PERFORMANCE EVALUATION AND COMPARISON OF ITU-T/ETSI VOICE ACTIVITY DETECTORS

F. Beritelli, S. Casale, G. Ruggeri

Dipartimento di Ingegneria Informatica e delle Telecomunicazioni - University of Catania
Viale Andrea Doria 6 95125 Catania
email: (beritelli, gruggeri, scasale)@iit.unict.it

ABSTRACT

The paper proposes a performance evaluation and comparison of recent ITU-T and ETSI voice activity detection algorithms. The comparison was made using both objective and psychoacoustic parameters, so as to have reliable judgements that were close to subjective ones. A highly varied speech database was also set up to evaluate the extent to which VADs depend on language, the signal to noise ratio, or the power level.

1. INTRODUCTION

A Voice Activity Detector (VAD) with a comfort noise generator (CNG), achieves silence compression, which is very important in modern telecommunication systems [1]. In multimedia communications a VAD guarantees simultaneous voice and data applications; in Universal Mobile Telecommunication Systems (UMTS), it reduces the average bit rate; finally, in a cellular radio system using the Discontinuous Transmission (DTX) mode, it reduces co-channel interference and power consumption in portable equipment.

The paper presents a performance evaluation and comparison of recent ITU-T and ETSI voice activity detection algorithms. The last ITU-T VAD standard is Rec. G.729 Annex B [2], developed for fixed telephony and multimedia communications. More recently the ETSI has standardized two VAD [3] standards (options 1 and 2) for the Adaptive Multi-Rate (AMR) codec developed for third generation mobile communication systems. In addition, in view of a new standardization phase for a more efficient voice activity detector for the next ITU-T 4 kbit/s speech coding standard [4], the paper also considers the Fuzzy VAD [5] proposed to the ITU-T Study Group 1 , in that it represents a good enhanced solution for the G.729 VAD. VAD performance will be compared in various signal to noise ratio conditions, using various languages and signal power levels.

2. THE CHARACTERISTICS OF THE VADS CONSIDERED

Due to the different applications for which the VADs have been designed for, they operate on frames of different lengths: 10 ms for G.729 and FVAD, 20 ms for the two AMR VADs. Both G.729 and FVAD use the following four classification parameters: the differential power in the 0-1kHz band, the differential power over the whole band, the differential zero crossing rate, and spectral distortion. The G.729 VAD uses a multi-boundary decision region in the space of the four parameters. The FVAD uses a set of six fuzzy rules. The AMR Option 1 VAD computes the SNR in 9 bands and the decision is based on a comparison between the SNRs and a threshold, which is different for each band. The thresholds are then adapted according the the absolute noise level. The AMR Option 2 VAD divides the 20ms frames into two subframes of 10 ms and calculates the following parameters for each of them: channel power, voice metrics, and noise power. The decision is made by comparing the voice metrics with a threshold that varies according to the estimated SNR. A frame is judged to be active is at least one subframe is active.

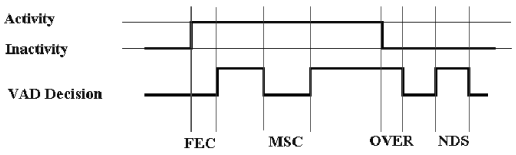

Fig. 1. Objective parameters

3. THE PARAMETERS USED FOR THE COMPARISON

3.1. Objective parameters

In order to evaluate the amount of clipping and how often noise is detected as speech, the VAD output is compared with those of an ideal VAD. The performance of a VAD is evaluated on the basis of the following four traditional parameters [5]:

- FEC (Front End Clipping): clipping introduced in passing from noise to speech activity;
- MSC (Mid Speech Clipping):clipping due to speech misclassified as noise;
- OVER: noise interpreted as speech due to the VAD flag remaining active in passing from speech activity to noise;
- NDS (Noise Detected as Speech): noise interpreted as speech within a silence period.

These four parameters are illustrated in Fig.: 1. The FEC and MSC parameters give the amount of clipping introduced, whereas OVER and NDS give the increment in the activity factor.

3.2. Limits of the objective parameters

Although the method described above provides useful objective information concerning the performance of a VAD, it only gives an initial estimate as regards the subjective effect. For example, the effects of speech signal clipping can at times be hidden by the presence of background noise, depending on the model chosen for the comfort noise synthesis, so some of the clipping measured with objective tests is in reality not audible. All in all, the parameters outlined above:

Languages	Italian, French, German, English
Levels	$-16, -26, -36$ dBovl
SNR	0, 10, 20 dB, Clean
Noises	Car, Office, Train, Restaurant, Street

Table 1. The structure of the Multi-language Database

- do not give sufficient information about the perceptive contents of frames suppressed by the VAD;
- are not good indicators of quality and intelligibility.

For this reason we used also a new parameter, called Activity Burst Corruption (ABC) [6], which is able to provide a close correlation with subjective judgements, so as to make a good prediction of the performance levels of a VAD.

3.3. The psychoacoustic parameter

The psychoacoustic parameter we recently introduced in[6] takes in account two different phenomena: hearing, the way in which human hear modifies the incoming sound, and judgement, the way in which human brain decides that a sound is better then another. In order to compute the ABC parameter a simple but effective auditory model is considered. In particular the non uniform frequency resolution and the non uniform loudness perception are the most important properties modeled. The mathematical we adopted makes it possible to pass from the power spectral density to analysis of the Subjective Loudness Density. Once taken in account the hearing the most significant effect of the clips consist in the loss of loudness. So the judgement parameter adopted, the Activity Burst Corruption, is defined in equation 1. Given the generic activity burst in which the VAD has introduced K cuts, the ABC is defined as:

$$ABC = \sum_{K=1}^{k=K} \frac{S_{Clip}^{k}}{S_{Burst}} \cdot 100 \quad (1)$$

where S_{Clip}^{K} is the total loudness suppressed by the k-th cut and $S_{Burst}$is the total loudness of the activity burst.

4. SPEECH DATABASE CONSIDERED

In order to compare the performance of the VADs being investigated, we created a speech database containing sequences uttered by both male and female speakers, linear quantized at 16 bits and sampled at 8kHz. Each sequence lasts 3 minutes, and has 40% speech activity (active frames), which is on average the typical activity percentage in a telephone conversation. To assess the behaviour of the various VADs when different languages are spoken the sequences were uttered by native speakers in Italian, English, French and German. Three different signal power levels (-16, -26, -36 dBovl), different types of noise (Car, Office, Train, Restaurant, Street) and different signal to noise ratios (SNR) (0,10,20) were also used, giving a total of 648 minutes of speech.

5. RESULTS

The various VADs were compared using the database illustrated in Section 4, and both the traditional objective parameters and the psychoacoustic parameter ABC introduced in Section 3. The comparison also considered all the different conditions for VAD use contemplated in the speech database. The results of

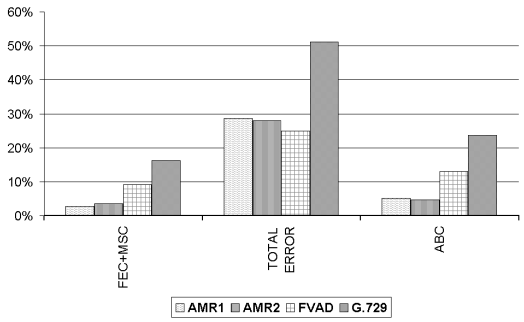

Fig. 2. Comparison using whole database

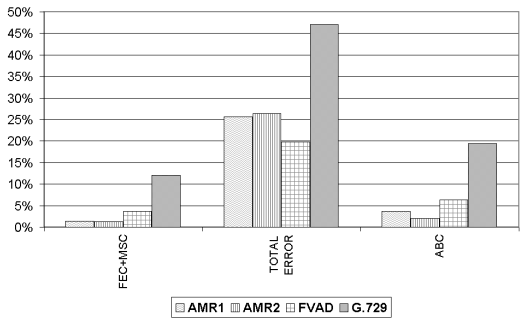

Fig. 3. Comparison with SNR 10dB and SNR 20dB

the comparison are given in Figs. 2, 3, 4, 5 6, 7, 8, 9 and 10. First of all, it is evident from the graphs that the performance of the G.729 VAD is worse in terms of both total error, given by the sum FEC+MSC+OVER+NDS, and ABC. The FVAD, on the other hand, proves to be the most efficient device in terms of total error (24,9 %). However, closer examination of the kind of errors reveals that its performance in terms of clips introduced (FEC+MSC) is worse than that of the two VADs standardized for the AMR codec. This is confirmed by the fact that the performance in terms of perceived quality, ABC, is also lower than that of the AMR VADs. Although AMR1 introduces fewer cuts than AMR2, however, the cuts introduced by the latter have less impact in terms of loss of loudness. If only non-extreme operating conditions are considered, specifically sequences with SNRs of 10 dB and 20 dB, it can be seen (fig.: 3) that the performance of AMR2 is even better in terms of both cuts and ABC. Once again the impact on perceived quality is less for AMR2. Although AMR2 is the VAD that introduces less distortion in terms of ABC, it is very sensitive to the presence of background noise. Its performance, in fact, deteriorates when the whole database is used, i.e. with an SNR of 0 dB (Fig.3), whereas the degradation in performance for both AMR1 and the FVAD when the noise increases is smoother, showing that they are more robust. If the VADs are compared using different languages, fig.: 4, 5 and fig.: 6, it can be seen that their performance, in particular as regards the number of cuts introduced, is better for languages featuring greater vocalization, i.e. Italian and French, than for English and above all German. The AMR2 VAD is, however, the device with the greatest oscillations,

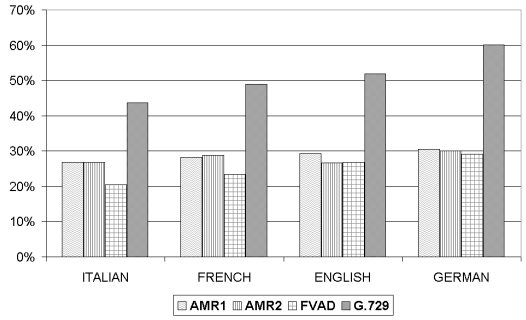

Fig. 4. Total error for different languages

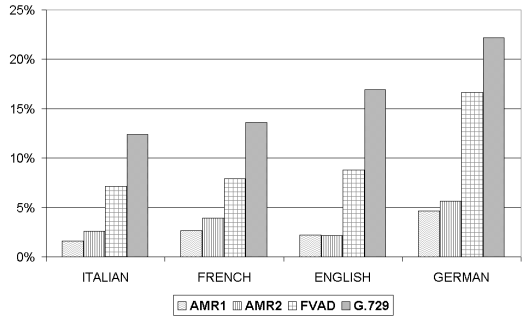

Fig. 5. Fec+MSC for different languages

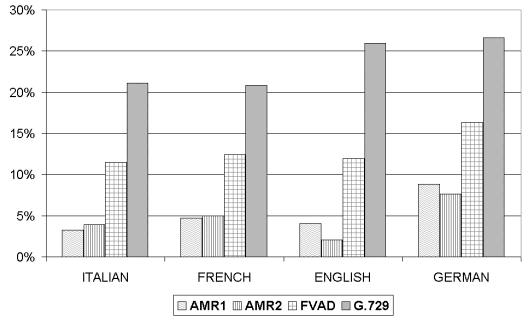

Fig. 6. ABC for different languages

in terms of ABC, when the language changes. In addition, it introduces a very slight degradation (2,07%) when the language spoken is English, where it is much more efficient than its competitors, while the degradation introduced when it operates on other languages is greater. Another series of measures referred to the behaviour of the various VADs when the level of the input signal varied (-16, -26, -36dBov).

Figs. 7, 8 and 9 show the results of the tests, grouped according to input signal level. It is interesting to note that, in terms of total error, the performance of AMR2, FVAD and G.729 is much the same when the level varies; for AMR2 there is a slight increase in cuts as the signal decreases. AMR1, on the other hand, improves its performance in terms of total error as the signal level decreases. This improvement, however, is accompanied by a deterioration in terms of cuts which, unlike the total error, increase, the increase being even more marked if measured in terms of ABC.

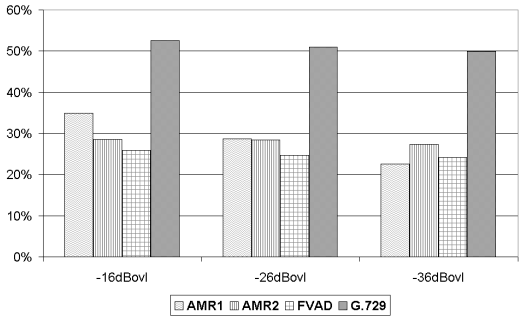

Fig. 7. Total error for different power levels

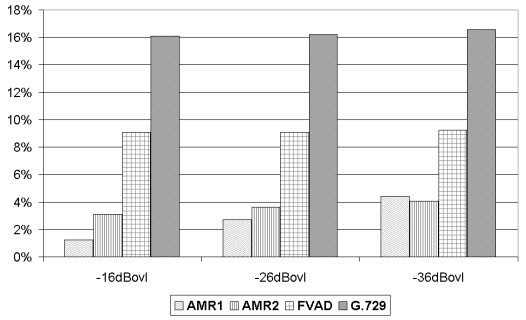

Fig. 8. Fec+MSC for different levels

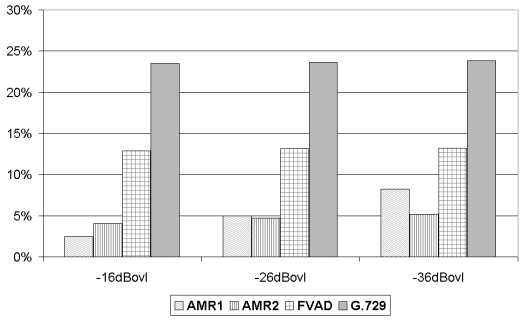

Fig. 9. ABC for different levels

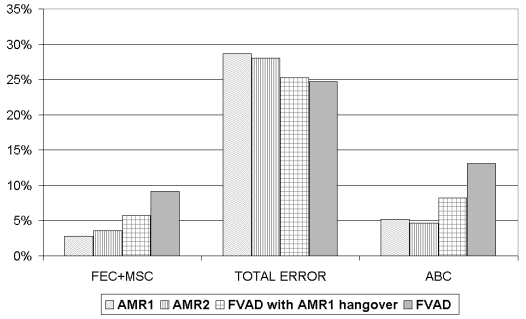

Fig. 10. Comparison when the FVAD uses the same hangover mechanism as the AMR1 VAD

	Performance	Sensitivity to noise	Sensitivity to level	Complexity
AMR1	excellent	low	high	medium
AMR2	excellent	high	low	high
FVAD	good	low	low	low
G.729	poor	low	low	low

Table 2. Summary of VAD features

This derives from the fact that the AMR1 decision is essentially based on a multi-boundary comparison between the SNRs calculated in different bands and thresholds that are only adapted according to noise level and not speech level.
Finally, to evaluate the improvement margins for FVAD we decided to assess its performance when instead of the native hangover routine,the same routine as the AMR1 is used. The results are given in Fig. 10. As can be seen, although the performance of FVAD is slightly worse when the AMR1 hangover is used, there is a greater improvement in terms of cuts and above all ABC. It should be pointed out that FVAD performance in terms of total error is still better when the AMR1 hangover is used than that of the AMR VADs. Table 2 summarizes the features of the VADs considered. The overall performance of FVAD is lower than that of the AMR VADs, but it is less sensitive to the presence of background noise and variations in speech level. The last column in the Table compares the VADs in terms of computational complexity. Qualitative analysis shows that the most complex algorithm is the AMR2 one, its complexity being three times that of AMR1. The complexity of G.729 and FVAD, on the other hand, is about one-tenth of that of AMR2.

6. CONCLUSIONS

In conclusion, we have presented a comparison between recent voice activity detection algorithms. In particular, the paper compares the performance of four VADs: the G.729 VAD, the Fuzzy VAD and the two options for the AMR VAD. The main issues pointed out by the tests, performed on a large multi-language database, can be summarized as follows. All the VADs considered perform slightly better when the language of the speakers is more vocalized, for example Italian and French. The G.729 VAD performs poorly in terms of both total error and the degradation introduced. Although the FVAD is designed on the basis of the G.729 VAD, due to a more sophisticated matching procedure, it performs better than the G. 729. If we look at total error, the FVAD obtains the best results in almost all the testing conditions considered. Furthermore, as shown in Fig. 10, it can be improved by the use of a more efficient hangover mechanism. The AMR VADs provide the best performance in terms of the degradation introduced. The performance of the two AMR VADs is very close, but there are some differences between them. The AMR2 VAD provides the best performance in many environments but it shows a high sensitivity to noise level and to the language spoken. Furthermore, it is computationally more complex than the AMR1 VAD. Likewise, the performance of the AMR1 VAD relies heavily on the level of the input signal.

7. REFERENCES

[1] R.V.Cox and P.kroon, "Low Bit-Rate Speech Coders for Multimedia Communications," *IEEE Communication Magazine*, vol. 34, no. 12, pp. 34–41, December 1996.

[2] Rec. G.729 Annex B, "A Silence Compression Scheme for G.729 Optimized for Terminal Conforming," Tech. Rep., ITU-T, Oct. 1996.

[3] GSM 06.94, "Digital cellular telecommunication system (Phase 2+); Voice Activity Detector VAD for AdaptiveMulti Rate (AMR) speech traffic channels; General description," Tech. Rep. V.7.0.0, ETSI, February 1999.

[4] Poul Barret, "Terms of Reference and Schedule for 4-Kbit/s speech coder," Contribution question 21/16, document 12-E (WP 3/16), ITU-T, 26 January-6 February 1998.

[5] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected areas in Communications (JSAC)*, vol. 16, no. 9, pp. 1818–1829, December 1998.

[6] F. Beritelli, S. Casale, and G. Ruggeri, "A Psychoacoustic Auditory Model to Evaluate the Performance of a Voice Activity Detector," *Signal Processing*, vol. 11, no. 11, pp. 11, 11 2000.