Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors

F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano

***Abstract*—The paper proposes a performance evaluation and comparison of G.729, AMR, and fuzzy voice activity detection (FVAD) algorithms. The comparison was made using objective, psychoacoustic, and subjective parameters. A highly varied speech database was also set up to evaluate the extent to which VADs depend on language, the signal-to-noise ratio (SNR), or the power level.**

***Index Terms*—Discontinuous transmission, speech quality evaluation, voice activity detector.**

I. Introduction

As is well-known, a voice activity detector (VAD) achieves silence compression, which is very important in both fixed and mobile modern telecommunication systems [1]. In communications systems based on variable bit rate speech coders (e.g., in multimedia or VoIP applications), it represents the most important block, reducing the average bit rate; in a cellular radio system using the discontinuous transmission (DTX) mode (e.g., GSM or UMTS systems), a VAD is able to increase the number of users and power consumption in portable equipment. Unfortunately, a VAD is far from efficient, especially when it is operating in adverse acoustic conditions (e.g., when the conversation takes place in noisy environments).

In order to evaluate the impact of background noise on recent voice activity detectors, this paper presents a performance evaluation and comparison of recent ITU-T and ETSI VAD algorithms. The latest 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 VADs [3] (options 1 and 2) for the adaptive multirate (AMR) codec developed for third-generation mobile communication systems. This paper also considers the fuzzy VAD (FVAD) [4] proposed by ITU-T Study Group 16, in that it represents a good enhanced solution for the G.729 VAD. VAD performance will be compared in various signal-to-noise ratio (SNR) conditions, using various languages and signal power levels. In order to overcome the limits of the traditional performance evaluation criteria of a VAD, several objective and subjective evaluation criteria were considered. More specifically, a large database with a 40% voice activity factor (VAF) was adopted for VAD comparison, using various speech and background noise conditions.

Manuscript received April 25, 2001; revised January 3, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andreas S. Spanias.

The authors are with the Department of Informatics and Telecommunication Engineering, University of Catania, 95125 Catania, Italy (e-mail: beritelli@diit.unict.it; scasale@diit.unict.it; gruggeri@diit.unict.it; sserrano@diit.unict.it).

Publisher Item Identifier S 1070-9908(02)04541-8.

II. Functioning Principles of the VADs Considered

Due to the different application scenarios, the VADs considered 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:

1) differential power in the 0–1 kHz band;
2) differential power over the whole band;
3) differential zero crossing rate;
4) spectral distortion.

The G.729 VAD uses a multiboundary decision region in the space of the four parameters [2]. In the pattern matching block, the FVAD uses a set of six fuzzy rules [4]. The AMR Option 1 VAD computes the SNR in nine bands and the decision is based on a comparison between the SNRs and a threshold, which is different for each band [3]. The thresholds are then adapted according to the absolute noise level. The AMR Option 2 VAD divides the 20-ms 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 if at least one subframe is active.

III. Parameters Used for the Comparison

A. Objective Parameters

In order to evaluate the amount of clipping and how often noise is detected as speech, the VAD output is compared with that of an ideal VAD, i.e., one obtained by manual marking of the database. The performance of a VAD is evaluated on the basis of the following four traditional parameters [4].

- Front End Clipping (FEC): Clipping introduced in passing from noise to speech activity.
- Mid Speech Clipping (MSC): 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.
- Noise Detected as Speech (NDS): Noise interpreted as speech within a silence period.

The FEC and MSC parameters give the amount of clipping introduced, whereas OVER and NDS give the increment in the activity factor.

B. 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. Hence, the parameters outlined above do not give sufficient information about the perceptive contents of frames suppressed by the VAD and are not good indicators of quality and intelligibility. For this reason, we also used a new parameter, called activity burst corruption (ABC) [5], which is able to provide a close correlation with subjective judgments, and to make a good prediction of the performance levels of a VAD.

C. Psychoacoustic Parameter

The psychoacoustic parameter we recently introduced in [5] takes two different phenomena into account: hearing—the way in which the human ear modifies the incoming sound and judgment—the way in which the human brain decides that one sound is better than another. In order to compute the ABC parameter, a simple yet effective auditory model is considered in which nonuniform frequency resolution and nonuniform loudness perception are the most important properties modeled. The mathematical model we adopted makes it possible to pass from the power spectral density to the analysis of the subjective loudness density. Once hearing is taken into account, the most significant effect of a clip is a loss of loudness. Let us define $S^{(k)}$ as the total loudness of a frame k; the total loudness of an activity burst containing N frames is

$$S_{Burst} = \sum_{k=1}^{N} S^{(k)}. \tag{1}$$

Considering a generic activity burst m comprising N frames in which the VAD has introduced K cuts, let $c(1), c(2), \ldots, c(K)$ be the frames cut and $S^{c(1)}, S^{c(2)}, \ldots, S^{c(K)}$ their loudness. We define the ABC parameter of the burst as follows:

$$ABC_m = 100 \cdot \sum_{k=1}^{K} \frac{S^{c(k)}}{S_{Burst}}. \tag{2}$$

If we have M activity bursts during a conversation, the ABC of the total sequence is defined as the average of the ABC of the single bursts

$$ABC = \frac{1}{M} \cdot \sum_{m=1}^{M} ABC_m. \tag{3}$$

D. Subjective Parameters

Although subjective tests require great effort in terms of both time and money, they represent a valid method for evaluation of the efficiency of a VAD. Listening tests were conducted using the comparison category rating (CCR) technique proposed by ITU-T [6]. In order to evaluate only the degradation introduced by the VAD, we measured the difference in mean opinion score (MOS) scores between phrases encoded with a real VAD and an ideal VAD.

For the subjective tests, 24 native listeners were used, equally divided between the two sexes. Degradation is due to synthesis via a simple comfort noise (CN) system of talkspurt frames cut by the VAD, i.e., erroneously considered to be background noise frames [7], [8]. Of course, to evaluate the degradation introduced by the AMR VAD, we used the AMR coder and its CN system in the 7.95 kbit/s coding mode, while for VAD G.729 and FVAD we used the G.729 8 kbit/s codec and its CN system.

IV. Speech Database Considered

A. Database for Objective Tests

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, linearly quantized at 16 bits and sampled at 8 kHz. Each sequence lasts 3 min, and has 40% speech activity (active frames), which is on average the typical activity percentage in a telephone conversation. To assess the behavior 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, and -36 dBov), different types of noise (car, office, train, restaurant, and street) and different SNRs (0, 10, 20 dB) were also used, giving a total of 576 min of speech.

B. Database for Subjective Tests

The database used for the subjective tests contains three different sequences in Italian lasting 10 s [9]. Each sequence consists of a sentence uttered by a male speaker and one uttered by a female speaker. A different pair of speakers was used for each sequence, giving a total of six different speakers. The sequences were sampled at a frequency of 8000 Hz and linearly quantized using 16 bits per sample. The signal level for each sequence was then digitally converted to -16 dBov, -26 dBov, and -36 dBov. Three different types of background noise (car, office, and street) were digitally added to the original sequences, with three different SNRs (20 dB, 10 dB, and 0 dB), thus obtaining a total of 18 sequences.

V. Results

The various VADs were compared using the two databases illustrated in Section IV, and all the objective and subjective evaluation parameters introduced in Section III. The results of the comparison are given in Figs. 1–6. 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. However, closer examination of the types of error reveals that its performance in terms of the clips introduced (FEC+MSC) is worse than 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 nonextreme operating conditions are considered, specifically sequences with SNRs of 10 dB and 20 dB, it can be seen that

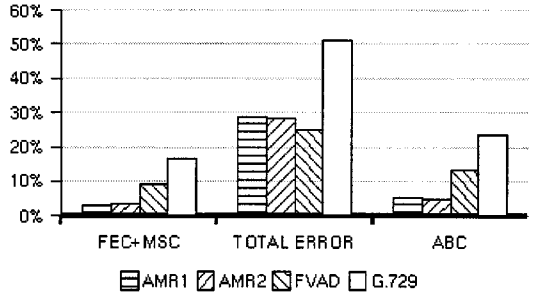

Fig. 1. Comparison using whole database.

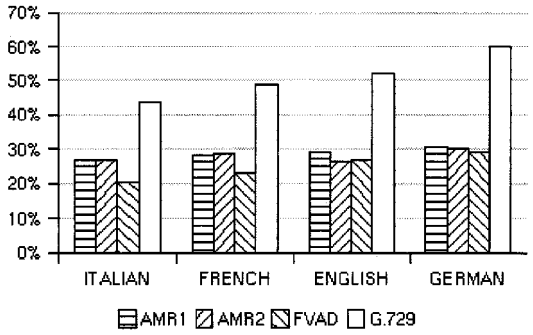

Fig. 2. Total error for different languages.

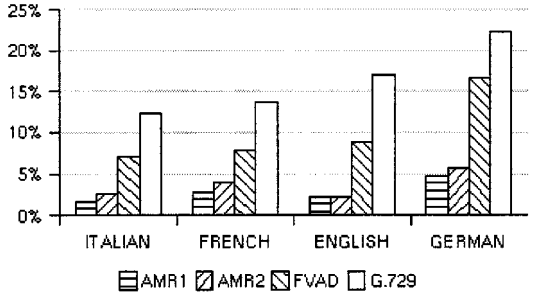

Fig. 3. FEC+MSC for different languages.

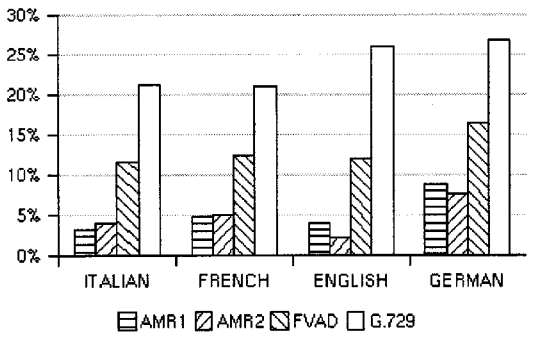

Fig. 4. ABC for different language.

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 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. 2), whereas the degradation in performance for AMR1 when the noise increases is smoother, showing that it is more robust. If the

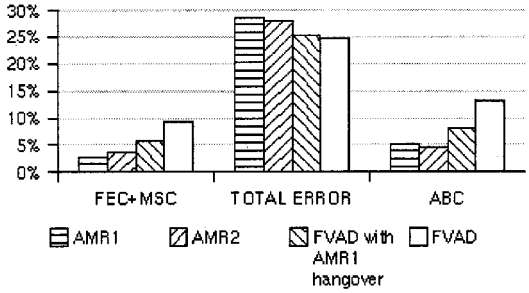

Fig. 5. Comparison when the FVAD uses the same hangover mechanism as the AMR1 VAD.

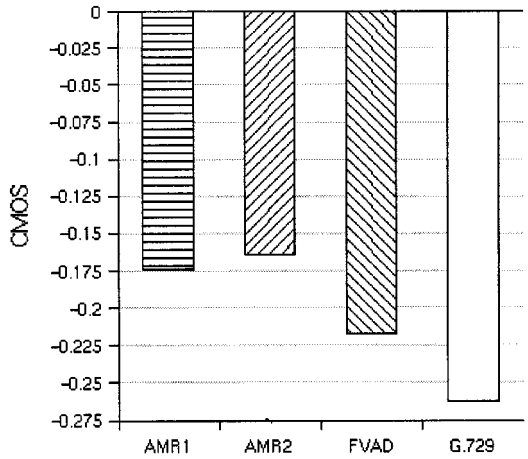

Fig. 6. Results of subjective tests.

VADs are compared using different languages, Figs. 2–4, 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 language dependency in terms of ABC. In addition, it introduces a very slight degradation 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 behavior of the various VADs when the level of the input signal varies (-16, -26, and -36 dBov).

In terms of total error, the performances of AMR2, FVAD, and G.729 are 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. This derives from the fact that the AMR1 decision is essentially based on a multiboundary comparison between the SNRs calculated in different bands and thresholds that are only adapted according to noise level and not speech level.

Fig. 6 gives the results of the subjective tests in terms of comparison mean opinion scores (CMOS) when phrases coded by an ideal VAD are taken as terms of reference. The comparison confirms the results obtained using the ABC parameter, at least

from a qualitative viewpoint. Once again, the AMR VADs perform better than FVAD and the G.729 VAD, the latter once again being the worst performer.

To evaluate the improvement margins for FVAD, we decided to assess its performance when it uses the AMR1 hangover routine instead of its original one. The results are given in Fig. 5. As can be seen, although the performance of FVAD is slightly worse when the AMR1 hangover is used, there is a greater improvement in cuts and, in particular, the ABC score. It should be pointed out that FVAD performance in terms of total error is still better than that of the AMR VADs when the AMR1 hangover is used.

VI. Conclusions

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. 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. 5, 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. Likewise, the performance of the AMR1 VAD relies heavily on the level of the input signal.

References

[1] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communications," *IEEE Commun. Mag.*, vol. 34, pp. 34–41, Dec. 1996.

[2] A. Benyassine, E. Shlomot, and H.-Y. Su, "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 application," *IEEE Commun. Mag.*, vol. 35, pp. 64–73, Sept. 1997.

[3] GSM 06.94. (1999, Feb.) Digital cellular telecommunication system (Phase 2+); voice activity detector VAD for adaptive multi rate (AMR) speech traffic channels; general description. ETSI, Tech. Rep. V.7.0.0. [Online]. Available: http://www.etsi.org.

[4] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1818–1829, Dec. 1998.

[5] F. Beritelli, S. Casale, and G. Ruggeri, "A psychoacoustic auditory model to evaluate the subjective performance of a voice activity detector," *Signal Process.*, vol. 80, no. 7, pp. 1393–1397, June 2000.

[6] "Methods for subjective determination of transmission quality," Rec. P.800 Tech. Rep., ITU-T, Aug. 1996.

[7] F. Beritelli, S. Casale, and A. Cavallaro, "New performance evaluation criteria and a robust algorithm for speech activity detection in wireless communications," in *IEEE Int. Conf. Telecommunications (ICT'98)*, vol. 1, June 1998, pp. 223–227.

[8] F. Beritelli, S. Casale, and G. Ruggeri, "New speech coding issues and algorithms for adaptive IP telephony," *Int. J. Speech Commun.*, to be published.

[9] A. Cavallaro, "Advanced techniques for voice activity detection," Ph.D. dissertation, University of Palermo, Palermo, Italy, Dec. 1999.