International Journal of Grid Distribution Computing
Vol. 8, No.6, (2015), pp.1-8
http://dx.doi.org/10.14257/ijgdc.2015.8.6.01

A Feature Parameter Modification Algorithm for Voice Activity Detection Based on Support Vector Machine

Qian Liu[1,2], Jinxiang Wang[1] and Mingjiang Wang[1]

[1]*Microelectronics Center, Harbin Institute of Technology, Harbin, P. R. China*
[2]*Department of Electronic Science and Technology, Harbin University of Science and Technology, Harbin, P. R. China*
jxwang@hit.edu.cn, liuqian0428@126.com

Abstract

A feature parameter modification algorithm is proposed to increase the accuracy of voice activity detection that based on support vector machine and energy acceleration parameters. The three energy acceleration parameters should have the equal importance for voice activity detection, in that all the three parameters can suitably characterize the classification features of speech and non-speech frames. For the three energy acceleration parameters, its minimum values vary a little; but its maximum values are greatly different. When radial basis function is chosen as the kernel function for voice activity detection, the coordinate values in Hilbert space is dominative determined by the energy values over a sub-region spectrum, the other two parameters only have a few contributions to it. The proposed algorithm extends the three energy acceleration parameters into the nearby or same order of magnitude. It make the three energy acceleration parameters at the same importance level in the calculation of coordinate values in Hilbert space, and can finally increase the accuracy of voice activity detection. The experimental results show that this algorithm can increase the accuracy of voice activity detection in the absence of noise and noise conditions.

Keywords: *feature parameter modification, support vector machine, energy acceleration parameter, radial basis function kernel*

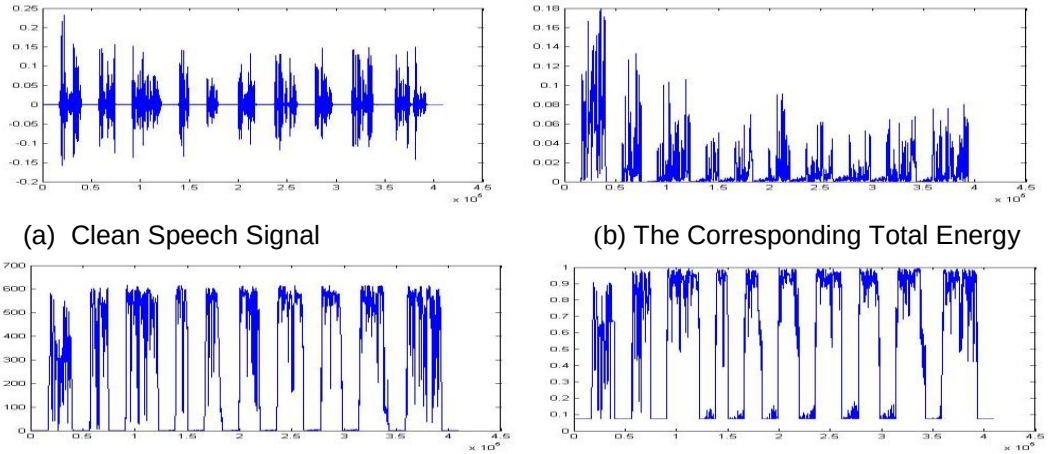
1. Introduction

Support vector machine is a pattern recognition method based on statistical learning theory. It has good ability to deal with small sample learning problems [1]. In recent years, some researchers introduce support vector machine to voice activity detection [2-6]. They use the following three kinds of feature parameters for voice activity detection: the feature parameters from time domain analysis, the feature parameters from frequency domain analysis, and the feature parameters from statistical model. Energy acceleration parameters [7] are the feature parameter used for voice activity detection in ETSI AFE standard. It includes three parameters: 1) the total energy of whole spectrum of each frame, 2) the sub-region energy over a sub-region, which contains the fundamental pitch of the spectrum of each frame, 3) the variance energy within the lower half of the spectrum of each frame. Figure 1 gives a clean speech signal and its corresponding three energy acceleration parameters. The three energy acceleration parameters can describe the voice/silent features of speech signal very well, but the range of the three energy acceleration parameters are totally different. Their minimum values are all near zero, but their maximum value is around 0.18, 600 and 1respectively.

It is found that the changes of the range of energy acceleration parameters can affect the results of voice activity detection, when impleme nting a voice activity detection system based on energy acceleration parameters and support vector machine by using radial basis function as its kernel function. Parts 2 analyzes why the influence happens by

ISSN: 2005-4262 IJGDC

analysis the theory of support vector machine, and find out how to modify the energy acceleration parameters to make the voice activity detection achieve its optimal classification results. Part 3 gives the detailed modification algorithm of energy acceleration parameters. Part 4 proves the modification algorithm by the experimental results.



(a) Clean Speech Signal (b) The Corresponding Total Energy

(c) The Corresponding Sub-region Energy (d) The Corresponding Variance Energy

Figure 1. Clean Speech Signal and Its Corresponding Three Energy Acceleration Parameters

2. Influence Analysis of Energy Parameter in SVM Method

Support vector machine has two categories: linear support vector machine and nonlinear support vector machine [8]. A linear support vector machine is shown in Figure 2 (a), its goal is to find the full line who represents the optimal separating hyperplane in order to separet ate the two different classes of samples, and makes the margin between the two dashed lines is maximum. The samples on the two dashed lines named support vectors. For nonlinear support vector machine, as shown in Figure 2 (b), it is need to map the feature vector into Hilbert space by using kernel function. The samples in Hilbert space is then converted to a linear separable case.

(a) Linear Support Vector Machine (b) Nonlinear Support Vector Machine

Figure 2. Examples of Support Vector Machine

Assume x are the energy acceleration parameters, which $x \in R^n$. There is a transform $\phi : x \to \phi(x)$, which makes $\phi(x) \in H$. $\phi(x)$ is the coordinate of x in Hilbert space. The two separating hyperplanes in H space are [8]:

$$(w \cdot x) + b = \pm 1 \tag{1}$$

The margin between the two hyperplanes is $2/\|w\|$, the problem of optimal separating hyperplane turns to the problem of convex quadratic programming[9]:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\phi(x_i)\cdot\phi(x_j)) - \sum_{j=1}^{n}\alpha_j \quad s.t. \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \le \alpha_i \le C \tag{2}$$

In Formula (2), n is the number of training samples, α is the optimal Lagrange coefficient who meets Formula (2), C is the punish factor. The w and b of separating hyperplane can be computed by the optimal Lagrange coefficient [10].

$$w^* = \sum_{i=1}^{n} y_i\alpha_i\phi(x_i) \tag{3}$$

$$b^* = -\frac{\min\limits_{y_i=-1}\{w\cdot x_i\} + \max\limits_{y_i=1}\{w\cdot x_i\}}{2} \tag{4}$$

The Formula (2) shows that the optimal separating hyperplane is determined by the feature parameters of the training set. The value of feature parameter in training set determines the position of optimal separating hyperplane. It thus influences the classification results of the test set.

The operator $(\cdot)$ in formula (2) means the inner product in H space. Define the kernel function as $K(x_i, x_j)$, so $K(x_i, x_j) = (\phi(x_i)\cdot\phi(x_j))$. Thus, the transform ϕ is accomplished by inner product.

By using energy acceleration parameters and support vector machine for voice activity detection, radial basis kernel function maps the energy acceleration parameters to H space. The radial basis kernel function is as follow:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{5}$$

It can expand as:

$$K(x_i, x_j) = \exp\left(-\frac{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + (x_{i3}-x_{j3})^2}{2\sigma^2}\right) \tag{6}$$

x_i and x_j in Formula (5) are the energy acceleration parameters which is computed by speech signal. x_{i1} and x_{j1} in Formula (6) are the total energy of whole spectrum; x_{i2} and x_{j2} are the sub-region energy; x_{i3} and x_{j3} are the variance energy.

It can be seen from the above formulas that the coordinate in H space is determined by the coordinate of energy acceleration parameters in the real space. If there is no modification on energy acceleration parameters, when using energy acceleration parameters and radial basis kernel function for voice activity detection, the position of optimal separating hyperspaces are almost determined by sub-region energy, the other two parameters only have a few influences on it. The three energy acceleration parameters are all used to describe the voice/silence feature of a speech segment, so the three energy acceleration parameters should at the same importance level in voice activity detection. The only way to make the three energy acceleration parameters equally contributed to voice activity detection is that magnifying the three energy acceleration parameters with different coefficients, and let the range of the three energy acceleration parameters at the nearby or same order of magnitude.

During voice activity detection based on energy acceleration parameters and support vector machine, there are some premises for the modification of the energy acceleration parameters: the three energy acceleration parameters, which are independent of each other, are positive real numbers; the values of other parameters won't change when a certain parameter is enlarged; anyone of the three energy acceleration parameters can always describe the voice/silence feature of the speech segment, even after it is enlarged by a coefficient.

Assume the corresponded modification coefficients of energy acceleration parameters are $cof\ 1$, $cof\ 2$, $cof\ 3$, the kernel function formula turns to:

$$K'(x_i, x_j) = \exp\left(-\frac{cof\ 1^2(x_{i1} - x_{j1})^2 + cof\ 2^2(x_{i2} - x_{j2})^2 + cof\ 3^2(x_{i3} - x_{j3})^2}{2\sigma^2}\right)$$

$$s.t. \quad cof\ 1, cof\ 2, cof\ 3 \geq 1 \tag{7}$$

Formula (7) shows that the result of kernel function is reduced after the modification of energy acceleration parameters. Name $K'(x_i, x_j) = \lambda_{ij} K(x_i, x_j)$, then $0 < \lambda_{ij} < 1$.

The problem of convex quadratic programming in Formula (2) turns to:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \lambda_{ij} (\phi(x_i) \cdot \phi(x_j)) - \sum_{j=1}^{n} \alpha_j$$

$$s.t. \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, 0 < \lambda_{ij} < 1 \tag{8}$$

The w^* of optimal separating hyperplane turns to:

$$w' = \sum_{i=1}^{n} y_i \alpha_i \lambda_i \phi(x_i) \tag{9}$$

Because the y_i in Formula (9) is the same as in Formula (3), the range of α_i are still the same, it is obtained that $w' < w^*$. It means margin between the two hyperplanes is increased after energy acceleration parameters are modified, the classification accuracy rate of test set will increase also.

3. Modification Algorithm of Energy Acceleration Parameters

The three energy acceleration parameters are of equal importance in voice activity detection, so the range of the three energy acceleration parameters should at the same or nearby orders of magnitude to make the voice activity detection achieves its optimal performance. By observing energy acceleration parameters from a set of speech database, it is found that its minimum values are at the same orders of magnitude, but its maximum values are at different orders of magnitude. Thus, the modification algorithm uses the maximum values and average values of energy acceleration parameters as its basis. The modification algorithm is shown in Figure 3.

As shown in Figure 3, the modification algorithm firstly needs to compute the maximum values of three energy acceleration parameters max1, max2, max3. Then, find out the maximum values of max1, max2, max3, which is named as mmax, at the same time set the corresponding modification coefficient to 1. Name the order of mmax as lmmax, and computer magnification between the other two maximum values of energy acceleration parameters and lmmax, which is named as tmax1 and tmax2. The corresponding modification coefficients of the other two energy acceleration parameters are computed by tmax1 and tmax2. The next step is check the average values of the modified energy acceleration parameters by the obtained coefficients. If the magnification of any two average values (the large one divide the small one) is larger than 100, then the modification coefficient corresponded to the parameters with smaller average values

should time 10. Otherwise, the three modification coefficients keep the same value as before.

The order of mmax is computed as below, its flowchart is shown in Figure 4:

1. check whether mmax is larger than 1: if it is greater than 1, go to step 2; otherwise, go to step 5;
2. let t=1;
3. compute the magnification of mmax and t;
4. Check whether the magnification is larger than 10: if it is, t=t*10, go to step 3; otherwise go to step 8;
5. let t=0.1;
6. compute the magnification of mmax and t;
7. Check whether the magnification is smaller than 10: if it is, t=t/10, go to step 6; otherwise go to step 8;
8. let t=lmmax;

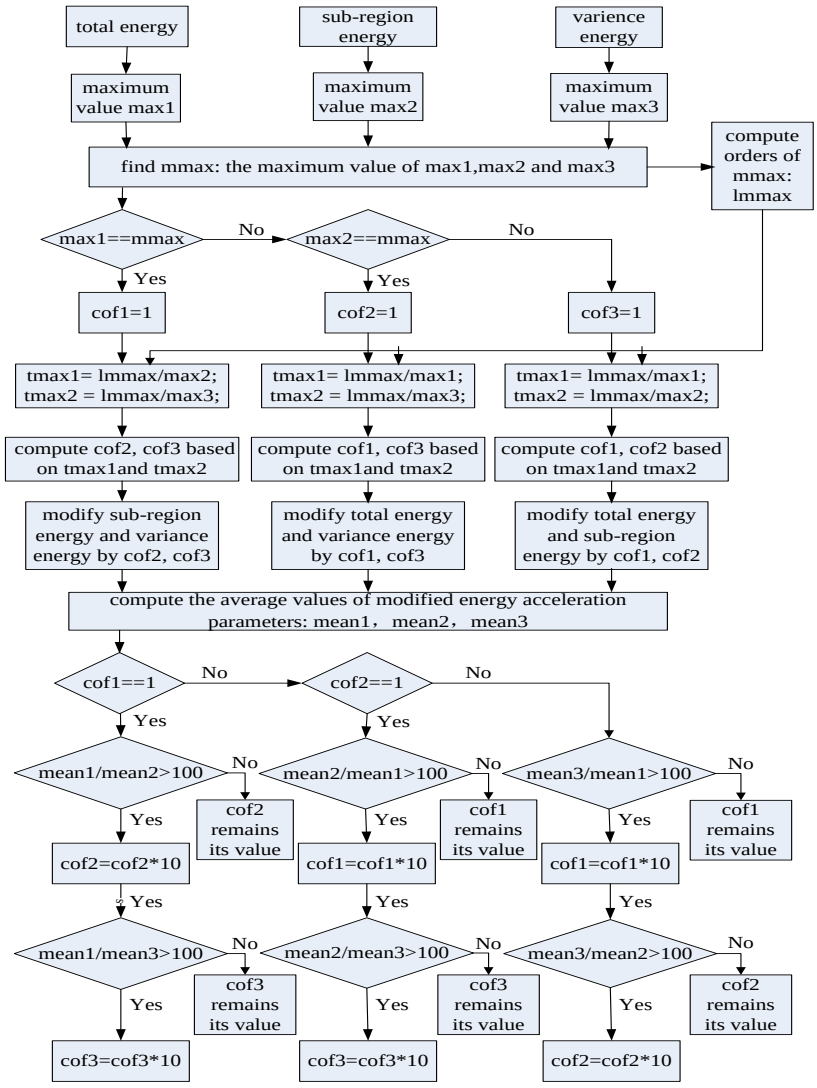


Figure 3. Feature Modification Algorithm

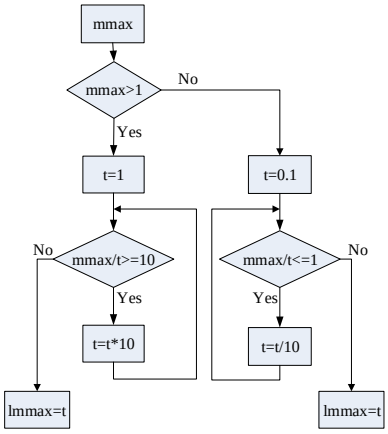


Figure 4. Computing the Orders of MMAX

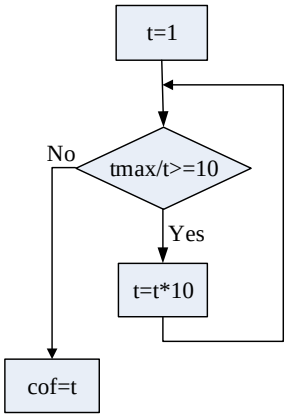


Figure 5. Computing Modification Coefficient

The modification coefficients, which are determined by tmax1 and tmax2, are computed as below. Its flow chart is shown in Figure 5.

1. let t=1;
2. compute the magnification of tmax and t;
3. Check whether the magnification is larger than 10: if it is, t=t*10, go to step 2; otherwise, t=lmmax;

4. Experimental Results

In order to verify the performance of this algorithm, two segment of speech are selected from TIMIT standard speech library [11]. Its lengths are 450.79 s and 239.18 s, the proportion of speech frame are 43.54% and 43.60%. The energy acceleration parameters extracted from speeches of 450.79s are used as training set; the energy acceleration parameters extracted from speeches of 239.18s are used as testing set. Table 3 shows the minimum values, maximum values and average values of energy acceleration parameters.

Table 3. Energy Acceleration Parameters With and Without Modification

Input speech	minimum values of energy acceleration parameters		maximum values of energy acceleration parameters		average values of energy acceleration parameters	
	without modification	with modification	without modification	with modification	without modification	with modification
train set	[0, 0.3845, 0.0736]	[0, 0.3845, 0.7359]	[0.1632 492.8918 0.9215]	[163.1730 492.8918 92.1475]	[0.0075 57.8443 0.2861]	[7.5355 57.8443 28.6103]
Test set	[0, 0.3845, 0.0736]	[0, 0.3845, 0.7359]	[0.1636, 469.5346, 0.9297]	[163.6360, 469.5346, 92.9656]	[0.0318, 56.0050, 0.2900]	[31.8309, 56.0050, 28.9978]

The result of voice activity detection can be evaluated by ROC curve [12]. ROC curve [13] is a commonly used method to measure the quality of two class classification problems. Its X-axis is the false alarm rate; its Y-axis is speech detection rate. During voice activity detection, classification result of each energy acceleration parameters can only belongs to one of the two categories: speech frame or non-speech frame. The false alarm rate is the ratio of false speech frame number to all non-speech frame numbers. The speech detection rate is the ratio of true speech frame numbers to the total speech frame numbers. In ROC curve analysis, a better ROC curve always with a smaller false alarm rate and a higher speech detection rate. ROC curve, which is widely used in researches,

can greatly represent the performance of voice activity detection. At the same time, the area under the ROC curve (AUC) is also used to measure the performance of voice activity detection.

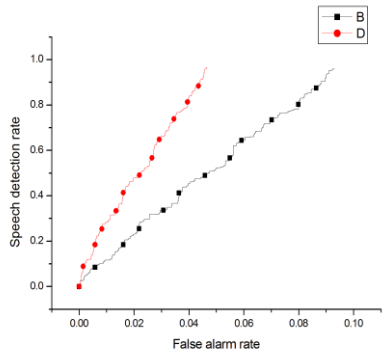


Figure 6. ROC Curves with No Noise

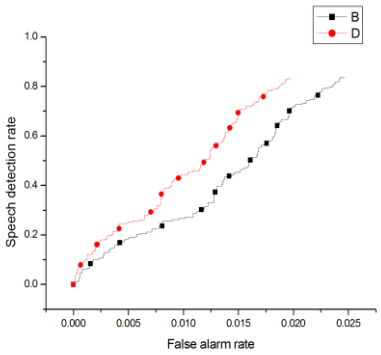


Figure 7. ROC Curves under 5db Noise

Figure 6, shows ROC curves of voice activity detection results with no noise. The full line D represents the ROC curve by the energy acceleration parameters without modification. The full line B represents the ROC curve by the energy acceleration parameters after modification. The maximum speech detection rate increases from 95.92% to 96.94%; the maximum false alarm rate fall to 4.68% from 9.30%. The speech detection rate of test set increases, and at the same time false alarm rate of test set falls. AUC of line D is larger than line B. It means the performance of voice activity detection is improved after energy acceleration parameters are modified.

In order to test the performance voice activity detection under noise environments, the train and test set are separately added three level of white gaussian noise [14]: 5db, 10db and 15db. Figure 7 shows ROC curves of voice activity detection results with 5db white gaussian noise. The full line D represents the ROC curve by the energy acceleration parameters without modification. The full line B represents the ROC curve by the energy acceleration parameters after modification. The maximum speech detection rate increases from 83.16% to 83.70%; the maximum false alarm rate fall to 1.98% from 2.47%. AUC of line D is still larger than line B; it means the performance of voice activity detection is improved after energy acceleration parameters extracted from the speeches with 5db noise are modified. The maximum values of speech detection rate are decreased in that the classification becomes more difficult under the noise environment.

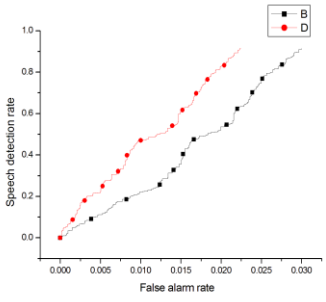


Figure 8. ROC Curves Under 10db Noise

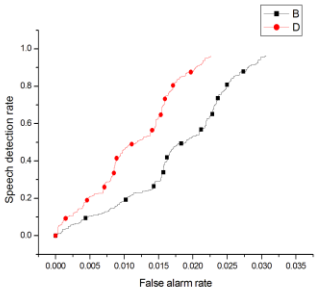


Figure 9. ROC curves Under 15db Noise

Figure 8, and Figure 9, show ROC curves of voice activity detection results with 10 db and 15 db White gaussian noise. The full line D and B represent the ROC curve by the energy acceleration parameters without and with modification. By analyzing figure 7 to figure 9, it is concluded that the performance of voice activity detection improves after energy acceleration parameters are modified. The maximum values of speech detection rate and false alarm rate vary a little at different signal noise ratio level. Its AUC also changes a little at different signal noise ratio level. It means this modification algorithm has good noise immunity and good robustness.

5. Conclusions

This paper proposed a modification algorithm of energy acceleration parameters. The premises for this modification algorithm are that: the three energy acceleration parameters are independent of each other; when a certain parameter is enlarged, it can still describe the voice/silence feature of speech segments. The property of this modification algorithm is that to make the three energy acceleration parameters at the same importance level in the calculation of coordinate values in Hilbert space. The performance of voice activity detection improves after energy acceleration parameters are modified by the algorithm. The experimental results show that this algorithm can improve the performance of voice activity detection both with clean speech and speech with white gaussian noise.

References

[1] P. Gang, S. William and Y. Wang, "Tone recognition of continuous Cantonese speech based on support vector machines", Speech Communication 45, **(2005)**, pp. 49-62.
[2] A. Q.-H. Jo, J.-H. Chang, J. W. Shin and N. S. Kim, "Statistical model-based voice activity detection using support vector machine", IET Signal Processing, vol. 3, no. 3, **(2009)**, pp. 205-210.
[3] B. Tomi Kinnunen, E. Chernenko, M. Tuononen, P. Fränti and H. Li, "Voice Activity Detection Using MFCC Features and Support Vector Machine", International Conference on SPECOM, vol. 2, **(2007)**, pp. 556-561.
[4] S.-H. Chen, S.-H. Chen and B. Rong Chang, "A Support Vector Machine Based Voice Activity Detection Algorithm for AMR-WB Speech Codec System", Second International Conference on Innovative Computing, Information and Control, ICICIC, **(2007)**.
[5] J. Wu and X.-L. Zhang, "Efficient Multiple Kernel Support Vector Machine Based Voice Activity Detection", IEEE Signal Processing Letters, vol. 18, no. 8, **(2011)** August, pp. 466-469.
[6] S-H Chen, R. Capobianco Guido, T-K Truong and Y. Chang, "Improved voice activity detection algorithm using wavelet and support vector machine", Computer Speech and Language, Elsevier, vol. 24, **(2010)**, pp. 531-543.
[7] Speech processing, transmission and quality aspects(STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; compression algorithms, ETSI ES 202, **(2007)**.
[8] G. Zheng-yan, Z. Yu-shuang and W. Mu-kun, "Speaker Recognition Using KernelK-mean Clustering and SVM", Journal HARBIN UNIV.SCI. & TECH., vol. 13, **(2008)**, pp. 40-46.
[9] V. N. Vapnik, "The Nature of Statistical Learning Theory[M]", New York:Springer-Verlag, **(1995)**.
[10] J. Chen and L. Jiao, "Classification Mechanism of Support Vector Machines", Proceedings of ICSP, **(2000)**, pp. 1556-1559.
[11] D. C. Enqing, L. Guizhong, Z. Yatong and Z. Xiaodi, "Applying support vector machines to voice activity detection", Proceeding of International Conference on Signal Process, vol. 2, **(2002)**, pp. 1124-1127.
[12] J. Saeedi, S. Mohammad Ahadi and K. Faez, "Robust voice activity detection directed by noise classification", SIViP, Signal, Image and Video Processing,Springer, **(2013),** pp. 1-12.
[13] T. Fawcett, "ROC Graphs: notes and practical considerations for researchers", HP laboratories, **(2004)**.
[14] D. Enqing, L. Guizhong, Z. Yatong and Z. Xiaodi, "Applying support vector machines to voice activity detection", Proceedings on International Conference on Signal Process, vol. 2, **(2002)**, pp. 1124-1127.

