Second and Third Order Adaptable Threshold for VAD in VoIP

Abhijeet Sangwan*, H.S.Jamadagni#, Chiranth M.C.', Rahul Sah*, R. Venkatesha Prasad#, Vishal Gaurav'

*Dept. of Electronics & Communication, PESIT, #Center for Electronic Design Technology, IISc, Bungalore
Email: sangwan-a@yahoo.com. chiranth@symonds.net, rahulsah79@mail.com., vishalgaurav78@yahoo.co.uk,
{hsjam, vprasad}@cedt.iisc.ernet.in

Abstract

Reducing the Internet traffic is a requirement to share the bandwidth wirh many sessions. When circuit switched applications are ported on to the Internet, bandwidth saving becomes all the more prominent. An efficient adaptable threshold VAD (Voice Activity Detection) for Voice over IP systems are presented in rhispaper. The discussion includes two energy-based algorithm with higher order threshold refreshing schemes. The performance of all the schemes nrepresented, the results bring out clearly the usefulness of the schemesfor VoIP applications with saving in bandwidth.

1 Introduction

Services offered by Public Switched Telephone Networks (PSTN) systems are expensive when the distance between the calling and called subscriber is large. The current trend is to provide this service on data networks. Data networks work on the best effort delivery and resource sharing through statistical multiplexing. Hence the cost of services compared to circuit-switched networks is considerably less. However, these networks do not guarantee faithful voice transmission. Voice over IP (VoIP) systems have to ensure that voice quality does not significantly deteriorate due to network conditions such as packet-loss and delays. Providing Toll Grade Voice Quality [4] through VoIP systems remains a challenge.

In this paper we concentrate on the problem of reducing the required bandwidth for a telephone-like voice connection on Internet using Voice Activity Detection (VAD), while maintaining the voice quality. VAD is used in Voice Recognition systems, Compression and Speech coding [3,11,5] which are non real-time applications.

In VoIP systems the voice data (or payload for packet) is transmitted along with a header on a network. The header size in case of Real Time Protocol (RTP, [8]) is 12 **bytes.** The ratio of header to payload size is an important factor for selecting the payload size for a better throughput from the **network.** Lower size payload helps in a better real-time quality, but decreases the throughput. Alternately, higher size payload **gives** more throughput but performs poorly in real-time. A constant payload size representing a segment of speech is referred to as a 'Frame' in this paper. The frame size is determined by the above considerations. If a frame does not contain a voice signal it need not be transmitted. The VAD for VoIP has to determine if *a* frame contains a voiced signal. The decision by VAD algorithms for VoIP is always on a frame-by-frame basis.

In this paper, time-domain energy-based algorithms are presented with varied complexity and quality of speech. Mainly, time domain techniques are discussed with the intention of lowering the computational complexity .involved. Results obtained, and an exhaustive comparison of various algorithms with quantitative measurements of speech quality is presented. We focus on improving the performance by using higher order systems in place of the first-order systems used in [9]. There are many previous studies on VAD that dealt with energy-based algorithms such as [7]. In this paper, a procedure for choosing the scaling parameter 'p' [7] is also' given for higher order system. We restrict our study to time domain algorithms as it is faster for implementation in real time systems compared to frequency domain solutions [1].

1.1 Characteristics of Speech

Conversational speech is a sequence of contiguous segments of silence and speech [2]. VAD algorithms employ some form of speech pattern classification to differentiate between voice and silence periods. Identifying and rejecting transmission of silence periods helps reduce Internet traffic.

1.2 Desirable aspects of VAD algorithms:

- A Good Decision Rule: A physical property of speech that can be exploited to give consistent judgement in classifying segments of the signal into silence or otherwise.
- Adaptability to changing background noise: Adapting to non-stationary background noise improves robustness, especially in wireless telephony.
- Low computational complexity: Internet telephony is a real-time application. Therefore the complexity of VAD algorithm must be low to suit real-time applications.

2. Parameters for VAD Design

The differentiation of the voiced signal into speech and silence is done on the basis of speech characteristics. The signal is sliced into contiguous frames. A real-valued non-negative parameter, the average energy content, is associated with each frame. If this parameter exceeds a certain threshold, the signal frame is classified as **ACTIVE;** else it is INACTIVE. We refer to these INACTIVE frames also as noise frames.

2.1 Choice of Frame Duration

VoIP receivers may queue up incoming packets in a packet-buffer that allows them to play audio even if incoming packets are delayed due to network conditions.
Consider a VoIP system having a buffer of 3-4 packets. Having frame duration of 10ms allows the VoIP system to start playing the audio at the receiver's end after 30 to 40ms from the time the

0-7803-7488-6/02/$17.00 © 2002 IEEE.

queue started building **up.** If the frame duration is 50ms, there would be an initial delay of 150-200ms, which is unacceptable. Therefore. the frame duration must be chosen properly. Current VoIP systems use 20-40ms frame sizes.

.The specifications for,encoding speech for all VAJJ algorithms are that of Toll Grade Quality [4]:

- 8 kHz sampling frequency
- 256 levels of linear quantization (8 Bit PCM) [10]
- Single channel (mono) recording.

Advantage of using linear PCM *is* that the voice data can be transformed to any other compressed code (G711, G723, G729). Frame duration of 10ms, corresponding to 80 samples is used.

2.2 Energy *of* a Frame

Let X(i) be the i^{th} sample of speech. If the length **of** the frame were k samples, then the j^{th} frame can be represented in time domain by a sequence as

$$f_j = \{x(i)\}_{i=(j-1)k+1}^{jk} \tag{1}$$

We associate energy E_j with the j^{th} frame as

$$E_j = \frac{1}{k} \sum_{i=(j-1)k+1}^{jk} x^2(i) \tag{2}$$

where,

E_j is the energy of the j^{th} frame and

f_j is the j" frame that is under consideration

2.3 Initial Value *of* Threshold

A sample that contains only background noise is used to compute the initial threshold value parameter. For example, the initial estimate of energy is obtained by taking the mean of the energies of each frame as in

$$E_r = \frac{1}{v} \sum_{m=0}^{v} E_m \tag{3}$$

where,

E_r = initial threshold estimate,

v = number of frames in prerecorded sample.

We have taken a prerecorded sample of **5** secs, i.e., 500 frames. Alternately, it may be assumed that the first 20 frames of the given speech sample are INACTIVE and compute E_r from **(3)** taking v =20

3. VAD Algorithms

Energy of a frame is a reasonable parameter **on** the basis **of** which frames may be classified as ACTIVE or INACTIVE. The energy of ACTIVE frames is higher than that of INACTIVE frames [2]. The classification rule is,

IF $E_j > kE_r$ where k > 1 **(4)**

Frame is ACI'IVE

ELSE . Frame is INACTIVE

In this equation, E_r represents the energy of noise frames, while kE_r is the 'Threshold' being **used** in the decision-making. Having a scaling factor, k allows a safe band for the adaptation of E_r, and hence, the threshold. We use k=4 in all our studies as this gives a fair amount of margin for adaptability.

ACTIVE frames are transmitted; INACTIVE frames are not. The following algorithms use Eq **(4)** as the decision rule. The first order system *is* discussed in detail in [9], we discuss it briefly here in connection with the second and third order systems for the sake of completeness.

3.1 SED Simple Energy-Based Detector

It is now sufficient to specify the reference noise energy, ***E,*** for use in Eq **(4)** to formulate the schemes completely. Since background disturbance is non-stationary an adaptive threshold *is* more appropriate. The rule to update the threshold value can be found in [7] as,

$$E_{rnew} = (1-p)E_{rold} + pE_{silence} \tag{5}$$

Here,

E_{rnew} is the updated value of the threshold,

E_{rold} is the previous energy threshold, and

$E_{silence}$ is the energy of the most recent noise frame.

The reference E_r is updated as a convex combination of the old threshold and the current noise update. p (such that 0<p<1) is chosen considering the impulse response of Eq.(5) as a first order filter to he 0.2 [1].

Remarks

- This algorithm is simple to implement. It gave an acceptable quality of speech after compression.
- This algorithm did **not** give a good speech quality under varying background noise. This was because the threshold of Eq. (5) is incapable of keeping pace with rapidly changing background noise leading to undesirable speech clipping, especially at the beginning and end of speech bursts.

3.2 AED: Adaptive Energy-Based Detector

The sluggishness of SED is a consequence of **p** in Eq. *(5)* being insensitive to the noise statistics. We compute E_r based on second order statistics of INACTIVE frames. A buffer (linear queue) **of** the most recent '*m*' noise frames is maintained. The buffer contains the value **of $E_{silence}$** rather than the voice packet itself. Therefore the buffer is an array ***of*** m double values. Whenever a **new** noise **frame** is detected, it is added to the queue and the oldest one is removed. The variance **of** the buffer, in terms of energy is given as

$$\sigma = \text{var}[E_{silence} 1 \tag{6}$$

A change in the background noise is reckoned by comparing the energy of the new INACI'IVE frame with a statistical measure of the energies **of** the past 'm' INACTIVE frames. The variance. just before the addition of a new LNAMVE frame is denoted by σ_{old}. After the addition of the new INACTIVE frame the variance *is* σ_{new}. A sudden change in the background noise implies

$$\sigma_{new} > \sigma_{old} \tag{7}$$

$\frac{\sigma_{new}}{\sigma_{old}} \geq 1.25$	0.25
$1.25 \geq \frac{\sigma_{new}}{\sigma_{old}} \geq 1.10$	0.20
$1.10 \geq \frac{\sigma_{new}}{\sigma_{old}} \geq 1.00$	0.15
$1.00 \geq \frac{\sigma_{new}}{\sigma_{old}}$	0.10

Table 1: Value of p dependent on $\frac{\sigma_{old}}{\sigma_{new}}$

We set a new rule to vary **p** in Eq *(5)* in steps as per Table 1. As the value of p is varied the adaptation was more profound. The convex combination (Eq.5) now has its coefficients dependent on variance of energies of INACTIVE frames. We are able to make the otherwise sluggish E_r respond faster to sudden changes in the background noise. The classification rule for the signal frames continues to be Eq(4). Hence, detection of ACTIVE frames is still energy-based. The obvious limitations are (a) Inability to detect non-plosive phonemes and (b) Low **SNR** conditions caused undue clippings in the compressed signal, as in **SED** Algorithm.

3.3 System Response

The Z-Transform **of** Eq *(5)* is,

$$E_r(Z) = (1-p)Z^{-1}E_r(Z) + pE_{noise}(Z) \quad (8)$$

The Transfer Function may be determined as,

$$H(Z) = \frac{E_r(Z)}{E_{noise}(Z)} = \frac{p}{1-(1-p)Z^{-1}} \quad (9)$$

which represents a first order system. The impulse response of this first order system is given below

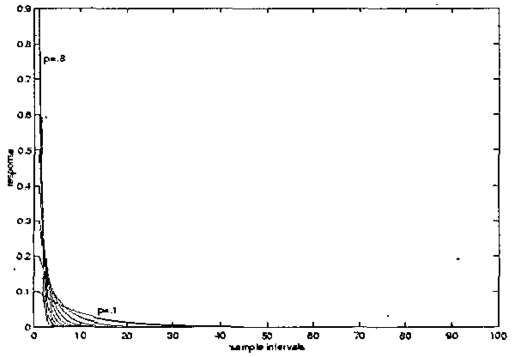


Fig. 1 : Impulse response of Eq **(9)**

In order to improve the performance, we take higher-order equation, as in **Eq** (10) second-order and Eq (11) thud order systems for updating E_r.

$$E_{rnew} = \frac{(1-p^2)E_{rold} + (1-p)E_{silence-1} + pE_{silence}}{(2-p^2)} \quad (10)$$

$$E_{rnew} = \frac{(1-p^2)E_{rold} + (1-p^2)E_{silence-2} + (1-p)E_{silence-1} + pE_{silence}}{(3-p^3-p^2)} \quad (11)$$

where, $E_{silence}$, $E_{silence-1}$ and $E_{silence-2}$ are energies **of** the last three INACTIVE frames, with E_{rnew} and E_{rold} having the same meaning as before. In **Eq.** *(5)*.

$$H(Z) = \frac{E_r(Z)}{E_{silence}(Z)} = \frac{p+(1-p)Z^{-1}}{Z-(1-p^2)Z^{-1}} \quad (12)$$

$$H(Z) = \frac{E_r(Z)}{E_{silence}(Z)} = \frac{(1-p^2)Z^{-2}+(1-p)Z^{-1}+p}{Z-(1-p^3)Z^{-1}} \quad (13)$$

The impulse response of the second and third order equation is shown in Fig. 2 and Fig. **3.**

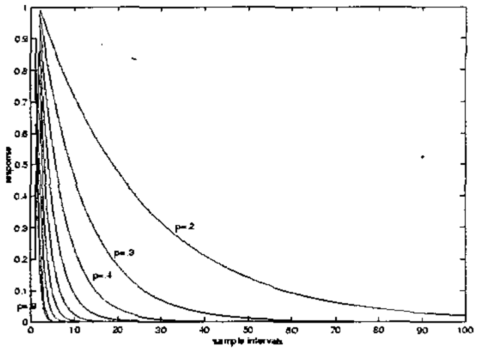


Fig. **2:** Impulse response of Eq. (10)

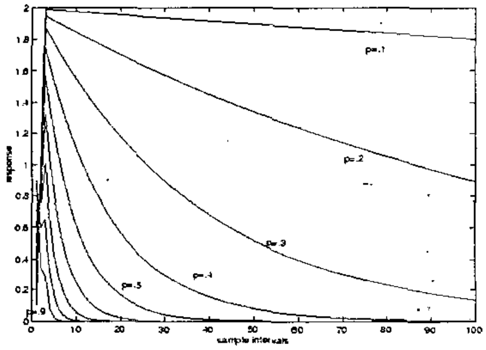


Fig **3.** Impulse response of **Eq.** (11)

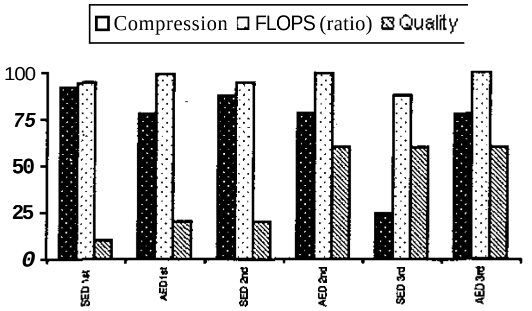
As we can see from the graphs, the first order system response is very fast, causing the threshold to vary beyond desired limits, whereas the second order system response is optimum, making it a good choice. The third order system response is sluggish in nature, making the threshold less adaptive. Incorporating these changes, we have'four new algorithms for the SED and AED.

4. Results and Discussions

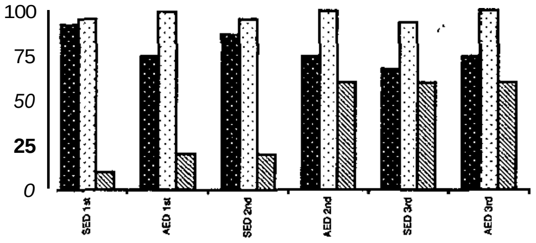
The algorithms were tested on various samples. The test templates varied in Loudness, speech continuity, background noise and accent. Both male and female voices have been used. The performance of the algorithms was studied on the basis of the following parameters:

1. Floating Point Operations (FLOPS) required: This parameter is useful in comparing algorithms of their applicability for real-time implementation.
2. Percentage compression: The ratio of total INACTIVE frames detected to the total number of frames formed expressed as a percentage. A good VAD should have high percentage compression.

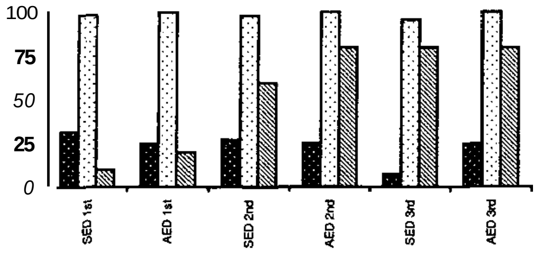
- Subjective Speech Quality: The quality of the samples was rated on a scale **of** 1 (poorest) to 5 (best) where **4** represents toll grade quality. The input signal was taken to have speech quality ***5.*** The speech samples after compression



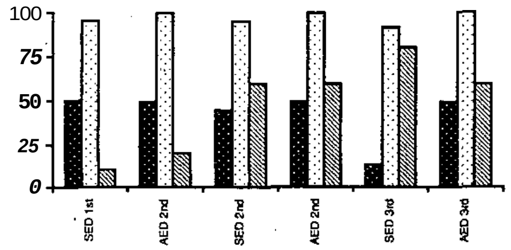
Sample 1: Dialogue (male voices)



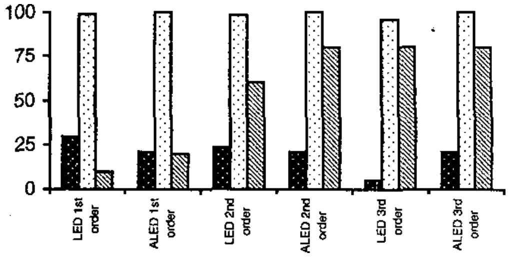
Sample 2 Monologue (male voice)



Sample 3: Monologue (female voice)



Sample 4: Dialogue (female voices)



Sample 5: Accented monologue (female voice)

5. Conclusions

VoIP has become a reality, yet not in common use. This is predominantly due to existing systems being 'not very satisfactory or dependable. A practical solution lies in efficient VAD schemes. The time domain VAD algorithms are found to be computationally less complex. With these schemes, good speech detection and noise immunity were observed. There is still a degradation of performance under low SNR conditions. Compared to first order threshold-refreshing system, second and third order performs better. The algorithms presented in this paper are found to be suitable for real-time applications, with a reasonable quality of speech. For better quality of speech, frequency domain solutions are necessary but it comes with added computational complexity.

References

[1] A. Sangwan, Chiranth M. C. R. Shah, V. Gaurav, R. Venkatesha Prasad 'Voice Activity Detection for VoIP- Time and Frequency domain Solutions". Tenth annual LEEE Symposium on Multimedia Communications and Signal Processing, Bangalore, Nov 2001, pp 20-24.

[2] B. Gold and N. Morgan, Speech and Audio Signal Processing, John Wiley Publications.

[3] Jongseo Sohn. Nam Soo Kim and Wonyong Sung, "A statistical model-based voice activity detection". IEEE signal processing letters, vol 6, no. 1, January 1999

[4] Kamilo Feher, Wireless Digital Communications, Prentice Hall India, 2001

[5] Khaled El-Maleh and Peter Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems". IEEE Canadian Conference on Electrical and Computer engineering". May 1997. pp 470-473

[6] L.R Rabiner and M.R. Sambur, An Algorithm for determining end-points of isolated utterances, Bell Technical Journal. Feb 1975, pp 297-315.

[7] Petr Pollak and Pavel Sovka, and Jan Uhlir," Noise Suppression System for a Car". proc. of the third European Conference on Speech, Communication and Technology -EUROSPEECH'93, Berlin, Sept 1993. pp 1073-1076

[8] RTP, Real Time Protocol. RFC 1889. www.ietf.org/rfc/rfc1889.txt

[9] Venkatesha Prasad. et. al. "Comparison of Voice Activity Detection Algorithms for VoIP", accepted fur publication at the Seventh IEEE Symposium on Computers and Communication, July 2002, Taormina, Italy.

[10] Xie and Reddy - Enhancing VoIP designs with PCM Coders. Communication System Design Magazine. San Francisco, California.

[11] Y.D.Cho, K.Al-Naimi and A.Kondoz, "Mixed Decision-Based Noise Adaption for Speech Enhancement". LEEE Electronics Letters Online No. 20010368, 6 Feb 2001