INTERSPEECH 2009 BRIGHTON

Voice Activity Detection Using Singular Value Decomposition-based Filter

Hwa Jeon Song, Sung Min Ban, Hyung Soon Kim

School of Electrical Engineering, Pusan National University,
Geumjeong-gu Jangjeon 2-dong, Busan 609-735, Korea
{hwajeon, bansungmin, kimhs}@pusan.ac.kr

10.21437/Interspeech.2009-632

Abstract

This paper proposes a novel voice activity detector (VAD) based on singular value decomposition (SVD). The spectro-temporal characteristics of background noise region can be easily analyzed by SVD. The proposed method naturally drops hangover algorithm from VAD. Moreover, it adaptively changes the decision threshold by employing the most dominant singular value of the observation matrix in the noise region. According to simulation results, the proposed VAD shows significantly better performance than the conventional statistical model-based method and is less sensitive to the environmental changes. In addition, the proposed algorithm requires very low computational cost compared with other algorithms.

Index Terms: voice activity detection, singular value decomposition

1. Introduction

The voice activity detector (VAD) in the noisy signal plays a very important role in various fields, such as efficient speech transmission and speech recognition. Most of VAD algorithms assume that the statistics of background noise are stationary over the period of time, which is longer than that of speech [1]. This assumption makes it possible to robustly estimate the spectral characteristics of slowly time-varying noise.

The statistical model-based VAD [1] is very effective method and many algorithms have been developed based on this. However, this has many detection errors at the offset region of speech whose energy decreases. To solve this problem, the hangover scheme in [1] and the smoothed likelihood ratio (SLR) in [2] have been proposed. In most of VADs, a kind of hangover algorithm is normally added to smooth the VAD decision and it is generally based on heuristic algorithms. Later it has been studied that the introduction of the long-term speech information, which is a major factor in hangover module, gives many advantages for speech/pause discrimination in the high noise environments [3].

This paper proposes a novel VAD based on applying the singular value decomposition (SVD) [4] on the observation matrix composed of adjacent multiple frames, in which possess long-term information, in the spectral domain. Our method has two great advantages over the statistical model-based VAD. First, the constraint of embedment of heuristic hangover algorithm is removed in the proposed method due to usage of multiple observation frames. Second, it is possible to drastically reduce the computational cost by properly retaining the number of eigenvectors based on the property of SVD since a set of the basis vectors in the proposed VAD is estimated and updated in the noise region only.

As a related work, KLT-based VAD in [5] has been proposed as a module in the speech enhancement system and that has been worked with the multiple frames of the noisy speech in time domain. Though it represented good results for high SNR, it has not been worked well in low SNR. Also, it is computationally very expensive [5] since basis has to be updated in every frame.

This paper is organized as follows. In Section 2, the general VAD scheme is described. In Section 3, the proposed algorithm based on SVD is developed and our proposed method is generalized by employing the adaptive threshold technique which reflects the time-varying background noise characteristic. The performance of the proposed algorithm is evaluated in Section 4 and then a conclusion is drawn.

2. VAD summary

This section briefly states the general VAD process and the overlooked or redundant process in VAD is also described. First, let's assume that the input speech signal $y(t)$ in time domain is corrupted by additive noise $n(t)$ which is uncorrelated with clean speech $x(t)$ by following equation

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

Here, there is no consideration of the channel distortion for convenience.

The input speech signal has to be segmented into frames to use as input feature of a certain VAD. Thus, a set of input data in one frame is converted to a specific domain through a specific operation as follows:

$$\mathbf{y}_s^{(i)} = \mathbf{x}_s^{(i)} + \mathbf{n}_s^{(i)} \tag{2}$$

where $\mathbf{y}_s^{(i)}$, $\mathbf{x}_s^{(i)}$ and $\mathbf{n}_s^{(i)}$ denote the vectors in a specific domain s transformed by an operation from $y(t)$, $x(t)$ and $n(t)$ at frame index i, respectively. Discrete fourier transform (DFT) is a representative operation. Of course, it is possible that no operation is performed on the input data in time domain, namely bypass. Then the feature parameter extraction process for the robust decision in VAD is followed as

$$f(\mathbf{y}_s^{(i)}) = f(\mathbf{x}_s^{(i)} + \mathbf{n}_s^{(i)}) = f(\mathbf{x}_s^{(i)}) + f(\mathbf{n}_s^{(i)}) \tag{3}$$

where $f(\cdot)$ is a linear function and $f(\mathbf{y}_s^{(i)})$ is used as the feature parameter for VAD decision. Examples of $f(\cdot)$ include the frame energy in time domain and the likelihood (or likelihood ratio) of a statistical model in frequency domain.

Final decision in VAD is generally accomplished by following process. For convenience, let's assume that the statistics of $f(\mathbf{n}_s^{(i)})$ is stationary and every frame is independent of each other. First, we should estimate the reference point from the initial signal segments (e.g., $f(\mathbf{y}_s^{(1)})$) where the noise signal is presented and the speech signal is absented (e.g., $\mathbf{y}_s^{(1)}$=$\mathbf{n}_s^{(1)}$).

Let's this be $f(\mathbf{n}_s^{init})$. Second, from a series of $f(\mathbf{y}_s^{(i)})$ along frame index, the speech/nonspeech decision on every frame is made by following decision rule:

$$d(f(\mathbf{y}_s^{(i)}), f(\mathbf{n}_s^{init})) \lessgtr TH \tag{4}$$

where $d(a, b)$ is a distance (or similarity) between a and b and frame i is decided as the speech when this measure is larger (smaller when the similarity is used) than the threshold (TH). The Euclidian distance between the feature parameter and reference point and the likelihood ratio based on the statistical model of the feature parameters are representatives of $d(\cdot)$. However the speech/nonspeech decision on every frame is very sensitive. To solve this problem, a hangover algorithm is usually attached to VAD and then more robust decision is obtained by using the results of adjacent frames. Its main idea is based on the strong correlation among adjacent speech frames. That is, the assumption of independence among frames actually operates as a weak point for the robust decision in VAD.

Most of VADs under a noise environment perform with both temporal and/or spectral information of input signal. This spectro-temporal characteristic of the background noise can be observed well in the spectrogram. Based on this, the correlation among the spectral bins over a certain period of time can be analyzed and this may be utilized as one effective feature parameter for VAD decision. Thus the additional hangover algorithm becomes naturally unnecessary because the correlation factor across multiple frames presents the very smooth trajectory. But the employment of multiple frames causes the computation increase. This paper proposes an efficient VAD algorithm based on the multiple frames with low computational complexity. The detail explanation is following section.

3. SVD-based VAD

3.1. Initial configuration by SVD

VAD necessarily requires an initial processing to setup a threshold value for the final decision which is generally estimated from the background noise regions only. To identify this noise characteristic along both frequency and time axes, the spectrogram is very useful. In this case, the operation in (3) for VAD corresponds to DFT. Of course, the mel scaled filter bank output may be taken as an alternative converted data to reduce the dimension of DFT data. For convenience, this converted data is called input feature vector of VAD. First, let $\mathbf{Y} = [\mathbf{y}^{(1)} \cdots \mathbf{y}^{(i)} \cdots \mathbf{y}^{(T)}]$ be the total input feature vector sequence composed of the T frames, where $\mathbf{y}^{(i)}$ denotes M-dimensional vector defined as $\mathbf{y}^{(i)} = \mathbf{x}^{(i)} + \mathbf{n}^{(i)}$ in frame i from (2). Second, let's assume that the initial K frames are composed of only the background noise and it is defined as

$$\mathbf{N} = \left[\mathbf{n}^{(1)}\mathbf{n}^{(2)} \cdots \mathbf{n}^{(K)}\right] \tag{5}$$

where $\mathbf{N}$ is $M \times K$ matrix which represents the specro-temporal characteristic of background noise in a specific noise region. That is, the column space basis reflects the correlation during some duration between noise spectral bins along the frequency axis, while the row space basis reflects the correlation between the temporal changes along the time axis in each bin. Thus it is very important to analyze these two correlations simultaneously. Among many analysis tools, SVD can be regarded as a powerful tool for this operation. From this, the observation matrix $\mathbf{N}$ is decomposed by SVD as follows:

$$\mathbf{N} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{6}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $M \times M$ and $K \times K$ matrices, respectively. The columns of both $\mathbf{U}$ and $\mathbf{V}$ matrices are orthogonal bases which span the row space and column space of the $\mathbf{N}$ matrix, respectively. Actually, $\mathbf{U}$ is a set of the eigenvectors of $\mathbf{N}\mathbf{N}^T$ and $\mathbf{V}$ is a set of the eigenvectors of $\mathbf{N}^T\mathbf{N}$. On the other hand, $\mathbf{S}$ is $M \times K$ diagonal matrix whose diagonal entries are known as the singular values of $\mathbf{N}$. Also it is the alternative of eigenvalues corresponding to $\mathbf{U}$ and $\mathbf{V}$. These singular values can be thought of as the weights of each basis vectors. In general they are sorted as ascending order of their values. Actually, the singular values correspond to the power density of $\mathbf{N}$ in this case. Therefore, a set of $\mathbf{U}$ and $\mathbf{V}$ can be regarded as the spectro-temporal basis of background noise underlying both time and frequency simultaneously.

3.2. SVD-based Filter

Since both of $\mathbf{U}$ and $\mathbf{V}$ are invertible, SVD result in equation (6) can be rewritten as

$$\mathbf{S} = \mathbf{U}^T\mathbf{N}\mathbf{V} \tag{7}$$

Now, let's briefly describe the motivation of SVD-based VAD based on (7). If another segment in the input vector sequence has statistics and energy level similar to those of $\mathbf{N}$ and substitutes for $\mathbf{N}$ in (7), the $\mathbf{S}$ hardly changes. But if the property of an input segment is very different from that of $\mathbf{N}$, the result is so different from the $\mathbf{S}$. Moreover there is no guarantee that it is the diagonal matrix.

Based on the above explanation, let the relation between factors in (7) be generalized to establish a new SVD-based VAD. Based on $\mathbf{y}^{(i)} = \mathbf{x}^{(i)} + \mathbf{n}^{(i)}$, the segment i in input sequence is defined as follows:

$$\mathbf{Y}^{(i)} = \mathbf{X}^{(i)} + \mathbf{N}^{(i)}. \tag{8}$$

where $\mathbf{Y}^{(i)} = [\mathbf{y}^{(i)} \cdots \mathbf{y}^{(i+K)}]$ denotes the segment i of the input observation sequence, which is composed of K adjacent frames. Of course, $\mathbf{y}^{(i)}$ denotes the location of reference frame in $\mathbf{Y}^{(i)}$ and its point is movable backward and forward. Moreover, it is assumed that $\mathbf{N}^{(i)}$ is stationary. That is, it's statistics is the same as $\mathbf{N}$ in (5). Under this assumption, $\mathbf{Y}^{(i)}$ is converted by $\mathbf{U}$ and $\mathbf{V}$ matrices in (7) as follows:

$$\begin{aligned}
\mathbf{\Sigma}^{(i)} &= \mathbf{U}^T\mathbf{Y}^{(i)}\mathbf{V} && (9)\\
&= \mathbf{U}^T\mathbf{N}^{(i)}\mathbf{V} + \mathbf{U}^T\mathbf{X}^{(i)}\mathbf{V} && (10)\\
&= \mathbf{S} + \mathbf{U}^T\mathbf{X}^{(i)}\mathbf{V} && (11)\\
&\cong \mathbf{S} + diag\left[\mathbf{U}^T\mathbf{X}^{(i)}\mathbf{V}\right] && (12)
\end{aligned}$$

where $diag(\cdot)$ is to take only diagonal elements of any matrix so it is assumed that $\mathbf{\Sigma}^{(i)}$ is diagonal.

Therefore, as aforementioned, the values of diagonal terms in $\mathbf{\Sigma}^{(i)}$ is determined depending on whether input segment is composed of only background noise or not. If $\mathbf{X}^{(i)}$ is absented in $\mathbf{Y}^{(i)}$, $\mathbf{\Sigma}^{(i)}$ becomes $\mathbf{S}$ under the above assumption. If $\mathbf{X}^{(i)}$ is presented in $\mathbf{Y}^{(i)}$, the values in $\mathbf{\Sigma}^{(i)}$ become larger than those of $\mathbf{S}$. However it is computationally expensive to perform all the matrix operations in $\mathbf{U}^T\mathbf{Y}^{(i)}\mathbf{V}$. Actually, the number of multiplications and additions is approximately $2M(M+K)K$. Fortunately, both $\mathbf{U}$ and $\mathbf{V}$ are composed of eigenvectors, the dimension reduction can be performed by retaining appropriate number of the eigenvectors, and thereby the computation cost can be drastically decreased. Now, how to determine the number of eigenvectors is described. For this, the ratio of the

individual singular value to the sum of the total singular values is first considered. From (7), the contribution to coverage of population of the initial noise segment is calculated as follows:

$$SVR_d = \frac{s_d}{\sum_{l=1}^{L} s_l} \tag{13}$$

where s_d is the dth diagonal term of $\mathbf{S}$; $L = min(M, K)$, namely the minimum value of M and K; SVR_d denotes how much the dth singular value contributes in the total sum of the singular values. In the majority of cases, the first few values can explain almost all population of input segment. Fig. 1 shows a set of SVR_1 in several background noise types where SVR_1 is computed only on the nonspeech region of all input frames sequence $\mathbf{Y}$. From the figure, the coverage of SVR_1 is over 95% of the total singular values. Thus, if only one basis vector is used in (9), the computation cost becomes drastically reduced.

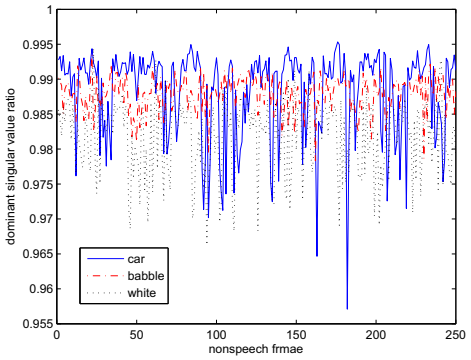

Figure 1: The ratio of maximum singular value to the sum of the total singular values

Based only on the first singular value s_1 and the first eigenvectors of $\mathbf{U}$ and $\mathbf{V}$, (11) can be simplified as follows:

$$\sigma_1^{(i)} = \mathbf{u}_1^T \mathbf{Y}^{(i)} \mathbf{v}_1 \tag{14}$$

$$= s_1 + \mathbf{u}_1^T \mathbf{X}^{(i)} \mathbf{v}_1 \tag{15}$$

where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the first column vector of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\sigma_1^{(i)}$ is the first diagonal term of $\mathbf{\Sigma}^{(i)}$. To calculate the computation cost in this case, (14) can be expressed as follows:

$$\mathbf{u}_1^T \mathbf{y}^{(i)} = u_{11} y_1^{(i)} + u_{12} y_2^{(i)} + \cdots + u_{1M} y_M^{(i)} \tag{16}$$

$$\mathbf{u}_1^T \mathbf{Y}^{(i)} \mathbf{v}_1 = [\mathbf{u}_1^T \mathbf{y}^{(i)}] v_{11} + \cdots + [\mathbf{u}_1^T \mathbf{y}^{(i+K)}] v_{1K} \tag{17}$$

where u_{1m}, $y_m^{(i)}$ and v_{1k} denote mth element of $\mathbf{u}_1$, mth element of $\mathbf{y}^{(i)}$ and kth element of $\mathbf{v}_1$, respectively. Note that the ith segment of input sequence, $\mathbf{Y}^{(i)}$, can be recursively computed. To make $\mathbf{Y}^{(i)}$ from $\mathbf{Y}^{(i-1)}$, the first column vector $\mathbf{y}^{(i-1)}$ in $\mathbf{Y}^{(i-1)}$ discarded; the rest of columns in $\mathbf{Y}^{(i-1)}$ is shifted left in turns; the new input vector with $i + K$ index $\mathbf{y}^{(i+K)}$ is inserted at Kth column location. In (16), u_{1m} can be regarded as FIR filter coefficients with respect to the observation vectors. So $\mathbf{u}_1^T \mathbf{y}^{(i)}$ is the filtered result of a input vector in observation matrix by $\mathbf{u}_1$. Thus a new K-dimensional feature vector is obtained as a result of filtering $\mathbf{Y}^{(i)}$ with $\mathbf{u}_1$. These K resulting values are again converted through the filter $\mathbf{v}_1$ composed of v_{1k}. Then the final feature parameter for the our proposed VAD is obtained. Therefore, in (17), if K-1 dimensional vector whose element is $\mathbf{u}_1^T \mathbf{y}^{(i)}$ is saved in a buffer, additional computation is required only on $\mathbf{y}^{(i+K)}$. As a result, the computation cost of one input segment is only about $2(M + K)$ multiplications and additions.

3.3. Decision operation

From (13), since the first eigenvector reflects the most statistic of background noise, the threshold can be naturally associated with the first singular value. Therefore, based on (15), an ingenious decision rule is defined as follows:

$$\sigma_1^{(i)} \begin{cases} \geq \eta, & \text{when } \mathbf{Y}^{(i)} \text{ is speech} \\ < \eta, & \text{otherwise.} \end{cases} \tag{18}$$

where $\eta = \beta s_1$ is the decision threshold and β ($\geq$1) is empirically tuned for the best tradeoff between speech and nonspeech classification errors.

3.4. Adaptive SVD-based VAD

Since the noise varies with time, the empirically tuned threshold η in (18) is not adequate in some regions. That is, when nonspeech is continued during some duration and its spectro-temporal characteristic is different from that in previous nonspeech region, the $\mathbf{U}$, $\mathbf{V}$ and the threshold value should be adaptively updated. This adaptive VAD is easily established by pseudo-code except for detailed sub-operations as follows:

Initialize $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_K] \leftarrow \mathbf{Y}^{(1)}$; $\mathbf{Y} = \mathbf{U}^{(1)} \mathbf{S}^{(1)} \mathbf{V}^{(1)^T}$
$N_c = 0$, $\eta \leftarrow \beta s_1^{(1)}$, $\mathbf{u}_1 \leftarrow \mathbf{u}_1^{(1)}$, $\mathbf{v}_1 \leftarrow \mathbf{v}_1^{(1)}$
for i = 2 to T **do**
  $\mathbf{Y} \leftarrow \mathbf{Y}^{(i)}$
  $\sigma_1^{(i)} \leftarrow \mathbf{u}_1^T \mathbf{Y} \mathbf{v}_1$
  if $\sigma_1^{(i)} \leq \eta$ **then**
    present frame is speech, $N_c = 0$
  else
    N_c++
    if $N_c == D$ **then**
      $\mathbf{Y} = \mathbf{U}^{(i)} \mathbf{S}^{(i)} \mathbf{V}^{(i)^T}$
      $\eta \leftarrow \beta s_1^{(i)}$, $\mathbf{u}_1 \leftarrow \mathbf{u}_1^{(i)}$, $\mathbf{v}_1 \leftarrow \mathbf{v}_1^{(i)}$, $N_c = 0$
    end if
  end if
end for

Here, $\mathbf{Y}$, $\mathbf{u}_1$ and $\mathbf{v}_1$ denote the temporary buffers; N_c denotes a counter variable for the duration of noise region; D denotes the maximum duration for updating the parameters. An example of the effect by the adaptive scheme is shown in Fig. 2, where the clean speech is corrupted by the babble noise at 10 dB SNR. Fig. 2(a) represents the original noisy speech and Fig. 2(b) and (c) represent the trajectory of the features under the constant and adaptive threshold, respectively. And the dash line indicates the true speech region. From the figure, it can be seen that the adaptive SVD-based VAD detects well the speech frame which is missed by the constant threshold.

4. Experimental results

The proposed VAD is compared with the statistical model-based VAD (ST-VAD) with the hangover [1] and G.729 VAD based on the speech detection probability P_d and false-alarm probability P_f. To obtain P_d and P_f, the reference decisions are made from a clean speech material with duration of 115 seconds by

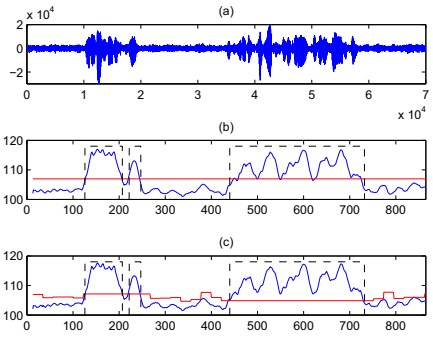

Figure 2: The effect of adaptive threshold and basis (a) noisy speech signal (b) constant threshold (c) adaptive threshold

manual labeling at every 10ms. Here, the VAD algorithms were applied to the noisy speech samples corrupted by adding white noise and vehicular noise from NOISEX-92 database to clean speech at various SNR. For the input vector $\mathbf{y}^{(i)}$ in our VAD, the mel-scaled filter bank outputs (M=23) are extracted every 10ms over frames with 20ms size. Moreover, though K=21 in $\mathbf{Y}^{(i)}$ is setup for best results in this paper, it was observed that the proposed method is less sensitive on K through a series of experiments.

The receiver operating characteristic (ROC) curves, which show the trade-off characteristics between P_d and P_f, are shown in Fig. 3 and Fig. 4 and it is seen that our proposed VAD performs better than ST-VAD and G.729 VAD.

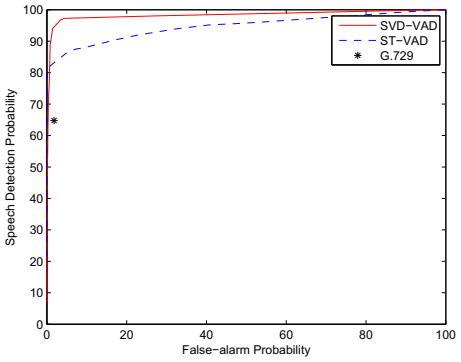

Figure 3: ROC curves of SVD and statistical model based decision rules for white noise at 5dB SNR

In Fig. 5, the proposed VAD shows more smoothed result than the ST-VAD. This is because the proposed VAD is based on multiple frames. Therefore, the proposed VAD can work without employing any hangover algorithm. Moreover, as aforementioned, Fig. 5(b) shows the weak point of ST-VAD in the the offset region of speech whose energy decreases, even though a hangover algorithm is attached, while our proposed VAD is more robust in that region.

5. Conclusions

This paper proposes a new effective SVD-based VAD with very low computational cost. In the proposed VAD, the projection of multiple frames into the eigenspace for background noise yields the smooth trajectory of feature parameter. Moreover the use

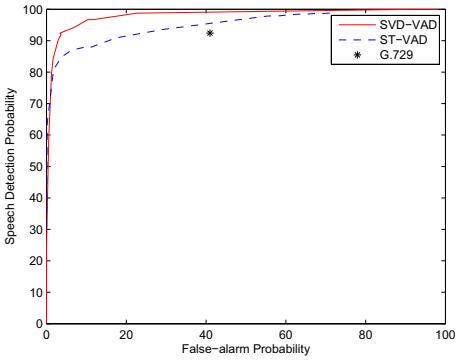

Figure 4: ROC curves of SVD and statistical model based decision rules for vehicular noise at 5dB SNR

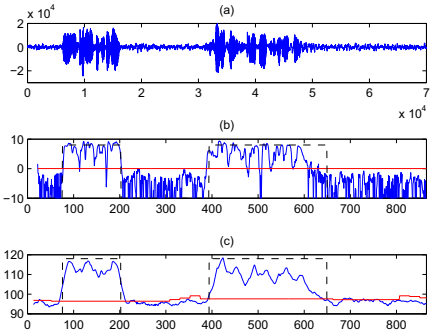

Figure 5: Comparison of the proposed VAD and ST-VAD (a) noisy speech signal (b) result of ST-VAD (c) result of proposed VAD

of the adaptive threshold based on the singular value estimated from background noise region also increases the speech detection probability for a given false-alarm rate. Future work will look at the statistical method to determine an adaptive threshold for more robust decision in various environmental conditions.

6. Acknowledgements

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea.

7. References

[1] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation”, in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 1998, pp. 365-368.

[2] Y. D. Cho and A. Kondoz, “Analysis and improvement of a statistical model-based voice activity detector”, IEEE Signal Process. Lett., vol. 8, no. 10, pp. 276-278, Oct., 2001.

[3] J. Ramirez, J. C. Segura, M. C. Benitez, A. de la Torre, and A. Rubio, “Efficient voice activity detection algorithms using long-term speech information”, Speech Comm., vol. 42, no. 3-4, pp. 271-287, 2004

[4] G. Strang, Linear Algebra and Its Applications, 3rd ed. New York: Harcourt Brace Jovanonich, 1988.

[5] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement”, IEEE Trans. Speech Audio Processing, vol. 9, pp. 87-95, Feb., 2001