ETSI AMR-2 VAD: EVALUATION AND ULTRA LOW-RESOURCE IMPLEMENTATION

E. Cornu[1], H. Sheikhzadeh[1], R. L. Brennan[1], H. R. Abutalebi[2] , E.C. Y. Tam[1], P. Iles[1], and K. W. Wong[1]

[1]Dspfactory Ltd, Kumpf Drive, Unit 200, Waterloo, Ontario, Canada N2V 1K8
[2]Amirkabir University of Technology, Tehran, Iran

e-mail: {etienne.cornu, hsheikh, robert.brennan, reza.abutalebi}@dspfactory.com

ABSTRACT

The ETSI AMR-2 VAD is rigorously evaluated in clean and noisy conditions. The VAD is then simplified and optimized for porting to an ultra low-resource DSP system using a fast oversampled DFT filterbank. The parameters of the low-resource VAD are optimized using two speakers and 6 types of noise at SNRs from –10 to 20 dB. The VAD is then tested by employing sentences from two other speakers and 12 different types of noise. Results show that the low-resource VAD offers a performance comparable to that of the ETSI VAD in both clean and noisy conditions. When deployed on a custom DSP running at a clock speed of 1.28 MHz and consuming less than 1 milliWatt of power, the low-resource VAD uses less than 30% of the available system resources.

1. INTRODUCTION

In many speech processing systems VADs (Voice Activity Detectors) are crucial components. VADs are employed for data compression and bandwidth reduction (in codecs), for controlling the adaptation process (in adaptive echo/noise cancellation), and for noise profile estimation (in speech enhancement). A useful VAD should be able to perform well in both clean and noisy conditions. Typically VADs are required to detect all the speech activity and at the same time, detect pauses as frequently as possible. In noisy conditions however, there is a trade-off between the two requirements; as the VAD attempts to detect more portions of speech activity, it misclassifies more and more of pauses as speech (false alarm). Also, speech can be misclassified as a pause (speech cut-off). While both of the misclassification errors should be minimized, errors in one direction can be more disruptive than errors in the other direction for certain applications. For example in codecs, false alarms are more tolerable than speech cut-off, while adaptive echo cancellation is more prune to errors due to false alarms (see [1][2] for a review on VADs).

For real-time compact and portable speech processing systems, VAD requirements are more restrictive. Most notably, VADs should have low delay, fast and optimal computation, and reliable performance under marginal conditions. The European Telecommunication Standard Institute (ETSI) has released a standard on two VAD options [3] recommended for use with their Adaptive Multirate (AMR) codecs. VAD Option 2 is considered for analysis, optimization, and hardware implementation in this research since it offers a subband-based frequency-domain algorithm that can be efficiently implemented on a DFT filterbank. In this research, it is first shown that the long-term prediction flag (LTP_flag) can be removed without performance degradation. Next, the ETSI AMR-2 VAD (called ETSI-VAD in this paper) is ported to an ultra low-resource hardware platform employing an oversampled DFT filterbank. To do so, some of the VAD parameters needed re-optimization. The performance of the low-resource VAD (LR-VAD) is evaluated both objectively and subjectively by listening to segments classified as speech or pause. Six different noise types and long passages from two speakers are used to compare the performance of the LR-VAD and the ETSI-VAD and to do parameter optimization. Performance evaluations of the two VADs (using passages from two new speakers and 12 different noise types) show that the LR-VAD offers reliable performance comparable to that of ETSI-VAD.

The ETSI AMR-2 VAD is briefly described in Section 2. In Section 3 the employed DSP system is introduced. Modifications to the ETSI-VAD for porting it to the DSP system are given in Section 4. In Section 5, experimental set-up and the results of performance evaluation of the ETSI-VAD and LR-VAD are described. Finally, conclusions of the work and future steps are discussed.

2. ETSI AMR-2 VAD

Figure 1 shows a simplified block diagram of the ETSI AMR-2 VAD. The block diagram is based on the C-code of the VAD [4]. As shown, input speech signal is first converted to frequency domain. The frequency bands are then clustered as a number of channels (Nc=16) and the energies of the channels ($E_{ch}(i,m)$, $i=1,2,…,Nc$, m=frame index) are estimated. Given an estimate of the background noise ($E_n(i,m)$, $i=1,2,…,Nc$), channel SNR ($\sigma(i)$, $i=1,2,…,Nc$) is estimated. A non-linear function maps the channel SNR to a voice metric, $V(m)$. Channel SNR is also used to calculate a frame SNR and a long-term SNR ($SNR_q(m)$). The voice metric ($V(m)$) and the long-term SNR provide primary parameters for VAD decision. There is also a *hangover* mechanism in the VAD.

A separate logic decides on the noise update as follows. The ratio of the peak subband energy to the average subband energy is calculated to detect sine waves (resulting in a "sinweave_flag"). A spectral deviation estimator measures the deviation between the frame subband energies and the long-term subband energies. When the deviation (averaged over subbands,
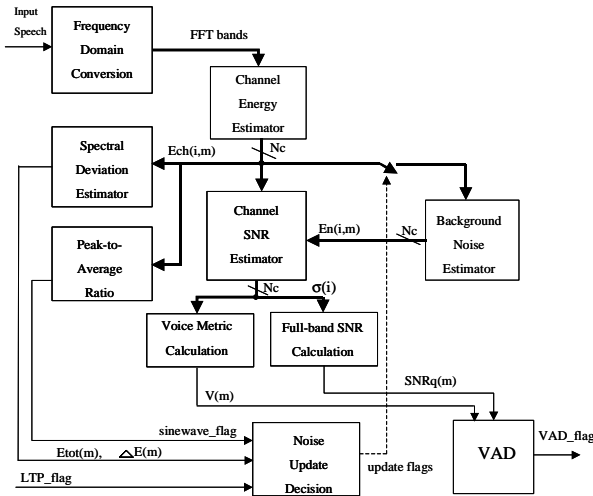
Figure 1: Simplified block diagram of the AMR-2 VAD.

$\Delta E(m)$), becomes very small, it may trigger a noise update under certain circumstances. An estimate of frame total energy ($E_{tot}(m)$) is also provided by the spectral deviation estimator. The noise energies are updated if the voice metric is less than a threshold. Otherwise, the spectral deviation, the sinewave_flag and the LTP_flag are used to decide on a "forced noise update".

3. THE DSP SYSTEM

Figure 2 shows a block diagram of the DSP system [5]. The DSP portion consists of three major components: a weighted overlap-add (WOLA) filterbank coprocessor, a 16-bit DSP core, and an input-output processor (IOP). The DSP core, WOLA coprocessor, and IOP run in parallel and communicate through shared memory. The parallel operation of the system enables the implementation of complex signal processing algorithms in low-resource environments with low system clock rates. The system is especially efficient for subband processing. The configurable WOLA coprocessor calculates the FFT coefficients leaving the core free to do the rest of the calculations of the VAD.

The core has access to two 4-kword data memory spaces, and another 12-kword memory space used for both program and data. The core provides 1 MIPS/MHz operation and has a maximum clock rate of 4 MHz at 1 volt. At 1.8 volts, 33 MHz operation is also possible. The system operates on 1 volt (i.e., from a single battery). For the LR-VAD application, the system is clocked at 1.28 MHz, consumes less than 1 mW of power, and operates with an input sampling rate of 8 kHz.

The input-output processor is responsible for management of incoming and outgoing samples. It takes as input the speech signal sampled by the 14-bit A/D converter on the analog portion of the chip at a frequency of 8 kHz. The analog portion of the chip also applies a DC-cancellation filter to the speech signal. The IOP creates frames of 200 samples, representing 25 milliseconds of speech. The frames overlap for 80 samples (10 milliseconds). The FFT calculation is launched on the WOLA coprocessor when the input-output processor indicates that 80 new samples are available for processing. The operations are exactly as specified by the ETSI standard [3]. The WOLA co-processor first applies a 200-point Hamming window and then
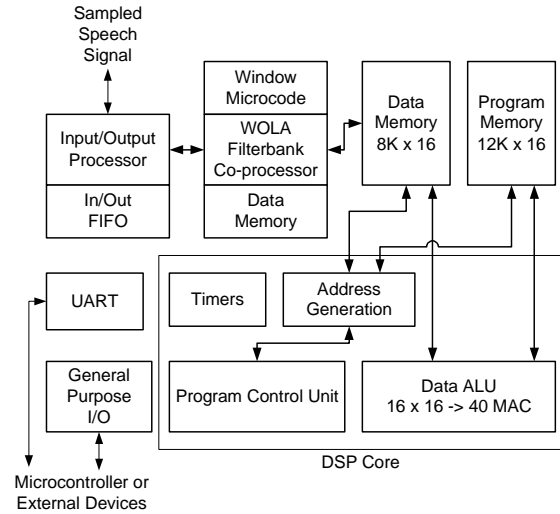

Figure 2: The DSP system block diagram

performs a zero-padded 256-point FFT. Subsequent VAD-related operations are performed at the DSP core. No data movement between the three processors is necessary at this stage because the data resides in the shared memory.

The data used in the DSP core is represented as 16 or 32-bit numbers. All 16-bit arithmetical operations are executed in a single CPU cycle. This is also the case for 32-bit add, subtract and compare operations. In a few cases, the multiplication of a 16-bit number by a 32-bit number is necessary requiring 15 CPU cycles. For SNR calculations, the on-chip math library provides a base-2 log function which uses a 32-point look-up table, executes in 9 cycles and has ± 3% accuracy.

4. MODIFICATIONS TO ETSI-VAD

To be able to port the ETSI-VAD to the hardware platform, some modifications are necessary. The goal is to simplify the VAD as much as possible, without compromising the performance. Another goal of the optimizations is to improve the performance of the VAD in noise, and to increase its stability under extreme noise conditions.

4.1. Removing the long-term prediction flag

The LTP_flag is a by-product of the AMR speech encoder and requires correlation analysis and a pitch estimate. It is the only parameter in the ETSI-VAD that cannot be obtained from a simple frequency-domain analysis. A quick review of the VAD shows that the LTP_flag is used as a safeguard in the forced-update logic of the background noise update decision. We repeated all of the evaluations (reported in Section 5) with the LTP_flag set to FALSE for all frames. Careful examination of the results showed that there was no performance degradation for clean or noisy speech. Thus, considering our resource constraints, we removed the LTP_flag and its associated computations from the LR-VAD.
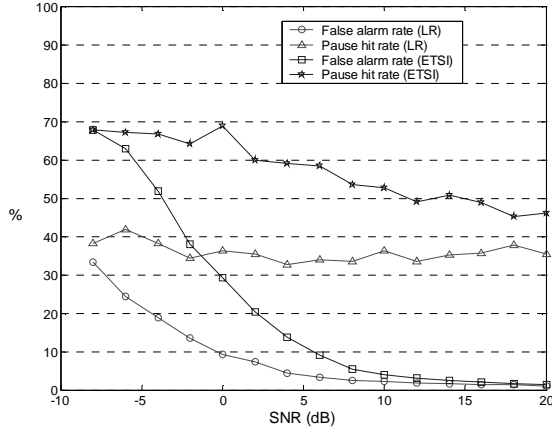
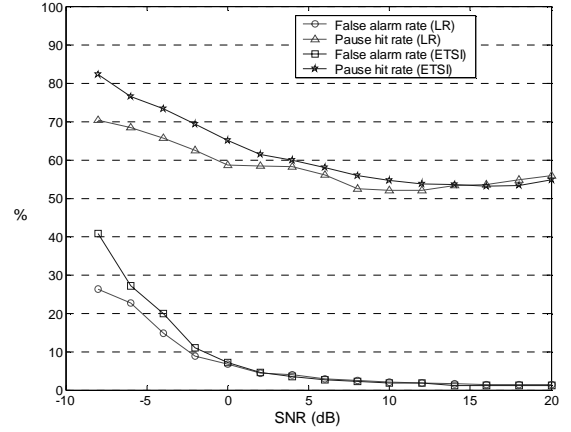Figure 3: False alarm rate and pause hit rate for LR-VAD and ETSI-VAD in babble noise.

Figure: 6 False alarm rate and pause hit rate for LR-VAD and ETSI-VAD in factory-2 noise.
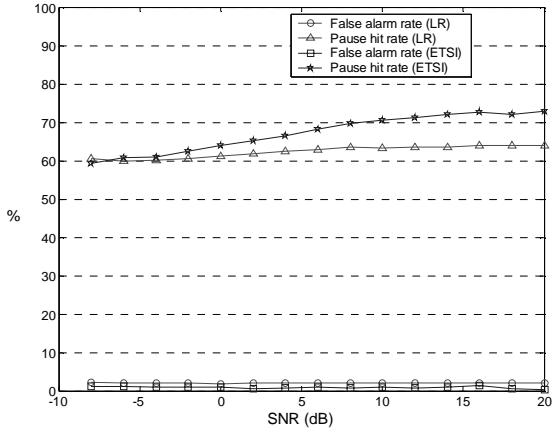
Figure 4: False alarm rate and pause hit rate for LR-VAD and ETSI-VAD in car noise.
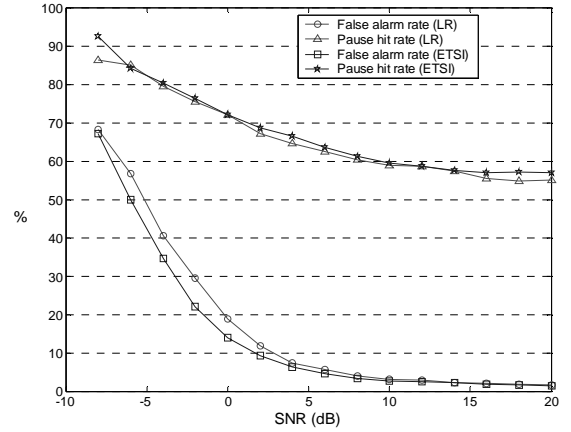
Figure 7: False alarm rate and pause hit rate for LR-VAD and ETSI-VAD in pink noise

Figure 5: False alarm rate and pause hit rate for LR-VAD and ETSI-VAD in factory-1 noise.

Table 1: FFT-Band to VAD channel mapping in the LR-VAD.

BAND	1	2	3	….	8	9	10-11	12-13	14-15
CHANNEL	0	1	2	….	7	8	9	10	11

4.2. Mapping FFT-bands to LR-VAD channels

In order to keep delay sufficiently low, a small number of bands are used on our low-resource oversampled filterbank [5]. For a typical low group-delay application at 8 kHz sampling rate, the parameter setup is: time-window length L=128 (samples), FFT length N=32, oversampling factor OS=2, and frame shift R=N/OS=16 (samples). With N/2=16 real FFT bands, we optimized the grouping of the available FFT bands to calculate channel energies for the LR-VAD. Table 1 shows that in the proposed FFT band grouping, FFT band zero is ignored and there are 12 channels in the LR-VAD. Accordingly, some minor parameter optimizations had to be applied to the VAD. For example, the sine wave flag had to be removed to avoid false alarms generated due to the wider channel bandwidths (compared to the ETSI-VAD) at low frequencies.
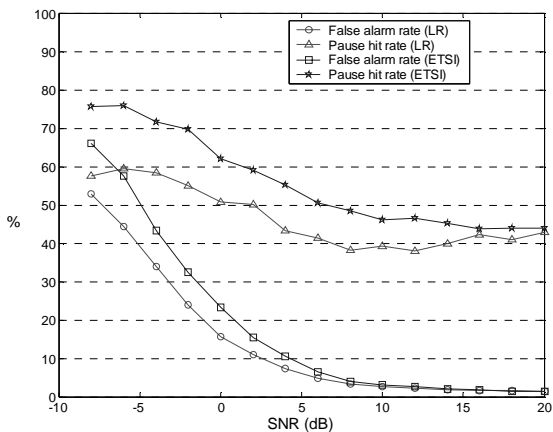
II - 587

5. PERFORMANCE EVALUATION

5.1. Methods

Readings of short stories (by a male and a female speaker) in a clean environment was recorded at a sampling rate of 22 kHz. Recordings were then down-sampled to 8 kHz. 180 seconds of recorded speech (including about 40 seconds of naturally occurring pause) from each speaker were used for LR-VAD optimization. Readings from two different (male and female) speakers were next used for VAD performance evaluations.

Six different noise types (babble, car, pink and white noises and two different factory floor noises) from the Noisex-92 database[6]-[7] were added to speech at SNRs from –10 dB to +20 dB. The ETSI-VAD was employed to detect speech activity for the recorded clean speech. The obtained VAD decisions were then used as a reference to verify all other tests.

Similar to [1], two measures of performance were defined: "false alarm rate" (defined as the fraction of all real speech frames that were erroneously detected as pause) and "pause hit rate" (defined as the fraction of all real speech pauses that were correctly detected as pauses). The VAD performance was also subjectively evaluated through extensive listening to segments classified as speech or pause at various SNRs. The results for white noise are not reported since they were very similar to those for pink noise. We used the original C-code [4] for the ETSI-VAD evaluations.

5.2. Results

The performances of the ETSI-VAD and the LR-VAD were extensively evaluated and compared. Figures 3 to 7 show the false alarm rate and the pause hit rate for the ETSI-VAD and the LR-VAD with various noises employed for one of the two test speakers. As shown, the LR-VAD has a lower false alarm rate at low SNRs, particularly for babble noise at the cost of a lower pause hit rate. This trade-off is expected and is consistent with similar results reported in the literature [1]. Careful listening to the segments classified as pause showed that for our experimental set-up, there was no voiced or high-energy speech segment misclassified as pause when the false alarm rate was 10% or less.

We further tested the ETSI-VAD and LR-VAD on noise sources not used in the optimizations. Various real-life background noises (Airport terminal, Car interior, Shopping center, Supermarket, Street, and City park) from the BBC Sound Effects Library [8] at SNRs from –10 to 20 dB were employed. The results were quite consistent with those reported in this section. Considering the (ETSI and LR) VAD performance, various noises could be classified into three major groups. Stationary narrowband noises (e.g. car noise), stationary wideband noises (e.g. pink, white and factory-2 noises) and nonstationary wideband noises (e.g. babble and factory-1 noises). The best performance (in terms of high hit-rate and low false alarm) was achieved for the first group since the effect of noise was limited to very few frequency bands. The lowest performance was obtained for the third group since wideband nonstationary noises contaminate most of the frequency bands, and noise profile may substantially change between noise model updates. Testing the VADs with another speaker and 12 various noise types led to very similar results. Generally the test results

were consistent across the optimization and test sets, each comprised of two (male and female) speakers.

6. CONCLUSIONS AND FUTURE WORK

The performance of the ETSI AMR-2 VAD is extensively evaluated in noisy environments. It was shown that the computationally expensive long-term prediction flag could be removed from the VAD without performance degradation. This simplified the VAD sufficiently for implementation on a low-resource DFT filterbank. The VAD was next ported to a low-resource hardware platform, performing very similarly to the ETSI-VAD.

Even at 1.28 MHz clock rate, the LR-VAD used less than 30% of the DSP system resources (WOLA co-processor, DSP core, and the IOP). This implied that the VAD could be integrated with other low-resource real-time subband signal processing applications. We are now in the process of employing the LR-VAD in applications such as subband adaptive filters, echo cancellation, and speech enhancement.

The LR-VAD is also being optimized to perform as a pause detector, typically needed in real-time speech enhancement. Such pause detectors typically require reliable and frequent detection of pauses in very low-SNR environments (SNRs around zero or less). The false alarm rate must be kept very low even at the expense of a low pause hit rate (see [2] for example). The subband-SNR approach employed in the ETSI-VAD offers great potential for use as a pause detector.

7. REFERENCES

[1]   Mark Marzinzik, Noise reduction schemes for digital hearing aids and their use for the hearing impaired, Ph.D. thesis, University of Oldenburg, Dec. 2000.

[2]   Mark Marzinzik and Birger Kollmeier, "Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics", *IEEE Trans. On Speech and Audio Processing, Vol. 10, No. 2, Feb. 2002,* pp. 109-118.

[3]   ETSI TS 126 094 V4.00 (2001-03) "Universal Mobile Telecommunication Systems (UMTS); Mandatory Speech Codec speech processing functions, AMR speech codec; Voice Activity Detector (VAD) (3GPP TS 26.094 version 4.0.0 Release 4)".

[4]   ETSI TS 126 073 V4.1.0 (2001-12) "Universal Mobile Telecommunication Systems (UMTS); ANSI-C code for the Adaptive Multi Rate speech codec (3GPP TS 26.073 version 4.1.0 Release 4)".

[5]   R. Brennan and T. Schneider, "A Flexible Filterbank Structure for Extensive Signal Manipulations in Digital Hearing Aids", *Proc. IEEE Int. Symp. Circuits and Systems*, pp.569-572, 1998.

[6]   Steeneken, H. J. M. and Geursten, F. W. M., "Description of the RSG.10 noise database", Technical report IZF 1988-3, TNO Institute for Perception, Soesterberg, Netherlands, 1988.

[7]   http://svr-www.eng.cam.ac.uk/comp.speech/Section1/ Data/noisex.html.

[8]   BBC Sound Effects Library, BBC Enterprises Ltd., 1991.