VAD Techniques for Real-Time Speech Transmission on the Internet

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

[*]Department of Electronics & Communication, PESIT , Email: sangwan_a@yahoo.com, chiranth@symonds.net, rahulsah79@mail.com, vishalgaurav78@yahoo.co.uk,

[#]Center for Electronic Design Technology, IISc, Bangalore
Email: {hsjams, vprasad}@cedt.iisc.ernet.in

ABSTRACT

We discuss techniques for Voice Activity Detection (VAD) for Voice over Internet Protocol (VoIP). VAD aids in reducing bandwidth requirement of a voice session thereby using bandwidth efficiently. Such a scheme would be implemented in the application Layer. Thus the VAD is independent of the lower layers in the network stack [3]. In this paper, we compare four time-domain VAD algorithms in terms of speech quality, compression level and computational complexity. A comparison of the relative merits and demerits along with the subjective quality of speech after pruning of silence periods is presented for all the algorithms. A quantitative measurement of speech quality for different algorithms is also presented.

1. Introduction

Traditional voice-based communication uses Public Switched Telephone Networks (PSTN) [3]. Such systems are expensive when the distance between the calling and called subscriber is large because of dedicated connection. The trend is shifting towards providing 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 influences such as packet-loss and delays. Providing Toll Grade Voice Quality through VoIP systems remains a challenge. In this paper we concentrate on the problem of reducing the bandwidth required for a 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 [4,14,6]. These are non real-time applications. VAD is also useful in VoIP, where the required accuracy of detection needed is less stringent.

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, [9]) is 12 bytes. The ratio of header to payload size is an important factor for selecting the payload size for better throughput. Lower size payload helps in better real-time quality, but decreases the throughput. Correspondingly, 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. 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.

The requirements of VAD algorithms for VoIP applications are:

- Lesser computational requirements (not more than one frame time)
- Toll-grade voice quality
- Saving in bandwidth to be maximized

In this paper, various VAD algorithms are presented with varied complexity and speech quality. Results obtained, and an extensive comparison of several algorithms with quantitative measurements of speech quality are presented. Improvement over a similar work [1,12] is demonstrated There are many previous studies on VAD that dealt with energy-based algorithms such as [8].

1.1 Speech Characteristics

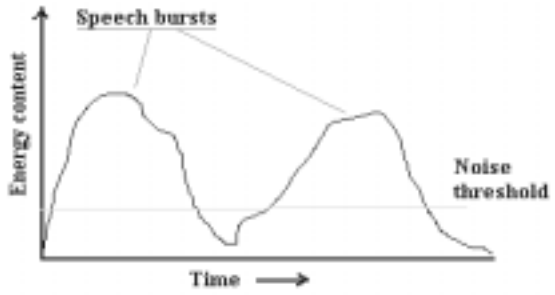

Fig. 1: A typical speech signal

Conversational speech is a sequence of contiguous segments of silence and speech (Fig. 1) [2]. VAD algorithms take recourse to some form of speech pattern classification to differentiate between voice and silence

periods. Thus, 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 and accurate judgement in classifying segments of the signal into silence or otherwise.
- *Adaptability to Background Noise:* Adapting to non-stationary background noise improves robustness, especially in wireless telephony where the user is moving.
- *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

Voiced signal is differentiated into speech or silence based on speech characteristics. The signal is sliced into contiguous frames. A real-valued non-negative parameter is associated with each frame. If this parameter exceeds a certain threshold, the signal frame is classified as ACTIVE; else INACTIVE. This threshold is updated when an INACTIVE frame is encountered, using a Periodic Noise Update (PNU). The PNU is specific to the algorithm being discussed.

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 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 VAD algorithms are that of Toll Grade Quality [5]:

- 8 kHz sampling frequency
- 256 levels of linear quantization (8 Bit PCM) [13]
- Single channel (mono) recording.

The advantage of using linear PCM is that the voice data can be transformed to any other coding (such as G711, G723, G729) for compressing the voice data packet. Frame duration of 10ms, corresponding to 80 samples is used.

2.2 Energy of a Frame

Energy of a frame indicates possible presence of voice data and is an important parameter for VAD algorithms.

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

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

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

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

2.3 Initial Value of Threshold

Two methods are proposed to set a starting value for the threshold.

Method 1: The VAD algorithm is trained for a small period by a prerecorded sample that contains only background noise. The initial threshold level for various parameters is computed from these samples. For example, the initial energy threshold is obtained by taking the mean of the energies of each sample as in

$$E_r = \frac{1}{\upsilon}\sum_{m=0}^{\upsilon} E_m \qquad (3)$$

E_r = initial threshold estimate,

υ = number of frames in prerecorded sample.

We have taken a prerecorded sample of 500 frames.

Method 2: Though similar to the previous method, here we assume that the initial 200ms of the sample does not contain any speech; i.e., these initial 20 frames are considered INACTIVE. Their mean energy is calculated as per Eq.(3). We set υ = 20.

3. Energy Based Speech Detection

The classification rule for speech is as follows [2]

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

Frame is ACTIVE

ELSE Frame is INACTIVE

Here, 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.

ACTIVE frames are transmitted; INACTIVE frames are not. The following algorithms use Eq (4) as the decision rule.

4 Exclusive Energy-Based VADs

4.1 Simple Energy Detector (SED)

It is now sufficient to specify the reference noise energy, E_r, for use in Eq (4) to formulate the schemes completely

Computation of $\boldsymbol{E_r}$

Since background disturbance is non-stationary an adaptive threshold is more appropriate. The rule to update the threshold value [8] is,

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

Here,

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

E_{rold} is the previous energy threshold, and

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

4.2 Adaptive Energy Detector (AED)

SED did not give a good speech quality under varying background noise because, the threshold, E_r of Eq. (5) is incapable of keeping pace with rapidly changing background noise. This leads to undesirable speech clipping, especially at the beginning and end of speech bursts. The reason for this sluggishness is that *p* in Eq. (5) is insensitive to noise statistics. To overcome this drawback, we make the value of *p* dependent on second-order statistics of noise frames.

A buffer (linear queue) of the most recent *'m'* noise frames is maintained. The buffer contains the value of energies rather than the voice packet itself. 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

$$\sigma = \mathrm{VAR}\left[E_{noise}\right] \tag{6}$$

In Eq. (6), unlike in Eq (5), E_{noise} represents the energies of the frames in the noise buffer.

A change in the background noise is reckoned by comparing the energy of the new INACTIVE frame with a statistical measure of the energies of the past ***m*** INACTIVE frames. Consider the instant of addition of a new INACTIVE frame to the noise-buffer. The variance, just before the addition is denoted by $\boldsymbol{\sigma_{old}}$. After the addition of the new INACTIVE frame, the variance is $\boldsymbol{\sigma_{new}}$. A sudden change in the background noise would mean

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

$\frac{\sigma_{new}}{\sigma_{old}} \geq 1.25$	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 $\frac{\sigma_{new}}{\sigma_{old}}$

Thus, we set a new rule to vary *p* in Eq (6) in steps as per Table 1. As the value of *p* is varied the adaptation was more acceptable.

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.

5 Recovery of Low Energy Phonemes

SED and AED are exclusively energy-based. Weak fricatives, such as "lo*v*", "hi*gh*", "flo*w*er" are sometimes silenced completely. It is observed that high energy voiced speech segments are detected in all VAD algorithms even under noisy conditions. However low energy unvoiced speech is commonly missed [8], reducing speech quality. The following algorithms are used to overcome this problem.

5.1 Zero Crossing Detector (ZCD)

Zero crossing for a signal is the number of times that it crosses the line of 'no disturbance' or 'zero line'. The number of zero crossings [7] for a voice signal lies in a fixed range. For example, for a 10ms frame, the number of zero crossings lies between 5 and 15. The number of zero crossings for noise is random and unpredictable. This property allows us to formulate a decision rule that is independent of energy and hence is able to detect low energy phonemes in quite a number of cases.

IF $N_{zcs}(f_j) \in \mathrm{R}$ (8)

Frame is 'ACTIVE'

Else Frame is 'INACTIVE'

$N_{zcs}(f_j)$ is the number of Zero Crosses detected in f_j.

R is the set of values {5, 6, 7,..., 15}, the number of Zero crossings for speech frames of 10ms.

This is incorporated in AED. ZCD checks the voice activity of the frames that were declared to be INACTIVE by AED. Thus, ZCD recovers almost all the low-energy speech phonemes that were otherwise silenced.

5.2 Weak Fricatives Detector (WFD) using Autocorrelation Vector Variance (AVV)

A drawback of ZCD is that it misclassifies noise frames as ACTIVE when the Zero Crossings of the noise frames satisfy Eq. (8). Such an eventuality is completely dependent on noise characteristics and may lead to failure of the algorithm. A classification feature is desired which can differentiate weak fricatives from noise independent of SNR or other noise characteristics.

One such feature that can be exploited is the high correlation found in speech signals [2,10]. A suitable measure of such characteristics is the Autocorrelation function.[5]. The *unbiased* Autocorrelation vector is given by

$$A[x] = \frac{1}{m} \sum_{m=-(N-1)}^{N-1} y[n] \times y[n-m] \tag{9}$$

where,

$\mathrm{A}[x]$ = the autocorrelation function vector.

$y[n]$ = vector under consideration

N = frame length

Each frame of incoming speech signal is divided into 20 sub-frames of four samples each. The energy of each sub-frame is computed as

$$E_{subframe} = \sum_{index=1}^{4} x^2((subframe-1)\times 4 + index) \qquad (10)$$

subframe takes values from 1 to the total number of sub-frames in the sample
index denotes each sample in a given sub-frame
Thus a vector of 20 such energy values is computed for each frame. We shall denote it as $E_{(j-1)\times 20ji\times 20}$ where *j* is the frame under consideration.

The classification parameter used in this algorithm is the variance of the above vector, the **Autocorrelation Vector Variance (AVV)** which is determined as

$$\rho_j = \mathrm{var}(E_{(j-1)\times 20:\, j\times 20}) \qquad (11)$$

A reference value of the AVV for *silence frames* is computed assuming the first 20 frames to be INACTIVE

$$\rho_{ref} = \frac{1}{20}\sum_{j=1}^{20} E_{(j-1)\times 20:\, j\times 20} \qquad (12)$$

We compare the AVV of subsequent frames with a scalar multiple of this reference value, to determine speech activity.

IF $\rho_i > k\rho_{ref}$ (13)
Frame is ACTIVE
ELSE Frame is INACTIVE

The value of *k* was set to 7 after trial and error. The reference value of AVV is not updated. Methods were designed to update the value of AVV during INACTIVE frames. However, there was no noticeable improvement in speech quality to warrant the updating of the reference AVV.
This algorithm currently has been implemented stand-alone. Work to integrate this with the energy-based VAD is underway in order to obtain better speech detection.

6. Results and Discussions

MATLAB was used to test the algorithms developed on various sample signals. The test templates used varied in loudness, speech continuity, background noise and accent. Both male and female voices have been used. The details of performance parameters and the results are as follows.

6.1 Criteria for Assessing VAD performance

Performance of the algorithms was studied on the basis of the following parameters:

- Floating Point Operations (FLOPS): The total FLOPS is calculated for all algorithms to compare their relative complexity. This parameter relates to real-time implementation of the algorithms.
- Percentage compression: The ratio of total INACTIVE frames detected to the total number of frames formed expressed as a percentage.
- 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 were played to independent jurors in a random manner for an unbiased decision.
- Percentage Misdetection: The number of frames which have speech content, but were classified as INACTIVE and number of frames without speech content but classified as ACTIVE are counted. The ratio of this count to the total number of frames in the sample explored as a percentage is **%MISDETECTION**.

An effective VAD algorithm should have high compression and a low number of FLOPS while maintaining an acceptable speech quality (and low misdetection). However, it is necessary to note that the percentage compression also depends on the speech sample being used. If the speech signal were continuous, without any breaks, it would be unreasonable to expect high compression levels.

6.2 Graphical representation of Results

Comparisons of the algorithms with respect to the above-mentioned criteria for five different speech samples (or templates) are presented. We have taken three types of templates for comparison namely, Dialogue, Monologue and Rapidly spoken Accented monologue. All data has been normalized and scaled to 100 or with respect to WFD whenever normalization can't be done. E.g., the parameter FLOPS will be always high for WFD, therefore the normalization is done with respect to WFD. Here, three standard speech templates are used for comparison of the algorithms. The results are tabulated for comparison of each algorithm with other. Each figure shows the response of all the above algorithms for a particular type of speech signal input (template).

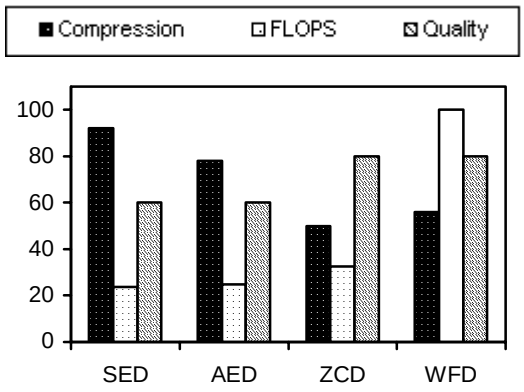

Figure 2: Dialogue (male voices)

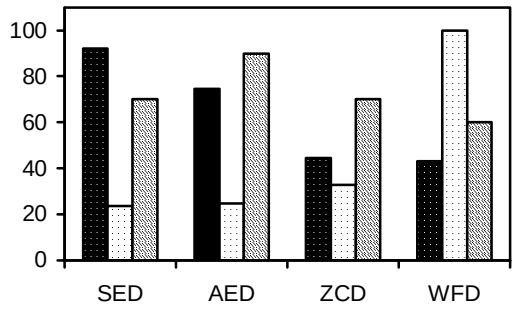

Figure 3: Discontinuous Monologue (male voice)

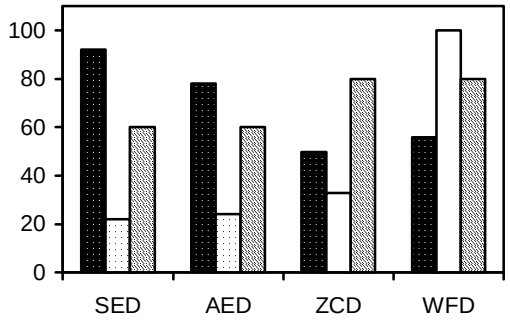

Figure 4: Rapidly spoken monologue (female voice)

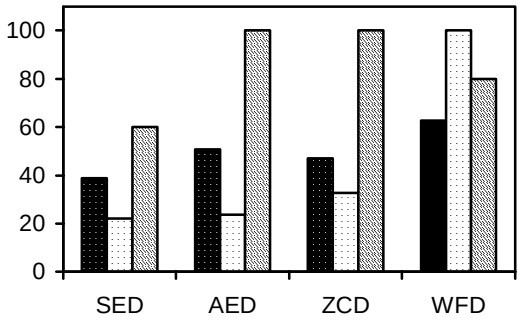

Figure 5: Dialogue (male & female voices)

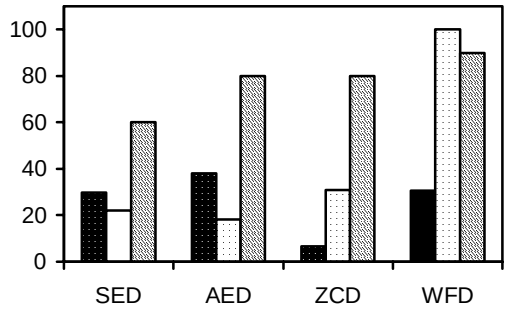

Figure 6: Accented monologue (female voice)

6.3 Trends Observed:

The following are some of the trends that were observed during the implementation and testing:

- The algorithms based solely on energy did not give an acceptable Speech Quality with all the test templates. The other techniques (Autocorrelation function and Zero Crossing Detection) gave better Speech Quality.
- The ZCD was used to recover some low energy phonemes that were rejected by the energy-based detector. However, it also picked up certain noise frames that matched the Zero Crossing criteria.
- WFD technique performed better than ZCD in detection of weak fricatives.

7. Conclusion

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. Time domain VAD algorithms are found to be computationally less complex. With these schemes, good speech detection and noise immunity were observed. The algorithms presented in this paper are found to be suitable for real-time applications, with acceptable quality of speech.

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 IEEE 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] J.E. Flood, *Telecommunications Switching - Traffic and Networks*, Prentice Hall India

[4] 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

[5] Kamilo Feher, "*Wireless Digital Communications*", Prentice Hall India, 2001

[6] 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

[7] L.R Rabiner and M.R. Sambur, "An Algorithm for determining End-points of Isolated Utterances", *Bell Technical Journal*, Feb 1975, pp 297-315.

[8] 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

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

[10] Simon Haykin -- *An Introduction to Analog & Digital Communication*, John Wiley & Sons, pp 202

[11] Tanenbaum, *Computer Networks*, Prentice Hall India, 3rd Edition

[12] V. Prasad, et. al. "Comparison of Voice Activity Detection Algorithms for VoIP", submitted to the *Seventh IEEE Symposium on Computers and Communication*, Taormina, Italy, July 2002.

[13] Xie and Reddy – "Enhancing VoIP designs with PCM Coders", *Communication System Design Magazine*, San Francisco, California.

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