Improving Voice Activity Detection Used in ITU-T G.729.B

H. Farsi, M. A. Mozaffarian, H. Rahmani
Department of Electrical Engineering
University of Birjand
P.O. Box: +98-97175-376
IRAN
h.farsi@birjand.ac.ir amozaffarian@birjand.ac.ir

Abstract: - In this paper, by using a new envelope estimation algorithm and a geometrical adaptive threshold method; we present a novel method to improve the performance of ITU-T G.729.B systems in various noisy environments. The proposed system has minimum change from this standard.
We compare the performance of the proposed method with G.729B, ETSI AMR option 1 and 2 using objective measures.

Key-Words: - voice activity detection, speech quality evaluation, computational complexity, low SNR, GAET, TE-LPC and LSPE.

1 Introduction

As is well-known, voice activity detection (VAD) achieves silence compression, which is important in both fixed and mobile modern telecommunication systems [1]. In communications systems based on variable bit rate speech coders, it represents the most important block, reducing the average bit rate; in a cellular radio system using the discontinuous transmission (DTX) mode, 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.

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. This method provides a poor performance, in noisy environments especially for non-stationary noise or low Signal-to-Noise Ratios (SNRs), compare with other standards (e.g., ETSI AMR option 1 and 2 [3]) and new methods (e.g., statistical or estimation methods [10], [11]). In this paper, we modify G.729 method by minimum change in its structure and compare its performance with other standards by objective measures.

2 VAD in G.729

Due to the different application scenarios, the VADs considered operate on frames of different lengths. In G.729 method it has 10 ms length. G.729 uses the following four classification parameters:

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

The G.729 VAD uses a multiboundary decision region in the space of the four parameters [4]. Unfortunately these parameters do not provide a good performance in various environments, that caused by parameters features. As an example the zero crossing rate has problems at low SNRs, especially in the presence of noise and speech with high zero crossing rates and the energy threshold method has problems in non-stationary noise and low SNRs. Therefore, in this paper, by changing some blocks in G.729 diagram, we try to increase the system performance in various environments also in low SNRs.

3 Parameters used for minimization

As we describe in section 2, the G.729B standard needs to have a set of modifications to receive the better performance. Before describe this modification, we indicate some new methods:

3.1 TE-LPC

The true envelope estimator and then using the band limit envelope to derive an all pole envelope model named TE-LPC. This proposition to improve the spectral envelope estimation is based on the *true envelope* estimator. The resulting estimation can be interpreted as a band limited interpolation of the observed sub-sampled spectral envelope [5].

Related to the speech signal, the resulting predictor is not optimal in the sense of the MSE criteria but it is supposed to fit closer the spectral envelope. A

comparison between LPC and TE-LPC is shown in Fig.1.

The results show that TE-LPC performs better Spectral-Peak Flatness Measure (SPFM) maximization in all the cases we measured. While the improvements are rather small if measured over the whole spectrum or the high frequency band, they are significant in the low frequency band which is more perceptually important. The improvement is bigger for high-pitched signals and is not very sensitive to the model order for the selected order values. Improvements found for the unvoiced cases could be due to voiced and mixed parts remaining in the unvoiced segments [6].

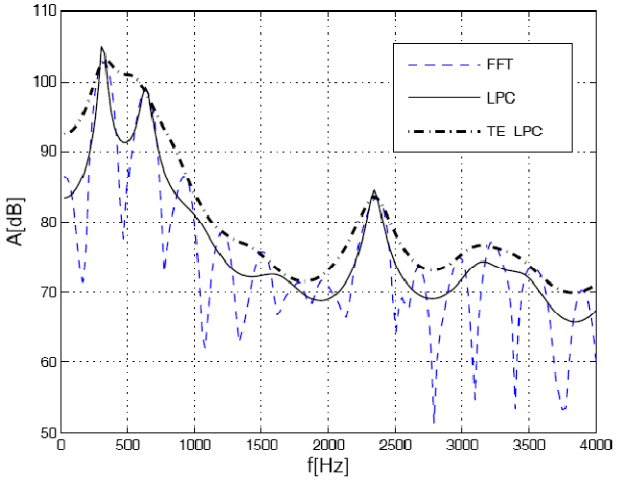

Fig.1. Example of LPC and TE-LPC spectral fitting

Since we try to make minimum change in G.729 standard and provide a better performance, therefore after TE-LPC calculation, we transform these coefficients to Line Spectral Frequency (LSF).

3.2 GAET

Recently Özer and Tanyer developed a new technique to estimate the optimum threshold for noise in the presence of speech accurately by using the amplitude probability distributions. The geometrically adaptive energy threshold (GAET) method is developed to set the threshold level adaptively without the need of voice-inactive segments by using the amplitude probability distributions of the speech signal [7]. The GAET method is robust to non-stationary noise but false triggering is often observed when noise has short burst.

3.3 LSPE

Tucker designed a VAD based on periodicity [8] named *Least-Square Periodicity Estimator* (LSPE). The major difficulty in designing a VAD based on periodicity is its sensitivity to any periodic signal which may well be interference or a background signal. Great care should be taken to avoid false triggering on non-speech periodic signals. If the speech signal contains non-periodic components, inaccurate values for endpoints of the voice-active segments could be obtained. Tucker used a preprocessor to detect and if possible remove, most of the expected types of interference. Different environments will have different interference, so the exact nature of the preprocessor will depend on the expected type of interference.

4 Modifying G.729B

In Fig.2 we introduce a modified version of G.729 Annex B standard, based on four differential parameter that describe in section 2.

The energy threshold methods (include of Low and Full-Band energy) have some problems in non-stationary and low values of SNR. As we seen in new methods to modify the performance of VAD systems in the case of energy threshold [10]-[11], we used an adaptive threshold. In this way, we use one of the best methods that work rapid and have a good performance especially in Low-SNR. It is the GAET method.

Zero-Crossing (ZC) method has a better performance. However, this method also has some problems in Low-SNRs especially in presence of periodic noise and speech with high Zero-Crossing Ratios. To solve the problem of periodic speech we use the LSPE method. By using this method, the ZC eliminate in periodic

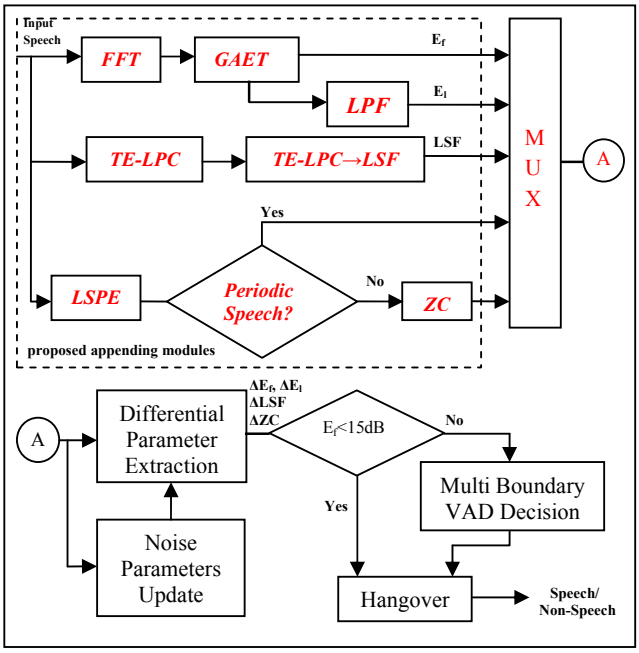

Fig.2. Block diagram of Modified ITU-T G.729B VAD

speech that cause some improvements in rate of the system in the case of Low-SNR.

To repair the LPC coefficients and estimate better spectrum shape of the speech signal, we use TE-LPC method. It has a good performance also its rate is two times better than LPC method.

We compare the complexity of commonly used VAD algorithms with our proposed method in Fig.3.

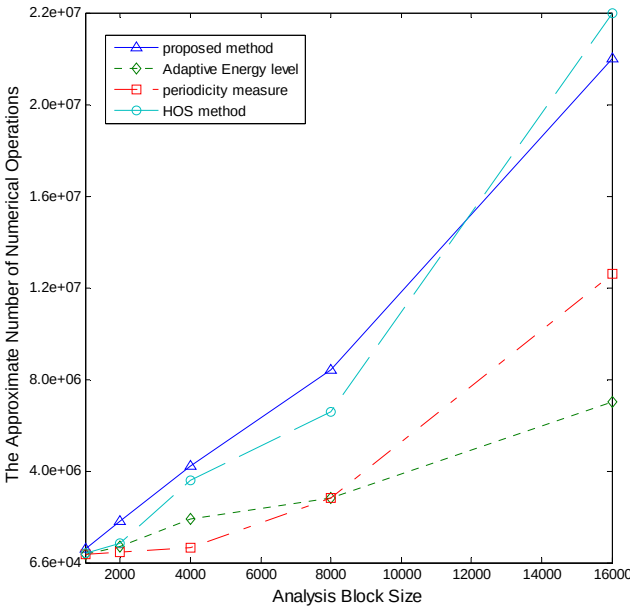

Fig.3. Approximate numbers of numerical operations by the algorithm as a function of analysis block size.

5 Parameters used for the comparison

Using the implemented system outlined in Section 4, the effectiveness of the proposed algorithm was evaluated. Surveying literature indicates two distinct schools pertaining to VAD evaluation, namely subjective and objective evaluation. In general, subjective evaluation methods attempt to determine the effect of erroneous VAD decisions on human perception [9]. Tests such as the ABC [9] however does not take into consideration the effect of false alarms and as such are inappropriate for a thorough evaluation of VAD performance. Therefore, in order to evaluate the performance of the proposed scheme objective evaluation was used.

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:

• *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.

• *Correct VAD decision* (Correct)*:* Correct decisions made by the VAD.

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

6 Results

VAD performance comparison is complicated and time consuming process. It should be considered carefully. Ideally, a VAD should maximize the correct value, and minimize all errors. However failing this, the affect different types of errors have on the discontinuous speech signal (speech signal with non-speech periods removed and comfort noise inserted) should be considered. The purpose of a VAD in the context of a telephone conversation is to enable data savings by not transmitting non-speech periods, while maintaining speech quality. Speech quality should be of utmost importance. Therefore, it is important to note the affect that each of the different errors have on speech quality.

In contrast, insertion errors such as NDS and OVER do not have any effect on speech quality. They do however result in reduced effectiveness of the VAD scheme. Here we will use the broad notion that clipping errors are less desirable than insertion errors.

In Tab.1 we show a full comparison of three standards, ITU-T G.729 Annex B, ETSI AMR option 1 and 2 with our proposed method. As you see, in the case of G.729B, excluding the OVER metric, it exhibit the worst average results over the test set. We can obviously see that clipping errors were generally worst in the Gaussian noise environment.

In summary, the proposed scheme presents a better alternative compare with standardized algorithms. It exhibits a consistent correct rate over a variety of noise environments and conditions. It has lower average NDS than all standardized algorithms over the test set and has low FEC and MSC while maintaining a low OVER rate. These characteristics make it a simple and reliable choice for many VAD applications. Further, the scheme requires only low computation time and memory.

The simplicity of the proposed VAD coupled with the encouraging results, mathematical tractability and high detection consistency make it a good alternative to current schemes. The behavior of the VAD is easily altered by changing one meaningful parameter, and as

such makes the VAD well suited to varying applications.

6 Conclusion

G.729B standard uses four parameter for VAD system. Unfortunately it has a poor performance in low SNRs. In this paper, whereas the rate and complexity of this standard are better than spectral shape (i.e. GSM-FR) and sub-band energy standards (i.e. IS-95, AMR 1 and 2), we try to modify G729B standard with minimum changes.

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] 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

[5] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100–102, 1996. For V.70 digital simultaneous voice and data application," *IEEE Commun. Mag.*, vol. 35, pp. 64–73, Sept. 1997.

[6] F. Villavicencio, A. Röbel and X. Rodet, "Improving LPC Spectral Envelope Extraction of Voiced Speech by True-Envelope Estimation," *ICASSP*, pp. 869–872, Sept. 2006.

[7] H. Özer and S. G. Tanyer, "A geometric algorithm for voice activity detection in nonstationary Gaussian noise," in *Proc. EUSIPCO'98*, Rhodes, Greece, Sept. 1998.

[8] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng.*, vol. 139, pp. 377–380, Aug. 1992.

[9] F. Beritelli, S. Casale, and G. Ruggeri, "A physcoacoustic auditory model to evaluate the performance of a voice activity detector," in *Proc. Int. Conf. Signal Processing*, vol. 2, Beijing, China, 2000, pp. 69–72.

[10] A. Davis, S. Nordholm and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold" *IEEE Trasns. On audio, speech and language processing, vol. 14, NO. 2, March 2006,* pp. 412-424.

[11] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE ICASSP'98*, vol. 1, Seattle, WA, 1998, pp. 365–368.

Table.1 VAD Performance comparison for various SNRs and noise environments

Environment		**ETSI AMR 1**					**ETSI AMR 2**				
Noise	SNR	Correct	FEC	MSC	NDS	OVER	Correct	FEC	MSC	NDS	OVER
Gaussian	-5dB	57.93	3.57	31.7	6.8	0.00	73.62	1.71	24.6	0.08	0.00
Gaussian	-2dB	69.25	2.3	22.31	6.1	0.05	83.79	1.19	14.5	0.12	0.4
Gaussian	0dB	75.32	1.7	17.27	5.62	0.09	88.39	0.93	9.72	0.15	0.81
Gaussian	2dB	80.23	1.27	13.1	5.3	0.1	91.7	0.72	6.2	0.17	1.21
Gaussian	5dB	85.72	0.85	8.4	4.68	0.35	94.59	0.5	2.92	0.21	1.78
Babble	-5dB	55.96	0.65	12.39	29.6	1.4	59.03	0.63	11.3	18.44	10.6
Babble	-2dB	61.34	0.47	7.49	28	2.7	64.07	0.48	7	17.95	10.5
Babble	0dB	63.78	0.37	5.13	27.17	3.55	66.39	0.39	4.87	17.75	10.6
Babble	2dB	65.82	0.29	3.39	26.4	4.1	68.18	0.3	3.2	17.62	10.7
Babble	5dB	68.24	0.2	1.71	25.04	4.81	69.79	0.19	1.58	17.54	10.9
Vehicle	-5dB	96.49	0.1	0.25	1.26	1.9	93.8	0.05	0.02	2.24	3.89
Vehicle	-2dB	96.95	0.08	0.62	0.85	1.5	94.56	0.03	0.02	1.59	3.81
Vehicle	0dB	97.08	0.08	0.76	0.69	1.39	94.95	0.02	0.01	1.29	3.73
Vehicle	2dB	97.26	0.07	0.83	0.58	1.26	95.24	0.02	0.01	1.1	3.63
Vehicle	5dB	97.41	0.07	0.88	0.51	1.13	95.61	0.02	0.01	0.94	3.42
Environment		**ITU-T G.729B**					**Modified G.729B**				
Gaussian	-5dB	57.49	3.95	33.57	4.99	0.00	78.11	1.23	18.34	2.32	0
Gaussian	-2dB	63.24	2.72	29.08	4.96	0.00	87.21	1.19	9.53	2.07	0
Gaussian	0dB	66.66	2.12	26.27	4.95	0.00	91.03	1.16	5.84	1.97	0
Gaussian	2dB	69.77	1.67	23.61	4.95	0.00	93.36	1.11	3.54	1.98	0.01
Gaussian	5dB	73.94	1.21	19.89	4.95	0.01	95.12	1.01	2.04	1.82	0.01
Babble	-5dB	55.32	1.34	23.11	20.02	0.21	71.3	0.93	11.84	15.44	0.49
Babble	-2dB	57.59	1.07	21.06	19.98	0.3	71.63	0.7	14.85	12.43	0.39
Babble	0dB	59.37	0.92	19.53	19.83	0.35	72.16	0.58	15.63	11.27	0.35
Babble	2dB	61.34	0.8	17.91	19.62	0.33	72.92	0.51	15.64	10.61	0.32
Babble	5dB	64.4	0.64	15.41	19.24	0.3	74.38	0.43	14.66	10.23	0.3
Vehicle	-5dB	58.65	0.16	8.26	32.23	0.7	93.93	0.28	3.84	1.28	0.67
Vehicle	-2dB	62.42	0.19	7.04	29.82	0.53	94.96	0.23	3.03	1.26	0.53
Vehicle	0dB	64.35	0.2	6.27	28.71	0.47	95.41	0.19	2.71	1.22	0.47
Vehicle	2dB	65.96	0.19	5.58	27.84	0.43	95.69	0.18	2.52	1.17	0.44
Vehicle	5dB	68.13	0.17	4.54	26.72	0.44	95.95	0.15	2.37	1.09	0.44