Robust Speech Detection for Noisy Environments

Óscar Varela, Rubén San-Segundo and Luis A. Hernández

ABSTRACT

This paper presents a robust voice activity detector (VAD) based on hidden Markov models (HMM) to improve speech recognition systems in stationary and non-stationary noise environments: inside motor vehicles (like cars or planes) or inside buildings close to high traffic places (like in a control tower for air traffic control (ATC)). In these environments, there is a high stationary noise level caused by vehicle motors and additionally, there could be people speaking at certain distance from the main speaker producing non-stationary noise. The VAD presented in this paper is characterized by a new front-end and a noise level adaptation process that increases significantly the VAD robustness for different signal to noise ratios (SNRs). The feature vector used by the VAD includes the most relevant Mel Frequency Cepstral Coefficients (MFCC), normalized log energy and delta log energy. The proposed VAD has been evaluated and compared to other well-known VADs using three databases containing different noise conditions: speech in clean environments (SNRs > 20 dB), speech recorded in stationary noise environments (inside or close to motor vehicles), and finally, speech in non stationary environments (including noise from bars, television and far-field speakers). In the three cases, the detection error obtained with the proposed VAD is the lowest for all SNRs compared to Acero's VAD (reference of this work) and other well-known VADs like AMR, AURORA or G729 annex b.

Index Terms— *robust voice activity detector (VAD), stationary and non-stationary noisy environments, voice detection inside motor vehicles or close to high traffic places, MFCC speech vs non-speech discrimination.*

I.- Introduction

The advantages of using Automatic Speech Recognition (ASR) are increasing for several types of applications, especially those where the subject wants to develop complementary actions (using ASR) when having his/her hands occupied performing the main task, as it is the case of a car driver, an air traffic controller or a pilot. Speech Recognition has important problems when the main speaker is embedded in noisy environments. These problems are related to the correct detection of the speech: there are false alarms (provoked by strong noises) and speech losses (when this speech is confused with noise). These factors degrade speech recognition rates producing an unsatisfactory experience for the user. If there are too many recognition mistakes, the user is forced to correct the system which takes too long, it is a nuisance, and the user will finally reject the system. A high error rate is not acceptable for critical tasks, such as in ATC environments, which is probably the main reason for the low use of speech interfaces in ATC. With the purpose of reducing these problems, this paper presents a robust voice activity detector (VAD) for

segmenting an audio signal into speech and non-speech frames. This segmentation is sent to the speech recognizer that will only process speech pronunciations. A good voice activity detector is important to reduce speech recognition errors caused by noise frames.

Nowadays, there is an increasing interest for developing robust voice activity detectors (VAD) for real time applications in adverse conditions. Similar to the VAD proposed in this work, Shon [1] uses a statistical model-based detector including an effective hang-over scheme which considers the previous observations by a first-order Markov process for modelling speech occurrences. This paper contributes with an analysis of the discrimination power of the different MFCCs and proposes a noise level adaptation process for increasing VAD robustness against different signal to noise ratios (SNRs).

Traditionally, log frame energy has been a very effective feature for detecting speech in any condition but it has the problem that it is necessary to adapt log energy thresholds for different SNRs. Increasing VAD robustness for different SNRs has been aimed in several works. In Ramirez et al. [2], authors face the problem of SNR independence by using the Kullback-Leibler divergence measure. In [3] authors train different noise or non-speech models for different SNRs and they propose an automatic decision module to choose the appropriate model based on SNR values estimated frame by frame. This solution has two main problems: it is cost-effectively expensive and complex to implement, and when the automatic decision is wrong, the VAD performance degrades rapidly. The proposed VAD presented in this paper uses only one model for speech and another for non-speech for all SNRs, reducing the complexity and avoiding performing any automatic decision from SNR estimation. This characteristic has been possible thanks to the noise adaptation process. On the other hand, improving Sheikhzadeh's work [4], Acero [5] proposed the idea of using normalized log energy (subtracting the average noise log energy) to avoid training different models depending on the SNR. Acero's work has been considered as the baseline for the study presented in this paper. Acero's VAD uses an HMM-based algorithm and a pulse detection mechanism using a simple post-process technique based on two thresholds instead of four, as Lamel [6] algorithm does. In this paper, authors propose a new front-end including an analysis about the discrimination power of the different MFCCs. Besides, the log energy normalization is an improved version of that included in AMR1 [6]: the noise level, necessary for normalized log energy calculation, is adapted online during noise frames (not during speech frames). An important aspect Acero did not consider in his VAD proposal was to consider normalized log energy calculation for HMMs training: Acero's VAD did perform normalization using post-training statistical information from HMMs. The same problem happens in Qi Li [7] that uses the detected endpoints to apply energy normalization sequentially. The proposed VAD has improved Acero's one based on three aspects: a better frond-end including the most discriminative MFCCs, an online level adaptation for log

energy normalization based on noise frames and the inclusion of log energy normalization into the training of the speech and non-speech HMMs.

Other endpoint detector including spectral information is Zhang [8] VAD. Zhang, considering the idea that linguistic information plays an important role in voice activity detection, presented a 5-state HMM-based VAD that uses MFCCs, short-term energy and zero-crossing rate into the feature vector, but without including normalized log energy and delta log energy information. Finally, in [9] two classification techniques, SVM and GMM, for VAD are presented using modified group delay. Two different models, speech model and non-speech model are considered by the classifiers, similar to our work but using a different feature vector.

This paper presents an improved VAD for robust voice detection in noisy environments with different SNRs. This improvement is based on three main contributions: 1) an improved front-end including a selection of the most discriminative MFCCs, 2) an improved reference level estimator for log energy normalization, 3) training the HMMs considering log energy normalization. The proposed VAD uses only two HMMs: one to represent speech frames and other to represent non-speech frames, but obtaining very good results in different conditions (SNRs).

The paper is organized as follows: the proposed VAD is described in section II. Section III shows the improvement of the new VAD compared to Acero's one for segmenting speech and non-speech frames (front-end comparison). Section IV presents global detection results when comparing our new approach to other well known VADs over three real mobile telephone databases. Finally, the main conclusions are presented in Section V.

II.- Voice Activity Detector Structure

The proposed VAD is composed of three main modules (Fig.1): The first one is the feature vector extraction, the second is the HMM-based algorithm, and finally the third one is the Speech Pulse Detector implemented as a state machine.

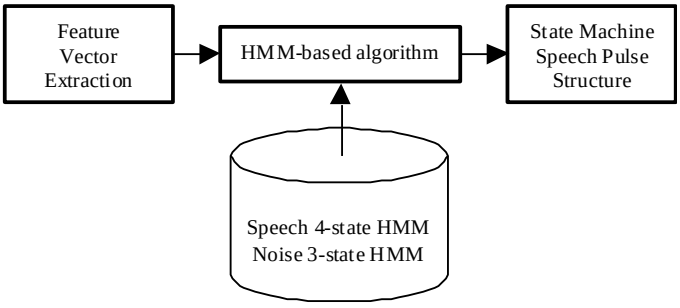

Fig. 1. Voice Activity Detector Block Diagram.

A. *Feature Vector Extraction*

The feature vector *v(n)* is composed by five features as shown in Fig.2.

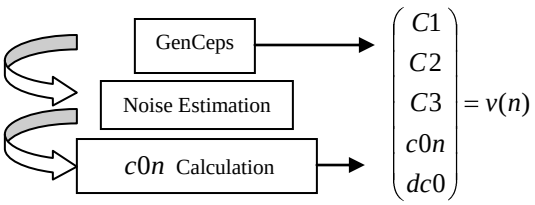

Fig. 2. Feature Extraction

The five features are the most discriminative MFCCs ($C1$, $C2$ and $C3$), obtained from a previous study which is developed and explained latter in this section (Table I), normalized log energy ($c0n$) and delta log energy ($dc0$) calculated at every frame. In this work, every frame includes audio samples during 24 ms with a 50% overlapping between consecutive frames. The GenCeps module computes MFCCs from a 12 Mel filter bank with pre-emphasis. If this specific feature extraction (MFCCs) is also used for speech recognition, front-end calculation will not increase the processing time

In order to find out the more discriminative MFCCs, that is those coefficients that produce bigger differences between the two acoustic classes (speech and non-speech), the speech and non-speech probability distribution functions for the first nine MFCCs ($C0$-$C8$) were computed and analysed. This analysis was done along the training database assuming independence between MFCCs. All the MFCCs were calculated for all the frames (speech and non-speech ones).

The discrimination power of a MFCC can be measured as the inverse of the uncertainty [10]. The uncertainty (1) is the probability of miss-classifying a frame according to only that coefficient.

$$UC_i = \int_{-\infty}^{th\_best} p_{sp}(x_i)dx_i + \int_{th\_best}^{\infty} p_{non-sp}(x_i)dx_i \quad (1)$$

where x_i represents the *i'th* MFCC. p_{sp} and p_{non-sp} denote the probability distributions of x_i MFCC for speech and non-speech frames respectively. For each coefficient independently, probability distributions in the training set were estimated for each acoustic class (speech and non-speech). The probability distributions were estimated without normalizing the histograms. The classification error (uncertainty) computed using (1) is based on an optimum threshold, th_best in (1), (x_i > *th_best* is speech otherwise non-speech) considering the probability distribution functions as continuous functions (without normalization). This *th_best* is the intersection point between the two probability distribution functions. Note that in this specific case discrete probability distributions are used and the *th_best* is the nearest discrete value to the intersection point between the two ideal continuous probability distribution functions.

Table I contains uncertainties for all MFCCs, sorted by uncertainty. The MFCCs selected to train the speech and non-speech acoustic models of the original VAD system are highlighted in bold. The uncertainty results show that the more discriminative MFCCs (lower uncertainty) are, in sequence, $C3$, $C0$, $C1$ and $C2$. As $C0$ will be used to calculate normalized log energy ($c0n$), $C3$, $C1$ and $C2$ were selected to be incorporated into the final feature vector. In the developing experiments, the use of more MFCCs (C4 for example), in addition to the three considered MFCCs, did not obtain better detection results. Because of this, in order to avoid increasing the VAD processing time, only C1, C2 and C3 were considered.

Index of the MFCC	Uncertainty
3	**0.3428**
0	**0.3606**
1	**0.3623**
2	**0.3686**
4	0.3765
5	0.3898
7	0.4137
6	0.4371
8	0.4495

Table I. Probability distributions uncertainty for each MFCC.

The next feature considered in the proposed front-end is the normalized log energy. In order to compute the normalized log energy, it is necessary to estimate the background noise log energy (*bg_n*). The noise estimator is based on an improved version of the AMR1 algorithm [6],

$$bg\_n[i+1] = (1.0-\alpha)\cdot bg\_n[i] + \alpha \cdot en[i-1] \tag{2}$$

where i denotes actual frame, *en* the energy and α takes values according to the next criterion:

$$\left.\begin{array}{l} if \quad bg\_n[i] < en[i-1], \quad \alpha = 1.0-\lambda \\ else \quad \alpha = \lambda \end{array}\right\} \tag{3}$$

In this study λ has been set to 0.85, getting in this way an 85% adaptation to energy falls due to silence or stationary background noise. Finally, normalized log energy is calculated frame by frame as the difference between the log energy at this frame ($C0$) and the background noise log energy estimated in this frame.

The last incorporated feature is delta log energy. This feature is calculated at frame i as the difference between log energy in frame i ($C0$) and log energy in previous frame (i-1).

B. *HMM-based algorithm*

This algorithm uses two acoustic models: a speech model and a noise or non-speech model. Model topology is represented in Fig.3. Both HMMs are left-to-right models with three and four emitting states for noise and speech model respectively, and one mixture per state (the exact number of states is not critical). Note that jumping states are allowed.

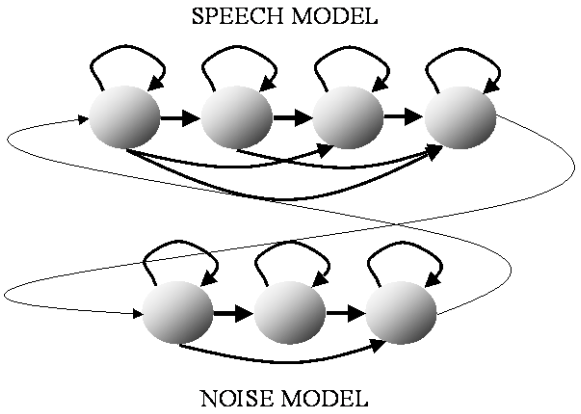

Fig. 3. HMMs structure.

The HMM-based algorithm consists of the calculation of a parameter named *score* for each frame, which is derived directly from the log likelihoods of one frame given speech/non-speech models (4).

$$score = \log(L(\varphi_{speech})) - \log(L(\varphi_{noise})) \tag{4}$$

where $L(\varphi) = prob(\varphi \mid \nu(n))$ symbolizes the likelihood of frame *n* given an acoustic model. Another important aspect is that speech and noise models are connected to each other: Fig. 3 presents a network where the noise model can be followed by the speech model and vice versa.

C. *Speech Pulse Detection*

The HMM based algorithm provides a preliminary frame classification into speech and non-speech frames. This classification is based on the speech/noise log likelihood ratio: *score*. If *score* is higher than zero, the frame is pre-classified as a speech frame; otherwise the frame is pre-classified as a noise or non-speech frame. After this decision, the speech pulse detection module adds additional information to detect speech pulses providing the final frame classification into speech or non-speech frames. This information is related to the pulse duration, silence between pronunciations and pulse extension:

- Pulse duration: If pulse duration is less than 168 ms (14 frames, considering 12 ms advance), is not considered as a speech pulse. With this condition, the VAD avoid detecting clicks, coughs or blows as speech. This value is the maximum delay of the VAD system.
- Silence between pronunciations: If the silence between consecutive speech pulses is less than a configuration parameter in ms, pulses are connected as only one. This value can be adjusted depending on the type of background noise.
- Pulse extension: the algorithm adds three frames before and after speech pulse in order to avoid losing low energy speech frames at the beginning and the end of pronunciations (fricative and occlusive sounds).

III.- Front-end comparison

In order to evaluate the improvement achieved with the new front-end proposed in this work, this section presents a comparison between the proposed VAD and Acero's VAD [4] considering only the frame segmentation proposed by the HMM-based algorithm (without considering the third module: speech pulse detector).

For this analysis, authors have considered a database consisting of 101.350 hand-labelled files from real conversations between users and real services recorded over GSM mobile phones. This database includes high speaker variability: 150 males and 148 females, aged between 21 and 43 year old and located in outdoors environments containing several kinds of noises. This database contains both stationary and non-stationary noise like hits, clicks, so that noise model considered this effect. The SNR average for the training database was around 20 dB and all audio files have a SNR higher than 18 dB. This database has been randomly divided in two sets: 90% for training the HMMs and performing the analysis of MFCCs discrimination power (see II.A), and 10% for developing the system (tuning the different thresholds).

For testing, authors have considered two new databases:

1. A stationary noise database (motor vehicle noise) composed of 2800 hand-labelled files that contains spontaneous spoken language over GSM mobile phones recorded while the main speaker is in different situations: the main speaker is in a bus stop, inside a car or a bus, or the main speaker is talking over his/her mobile phone while he/she is driving a car at different speeds. Different speakers, 11 males and 7 females, aged between 19 and 33 years old, were considered. This is a stationary noise database including mainly motor-vehicle noises: similar stationary noise appears in control tower for ATC. SNR ranges between 0 and 20 dB.
2. A non-stationary noise database: 2900 hand-labelled files containing conversational language over GSM mobile phones in airports, bars with television social gathering programs and far-field speakers. Different speakers, 11 males and 10 females, aged between 25 and 47 years old were considered. This non-stationary noise database contains files with different SNRs from 5 dB to 25dB.

Fig.4 presents detection error trade-off (DET) curves (varying the *score* threshold) for Acero's front-end and the new proposed front-end considering the stationary noise database with a 5dB SNR. As it is shown, the behaviour of the new VAD is much better than Acero's, in both cases, false alarm and miss rates.

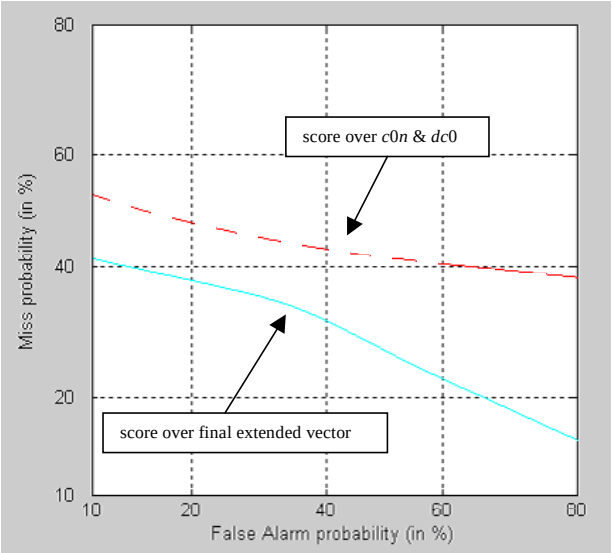

Fig. 4. DET Curve over *score* parameter. Stationary noise SNR=5dB. Dashed line for Acero's VAD and continuous line for the proposed VAD.

Extending results to other SNRs and fixing *score* threshold to zero (note that this decision means only one point in DET Curve), Table III presents false alarm and miss rates for different SNRs.

SNR (dB)	False alarm rate (Acero's VAD)	False alarm rate (New VAD)	Miss rate (Acero's VAD)	Miss rate (New VAD)
0	1.58%	1.55%	67.23%	55.67%
5	1.38%	1.27%	61.53%	43.72%
15	2.09%	1.83%	54.46%	32.88%
20	2.62%	2.34%	52.89%	31.25%

Table II. False alarm rate and miss rate for different SNRs considering the database with traffic noise.

As Table II shows, non-speech model allows obtaining a very good false alarm rate in both cases: the new robust VAD performs a little better than Acero's one. However, differences are bigger when comparing miss rates: the new robust VAD speech model performs much better than Acero's one. In conclusion, the global improvement is important with this type of noise. Table III shows the Equal Error Rate (EER) over the stationary noise database for different SNRs.

SNR (dB)	EER (Acero's VAD)	EER (considering only log energy + delta log energy)	EER (New VAD)
0	53.2%	72.8%	41.7%
5	42.9%	57.9%	33.4%
15	32.7%	38.3%	24.8%
20	25.1%	27.6%	18.3%

Table III. EER for Acero's vector and the new final extended vector for different SNRs.

Table III shows the new feature vector EER improvements, for all SNRs, the Acero's proposal. As expected EER decreases when SNR increases. Table III also includes the EER considering only log energy and delta log energy (without log energy normalization neither MFCCs).

Table IV presents false alarm and miss rates for different SNRs in non-stationary noise environments (Table IV):

SNR (dB)	False alarm rate (Acero's VAD)	False alarm rate (New VAD)	Miss rate (Acero's VAD)	Miss rate (New VAD)
5	15.88%	14.72%	82.12%	82.03%
10	14.70%	13.60%	77.63%	76.11%
15	12.65%	10.50%	72.61%	66.89%
20	10.30%	7.74%	66.63%	54.75%
25	5.41%	3.68%	55.95%	35.10%

Table IV. False alarm rate and miss rate for different SNRs considering the database with babble noise.

As it was expected, detection results are worse in the non-stationary noise database compared to the stationary noise database: the false alarm rate increases due mainly to the far-field speech included in this database. Even in this case, the proposed VAD obtains better results than Acero's VAD for all SNRs.

In order to improve speech vs noise frame decision, a new constrain over the normalized log energy was evaluated. The decision about the frame type was based on acoustic model log likelihood and the normalized log energy:

- $Score \geq 0$ and normalized log energy $\geq 0 \Rightarrow$ Speech frame.
- $Score \geq 0$ and normalized log energy $< 0 \Rightarrow$ Noise frame.
- $Otherwise \Rightarrow$ Noise frame

Considering the two constraints, the results showed a relative improvement of 28.3% for false alarm and a relative reduction of 11.6% for miss rate over the stationary noise database with a 0dB SNR. So, the second constrain, based on the normalized log energy, did not report better results. The global results depend on the false_alarm_rate/miss_rate ratio. If the ratio is close to one (equal error rate) the second condition will improve the global detection error, but if the proportion tends to zero (as in our experiments) there is no improvement: global detection error gets worse. So the second constrain depends on the VAD working point, so it was discarded.

It is important to remark that the results obtained in this section, do not include the "Speech pulse detection" module for none VAD: neither Acero's one nor the proposed in this paper. The main target has been to compare the front-end (feature extraction) module. Moreover, the "Speech pulse detection" module parameter adjustment depends on the kind of application in which the VAD is used. In the next section, global detection errors are shown with the complete scheme: including the Speech Pulse Detection module.

IV.- Global detection and evaluation results

This section presents the evaluation results considering the full proposed VAD (including the Speech Pulse Detector) and comparing the performance to others well-known VADs: AMR1 [6], AMR2 [6], AURORA(FD)

[11] and G729 annex b [12]. The “New HMM VAD” working point is set to score = 0 (the same used in section III). The working points for the reference VADs are those adjusted and considered by the standard, so no software modification has been done. Three hand-labelled databases have been considered in these experiments. The first one is a clean speech database that includes 2500 hand-labelled files containing short phrases over GSM mobile phones from 9 males and 8 females, aged between 23 and 41 years old. In this case, there is no specific noise, only the noise produced by channel: speakers are located in a quiet room. The SNR database average is around 25 dB (a clean speech database). The two next databases are the test databases described in the previous section: a stationary noise database and a non-stationary noise database.

These three databases include all possible environments in which a speaker can be located, including noise in control tower for ATC or similar applications.

Next figures show the “global detection error” (*GDE*) (equation 5): sum of normalized speech and non-speech frames detection errors (Fig.5.), so normalized false alarm rate and normalized miss detection rate.

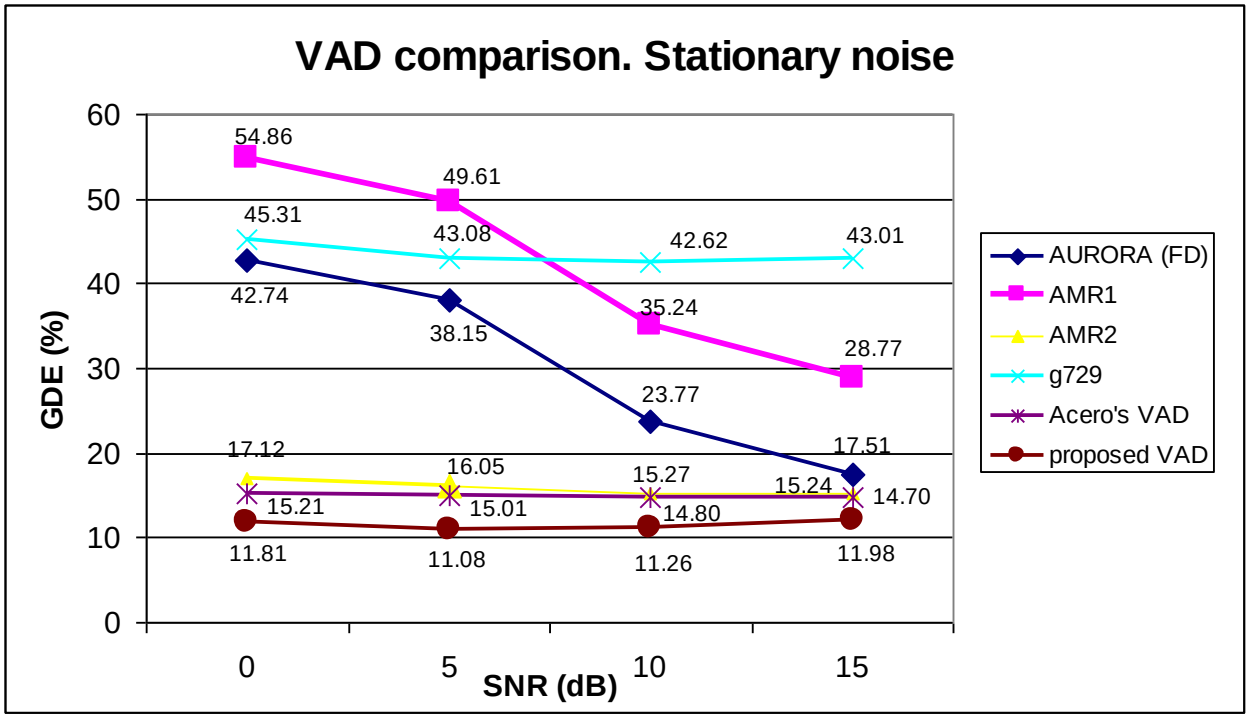

Fig.5. Global detection error for different SNRs considering the database with stationary noise.

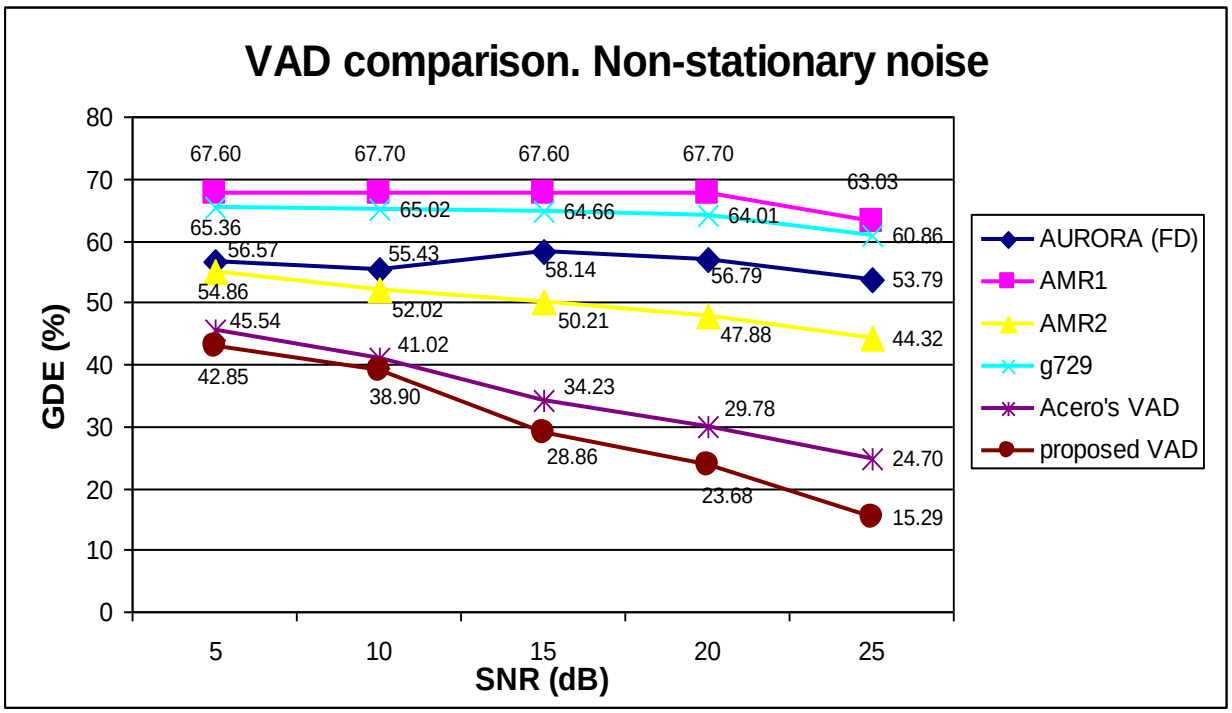

Fig.6. Global detection error for different SNRs considering the database with non-stationary noise

$$GDE(\%) = \frac{100}{2}\left(\frac{N_{Nf \to Sf}}{N_{Sf}} + \frac{N_{Sf \to Nf}}{N_{Nf}}\right) \tag{5}$$

In (5) Nf denote number of noise frames, Sf number of speech frames, $Nf \to Sf$ number of noise frames detected as speech, $Sf \to Nf$ number of speech frames considered as noise.

For clean speech, results are shown in Table V.

GDE	AURORA(FD)	AMR1	AMR2	g729b	Acero's VAD	New VAD
%	40.59	22.38	13.55	32.98	18.12	**13.34**

Table V. Global detection error (*GDE*) with clean speech.

The proposed VAD obtains the best results in the three databases, followed by AMR2 VAD. It is important to remark the flat behaviour of the proposed VAD over the stationary noise database for different SNRs and the error is very similar to that obtained for clean speech result. This behaviour demonstrates the robustness of the proposed VAD. This behaviour has been possible due to the use of a new front-end including the most discriminative MFCCS and normalized log energy computed after a voice level adaptation process. Nevertheless in non-stationary database, global detection error decreases when SNR increases.

Numerical results show that the proposed VAD is the best approach compared to other well-known VADs. For example, a relative overall detection error improvement of 26.41% and 31.02% have been obtained for SNR=0dB with stationary noise when comparing to Acero's VAD and Motorola VAD (AMR2) respectively. On the other hand, in babble noise conditions, the proposed VAD obtains a relative error improvement of 5.20%, 21.89%, 24.25%, 34.44% and 36.61% (for SNR=5dB) when comparing to Acero's VAD, AMR2, AURORA(FD), G729 annex b and AMR1 respectively.

V.- Conclusion

In conclusion, this paper presents an improved VAD for robust detection in noisy environments with different SNRs without the need of tuning. This improvement is based on three main contributions: an improved front-end including a selection of the most discriminative MFCCs, an improved reference level estimator for log energy normalization, and finally, the HMMs training considering the log energy normalization process. The proposed VAD uses only two HMMs: one to represent speech frames and other to represent non-speech frames.

The evaluation in noise conditions has been carried out using two noisy databases: considering stationary noise and non-stationary noise. In stationary noise database, noise model performs very well for all SNRs. As expected, this aspect is more difficult in presence of non-stationary noise. Final results show that the proposed VAD is the best approach compared to other well-known VADs. *GDE* is lower than 12% for all SNRs in stationary noise environment. Nevertheless VAD results in non-stationary noise are not very good as expected.

Future work will be focused on the study of incorporating new information to reject pulses coming from far-field speakers. This information will be included in the Speech Pulse Detection module.

References

[1] J. Sohn, Student Member, IEEE, N. S. Kim, Member, IEEE, and W. Sung, “A Statistical Model-Based Voice Activity Detection”, IEEE Signal Processing Letters, Vol. 6, No. 1, January 1999. pp 1-3

[2] J. Ramírez, Student Member, IEEE, J. C. Segura, Senior Member, IEEE, C. Benítez, Member, IEEE, A. Torre, and A. J. Rubio, Member, IEEE, “A New Kullback–Leibler VAD for Speech Recognition in Noise”, IEEE Signal Processing Letters, Vol. 11, No. 2, February 2004. pp. 266-269.

[3] H. Sheikhzadeh, R. L. Brennan and H. Sameti, “Real-Time implementation of HMM-Based MMSE Algorithm for speech enhancement in hearing aid applications”, 1995 IEEE. Pp. 808-811.

[4] A. Acero, C. Crespo, C. Torre and J. C. Torrecilla, “Robust HMM-Based endpoint detector”, EUROSPEECH 1993, 1551-1554.

[5] L. Lamel, L. Rabiner, A. Rosenberg and J. Wilpon, “An Endpoint Detector for Isolated Word Recognition”, IEEE Trans on Acoustics, Speech and Signal Processing. Vol.. 19, No 4.

[6] ETSI TS 126 094 V4.0.0 (2001-03).

[7] Q. Li, Senior Member, IEEE, J. Zheng, A. Tsai, and Q. Zhou, Member, IEEE, “ Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition”, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 3, March 2002. pp. 146-157.

[8] J. Zhang, W. Ward and B. Pellom, “Phone based Voice Activity Detection using online bayesian adaptation with conjugate normal distributions”, University of Colorado at Boulder. Boulder, Colorado 80309-0594, USA.

[9] R. Padmanabhan, S. H. Krishnan and H. A. Murthy, “A pattern recognition approach to VAD using modified group delay”, IEEE Proceedings NCC-2008.

[10] P. Tarapeik, J. Labuda and B. Fourest, “Measurement uncertainty distributions and uncertainty propagation by the simulation approach”, 3rd EURACHEM Workshop, September 1999, Bratislava.

[11] ETSI ES 202 050 V1.1.1 (2002-10).

[12] Recommendation G.729 Annex B (10/96): A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70.