A SIMPLE BUT EFFICIENT REAL-TIME VOICE ACTIVITY DETECTION ALGORITHM

M. H. Moattar and M. M. Homayounpour

Laboratory for Intelligent Sound and Speech Processing (LISSP), Computer Engineering and Information Technology Dept.,
Amirkabir University of Technology, Tehran, Iran
phone: +(0098) 2164542722, email: {moattar, homayoun}@aut.ac.ir

ABSTRACT

Voice Activity Detection (VAD) is a very important front end processing in all Speech and Audio processing applications. The performance of most if not all speech/audio processing methods is crucially dependent on the performance of Voice Activity Detection. An ideal voice activity detector needs to be independent from application area and noise condition and have the least parameter tuning in real applications. In this paper a nearly ideal VAD algorithm is proposed which is both easy-to-implement and noise robust, comparing to some previous methods. The proposed method uses short-term features such as Spectral Flatness (SF) and Short-term Energy. This helps the method to be appropriate for online processing tasks. The proposed method was evaluated on several speech corpora with additive noise and is compared with some of the most recent proposed algorithms. The experiments show satisfactory performance in various noise conditions.

1. INTRODUCTION

Voice Activity Detection (VAD) or generally speaking, detecting silence parts of a speech or audio signal, is a very critical problem in many speech/audio applications including speech coding, speech recognition, speech enhancement, and audio indexing. For Instance, the GSM 729 [1] standard defines two VAD modules for variable bit speech coding. VAD robust to noise is also a critical step for Automatic Speech Recognition (ASR). A well-designed voice activity detector will improve the performance of an ASR system in terms of accuracy and speed in noisy environments.

According to [2], the required characteristics for an ideal voice activity detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no prior knowledge of the noise. Among these, robustness against noisy environments has been the most difficult task to accomplish. In high SNR conditions, the simplest VAD algorithms can perform satisfactory, while in low SNR environments, all of the VAD algorithms degrade to a certain extent. At the same time, the VAD algorithm should be of low complexity, which is necessary for real-time systems. Therefore simplicity and robustness against noise are two essential characteristics of a practicable voice activity detector.

According to the state of art in voice activity detection, many algorithms have been proposed. The main difference between most of the proposed methods is the features used. Among all features, the short-term energy and zero-crossing rate have been widely used because of their simplicity. However, they easily degrade by environmental noise. To cope with this problem, various kinds of robust acoustic features, such as autocorrelation function based features [3, 4], spectrum based features [5], the power in the band-limited region [1, 6, 7], Mel-frequency cepstral coefficients [4], delta line spectral frequencies [6], and features based on higher order statistics [8] have been proposed for VAD. Experiments show that using multiple features leads to more robustness against different environments, while maintaining complexity and effectiveness. Some papers propose to use multiple features in combination with some modeling algorithms such as CART [9] or ANN [10], however these algorithms add up with the complexity of the VAD itself.

On the other hand, some methods employ models for noise characteristics [11] or they use enhanced speech spectra derived from Wiener filtering based on estimated noise statistics [7, 12]. Most of these methods assume the noise to be stationary during a certain period, thus they are sensitive to changes in SNR of observed signal. Some works, propose noise estimation and adaptation for improving VAD robustness [13], but these methods are computationally expensive.

There are also some standard VADs which have been widely employed as references for the performance of newly proposed algorithms. Some of these standards are the GSM 729 [1], the ETSI AMR [14] and the AFE [15]. For example, G.729 standard uses line spectrum pair frequencies, full-band energy, low-band energy and zero-crossing rate and applies a simple classification using a fixed decision boundary in the space defined by these features [1].

In This paper a VAD algorithm is proposed which is both easy-to-implement and real-time. It also possesses a satisfying degree of robustness against noise. This introduction follows by a discussion on short-term features which are used in the proposed method. In Section 3, the proposed VAD algorithm is explained in detail. Section 4 discusses the experiments and presents the results. Finally, the conclusions and future works are mentioned in Section 5.

2. SHORT_TERM FEATURES

In the proposed method we use three different features per frame. The first feature is the widely used short-term energy (E). Energy is the most common feature for speech/silence detection. However this feature loses its efficiency in noisy conditions especially in lower SNRs. Hence, we apply two other features which are calculated in frequency domain.

The second feature is Spectral Flatness Measure (SFM). Spectral Flatness is a measure of the noisiness of spectrum and is a good feature in Voiced/Unvoiced/Silence detection. This feature is calculated using the following equation:

$$SFM_{db} = 10\log_{10}(G_m / A_m) \qquad (1)$$

Where A_m and G_m are arithmetic and geometric means of speech spectrum respectively. Besides these two features, it was observed that the most dominant frequency component of the speech frame spectrum can be very useful in discriminating between speech and silence frames. In this paper this feature is represented by F. This feature is simply computed by finding the frequency corresponding to the maximum value of the spectrum magnitude, $|S(k)|$.

In the proposed method, these three features are applied in parallel to detect the voice activity.

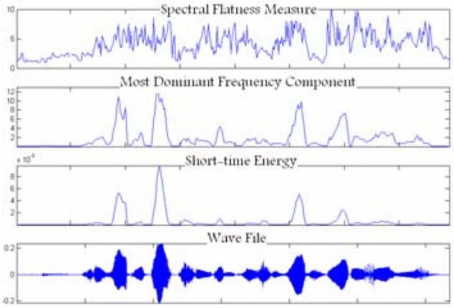

Figure 1: Feature values of clean speech signal

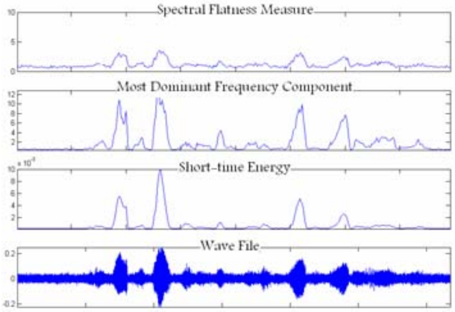

Figure 2: Feature values of speech signal corrupted with white noise

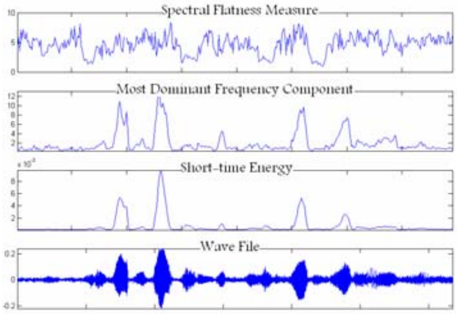

Figure 3: Feature values of speech signal corrupted with babble noise

Figures 1, 2 and 3 represent the effectiveness of these three features when speech is clean or is corrupted by white and babble noises.

3. PROPOSED VAD ALGORITHM

The proposed Algorithm starts with framing the audio signal. In our implementation no window function is applied on the frames. First N frames are used for threshold initialization. For each incoming speech frame the three features are computed. The audio frame is marked as a speech frame, if more than one of the feature values fall over the pre-computed threshold. The complete procedure of the proposed method is described below:

```
Proposed Voice Activity Detection Algorithm
1- Set $Frame\_Size = 10ms$ and compute number of frames ($Num\_Of\_Frames$)(no frame overlap is required)
2- Set one primary threshold for each feature {These thresholds are the only parameters that are set externally}
        • Primary Threshold for Energy $(Energy\_PrimThresh)$
        • Primary Threshold for F $(F\_PrimThresh)$
        • Primary Threshold for SFM $(SF\_PrimThresh)$
3- for $i$ from 1 to $Num\_Of\_Frames$
    3-1- Compute frame energy $(E(i))$.
    3-2- Apply FFT on each speech frame.
        3-2-1- Find $F(i) = \arg\max_k(S(k))$ as the most dominant frequency component.
        3-2-2- Compute the abstract value of Spectral Flatness Measure $(SFM(i))$.
    3-3- Supposing that some of the first 30 frames are silence, find the minimum value for $E$ $(Min\_E)$, $F$ $(Min\_F)$ and $SFM$ $(Min\_SF)$.
    3-4- Set Decision threshold for $E$, $F$ and $SFM$.
        • $Thresh\_E = Energy\_PrimThresh * \log(Min\_E)$
        • $Thresh\_F = F\_PrimThresh$
        • $Thresh\_SF = SF\_PrimThresh$
    3-5- Set $Counter = 0$.
        • If $((E(i) - Min\_E) >= Thresh\_E)$ then $Counter++$.
        • If $((F(i) - Min\_F) >= Thresh\_F)$ then $Counter++$.
        • If $((SFM(i) - Min\_SF) >= Thresh\_SF)$ then $Counter++$.
    3-6- If $Counter > 1$ mark the current frame as speech else mark it as silence.
    3-7- If current frame is marked as silence, update the energy minimum value:
            $Min\_E = \frac{(Silence\_Count * Min\_E) + E(i)}{Silence\_Count + 1}$
    3-8- $Thresh\_E = Energy\_PrimThresh * \log(Min\_E)$
4- Ignore silence run less than 10 successive frames.
5- Ignore speech run less than 5 successive frames.
```

In the above algorithm there are three parameters that should be set primarily. These parameters are found automatically on a finite set of clean validation speech signals to obtain the best performance. Table 1 shows these parameter values used during all of the experiments in this paper.

Table 1: Appropriate values for parameters

$Energy\_PrimThresh$	$F\_PrimThresh$ (Hz)	$SF\_PrimThresh$
40	185	5

The experiments show that these parameters are not much dependent on recording conditions and are the best choices for general applications.

4. EXPRIMENTS

For evaluating the proposed method, we used four different speech corpora. The first one is TIMIT Acoustic-Phonetic Continuous Speech Corpus [16] which is regularly used for speech recognition evaluation and contains clean speech data. We used the test data of this corpus in our evaluations. The second corpus is a Farsi microphone speech corpus known as Farsdat which contains speech data of over 300 native speakers [17]. The third corpus is a Farsi telephony speech corpus named TPersianDat collected in Laboratory for Intelligent Sound and Speech Processing (LISSP) in our department. This corpus is recorded for telephony speech and speaker recognition. This corpus is gathered in real world conditions and the speech files include background noise. Fourth dataset which is commonly used for evaluating VAD algorithms is the Aurora2 Speech Corpus. This corpus includes clean speech data as well as noisy speech.

To show the robustness of the proposed method against noisy environments, we added different noises with different SNRs to the clean speech signals in the first three corpora. No additional noise was added to Aurora2 Speech corpus.

To obtain a better viewpoint of the performance of the proposed method we compare it with two other VAD algorithms. The first one which is proposed in [13] finds an estimation of noise using Minimum Mean-Squared Error (MMSE) and is proposed for VAD in high variance vehicular noise. The other method which is mostly used as a reference method for VAD algorithms evaluation is the VAD used in the ITU G.728 Annex B standard [1].

Two common metrics known as Silence Hit Rate (HR0) and Speech Hit Rate (HR1) are used for evaluating the VAD performance. It is necessary to mention that mostly there is a trade-off between these two metrics and increasing one may lead to decreasing the other. To have a better metric for comparing two different VAD algorithm, we define a total performance metric (T) as the mean of HR0 and HR1. It is worth mentioning that in most applications increasing in HR1 is most important than increasing in HR0 so it is better to have a weighted mean between these two measures. But since we don't want to have a prejudgment for the possible applications of the method, we relinquish the weighted mean. However, the results demonstrate that the weighted mean is almost indifferent for the proposed evaluation.

The proposed method is first evaluated for three different datasets and five different noises. For these evaluations, white, babble, pink, factory and Volvo noises with 25, 15, 5 and -5 db SNRs are added to the original speech data. The proposed algorithm is also evaluated on the original speech. Table 2, shows the results achieved from these evaluations in term of HR0, HR1 and T.

Table 2: Experimental Results for the proposed algorithm on three different speech corpora

Corpus	Noise	SNR (db)	Accuracy %			Accuracy for Noise %
			HR0	HR1	T	
TIMIT	None	---	95.07	98.05	96.56	96.56
	white	25	90.41	99.77	95.09	86.27
		15	97.59	84.73	91.16	
		5	73.68	100	86.84	
		-5	44.01	100	72	
	babble	25	96.46	97.89	97.18	85.40
		15	91.81	96.81	94.31	
		5	70.17	95.61	82.89	
		-5	47.91	86.56	67.24	
	pink	25	90.64	99.77	95.20	83.22
		15	84.29	98.06	91.17	
		5	69.65	100	84.82	
		-5	23.40	100	61.70	
	factory	25	90.54	99.83	95.18	77.21
		15	84.06	95.31	89.68	
		5	75.31	76.20	75.75	
		-5	41.16	55.33	48.24	
	Volvo	25	97.83	97.84	97.83	72.29
		15	51.62	97.13	74.37	
		5	51.62	97.69	61.03	
		-5	15.14	96.72	55.93	
Farsdat	None	---	97.19	97.72	97.45	97.45
	white	25	99.13	98.47	98.90	92.74
		15	98.01	99.16	98.58	
		5	96.13	99.16	97.65	
		-5	54.13	97.58	75.86	
	babble	25	93.25	98.80	96.02	75.72
		15	54.13	99.23	76.78	
		5	35.95	99.80	67.88	
		-5	24.44	100	62.22	
	pink	25	98.43	99.04	98.74	93.96
		15	97.86	98.25	98.05	
		5	94.06	96.47	95.27	
		-5	71.43	96.16	83.79	
	factory	25	99.04	98.55	98.79	88.53
		15	90.06	98.28	94.17	
		5	80.43	95.76	88.10	
		-5	62.02	84.14	73.08	
	Volvo	25	99.13	98.70	98.92	84.07
		15	98.01	98.89	98.45	
		5	51.65	97.95	74.80	
		-5	29.30	98.97	64.13	
TPersianDat	None	---	86.83	92.93	89.88	89.88
	white	25	85.15	93.75	89.45	84.27
		15	80.17	94.2	87.19	
		5	74.21	94.96	84.58	
		-5	65.66	86.10	75.88	
	babble	25	86.36	93.49	89.93	77.15
		15	84.83	85.82	85.32	
		5	61.62	85.79	73.70	
		-5	38.47	80.83	59.65	
	pink	25	85.68	93.75	89.71	81.14
		15	80.70	94.05	87.37	
		5	69.03	95.14	82.09	
		-5	33.05	97.73	65.39	
	factory	25	85.79	93.52	89.65	80.31
		15	80.99	94.29	87.64	
		5	72.62	88.3	80.46	

		-5	51.86	75.12	63.49	
	Volvo	25	93.68	85.30	89.48	79.18
		15	81.49	86.85	84.17	
		5	57.48	91.56	74.52	
		-5	46.52	90.60	68.56	
Average			73.14	93.94	83.33	84.74

In the above Table the last column shows the average accuracy of the proposed method for a single noise in different SNRs. The average accuracy of the proposed method for the mentioned test set is 84.74%. For comparison, we evaluated the method proposed in [13] in the same conditions. From now on we simply call the method proposed in [13] as the MMSE method. The evaluation results are listed in Table 3 in the same order.

Table 3: Experimental Results for MMSE method presented in [13]

Corpus	Noise	SNR (db)	Accuracy %			Accuracy for Noise %
			HR0	HR1	T	
TIMIT	None	---	93.93	88.18	91.05	91.05
	white	25	95.84	84.18	90.01	80.05
		15	99.2	77.13	88.17	
		5	99.95	58.85	79.4	
		-5	99.97	25.27	62.62	
	babble	25	89.68	85.65	87.67	83.41
		15	84.01	86.01	85.01	
		5	81.27	83.16	82.21	
		-5	79.32	78.22	78.77	
	pink	25	93.12	83.68	88.4	78.56
		15	98.09	76.68	87.39	
		5	99.95	57.80	78.87	
		-5	99.97	19.23	59.60	
	factory	25	95.01	83.12	89.06	78.88
		15	97.20	76.38	86.79	
		5	97.80	58.77	78.28	
		-5	98.51	24.30	61.40	
	Volvo	25	94.29	88.00	91.15	90.49
		15	95.29	86.58	90.94	
		5	95.23	85.26	90.24	
		-5	95.77	83.55	89.66	
Farsdat	None	---	89.50	98.78	94.14	94.14
	white	25	94.54	95.73	95.14	87.91
		15	98.61	90.76	94.68	
		5	99.30	77.93	88.61	
		-5	99.86	46.62	73.24	
	babble	25	77.95	90.99	84.47	73.12
		15	63.1	90.28	76.69	
		5	60.10	82.23	71.17	
		-5	55.52	64.80	60.16	
	pink	25	97.10	94.29	95.70	82.7
		15	98.61	87.61	93.11	
		5	99.50	65.77	82.64	
		-5	99.86	18.86	59.38	
	factory	25	90.61	93.07	91.84	77.57
		15	92.43	84.75	88.59	
		5	94	58.66	76.33	
		-5	94.49	12.56	53.53	
	Volvo	25	98.80	94.57	96.68	95.69
		15	98.79	93.26	96.02	
		5	98.73	92.71	95.72	
		-5	97.02	91.73	94.37	
TPersianDat	None	---	93.73	90.93	92.33	92.33
	white	25	97.09	87.07	92.05	80.65
		15	98.71	79.00	88.85	
		5	92.71	64.52	78.61	
		-5	99.97	26.28	63.12	
	babble	25	89.23	88.95	89.09	78.79
		15	79.56	84.37	81.96	
		5	64.31	82.95	73.63	
		-5	61.27	79.72	70.49	
	pink	25	99.12	85.14	92.13	80.16
		15	98.32	76.23	87.27	
		5	99.55	60.16	79.85	
		-5	100	22.82	61.41	
	factory	25	99.95	76.97	88.46	76.69
		15	99.95	71.03	85.49	
		5	99.01	52.18	75.59	
		-5	99.18	15.29	57.23	
	Volvo	25	95.83	90.81	93.32	92.42
		15	95.93	89.23	92.58	
		5	96.89	88.20	92.55	
		-5	96.83	85.67	91.25	
Average			92.68	73.22	82.95	84.14

The results illustrated in Tables 2 and 3 show a slightly higher performance for the proposed method. But there are some other observations that should be highlighted. First of all these experiments show that in contrary to the proposed method, the MMSE method has a lower average HR1 and a higher Average HR0 which is not acceptable in most cases. The second point is that the main flaw of the proposed method is its low accuracy in Volvo noise which is the potency point of the MMSE method. As mentioned earlier the MMSE method is mainly proposed for VAD in vehicular noise environment which includes Volvo noise. If we exclude the Volvo noise from our evaluations the proposed method will highly outperform the method in [13]. Also the proposed method is about ten times faster than the MMSE method and is more applicable for real-time processing.

The second experiments are done on the Aurora speech corpus for evaluating the proposed method for three other noise conditions and comparing it with G. 729B VAD algorithm as a reference method. These experiments are done on subway, babble and car noises. Table 4 shows the resulted accuracy in terms of HR0, HR1 and T. The values indicating the performance of G. 729 have been derived from [18] where Aurora2 has also been used for performance evaluation.

Table 4: Comparison results for the proposed method and G. 729 standard

Method	Subway noise			Babble Noise			Car Noise		
	HR0	HR1	T	HR0	HR1	T	HR0	HR1	T
G. 729B	42.3	92.5	67.4	41.7	92.9	67.3	55.3	87.7	71.5
Proposed	63.11	85.8	74.4	68.0	90.5	79.28	72.68	87.73	80.2

The above table shows a considerable improvement in total VAD accuracy using the proposed method. But the main improvement in its accuracy is due to HR0 and HR1 is lower for the proposed method.

The third experiment is done to demonstrate the importance of each of the applied features in VAD algorithm independently. For this purpose we made a minor change in our algorithm so that a frame shall be marked as speech if only one of the features approve it. This experiment is done using TIMIT corpus in clean and noisy conditions. Only white and babble noises with four SNRs are used. The results are illustrated in Figure 4.

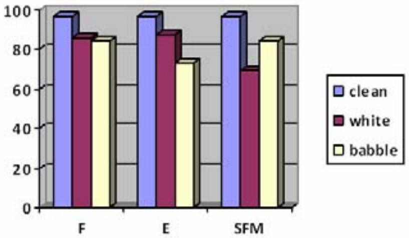

Figure 4: Effect of each selected feature on VAD Accuracy (The most dominant frequency (F), Short-term Energy (E) and Spectral Flatness (SF))

Figure 4 shows the role of each of these features in VAD performance. This illustrates that the good performance of the algorithm in babble noise is mostly due to SFM feature. Also it shows that for white noise, short-time energy and updating it as mentioned in the algorithm lead to better results. Also, the higher average results for F feature shows that the most dominant frequency component plays an important role for achieving the highest total performance.

5. CONCLUSIONS AND FUTURE WORKS

The main goal of this paper was to introduce an easy to implement Voice Activity Detection which is both robust to noise environments and computationally tractable for real-time applications. This method uses short-time features of speech frames and a decision strategy for determining speech/silence frames. The main idea is to vote on the results obtained from three discriminating features. This method was evaluated on four different corpora and different noise conditions with different SNR values. The results show promising accuracy in most conditions compared to some other previously proposed methods.

There are two minor deficiencies for the proposed method. First, this method is still vulnerable against certain noises such as car noise. This flaw can be solved by using some other features which are more robust to this condition.

The second defect of the proposed method which is also minor but can be decisive for some applications is its relatively lower average speech hit rate, specially compared to G. 729 VAD standard. This flaw may be solved with some revisions in VAD algorithm for example possibly by incorporating fuzzy terms in the decision process.

ACKNOWLEDGMENTS

The authors would like to thank Iran Telecommunication Research Center (ITRC) for supporting this work under contract No. T/500/14939.

REFERENCES

[1] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin and J. P. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine 35*, pp. 64-73, 1997.

[2] M. H. Savoji, "A robust algorithm for accurate end pointing of speech," *Speech Communication*, pp. 45–60, 1989.

[3] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan and R. Sarikaya, “Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system,” Proc. *ICASSP*, 1, pp. 53-56, 2002.

[4] T. Kristjansson, S. Deligne and P. Olsen, “Voicing features for robust speech detection,” Proc. *Interspeech*, pp. 369-372, 2005.

[5] R. E. Yantorno, K. L. Krishnamachari and J. M. Lovekin, “The spectral autocorrelation peak valley ratio (SAPVR) – A usable speech measure employed as a co-channel detection system,” Proc. *IEEE Int. Workshop Intell. Signal Process*. 2001.

[6] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” *IEEE Trans. Speech Audio Process*, 10, pp. 109-118, 2002.

[7] *ETSI standard document*, ETSI ES 202 050 V 1.1.3., 2003.

[8] K. Li, N. S. Swamy and M. O. Ahmad, “An improved voice activity detection using higher order statistics,” *IEEE Trans. Speech Audio Process.*, 13, pp. 965-974, 2005.

[9] W. H. Shin, "Speech/non-speech classification using multiple features for robust endpoint detection," *ICASSP*, 2000.

[10] G. D. Wuand and C. T. Lin, "Word boundary detection with mel scale frequency bank in noisy environment," *IEEE Trans. Speechand Audio Processing*, 2000.

[11] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari and K. Shikano, “Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs,” *Interspeech*, pp. 173-176, 2004.

[12] J. Sohn, N. S. Kim and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, pp. 1-3, 1999.

[13] B. Lee and M. Hasegawa-Johnson, "Minimum Mean Squared Error A Posteriori Estimation of High Variance Vehicular Noise," in Proc. *Biennial on DSP for In-Vehicle and Mobile Systems*, Istanbul, Turkey, June 2007.

[14] ETSI EN 301 708 recommendations, “*Voice activity detector for adaptive multi-rate (AMR) speech traffic channels,*” 1999.

[15] ETSI ES 202 050 recommendation, “*Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,*” 2002.

[16] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, 1993.

[17] M. Bijankhan, *Great Farsdat Database*, Technical report, Research center on Inteligent Signal Proccessing, 2002

[18] P.A. Estévez, N. Becerra-Yoma, N. Boric and J.A. Ramırez, "Genetic programming-based voice activity detection," *Electronics Letters*, Vol. 41, No. 20, 2005.