

Robust Speech Detection for Noisy Environments

Óscar Varela

Indra S.A.

&

Rubén San-Segundo & Luis A. Hernández

Universidad Politécnica de Madrid

ABSTRACT

This presents a robust voice activity detector (VAD) based on Hidden Markov Models (HMM) 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 herein 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

[4]) and other well-known VADs like AMR, AURORA, or G729 annex b.

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 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 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, 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, Sohn [1] uses a statistical model-based detector including an effective hang-over scheme which

Author's Current Address:

O. Varela, Telephone Platforms Group, Indra S.A., C/Anabel Segura, 28108-Alcobendas, Madrid, Spain; R. San-Segundo, (Speech Technology Group), and L.A. Hernández, (Applications and Signal Processing Group), Universidad Politécnica de Madrid, Madrid, Spain.

Manuscript received August 3, 2010; revised January 2011.

Review was handled by M. Greco.

0885/8985/11/ \$26.00 © 2011 IEEE

considers the previous observations by a first-order Markov process for modelling speech occurrences. This 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, VAD performance degrades rapidly. The proposed VAD presented herein uses only one model for speech and another for non-speech for all SNRs, reducing the complexity and avoiding performing any automatic decision from a SNR estimation. This characteristic has been possible thanks to the noise adaptation process. On the other hand, improving Sheikhzadeh's work [3], Acero [4] 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 herein. 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 [5] algorithm does. Herein, 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 performed 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.

Another 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 work 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; and*
- 3) *training the HMMs considering log energy normalization.*

The proposed VAD uses only two HMMs: one to represent speech frames and the other to represent non-speech frames, but obtaining very good results in different conditions (SNRs).

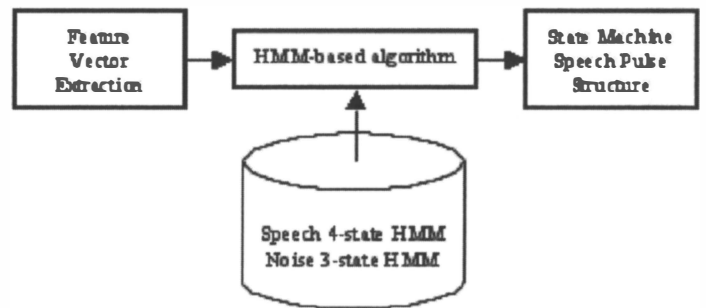


Fig. 1. Voice Activity Detector Block Diagram

This is organized as follows: the proposed VAD is described in the Voice Activity Detector Structure section; The Front-End Comparison section shows the improvement of the new VAD compared to Acero's for segmenting speech and non-speech frames (front-end comparison); the Global Detection Results section presents global detection results when comparing our new approach to other well-known VADs over three real mobile telephone databases. Finally, conclusions are presented.

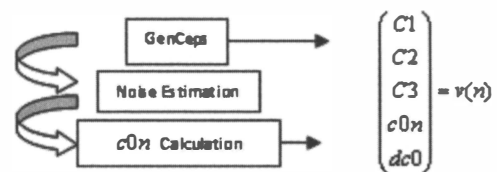


Fig. 2. Feature Extraction

VOICE ACTIVITY DETECTOR STRUCTURE

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

Feature Vector Extraction

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

The five features are the most discriminative MFCCs ($C1$, $C2$ and $C3$), obtained from a previous study which is developed and explained later in this section (Table 1), normalized log energy ($c0n$) and delta log energy ($dc0$) calculated at every frame. In this work, the frame length is 24 ms with a 50% overlap. 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.

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 of the MFCCs were calculated for all of the frames (speech and non-speech).

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 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 1 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

Table 1. Probability Distributions Uncertainty for Each MFCC

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

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.

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

$$bg_n[i+1] = (1.0 - \alpha) \cdot bg_n[i] + \alpha \cdot en[i-1] \quad (2)$$

based on an improved version of the AMR1 algorithm [6], where i denotes actual frame, en the energy and α takes values according to the next criterion:

$$\left. \begin{array}{l} \text{if } bg_n[i] < en[i-1], \alpha = 1.0 - \lambda \\ \text{else } \alpha = \lambda \end{array} \right\} \quad (3)$$

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 the current 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 ($i-1$).

HMM-Based Algorithm

This algorithm uses two acoustic models: a speech model and a noise or non-speech model. Model topology is

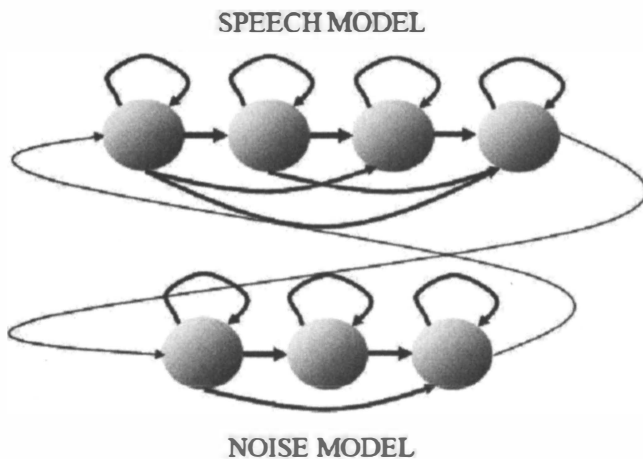


Fig. 3. HMMs Structure

represented in Figure 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.

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).

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

where $L(F) = \text{prob}(F | v(n))$ symbolizes the likelihood of frame n given an acoustic model. Another important aspect is that speech and noise models are connected to each other. Figure 3 presents a network where the noise model can be followed by the speech model, and vice versa.

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).

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, these 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, these 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, 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, 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 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.

Figure 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 5 dB SNR. As shown, the behaviour of the new VAD is much better than Acero's, in both cases, false alarm and miss rates.

Extending results to other SNRs and fixing score threshold to zero (Note that this decision means only one point in DET Curve), Table 2 presents false alarm and miss rates for different SNRs.

As Table 2 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. However, differences are bigger when comparing miss rates: the new robust VAD speech model performs much better than Acero's. In conclusion, the global improvement is important with this type of noise. Table 3 shows the Equal Error Rate (EER) over the stationary noise database for different SNRs.

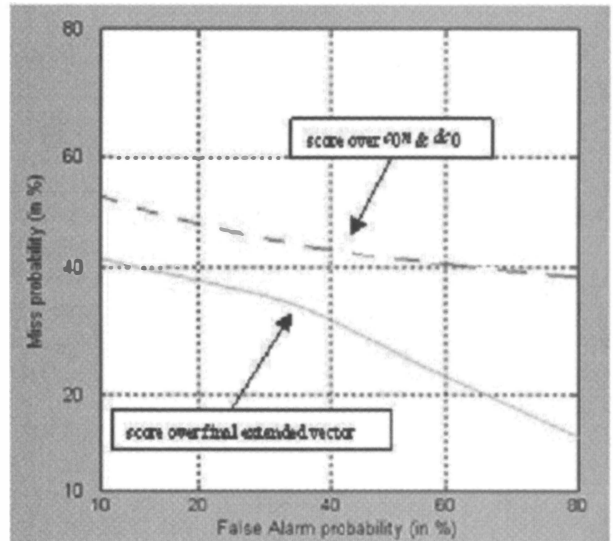


Fig. 4. DET Curve over score parameter. Stationary noise SNR = 5 dB. Dashed line for Acero's VAD and continuous line for the proposed VAD

Table 3 shows the EER improvements of the new VAD over Acero's VAD for all SNRs. As expected EER decreases when SNR increases. Table 3 also includes the EER

Table 2. False Alarm Rate and Miss Rate for Different SNRs Considering the Database with Traffic Noise

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 3. EER for Acero's Vector and the New Final Extended Vector 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 4. False Alarm Rate and Miss Rate for Different SNRs Considering the Database with Babble Noise

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%

considering only log energy and delta log energy (without log energy normalization and MFCCs).

Table 4 presents false alarm and miss rates for different SNRs in non-stationary noise environments:

As 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 the non-stationary 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 constraint 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 0 dB SNR. So, the second constraint, 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.

It is important to remark that the results obtained in this section do not include the "Speech pulse detection" module. The main goal is 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,

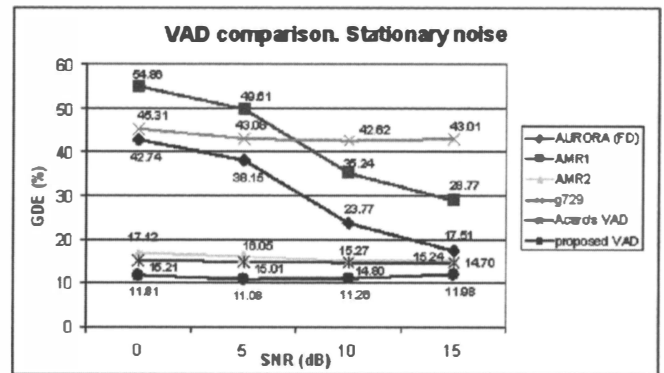


Fig. 5. Global Detection Error for Different SNRs considering the Database with Stationary Noise global detection errors are shown with the complete scheme: including the Speech Pulse Detection module.

GLOBAL DETECTION AND EVALUATION RESULTS

This presents 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 the previous section). 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 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. 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 next two databases are the test databases described in the previous

Table 5. Global detection error (GDE) with clean speech

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

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.

The next figures show the "Global Detection Error" (GDE) (Equation 5 below): sum of normalized speech and non-speech frames detection errors (Figure 5), so normalized false alarm rate and normalized miss detection rate.

In (5) N_f denotes number of noise frames, S_f number of speech frames, $S_f \rightarrow N_f$ number of noise frames detected as speech, number of speech frames considered as noise. For

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

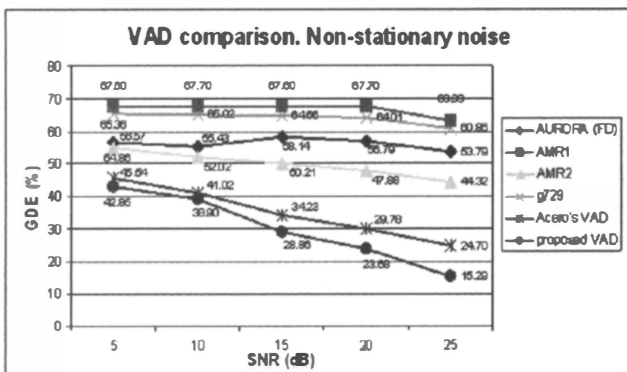


Fig. 6. Global detection error for different SNRs considering the database with non-stationary noise

clean speech, results are shown in Table 5.

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 a 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 = 0 dB 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 = 5 dB) when compared to Acero's VAD, AMR2, AURORA(FD), G729 annex b, and AMR1, respectively.

CONCLUSION

This 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 the 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 a stationary noise database, noise model performs very well for all SNRs. As expected, this aspect is more difficult in the presence of a 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 a stationary noise environment. Nevertheless VAD results in non-stationary noise are not very good, as expected.

Future work will focus on incorporating new information to reject pulses coming from far-field speakers. This 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. //