2014 IEEE 28-th Convention of Electrical and Electronics Engineers in Israel

# Voice Activity Detection In Presence Of Transients Using The Scattering Transform

David Dov and Israel Cohen
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{davidd@tx, icohen@ee}.technion.ac.il

***Abstract*—Voice activity detection in the presence of highly non-stationary noise and transient interferences is an open problem. State-of-the-art voice activity detectors which are based on statistical models usually assume that noise is slowly varying with respect to speech. This assumption does not hold for transient interferences which are short time interruptions, and the performance of these detectors significantly deteriorates. In this paper, we propose a supervised learning algorithm for voice activity detection which is designed to perform in the presence of transients. We consider a labeled training set which comprises speech, background noise and transients, and propose a continuous measure for voice activity based on the Support Vector Machine (SVM) classifier. The measure of voice activity is constructed in a features domain, where the features are based on the scattering transform, include noise estimation, and are designed to separate speech and non-speech frames. Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art detectors for different types of background noises, and in particular accurately classifies frames which contain transient interferences.**

***Index Terms*—Voice activity detection, impulse noise, transient noise, Scattering transform.**

## I. Introduction

Accurate voice activity detection is necessary for a variety of speech processing applications such as speech recognition and coding, and dominant speaker identification [1]. Early methods for voice activity detection are based on straightforward features such as the energy of the signal and zero-crossing rate [2]. Although these methods perform well for clean signals, their performances deteriorate in the presence of background noise since for example the zero-crossing rate may be increased due to the background noise, wrongly indicating voice activity. To overcome this problem, several Voice Activity Detectors (VADs) which assume statistical models for the input signal and the noise, and are based on Likelihood Ratio Test (LRT) were presented in recent years. Among the different statistical models are the Gaussian model [3], [4], the Laplacian model [5], [6], [7] and the generalized Gamma model [8]. These methods typically assume that the noise is slowly varying with respect to speech and perform well in the presence of stationary noise. This assumption does not hold for highly non-stationary noise and transient interferences, which are short time interruptions such as keyboard typing and door knocking [9], [10], [11], and hence the performances of these methods significantly deteriorate.

This research was supported by the Israel Science Foundation (grant no.1130/11).

Recently, A supervised learning algorithm for voice activity detection in the presence of highly non-stationary noise was presented in [12]. The algorithm comprises features selection procedure, where the feature are based on the Mel-Frequency Cepstral Coefficients (MFCC) and noise estimation. Then, a spectral clustering method is applied in the features domain for the classification process. Although the algorithm was shown to outperform several state-of-the-art methods, its performance are still limited in the presence of transients since the transients and speech are only partially separated in the features domain.

In this paper, we present a supervised learning algorithm for voice activity detection in the presence of highly non-stationary noise and transients. To train the algorithm, we consider a labeled training data set of speech signals contaminated with noise and transients. The algorithm is based on the representation of the noisy signal by features which are based on the scattering transform [13], [14], [15], include noise estimation, and are specifically designed to separate speech and non-speech frames. For voice activity detection, we propose a continuous measure which is constructed in the features domain, and rely on the SVM classifier. The algorithm is evaluated for different types of background noises and transients, and experimental results demonstrate improved voice activity detection compared to state-of-the-art methods.

The remainder of the paper is organized as follows. In Section II we formulate the problem. The proposed algorithm is described in Section III. Experimental results demonstrating the performance of the algorithm are presented in Section IV.

## II. Problem Formulation

Let $y(\tau)$ denote a speech signal contaminated with additive background noise and an additive transient interference, given by:

$$y(\tau) = x(\tau) + d(\tau) + z(\tau) \tag{1}$$

where $x(\tau)$,$d(\tau)$ and $z(\tau)$ are speech, the background noise and the transient interference, respectively. The signal is processed in overlapping time frames of length $T$, such that time frame $t$ is denoted by $y_t$ and is given by $y(\tau)$; $\tau \in [t - T/2\,,\, t + T/2]$. Let $\mathbb{1}_{\mathbf{s}}(t)$ denote a speech indicator of

frame $t$, given by:

$$\mathbb{1}_{\mathbf{s}}(t) = \left\{ \begin{array}{lcl} 1 & ; & t \in \mathcal{H}_1 \\ 0 & ; & t \in \mathcal{H}_0 \end{array} \right\} \quad (2)$$

where $\mathcal{H}_1$ and $\mathcal{H}_0$ are two hypotheses denoting speech presence and absence, respectively. The goal in this paper is to estimate the speech indicator in (2) for each frame. The algorithm is based on a supervised learning procedure, and we consider a training data set which consists of speech signals contaminated with noise and transients, and is labeled according to the speech absence and presence hypotheses.

## III. Proposed Algorithm

### A. The Features

The proposed features are based on the scattering transform which is a cascade of wavelet convolutions and modulus operators [13], [14]. Let a wavelet $\psi(\tau)$ be a band pass filter with a central frequency normalized to 1, and let $\{\psi_\lambda(\tau)\}_\lambda$ be a wavelet filter bank, which is constructed by dilating the wavelet:

$$\psi_\lambda(\tau) = \lambda\psi(\lambda\tau), \quad (3)$$

where $\lambda = 2^{j/Q}$, $\forall j \in \mathbb{Z}$ and $Q$ is the number of wavelets per octave. The bandwidth of the wavelet $\psi(\tau)$ is of the order of $1/Q$, and as a result, the filter bank is composed of band pass filters which are centered in the frequency domain in $\lambda$ and have a frequency bandwidth $\lambda/Q$, i.e., they are logarithmically spaced in the frequency domain. The first and the second orders of the scattering transform are denoted by, $S_1(\tau, \lambda_1)$ and $S_2(\tau, \lambda_1, \lambda_2)$, respectively, and are given by:

$$S_1(\tau, \lambda_1) = |y * \psi_{\lambda_1}| * \phi(\tau), \quad (4)$$

and:

$$S_2(\tau, \lambda_1, \lambda_2) = ||y * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(\tau), \quad (5)$$

where $\phi(\tau)$ is a low pass filter with a frequency bandwidth $2\pi/T$. A scattering vector of frame $t$ is denoted by $\mathbf{y}_t^S$ and is given by concatenating the first and the second orders of the scattering transform calculated at time $t$ for each filter.

The scattering transform is invariant to time shifts and is stable to time-warping due to the logarithmically spaced filter bank, making it useful for classification (see more details in [13], [14]). These properties are also held for the Mel-Frequency Spectral Coefficients (MFSC) obtained by averaging the signal in the Short-Time Fourier Transform (STFT) domain with Mel-scale filters which are also logarithmically spaced in the frequency domain for high frequencies [14]. In addition, it is shown in [14] that the MFSCs are similar to the coefficients of the first order of the scattering transform. The MFCCs are given by applying a cosine transform on the log of the MFSCs, they are widely used in speech recognition [16], and were recently exploited for voice activity detection in [12]. However, the averaging in the frequency domain in the construction of the MFSCs and MFCCs removes information over small time scales [14]. Similarly, the convolution with the low pass filter in (4) causes loss of information. In particular, the representation of transients which are usually short in time, may be similar to the representation of speech, and may lead to false voice activity detection in the presence of transients. The second order of the scattering transform recovers the lost information using a new set of wavelet filters and the modulus operator [13], [14]. Therefore, the representation of signals using the first and the second orders of the scattering transform extends the MFCC representation and better separates between speech and transients.

Yet, non-speech frames which are contaminated with background noise may be similar to speech frames. In order to improve the separation between speech and noise frames, the scattering vector $\mathbf{y}_t^S$ is weighted with a scalar which incorporates noise estimation in the STFT domain [12]. Let $Y(t, \omega)$ be the STFT of $y(\tau)$, and let $p_r(Y(t, \omega); \mathcal{H}_0)$ and $p_r(Y(t, \omega); \mathcal{H}_1)$ be Probability Density Functions (PDF) of the noisy signal conditioned on the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. The log of the likelihood ratio between the conditional PDFs is given by:

$$\Lambda_t(\omega) = \log\left(\frac{p_r(Y(t, \omega); \mathcal{H}_1)}{p_r(Y(t, \omega); \mathcal{H}_0)}\right). \quad (6)$$

$\Lambda_t$ is a scalar obtained by averaging the log of the likelihood ratio in (6) over the frequency scale, and is used to weight the features. The weight of frame $t$ is denoted by $w_t$ and is given by:

$$w_t = 1 - e^{-\frac{\Lambda_t}{\epsilon}}$$

where $\epsilon$ is a normalization parameter. Accordingly, the feature vector of frame $t$ is given by:

$$\mathbf{y}_t = w_t \mathbf{y}_t^S. \quad (7)$$

In frames which contain merely background noise, $\Lambda_t$ receives low values since $p_r(Y(t, \omega); \mathcal{H}_1) \to 0$ in (6), and $\epsilon$ is set such that $w_t$ receives values close to 0. In speech frames, $\Lambda_t$ in (6) receives high values, and $w_t$ receives values close to 1. $\Lambda_t$ is estimated according to [3] and incorporates noise estimation procedure which is based on the assumption that noise is (quasi) stationary and is slowly varying with respect to speech [17], [18]. Since transients are highly non-stationary signals and are varying faster than speech, high values of $\Lambda_t$ are obtained also in presence of transients. As a result, $w_t$ receives values close to 1 both in the presence of speech and transients, and is used in this work to separate noise from the non-stationary part of the signal, i.e. speech and transients. Therefore, the proposed features allows for the separation of speech from noise using the weighting scalar and from transients, using the second order of the scattering transform.

### B. Voice Activity Detection

We base the estimation scheme on the SVM procedure. Originally, this procedure provides a binary classification of feature vectors according to their position with respect to a hyperplane, which is optimized using the labeled training data

to maximize inter class separation. In this paper, we propose a continuous measure for voice activity which is based on the distance of the tested features to the hyperplane such that the classification is given by comparing the measure to a threshold. The advantage of a continuous measure over a binary classification is that the threshold value, which controls the tradeoff between false alarm and correct detection rates, may be adjusted to a specific application. Let $\mathbf{n} \in \mathbb{R}^{\mathbb{K}}$ be the normal vector (not necessarily normalized) to the hyperplane, and let $b$ be a parameter such that $b/||\mathbf{n}||$ is the offset of the hyperplane from the origin, where $||\cdot||$ is the $L_2$ norm. The distance of a tested feature vector $\mathbf{y}_t$ from the hyperplane, denoted by $L_t$, is given by:

$$L_t = \frac{<\mathbf{y}_t, \mathbf{n}> + b}{||\mathbf{n}||}.$$

Note that for simplicity, we relate to a linear SVM, while the extension to a kernel SVM is straightforward. In a binary classification, $\mathbf{y}_t$ is classified according to the sign of $L_t$ such that $\mathbf{y}_t$ is considered as a speech frame if (say) $L_t > 0$ and as a non-speech frame otherwise. In this work we propose a continuous measure for voice activity which exploits the dynamical range of $L_t$ rather than its sign, and in particular, we assume that large values of $L_t$ indicate on high probability of voice activity in frame $t$. To define the voice activity measure, we first reduce the dynamical range of $L_t$ by applying a soft threshold. The distance with a reduced dynamical range is denoted by $\acute{L}_t$ and is given by:

$$\acute{L}_t = \left\{ \begin{array}{ccc} L_{\min} & ; & L_t < L_{\min} \\ L_t & ; & L_{\min} < L_t < L_{\max} \\ L_{\max} & ; & L_t > L_{\max} \end{array} \right\}, \quad (8)$$

where $L_{\min}$ and $L_{\max}$ are constant distances from the hyperplane such that beyond them speech is assumed to be absent and present, respectively. $L_{\min}$ and $L_{\max}$ are empirically set to be half of the maximal negative and positive distances from the hyperplane in the training set, respectively. Then, $\acute{L}_t$ is normalized to provide values in the range of $0 \div 1$, and the normalized distance, denoted by $\tilde{L}_t$, is given by:

$$\tilde{L}_t = \frac{\acute{L}_t - L^{\min}}{L^{\max} - L^{\min}}. \quad (9)$$

The voice activity measure, denoted by $P_t$, is given by averaging $\tilde{L}_t$ over $2J+1$ temporally neighboring frames:

$$P_t = \frac{1}{2J+1} \sum_{j=t-JT}^{t+JT} \tilde{L}_j, \quad (10)$$

where $J$ is a non-negative parameter that defines the temporal neighborhood. The value of $P_t$ is in the range of $0 \div 1$, and the higher $P_t$ the higher the probability for speech presence in frame $t$. By taking into account several consecutive frames in (10), the effect of transients on the voice activity measure is attenuated since their length is assumed to be of the order of a single frame. The speech presence indicator defined in (2) is estimated by comparing the speech presence measure $P_t$ to a threshold $\alpha$ such that the estimated indicator, denoted by $\hat{\mathbb{1}}_{\mathbf{s}}(t)$, is given by:

$$\hat{\mathbb{1}}_{\mathbf{s}}(t) = \left\{ \begin{array}{ccc} 1 & ; & P_t > \alpha \\ 0 & ; & \text{otherwise} \end{array} \right\}. \quad (11)$$

## IV. Experimental Results

In this section we evaluate the performance of the proposed algorithm and compare it to the methods presented in [3], [4] and [12], which are called "Sohn", "Ramirez" and "Mousazadeh" in the plots, respectively. The algorithm is evaluated for different types of background noises, including white Gaussian noise, colored Gaussian noise and babble noise, and different types of transients e.g. door knocks and keyboard taps. The SNR is defined as the ratio between the speech energy and the background noise energy such that the latter is calculated in frames where speech is present. The transients are normalized to have the same maximal amplitude as speech. This is a common setup rather than defining a signal to transient ratio due to the short duration of the transients [12].

The simulated signals are sampled at 16 kHz and are processed in consecutive time frames of length $T = 32$ ms (512 samples) with $50\%$ overlap. The speech utterances used in the experiments are taken from TIMIT database [19]. The training set is composed of 20 speech utterances, and the test set, is composed of different 30 speech utterances. Each utterance is approximately of 9 s long and following the experimental setup in [12] is composed of three parts. The first part contains speech and background noise (without transients), the second part contains background noise and transients (without speech) and the third part contains the all three signals- speech, background noise and transients.

For the implementation of the proposed algorithm, we use the scattering transform library available in [20]. We exploit the Morlet wavelet similarly to [14] and set the quality factor to a small value $Q = 1$ for both the first and the second orders of the transform. Note that this choice of the quality factor $Q$ provide filters with a small time support and they better characterize transients which are assumed to be of a short length. In addition, we use filters with a central frequency $\lambda > 2\pi/T$ such that the filter bank $\{\psi_\lambda\}_\lambda$ adequately covers the frequency axis. The number of the coefficients of a single frame for this setting is 9 and 36 for the first and the second orders of the scattering transform, respectively. The normalization parameter in (7) is set to $\epsilon = 3$, as was proposed in [12]. For the voice activity measure, the hyperplane of the SVM is optimized using standard MATLAB software using a Gaussian kernel with a variance $\sigma^2 = 1$ and the soft margin parameter is set to 1. We remark that these parameters are set to the default values of the software, they may be further optimized using a validation set to improve the classification results, and their optimization is not in the scope of this paper. In addition, we empirically set the smoothing parameter in (10) to $J = 2$, which induce a lag of 32 ms.

Both for training the algorithm and for evaluating its performance on the test set, a ground truth is set according to

the clean speech signal. A frame is considered as a speech frame if the energy of the clean signal in the frame is above a certain threshold $\tilde{\alpha}$. Namely, the speech indicator defined in (2) is given by:

$$\mathbb{1}_{\mathbf{s}}(t) = \left\{ \begin{array}{ccc} 1 & ; & ||x_t||^2 > \tilde{\alpha} \\ 0 & ; & \text{otherwise} \end{array} \right\} \tag{12}$$

where $x_t$ is the clean speech signal in frame $t$. The threshold $\tilde{\alpha}$ is set as the maximal threshold such that thresholding the speech signal has negligible auditory effect [12].

The performance of the algorithms is evaluated in the form of Receiver Operating Characteristic (ROC) curves, i.e. plots of probability of detection versus the probability of false alarm. The ROC curves are generated by sweeping the threshold over all possible values of the voice activity measure $P_t$ in (10). We use two types of probabilities of false alarm as in [12]. The first is denoted by $\mathrm{P_{fa}}$ and is defined as the probability that a non-speech frame (which may contain a transient or may not) is detected as a speech frame. The second is denoted by $\mathrm{P_{fatr}}$ and is defined as the probability that a non-speech frame that contains a transient is wrongly detected as a speech frame. Namely, $\mathrm{P_{fa}}$ allows for evaluating the general performance of the algorithms, while $\mathrm{P_{fatr}}$ provides an insight on the performance of the algorithms in frames where transients are present. Note that the ground truth for the transients which is used for the evaluation of $\mathrm{P_{fatr}}$ is set in a similar way to the speech presence ground truth in (12). In addition, we evaluate the performance of the algorithm in terms of the Area Under the Curve (AUC) score, which is a scalar measure given by integrating the probability of detection over all values of false alarms. The AUC score of each method is given in percents in the legend box of each plot, and the higher the AUC the better the performance of the algorithm.

The experimental results are presented in Figures 1 to 3. It can be seen that both the proposed method and the method presented in [12], which are specifically designed to perform in a highly non-stationary acoustic environment, significantly outperform the methods presented in [3] and [4]. In addition, the proposed method outperforms the method presented in [12], and in particular provides higher classification results in the presence of transients as demonstrated by the plots with the second type of false alarm $\mathrm{P_{fatr}}$.

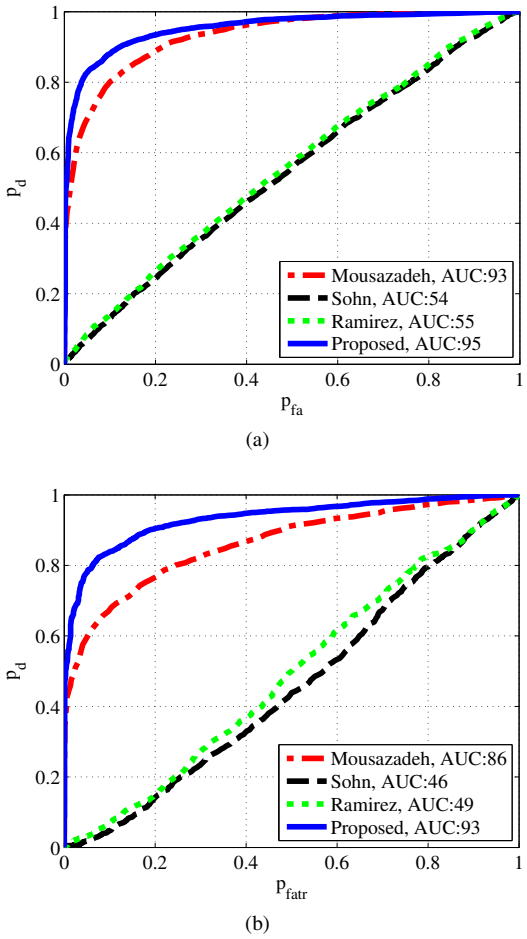



Fig. 1: (a) Probability of detection vs probability of false alarm $(\mathrm{P_{fa}})$, and (b) Probability of detection vs probability of false alarm in transient frames $(\mathrm{P_{fatr}})$. Test for a Gaussian noise with 0 dB SNR and keyboard typing transients.

## V. Conclusions

We have presented a supervised learning algorithm for voice activity detection. The algorithm incorporates features extraction procedure, where the features are based on the scattering transform, and allow for a good separation between speech frames and non-speech frames which contain transients. In addition, the features incorporate noise estimation procedure and low weights are assigned to non-speech frames which contain background noise, separating them from the speech frames. The features are used to define a continuous measure for voice activity based on the SVM classifier. The proposed algorithm outperforms state-of-the-art VADs and in particular provides enhanced voice activity detection in presence of transients.

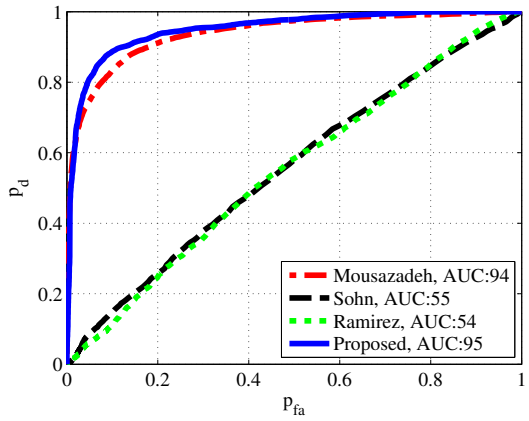
## References


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

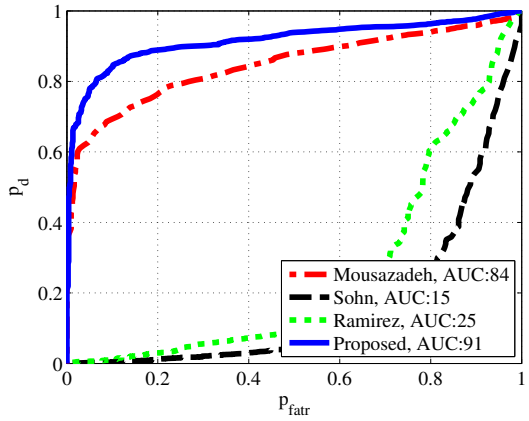
[2] A.l Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin, and J.P. Petit, "Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[3] 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.

[4] 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.

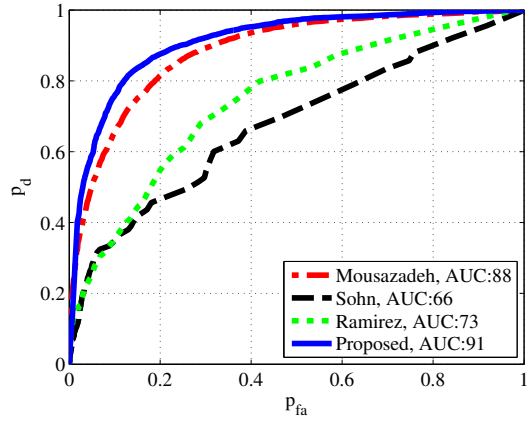



(a)

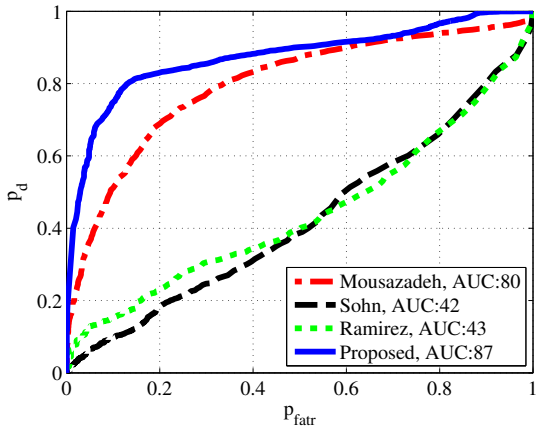



(b)

Fig. 2: (a) Probability of detection vs probability of false alarm $(\mathrm{P_{fa}})$, and (b) Probability of detection vs probability of false alarm in transient frames $(\mathrm{P_{fatr}})$. Test for a colored Gaussian noise with 0 dB SNR and scissors transients.




(a)




(b)

Fig. 3: (a) Probability of detection vs probability of false alarm $(\mathrm{P_{fa}})$, and (b) Probability of detection vs probability of false alarm in transient frames $(\mathrm{P_{fatr}})$. Test for a babble noise with 0 dB SNR and door knocks transients.


[5] 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.

[6] J. H. Chang, J. W0 Shin, and N. S. Kim, "Likelihood ratio test with complex laplacian model for voice activity detection.," in *INTERSPEECH*, 2003.

[7] 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.

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

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

[10] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the om-lsa algorithm," in *Proceedings of IWAENC 2012, International Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[11] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *Submitted to IEEE Transactions on Audio, Speech, and Language Processing*, 2014.

[12] 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.

[13] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[14] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug 2014.

[15] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, "Representing environmental sounds using the separable scattering transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 2013, pp. 8667–8671.

[16] B. Logan, "Mel frequency cepstral coefficients for music modeling.," in *ISMIR*, 2000.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] 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.

[19] 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.

[20] *[Online]. Available: http://www.cmap.polytechnique.fr/scattering*.


-