VOICE ACTIVITY DETECTION IN TRANSIENT NOISE ENVIRONMENT USING LAPLACIAN PYRAMID ALGORITHM

Nurit Spingarn, Saman Mousazadeh and Israel Cohen

Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel
E-mail addresses: (nurits@tx , smzadeh@tx , icohen@ee).technion.ac.il

ABSTRACT

Voice activity detection (VAD) has attracted significant research efforts in the last two decades. Despite much progress in designing voice activity detectors, voice activity detection in presence of transient noise and low SNR is a challenging problem. In this paper, we propose a new VAD algorithm based on supervised learning. Our method employs Laplacian pyramid algorithm as a tool for function extension. We estimate the likelihood ratio function of unlabeled data, by extending the likelihood ratios obtained from the labeled data. Simulation results demonstrate the advantages of the proposed method in transient noise environments over conventional statistical methods.

Index Terms— Voice activity detection, Likelihood ratio function, transient noise, Laplacian pyramid algorithm

1. INTRODUCTION

Voice activity detection (VAD) is required in many speech communication applications, such as speech recognition, speech coding, hands-free telephony, speech enhancement and echo cancellation. Elementary VAD algorithms rely on averaged parameters over frames such as zero crossing rate, pitch period, autocorrelation coefficients, energy levels, etc. These methods have applicable results for clean speech signals, but in noisy environments their performance severely degrades. To overcome this shortcoming, several statistical model based VAD algorithms have been proposed in the last two decades. Sohn et al. [1] assumed that the spectral coefficients of the noise and speech signal can be modeled as complex Gaussian random variables, and developed a VAD algorithm based on the likelihood ratio test (LRT). Following their work, researchers tried to improve the performance of model-based VAD algorithms by assuming different statistical models for speech signals, see [2, 3, 4, 5, 6]. While these methods perform well in stationary noisy environments, as long as the signal-to-noise ratio (SNR) is not too low, their performances degrade significantly in presence of transient noise, such as coughing, sneezing, keyboard strokes and door knocking sounds.

In this paper, we present a new supervised learning VAD algorithm based on the Laplacian pyramid algorithm. Training data is used for estimating the parameters of the similarity matrix kernel and for the Laplacian pyramid representation. The training data is also used in finding two Gaussian mixture models for modeling the first two eigenvectors of the Laplacian of the similarity matrix corresponding to the first two leading eigenvalues of the normalized Laplacian matrix. Upon receiving new unlabeled data, the Laplacian pyramid algorithm is used for evaluating the likelihood ratio. The final VAD is obtained by comparing that likelihood ratio to a threshold.

The rest of this paper is organized as follows. In Section 2, we formulate the voice activity detection problem in transient noisy environments and introduce our VAD. Simulation results and performance evaluation are presented in Section 3. Finally, we conclude the paper in Section 4.

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

2. PROBLEM FORMULATION

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

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

The goal is to determine whether there exists a speech signal in a given time frame (each approximately 16-20 msec long).

2.1. Feature Selection

The main purpose in feature representation is to encapsulate the relevant characteristics of a speech signal for the detection process. Here we choose absolute value of Mel-frequency cepstrum coefficient (MFCCs) and the arithmetic mean of the log-likelihood ratios for the individual frequency bins as our feature space. The likelihood ratio has been long exploited as a feature for voice activity detection in presence of stationary noise [1, 2, 3, 4, 5]. MFCC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. MFCCs are commonly used as features in speech recognition systems. Combining these two features appropriately, would be a suitable feature space for voice activity detection in presence of transient noise. See [7] for a discussion on this issue. The feature vector for frame t is defined as column concatenation of two components as follows

$$\boldsymbol{Y}(:,t) = \begin{bmatrix} \boldsymbol{Y}_{\mathrm{m}}(:,t) \\ \Lambda_{\mathrm{t}} \end{bmatrix} \tag{2}$$

where $\boldsymbol{Y}_{\mathrm{m}}(:,t)$ is the absolute value of MFCCs and Λ_{t} is the arithmetic mean of the log-likelihood ratios for the individual frequency bands in frame t.

2.2. Training Stage

The first stage in our learning algorithm is calculating the likelihood ratio function for training data. The second stage is extraction of essential parameters needed for the Laplacian pyramid algorithm. Suppose that we have a database of clean speech signal, a database

of transient noise and a database of stationary noise. We choose L different signals from each database and combine them as follows. Let $x^{\ell}_{\rm sp}(n), x^{\ell}_{\rm tr}(n), x^{\ell}_{\rm st}(n)$ be the ℓ-th speech signal, transient noise, and stationary noise, respectively. Without loss of generality, we assume that all of these signal are of the same length (i.e. N_ℓ). We build the ℓ-th training sequence, $\boldsymbol{Y}^{\ell}$, as follows. Let

$$x^{\ell}_1(n) = x^{\ell}_{\rm sp}(n) + x^{\ell}_{\rm st}(n), \quad (3)$$

$$x^{\ell}_2(n) = x^{\ell}_{\rm tr}(n) + x^{\ell}_{\rm st}(n), \quad (4)$$

$$x^{\ell}_3(n) = x^{\ell}_{\rm sp}(n) + x^{\ell}_{\rm tr}(n) + x^{\ell}_{\rm st}(n), \quad (5)$$

and let $\boldsymbol{Y}^{\ell}_1$, $\boldsymbol{Y}^{\ell}_2$, $\boldsymbol{Y}^{\ell}_3$ be the feature matrix extracted using (2) from $x^{\ell}_1(n), x^{\ell}_2(n), x^{\ell}_3(n)$, respectively, and $\boldsymbol{Y}^{\ell}$ be the row concatenation of these matrices. For each frame t, in the training sequence l we compute an indicator vector as follows

$$\boldsymbol{C}^{\ell}_{\rm t} = \begin{cases} 1 & P(\boldsymbol{X}^{\ell}_{\rm sp}(:,t)) > \delta_{\rm sp} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $\boldsymbol{X}^{\ell}_{\rm sp}(:,i)$ is short time fourier transform (STFT) of $x^{\ell}_{\rm sp}(n)$, $\delta_{\rm sp}$ is a threshold and $P(.)$ is the power calculation operator. The most important issue in every kernel based method (e.g. spectral clustering) is defining an appropriate kernel which preserves the similarity between points. For our problem we define the parametric kernel as follows:

$$\boldsymbol{W}^{\ell}_{\theta}(i,j) = \exp\left(\sum_{p=-P}^{P} -\alpha_{\rm p}\boldsymbol{Q}(i+p, j+p)\right) \quad (7)$$

$$\boldsymbol{Q}(i,j) = \|\boldsymbol{Y}^{\ell}_{\rm m}(:,i)\left(1-\exp(-\Lambda^{\ell}_{\rm i}/\epsilon)\right) - \boldsymbol{Y}^{\ell}_{\rm m}(:,j)\left(1-\exp(-\Lambda^{\ell}_{\rm j}/\epsilon)\right)\|^2_2 \quad (8)$$

where $\boldsymbol{\theta} = [\epsilon, \alpha_{-P}, \alpha_{-P+1}, \cdots, \alpha_{P-1}, \alpha_P] \in \mathbb{R}^{2P+2}$ is a vector of parameters, $\boldsymbol{Y}^{\ell}_{\rm m}(:,i)$ and $\Lambda^{\ell}_{\rm i}$ are the absolute value of the MFCC and the arithmetic mean of log likelihood ratios for the individual frequency bands of the ℓ-th training sequence in frame i, respectively. The choice of weight matrix is taken into account the following issues. The first one is the similarity between two individual frames, and the second one is the effect of neighboring frames on deciding whether a specic frame contains speech or transient noise. By Combining the two features (MFCC and likelihood ratio) as in (7)-(8) lead to a good metric utilized as a similarity notion between two frames for voice activity detection in noisy environment. Moreover, it can be seen from (8) that when the current frame is a non speech frame, the value of the likelihood ratio is around zero. Hence, the term $\left(1-\exp(-\Lambda^{\ell}_{\rm j}/\epsilon)\right)\|^2_2$ tend to 1 and the similarity matrix is depend only on the MFCC. The kernel parameters can be obtained by solving the following optimization problem [8]

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

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

where L is the number of training sequences, $(\cdot)^T$ denotes transpose of a vector or a matrix, $\boldsymbol{\Upsilon}$ is an approximate orthonormal basis of the projections on the second principal subspace of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ obtained by classical *orthogonal iteration* [9]. In practice we use the gradient method (e.g. *fminunc* or *fmincon* functions in Matlab if there exists any constraint on the parameters) to solve this minimization problem. Let $\boldsymbol{W}^{\ell}_{\boldsymbol{\theta}^{opt}}$ 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 $\sum_{j=1}^{N}\boldsymbol{W}(i,j)$. Suppose we have L training sequences, let the column concatenation of $\boldsymbol{U}_1$ through $\boldsymbol{U}_L$ be $\boldsymbol{U}$. In fact, the matrix $\boldsymbol{U}$ is a new representation of the training data such that each row of U corresponding to a specific training frame. For further information, see [7]. Now, we use Gaussian mixture modeling to model each cluster (i.e. speech presence or absence) with a different Gaussian Mixture Model (GMM). For each cluster we find the rows of the matrix $\boldsymbol{U}$ corresponding to that cluster using an indicator vector. Then by exploiting the Expectation-Maximization (EM) algorithm and Bayesian information criterion (BIC), we fit GMMs in those clusters. This means that we model the low dimensional representation of the original data using two different GMMs, one for each cluster. The likelihood ratio for each labeled frame t is then obtained by

$$\Gamma^{\rm train}_{\rm t} = \frac{f(\boldsymbol{U}(t,:);\mathcal{H}_1)}{f(\boldsymbol{U}(t,:);\mathcal{H}_0)} \quad (10)$$

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. The second stage as mentioned earlier is an implementation of the Laplacian pyramid algorithm. The algorithm used here is based on [10]. The Laplacian pyramid is a multi-scale algorithm for extending an empirical function f, which is defined on a dataset A, to new data points. In our case, this function is the likelihood ratio function ($\boldsymbol{\Gamma}^{\rm train}$), and the dataset A is the collection of the labeled feature vectors.

First, we represent our likelihood function in different resolutions approximated by mutual distances between the data points. In order to emphasize the mutual distances between the feature vectors we use the same kernel as formulated in (7)-(8). The similarity matrix for each resolution level $v = 0, 1, ..$ is given by

$$\boldsymbol{W}^{\rm v}_{\theta}(i,j) = \exp\left(\sum_{p=-P}^{P} -\alpha_{\rm p}\boldsymbol{Q}^{\rm v}(i+p, j+p)\right) \quad (11)$$

$$\boldsymbol{Q}^{\rm v}(i,j) = 2^v\|\boldsymbol{Y}_{\rm m}(:,i)\left(1-\exp(-\Lambda_{\rm i}/\epsilon)\right) - \boldsymbol{Y}_{\rm m}(:,j)\left(1-\exp(-\Lambda_{\rm j}/\epsilon)\right)\|^2_2 \quad (12)$$

and the smoothing operator $\boldsymbol{K}_{\mathbf{v}}(i,j)$ is defined by

$$\boldsymbol{K}_{\rm v}(i,j) = \frac{\boldsymbol{W}^{\rm v}_{\theta}(i,j)}{\sum_{j=1}^{n}\boldsymbol{W}^{\rm v}_{\theta}(i,j)} \quad (13)$$

where n is the number of training frames. The Laplacian pyramid representation is calculated iteratively as follows

$$\boldsymbol{s}_0(t) = \sum_{j=1}^{n}\boldsymbol{K}_{\mathbf{0}}(t,j)\Gamma^{\rm train}_{\rm j} \quad (14)$$

$$\boldsymbol{s}_{\rm v}(t) = \sum_{j=1}^{n}\boldsymbol{K}_{\rm v}(t,j)\boldsymbol{d}_{\rm v}(j). \quad (15)$$

The differences are given by

$$\boldsymbol{d}_{\rm v} = \boldsymbol{\Gamma}^{\rm train} - \sum_{i=0}^{v-1}\boldsymbol{s}_{\rm i} \quad (16)$$

and are used as inputs to the Laplacian pyramid algorithm at level v. The iterations in (15) stop when $|\boldsymbol{d}_v|$ is smaller than a certain threshold.

2.3. Testing Stage

During testing, our goal is to decide whether a given unlabeled frame contains speech or not. Mousazadeh and Cohen [7] suggested to represent the unlabeled data in terms of eigenvectors of the normalized Laplacian of the similarity matrix of the training data using Nysröm extension and computing the likelihood ratio as follows

$$\Gamma_{\mathrm{t}}^{\mathrm{test}} = \frac{f(\tilde{\boldsymbol{U}}(t,:);\mathcal{H}_1)}{f(\tilde{\boldsymbol{U}}(t,:);\mathcal{H}_0)} \tag{17}$$

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 our algorithm, we propose to approximate the likelihood ratio function using the Laplacian pyramid algorithm which might be more accurate, requires less training files and has better performance in low SNR. Let $\boldsymbol{Z}(:,t)$ be the feature matrix for the t-th unlabeled frame. For each resolution level v, the similarity matrix between the new data and labeled data is computed as follows

$$\boldsymbol{W}_{\theta}^{\mathrm{v}}(i,j) = \exp\left(\sum_{p=-P}^{P} -\alpha_{\mathrm{p}}\boldsymbol{Q}^{\mathrm{v}}(i+p,j+p)\right) \tag{18}$$

$$\boldsymbol{Q}^{\mathrm{v}}(i,j) = 2^{v}\|\boldsymbol{Y}_{\mathrm{m}}(:,i)\left(1-\exp(-\Lambda_{\mathrm{i}}/\epsilon)\right) - \boldsymbol{Z}_{\mathrm{m}}(:,j)\left(1-\exp(-\Lambda_{\mathrm{j}}/\epsilon)\right)\|_2^2. \tag{19}$$

Yet, the likelihood function of labeled frames in different resolutions is given by

$$\boldsymbol{s}_0^{\mathrm{un}}(t) = \sum_{j=1}^{n} \boldsymbol{K}_{\mathbf{0}}(t,j)\Gamma_{\mathrm{j}}^{\mathrm{train}} \tag{20}$$

$$\boldsymbol{s}_{\mathrm{v}}^{\mathrm{un}}(t) = \sum_{j=1}^{n} \boldsymbol{K}_{\mathrm{v}}(t,j)\boldsymbol{d}_{\mathrm{v}}(j). \tag{21}$$

when the smoothing operator is calculated as in (13). The likelihood ratio of the t-th frame is then evaluated by

$$\Gamma_{\mathrm{t}}^{\mathrm{test}} = \sum_{k=0}^{v-1} \boldsymbol{s}_{\mathrm{k}}^{\mathrm{un}}(t). \tag{22}$$

Using the fact that frames containing speech are usually followed by a frame which also contains speech the transient signals usually last for few time frames, the decision rule for an unlabeled frame is given by

$$V_A = \sum_{j=-J}^{J} \Gamma_{\mathrm{t+j}}^{\mathrm{test}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} T_h \qquad t = 1,2,\cdots,T \tag{23}$$

where T_h is the threshold which controls the tradeoff between probability of detection and false alarm. Increasing (decreasing) this parameter leads to decrease (increase) of both the probability of false alarm and the probability of detection.

3. SIMULATIONS RESULTS

In this section, we examine the performance of the proposed method using several simulations. We compare the performance of our method to those of conventional statistical model-based methods presented in [1, 2, 3, 6] and VAD based spectral clustering presented in [7]. The simulation setup is as follows. We perform our simulation for different types of stationary and transient noise and for different SNR levels. We use different data within the training and testing phases, 5 training sequences and 30 testing sequences. Speech signals are taken from the TIMIT database [11], and transient noise signals are taken from [12]. The sampling frequency is 16kHz. We use STFT with frame length of 512 samples, with 50% overlap and a hamming window. We compute the MFCC in $K_m = 24$ Mel frequency bands and the a-priori threshold used in the Laplacian pyramid algorithm is set to be 10^{-5}. In order to compare our method to a conventional statistical based method, we introduce two different kinds of false alarm probabilities. The first one denoted by P_{fa} is defined as the probability that a speech free frame is detected as a speech frame. The second one denoted by P_{fatr} is defined as the probability that a frame consisting of stationary and transient noise is detected as a speech frame. We need these two concepts to show the advantage of the proposed method over conventional statistical model-based methods. The number of frames which contain transient noise (which are mostly detected as speech in statistical model-based methods) is small with respect to the total number of frames. Such frames do not affect the probability of false alarm significantly if it is defined as the probability that a noise frame is detected as a speech frame. In all of the following simulations we set the parameter vector to $\boldsymbol{\theta} = .001 \times [300\ \ 0.4\ \ 0.75\ \ 1\ \ 0.75\ \ 0.4]$.

The simulation results are presented in Figure 1. Although different statistical model based methods have different performances in different situations, the proposed method has superior performance in all simulations over the compared statistical model based methods, particulary for low false alarm rates. When compared to the method presented in [7], our method has better performance when the number of training sequences is small and when the SNR is low.

4. CONCLUSIONS

We have proposed a new voice activity detector based on Laplacian pyramid. Our main concern was dealing with low SNR transient noise conditions, which are difficult to handle. Conventional statistical model based methods fail in these situations. Simulation results have demonstrated the improved performance of the proposed method and particularly its advantage in treating transient noise using only few training sequences.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[5] 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.

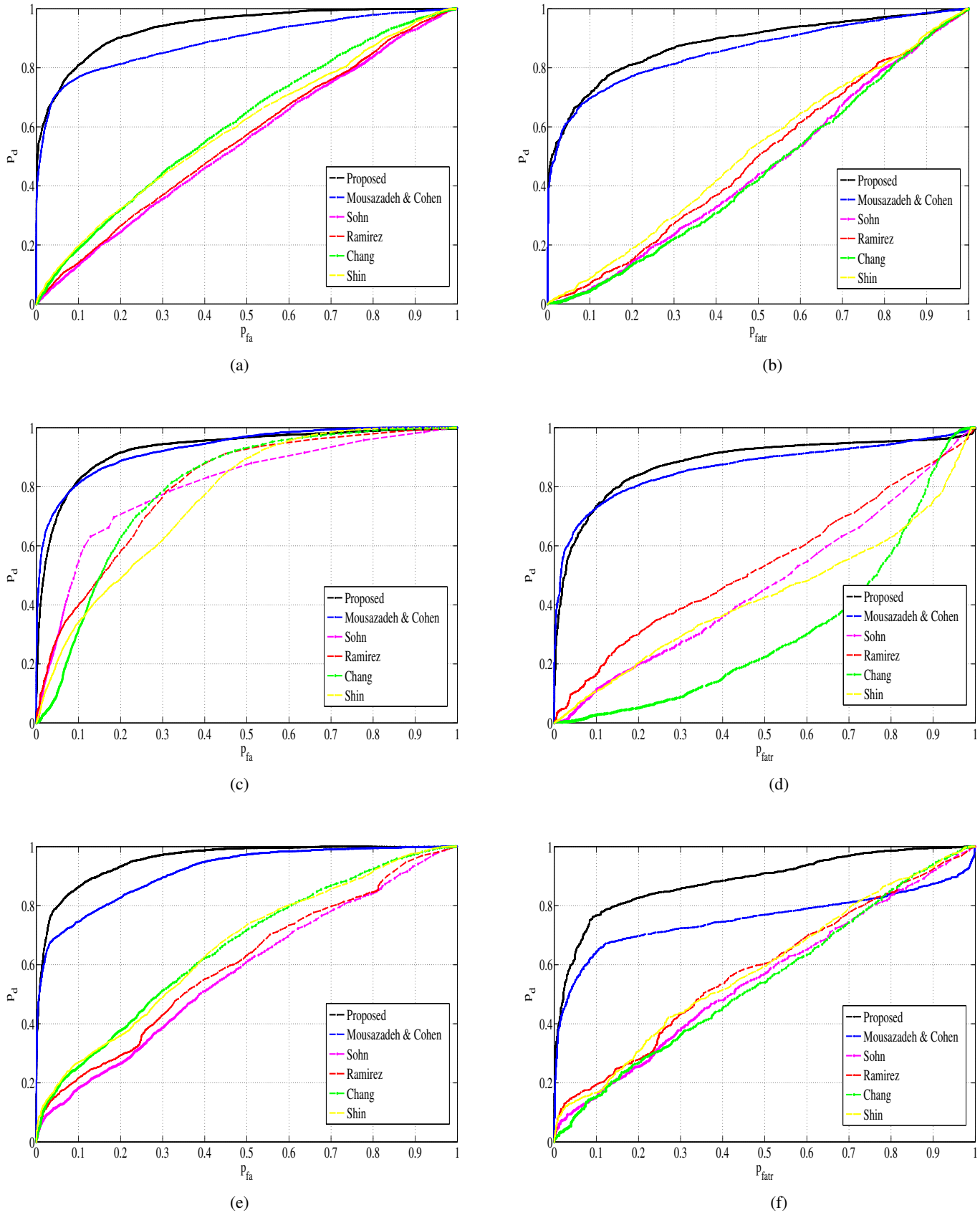

Fig. 1: P_{d} versus P_{fa} (left column), P_{d} versus P_{fatr} (right column) for different noise environments. Figures (a)-(b): SNR = 0dB, stationary noise - white Gaussian, transient noise - keyboard stroke. Figures (c)-(d): SNR = 10dB, stationary noise - babble noise, transient noise - keyboard stroke. Figures (e)-(f): SNR = 20dB, stationary noise - colored noise, transient noise - door knocks.

[6] 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.

[7] 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*.

[8] 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.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1996.

[10] N.Rabin and R.R. Coifman., "Heterogeneous datasets representation and learning using diffusion maps and laplacian pyramids," in *Proc. 12th SIAM International Conference on Data Mining., Speech Date Mining .,(SDM)*, 2012.

[11] 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.

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