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

VOICE ACTIVITY DETECTION IN PRESENCE OF TRANSIENT NOISE USING SPECTRAL CLUSTERING AND DIFFUSION KERNELS

Oren Rosen, Saman Mousazadeh and Israel Cohen
Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel
(roseno@tx , smzadeh@tx , icohen@ee).technion.ac.il

Abstract—**In this paper, we introduce a voice activity detection (VAD) algorithm based on spectral clustering and diffusion kernels. The proposed algorithm is a supervised learning algorithm comprising of learning and testing stages: A sample cloud is produced for every signal frame by utilizing a moving window. Mel-frequency cepstrum coefficients (MFCCs) are then calculated for every sample in the cloud in order to produce an MFCC matrix and subsequently a covariance matrix for every frame. Utilizing the covariance matrix, we calculate a similarity matrix using spectral clustering and diffusion kernels methods. Using the similarity matrix, we cluster the data and transform it to a new space where each point is labeled as speech or non-speech. We then use a Gaussian Mixture Model (GMM) in order to build a statistical model for labeling data as speech or non-speech. Simulation results demonstrate its advantages compared to a recent VAD algorithm.**

I. INTRODUCTION

Voice activity detection (VAD) is an important component in echo cancellation [1], speech recognition[2] and telephony [3] applications. Common VAD implementations include the G.729 [4] and Global System for Mobile Communications (GSM) [5] standards. GSM includes two VAD operations: First, computation of the Signal-to-Noise Ratio (SNR) in nine bands and applying a threshold to these values. Second, calculation of various parameters such as noise and channel power and voice metrics. The algorithm then thresholds the voice metrics using a varying threshold which varies according to the estimated SNR. These standards demonstrate a fair performance and therefore are widely used in communication applications. Nevertheless, their performances degrade in the presence of environmental noise, even for relatively high SNR values. To overcome this shortcoming, several statistical model based VAD algorithms have been proposed in the last two decades. Sohn et al. [6] assumed that the spectral coefficients of the noise and speech signal can be modeled as complex Gaussian random variables, and proposed a VAD algorithm based on the likelihood ratio test (LRT). Although much progress has been made [7], [8], [9], [10], [11] in improving VAD algorithms performance in the presence of environmental noise, overcoming transient noise still remains a big obstacle. Transient noises such as keyboard strokes and door knocks are characterized as fast varying signals and often labeled as speech. Furthermore, using a supervised learning algorithm implies using a database for learning proposes. Large databases require substantial storage space and increase the computational complexity [12].

In this paper, we propose a voice activity detection algorithm using both spectral clustering [13] and diffusion kernel [14] methods. First, a sample cloud is created: Once a frame is sampled, we utilize a moving window in order to calculate MFCCs for a small part of the frame. The algorithm generates an MFCC matrix for the entire frame in this fashion. We then calculate a covariance matrix for the MFCCs in order to determine the correlation between samples in the sample cloud. Using the above, we calculate a similarity matrix between frames and cluster the input signal into two classes, speech or non-speech. The online testing stage receives a new input signal, it utilizes both the output information of the offline training stage and a new similarity matrix calculated in the same fashion as the above in order to cluster the data as speech and non-speech.

The rest of this paper is organized as follows. In Section 2 we formulate the problem. Section 3 presents simulation results of the proposed VAD algorithm. Finally, Section 4 concludes the paper.

II. PROBLEM FORMULATION

In this section, we elaborate the discussion on the theory of the proposed algorithm. The proposed VAD system utilizes spectral clustering [13] and diffusion kernel [14] methods in order to find a novel way of calculating a similarity matrix.

Let $x_{\mathrm{sp}}(n)$ denote a speech signal and let $x_{\mathrm{tr}}(n)$ and $x_{\mathrm{st}}(n)$ denote additive interfering transient and stationary noise signals, respectively. The microphone input signal is given by

$$y(n) = x_{\mathrm{sp}}(n) + x_{\mathrm{tr}}(n) + x_{\mathrm{st}}(n). \tag{1}$$

The proposed VAD algorithm generates a cloud of samples for each short frame (approximately 20 ms long) of the input signal, calculates MFCCs for each sample in order to from an MFCC matrix. Then, the algorithm calculates a covariance matrix for each frame. The covariance matrix adds additional factor of similarity between frames, which is utilized for the calculation of a similarity matrix.

A. Sample Cloud

Given an audio input signal, the algorithm divides it into N frames, approximately 20 ms long. A moving window of size M is then utilized in order to generate i new samples of the frame. Every sample of the frame is regarded as an iteration of the generation process for the sample cloud given by

$$\hat{y}_j(n-k) = y(n) \cdot w_M(n-k-M\cdot(j-1)+\frac{i}{2}),\ k=1,...,i \tag{2}$$

where $\hat{y}_j(n-k)$ is the generated sample cloud of the j-th frame (out of N), $y(n)$ is the original signal, w_M is the moving window of size M, i is the desired number of samples per frame and k is the current iteration index. For every iteration k of the sample cloud, the algorithm calculates m MFCCs. MFCCs are coefficients that form a representation of the short-term power spectrum of a sound. MFCC is based on a linear cosine transformation of a log power spectrum on a non-linear Mel scale frequency, thus convenient for human auditory applications. An $m \times i$ matrix of MFCCs is created in this fashion for each frame of the N frames. Finally, for each of the MFCC matrices the algorithm calculates a $m \times m$ covariance matrix.

Let $\boldsymbol{X}_m^j$ be the $m \times i$ matrix of MFCCs of the j-th frame. The covariance matrix for the j-th frame is given by

$$\boldsymbol{\Sigma} = \mathbb{E}\left(\left(\boldsymbol{X}_m^j - \mathbb{E}\left(\boldsymbol{X}_m^j\right)\right)\left(\boldsymbol{X}_m^j - \mathbb{E}\left(\boldsymbol{X}_m^j\right)\right)^T\right) \tag{3}$$

Where $\mathbb{E}$ denotes the expected value of a matrix. With the covariance matrix $\boldsymbol{\Sigma}$, we find the correlation between samples in the sample cloud.

B. Similarity Matrix

The most important part of the proposed VAD algorithm is the similarity matrix. The similarity matirx is utilized in order to effectively cluster the data and label it as speech or non-speech. Given an audio input signal composed of a combination of speech, stationary noise and transient noise components (i.e., $x_{\rm sp}(n)$, $x_{\rm st}(n)$ and $x_{\rm tr}(n)$, respectively), we choose absolute value of MFCCs and the arithmetic mean of the log-likelihood ratios for the individual frequency bins as the feature space, as in [13].

Let $\boldsymbol{Y}_m(t,k)$ $(t = 1,...,N; k = 1,...,K_m)$ and $\boldsymbol{Y}_s(t,k)$ $(t = 1,...,N; k = 1,...,K_s)$ be the absolute value of the MFCC and the STFT coefficients in a given time frame, respectively. Both MFCC and STFT coefficients are computed in K_m and K_s frequency bins, respectively. Then, each frame is represented by a column vector of dimension $(K_m + 1)$ defined as follows

$$\boldsymbol{Y}(:,t) = \begin{bmatrix} \boldsymbol{Y}_{\rm m}(:,t) \\ \Lambda_{\rm t} \end{bmatrix} \tag{4}$$

where $\boldsymbol{Y}_{\rm m}(:,t)$ denotes the absolute value of MFCCs in a specific time frame t. $\Lambda_{\rm t}$ denotes the arithmetic mean of the log-likelihood ratios for frame t. The expression combines various statistical calculations on the noise in the training stage as well as STFT coefficients of the input audio signal. $\Lambda_{\rm t}$ is given by

$$\Lambda_t = \frac{1}{K_s}\sum_{k=1}^{K_s}\left(\frac{\gamma_k\left(t\right)\xi_k\left(t\right)}{1+\xi_k\left(t\right)} - \log\left(1+\xi_k\left(t\right)\right)\right) \tag{5}$$

where $\xi_k(t) = \lambda_s(t,k)/\lambda_N(t,k)$ and $\gamma_k(t) = |\boldsymbol{Y}_s(t,k)|^2/\lambda_N(t,k)$ denote the a-priori and a-posteriori SNR [15], respectively. $\lambda_s(t,k)$ denotes the variance of speech signal in the k-th frequency bin of the t-th frame and $\lambda_N(t,k)$ denotes the variance of stationary noise in t-th time frame and k-th frequency bin.

Combining (3)-(5), we can now define the expression for the similarity matrix

$$\boldsymbol{W}_\theta^\ell(i,j) = \exp\left(\sum_{p=-P}^{P} -\alpha_{\rm p}\boldsymbol{Q}(i+p,j+p)\right) \tag{6}$$

$$\begin{aligned}
&\boldsymbol{Q}\left(i,j\right) = \\
&\left[\boldsymbol{Y}_{\rm m}^\ell(:,i)\left(1-\exp\left(-\Lambda_{\rm i}^\ell/\varepsilon\right)\right) - \boldsymbol{Y}_{\rm m}^\ell(:,j)\left(1-\exp\left(-\Lambda_{\rm j}^\ell/\varepsilon\right)\right)\right] \\
&\cdot\left(\boldsymbol{\Sigma}_i^\ell + \boldsymbol{\Sigma}_j^\ell\right)^\dagger \\
&\cdot\left[\boldsymbol{Y}_{\rm m}^\ell(:,i)\left(1-\exp\left(-\Lambda_{\rm i}^\ell/\varepsilon\right)\right) - \boldsymbol{Y}_{\rm m}^\ell(:,j)\left(1-\exp\left(-\Lambda_{\rm j}^\ell/\varepsilon\right)\right)\right]^T
\end{aligned} \tag{7}$$

where $\boldsymbol{\theta} = [\epsilon, \alpha_{-P}, \alpha_{-P+1}, \cdots, \alpha_{P-1}, \alpha_P] \in \mathbb{R}^{2P+2}$ is a vector of system parameters, $\boldsymbol{Y}_{\rm m}^\ell(:,i)$, $\Lambda_{\rm i}^\ell$ and $\boldsymbol{\Sigma}_{\rm i}^\ell$ are the absolute value of the MFCC, the arithmetic mean of the log-likelihood ratio and the covariance matrix of the i-th frame in the ℓ-th sequence, respectively, ϵ is the kernel width obtained during the training stage and † denotes the pseudo-inverse of a matrix. The main motivation behind the proposed similarity matrix calculation in (7) is finding a model for the signal generating system, i.e. the speech system of the speaker. With the new representation, we gain a smaller degree of freedom for the system model. We tag the system as a "black box" and try to find a model for the system by viewing its' outputs. In fact, a second order approximation is applied on the parameters in order to receive random Gaussian perturbations. A covariance matrix is then calculated and used in order to express a Jacobian matrix. Finally, the Jacobian is used in order to find a similarity matrix. In order to calculate the pseudo-inverse of the expression $\boldsymbol{\Sigma}_{\rm i}^\ell + \boldsymbol{\Sigma}_{\rm j}^\ell$ in (7), we use the first three largest eigenvectors received in singular vector decomposition (SVD).

Let $\boldsymbol{\Sigma}$ be a covarince matrix as in (3), applying SVD yields

$$\boldsymbol{\Sigma}^\dagger = \boldsymbol{V}\boldsymbol{S}^\dagger\boldsymbol{\Delta}^T \tag{8}$$

Where $\boldsymbol{\Delta}$ is an orthogonal matrix of size $3 \times N$, the columns of $\boldsymbol{\Delta}$ are the eigenvectors of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$. $\boldsymbol{S}$ is a diagonal matrix at the same size of $\boldsymbol{\Sigma}$, its' values are the square roots of the non-zero eigenvalues of both $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$ and $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$. $\boldsymbol{V}$ is an orthogonal matrix, the same size of $\boldsymbol{\Delta}$. The columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$.

C. Training Stage

The training algorithm in our paper is based on [13]. Given databases of clean speech, transient noise and stationary noise signals, We choose L different signals from each database. Without loss of generality, we take the ℓ-th speech signal, transient noise and stationary noise and combine them as in Fig. 1. We assume that all of these signal are of the same length of N_ℓ. With the new database and by utilizing (4) and (5), we extract the feature matrix $\boldsymbol{Y}_1^\ell$, $\boldsymbol{Y}_2^\ell$, $\boldsymbol{Y}_3^\ell$. By concatenating the feature matrix, we build the ℓ-th training sequence, $\boldsymbol{Y}^\ell$, as

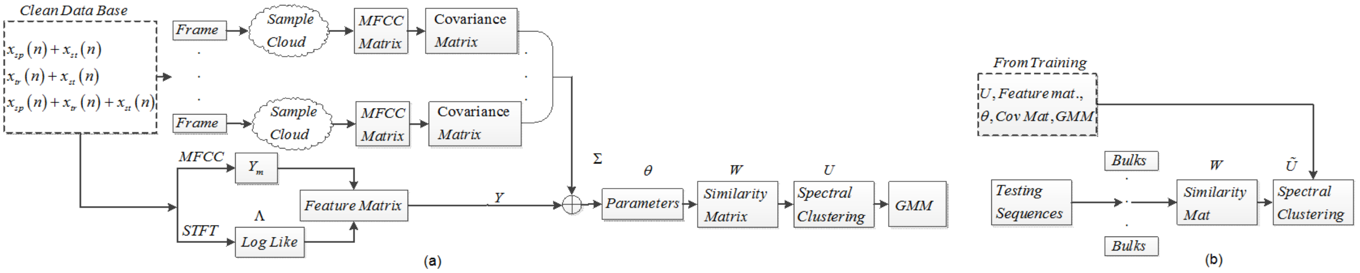


Fig. 1: A block scheme of the proposed (a) training, and (b) testing stages.

shown in Fig. 1. For each frame t, in the training sequence ℓ we compute an indicator matrix $\boldsymbol{C}_{\mathrm{t}}^{\ell}$ in order to indicate a speech containing frame. For further discussion, see [13].

Next, we define a kernel which preserves the similarity between points, as the similarity matrix in (6) and (7). This metric guarantees small similarity between two frames of different classes, i.e., speech and transient noise, even if they are very similar to each other (in the Euclidean sense). This is enabled due the large distance between neighboring frames. Upon defining the parametric weight function, the parameters can be obtained by solving the following optimization problem [16]:

$$\boldsymbol{\theta}^{opt} = \arg\min_{\theta} \frac{1}{L} \sum_{\ell=1}^{L} F(\boldsymbol{W}_{\boldsymbol{\theta}}^{\ell}, \boldsymbol{C}^{\ell}) \tag{9}$$

$$F(\boldsymbol{W}, \boldsymbol{C}) = \frac{1}{2} \left\| \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T - \boldsymbol{D}^{1/2}\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{D}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{D}^{1/2} \right\|_F^2 \tag{10}$$

where $\boldsymbol{\Upsilon}$ is an approximate orthonormal basis of the projections on the second principal subspace of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$.

Let $\boldsymbol{W}_{\boldsymbol{\theta}^{opt}}^{\ell}$ be the similarity matrix of the ℓ-th training sequence and $\boldsymbol{U}_{\ell}$ be a matrix consisting of the two eigenvectors of $\boldsymbol{D}^{\ell^{-1/2}}\boldsymbol{W}^{\boldsymbol{\ell}}\boldsymbol{D}^{\ell^{-1/2}}$ corresponding to the first two largest eigenvalues, where $\boldsymbol{D}$ is a diagonal matrix whose i-th diagonal element equals to $\sum_{j=1}^{N} \boldsymbol{W}(i,j)$. We then define $\boldsymbol{U}$ as the column concatenation of $\boldsymbol{U}_1$ through $\boldsymbol{U}_L$. $\boldsymbol{U}$ is a new representation of the training data such that each row of U corresponding to a specific training frame. For further information, see [13].

We use Gaussian mixture modeling to model each cluster, i.e., label as speech presence or absence, with a different Gaussian Mixture Model (GMM). This means that we model the low dimensional representation of the original data using two different GMMs, one for each cluster. Let $f(\cdot;\mathcal{H}_0)$ and $f(\cdot;\mathcal{H}_1)$ be the probability density function corresponding to speech absence and presence frames, respectively. The likelihood ratio for each labeled frame t is then obtained by

$$\Gamma_{\mathrm{t}}^{\mathrm{train}} = \frac{f(\boldsymbol{U}(t,:);\mathcal{H}_1)}{f(\boldsymbol{U}(t,:);\mathcal{H}_0)} \tag{11}$$

where $\boldsymbol{U}(t,:)$ is the t-th row of the matrix $\boldsymbol{U}$, and $\mathcal{H}_1$ and $\mathcal{H}_0$ are the speech presence and absence hypotheses, respectively.

D. Testing Stage

The main goal of the testing stage is to to cluster the unlabeled data and decide whether a given unlabled frame contains speech or not. In order to compute the likelihood ratio for a new unlabled frame, [13] utilizes GMM to model the eigenvectors of normalized Laplacian matrix.

$$\Gamma_{\mathrm{t}}^{\mathrm{test}} = \frac{f(\tilde{\boldsymbol{U}}(t,:);\mathcal{H}_1)}{f(\tilde{\boldsymbol{U}}(t,:);\mathcal{H}_0)} \tag{12}$$

where $\tilde{\boldsymbol{U}}(t,:)$ is the t-th row of the new representation of the unlabeled data in terms of eigenvectors of the normalized Laplacian of the similarity matrix. In [11] it was shown that using the information supplied by neighboring frames can improve the performance of VAD algorithms. The improvement is enabled due to the fact that frames containing speech signal are usually followed by a frame that contains speech signal as well. In the contrary, transient signals usually last for a single time frame. Using this fact, the decision rule for an unlabeled time frame is obtained by:

$$\mathrm{VA}_t = \sum_{j=-J}^{J} \Gamma_{t+j} \underset{\boldsymbol{\mathcal{H}}_0}{\overset{\boldsymbol{\mathcal{H}}_1}{\gtrless}} T_h \qquad t = 1, 2, \cdots, T \tag{13}$$

where T_h is a threshold which controls the tradeoff between probability of detection and false alarm. Increasing (decreasing) this parameter leads to a decrease (increase) of both the probability of false alarm and the probability of detection. Both the training and testing stages are summarized in Table I. The block schemes of both learning and testing stage are depicted in figure 1.

III. Simulations Results

In this section we demonstrate the performance and advantages of the proposed VAD algorithm via several simulations. We compare the results acquired to the results of the VAD algorithm proposed in [13]. We run the simulations for various SNR values, stationary noises and transient noises. We confiugre the number of sequences to 4 training sequences and 20 testing sequences. Speech signals are taken from the TIMIT database [17], and transient noise signals are taken from [18]. The sampling frequency is set to 16kHz. Furthermore, we pick a window of size 512 for STFT calculations, $m = 14$ mel-frequency bands, $M = 257$ as the size of the moving window

TABLE I: Proposed Voice Activity Detection Algorithm Based on Spectral Clustering Method.

```
Learning algorithm:
  1. Construct a training data set consisting of
     $L$ training signals $\{\boldsymbol{Y}^\ell \in \mathbb{R}^{K_m+1\times 3N^\ell}; \ell = 1, ..., L\}$
     and $L$ indicator vectors $\{\boldsymbol{C}^\ell \in \mathbb{R}^{3N^\ell\times 4}; \ell = 1, ..., L\}$.
  2. Solve the optimization problem given in (9), to find
     the optimum value of the parameters (i.e. $\boldsymbol{\theta}^{opt}$).
  3. Construct $\boldsymbol{U}$ by concatenation of $U^1$ through $U^L$
     $K$ largest eigenvectors of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ .
  4. Fit a GMM model to the rows of $\boldsymbol{U}$ for
     each cluster (see ([13])).
Output:
  $\boldsymbol{U}$, $f(\cdot;\boldsymbol{\mathcal{H}}_1)$ and $f(\cdot;\boldsymbol{\mathcal{H}}_0)$

Testing Procedure:
Let $z_t(n)$ be the test sequence and $\boldsymbol{Z}_t$
the feature matrix of $t$-th unlabeled frame obtained by (5).
for $t = 0 : T : N^z - T \; (T \ll N^z)$
  1. $\boldsymbol{Z} = \boldsymbol{Z}_t(:, t+1 : t+P)$.
  2. Compute $\boldsymbol{B}$ by
     $\boldsymbol{B}^\ell_{\boldsymbol{\theta}^{opt}}(i,j) = \exp\left(\sum_{p=-P}^{P} -\alpha_p^{opt}\boldsymbol{Q}^\ell(i+p, j+p)\right)$
     $\boldsymbol{B} = \left[(\boldsymbol{B}^1_{\boldsymbol{\theta}^{opt}})^T, (\boldsymbol{B}^2_{\boldsymbol{\theta}^{opt}})^T, \cdots, (\boldsymbol{B}^L_{\boldsymbol{\theta}^{opt}})^T\right]^T$
  3. Compute the new representation of the unlabeled data (12)
     $\tilde{\boldsymbol{U}} = \text{diag}\left((\boldsymbol{1}\boldsymbol{B}_{k_{nn}})^{-1}\right)\boldsymbol{B}^T_{k_{nn}}\boldsymbol{U}$.
  4. Compute the likelihood ratio for a new unlabeled frame
     $\Gamma_t = \frac{f(\tilde{U}(t,:);\boldsymbol{\mathcal{H}}_1)}{f(\tilde{U}(t,:);\boldsymbol{\mathcal{H}}_0)}$.
  5. The decision rule for an unlabeled time frame is given
     $\text{VA}_t = \sum_{j=-J}^{J}\Gamma_{t+j} \underset{\boldsymbol{\mathcal{H}}_0}{\overset{\boldsymbol{\mathcal{H}}_1}{\gtrless}} T_h$ .
  6. Use $\text{VA}_t$ to obtain the final VAD decision.
end
```

to utilize in order to create the sample cloud and $i = 45$ as the number of samples for each frame. The graphs of probability of detection, P_d, vs. probability of false alarm, P_{Fa}, are depicted in Fig 2. We use identical experiment conditions with both the proposed algorithm and [13] in every simulation. In Fig 3, we provide the clustering results, i.e. the U space representation where speech labeled data is marked with blue rings and non-speech data is marked with red crosses. The proposed VAD algorithm has superior performance in the entire SNR range, especially for low SNR values. Moreover, the proposed algorithm performs better in cases of very small training sets.

IV. Conclusions

We have presented a VAD algorithm based on spectral clustering and diffusion kernel methods. The main challenge was providing good results in presence of environmental noise and transient noise in particular. The key features of the proposed algorithm are the covariance matrix calculations via sample clouds and the novel similarity matrix computations. We demonstrated better results compared to a work that has already been proven to be superior to conventional methods of dealing transient noises, especially in cases of low SNR and small data bases. The goal in the near future would be trying to improve the algorithm's results using enhanced features. Another possible research direction would be choosing better parameters for the GMM.

Acknowledgment

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

References

[1] AM Kondoz and BG Evans, "A high quality voice coder with integrated echo canceller and voice activity detector for vsat systems," in *Satellite Communications-ECSC-3, 1993., 3rd European Conference on*. IET, 1993, pp. 196–200.

[2] Javier Ramírez, José C Segura, Carmen Benítez, Angel De la Torre, and Antonio Rubio, "An effective subband osf-based vad with noise reduction for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 6, pp. 1119–1129, 2005.

[3] Huang Hai, Yu Hong-Tao, and Feng Xiao-Lei, "A spit detection method using voice activity analysis," in *Multimedia Information Networking and Security, 2009. MINES'09. International Conference on*. IEEE, 2009, vol. 2, pp. 370–373.

[4] Adil Benyassine, Eyal Shlomot, Huan-Yu Su, Dominique Massaloux, Claude 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," *Communications Magazine, IEEE*, vol. 35, no. 9, pp. 64–73, 1997.

[5] Jon E Natvig, Stein Hansen, and Jorge de Brito, "Speech processing in the pan-european digital mobile radio system (gsm)-system overview," in *Global Telecommunications Conference and Exhibition'Communications Technology for the 1990s and Beyond'(GLOBECOM), 1989. IEEE*. IEEE, 1989, pp. 1060–1064.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 16, pp. 1–3, 1999.

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

[8] J. W. Shin, J. H. Chang, H. S. Yun, and N. S. Kim, "Voice activity detection based on generalized gamma distribution," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. I781–I784, 2005.

[9] Youngjoo Suh and Hoirin Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection.," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 507–510, 2012.

[10] S. Mousazadeh and I. Cohen, "AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 916–926, 2011.

[11] J. Ramirez and J. C. Segura, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689–692, 2005.

[12] Dan Levi and Shimon Ullman, "Learning model complexity in an online environment," in *Computer and Robot Vision, 2009. CRV'09. Canadian Conference on*. IEEE, 2009, pp. 260–267.

[13] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering," *Accepted for publication in IEEE Trans. Audio, Speech and Signal Processing*.

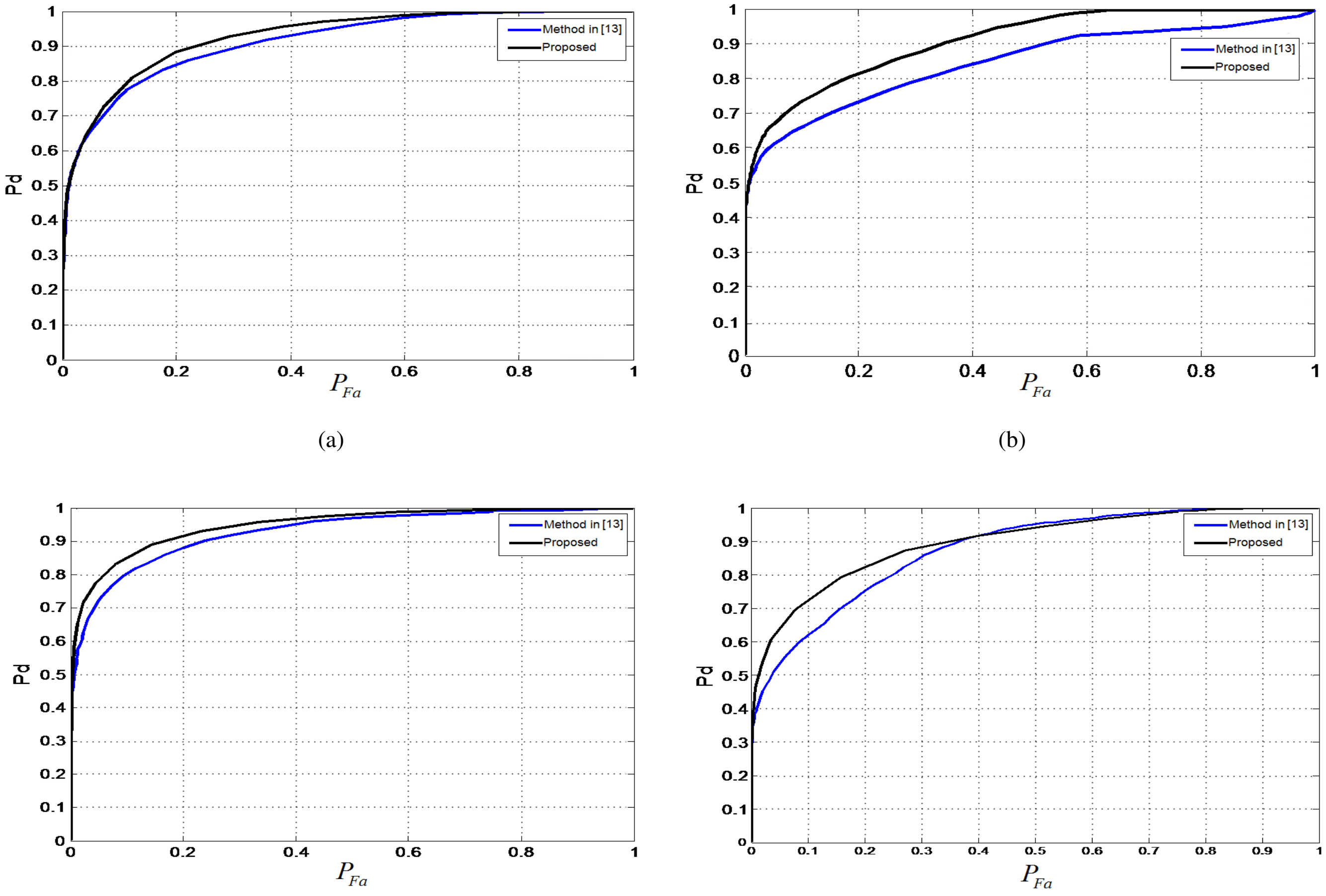
[14] Ronen Talmon, Dan Kushnir, Ronald R. Coifman, Israel Cohen, and Sharon Gannot, "Parametrization of linear systems using diffusion kernels," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1159–1173, 2012.

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

[16] Francis R. Bach and Michael I. Jordan, "Learning spectral clustering, with application to speech separation," *Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, 2006.

[17] 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.

[18] "[online]. available: http://www.freesound.org," .



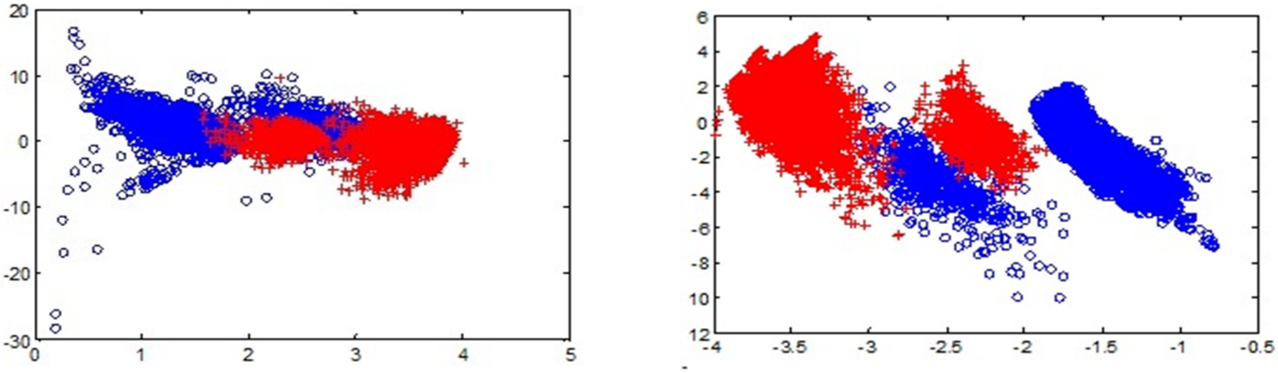
(a)

(b)

(c)

(d)

Fig. 2: Probability of detection (P_{d}) versus probability of false alarm (P_{fa}), for various noise environments. (a) Babble noise with SNR of 5dB, and transient noise of door knocks. (b) White noise with SNR of 5dB, and transient noise of door knocks. (c) White noise with SNR of 10dB, and transient noise of typing. (d) Babble noise with SNR of 5dB, and transient noise of typing.



(a)

(b)

Fig. 3: Clustering results of training stage for babble noise with SNR of 5dB, and transient noise of typing of (a) the proposed algorithm, and (b) the algorithm proposed in [16].