A Voice Activity Detector for the ITU-T 8kbit/s Speech Coding Standard G.729

S. D. Watson, B. M. G. Cheetham*, P. A. Barrett#, W. T. K. Wong# and A. V. Lewis#*
**Department of Electrical Engineering and Electronics, The University of Liverpool, LIVERPOOL. L69 3BX. U.K.*
#BT Laboratories, Martlesham Heath, IPSWICH. IP5 7RE. U.K

ABSTRACT

Voice Activity Detectors (VAD's) are widely used in speech technology applications where available transmission or storage capacity is limited (e.g. mobile, DCME, etc.) and must be utilised with maximum economy. Modern day digital speech coding algorithms can provide toll quality speech at bit-rates as low as 8kbit/s (e.g. ITU-T G.729) and the use of a VAD can achieve further economy in average bit-rate. This paper presents a modified version of the GSM VAD, for use with the ITU-T 8kbit/s speech coding algorithm CS-ACELP, which makes an active/inactive decision for every 10 ms coding frame. The performance of the proposed voice activity detector is compared to that of the GSM coder in terms of VAD errors and subjective quality. Results indicate that the modified VAD has similar performance to the standardised GSM VAD while operating with G.729 parameters and coding frame size.

1. INTRODUCTION

The use of a voice activity detector with the ITU-T 8kbit/s speech coding standard G.729 [1] converts the fixed bit-rate codec into a variable bit-rate version that is able to exploit the redundant silence exhibited by conversational speech. The ITU-T have recently defined a VAD algorithm [2] for G.729 that is optimised for simultaneous voice and data applications and is not specifically designed to deal with high levels of acoustical background noise. The aim of the work reported in this paper was to develop a VAD algorithm primarily for speech only applications where high acoustical background noise levels may frequently be encountered (e.g. mobile, speech over Asynchronous Transfer Mode (ATM) [3], etc). The VAD can be used in mobile radio applications to reduce the mean RF interference to other users or to reduce the power consumption of hand-held terminals. It can also be used in Digital Circuit Multiplication Equipment (DCME) [4] to detect the speech activity on the incoming trunks. Due to the need for robust operation in high levels of noise the VAD was based on the GSM voice activity detector [5], which works well, in conjunction with the GSM codec, under these conditions. The application of the GSM VAD directly to G.729 is not possible since it uses GSM coder parameters and threshold values based on the 20ms frame length of GSM. The modified VAD uses raw parameters from the 10ms coding frame of G.729.

Voice activity detectors can exhibit four types of error and these are Noise Detected as Speech (NDS), Hang Over (HO), Front End Clipping (FEC), and Mid Speech Clipping (MSC). NDS occurs when the VAD incorrectly classifies the background noise as speech. HO is where the VAD delays its silence decision in order to stop very low amplitude mid speech sections being classified as silence, and thus clipping the speech. FEC occurs at the start of active speech while MSC occurs mid-way through, or at the end of, active speech bursts, and both occur when frames are incorrectly classified as silence. FEC and MSC occurs due to very low amplitude speech being swamped by relatively high levels of background noise. The first two errors, NDS and HO, quantify the accuracy of the algorithm while the final two errors, FEC and MSC, relate directly to the subjective quality of the VAD, and as such should be minimised. The VAD algorithm cannot afford to compromise quality and is therefore conservatively designed such that if the frame may contain speech then it is classified as speech.

Since the fixed bit-rate coder is inactive during frames classified as silence, the decoder receives no detailed information about these. This can be perceptually disturbing due to modulation of the background noise [6], especially in environments where the noise levels are high. To combat this, comfort noise injection is used at the decoder to produce frames which have similar characteristics to those of the actual background noise. By approximating the background noise, during frames classed as non speech by the VAD, there is a continuity of background noise power and the problem of noise modulation, or noise pumping, is removed.

2. THE VAD ALGORITHM

A simple way to implement a voice activity detector is energy thresholding, that is to say that if the energy of the current frame exceeds a threshold then it is classed as speech. This technique is inadequate for applications where high levels of acoustical background noise are encountered, since the high energy during non speech sections can be mis-classified as speech. The VAD used

in this work improves on simple thresholding by firstly attempting to adaptively remove a large proportion of the background noise energy from the corrupted speech and then comparing the energy of the output against an adaptive energy threshold.

The VAD algorithm comprises of two separate blocks. The main block uses time averaged Linear Prediction (LP) parameters, determined during the spectrally stationary noise, to inverse filter the noisy speech signal, removing most of the noise energy. The energy of the residual, or 'cleaned', signal is compared with an adaptive energy threshold with the result being input to a hangover module which makes the activity decision for the current frame. The second block comprises of a secondary VAD and a parameter adaptation module. The primary function of this block is to determine if the current frame is definitely non speech, updating the inverse filter parameters and the adaptive energy threshold if this is the case. A simplified block diagram of the VAD is shown in Figure 1 [7].

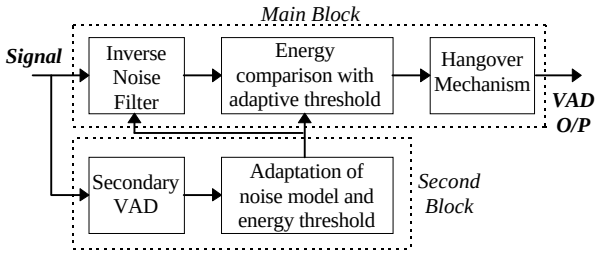

Figure 1 : Simplified Block Diagram of the Voice Activity Detector described in the paper.

The inverse filter attempts to flatten the spectrum of the corrupted speech signal in the areas where the energy of the background noise is greatest, removing most of the energy due to noise. The inverse filter uses the ten LP parameters from each G.729 frame, averaged over time, to model the noise spectrum. This is in contrast to the eight LP coefficients used every 20ms in GSM. The LP coefficients are updated when the secondary VAD signifies a non speech frame. Since it is assumed that noise is stationary, inverse filtering will remove approximately the same amount of noise energy from each successive frame. The energy of this *cleaned* speech is calculated and compared with an adaptive energy threshold, which is updated at the same time as the LP coefficients. If the energy of the inverse filtered speech is greater than the adaptive energy threshold for a number of successive frames, known as the hangover period, then speech activity is indicated. Hangover prevents low level mid-speech being classified as silence. It was found that fifteen 10ms frames of overhang produced adequate results, in contrast to the 5 overhang frames used in GSM VAD.

The secondary VAD determines when parameters used by the main algorithm are updated and must therefore have a very high confidence that frames it classifies as silence contain no speech. Using the assumption that acoustical background noise is stationary and non periodic, two different measures can be used for robust noise classification. The pitch of the current frame, supplied by the open loop pitch search of G.729, is used to determine if the local signal is periodic. If there is pitch continuity, i.e. the variation in pitch falls within a set tolerance, for the previous three frames then the parameters of the noise model and the adaptive energy threshold are not updated. The stationarity of consecutive frames is determined by comparing a distortion measure for those frames. The distortion measure calculates the amount of energy removed from the signal when it is inverse filtered using the LP coefficients derived during noise. For spectral stationarity the distortion measures of consecutive frames will be very close. If there is no spectral stationarity the VAD parameters are not updated. The secondary VAD will update the main VAD parameters after several strict conditions have been satisfied for several consecutive frames, ensuring that the update only occurs during pure background noise. The modified VAD uses threshold values, for the secondary VAD stage, based on the 10ms coding frame length of G.729 as opposed to the 20ms frame length of GSM.

During parameter update the adaptive energy threshold level is set high enough so that noise frames with a relatively high energy are not classed as active and low enough so that low energy speech frames are not classed as inactive. An adequate level is achieved at approximately three times the energy of the inverse filtered noise. The VAD learns quickly if the background noise is highly stationary and non periodic, e.g. car noise, but more slowly for multiple speaker or babble noise where there may be some amount of periodicity and spectral change. The operation of the energy threshold adaptation is shown by Figure 2, with the solid line showing the energy of the noise after inverse filtering and the dotted line indicating the adaptive energy threshold used by the main VAD on a frame by frame basis.

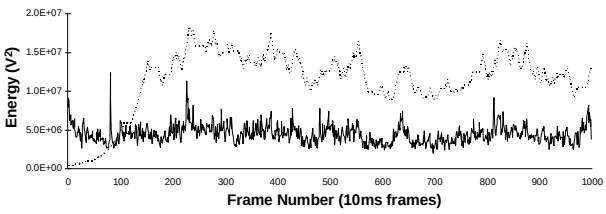

Figure 2 : Graph of adaptive energy threshold and short-term energy of the signal, after processing, for car noise.

It can be seen from figure 2 that secondary VAD learns fast in car noise and then tracks the noise floor closely at approximately three times its level. Other noise

types, such as babble and multiple speaker noise, can take substantially longer to adapt.

This adaptation of the energy threshold works well for speech contaminated by high levels of acoustical background noise. However, when the energy background noise, before inverse filtering, is very low (i.e. inaudible) the adaptation of the energy threshold is unnecessary and it is set to a fixed level, reducing the complexity of the algorithm.

3. COMFORT NOISE INJECTION

Comfort noise injection is a necessary part of the decoder when the encoder uses a VAD algorithm. The magnitude spectrum of background noise, modelled by the linear prediction coefficients, can be assumed to be relatively stationary. The excitation used to drive the spectral model of the background noise and the power of the resulting frames need to be determined. To validate that a stationary spectral model of the background noise is adequate to reproduce a replica of the noise, the background noise was inverse filtered using the current frames LP parameters, then the residual was used to drive a synthesis filter using static LP coefficients derived in noise. The generated signal closely represented the background noise. The only remaining difficulty was the representation of an excitation and frame power that would still retain the characteristics of the noise making the transition to comfort noise least noticeable.

The residual excitation for three different types of background noise were found to have Gaussian distribution with zero mean and varying values for standard deviation (which relates to frame power). Such an excitation was generated at the decoder and used to drive the synthesis filter, which used LP coefficients sent after the last active frame. The output from the filter was then power scaled so that it matched the average power of the last ten noise frames. This averaged power matching reduced the effect of spurious high energy noise frames. This simple comfort noise injection scheme is simpler than that used by GSM but still masks noise modulation.

4. TESTING AND RESULTS

Testing was carried out using speech corrupted with several types of acoustical background noise added at varying levels. The noise consisted of car, multiple speaker and babble noise and were added to the speech at 0, 10, 20, and 30 decibels below the average power of the speech. Each test vector was 36 seconds long and consisted of male and female conversational speech.

Results from the VAD were compared with those from the standardised GSM voice activity detection algorithm in order to asses its performance. Several test were performed to gain statistics for the two VAD's. Firstly, the voice activity factors for both VAD's were calculated when driven by each test vector. The Voice Activity Factor (VAF) is the percentage of time that the speech is classed as active. These results are given in Figure 3.

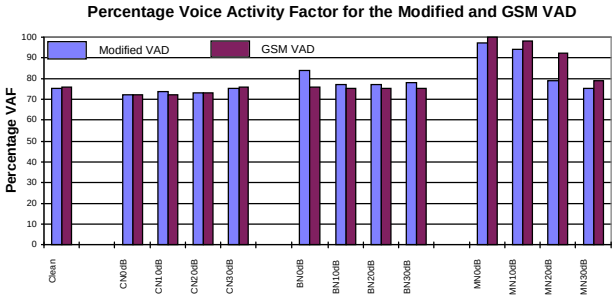

Figure 3 : Percentage of time that the modified and GSM VAD's classed the test vectors as active.

From figure 3 it can be seen that VAF results for both voice activity detection algorithms are approximately the same. Superficially, this shows that the algorithms have similar efficiency. However, the modified VAD could, hypothetically, be classifying more silence as speech while severely clipping the actual speech and still maintaining the same VAF as the GSM algorithm. In order to obtain an accurate measure of the performance of the modified VAD the four voice activity detector errors must be examined more closely.

A reference classification was produced using clean speech which was hand classified into active and inactive frames. The reference classification was then used to check the activity decision produced by both the GSM and modified VAD algorithms when driven by each test vector. By examining the total percentage of NDS and HO the efficiency of the VAD was determined, where percentages greater than those obtained for clean speech show a reduced efficiency of the VAD algorithm from the optimum. The total percentage of FEC and MSC gave an objective measure of the expected quality, however, this objective measure was not definitive and subjective tests were still required.

The combined percentage of noise detected as speech and hangover errors, i.e. the efficiency of the algorithm for the two VAD's is shown in Figure 4.

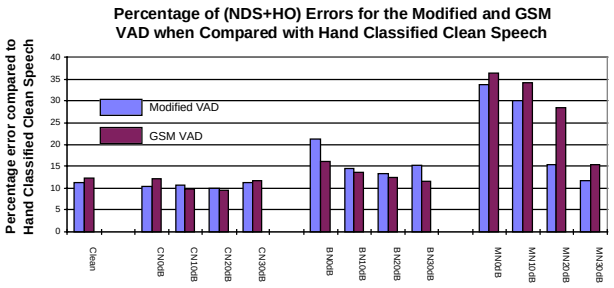

Figure 4 : VAD efficiency statistics for both the modified and GSM voice activity detectors with several different test files.

The combined percentage of front end clipping and mid speech clipping errors, i.e. a possible measure of the objective quality for the two VAD algorithms, are shown in Figure 5.

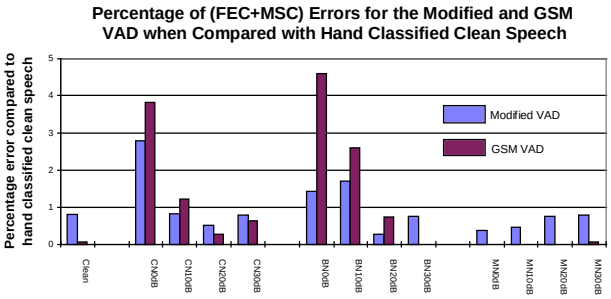

Figure 5 : Objective quality statistics for the modified and GSM voice activity detectors using several different test files.

From figure 4 it can be seen that both VAD's work well for car noise and babble noise but not so well for multiple speaker noise. This argument is strengthened by examining the data in figure 3 which shows very high voice activity factors for multiple speaker noise. The statistics for clean speech, in figure 4, mostly represent the fixed hangover error, and thus values for other files that are near this same percentage show that the algorithm worked well in discriminating between frames containing speech and those which did not. Both VAD algorithms found multiple speaker background noise the most difficult to deal with. On aggregate the two VAD's were comparable in terms of efficiency.

Figure 5 shows results which give an idea of a possible objective quality measure. The results obtained from the modified VAD, compared with those of GSM, look disappointing for multiple speaker noise and other types of noise applied at low levels to the speech. This is most evident when examining the results for clean speech. The modified VAD has approximately 0.3% (11 out of 3600 frames) of FEC and 0.5% (18 out of 3600 frames) of MSC in clean speech. Multiple MSC or FEC errors were all in areas where the speech signal was very low amplitude. In informal listening tests it was not possible to distinguish between original and VAD processed speech at the areas of FEC and MSC in clean speech. Other values of FEC+MSC around 0.8% also had the errors in the same places as the clean speech, and as a consequence were also imperceptible. In fact any value of FEC+MSC below 1.0% produced speech that had no perceptual clipping distortion.

With the addition of comfort noise injection at the decoder the modified VAD and G.729 produced good subjective quality speech for all vectors used in the test. The comfort noise was matched to average power of the noise and so no power discontinuities were perceived. At high levels of background noise a distinction between the real noise and the comfort noise was apparent but not uncomfortable to listen to. At lower levels of background noise the comfort noise blended well with the actual acoustical background noise.

5. CONCLUSIONS

Both voice activity detectors exhibited similar efficiency results for all test vectors. The subjective quality of the two VAD's was also found to be very similar, both in terms of the objective measurements of FEC + MSC and subjective listening tests.

A voice activity detection algorithm has been described that is based on the GSM VAD. By tailoring this algorithm to operate specifically with the ITU-T G.729 8kbit/s speech coder, a VAD was developed that is of comparable quality and efficiency to that of the GSM VAD while still capable of working within the 10ms coding frame length of G.729. The VAD is optimised for speech applications which are likely to be contaminated by high levels of acoustical background noise and thus its development opens up many new possibilities for newest of the ITU's speech coding standards.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of B.T. Laboratories and EPSRC in this work.

REFERENCES

[1] ITU Recommendation G.729, "Coding of speech at 8kbit/s using Conjugate-Structure Algebraic-Code-Excited-Linear-Predictive (CS-ACELP) coding".

[2] ITU Recommendation G.729 - Annex B : A Silence Compression Scheme for G.729 Optimised for Terminals Conforming to ITU-T V.70.

[3] S. D. Watson et al. "Low and Variable Bit-Rate Speech Coding for ATM Networks", EUROSPEECH'95, September 1995.

[4] CCITT Recommendation G.763, "Digital Circuit Multiplication Equipment using 32 kbit/s ADPCM and Digital Speech Interpolation".

[5] D. K. Freeman, G. Cosier, C. B. Southcott and I. Boyd, "The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service", IEEE ICASSP'89, pp. 369-372.

[6] ETSI Recommendation GSM 06.12, "Comfort noise aspects for full-rate speech traffic channels", June 1989.

[7] P. A. Barrett et al., "Speech Transmission over Digital Mobile Radio Channels", BT Technol J, 14, No. 1, pp 28-44, January 1996.