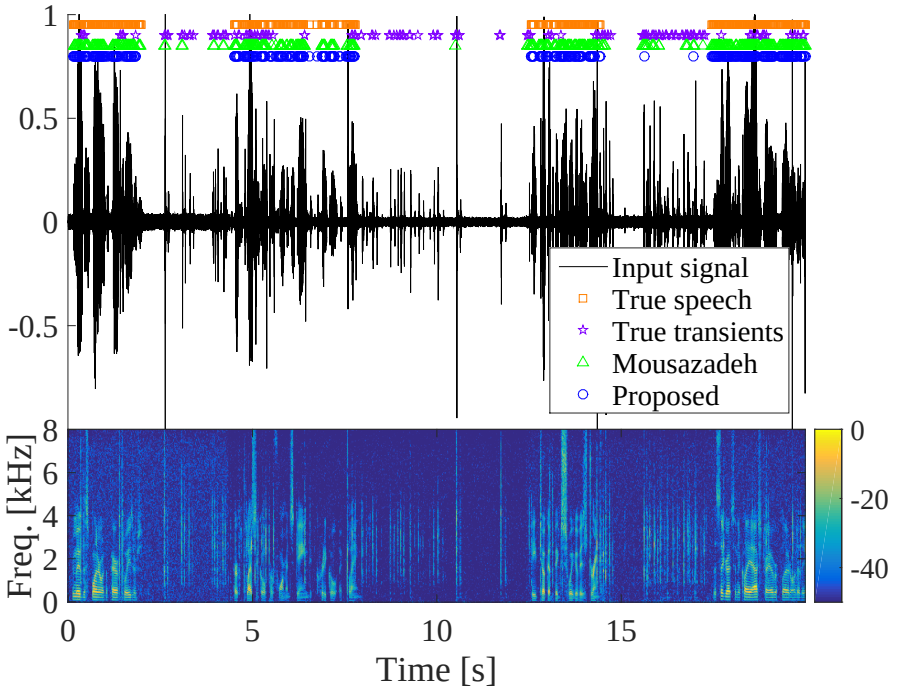


Fig. 2: Qualitative assessment of the proposed VAD, with a keyboard taps transient. (Top) Time domain, input signal- black solid line, true speech- orange squares, true transients- purple stars, Mousazadeh with a threshold set for 90 percents correct detection rate- green triangles, proposed algorithm with a threshold set for 90 percent correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

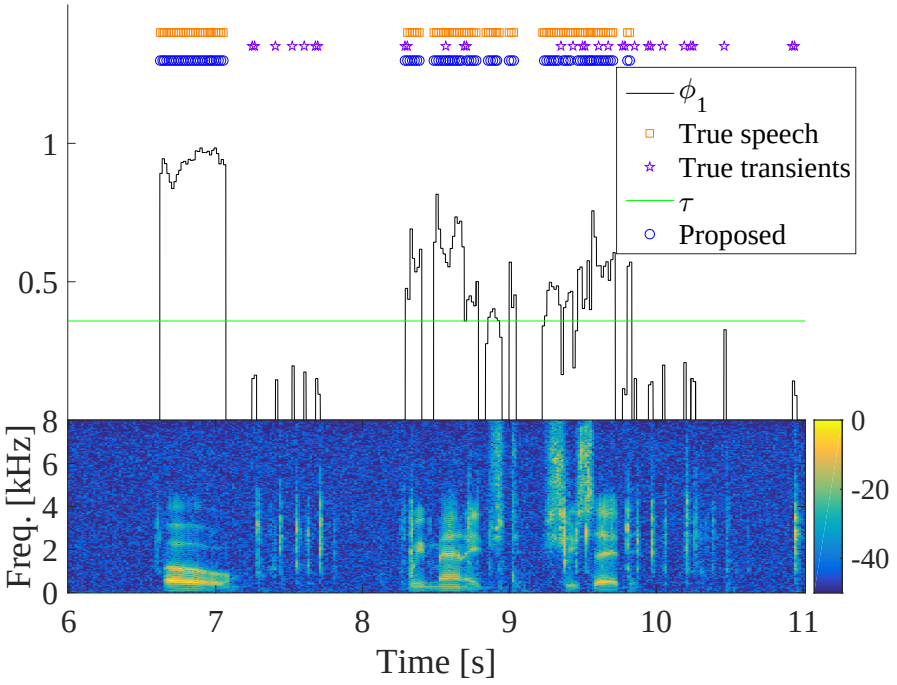


Fig. 3: Qualitative assessment of the proposed VAD, with a keyboard taps transient. (Top) Time domain, the voice activity measure, i.e.,ϕ_1- black solid line, true speech- orange squares, true transients- purple stars, a threshold value τ providing 90 percents correct detection rate- green line, proposed algorithm with a threshold set for 90 percent correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

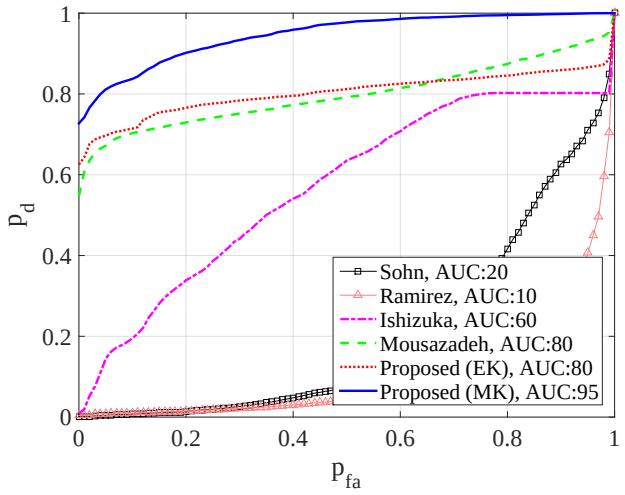


Fig. 5: Probability of detection vs probability of false alarm. Test for a hammering transient.

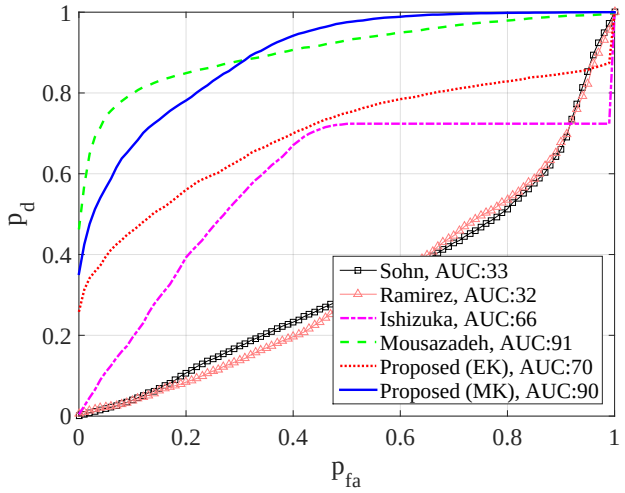


Fig. 6: Probability of detection vs probability of false alarm. Test for a door-knocks transient.

experiments based on supervised learning, is trained only for transient to speech ratio of 1 such that the transient to noise ratio is assumed to be unknown for all methods. We observe in Table I that for transient to speech ratio of 0.5, the proposed method based on the Mahalanobis distance provides performance, which is comparable to Mousazadeh and outperforms the competing detectors. Moreover, the proposed method provides the best performance in most of the experiments for transient to speech ratios of 1 and 2. The improved performance of the proposed method for high transient to speech ratio demonstrates its ability to reduce the dominance of the transients. In Table I (d) we summarize the average performance of the different methods for the different transients and transient to speech ratios demonstrating that the proposed method based on the modified Mahalanobis distance outperforms the competing detectors.

VI. Conclusions

We have addressed the problem of voice activity detection in the presence of transients and have proposed a modified version of the Mahalanobis distance, which better distinguishes between speech and transients. To motivate the use of the modified Mahalanobis distance, we have presented a model in which speech and transients are represented by two independent sets of generating variables. The generating variables represent the content of the signal, i.e., speech and transients, and, as a result, speech and transients are successfully distinguished. Although the generating variables are not directly accessible, we have shown that distances between them can be approximated by the modified Mahalanobis distance. Moreover, we have shown that the Mahalanobis distance approximates the Euclidean distance between re-scaled variables for which the dominance of the transients is reduced; therefore, it is especially suitable for voice activity detection since it allows us to cluster the signal according to the presence of speech rather than according to the presence of transients. The main limitation in the use of the Mahalanobis distance is that the approximation of the Euclidean distance between the generating variables holds only for small distances. To overcome this problem, we have proposed to exploit a kernel-based manifold learning approach that integrates short Mahalanobis distances into a global canonical representation of the signal. We have shown that the canonical representation successfully divides the signal into speech and non-speech clusters. Based on the canonical representation we have proposed a measure of voice activity providing improved performance compared to competing detectors.

Acknowledgment

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

Appendix I

Second Order Taylor Expansion at The Mid-point

Recall that the mid-point $\mathbf{y}_p$ is given by $\mathbf{y}_p = \frac{\mathbf{y}_n+\mathbf{y}_m}{2}$; by a second order Taylor expansion at the mid-point, the ith re-scaled generating variable, denoted by $\boldsymbol{\psi}_n(i)$, is given by [39]:

$$\begin{aligned}\boldsymbol{\psi}_n(i) = \boldsymbol{\psi}_p(i) + \tfrac{1}{2}\textstyle\sum_j g_{i,j}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(j)-\mathbf{y}_m(j)\right)\\ +\tfrac{1}{8}\textstyle\sum_{kl} g_{i,kl}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(k)-\mathbf{y}_m(k)\right)\left(\mathbf{y}_n(l)-\mathbf{y}_m(l)\right)\\ +O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^3\right).\end{aligned} \tag{22}$$

where g_i is the ith element in g, $g_{i,j} \triangleq \frac{\partial g_i}{\partial \mathbf{y}_n(j)}$ and $g_{i,kl} \triangleq \frac{\partial g_i}{\partial \mathbf{y}_n(k)\partial \mathbf{y}_n(l)}$. Using a similar expansion of $\boldsymbol{\psi}_m(i)$ around the mid-point, the Euclidean distance between the re-scaled variables is given by:

$$\begin{aligned}\left\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\right\|^2 = \textstyle\sum_i\left(\boldsymbol{\psi}_n(i)-\boldsymbol{\psi}_m(i)\right)^2 =\\ \textstyle\sum_{ijk} g_{i,j}\left(\mathbf{y}_p\right) g_{i,k}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(j)-\mathbf{y}_m(j)\right)\left(\mathbf{y}_n(k)-\mathbf{y}_m(k)\right)\\ +O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right),\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

	Keyboard taps	Hammering	Door-knocks	Metronome	Scissors	Crackles
Sohn	71	20	33	28	68	59
Ramirez	71	10	32	14	65	60
Ishizuka	65	60	66	57	74	49
Mousazadeh	88	80	**91**	80	87	**93**
Proposed (EK)	88	80	70	81	**93**	79
Proposed (MK)	**97**	**95**	90	**91**	**93**	88

(a)

	Keyboard taps	Hammering	Door-knocks	Metronome	Scissors	Crackles
Sohn	52	14	26	16	48	39
Ramirez	49	6	25	6	42	38
Ishizuka	63	53	64	57	73	51
Mousazadeh	71	58	81	58	64	76
Proposed (EK)	91	79	72	82	**96**	78
Proposed (MK)	**97**	**91**	**87**	**91**	92	**86**

(b)

	Keyboard taps	Hammering	Door-knocks	Metronome	Scissors	Crackles
Sohn	86	29	43	48	83	79
Ramirez	87	20	45	35	85	81
Ishizuka	66	63	66	58	74	45
Mousazadeh	94	86	**95**	**90**	**92**	**96**
Proposed (EK)	86	82	77	81	90	80
Proposed (MK)	**95**	**97**	92	86	91	90

(c)

Sohn	Ramirez	Ishizuka	Mousazadeh	Proposed (EK)	Proposed (MK)
47	43	61	82	83	**92**

(d)

TABLE I: (a) AUC scores; transient to speech ratio: 1. (b) AUC scores; transient to speech ratio: 2. (c) AUC scores; transient to speech ratio: 0.5. (d) Average AUC scores.

because all multiplications terms comprising the second order of the Taylor expansion in (22) are of the order of $O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right)$ due to symmetry. In a matrix notation, we have:

$$\left\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\right\|^2=\left(\mathbf{y}_n-\mathbf{y}_m\right)^T\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n-\mathbf{y}_m\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right),$$

where $\boldsymbol{\Lambda}\left(\boldsymbol{\psi}_p\right)\triangleq\mathbf{J}\mathbf{J}^T\left(\boldsymbol{\psi}_p\right)\in\mathbb{R}^{L\times L}$, and $\mathbf{J}\left(\boldsymbol{\psi}_p\right)\in\mathbb{R}^{L\times d}$ is the Jacobian of the function h at the mid-point.

APPENDIX II
JACOBIAN AT THE MID-POINT

Let $\gamma_{ij}\left(\mathbf{y}_n\right)$ be the (i,j)th entry of $\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)$; the first order Taylor expansions of $\gamma_{ij}\left(\mathbf{y}_n\right)$ and $\gamma_{ij}\left(\mathbf{y}_m\right)$ at the mid-point are given by:

$$\gamma_{ij}\left(\mathbf{y}_n\right)=\gamma_{ij}\left(\mathbf{y}_p\right)+\tfrac{1}{2}\textstyle\sum_k\gamma_{ij,k}\left(\mathbf{y}_p\right)\left(\mathbf{y}_n(k)-\mathbf{y}_m(k)\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^2\right),$$

$$\gamma_{ij}\left(\mathbf{y}_m\right)=\gamma_{ij}\left(\mathbf{y}_p\right)+\tfrac{1}{2}\textstyle\sum_k\gamma_{ij,k}\left(\mathbf{y}_p\right)\left(\mathbf{y}_m(k)-\mathbf{y}_n(k)\right)+O\left(\left\|\mathbf{y}_m-\mathbf{y}_n\right\|^2\right),$$

where $\gamma_{ij,k}\left(\mathbf{y}_n\right)=\frac{\partial\gamma_{ij}}{\partial\mathbf{y}_n(k)}$. The summation of this two equations yields:

$$\gamma_{ij}\left(\mathbf{y}_p\right)=\frac{1}{2}\gamma_{ij}\left(\mathbf{y}_n\right)+\frac{1}{2}\gamma_{ij}\left(\mathbf{y}_m\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^2\right).$$

Hence, in a matrix form, we have:

$$\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_p\right)=\frac{1}{2}\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)+\frac{1}{2}\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^2\right),$$

and by substituting the last equation into (13), we have:

$$\left\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\right\|^2=\tfrac{1}{2}\left(\mathbf{y}_n-\mathbf{y}_m\right)^T\left(\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_n\right)+\boldsymbol{\Lambda}^{-1}\left(\mathbf{y}_m\right)\right)\left(\mathbf{y}_n-\mathbf{y}_m\right)+O\left(\left\|\mathbf{y}_n-\mathbf{y}_m\right\|^4\right).$$

APPENDIX III
LOCAL JACOBIAN VS LOCAL COVARIANCE

We estimate the covariance matrix $\mathbf{C}_n$ of the observable signal $\mathbf{y}_n$ at a small temporal neighborhood of frame n, and assume a set of frames in a small temporal neighborhood of $\mathbf{y}_n$ for which $\mathbf{y}_n$ is the mean value. The first order Taylor expansion of an arbitrary frame, denoted by $\mathbf{y}$, around $\mathbf{y}_n$ is given by:

$$\mathbf{y}=\mathbf{y}_n+\mathbf{J}\left(\boldsymbol{\psi}_n\right)\left(\boldsymbol{\psi}-\boldsymbol{\psi}_n\right)+O\left(\left\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\right\|^2\right). \quad (23)$$

The relation between the covariance matrix $\mathbf{C}_n$ and the Jacobian $\mathbf{J}\left(\boldsymbol{\psi}_n\right)$ in frame n is given by:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

$$\begin{aligned}\mathbf{C}_n &= \mathbb{E}\left[(\mathbf{y}-\mathbf{y}_n)(\mathbf{y}-\mathbf{y}_n)^T\right] \\ &= \mathbf{J}(\boldsymbol{\psi}_n)\,\mathbb{E}\left[(\boldsymbol{\psi}-\boldsymbol{\psi}_n)(\boldsymbol{\psi}-\boldsymbol{\psi}_n)^T\right]\mathbf{J}^T(\boldsymbol{\psi}_n) \\ &\quad + O\left(\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\|^3\right) \\ &= \boldsymbol{\Lambda}(\boldsymbol{\psi}_n) + O\left(\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\|^3\right),\end{aligned} \tag{24}$$

where assuming that $\boldsymbol{\psi}_n$ is the mean value of the generating variables, $\mathbb{E}\left[(\boldsymbol{\psi}-\boldsymbol{\psi}_n)(\boldsymbol{\psi}-\boldsymbol{\psi}_n)^T\right]$ is the covariance of $\boldsymbol{\psi}$, which is the identity matrix due to the normalization in (8) and (9). We note that the error term in (24) is neglected since we assume that frames in a small temporal neighborhood tend to be more similar to $\mathbf{y}_n$ compared to an arbitrary frame $\mathbf{y}_m$. Moreover, assuming a symmetric distribution of $\boldsymbol{\psi}$ around $\boldsymbol{\psi}_n$, e.g. a Gaussian distribution, the error term becomes of the order of four since odd moments of the distribution equal zero. By following the derivation presented in [39], it may be further shown that $\left[\boldsymbol{\Lambda}(\boldsymbol{\psi}_n) + O\left(\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\|^3\right)\right]^{-1} = \left[\boldsymbol{\Lambda}^{-1}(\mathbf{y}_n) + O\left(\|\mathbf{y}-\mathbf{y}_n\|^3\right)\right]$. This result is obtained by further assuming that the function f in (4) is bi-Lipschitz such that distances between frames in the observable domain $\|\mathbf{y}-\mathbf{y}_n\|$ are of the same order as in the domain of the generating variables $\|\boldsymbol{\psi}-\boldsymbol{\psi}_n\|$. Hence, by setting $\mathbf{C}_n^{-1} \approx \boldsymbol{\Lambda}^{-1}(\mathbf{y}_n)$ in (13) we have:

$$\|\mathbf{y}_n-\mathbf{y}_m\|_M^2 = \|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\|^2 + O\left(\|\mathbf{y}_n-\mathbf{y}_m\|^4\right).$$

APPENDIX IV
MODIFIED MAHALANOBIS DISTANCE FOR DIFFERENT HYPOTHESES

The derivation of (6) was made in Section III under the assumption that hypotheses $\mathcal{H}_1^x$ and $\mathcal{H}_1^t$ hold for both frames $\mathbf{y}_n$ and $\mathbf{y}_m$. We now address the other hypotheses, starting from the case in which only speech is present in both $\mathbf{y}_n$ and $\mathbf{y}_m$, and as a result, in the mid-point $\mathbf{y}_p$ as well. Since both frames are independent of transients, the partial derivatives of entries of the function g with respect to the generating variables of transients equal zero, i.e., $\forall j : g_{i,j} = \frac{\partial g_i}{\partial y_n(j)} = 0$. Accordingly, the Jacobian of g is reduced to:

$$\mathbf{J}(\boldsymbol{\psi}_n) \triangleq \begin{bmatrix} \mathbf{J}_x(\boldsymbol{\psi}_n^x) \\ \mathbf{J}_t(\boldsymbol{\psi}_n^t) \end{bmatrix} = \begin{bmatrix} \mathbf{J}_x(\boldsymbol{\psi}_n^x) \\ 0 \end{bmatrix},$$

where $\mathbf{J}_x(\boldsymbol{\psi}_n^x) \in \mathbb{R}^{d^x \times L}$ and $\mathbf{J}_t(\boldsymbol{\psi}_n^t) \in \mathbb{R}^{d^t \times L}$ are the parts of the Jacobian associated with the generating variables of the speech and of the transients, respectively, and hence:

$$\boldsymbol{\Lambda}(\boldsymbol{\psi}_n) \triangleq \mathbf{J}\mathbf{J}^{\mathbf{T}}(\boldsymbol{\psi}_n) = \mathbf{J}_x\mathbf{J}_x^T(\boldsymbol{\psi}_n^x). \tag{25}$$

By substituting (25) into (13) and using a similar derivation as in Appendix II, we obtain a result similar to (14):

$$\begin{aligned}&\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\|^2 = \\ &\tfrac{1}{2}(\mathbf{y}_n-\mathbf{y}_m)^T\left(\boldsymbol{\Lambda}_x^{-1}(\mathbf{y}_n)+\boldsymbol{\Lambda}_x^{-1}(\mathbf{y}_m)\right)(\mathbf{y}_n-\mathbf{y}_m) \\ &\qquad + O\left(\|\mathbf{y}_n-\mathbf{y}_m\|^4\right),\end{aligned}$$

where $\boldsymbol{\Lambda}_x(\boldsymbol{\psi}_p) \triangleq \mathbf{J}_x\mathbf{J}_x^T(\boldsymbol{\psi}_p) \in \mathbb{R}^{L\times L}$. Since transients are absent for frames $\mathbf{y}_n$ and $\mathbf{y}_m$, the estimated statistics of the observable signal are related to the generating variables of speech, and, by revisiting Appendix III, we have $\mathbf{C}_n^{-1} \approx \boldsymbol{\Lambda}_x^{-1}(\mathbf{y}_n)$ and $\mathbf{C}_m^{-1} \approx \boldsymbol{\Lambda}_x^{-1}(\mathbf{y}_m)$. Therefore, (6) holds under the hypothesis that only speech is present in both $\mathbf{y}_n$ and $\mathbf{y}_m$. The derivation of (6) is analogous in case $\mathbf{y}_n$ and $\mathbf{y}_m$ comprising only transients.

Our derivation of (6) concludes by addressing the case where $\mathbf{y}_n$ contains only speech and $\mathbf{y}_m$ contains only transients. In this case, we exploit the introduction of the mid-point, which comprises both speech and transients. As a result, the derivation of (13) remains unchanged, and by revisiting Appendix II, we have

$$\begin{aligned}&\|\boldsymbol{\psi}_n-\boldsymbol{\psi}_m\|^2 = \\ &\tfrac{1}{2}(\mathbf{y}_n-\mathbf{y}_m)^T\left(\boldsymbol{\Lambda}_x^{-1}(\mathbf{y}_n)+\boldsymbol{\Lambda}_t^{-1}(\mathbf{y}_m)\right)(\mathbf{y}_n-\mathbf{y}_m) \\ &\qquad + O\left(\|\mathbf{y}_n-\mathbf{y}_m\|^4\right),\end{aligned} \tag{26}$$

where $\boldsymbol{\Lambda}_t(\boldsymbol{\psi}_m) \triangleq \mathbf{J}_t\mathbf{J}_t^T(\boldsymbol{\psi}_m) \in \mathbb{R}^{L\times L}$. Recall that according to Appendix III, $\mathbf{C}_n^{-1} \approx \boldsymbol{\Lambda}_x^{-1}(\mathbf{y}_n)$ and $\mathbf{C}_m^{-1} \approx \boldsymbol{\Lambda}_t^{-1}(\mathbf{y}_n)$. Thus, by substituting them in (26), we obtain (6).

REFERENCES

[1] I. Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands," in *Proc. IEEE Workshop on Hands-Free Speech Communication, HSC'01, 2001*, pp. 95–98.

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

[3] 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, 2002.

[4] 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, 2003.

[5] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[6] J.H. Chang and N.S. Kim, "Voice activity detection based on complex laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, 2003.

[7] J. H. Chang, J. W0 Shin, and N. S. Kim, "Likelihood ratio test with complex laplacian model for voice activity detection.," in *Proc. the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2003.

[8] J. Ramìrez, J.C. Segura, C. Benìtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[9] J.W. Shin, J.H. Chang, H.S. Yun, and N.S. Kim, "Voice activity detection based on generalized gamma distribution," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 781–784.

[10] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[11] J. Ramírez, J. C. Segura, and J. M. Górriz, "Revised contextual LRT for voice activity detection," in *Proc. 32th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 4, pp. IV–801.

[12] J.W. Shin, H.J. Kwon, S.H. Jin, and N.S. Kim, "Voice activity detection based on conditional map criterion," *IEEE Signal Processing Letters*, vol. 15, pp. 257–260, 2008.

[13] I. Volfin and I. Cohen, "Dominant speaker identification for multipoint videoconferencing," *Computer Speech & Language*, vol. 27, no. 4, pp. 895–910, 2013.

[14] R. Talmon, I. Cohen, and S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," in *Proc. 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5084–5087.

[15] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the OM-LSA algorithm," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2566919, IEEE/ACM Transactions on Audio, Speech, and Language Processing

[16] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 132–144, 2013.

[17] D. Dov and I. Cohen, "Voice activity detection in presence of transients using the scattering transform," in *Proc. IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, 2014, pp. 1–5.

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

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

[20] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 6, pp. 1261–1271, 2013.

[21] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 732–745, 2015.

[22] D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *arXiv preprint arXiv:1604.02946*, 2016.

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

[24] M. Balasubramanian, E. L. Schwartz, Tenenbaum J. B., de Silva V., and J. C. Langford, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

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

[26] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[27] R.R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[28] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[29] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3231–3254, 2005.

[30] C. J. Dsilva, R. Talmon, C. W. Gear, R. R. Coifman, and I. G. Kevrekidis, "Data-driven reduction for multiscale stochastic dynamical systems," *arXiv preprint arXiv:1501.05195*, 2015.

[31] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. 1st International Conference on Music Information Retrieval (ISMIR)*, 2000.

[32] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[33] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[34] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, 2007, vol. 2, pp. 556–561.

[35] A. Singer and R.R. Coifman, "Non-linear independent component analysis with diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 25, no. 2, pp. 226–239, 2008.

[36] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[37] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[38] A. Singer, R. Erban, I.G. Kevrekidis, and R.R. Coifman, "Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps," *Proceedings of the National Academy of Sciences*, vol. 106, no. 38, pp. 16090–16095, 2009.

[39] D. Kushnir, A. Haddad, and R. 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.

[40] R. Talmon and R. R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12535–12540, 2013.

[41] Y. Keller, R. R. Coifman, S. Lafon, and S. W. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 403–413, 2010.

[42] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

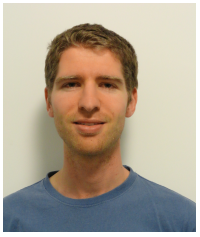
[43] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database," National Inst. of Standards and Technology (NIST), Gaithersburg, MD, Feb 1993.

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

[45] O. Rosen, S. Mousazadeh, and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering and diffusion kernels," in *IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI), 2014*. IEEE, 2014, pp. 1–5.

[46] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, 2004.

[47] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.



David Dov received the B.Sc. (Summa Cum Laude) and M.Sc. (Cum Laude) degrees in electrical engineering from the Technion - Israel Institute of Technology, Haifa, Israel, in 2012 and 2014, respectively. He is currently pursuing the PhD degree in electrical engineering at the Technion - Israel Institute of Technology, Haifa, Israel.

From 2010 to 2012, he worked in the field of Microelectronics in RAFAEL Advanced Defense Systems LTD. Since 2012, he has been a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion.

His research interests include geometric methods for data analysis, multi-sensors signal processing, speech processing, and multimedia.

David Dov is the recipient of the IBM PhD Fellowship for 2016-17, the Jacobs Fellowship for 2014, the Excellence in Teaching Award for outstanding teaching assistants in 2013, the Meyer Fellowship and the Cipers Award for 2012, the Excellent Undergraduate Project Award from the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, Technion for 2012, and Intel Award for excellent undergraduate students for 2009.

Ronen Talmon is an Assistant Professor of electrical engineering at the Technion – Israel Institute of Technology, Haifa, Israel. He received the B.A. degree (Cum Laude) in mathematics and computer science from the Open University in 2005, and the Ph.D. degree in electrical engineering from the Technion 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 at the Department of Electrical Engineering, Technion. From 2011 to 2013, he was a Gibbs Assistant Professor at the Mathematics Department, Yale University, New Haven, CT. In 2014, he joined the Department of Electrical Engineering of the Technion.

His research interests are statistical signal processing, analysis and modeling of signals, speech enhancement, biomedical signal processing, applied harmonic analysis, and diffusion geometry.

Dr. Talmon is the recipient of the Irwin and Joan Jacobs Fellowship, the Andrew and Erna Fince Viterbi Fellowship, and the Horev Fellowship.

Israel Cohen (M'01-SM'03-F'15) is a Professor of electrical engineering at the Technion – Israel Institute of Technology, Haifa, Israel. He received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion – Israel Institute of Technology, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT, USA. In 2001 he joined the Electrical Engineering Department of the Technion. He is a coeditor of the Multichannel Speech Processing Section of the *Springer Handbook of Speech Processing* (Springer, 2008), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), a Coeditor of *Speech Processing in Modern Communication: Challenges and Perspectives* (Springer, 2010), and a General Cochair of the 2010 International Workshop on Acoustic Echo and Noise Control (IWAENC). He served as Guest Editor of the *European Association for Signal Processing Journal on Advances in Signal Processing* Special Issue on Advances in Multimicrophone Speech Processing and the *Elsevier Speech Communication Journal* a Special Issue on Speech Enhancement. 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.

Dr. Cohen was a recipient of the Alexander Goldberg Prize for Excellence in Research, and the Muriel and David Jacknow Award for Excellence in Teaching. He serves as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as a member of the IEEE Speech and Language Processing Technical Committee.