A FUZZY LOGIC-BASED SPEECH DETECTION ALGORITHM FOR COMMUNICATIONS IN NOISY ENVIRONMENTS

A. Cavallaro, F. Beritelli, S. Casale
Istituto di Informatica e Telecomunicazioni - University of Catania
V.le A. Doria 6, 95125 Catania - Italy

ABSTRACT

In the field of mobile communications correct Voice Activity Detection (VAD) is a crucial point for the perceived speech quality, the reduction of co-channel interference, the power consumption in portable equipment. This paper shows that a valid alternative to deal with the problem of activity decision is to use methodologies like fuzzy logic, which are suitable for problems requiring approximate rather than exact solutions, and which can be presented through descriptive or qualitative expressions. The Fuzzy Voice Activity Detector (FVAD) proposed uses the same set of parameters adopted by the VAD in Annex B to ITU-T G.729 and a set of six fuzzy rules automatically extracted through supervised learning. Objective and listening tests confirm a significative improvement respect the traditional methods above all for low signal-to-noise ratios.

1. INTRODUCTION

A Voice Activity Detector (VAD) aims to distinguish between speech and several types of acoustic background noise even with low signal-to-noise ratios (SNRs). Therefore, in a typical telephone conversation, a VAD, together with a comfort noise generator (CNG), achieves a silence compression. In the field of multimedia communications, silence compression allows the speech channel to be shared with other information, thus guaranteeing simultaneous voice and data applications [1]. In a cellular radio system that uses the Discontinuous Transmission (DTX) mode, such as the Global System for Mobile communications (GSM), a VAD reduces co-channel interference (increasing the number of radio channels) and power consumption in portable equipment. Moreover, a VAD is vital to reduce the average bit rate in future generations of digital cellular networks, such as the Universal Mobile Telecommunication Systems (UMTS), which provide for variable bit-rate (VBR) speech coding. Most of the capacity gain is due to the distinction of speech activity and inactivity.

The performance of a speech coding approach based on phonetic classification, however, strongly depends on the classifier which must be robust to every type of background noise [2]. As is well known, for example, above all with low SNRs, the performance of a VAD is critical for the overall speech quality. When some of speech frames are detected as noise, intelligibility is seriously impaired due to speech clipping in the conversation. If, on the other hand, the percentage of noise detected as speech is high, the potential advantages of silence compression are not obtained. In the presence of background noise it may be difficult to distinguish between speech and silence, so for voice activity detection in wireless environments more efficient algorithms are needed [3][4].

The activity detection algorithm proposed in this paper is based on a pattern recognition approach in which the feature extraction module uses the same set of acoustic parameters adopted by the VAD recently standardized by ITU-T in Rec. G.729 annex B [5], but basing the matching phase on fuzzy logic. Through a series of performance comparisons with the ITU-T G.729 Annex B VAD and the VAD standardized by ETSI for the Full Rate GSM codec [6], varying the type of background noise and the signal-to-noise ratios, we outline the validity of the new methodology in terms of both communication quality improvement and bit-rate reduction as compared with the traditional solution.

2. THE FUZZY VAD ALGORITHM

The functional scheme of the Fuzzy Voice Activity Detector (FVAD) is based on a traditional pattern recognition approach. The four differential parameters used for speech activity/inactivity classification are the same as those used in G.729 Annex B [5] and are: the full-band energy difference ΔE_f , the low-band energy difference ΔE_l, the zero-crossing difference ΔZC and the spectral distortion ΔS. The matching phase is performed by a set of fuzzy rules obtained automatically by means of a new hybrid learning tool [7]. As is well known, a fuzzy system allows a gradual, continuous transition rather a sharp change between two values. So, the Fuzzy VAD proposed returns a continuous output ranging from 0 (Non Activity) to 1 (Activity), which does not depend on whether the single inputs have exceeded a threshold or not, but on an overall evaluation of the values they have assumed. The FVAD translates several individual parameters into a single continuous value which, in our case, indicates the degree of membership in the Activity class and the complement of the degree of membership in the Non Activity class. The final decision is made by comparing the output of the fuzzy system, which varies in a range between 0 and 1, with a fixed threshold experimentally chosen by minimizing the sum of Front End

Clipping (FEC), Mid Speech Clipping (MSC), OVER, Noise Detected as Speech (NDS) [8] and the standard deviation of the MSC and NDS parameters. In this way we found an appropriate value for the hangover module that satisfies the MSC and NDS statistics, reducing the total error. The hangover mechanism chosen is similar to that adopted by the GSM [6].

2.1 Speech database

The speech database used to obtain the learning and testing patterns contains sequences recorded in a non-noisy environment (Clean sequences, SNR=60 dB), sampled at 8000 Hz and linear quantized at 16 bits per sample. It consists of 60 speech phrases (in English and Italian) spoken by 36 native speakers, 18 males and 18 females. The database was then subdivided into a learning database and a testing database, which naturally contains different phrases and speakers from the first one. The two databases were marked manually as active and non-active speech segments. In order to have satisfactory statistics as regards the languages and the speakers, the male and female speakers and the languages were equally distributed between the two databases. Further, to respect the statistics of a normal telephone conversation (about 40% of activity and 60% of non-activity), we introduced random pause segments, extracting from an exponential population the length of talkspurt and silence periods.

In order to evaluate the effects of changes in the speech level we considered 3 different levels in the testing database: 12, 22, 32 dB Below Codec Overload (BCO), i.e. from the overload point of 16 bit word length, whereas the effects of background noise on VAD performance was tested by adding various types of stationary and non-stationary background noise (Car, White, Traffic and Babble), made available by CSELT, to the clean testing sequence at different signal-to-noise ratios (20, 10, 0 dB). The learning database consists of only clean sequences, so the trained fuzzy system used for the matching phase is independent of any type of background noise.

To summarize, the learning database comprises clean speech sequences at 22 dB below codec overload lasting about 4 minutes, whereas the testing database includes clean speech and noisy sequences corresponding to about 342 minutes of signal, divided in 57 files of 6 minutes each (6 types of superimposed noise, white, car, street, restaurant, office and train noise, with 3 different SNRs and 3 different levels, plus 3 clean files at different levels).

2.2 Fuzzy rules

After the training phase we obtained a knowledge base of only six fuzzy rules. Figure 1 shows the six fuzzy rules the tool extrapolated from the examples. In the rows we have the rules, and in the first four columns the four fuzzy system inputs. Each of the fuzzy sets represented has the Universe of Discourse corresponding to the relative input on the abscissa and the truth values on the ordinates. The crisp value of the output singleton is presented in the last column. More specifically, we say that 0 is inactivity and 1 is activity. We adopted the Weighted Mean defuzzification method which offers better results in classification problem [7].

If we neglect very large fuzzy sets we can give a linguistic representation of the six fuzzy rules:

Rule 1 :	**IF** (ΔS is medium-low) **THEN** (Y is active)
Rule 2 :	**IF** (ΔE_f is very high) **THEN** (Y is inactive)
Rule 3 :	**IF** (ΔE_l is low) **AND** (ΔS is very low) **AND** (ΔZC is high) **THEN** (Y is active)
Rule 4 :	**IF** (ΔE_l is low) **AND** (ΔS is high) **AND** (ΔZC is medium) **THEN** (Y is active)
Rule 5 :	**IF** (ΔE_l is high) **AND** (ΔS is very low) **AND** (ΔZC is low) **THEN** (Y is active)
Rule 6 :	**IF** (ΔE_l is high) **AND** (ΔS is not low) **AND** (ΔZC is very high) **THEN** (Y is active)

Of course, the output of the fuzzy system, which indicates the degree of membership in the Activity/Inactivity classes. depends on an overall evaluation of the input parameter values by means the defuzzyfication process. For example, we have a high output (i.e. the frame is detected as active) if ΔE_f is not high and ΔS is medium-low, whereas we have a low output (i. e. the frame is detected as inactive) if ΔS is medium and ΔE_f is very high and ΔZC is not high; in this last case, in fact, only the degree of truth of rule 2 is high, while the degree of truth of the other rules is low.

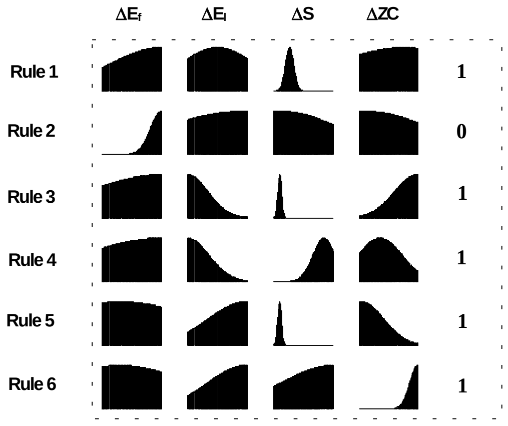

Fig. 1 FVAD rules

2.3 Decision module

In order to establish an optimal threshold value with which to compare the fuzzy system output, we analyzed the total misclassification error with respect to a threshold value , F_{th}, ranging between 0 and 1. The threshold was chosen in such a way as to achieve a trade-off between the values of the four parameters FEC+MSC+OVER+NDS. Although some of them (specifically MSC and FEC) can be improved by introducing a successive hangover mechanism, which delays the transitions from 0 to 1, the presence of a hangover block makes the values of the OVER and NDS parameters worse. The latter were therefore given priority over MSC and FEC in choosing the threshold.

The minimum total error is achieved with about F_{th} =0.21. We chose F_{th}=0.25 , so as to reduce the value of OVER and NDS; as mentioned previously, the corresponding increase in FEC and MSC can be solved by introducing a hangover mechanism. The threshold F_{th} was also chosen so as to minimize the variance of the parameters affected by the hangover: this then allows us to design a suitable hangover for our VAD. We used a VAD hangover to eliminate mid-burst clipping of low levels of speech. The mechanism is similar to the one used by the GSM VAD.

3. EXPERIMENTAL RESULTS

In this Section we compare the performance of the ITU-T G.729 standard VAD, the Full Rate GSM VAD and the FVAD proposed in this paper. All results were averaged on the six types of background noise: white, car, street, restaurant, office and train noise. The results were analyzed considering the percentage of FEC and MSC in active voice frames, to calculate the amount of clipping introduced, and the percentage of OVER and NDS in non-active voice frames, to calculate the increase in activity.

Figs. 2 (a-b) show a performance comparison in the case of a signal level of 22 dB below codec overload. Both in terms of clipping introduced (FEC+MSC) and in terms of increase in activity (OVER+NDS), the FVAD performs better than the G.729 except in the clean case for which performance is similar. At SNR=10 dB, for example, we halved both misclassification errors. We also observed that on average FVAD performance is similar to that of the GSM VAD, which in turn performs better than the ITU-T standard. In Figs. 2 (c-d) we compare performance in the case of a signal level of 32 dB below codec overload. The performance of the FVAD and G.729 VAD is substantially unchanged whereas we observed an improvement in that of the GSM VAD in terms of the activity factor but a deterioration in terms of clipping, above all with very high and very low SNRs. Finally, Figs. 2 (e-f) show a comparison in the case of a signal level of 12 dB below codec overload. In terms of FEC+MSC, the FVAD still performs better than the G.729 (in fact the performance is substantially unchanged with respect to the 22 dB BCO case), whereas we observed a slightly an improvement in the GSM VAD performance in terms of clipping. In terms of OVER+NDS the GSM VAD presents worse performance when the SNR is below 10 dB due to the high signal level. We observed a deterioration in the performance of both the FVAD and the G.729 in the clean case, whereas below SNR=20 dB FVAD performance is better than that of both the G.729 and the GSM VAD.

A performance evaluation in terms of FEC+MSC+OVER+NDS with varying types of background noise is shown in Fig. 3. The FVAD results are always better than those of the G.729. More specifically, we have a significant improvement in the case of car, train and street noises. Further, for non-stationary background noise, FVAD performance is also better than that of the GSM VAD, whereas for stationary noise, performance is similar except for the car noise case.

We also made comparisons considering several sequences of modern and classical music, sampled at 8 kHz. More specifically, we calculated the percentage of clipping introduced by the 3 different VADs. The results indicate that the GSM VAD introduces about 5 % of clipping, the G.729 VAD 20 % and the FVAD 14 %.

To evaluate the efficiency of new VAD in terms of perceived speech quality and the effect on listeners of the clipping introduced we carried out a series of listening tests. We used the Comparison Category Rating method, in the same conditions adopted in [9] extending the requirements about the SNR up to 0 dB. Fig 4 gives the results in terms of CMOS values, i.e. the differences in MOS scores between the FVAD and the ITU-T VAD. In average the FVAD presents similar performance than the G.729 VAD. For car and office noise at SNR=0 dB FVAD performs better of about 0.2 MOS scores.

4. CONCLUSION

In conclusion, we have presented a new voice activity detector based on fuzzy logic. The new approach is more efficient than the traditional threshold method since it exploits all the information and the non-linearity in the input pattern of parameters. The six fuzzy rules, on which the matching phase is based, were obtained through a training phase performed by means of a new hybrid learning tool, without any need for a priori knowledge of the problem. The results obtained show a clear improvement in fuzzy VAD performance as compared with the traditional solution, with a negligible increase in complexity. On average, FVAD allows an improvement of about 25 % in bit rate reduction and of about 43 % in clipping reduction. Formal listening tests, based on Comparison Category Rating method, show also a slightly improvement in the perceived speech in terms of clipping audibility above all with high level of background noise.

5. REFERENCES

[1] R. V. Cox, P. Kroon, “Low Bit-Rate Speech Coders for Multimedia Communication”, *IEEE Communication Magazine*, Dec. 1996, pp. 34-41.

[2] F. Beritelli, S. Casale, “Robust Voiced/Unvoiced Speech Classification using Fuzzy Rules”, *Proc. IEEE Workshop on Speech Coding*, Pennsylvania, USA, Sept. 1997, pp. 5-6.

[3] K. Srinivasan, A. Gersho, “Voice Activity Detection for Cellular Networks”, *IEEE Workshop on Speech Coding for Telecommunications*, Oct. 1993, pp. 85-86.

[4] J. Stegmann, G. Schroeder, “Robust Voice Activity Detection Based on the Wavelet Transform”, *Proc. IEEE Workshop on Speech Coding*, Pennsylvania, USA, Sept. 1997, pp. 99-100.

[5] Rec. ITU-T G.729 Annex B, 1996 .

[6] ETSI GSM 06.32 (ETS 300-580-6), September 1994.

[7] M. Russo, “FuGeNeSys: A Fuzzy Genetic Neural System for Fuzzy Modeling”, to be appear in *IEEE Transaction on Fuzzy Systems*.

[8] C.B. Southcott et al. “Voice Control of the Pan-European Digital Mobile Radio System” *ICC ’89*, pp. 1070-1074.

[9] D. Pascal “Results of the Quality of the VAD/DTX/CNG of G.729 A (CCR Method)” *ITU-T contribution*, Geneva, 27 May - 6 June, 1996.

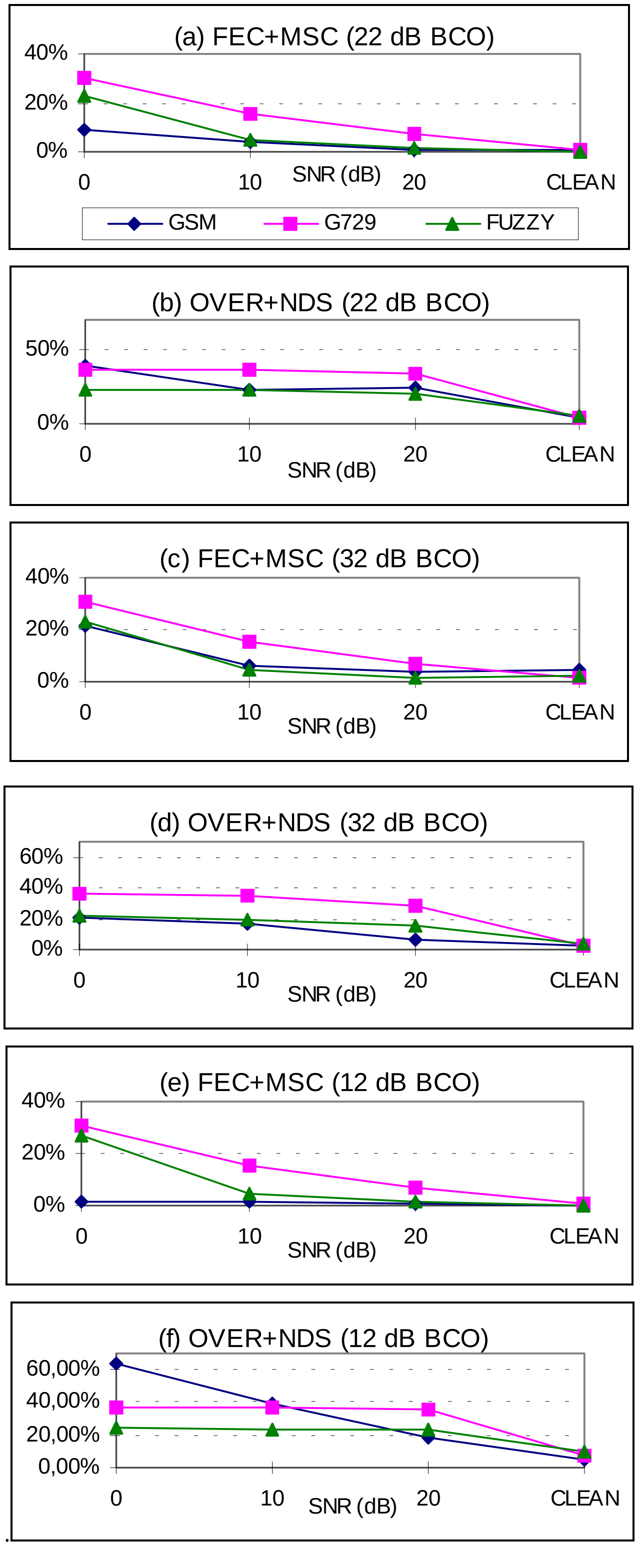

Fig. 2 (a-f) Clipping and increase of activity varying speech level

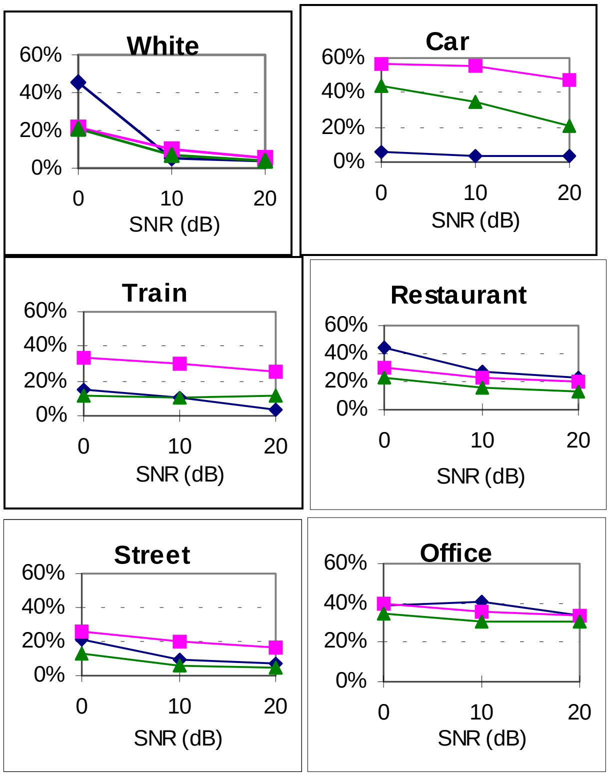

Fig. 3 Results varying types of background noise

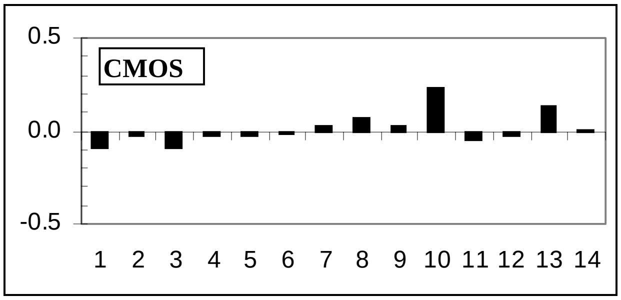

Fig. 4 Comparison Mean Opinion Score varying acoustic condition (a minus sign means that VAD proposed is worse than the ITU-T VAD)

1 - Clean
2 - Car SNR =20 dB
3 - Babble SNR = 20 dB
4 - Traffic SNR = 20 dB
5 - Office SNR = 20 dB
6 - Car SNR =10 dB
7 - Babble SNR = 10 dB
8 - Traffic SNR = 10 dB
9 - Office SNR = 10 dB
10 - Car SNR =0 dB
11 - Babble SNR = 0 dB
12 - Traffic SNR = 0 dB
13 - Office SNR = 0 dB
14 - Mean value