A Multi-Channel Speech/Silence Detector based on Time Delay Estimation and Fuzzy Classification

Francesco Beritelli, Salvatore Casale, Alfredo Cavallaro
Istituto di Informatica e Telecomunicazioni - University of Catania
V.le A. Doria 6, 95125 Catania - Italy
e-mail: {beritelli, casale, acavallaro}@iit.unict.it

ABSTRACT

Discontinuous transmission based on speech/pause detection represents a valid solution to improve the spectral efficiency of new-generation wireless communication systems. In this context, robust Voice Activity Detection (VAD) algorithms are required, as traditional solutions present a high misclassification rate in the presence of the background noise typical of mobile environments. The Fuzzy Voice Activity Detector (FVAD) recently proposed in [1], shows that a valid alternative to deal with the problem of activity decision is to use methodologies like fuzzy logic. In this paper we propose a multichannel approach to activity detection using both fuzzy logic and time delay estimation. Objective and subjective tests confirm a significant improvement over traditional methods, above all in terms of a reduction in activity increase for non stationary noise.

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. 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 between speech activity and inactivity [2]. 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 the performance of a VAD is critical for the overall speech quality, above all with low SNRs. 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 [1][2][3]. Though the Fuzzy Voice Activity Detector (FVAD) recently proposed in [1] performs better than the solutions presented in literature [4][5], it exhibits an activity increase, above all in the presence of non-stationary noise. This paper therefore deals with the problem of speech activity detection in noisy environments, using parameters available only in a multichannel system, such as delay estimation and the difference in the power level between the channels. Referring to systems which use arrays of microphones for various purposes, such as privileging a certain incoming signal direction or detecting and locating a signal source to be found in the reception area [6][7][8][9][10], we thought that a multi-channel system could be useful to obtain new information to be used in the detection of speech activity in noisy environments. This idea is also confirmed by recent studies on the central nervous system. It has been observed, in fact, that the *medial nucleus of the superior olive*, the structure whose neurons receive signals from both ears, is related with the location of sounds, which are said to be distinguished according to the differences in their arrival times. These time differences, in fact, are one of the elements used to locate sounds in space: a sound coming from a certain side reaches the ear on that side first and then, a few tens of milliseconds later, the other ear. Starting from this consideration, VADs have been developed which use a fuzzy system whose input is the Time Delay Estimation (TDE), the difference in the levels of the signals on the two microphones, and the continuous output of the FVAD, thus obtaining greater precision in detecting between activity and non-activity. Finally, objective and subjective tests demonstrate that the new approach, with a negligible increase in complexity, considerably increase the spectral resources, maintaining the same perceptive quality as the ITU-T G.729 VAD.

2. THE FUZZY VAD (FVAD)

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 [4] and are: the full-band energy difference, the low-band energy difference, the zero-crossing difference and the spectral distortion. The matching phase is performed by a set of fuzzy rules obtained automatically by means of a new hybrid learning tool [11]. 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 (defuzzyfication process). 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) [12] 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 [5].

3. THE TIME DELAY FVAD

The use of more than one microphone in a speech activity detection system gives the opportunity to exploit the parameters obtained by two registrations of the same signal, which would not be available if only one microphone were used. Below we present some new solutions for activity/non-activity detection, either using two parameters (delay estimate between the two signals and FVAD continuous output) or combining them with others such as the difference in the signal level between the two microphones. The delay between two signals is generally determined by plotting the x-axis on which the maximum cross-correlation function between the two signals is obtained, or by pre-filtering the signals before calculating the cross-correlation, thus obtaining the so-called generalised cross-correlation [13][14][15][16]. An algorithm we used for delay estimation is based on calculation of the Euclidean DIstance (EDI) between the two displaced signals, using 80-sample windows [17]. In the first system proposed, D_F, we trained a fuzzy network inputting the continuous output of the FVAD with no hangover and the delay between the two channels, . The output of the fuzzy system is a continuous value which ranges from 0 to 1. 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 and 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 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. In the second system, DSL_F, we trained a fuzzy network inputting the continuous output of the FVAD with no hangover, the delay between the two channels, and the power level on the two microphones. Bearing in mind that the voice source is closer to the main microphone, it is plausible that in speech activity periods the power level of the signal recorded on the main microphone will be higher than that on the other microphone. For the threshold and hangover we use the same approach of the D_F VAD. A scheme of how the D_F and DSL_F work is shown in Fig. 1. Of course an approach of this kind means a slight increase in the computational load.

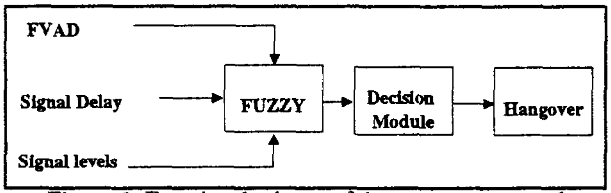


Figure 1. Functional scheme of the systems proposed

4. PERFORMANCE EVALUATION

As regards the objective parameters, the performance was evaluated considering the total contribution of FEC (Front End Clipping) and MSC (Mid Speech Clipping) in active voice frames to calculate the amount of clipping introduced, and the total contribution of OVER and NDS (Noise Detected as Speech) [12] in non-active voice frames to calculate the increase in activity. To estimate the amount of clipping which causes most perceived degradation in quality we propose considering the following three subclasses of objective parameters as well:

- VDN (Voiced Detected as Noise), which gives the number of voiced activity frames detected as noise;
- MDN (Mixed Detected as Noise), which gives the number of mixed activity frames detected as noise;
- UDN (Unvoiced Detected as Noise), which gives the number of unvoiced activity frames detected as noise;

The distribution of clipping length only over voiced and mixed frames by means of outlier measurements is also given in this paper. In this way we have an estimate both of the amount of short clipping, which produces little perceived degradation, and of the amount of long clipping, which produces a more relevant perceived degradation. The total percentage of reduction in the increase in activity and the total clipping reduction as regards a reference VAD are given using the following definition:

Total percentage of reduction in the activity increase =

$$= \left\{ 1 - \left[\frac{(\mathrm{OVER} + \mathrm{NDS} - \mathrm{FEC} - \mathrm{MSC})_{\mathrm{New\_VAD}}}{(\mathrm{OVER} + \mathrm{NDS} - \mathrm{FEC} - \mathrm{MSC})_{\mathrm{Ref\_VAD}}} \right] \right\} \cdot 100$$

Total clipping reduction percentage =

$$= \left\{ 1 - \left[\frac{(\mathrm{FEC} + \mathrm{MSC})_{\mathrm{New\_VAD}}}{(\mathrm{FEC} + \mathrm{MSC})_{\mathrm{Ref\_VAD}}} \right] \right\} \cdot 100$$

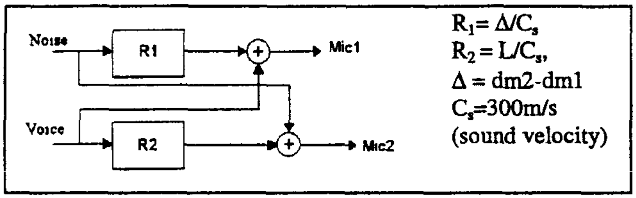
As regards the subjective tests, in order to evaluate the perceived quality degradation due to Discontinuous Transmission (DTX), we followed the approach proposed in [18] based on the Comparison Category Rating method (CCR).

5. SPEECH DATABASE

In order to make objective test the proposed algorithms, two different databases were used, one simulated and one real. The first was obtained by simulating a real situation in which 2 spatially separated microphones pick up two displaced signals in the presence of various types of noise, and the second was made up of telephone conversation recordings made in different environments using a telephone with two microphones. Both databases were then subdivided into a learning and testing database, the latter naturally containing different phrases and speakers from the former. The speech material used in the subjective tests consisted of simple, meaningful, short sentences arranged in pairs and lasting 10 seconds. The two speech sentences within a pair are separated by a pause of 3-4 seconds. All factors and reference conditions are reported in [18].

5.1 Simulated Database

The simulation system has to reproduce a real telephone conversation on a telephone with two microphones placed at a distance of about 10 cm from each other. The two signals recorded on the two microphones will be the sum of the desired speech signal and background noise, with various suitable delays. [17]. The scheme used to obtain the simulated database is shown in Fig. 2. L indicates the distance between the two microphones, dm1 the distance between mic1 and the noise source and dm2 that between the second microphone (mic2) and the noise source. The distance L between the two microphones was set to 10 cm, and dm1 was set to 10m. The speech phrases used to obtain the simulated database contains sequences recorded in a non-noisy environment (Clean sequences, SNR=60 dB), sampled at 8000 Hz and linearly quantized at 16 bits per sample. The database was 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 in the database. 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. We considered speech sequences at 22 dBovl 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). Movement of the noise source is simulated by generating a random number every 1000 frames, on the basis of which the position of the noise source may or may not vary. Any variation causes an increase in the angle ψ by steps of 30°, while the distance from the origin of the axes remains unaltered. The initial value of the angle ψ is set to 0°.



$R_1 = \Delta/C_s$
$R_2 = L/C_s$,
Δ = dm2-dml
C_s=300m/s
(sound velocity)

Figure 2. Scheme used to obtain the simulated database.

5.2 Database Obtained

The database obtained comprises speech sequences sampled at 8000 Hz and linearly quantized at 16 bit/sample, recorded by means a telephone with 2 microphones. One of the two microphones is located in the classical position, while the other is placed at a distance of about 10 cm, on top of the telephone. By means of the two microphones, the telephone picks up two signals that are a combination of a speech signal and environmental noise. Each channel is then amplified and then sent as input to the audio board. The signals of the two channels at the system output are two recordings of the same signal which are different due to the different distances between the sources and the microphones. The database contains 18 telephone conversations made by 6 different speakers, 3 males and 3 females. Each speaker recorded 3 conversations in different environments; more specifically, a noiseless environment, one with traffic noise and one with babble noise. Each conversation lasted for about two minutes. The marking needed for VAD performance evaluation was performed manually, detecting the active speech and non-active speech segments. These were divided into 10-ms frames and were given a flag (0 or 1) according to the class they belonged to (Activity/Non-Activity). Table 1 shows the structure of the database.

Speakers	**Language**	**# Frame (10 ms)**
Male	Italian	97380
Female	Italian	95580
Total Activity		46224
Total Non-Activity		146736

Table 1: Structure of the Database Obtained

6. EXPERIMENTAL RESULTS

In this Section we compare the performance of the standard VAD ITU-T G.729 [4], the FVAD [1] and the new solutions proposed in this paper, distinguishing between applications using the simulated database and those using the one based on recordings of real conversations. In tables 2, 3 and 4 we compare the performance of the algorithms D_F with the performance of the G.729 Annex B and FVAD. All results were averaged on the six types of background noise: white, car, street, restaurant, office and train noise. Table 2 gives the results obtained with the simulated database using various SNRs, in terms of the amount of clipping introduced, FEC+MSC, and in the increase of activity, OVER+NDS, total error (FEC+MSC+OVER+NDS). In terms of clipping introduced, increase of activity and total error, the D_F performed better than the G.729 and the FVAD at all SNR values. Table 3 gives the results in terms of VDN, UDN and MDN to estimate the voiced to silence, unvoiced to silence and mixed to silence misclassification. We can observe a better performance of D_F on all parameters. Table 4 shows the amount of clipping introduced only on voiced and mixed frames, of a length greater than X frames, where X=0, 10, 40. As the values show, the D_F introduce a lower amount of clipping than the G.729 and FVAD solution; it is comparable with FVAD for SNR=20 dB, but it performs better than G.729. In all other cases it performs better thna G.729 and FVAD, except for outliers greater 40 where the performances are comparable.

SNR	FEC+ MSC(%)			OVER+NDS (%)			TOTAL ERROR (%)		
(dB)	G.729	FVAD	D_F	G.729	FVAD	D_F	G.729	FVAD	D_F
00	30,23	23,11	15,68	36,78	23,44	10,94	67,01	46,55	26,62
10	15,38	4,53	3,86	35,91	22,23	10,34	51,29	26,75	14,20
20	7,32	1,24	0,97	33,50	20,53	7,21	40,82	21,77	8,18

Table 2: Objective comparisons in terms of FEC+MSC, OVER+NDS, Total Error

SNR	VDN (#)			UDN (#)			MDN (#)		
(dB)	G.729	FVAD	D_F	G.729	FVAD	D_F	G.729	FVAD	D_F
00	3056	1969	1813	4609	1765	1425	3403	1816	1316
10	870	113	105	2619	459	351	1853	748	571
20	5	7	5	39	21	18	40	43	35

Table 3: Comparisons in terms of misclassification

SNR	Outliers>0 (#)			Outliers>10 (#)			Outliers>40 (#)		
(dB)	G.729	FVAD	D_F	G.729	FVAD	D_F	G.729	FVAD	D_F
20	24	4	4	7	2	2	6	2	2
10	1074	212	178	70	10	8	2	2	2
00	2112	1084	956	183	103	95	3	2	2

Table 4: Comparisons in terms of outliers

Table 5 gives the results obtained applying the FVAD and the DSL_F to the acquired (real) database. In this case, all results were averaged on the three types of background noise, present in the acquire database. As far as activity clipping is concerned, FVAD and DSL_F performance is comparable, as was to be expected. In terms of activity increase, on the other hand, the DSL_F performs decidedly better than the FVAD. Finally, considering the total misclassification error, the DSL_F gave the lowest values. Table 6 gives the results in terms of VDN, UDN and MDN to estimate the voiced to silence, unvoiced to silence and mixed to silence misclassification. We can observe a better performance of D_F on all parameters. Table 7 shows the amount of clipping introduced only on voiced and mixed frames, of a length greater than X frames, where X=0, 10, 40. As the values show, the D_F introduce a lower amount of clipping than the FVAD solution. To summarise, the objective performance of the D_F is better than that of the FVAD in that provides a total percentage of reduction in the increase in activity of about 65 % and a total clipping reduction percentage of 29 %, while the DSL_F provides a total percentage of reduction in the increase in activity of about 60 % and a total clipping reduction percentage of 11 %, (Table 8) [10]. We must remember that these results were calculated on different database for the D_F and the DSL_F. Informal listening tests done comparing the D_F with the G.729 VAD and the DSL_F with the G.729 VAD, indicate that the new solutions can be considered statistically equivalent to G.729 with the advantage of a strong improvement in the spectral efficiency.

FEC + MSC (%)		OVER + NDS (%)		TOTAL ERROR (%)	
FVAD	DSL_F	FVAD	DSL_F	FVAD	DSL_F
2,71	2,40	8,73	4,14	11,44	6,54

Table 5: Objective comparisons in terms of FEC+MSC, OVER+NDS, Total Error

VDN (#)		UDN (#)		MDN (#)	
FVAD	DSL_F	FVAD	DSL_F	FVAD	DSL_F
45	32	154	111	76	48

Table 6: Comparisons in terms of misclassification

Outliers>0 (#)		Outliers>10 (#)		Outliers>40 (#)	
FVAD	DSL_F	FVAD	DSL_F	FVAD	DSL_F
62	41	6	4	2	2

Table 7: Comparisons in terms of outliers

	D_F	DSL_F
Total percentage of reduction in the increase in activity	65,28 %	60,59 %
Total clipping reduction percentage	28,98 %	11,43 %

Table 8: Average improvement as regards FVAD

7. CONCLUSION

In conclusion, we have presented a new voice activity detector based on time delay estimation and fuzzy logic. The new approach is more efficient than the traditional threshold method since it exploits all the information in the pattern of input parameters due to the fuzzy logic approach and the multichannel information due to the presence of two microphones. When a fuzzy network trained with input given by the FVAD output and delay estimates is used, the results obtained reflect the positive features of the FVAD and delay estimate techniques, giving a great reduction in activity increase which is fundamental for more efficient use of the communication channel. A series of subjective tests demonstrate that, in clean speech conditions at all input levels and for all types of noisy environments, the algorithms are statistically equivalent to the recent VAD standardized by ITU-T.

6. REFERENCES

[1] F. Beritelli, S. Casale, A. Cavallaro, "Improved VAD G.729 Annex B for Mobile Communications Using Soft Computing", Contribution ITU-T, Study Group 16, question 19/16, Washington, 2-5 September 1997.

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

[3] J. Stegmann, G. Schroeder, "Robust Voice Activity Detection Based on the Wavelet Transform", Proc. IEEE Workshop on Speech Coding, Pocono Manor, Pennsylvania, USA, September 7-10, 1997, pp. 99-100.

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

[5] ETSI GSM 06.32 (ETS 300-580-6) "European digital cellular telecommunications system (Phase 2); Voice Activity Detection (VAD)", September 1994.

[6] M. S. Brandstein and H. F. Silverman. A new time-delay estimator for finding source locations using a microphone array. LEMS Technical Report 116, LEMS, Division of Engineering, Brown University, Providence, RI 02912, March 1993.

[7] Knapp, C.H., and Carter, G.C., "Estimation of time delay in the presence of source or receiver motion," J. Acoust. Soc. Amer., v.61, n.6, pp. 1545-1549, 1977.

[8] Martin Drews. Time Delay Estimation For Microphone Array Speech Enhancement Systems. Esca. Eurospeech 95. 4° European Conference on Speech Communication and Technology. Madrid, September 1995. ISSN 1018-4074

[9] M. Omologo, P. Svaizer, "Talker Localization and Speech Enhancement in a Noisy Environment using a Microphone Array based Acquisition System", Proceedings Eurospeech, Berlin, September 1993, pp. 605-609.

[10] M. Omologo, P. Svaizer "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", Proc. ICASSP, Adelaide 1994, pp. 11273-11276.

[11] M. Russo, "FuGeNeSys: Fuzzy Genetic Neural System for Fuzzy modelling", to appear in IEEE Transaction on Fuzzy Systems.

[12] C.B. Southcott et al. "Voice Control of the Pan-European Digital Mobile Radio System" ICC '89, pp. 1070-1074.

[13] IEEE Trans. Acoust., Speech, Signal Processing. Special Issue on Time-Delay Estimation, volume ASSP-29, June 1981.

[14] Knapp, C.H., and Carter, G.C., "The generalized correlation method for estimation of time delay," IEEE Trans. Acoustics, Speech and Signal Processing, v.ASSP-24, n. 4, pp. 320-327, 1976.

[15] G. C. Carter, Coherence and Time Delay Estimation. IEEE Press, 1993.

[16] Segal, M., Weinstein, E., and Musicus, BR., "Estimate-maximize algorithms for multichannel time delay and signal estimation," IEEE Trans. Signal Processing, v.39, n.1, pp. 1-15, 1991.

[17] F. Beritelli, S. Casale, A. Cavallaro, "A Multi-Channel Approach to Voice Activity Detection in Noisy Environments based on Time Delay Estimation", IEEE 10th Tyrrenian International Workshop on Digital Communications, Ischia (Napoli) Italy, Sept. 15-18

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